A BIOINFORMATICS AND MOLECULAR ANALYSIS OF ANTIGENIC VARIATION IN AFRICAN TRYPANOSOMES

LUCIO MARCELLO

Wellcome Centre for Molecular Parasitology
Glasgow Biomedical Research Centre
University of Glasgow

Supervisor: Dave Barry

Submitted for the degree of Doctor of Philosophy
July 2006

Abstract

African trypanosomes are blood dwelling eukaryotic parasites infecting a wide range of mammals in subsaharan Africa, causing significant disease in humans and cattle. The predominant route of transmission is by the bite of an infected tsetse fly, which represents the main vector between mammalian hosts. The disease, termed African Sleeping Sickness in humans and Nagana in cattle, is often characterised by a chronic infection profile, and the chronicity is due largely to a dedicated system of immune evasion termed antigenic variation. African trypanosomes are shielded by a homogeneous glycoprotein coat whose main component is the Variant Surface Glycoprotein (VSG). This protein is highly immunogenic and targeted by host antibodies. Escape is achieved by switching to a non-cross reactive VSG isoform, by either transcriptional or recombinational mechanisms. Repeated switching allows continual escape from antibodies, and the "directionality" of this process is achieved by interaction with the immune response: VSGs have been shown to have different activation probabilities depending on their locus, so in the absence of antibodies only antigens with the highest activation probabilities would be detected. VSGs are transcribed from telomeric expression sites, and telomere-proximal silent alleles (~200) are known to be expressed early in infection: ease of sequence interaction and extensive homology results in such gene conversion events to be favoured. Most alleles (~1000) are located on megabase chromosomes, in arrays that have been recently defined to be subtelomeric. These alleles have lower activation probabilities and appear later in infection, enhancing the chronicity of the disease and therefore its transmission and persistence.

The aim of this thesis was to further our knowledge about the contribution of silent alleles on megabase chromosomes to the late stages of trypanosome infection and test the hypothesis that this contribution takes shape in a hierarchy of expression due to differences between alleles in terms both of flanking regions and coding sequence. This was achieved through a combination of bioinformatics and molecular studies. The initial approach was to undertake an extensive manual curation of the available *VSG* archive; this endeavour resulted in establishment of a fertile collaboration with the *Trypanosoma brucei* genome sequencing project (http://www.sanger.ac.uk/Projects/T_brucei/), and in creation, with the aid of P. Ward and S. Menon, of a dedicated web-based tool to handle and query curated *VSG* genes (http://www.vsgdb.org/). Out of an updated estimate of ~1600 *VSG* genes, 940 (between half and three quarters) were annotated and shown to be arranged in subtelomeric arrays and to be largely present as pseudogenes (~90%). By considering separately the

hypervariable N-terminal domain (three types, A, B and C) and the more conserved C-terminal domain (types 1 to 4, with two additional types identified in this study), it appeared that most of the degeneracy lied in the C-terminal domain. This suggested that N-terminal domains (one third of them being intact) would be utilised by a process of segmental gene conversion yielding hybrid genes, by recombination with functional C-terminal ends resident at the expression site. Under the assumption that "order" within the genome (the presence of patterns within the *VSG* archive) helps inform "order" in *VSG* expression (a hierarchy based on different activation probabilities), it was somewhat surprising to detect little evidence of clear substructuring within the archive: no "classes" of *VSG*s could be identified, based on coding sequence and flanking sequence features. In keeping with the observed high level of divergence within the *VSG* archive, clear orthologue groups (here defined as alleles sharing >60% amino acid sequence identity) were found not to include more than three to four members and to be scattered at random across the arrays. Putative functional genes could not be separated into groups based on expected differences in activation probabilities, such as a different number of upstream 70-bp repeats, shown to be involved in copying silent alleles to the expression site.

Due to the lack of order found in the *VSG* archive, the "simple" hypothesis that led to the planning of an *in vivo* experiment was that mosaic gene formation would give rise to a significant fraction of late appearing variants. A chronic mouse infection involving eleven animals was conducted, and this allowed gDNA and RNA from different timepoints to be obtained, between day 3 and day 28 (each mouse was terminally sampled). Expressed *VSG* genes were amplified from cDNA, then cloned and sequenced; the sequences obtained were then queried against the *T. brucei* genome database (http://www.genedb.org/genedb/tryp/index.jsp). Indeed, it appeared that mosaic genes represented a high proportion (five out of eight) of variants detected in the two infections at day 28. Two of the putative mosaics were confirmed experimentally by PCR. In addition, a significant number of point mutations were observed between putative donor and cDNA sequences and found to correlate with regions of reduced protein secondary structure potential, possibly representing surface loops and epitopes. Mutations were more numerous later in the infection, showing a timing of appearance similar to that of mosaic genes. These novel findings pose two major questions as to the kinetics of late stage antigenic variation: firstly, to what extent is a hierarchy present late in infection, and to what extent order is abandoned, allowing a different infection profile to emerge. Secondly, whether further variation within expressed mosaic genes and accumulation of point mutations contribute to the process of antigenic variation or are important only when considering long-term archive diversification. Detailed discussion of the data leads to the

proposal that a mutator phenotype might be selected for late in infection, as this might increase the frequency of these heretofore underestimated sequence variation mechanisms.

An additional and unexpected finding was the detection of the novel gene family of *VSG*-related genes. Bioinformatic analysis showed that its ~30 alleles are genetically and physically separate from *T. brucei VSG*s. Although N-terminally they resemble very closely type B *VSG* N-terminal domains, their C-terminal end is divergent, bearing closer similarity to that of *T. congolense VSG*s. Their expression was shown to be unlinked to the monoallelic expression of *VSG*s, and it appears that several members of this family are expressed at the same time, in all likelihood from their own loci. The presence of a signal peptide and the potential for GPI anchoring (at least in some of the alleles) suggests that *VSG*-related genes might encode surface receptors, adding to the diversity of the *VSG* superfamily. This already includes, apart from *VSG*s themselves, the receptor for host transferrin (*ESAG6* and *ESAG7* genes) and a gene associated with human serum resistance in a subset of human infective *T. brucei* strains (*SRA* gene). Whether a similarly important role can be ascribed to *VSG*-related genes remains to be established.

Table of Contents

List of Tables

List of Figures

Acknowledgements

My thanks go to:

Dave, for true guidance and inspiration, allowing me the freedom to find my own way

Peter, Liam and Richard, for showing me how to joyously push the boundaries of science

All lab members, for reminding me at various points where the laboratory was located

Mel, Peter and Martyn for opening beautiful unexpected musical avenues

Rui for our lengthy Chinese conversations

The Wellcome Trust for funding

Rita the Saint of the Impossible for luxurious accomodation

Glasgow friends and Glasgow people for the constant good times

Giovanni, Neil, Kim, Sambayabamba and Mão No Chão for magic times

My parents and sister and cousin and all the family for keeping it real

Author's Declaration

The bioinformatics annotation of *VSG* genes presented in Chapter 3 was conducted in collaboration with Christiane Hertz-Fowler, Hubert Renauld and Matt Berriman at the Sanger Institute (Cambridge) and with Dr. Mark Carrington at University of Cambridge.

The *VSG* database VSGdb was developed together with bioinformaticians Pauline Ward and Suraj Menon.

I declare that this thesis and the results presented in it are entirely my own work, except where otherwise stated.

Lucio Marcello

CHAPTER 1

INTRODUCTION

1. Introduction

1.1 African trypanosomes

"The trypanosomes of man have neither abandoned their evolutionary cradle nor reached their ultimate parasitological destination; they still pursue the protracted odyssey which a parasite must follow during the hazardous traverse from an ancient host to a new one, and the consequences of human infection betray the insecurity of an undecided relationship" (Duggan, 1970)

African trypanosomes are unicellular eukaryotes constituting the Salivarian branch of trypanosomes, a monophyletic group branching from the order kinetoplastida[1] 400 mya (Simpson *et al.*, 2006). One of their distinguishing features is the evolution of a complex form of immune evasion termed antigenic variation, which enables them to sequentially activate different non cross-reactive copies of their surface antigen (VSG, Variant Surface Glycoprotein), thereby establishing a chronic infection in their vertebrate hosts, causing an often fatal disease that in humans is termed African Sleeping Sickness and in cattle is termed Nagana. Many different aspects of antigenic variation have been studied. Broadly speaking, there are biological studies, in which the course of an infection is monitored and the growth patterns of trypanosome variants are studied in relation to the immune response. There are also biochemical and molecular studies concerning VSG protein structure and organisation of *VSG* genes, their transcription and the mechanisms of switching from one gene copy to another (Turner, 1999). All these studies are interrelated and overlapping, with molecular and biochemical results feeding into the overall understanding of infection, and general biological considerations providing the framework for testing hypotheses at the molecular level.

Firstly, the species and life cycle of African trypanosomes will be presented, followed by general features of genome organisation, replication, transcription and translation. Secondly, an overview of the cell surface of these parasites will be given, leading to the VSG system, looking at the *VSG* coding sequence, its diversity and its features at the protein level. Thirdly, the mechanisms by which *VSG*s are expressed and expressed *VSG*s are replaced by other *VSG* copies will be considered, with the aim of conveying the current

[1] Kinetoplastids are defined as flagellated protozoa with an unusually organised mitochondrial genome, termed kinetoplast or kDNA (Lukes *et al.*, 2005). Within this group other important organisms are *Trypanosoma cruzi* and *Leishmania* species.

understanding of antigenic variation at the molecular level. Lastly, the broader issue of *VSG* divergence and evolution will be addressed, and antigenic variation strategies of other parasites will be considered, to put African trypanosomes into perspective.

1.1.1 Species

There are three main species of African trypanosome: *Trypanosoma brucei*, *Trypanosoma congolense* and *Trypanosoma vivax*; most research has been conducted on *T. brucei*, as it is the only one to infect humans and grows more easily than the others in laboratory conditions. The remaining two are important livestock parasites, *T. congolense* being the most widespread and *T. vivax* being the most pathogenic. While keeping *T. brucei* as the main focus, information on antigenic variation will be given, where appropriate and if available, on studies conducted on *T. vivax* and *T. congolense*. Table 1.1 gives a brief summary of the species within the Salivaria.

Table 1.1: African trypanosomes (Hoare, 1970).

SPECIES	VERTEBRATE HOST	VECTOR /TRANSMISSION	DISTRIBUTION
A) *Duttonella*		Development only in proboscis[2]	
T. vivax	Cattle, sheep, goats, equines, camels	Tsetse fly (Africa); tabanid flies (South America)	Africa, South America
B) *Nannomonas*		Development in midgut and proboscis[2]	
T. congolense	Cattle, sheep, goats	Tsetse fly	Equatorial Africa
T. simiae	Pigs	Tsetse fly	Central and East Africa
C) *Trypanozoon*		Development in midgut and salivary glands[2]	
T. brucei	Equines, sheep, goats, humans	Tsetse fly	Equatorial Africa
T. evansi[3]	Equines, camels	Tabanid flies (mechanical transmission)	Africa, Middle East, Asia and Latin America
T. equiperdum	Equines	Transmitted by venereal contact	Africa, Middle East

[2] This refers only to tse tse infections, mechanical transmission by tabanid flies does not involve any development of parasites in the vector (see Section 1.1.2).

[3] *T. evansi* is thought to have evolved recently from *T. brucei* on the northern extremity of the tsetse belt following infection of camels by tsetse, then to have been propagated by mechanical transmission by tabanid flies. It has spread to the Middle East, Asia and Latin America.

Within the subgenus *Trypanozoon* are *T. brucei* together with its close relatives *Trypanosoma evansi* and *Trypanosoma equiperdum*, which have separated by developing alternative modes of transmission, allowing them to spread beyond the tsetse equatorial habitat (see Table 1.1). *T. brucei* has been classified into three subspecies, *T. brucei brucei*, *T. brucei gambiense* and *T. brucei rhodesiense*. The differences between subspecies are thought to be relatively small: the human-infective subspecies can be identified because, unlike *T. brucei brucei*, they are resistant to lysis *in vitro* by human serum. The distinction has been found to be not so clear-cut, as together with sensitive and resistant *T. brucei*, a third groups of strains with an intermediate phenotype has recently been characterised: this latter group, of which strain *T. brucei* 927 is an example, shows slower kinetics of lysis, compatible with growth in the presence of human serum (at least *in vitro*), and, unlike sensitive strains, has the ability to develop resistance upon prolonged exposure to human serum (Turner *et al.*, 2004). A further complication in defining these subspecies is due to the fact that genetic exchange can readily occur between them (see section 1.1.2), which in turn could cause spread and modification of serum resistance traits. The extent as to which this happens in the wild has not been ascertained, as population studies have so far established the pattern of exchange within, rather than between subspecies (A. MacLeod, pers. comm.). The three phenotypes detected with regards to lysis in the presence of human serum have been ascribed to the action of a limited number of genes (Turner *et al.*, 2004), one of which, *SRA* (Serum Resistance-Associated gene), present in *T. rhodesiense*, has been extensively characterised (Oli *et al.*, 2006). Basic features of the three *T. brucei* subspecies are given in Table 1.2.

Table 1.2: *T. brucei* species (Hoare, 1970).

Species	VERTEBRATE HOST/ VECTOR	DISTRIBUTION	COMMENTS
T. brucei brucei	Cattle, pigs, horses, sheep, goats; game animals / *Glossina morsitans* (main vector)	All equatorial Africa where tsetse flies are present	Lack of human infectivity is main distinguishing trait[4]
T. brucei rhodesiense	Humans, reservoir host overlapping with *T. brucei brucei* / *Glossina morsitans* (dry fly)	First appeared in Zambia; spread across East Africa (savanna)	Anthropozoonosis; Causes acute Sleeping Sickness
T. brucei gambiense	Humans; non-human reservoir secondary / *Glossina palpalis* (wet fly)	West and central Africa (rivers)	Anthroponosis; causes chronic Sleeping Sickness

[4] See text above and (Turner *et al.*, 2004) for caveat on this statement.

1.1.2 Life cycle

Trypanosomes are transmitted to their vertebrate host by the bite of an infected tsetse fly, injected below the skin tissue as the fly takes a blood meal. The developmental stage of the initial infecting trypanosome population (10-10000 trypanosomes) is termed metacyclic: it develops in the salivary glands of the tsetse and is already expressing a VSG coat (Tetley *et al.*, 1987) (see Figure 1.1 for a life cycle outline). Non-dividing metacyclic trypanosomes evolve into what has been described as the long slender bloodstream form (BSF), the growth of which results in peaks of parasitaemia corresponding to antigenically distinct VATs[5] emerging (Capbern *et al.*, 1977). The mechanics of the process will be explained later in this chapter, for now suffice it to say that at each peak of growth a density-dependent differentiation into the cell-cycle-arrested short stumpy BSF occurs, the latter being responsible for infecting the tsetse fly (Gruszynski *et al.*, 2006). As demonstrated by laboratory studies, only a small proportion of flies becomes infected (*T. brucei* has the lowest fly infection rate amongst the Salivaria, *T. congolense* and *T. vivax* showing a higher adaptation to the insect vector), and trypanosomes undergo several population bottlenecks as they develop from the midgut of the fly to the salivary glands, over the course of at least two weeks (Gibson and Bailey, 2003). Both infection rate and length of cycle have been shown to be temperature dependent, higher physiological temperatures (up to 37°C) resulting in a higher infection rate and a faster cycle through the fly (Hoare, 1970). In the midgut of the fly, the trypanosome becomes the procyclic form (PF), with a coat presumably more suitable to the new environment, made of a small set of proteins called procyclins (Liniger *et al.*, 2004). The procyclic trypanosome then undergoes a process of maturation, involving further rounds of differentiation, as it makes its way to the salivary glands (Urwyler *et al.*, 2005), to the epithelium of which it becomes attached. Attachment is accompanied by development into the epimastigote stage, a replicating stage that gives rise to the next metacyclic trypanosome population, leading to completion of the life cycle (see Figure 1.1).

[5] Variable antigen type (VAT) is referred to as a population of trypanosomes expressing the same VSG at a given time. Different subpopulations express different *VSG*s during the course of infection. MVAT is a population of metacyclic trypanosomes expressing a given *VSG*. A BVAT is the bloodstream equivalent.

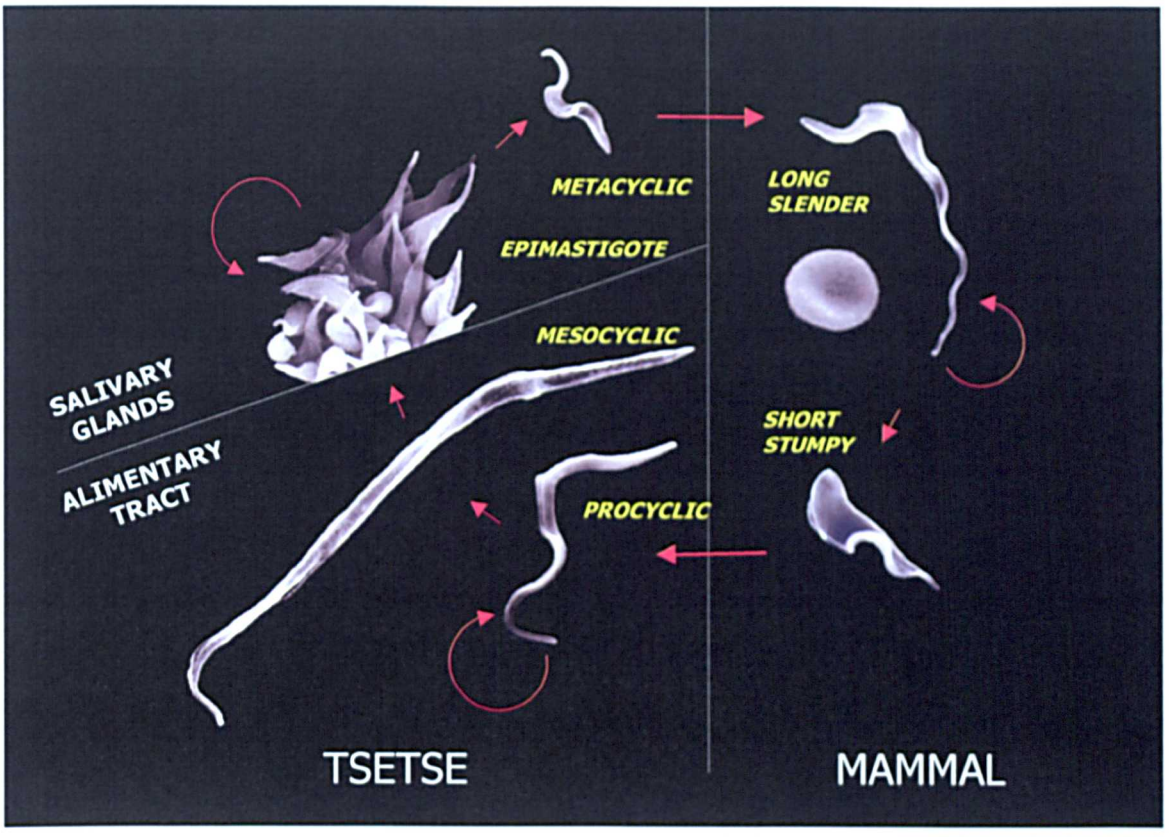


Figure 1.1: Life cycle of *T. brucei* (Barry and McCulloch, 2001).

Maturation through the fly is thought to amplify the typically low parasitaemia found in the mammalian host, ensuring infection of one or more hosts by a single fly, when compared with simple mechanical transmission by contamination of the mouthparts with infected blood (Gibson and Bailey, 2003). In addition, different *T. brucei* strains have the opportunity for sexual reproduction in the fly (possibly at the epimastigote stage), an added diversification tool for the parasite (Schweizer *et al.*, 1988). Some laboratory strains have lost the ability to be transmitted through the fly and have been termed monomorphic (only the long slender bloodstream form is present in the vertebrate stage), whereas trypanosome strains that are fully competent in completing the life cycle have been termed pleomorphic (in the bloodstream form, the full range of forms, from long slender to short stumpy, can be detected). Other differences between these two types of strain have proven to be important in relation to antigenic variation, as will be discussed in section 1.4.

1.1.3 Genome organisation

The diploid genome of *T. brucei* strain 927/4 (haploid size 26 Mb) has been shown to contain 9068 predicted genes, including ~900 pseudogenes and ~1700 *T. brucei*–specific genes, the remainder being shared with the other sequenced kinetoplastids, *Leishmania* and *Trypanosoma cruzi* (Berriman *et al.*, 2005). The genes of *T. brucei* are tightly packed and oriented unidirectionally over long regions, as it is probable that they are transcribed polycistronically. These polycistrons are found to be very conserved across kinetoplastids, highlighting the fact that key housekeeping functions have been retained. In addition, regions of discontinuity appear at strand switches between polycistrons, suggesting in the case of *T. brucei* that the current chromosomes have originated from several chromosome fusion events (Ghedin *et al.*, 2004). Early analysis of cosmid clones suggested clustering of *VSG*s in arrays in the genome (Van der Ploeg *et al.*, 1982b): this has been confirmed bioinformatically, and the location of these clusters has been shown to be predominantly subtelomeric (see Chapter 3).

The diploidy of *T. brucei* is not typical, as it has been demonstrated so far only for the 11 megabase (1-6 Mb) chromosomes and not for the several (between one and five[6]) intermediate (200-900 kb) chromosomes and around 100 minichromosomes (50-150 kb) (Ersfeld *et al.*, 1999); triploidy occasionally arises as a result of mating in the fly, but is considered to be rare and to occur mostly between subspecies (Gibson and Stevens, 1999; Hope *et al.*, 1999). Furthermore, *VSG* genes and *VSG* expression sites appear to be haploid, as often only single copies of *VSG* genes are recovered (Morrison *et al.*, 2005); haploidy has been shown for the expression site (*VSG* transcription unit, see section 1.3.1) on chromosome (chr) 1 (El Sayed *et al.*, 2000) and for the *VSG* arrays on chr 8 (Berriman *et al.*, 2005), chr 5 and chr 11 (Berriman, pers. comm.). Intermediate chromosomes and minichromosomes have been shown to harbour only *VSG*s and *VSG* expression sites (which contain also other genes, see section 1.3.1) and to be present only in small numbers in *T. vivax,* suggesting a potential expansion of the *VSG* archive in *T. brucei* and *T. congolense* (Barry and McCulloch, 2001; Wickstead *et al.*, 2004).

Strains of *T. brucei* differ by as much as 25% in their haploid DNA content (compared with the genome strain) (El Sayed *et al.*, 2000) and within each strain there are large size differences between pairs of homologues (up to fourfold, assessed by pulsed field gel electrophoresis) (Gottesdiener *et al.*, 1990). Whereas comparison of chr 1a of TREU 927/4

[6] two in the 927/4 genome strain (Horn and Barry, 2005).

with chr 1 of STIB 247 showed that differences were also due to the presence of variation in the core region of the chromosome (source unknown, but possibly involving tandemly repeated genes[7] and variation at strand switch regions), it appears that differences between chr 1a and the homologue 1b were largely due to differences in the subtelomeric repetitive region (Melville *et al.*, 1999). This subtelomeric repetitive region had been associated with the retrotransposon *INGI* and can now be safely assumed to correspond to *VSG* arrays, as revealed by the genome project (further considerations on this matter will be made in Chapter 3, and section 3.2.1 in particular).

1.1.4 Transcription and translation

In all kinetoplastids studied, transcription is polycistronic, with up to 90 genes in a single transcription unit in the case of *T. brucei*; these polycistrons are composed of genes that appear largely unrelated in function (El Sayed *et al.*, 2003). It has been reported that genes within the same polycistron can differ in the levels of transcript they yield and in the developmental stage in which they are expressed (Belli, 2000). If each gene has an upper limit of transcript abundance dependent on basal polycistronic transcription, genes encoding abundant proteins are likely to be present in multiple copies, and this is indeed the case for the tubulin genes, for which an array of ~15 copies is present. Some tandem duplications have also led to diverged copies of the original genes, possibly when the constraints of high expression were not great; the current estimate is that unique genes represent around 50% of the total number, this figure broadly reflecting the common occurrence (and importance) of duplications (H. Renauld, pers. comm.).

In order to produce mature mRNA from the polycistronic transcript, each gene, as transcribed, is individually processed with a 39 nt leader sequence (miniexon) spliced in trans to the 5' end. This spliced leader sequence supplies the mRNA cap and its acceptor site is the first AG dinucleotide following an 8-25 nt polypyrimidine tract. Polyadenylation occurs as a coupled reaction, dependent on trans-splicing of the downstream gene, as no "autonomous" consensus signal for polyadenylation is to be found at the 3' end of the nascent transcript (Benz *et al.*, 2005). Trans-splicing results in what has been termed discontinuous transcription and has been found to occur also in nematodes and several other protozoa and lower chordates, but in kinetoplastids alone it is the only system available for gene expression (Das and Bellofatto, 2003).

---

[7] Polymorphism in copy number has been shown for several genes, including those encoding tubulin, actin and the glucose transporter THT-1 (Barrett *et al.*, 1996).

Gene regulation therefore seems to occur mostly at various post-transcriptional levels, with mechanisms possibly including mRNA processing (trans-splicing and polyadenylation), nucleocytoplasmic export, mRNA stability, antisense mRNA, translational control and post translational turnover of proteins[8] (D'Orso *et al.*, 2003). It has been speculated that the 3' UTR of each gene would be the most important cis element allowing developmental regulation of transcribed genes, based on studies conducted on EP-procyclin 3' UTR (Matthews *et al.*, 2004). As for trans-acting factors involved in post-transcriptional regulation, to date (literature search updated to 2006), only proteins involved in trans-splicing and mRNA turnover have been identified (D'Orso *et al.*, 2003).

Eukaryotic transcription is mediated by three types of polymerase (pol), of which polI generates rRNA, polII yields mRNA and some small nuclear RNA, and polIII synthesizes tRNA and some other small nuclear RNA. In *T. brucei* there is a lack of recognisable promoter elements for polII upstream of individual polycistrons, the only polII promoters found being those associated with the 39 nt spliced leader sequence (Palenchar *et al.*, 2006). PolII appears to be the sole polymerase responsible for constitutive transcription of polycistrons, as knockdown of one of its subunits resulted in transcription being restricted to the nucleolus, where polI operates (Devaux *et al.*, 2006). Another unusual feature of transcription in *T. brucei* is that the stage-regulated procyclic and bloodstream surface coat proteins (procyclins and *VSG*s) are transcribed from expression sites that are driven by polI. This is the only example of polI transcribing mRNA, possibly to allow high levels of transcription for these very abundant proteins (Gunzl *et al.*, 2003): *VSG* mRNA levels are comparable to those of the tubulin gene, which is polII-transcribed but is present in ~15 copies (Kooter and Borst, 1984).

An important aspect when considering transcription is also the level of compaction of the DNA substrate into organised chromatin. *T. brucei* appears not to have as wide a range of chromatin states as do other organisms. This possibly is due to the linker histone H1 protein being shorter than usual in trypanosomes, resulting in highly condensed chromosomes not being visible at mitosis (Horn, 2001). A role in chromatin condensation has been nevertheless ascribed to histone H1. The fact that chromatin appears to be more condensed in the bloodstream form than in the procyclic form positively correlates with a higher level of expression of histone H1 in the bloodstream form (Belli, 2000). Chromatin

[8] Transcriptional elongation has been shown to be an important stage of gene regulation in the case of *VSG* expression sites (see section 1.4.1), but it is not known whether it could be of importance for housekeeping genes.

is likely to play a role in the regulation of cell surface protein genes that are developmentally regulated at the level of transcription (see section1.4.1).

1.2 Antigenic variation: the *VSG* gene and the trypanosome cell surface

Having presented the core features of African trypanosomes, and before moving on to the mechanisms by which they conduct antigenic variation, it is important to discuss more closely the features of the antigen, the Variant Surface Glycoprotein, or VSG.

1.2.1 VSG *archive*

Considering the full potential of immune evasion of African trypanosomes leads to the question of how many *VSG* gene silent copies are present: this silent potential will be referred to as the *VSG* archive. Up to ~200 *VSG*s are expected to be located at telomeres in all chromosome classes (especially in minichromosomes), but the majority of them are to be found in long arrays located subtelomerically in the megabase chromosomes, accounting for ~20% of the whole genome sequence (Berriman *et al.*, 2005). The number of *VSG*s had previously been estimated to be around 1000 (Van der Ploeg *et al.*, 1982b), but is now thought to vary greatly depending on the trypanosome strain, due to the plasticity of *T. brucei* subtelomeres (see section 1.1.3 and 3.2.1 for more details). The archive composition of different strains can vary significantly, as only a few common *VSG*s have been found (Bernards *et al.*, 1986); the precise degree of overlap between *T. brucei* strains is currently unknown, and a larger set of genes would have to be considered and compared in order to draw more definite conclusions (see Chapter 5, section 5.3.1 for an additional discussion on this matter).

1.2.2 VSG *structure and function*

VSG domains, post-translational modifications and folding will now be considered, leading to a description of the trypanosome cell surface. The information below and in the rest of this chapter attempts to summarise the vast amount of data generated since the early 1980s, dealing with the molecular mechanisms of antigenic variation and, directly or indirectly, with VSG structure. It includes a survey of full-length *VSG* sequences deposited in the online databases (ncbi, http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi?itool=toolbar; expasy, http://us.expasy.org/, January 2006), which yielded 97 distinct variants, all but one

derived from mRNA transcripts. Eighty-six are from the *Trypanozoon* subgenus (*T. brucei, evansi, equiperdum*), 10 are from *T. congolense* and one is from *T. vivax*. A table (Table 7.1) with all sequences listed with accession number and key reference is provided in the Appendix, and Figure 1.2 summarises the primary sequence features and diversity of these sequences.

T. brucei VSGs are 400-500 amino acids (aa) long and are composed of an N-terminal (350-400 aa) and a C-terminal (100-150 aa) domain (see Figure 1.2). The domains are separated by a "hinge" region that is sensitive to proteolysis (Johnson and Cross, 1979), recently defined to be ~15 amino acids long (Chattopadhyay et al., 2005). There are three N-terminal domain types (A-C) and four C-terminal domain types (1-4) (Carrington *et al.*, 1991) and these can be identified by the pattern of cysteines along the length of the protein, forming an N- and a C-terminal cluster. A notable difference between the two domains is the level of sequence conservation, which is much higher in the C-terminal domain (40% identity between domains is common), while the N-terminal domain tends to be hypervariable (17-22% identity) (Rice-Ficht et al., 1981). *T. congolense* and *T. vivax* VSGs are shorter (between 350 and 410 aa long), mainly due to the absence of a defined C-terminal domain, and all known VSGs of these trypanosome species have homology only with the T. brucei type B N-terminal domain. The C-terminal end of *T. congolense* VSGs has been found to be proline-rich (Rausch et al., 1994), possibly resulting in formation of a rigid elongated structure that might increase the height of these shorter VSGs (Carrington and Boothroyd, 1996).

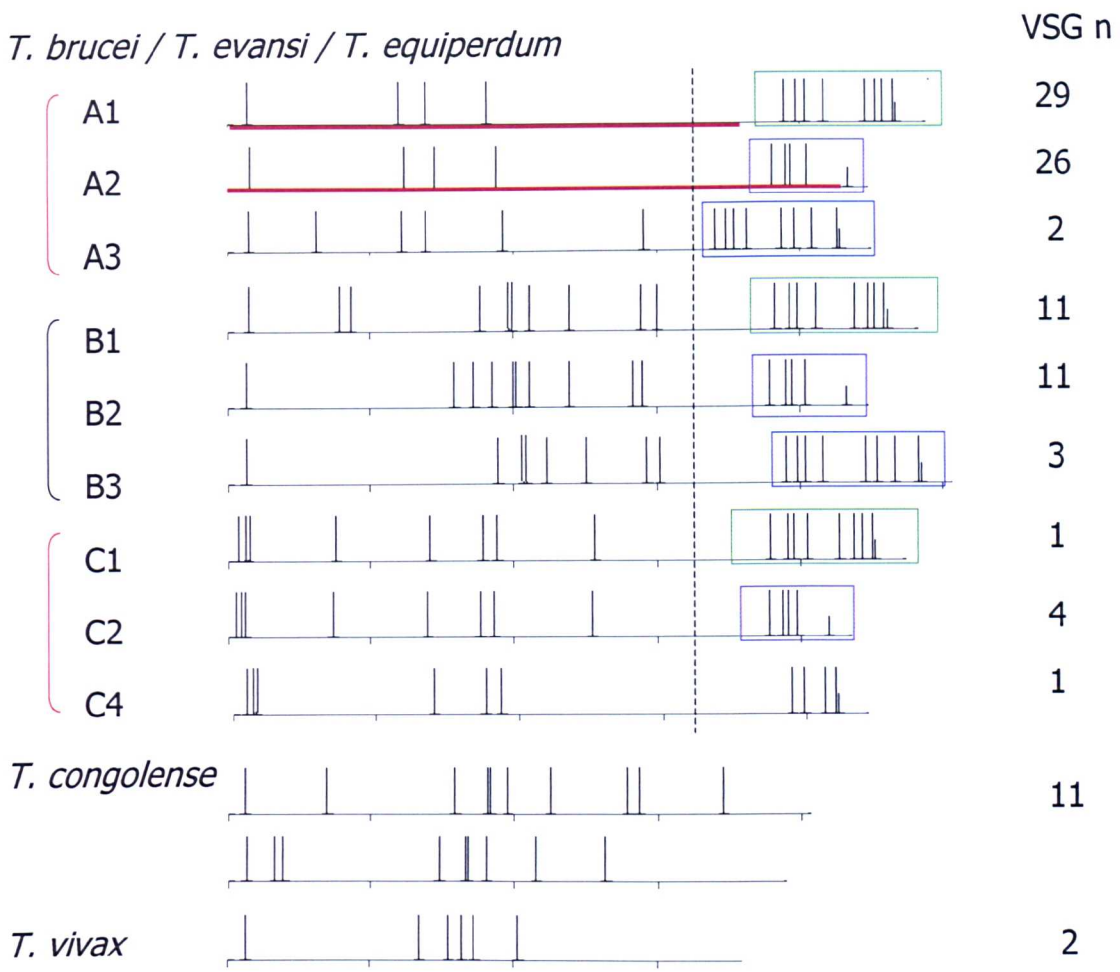


Figure 1.2: Primary structure features of known expressed VSGs, with cysteine pattern highlighted.

Cysteine residues are depicted as full-length vertical bars. Half-length vertical bars indicate the GPI signal cleavage site, if known. Example VSGs are given for each domain combination found, ordered based on the N-terminal domain. The number of VSGs falling into each category is given to the right. N- and C-terminal domains are divided by a vertical dashed line, apart from the case of *T. congolense* and *T. vivax*, which do not have such distinct C-terminal domains (see text for discussion). C-terminal domains type 1, 2 and 3 are boxed in green, purple and blue respectively. VSG sequences underlined in red correspond to the types for which tertiary structure has been determined by crystallographic studies.

There are several steps leading to the VSG becoming a surface glycoprotein. Initially, around 20 hydrophobic amino acids (signal peptide) are removed by a type I signal peptidase, directing the VSG to the secretory pathway through the endoplasmic reticulum (ER) and Golgi, where VSG folding and further post-translational modifications occur.

The tertiary structure of VSGs has been solved for two *T. brucei* type A N-terminal domains (ILTat 1.24 (Blum *et al.*, 1993); MITat 1.2 (Freymann *et al.*, 1990)), and for a type 2 C-terminal domain (MITat 1.2 (Chattopadhyay *et al.*, 2005)). It appears that VSGs

with very different primary sequence have a remarkably similar N-terminal domain fold, termed the VSG fold, consisting of two antiparallel alpha helical regions (helix A and helix B in Figure 1.3) followed by loops and smaller helices. The structure is held together by four cysteines forming two disulphide bridges, only one of which is thought to be essential, as there is an example of an expressed VSG (MVAT5) with only three cysteines (Blum *et al.*, 1993). A comparison of the two N-terminal domains for which the structure has been solved shows that structural identity reaches 60% (over 207 residues), and that residues with identical structure show only 16% sequence identity (Blum *et al.*, 1993). The N-terminal domain is also responsible for the dimerisation of the protein (Auffret and Turner, 1981). Type B domains appear to share with type A the position of the first cysteine at the beginning of the domain and sequence analysis (detection of hydrophobic heptad repeats) suggests the presence of two antiparallel alpha helices separated by a region containing glycine and/or proline; although the pattern of cysteines differs, it is therefore likely that type B domains share the same VSG fold as type A (Blum *et al.*, 1993). The same heptad repeat pattern has also been found in *T. congolense* VSGs (Rausch *et al.*, 1994). These findings rendered obsolete a previous hypothesis envisaging antigenic variation deriving from multiple folding patterns (see for example Strickler *et al.* (1987)).

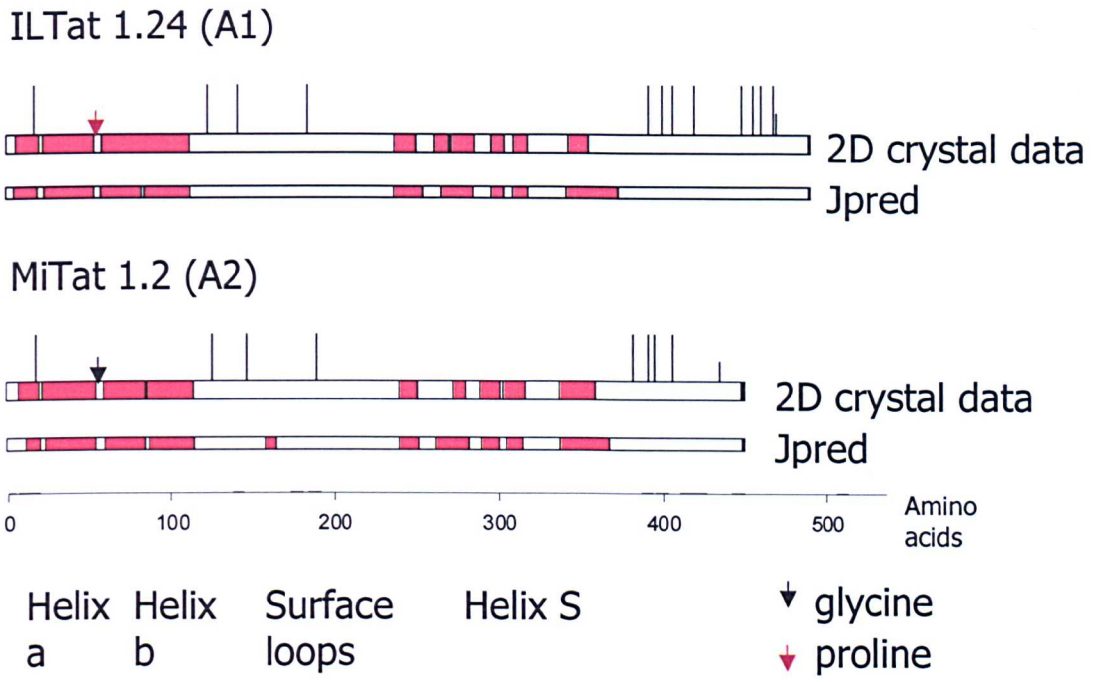


Figure 1.3: VSG secondary structure representation for MITat 1.2 (N-terminal domain type A, C-terminal domain type 2) and ILTat 1.24 (N-terminal type A), the two VSGs for which the crystal structure has been solved in detail.

Regions in red indicate alpha helices; crystallographic data is plotted into the first linear diagram for each of the two genes, whereas the second narrower plot represents the output of the secondary structure prediction program Jpred (http://www.compbio.dundee.ac.uk/~www-jpred/), used in Chapters 3 and 4 to present putative secondary structure of VSG sequences obtained in this study (the output is shown here for validation purposes).

The C-terminal domain has either four (domain types 2 and 4) or eight cysteines (types 1 and 3) (see Figure 1.2), with very few exceptions: four instances in three different strains of a domain with a type 2 GPI signal, but devoid of cysteines (accession numbers AAN78184, AF335471, AF335472 and ILTat 1.68, unpublished). The C-terminal domain is rich in lysine and glutamic acid and therefore hydrophilic, and this is thought to facilitate interactions with the polar head of phospholipids on the plasma membrane (Allen *et al.*, 1982). Modelling the known crystal structure of the MITat 1.2 C-terminal domain onto other type 2 domains suggests that, as in the case of N-terminal domains, C-terminal domains share more higher order structure than primary sequence (Chattopadhyay *et al.*, 2005). Their sequence variation is high (especially due to short indels), although not as high as between N-terminal domains. An extra cysteine is sometimes present upstream of the four or eight cysteines (hinge region), and this appears to be responsible for disulphide-linked dimerisation in solution and may have similar properties *in vivo*, strengthening the *VSG* dimer (Carrington *et al.*, 1991; Blum *et al.*, 1993). There appear not to be any major constraints in domain combinations between different N- and C-terminal domain types[9], as can be seen from the variety of domain combinations found in expressed *VSG*s (see Table 7.1 in Appendix) and by the finding that *VSG*s with closely related N-terminal domains may have different C-terminal domains (Hutchinson *et al.*, 2003).

In the process of polypeptide maturation, within a minute from initial polypeptide production (Ferguson *et al.*, 1986), at least one N-linked oligosaccharide is added to most C-terminal domains, and another one to three may be added to the N-terminal domain (Carrington *et al.*, 1991). The type of glycosylation varies and the pattern of glycosylation is determined mainly by tertiary structure of the "acceptor" VSG (Zitzmann *et al.*, 2000): comparison of tertiary structure of the two crystallised N-terminal domains suggested that N-glycosylation in MITat 1.2 had the same structural role as a helix in ILTat 1.2 (Blum *et al.*, 1993). The contribution of glycosylation to VSG structure stability varies: by use of the N-glycosylation inhibitor tunicamycin, this post-translational modification was shown to be a requirement for correct cell surface expression of VSG MITat 1.5, but not of MITat 1.2 and MITat 1.4 (Ferguson *et al.*, 1986). In addition, it was noted that six out of seven type A N-terminal domains that combined with type 1 C-terminal domains were devoid of N-glycosylation sites (Carrington *et al.*, 1991), further suggesting that N-glycosylation plays an important ancillary structural role and might be a factor in determining whether specific domain combinations are viable. Further experimental evidence tends to support this view, at least in one *in vitro* study where the N-glycosylation site was mutated (Wang

[9] different N-glycosylation patterns might constitute a constraint to domain combinations, see next paragraph.

et al., 2003). N-glycosylation is not present in the only studied *T. vivax* VSG (Gardiner *et al.*, 1996).

While still in the ER, the hydrophobic C-terminal extension (17-23 aa) is replaced via a transamidation reaction by a glycosylphosphatidylinositol (GPI) lipid anchor, which is in turn extensively modified by galactose. GPI signals in expressed VSGs appear to be very conserved, most being either 17 or 23 amino acids long (type 2 and types 1, 3 and 4 C-terminal domains respectively). Mutational analysis suggests that *in vitro* the sequence requirements may not be as stringent as might have been expected from known expressed VSGs (Bohme and Cross, 2002), although there is no *in vivo* proof to support this claim. On the other hand, the transamidation reaction is an absolute requirement even *in vitro*: failure to add the anchor due to GPI signal cleavage site mutations results in ER retention and degradation (Bohme and Cross, 2002). VSG secretion appears to be determined also by its structure, as GPI-reporter constructs using other proteins with VSG signals were not secreted (Bohme and Cross, 2002). There is evidence that *T. congolense* VSGs are GPI-anchored, but the anchor site has not been determined experimentally for any variant (Rausch *et al.*, 1994). A GPI anchor site was suggested for one VSG (YNat 1.1) (Strickler *et al.*, 1987) by comparing amino acid and cDNA data, but there appears not to be the degree of sequence conservation that occurs amongst *T. brucei* GPI signals, so it has been difficult to extend this finding to other *T. congolense* variants.

When the VSG reaches the surface it covers the trypanosome in its entirety, forming what has been defined as the "VSG coat" (see section 1.2.3 and Figure 1.4 for more information). The C-terminal domain is buried within the coat, whereas the N-terminal domain alpha helices project outwards, with the loops between helices predicted to occupy the uppermost part of the protein, exposed to the immune system of the host. The pressure for divergence therefore is acting most strongly on the N-terminal domain, with divergence operating within the constraints of maintaining 3D structure[10]. In contrast to the N-terminal domain, definition of the role of the C- terminal domain remains problematic, as it is not present in *T. congolense* or *T. vivax*. It was suggested to enhance the diffusion barrier properties of the *VSG* by elongating its structure (Ziegelbauer and Overath, 1993), although current modelling suggests that its compact structure may not warrant this role, but rather increase packing horizontally in the monolayer (Chattopadhyay *et al.*, 2005)[11].

---

[10] *VSG* divergence will be considered in more detail in subchapter 1.4.

[11] This statement was made with reference to the solved type 2 C-terminal domain structure, with only four cysteines. Whether packing alters significantly in the presence of an eight-cysteine domain remains to be determined.

In the same line of thought, N-glycosylation might have, together with its ancillary structural role, a space-filling role to hinder immunoglobulin penetration in the monolayer (Mehlert *et al.*, 2002).

1.2.3 The trypanosome cell surface and VSG recycling

The primary function of the trypanosome cell surface appears to be that of immune evasion, as VSGs account for 15-20% of total cell protein and >95% of surface proteins (Barry and Carrington, 2004). The VSG protein is thought to form dimers on the cell surface (Strickler and Patton, 1982) with 5.5 x 10^6 dimers per cell, resulting in the production of a 12-15 nm thick coat (Vickerman, 1969) (see Figure 1.4), extending also within the flagellar pocket, a specific membrane invagination at the base of the flagellum where receptor-mediated endocytosis takes place (Carrington *et al.*, 1991).

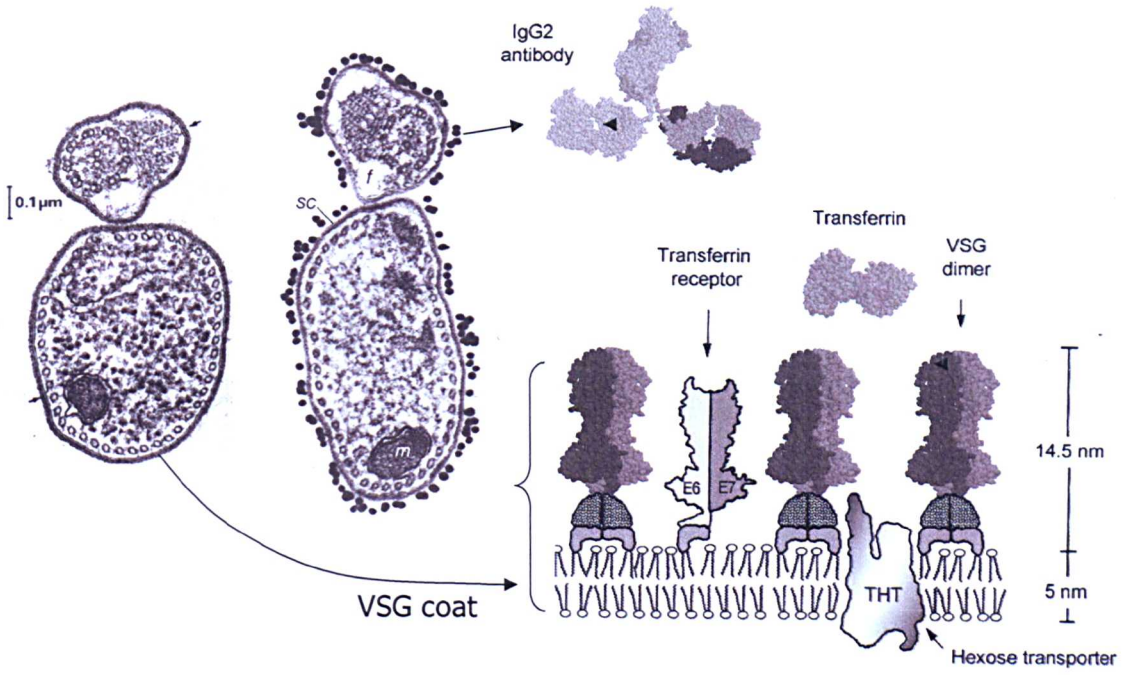


Figure 1.4: The trypanosome VSG surface coat.

To the left is a cross section of the parasite and its flagellum (f). The electron dense surface coat (sc) can be seen very clearly. The second cross section, to the right, shows antibody binding to the surface coat, by gold immunolabeling (20nm gold coupled to anti-VSG (AnTat7.1) antibodies, image kindly provided by L. Tetley). The relative size of coat components is given in the diagramatic illustration to the right, with the addition of the structure of the IgG2 antibody and host transferrin (adapted from Borst and Fairlamb (1998)).

Buried within the VSG coat, and therefore potentially shielded from the immune system, are invariant molecules. They generally are not detected by antibodies on living trypanosomes (Ziegelbauer and Overath, 1992) and include membrane transporters, proteases and receptors (Chung *et al.*, 2004). Two bloodstream form-specific invariant surface glycoproteins of unknown function, ISG65 and ISG75, are present at respectively 70 000 and 50 000 molecules/cell (Ziegelbauer and Overath, 1992; Ziegelbauer and Overath, 1993). An *in vitro* study has led to speculation that ISG65 could be involved in a signalling pathway, due to its very short half-life, a feature that is frequently associated with this type of function in other eukaryotes. In addition, it appears to be recognised by antibodies, although it is currently unknown whether this has any *in vivo* significance (Chung *et al.*, 2004). Another key surface molecule acting in the bloodstream stage is the transferrin receptor, a GPI-anchored heterodimer made up of proteins ESAG6 and ESAG7 (Salmon *et al.*, 1997), present at 3000 molecules/cell, around the flagellar pocket (Steverding *et al.*, 1995; Steverding *et al.*, 1994). It has been shown that both its steady state location in the flagellar pocket and its shorter half life (7 h) compared with that of VSGs (30 h) is due to the fact that it has a single GPI anchor on ESAG6, as opposed to the two GPI anchors of the VSG homodimer (Gruszynski *et al.*, 2006). Other surface molecules such as ISG65 are transmembrane proteins and are recycled through the same endosomal compartment as GPI-anchored proteins, suggesting the presence of multiple and overlapping mechanisms for surface molecule recycling (Chung *et al.*, 2004). It is remarkable that ISG65, ISG75, ESAG6 and ESAG7, together with the *PAG1* (procyclin associated gene) product, expressed in procyclic trypanosomes (Koenig-Martin *et al.*, 1992), probably all share the N-terminal domain type A VSG fold (Carrington and Boothroyd, 1996). It is suggested that tertiary structure provides an elongated scaffold for a variable active site at the tip of these molecules, reminiscent of the immunoglobulin gene family (Carrington and Boothroyd, 1996). The VSG type B N-terminal domain possibly also has a relative in PSSA1, a surface protein found in procyclic trypanosomes, but the crystal structure for type B VSGs is unknown, making this comparison more tentative. The C-terminal domain of VSGs does share similarity (although very localised) with part of the N-terminal region of a transmembrane surface protein, the hexose transporter (Carrington and Boothroyd, 1996); the significance of this similarity awaits tertiary structure comparison.

The question as to how invariant surface antigens are protected by the VSG coat is still at least partially unanswered and requires further investigation. The easiest explanation is that VSGs are more elongated in structure, hiding the other surface proteins, although this could be proven only by a crystal structure of these other molecules, as they appear to be of

similar molecular weight, and therefore possibly of similar height (Carrington and Boothroyd, 1996). In addition, more needs to be elucidated in terms of antibody responses against invariant antigens *in vivo.*

Another important aspect of VSG expression is that of VSG recycling, which occurs at the flagellar pocket, implicated as the only site of exocytosis and endocytosis. A steady state is reached between the internalised (endosomal) and surface VSG pools, with the endosomal pool amounting to ~9% of total VSG protein. Recycling is very rapid, 12 minutes allowing cycling of one coat equivalent (Engstler *et al.*, 2004). As the half-life of VSG molecules has been estimated to be ~30 hours (Seyfang *et al.*, 1990), this suggests repeated cycling of VSG protein with only limited replacement. Upon differentiation from bloodstream long slender to the non-dividing short stumpy form, *VSG* transcription is repressed (Amiguet-Vercher *et al.*, 2004): as the half life of the latter form is 48-72 hours (Turner *et al.*, 1995), there might also be compensatory mechanisms to enhance stability of the VSG coat (half life of 30 hours), such as reduced surface turnover, ensuring trypanosome survival. When short stumpy trypanosomes develop to the procyclic stage, procylin expression commences and *VSG* expression is further repressed: it takes 12 hours for the VSG coat to be completely replaced by procyclin, through an active process involving GPI hydrolysis and proteolytic cleavage (Gruszynski *et al.*, 2006). With regards to the suggestion that the rapid recycling could eliminate antibodies bound to the cell surface if antibodies were at low concentration, contributing to immune evasion (Engstler *et al.*, 2004), there is no conclusive evidence supporting this. One study analysed agglutination of trypanosomes in the presence of antibody and reported disaggregation of trypanosome-antibody complexes leading to fully motile dividing trypanosomes, likely to be due to endocytosis of the bound VSG-antibody complex (OBeirne *et al.*, 1998); this has been confirmed more recently and antibody degradation has been shown to proceed rapidly, in less than one hour (Pal *et al.*, 2003). While the explanation for the *in vitro* "survival" of trypanosomes exposed to antibodies was derived from what appears to be an internally consistent experiment, its relevance is unclear because immune lysis, which would be likely to kill trypanosomes during the incubation, would not occur in the reported absence of added complement (the labile nature of this factor requires exogenous complement to be added *in vitro*). Due to this shortcoming, it is not yet possible to support a model in which endocytosis of antibodies promotes immune evasion.

1.3 Antigenic variation: *VSG* gene activation and switching

1.3.1 VSG *expression sites and* VSG *transcription*

The process of *VSG* transcription is highly regulated, such that *VSG*s are exclusively transcribed when present in an active telomeric expression site (ES), and transcription is stage-specific and mediated by RNA polI (see section 1.1.4). There are two types of expression site (ES), metacyclic (MES) and bloodstream (BES): whereas the MES locus is relatively simple, the only documented example in kinetoplastids of monocistronic transcription, except for the case of the spliced leader genes, BESs are more complex, consisting of a 30-60 kb long polycistronic unit containing, besides the *VSG*, several *ESAG*s (expression site associated genes) (See Figure 1.5) (Xong *et al.*, 1998).

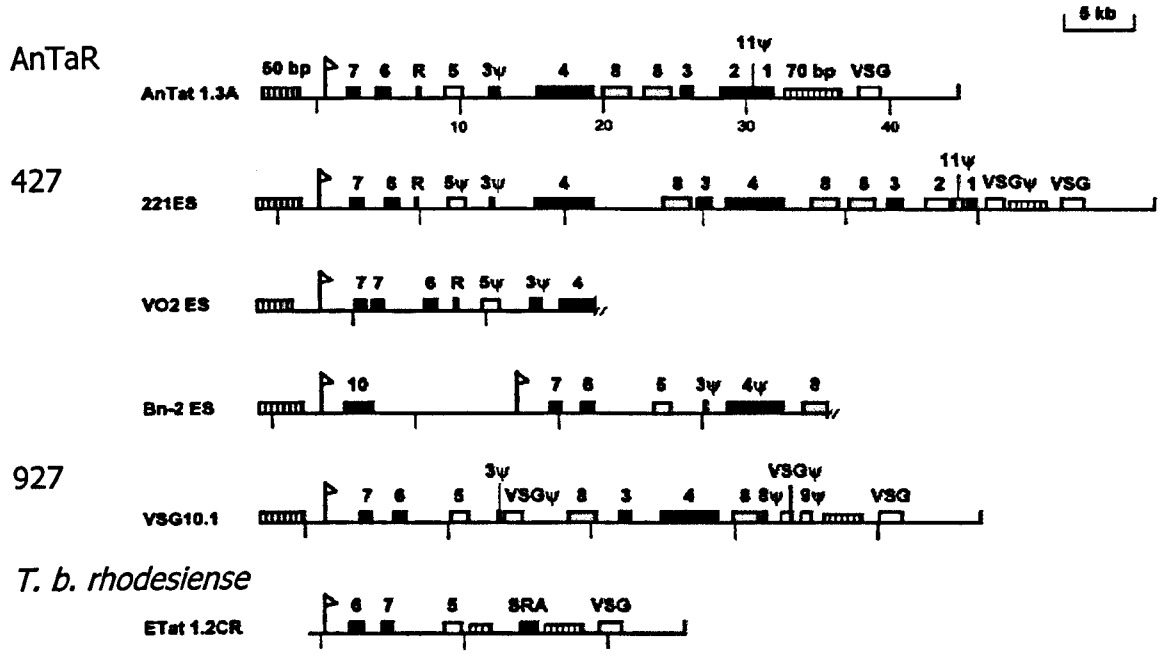


Figure 1.5: Sequenced *VSG* expression sites of *T. brucei*, modified from Berriman *et al.* (2002).

Promoters are indicated with flags, and *ESAG*s with numbered boxes (*ESAG* genes 1 to 11). The *VSG* is indicated with a white box. 50-bp and 70-bp repeat arrays (not drawn to scale) are indicated with striped boxes. Putative pseudogenes are indicated with ψ. The strain from which the expression sites were sequenced is indicated to the left. The SRA gene found in the *T.b. rhodesiense* expression site is labelled "SRA".

A puzzling feature of *VSG* expression sites is that they are present in multiple copies in the genome: ~20 BES are present in strain 427, as detected by probing at high stringency with sequences from the 221 *VSG* ES (Zomerdijk *et al.*, 1990). In a more recent study in the same strain, isolation and PCR analysis of telomere containing clones has confirmed the previous estimate, yielding a value of 17 BES (Becker *et al.*, 2004). There might be variation in BES numbers between strains, as the existence of as many as 40-50 BES has been suggested based on unique RT-PCR expression site products in strains AnTat 1 and ETat 1 (Vanhamme *et al.*, 2000). ES are mostly located at telomeres in the megabase chromosomes (de Lange and Borst, 1982; Raibaud *et al.*, 1983), but have also been found on a 225 kb and a 160 kb intermediate chromosome (Guyaux *et al.*, 1985; Shah *et al.*, 1987)[12]. In addition, in the case of *T. rhodesiense*, perhaps as many as 27 MES are present (Turner *et al.*, 1988).

1.3.1.1 Promoter

The minimal BES promoter is about 70 bp long (de Lange and Borst, 1982) and seems not to be influenced by specific surrounding upstream and downstream sequences, although deletion of all upstream sequences up to the 50 bp repeats (see Figure 1.5) was not possible, suggesting that a minimal distance between promoter and upstream repeats had to be maintained in order to have efficient transcription (Borst and Ulbert, 2001). All BES promoters are similar in sequence (around 90%) and share no similarity with MES promoters; they can be replaced by rRNA promoters, with which they do not have sequence homology, and still participate in *VSG* switching (Rudenko *et al.*, 1995). This suggests a lack of specific upstream control elements acting at the level of transcription initiation (see section 1.3.1.5). A difference was noticed, though, in promoter repression in the procyclic stage: rRNA promoters were not subject to the same level of repression as BES promoters, when inserted either at the expression site or within *VSG* arrays (Horn and Cross, 1997).

1.3.1.2 Associated genes

As mentioned above, BESs include 8-11 *ESAG*s (expression site associated genes) (Johnson *et al.*, 1987; Kooter *et al.*, 1987), their most studied products being the transferrin receptor (ESAG6 and ESAG7 heterodimer), responsible for iron binding and uptake, and an adenylate cyclase family (ESAG4), proposed to be responsible for cAMP-dependent

[12] In two different *T. brucei* strains, 427 and ILTAR respectively.

signal transduction. Some of the *ESAG*s (1, 4, 9, 11) have also been found upstream of the MES promoter (Barry *et al.*, 1998) and at other loci in the genome (Berriman *et al.*, 2005) (see Table 1.3 for a summary). Only *ESAG6* and *ESAG7* appear to be found in all BES analysed so far (Berriman *et al.*, 2002), whereas other *ESAGs* appear to be dispensable (at least in monomorphic trypanosomes), possibly also due to their alternative expression outside the expression site.

Table 1.3: Summary of ESAG product features.

Column (1) indicates whether the protein is predicted to be membrane-associated (Y/N). Column (2) gives the incidence of *ESAG* genes at non-ES locations for strain 927 (number of pseudogenes/total number of non-ES associated genes).

ESAG	FUNCTION (predicted or demonstrated)	(1)	(2)[13]	REFERENCES [14]
ESAG1	Surface glycoprotein of unknown function; expressed (at low levels) only in bloodstream.	Y	8/22	(Cully *et al.*, 1986; Carruthers *et al.*, 1996)
ESAG2	Putative surface protein	Y	3/17	
ESAG3	Potentially secreted or membrane-spanning protein; decreased expression in procyclic stage	Y	88/97	(Vanhamme *et al.*, 1999)
ESAG4	Adenylate cyclase	Y	14/71	
ESAG5	Putative intracellular soluble protein	N	0/5	
ESAG6	Transferrin receptor	Y	2/3	
ESAG7	Transferrin receptor	Y	1/2	
ESAG8	DNA binding protein	N	0/2	(Hoek *et al.*, 2002)
ESAG9	Surface protein of unknown function, possibly GPI anchored; can be transcribed from a non-ES location	Y	7/12	
ESAG10	Transmembrane protein of unknown function; often absent from ES; transcribed from other loci	Y	0/3	
ESAG11	Putative surface gene (signal peptide, possibly GPI anchor); N-glycosylation sites present	Y	6/12	(Redpath *et al.*, 2000)
SRA	Present in ES of some *T. rhodesiense* strains	Y	-	(Xong *et al.*, 1998)

An extreme example of post-transcriptional regulation, mRNA for *ESAGs* can be 100-700 fold less abundant than that of *VSG* mRNA (Cully *et al.*, 1985), despite being co-transcribed. Such low expression level makes analysis of these genes problematic. Another problem associated with studying *ESAG* function is their presence in multiple copies, expressed within and outside expression sites: an *ESAG1* knockout did not yield any phenotype, as other *ESAG1* transcripts were found to "complement" the knockout gene

[13] Data from supplementary material published online, accompanying Berriman *et al.* (2005).

[14] Information on *ESAG* genes with no specific reference was taken from introduction of Carruthers *et al.*, (1996).

(Carruthers *et al.*, 1996). Current evidence, summarised in Table 1.3, suggests that most ESAG products are surface-associated, although in most cases their specific function in bloodstream stage trypanosomes remains unknown. ESAG8, one of the few non-secreted ESAG products, has been shown to be a putative DNA binding protein (Hoek *et al.*, 2002), and has therefore been suggested possibly to have a role in BES transcription regulation. Interestingly, sequence variation within the *ESAG1* family has been shown to be greater than within other *ESAG*s, suggesting a comparison with the high divergence within the *VSG* family (Morgan *et al.*, 1996), although the significance of this comparison is currently unknown.

1.3.1.3 70-bp repeats

Both MES and BES contain (A+T)-rich 70-bp repeats upstream of the resident *VSG*. The repeats are thought to be important in *VSG* duplication and gene conversion into the active expression site, copying silent *VSG*s with similar 5' flanking repeats to the expression site (see section 1.3.4 for more details). There is a limited number (normally zero to two) of 70-bp repeats in MES, perhaps allowing limited chance for recombination, and this could explain the small size and relative stability of the M*VSG* repertoire (Matthews *et al.*, 1990). On the other hand, switching was shown to occur readily with a single 70-bp repeat (Matthews *et al.*, 1990), and a more recently described MES was found to harbour 13 repeats (LaCount *et al.*, 2001), arguing for a less genetically isolated status for *MVSG*s. In contrast, BESs harbour long arrays of repeats, between one and 20 kb long (Shah *et al.*, 1987) (see Figure 1.5 and Figure 1.6).

1.3.1.4 Switch from MES to BES

Metacyclic trypanosomes express a set of *MVSG*s (metacyclic *VSG*s) allowing establishment of infection in the mammalian host, a preadaptation developed in the salivary glands of the fly (Gray, 1965; Hajduk *et al.*, 1981). The metacyclic "archive" comprises a conserved set of up to 27 different *VSG*s (with 14 MVATs, metacyclic variable antigen types, representing 95% of population), but there is a gradual turnover in the content of this archive that is thought to prevent herd immunity in the wild: current evidence suggests that gene conversion introducing novel variants into a MES from the wider silent *VSG* archive occurs at a rate of 0.03 of the metacyclic "archive" per fly per transmission (Barry, 2006). The switch from MES to BES following transmission from fly to mammal is not understood, as it does not occur simultaneously with the differentiation

from metacyclic to bloodstream stage. Whereas the differentiation occurs within one day, bloodstream VATs start appearing 4-6 days after infection (Barry and McCulloch, 2001).

1.3.1.5 Single BES expression

It appears that BES (and procyclin) promoters are never completely inactive in any examined life cycle stage, instead being partially up or down regulated (Roditi, 1996), whereas MES are subjected to a tight transcriptional control (Barry *et al.*, 1998). While the mechanism by which only a BES is transcriptionally active is unknown, the regulation appears to be stringent, as stable maximal activation of two separate BESs by insertion of two drug resistance markers has not been possible (Borst and Ulbert, 2001), and previous examples of stable double expressors have been questioned (Cross *et al.*, 1998). Interestingly, transcription is only partially repressed in silent BESs, as *ESAG6* and *ESAG7* transcripts (the first genes in the polycistronic unit) have been found to be transcribed in different amounts from a number of silent BESs (Ansorge *et al.*, 1999). These different activity states appear to be stably inherited and have been considered epigenetic. Furthermore, the different activity levels seem not to correlate with the transcriptional "switching avidity" of silent BESs (Navarro *et al.*, 1999; Borst and Ulbert, 2001). In addition, it appears that primary transcription in the active expression site is not upregulated compared with that of inactive BESs, transcription regulation occurring at the elongation, rather than initiation, stage (Vanhamme *et al.*, 2000).

A positive lead to uncovering the mechanism allowing allelic exclusion to the advantage of a single expression site was the finding that transcription appears to occur in a non-nucleolar "expression site body" (ESB), from which silent BESs are excluded, so it is likely that this circumscribed region contains the factors that enable full processivity of RNA polymerase, by overcoming default silencing factors and/or allowing the opening of chromatin (Navarro and Gull, 2001). These results fit very well with the location of telomeres in the nucleus: it appears that in the procyclic form they are all sequestered to the nuclear periphery, whereas dividing bloodstream forms give also a telomere signal from a non-peripheral nuclear location, which might correspond to the ESB (Perez-Morga *et al.*, 2001).

An additional factor allowing silencing to occur has been considered to be the telomeric location of expression sites: a telomere position effect silencing mechanism has been proposed by analogy with yeast, where chromatin structure at telomeres influences their transcriptional status. The problem with the analogy is that silencing in yeast spreads

merely some 5 kb upstream of the telomere, whereas in trypanosomes it would have to spread over 40-60kb, in the case of BES (on the other hand, the MES promoter would be close enough to the telomere to envisage this silencing mechanism as a possibility) (Barry and McCulloch, 2001). The presence of a histone H3 variant that is enriched at subtelomeres argues that subtelomeres can have a chromatin context different from that of core chromosomal regions (Lowell and Cross, 2004), and that this context could be related to default repression in all but the ESB-associated active expression site telomere (Horn and Barry, 2005). In addition, a modified thymine base, J (Beta-D-glucosyl (hydroxy-methyl) uracil), has been found associated with repetitive regions in the genome of bloodstream stage trypanosomes, and in particular with telomeres. It is present in inactive expression sites and is not in the transcribed BES, its formation being associated with a chromatin remodelling protein, JBP2 (J-binding protein 2), which is thought to enable the two step enzymatic reaction leading to J biosynthesis from thymine residues (DiPaolo *et al.*, 2005). In the same study, JBP2 has been shown to have a punctate nuclear localisation, suggestive of an ability to associate with repetitive sequences, and it is currently thought to be essential in bloodstream form trypanosomes. Whether base J biosynthesis is an active player in BES repression remains to be elucidated.

In summary, current evidence tentatively suggests that inactive telomeric expression sites are sequestered at the nuclear periphery and associate with specific chromatin remodelling factors; both BES nucleosomes and DNA might be altered compared with the active expression site. For activation of a silent BES to occur, at least some of the telomeric factors would have to be displaced and epigenetic modifications would have to be removed, as the tethering to the nuclear periphery is disrupted, allowing access to the ESB.

Although different BESs can be activated, there seems normally to be a dominant BES that is used most of the time. As to why more than one BES is present, two non-mutually exclusive theories have been put forth. The first states that expression of a given BES might be selected in a host to maximise the affinity with the transferrin of that host (Bitter *et al.*, 1998), although this has recently been questioned, suggesting that the correlation between BES switching and host transferrin might not be direct, other stress inducing changes in the trypanosome medium contributing also to BES switching (Salmon *et al.*, 2005). Alternatively, silent BES can be considered to have originated as sites for mosaic *VSG* production, so that the process of splicing variants together could occur in the absence of transcription, preventing the lethal expression of a non-functional, imperfect mosaic *VSG* (Barry and McCulloch, 2001). Support for this theory comes from the detection of two silent telomeric pseudogene donors contributing respectively to mosaic genes VSG20

(Thon *et al.*, 1989) and VSG78 (Roth *et al.*, 1989), detected in a rabbit infection with *T. equiperdum* (this is further discussed in section 1.3.7.).

1.3.2 VSG *activation (switching)*

It has been show that switching between two *VSG*s is spontaneous and independent of the immune system, and occurs at a rate of 10^{-2}-10^{-3} switches/cell/population in pleomorphic trypanosomes, while in laboratory-adapted monomorphic trypanosomes the rate is much lower (10^{-6}-10^{-7}). The high switching trait is apparently dependent upon transmission through the fly, as a low-switching line obtained by syringe passaging could partially revert to high switching by completing the life cycle (Turner, 1997). One possible explanation is that loss of a labile epigenetic state might be responsible for the reduced rate of switching in monomorphic lines.

Several mechanisms of *VSG* activation have been described. These include transcriptional (*in situ*) switching between BES, telomere reciprocal exchange between BESs downstream of the activated promoter (Pays *et al.*, 1983a; Rudenko *et al.*, 1996) and various forms of duplicative transposition, which involve the copying of a silent *VSG* (basic copy, BC, or donor gene) into an expression site (creating an expression-linked copy, ELC) and the deletion of the previously expressed *VSG*[15] (see Figure 1.6 for main mechanisms). Transcriptional switching appears to be rapid: no transient state in which two BESs are active has been detected (Chaves *et al.*, 1999; Ulbert *et al.*, 2002). Although switching is very rapid at the gene expression level, it can be regarded as the trigger for a longer process at the organismal level: *VSG* mRNA has a half life of 4.5 hours and VSGs have a half life of 30 hours (Cross *et al.*, 1998), so it is possible that the complete replacement of a VSG coat with another non-cross reactive one might take around two days *in vivo,* unless there exists an as yet unidentified mechanism that speeds up the process of coat exchange upon switching. The switching from MVAT to BVAT, monitored by detection of dual expressors with monoclonal antibodies, would suggest two days as the upper estimate for coat replacement, as dual expressors were first detected at day four of infection and by day six only 6% of the population had MVATs on their surface (Esser and Schoenbechler, 1985).

[15] Hence the name clonal *phenotypic* variation to indicate the process of antigenic variation, as genotypic information is retained in the silent archive.

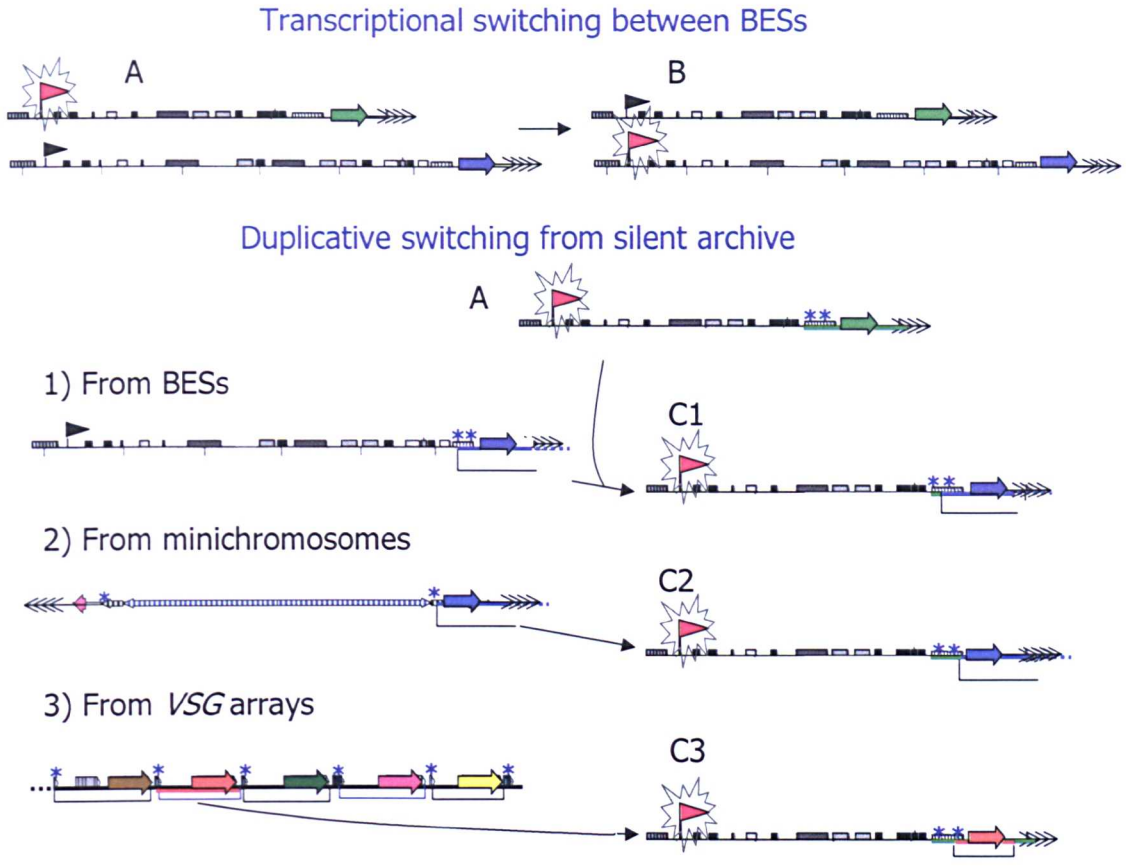


Figure 1.6: *VSG* switching mechanisms, transcriptional and recombinational.

"A" refers to the original active BES, "B" to the result of transcriptional switching, activating a second BES; "C1", "C2" and "C3" refer to the result of duplicative transposition from three genomic loci (BES, minichromosomes, *VSG* arrays). BES promoters are indicated as red flags in the active state, as black flags in the inactive state. The depiction of the two BESs is derived and adapted from (Berriman *et al.*, 2002), see also Figure 1.5. *VSG* genes are indicated as coloured blocs with direction of the gene highlighted. Asterisks indicate 70-bp repeats upstream of *VSG*s (see section 1.3.4 on gene conversion boundaries for a discussion of their role). The horizontal bracket underlying donor *VSG*s highlights the *VSG* cassette of array *VSG*s (see section 1.3.4). The bracket is left open in the case of telomeric conversions C1 and C2, as the conversion might include the whole downstream region and telomere (indicated with "empty" arrows). The sequence surrounding donor and recipient *VSG* sequences is given the same colour as that of the coding sequence, to highlight the extent of sequence that is duplicated.

1.3.2.1 *VSG* switching and hierarchy of expression

Order in the sequential expression of different variants has been associated with distinct switching reaction types occurring at different frequencies, leading to variants having a range of activation probabilities, depending on their locus. Telomeric *VSG*s are more easily activated than array *VSG*s, and this has been formally demonstrated by the observation that a late gene was expressed early when it was transferred to a telomeric position (Laurent *et al.*, 1984). This is possibly because the sequence environment of a telomeric *VSG* is more similar to that of the BES (allowing more sequence homology to

trigger recombination reactions) and because of the general capability of telomeres to recombine ectopically with each other (Horn and Barry, 2005). Early in infection, when telomeric genes are activated, in monomorphic strains around two thirds of switching is transcriptional (Liu *et al.*, 1985), while in pleomorphic trypanosomes more than 90% of switching is duplicative[16] (Robinson *et al.*, 1999); this implies that the lower switching rate of monomorphic trypanosomes is associated with a reduced level of duplicative transposition. Later in infection, on the other hand, duplicative transposition involving array *VSG*s is common to both pleomorphic (Morrison *et al.*, 2005) and monomorphic strains (Timmers *et al.*, 1987; Lee and Van der Ploeg, 1987).

1.3.3 Duplicative transposition

The process of recombination of a silent *VSG* into an expression site can occur in a variety of ways, as the boundaries of transposition depend on the homology between the active *VSG* and the incoming *VSG*. The *VSG*-encoding region may be replaced in its entirety from the upstream 70-bp repeats down to the C-terminal domain-encoding sequence (or to the telomere if the donor *VSG* is telomeric), or only parts of it can be replaced (segmental duplication leading to mosaic gene formation, discussed in section 1.3.7).

1.3.4 Gene conversion boundaries, sequence homology and hierarchy

A *VSG* recombination unit (also termed *VSG* cassette) is composed of 70-bp repeats, a 5' CTR (cotransposed region), the *VSG* itself and a 3' region of variable length (Liu *et al.*, 1983), with potential to extend to the telomere in the case of telomeric donor *VSG*s (see Figure 1.6). Unless related *VSG*s recombine with each other, habitually the regions of homology between *VSG* cassettes would be restricted to the 70-bp repeats at one end and, at the other end, to the final portion (encoding the GPI signal) of the *VSG* coding sequence and to downstream sequences. The intervening region, comprising the 5' cotransposed region and most of the *VSG* coding sequence, has been shown to be largely devoid of any conserved or repetitive element (Van der Ploeg *et al.*, 1982a). Both 5' and 3' flanks will now be described in some detail, before the molecular mechanisms and players that are thought to be associated with the use of these sequences are introduced.

[16] Data from Liu (1985) are derived from single relapse studies, in which the original variant is grown in a rodent, blood is taken from the first peak and treated with antibodies against the initial variant, then a small number of trypanosomes (4-5) are injected into a second rodent, amplifying switchers from the initial peak.

The 70-bp repeat sequence was first obtained by sequencing the 5' region of transposition in the activation of array gene 118 in strain 427 (Liu *et al.*, 1983). Although the precise point of transposition was not determined, it mapped to the region spanning a set of five 70-bp repeats. By southern blotting, the repeats were shown to be present in many copies across the genome, associated with all known *VSG* genes (Van der Ploeg *et al.*, 1982b), giving a first indirect estimate of the size of the *VSG* archive. A more extensive analysis was provided by the sequencing of a 4 kb array of 53 repeats present within a BES in the EATRO 795 (ILTAR) pleomorphic strain: comparison showed that repeats are imperfect and vary in size between 60 and 110 bp (Shah *et al.*, 1987), although longer repeat units (~400 bp) have also been isolated (Campbell *et al.*, 1984). The repeat unit is a tripartite element comprising a variable number of triplet TAA repeats, followed by a ~20 bp long GT-rich region and a ~20 bp long AT-rich region. All three components of 70-bp repeats have been advocated as potential contributors to recombination initiation: the triplet repeat has been shown to predispose the DNA helix to melting *in vitro* (Ohshima *et al.*, 1996); the GT-rich tract has been suggested to be a binding site for RAD51 and even to be recognised and cut by a restriction endonuclease (Michiels *et al.*, 1983; Matthews *et al.*, 1990), due to the possibility of it adopting a Z DNA conformation (Aline *et al.*, 1985); the final AT rich tract has been proposed to form a stem loop (Liu *et al.*, 1983) and act as a recognition signal for DNA binding proteins. As yet, no conclusive evidence has been produced for any of these putative events to operate in switching. The association of the 70-bp repeat with the upstream limit of recombination in *VSG* gene duplication reactions has been reported extensively (Shah *et al.*, 1987), and it appears that multiple initiation-termination points within the repeat are chosen "at random" (Florent *et al.*, 1987), the conversion therefore acting possibly by homologous, rather than site-specific, recombination. Conversions have also been found to occur in the absence of 70-bp repeats in monomorphic trypanosomes with an expression site engineered to be devoid of them (McCulloch *et al.*, 1997). This parallels the high numbers of duplications detected in monomorphic trypanosomes that do not use 70-bp repeats, and contrasts with the seemingly exclusive use of these sequences in pleomorphic trypanosomes (Barry, 1997), suggesting that switching rate and use of 70-bp repeats in gene conversion are related processes, both downregulated in monomorphic trypanosomes.

Although a significant number of indels are present (Field and Boothroyd, 1996), the 3' region downstream of the *VSG* coding sequence has been shown to be relatively well

---

Data from Robinson (1999) are derived from the natural first relapse peak (the second peak) of a chronic infection, so the two experiments might not be fully comparable.

conserved in outline (Liu *et al.*, 1983), providing the 3' end homology in the *VSG* cassette, together with the sequence encoding the conserved C-terminal end of the *VSG* itself (Rice-Ficht *et al.*, 1981). Telomeric *VSG* copies are also associated with several conserved repeats, providing more extensive blocks of homology between cassettes, in many cases extending to the telomere itself (Kooter *et al.*, 1988). A conserved 16-mer is present very close to the *VSG* end, and further downstream TTAGGG telomeric repeats are found, together with motifs termed Trpt and srpt2 (H. Renauld, pers. comm.). The 16-mer is present in the 3' UTR of all *T. brucei VSG* mRNA (Rice-Ficht *et al.*, 1981; Matthyssens *et al.*, 1981; Majumder *et al.*, 1981)[17]: it has been shown to contribute to the regulation of stage-specific abundance of *VSG* mRNA, apparently stabilising the mRNA transcript (Berberof *et al.*, 1995). The 16-mer has been found downstream of many silent *VSG* copies (see this study, Chapter 5, section 5.2.5) and, in addition, other telomeric repeats have been found 3' of some array *VSG*s (H. Renauld, pers. comm.), suggesting the possibility of low-frequency recombination from telomere to array, a possibility that has never been demonstrated experimentally.

1.3.5 A molecular model for duplicative transposition

Several mechanisms for duplicative transposition have been proposed, within the framework of a general model envisaging a double strand break at 70-bp repeats in the active BES, followed by repair of the break using a silent *VSG* cassette as a template, resulting in deletion of the BES-resident *VSG* and replacement with the silent donor *VSG* (Barry, 1997; Barry and McCulloch, 2001). The favoured mechanism is that of synthesis–dependent strand annealing (SDSA): after a double strand break in the 70-bp repeats is produced, either by a specific endonuclease or because of the intrinsic instability of the repeats, a 3' single strand would be revealed by the action of an exonuclease. This single strand would then conduct strand invasion at silent *VSG* cassettes, seeking homology. Once annealing with a homologous stretch of DNA has taken place, gene conversion would then proceed by branch migration and termination of conversion would occur when a short stretch of homology is found, usually at the 3' end of the *VSG* coding sequence. The problem associated with this mechanism is that it relies on analogy with a pathway discovered in yeast, involving the recombination enzyme RAD52, which does not seem to be present in trypanosomes. It nevertheless has advantages over other mechanisms such as double strand break repair (DSBR) and break induced replication (BIR): DSBR involves

[17] This is absent in *T. congolense VSGs*, which have very short (7-24 bp) or even absent untranslated regions before the poly[A] tail (Urakawa *et al.*, 1997).

formation of Holliday junctions and therefore the risk of lethal translocations occurring, BIR converts the double strand break into a replication fork, so it could be implicated only in telomeric conversion, where replication is allowed to run to the telomere, and conversion preserves BES architecture. If BIR involved array *VSG*s, other *VSG*s downstream could be copied, effectively resulting in translocation of a *VSG* array downstream of a BES, which is unlikely to be a viable conversion outcome. The presence of these (and other) pathways might not be mutually exclusive: as described in the next section, homologous recombination in trypanosomes relies on more than one strategy.

1.3.6 Homologous recombination pathways

A key player in homologous recombination is the RAD51 enzyme: it binds single- or double-stranded DNA in the presence of ATP and forms nucleoprotein filaments in which the DNA is extended and underwound relative to its normal structure, allowing strand invasion and homology searching to take place. In order to observe whether *VSG* gene conversion reactions relied on homologous recombination, the *T. brucei rad51 -/-* mutant was analysed for its ability to switch between *VSG*s. The reduced switching rate (2-130-fold) showed that RAD51 contributes to *VSG* switching and, indirectly, also uncovered a RAD51-independent recombination pathway that is able to catalyse switching (McCulloch and Barry, 1999). The substrate requirements for this pathway were tested by transformation of *T. brucei rad51 -/-* mutants with linear constructs and by observing the pattern of integration. A 10-30-fold reduction in transformation efficiency was observed, but this second pathway could nevertheless perform faithful homologous recombination, requiring as little as 24 bp of perfect sequence homology (Conway *et al.*, 2002). On the other hand, the minimal efficient processing sequence for RAD51 in *T. brucei* is 142 bp (Bell *et al.*, 2004), suggesting that the alternative pathway described above might provide scope for interaction between sequences with only localised homology, such as in intragenic recombination between *VSG* genes. Current evidence (R. Barnes, pers. comm.) suggests that recombination between substrates that are shorter than 50 bp is highly impaired and mismatches (5% divergence) are poorly tolerated. The imperfect nature of 70-bp repeats and the above-described limitations that recombination faces in allowing short sequences with mismatches to engage in gene conversion reactions lends strength to the argument that there must be sequence-specific features of these repeats (and/or factors that associate with them) enabling these limitations to be overcome, enhancing recombination at these loci.

Additional findings were reached by analysis of the *rad51 -/-* mutants. A construct that could recombine either at the BES or internally at the tubulin array gave corresponding levels of recombination into the two locus types in the wildtype background (RAD51-dependent pathway), whereas the *rad51 -/-* mutants (RAD51-independent pathway) appeared to prefer integrating at the BES. This suggests a preferential association of putative RAD51-independent recombination factors with subtelomeres and further implicates alternative recombination pathways with *VSG* switching (Conway *et al.*, 2002).

1.3.7 Mosaic genes and hierarchy

There have been two main *in vivo* studies in which variants were detected that could be ascribed to the presence of multiple donors, contributing together to form a mosaic gene. Both involved chronic rabbit infections, one using the *T. rhodesiense* WaTat strain (Barbet *et al.*, 1982), and the other using the *T. equiperdum* BoTat strain (Capbern *et al.*, 1977). The aim of the first study was to isolate variants that were similar (partially cross-reactive) to the initial variant WaTat1.1. Three separate expressed genes were found, WaTat 1.12, 1.13 and 1.14. It was possible to explain all four expressed copies by a combination of gene conversion reactions utilising four silent genes, at least two of which were pseudogenes (Kamper and Barbet, 1992). A limitation of this study is that expressed variants cannot be clearly related to the rabbit infection: clone WaTat 1.12 was obtained from passaging rabbit blood from day 14 into a rat and was then passaged in mice, where it was grown for seven weeks before being cloned into a second rat. Similar procedures (exact timing of gene isolation not described) were followed to isolate the other two variants. The importance of this study lies chiefly in the attempt to assess the degree of divergence amongst mosaic genes leading to the production of novel non-cross-reactive variants. It appears that closely related mosaics derived from silent copies that are around 90% identical at the DNA level produce mosaics that are largely cross-reactive.

Of more interest from the perspective of understanding the *in vivo* significance of mosaics is the *T. equiperdum* study. The study isolated 101 variants from 11 rabbit infections. Two variants, VSG20 (Thon *et al.*, 1989) and VSG78 (Roth *et al.*, 1986), were explainable as mosaic genes, derived from silent basic copies, at least some of which were pseudogenes. Mosaics that were identical or sufficiently similar to these two VSGs to be detected by specific antisera were observed in separate infections: VSG20 was detected in 8 rabbits between day 21 and 30, whereas VSG78 appeared between day 33 and 45 in four rabbits (60% of the rabbits remaining in the study by those times) (Capbern *et al.*, 1977).

This shows that mosaic gene formation is, at least in these cases, a reproducible and predictable event.

It has been suggested, following analysis of two mosaics related to *VSG*20, that the gene ends of *VSG*s are responsible for ordered appearance of variants. It was concluded that genes with similar 3' end could be found to be expressed at the same time, and that the level of homology in the rest of the gene would enable prediction of the probability for transition from one gene to another (Thon *et al.*, 1990). On the other hand, it had been shown previously that *VSG*s with the same end were expressed at different times during two rabbit infections with strain 427[18] (Michels *et al.*, 1983), so the notion of order imposed by 3' donor sequences needs to be tested further.

1.3.8 Modelling VSG *switching and hierarchy*

A hierarchy of expression of different VATs has been shown experimentally, VATs being grouped as early, middle and late, based on their mean timing of appearance (Capbern *et al.*, 1977); (Gray, 1965). Recent reanalysis of data published by Capbern (J.D. Barry, pers. comm.), suggests that VATs appear more as a continuum, in keeping with previous experimental work conducted in *T. vivax* (Barry, 1986). The order is semi-predictable, as the exact timing of appearance can vary between different infections of the same host species. Order is retained also when infections in different host species are compared, although progression through the *VSG* archive has been shown to be faster in large animals, such as cows, than in smaller animals, such as rodents, as a larger number of parasites can be accommodated in animals with a larger volume of blood, leading to a higher number of switching events (Barry, 1986). The faster VAT progression in larger animals might be the underlying cause for the reported cases of self-cure in cattle (Barry, 1986) and in pigs (Penchenier *et al.*, 2005), an event much less common in rodents (Barry, 1986). In this scenario, natural termination of the infection might be due to exhaustion of the entire *VSG* archive. At the population level, the hierarchy of expression of different *VSG*s is not a linear process, as switching is also divergent and one VAT gives rise to several VATs. On the other hand, "convergent" events represented by multiple duplicative activations of the same silent gene within an infecting trypanosome population[19] (Timmers *et al.*, 1987) and by a similar timing of appearance of the same VAT in different infections

[18] The three VSGs with related C-terminal domains are VSG 121a, 117a and MITat 1.196; in rabbit 1, VSG 121a and 117a appeared at day 20 and 28 respectively; in rabbit 5, VSG 121a was used to start the infection and MiTat 1.196 was isolated at day 45 (rabbit numbers refer to those used in the cited study).

[19] Also possibly due to a single switching event followed by repeated gene conversion between donor and expressed copies.

(Morrison *et al.*, 2005) have uncovered the probabilistic nature of *VSG* switching. The hierarchy is therefore the result of interplay between "fixed" different *VSG* activation probabilities and the pressure of the immune system against the re-emergence of VATs for which antibodies are already present: transferring trypanosome by syringe to a new host resets the order of expression to early *VSGs* (Gray, 1965), which are those with high activation probabilities. Recent models have therefore factored in that early variants are always activated throughout an infection, but are cleared by antibodies as the infection proceeds (Frank, 1999; Lythgoe *et al.*, 2006). This type of model has outlived that proposed by Seed in 1978 (Seed, 1978), according to which hierarchy derived from growth differences amongst VATs, such that the first VATs expressed are those that grow better. The latter model has not received significant experimental support and there is evidence against it from growth rates of different VATs (Seed, 1978; Miller and Turner, 1981). Recent modelling is considering ways of constructing likely "transition pathways" between variants to expand from the simpler idea of loose time groupings detected in the course of infection, using published data on *Borrelia* antigenic variation (Frank and Barbour, 2006). In *T. brucei*, statistical analysis and modelling of 30 chronic infections in mice and two in cattle following the onset of seven single copy genes suggests that locus type is associated with hierarchical expression early in infection, whereas possible "transition pathways" relying on coding sequence homology might occur at later stages of infection, when mosaic genes are produced (Morrison *et al.*, 2005).

1.4 Antigenic variation: *VSG* gene family divergence and evolution

1.4.1 VSG *divergence*

Overall, although some *VSG* families with more than 70% protein sequence identity have been detected (Pays *et al.*, 1985; Field and Boothroyd, 1996), on average two aligned *VSGs* would be only 20% identical, and less than that if aligned on the basis of tertiary structure (Carrington and Boothroyd, 1996), so the abundance and breadth of *VSG* families remain elusive. As mentioned in section 1.3.7, it appears that cross-reactivity has been detected between variants having ~90% identity (Thon *et al.*, 1990), whereas two related VSGs sharing 75% sequence identity (AnTat 1.1 and AnTat 1.10) were non cross-reactive (Carrington *et al.*, 1991); it therefore seems that minimal divergence for non-crossreactivity might lie between 75 and 90% sequence identity, although a much larger sample size would be needed to confirm this hypothesis. Variation is continuous along the

N-terminal domain and the reason for this non-specific divergence is unknown: it was suggested to minimise activation of T helper cells by processed peptides derived from *VSG*s that might otherwise occur during a persistent infection (Blum *et al.*, 1993), but no evidence confirming this has been gathered since. In addition, the extent of divergence amongst *VSG*s (protein sequence identity as low as 20%) suggests that the mechanism that generates *VSG* diversity might be producing more divergence than required, its only constraint at the level of selection being conservation of tertiary structure (Blum *et al.*, 1993).

1.4.2 VSG epitopes

Two studies have addressed the issue of *VSG* divergence by selecting with monoclonal antibodies for the emergence or detection of *VSGs* closely related to the initial expressed VSG. In the first, analysis of six related variants originating from VSG78 of *T. equiperdum* (cross-reactive using polyclonal antisera), experimentally selected for loss of reactivity to monoclonal antibodies (Baltz *et al.*, 1991), has revealed that a small number of point mutations (1-3) might be sufficient to generate different epitopes. Analogously, a second experiment selecting for VATs reacting to a monoclonal antibody against MVAT5 (WRATat strain) upon prolonged *in vivo* passaging (3-4 months) yielded three VATs, expressing *VSG*s closely related to MVAT5, with 35, 11 and 28 point mutations respectively (Lu *et al.*, 1994). Both experiments were questioned because of the artificial procedures utilised to select point mutated variants, far removed from the normal course of an infection: only one point mutation in three MVSG clones was detected from the first peak of a "standard" infection (Graham and Barry, 1996). What remains to be established is whether point mutations accumulate *in vivo* in the course of an infection for the 'purpose' of immune evasion, as the above study considered only the first peak of parasitaemia. Exposed epitopes were studied on the MITat 1.6 VSG with nine monoclonal antibodies grouped into five classes recognising different "regions" of the protein. All putative epitopes mapped to the N-terminal domain, but with no precise link to primary sequence; only one of the five antibody classes was found to bind to the surface of live trypanosomes (Miller *et al.*, 1984a; Miller *et al.*, 1984b). Further work to pinpoint epitopes on MITat 1.2 VSG, for which the crystal structure was available, did not yield a much more precise understanding of epitope location (Masterson *et al.*, 1988).

1.5 Antigenic variation in other systems

Antigenic variation is a common theme of both eukaryotic and prokaryotic parasites that develop chronic infections in their hosts. *T. brucei* represents one of the best studied systems, but other protozoa display this immune evasion strategy, including *Plasmodium* (Kraemer and Smith, 2006), *Pneumocystis* (Keely *et al.*, 2005) and *Giardia* (Nash, 2002). Amongst eubacteria, the antigenic variation system of the spirochetes *Borrelia hermsii* (Dai et al., 2006), and *B. burgdorferi* (Zhang *et al.*, 1997) has been well characterised, as is that of *Anaplasma marginale* (Futse et al., 2005) and *Neisseria gonorrhoeae* (Zhang *et al.*, 1992; Sechman *et al.*, 2005). Recombination between antigen alleles is a common theme amongst all species quoted above, and seems to function as the primary diversification and immune evasion method. Variations between species concern the size of the antigen archive, the mode of antigen activation (transcriptional or by duplication into an expression site) and the locus occupied by the archive (see Table 1.4). It is now also apparent that certain antigenic variation systems include the simultaneous variation of different surface protein gene families (see Table 1.4). In protozoan pathogens, a common theme emerging is the use of subtelomeres to house the surface antigen archives, and again (see section 1.3.2.1) this is likely to be due to the plasticity of chromosome ends and their proneness to ectopic recombination, involving homologous sequences on non-homologous chromosomes (Barry *et al.*, 2003).

Table 1.4: Antigenic variation in eukaryotic parasites (adapted and expanded from Borst and Ulbert (2001), see text for additional references).

Organism	Surface antigen	Number of genes (approximate.)	Mechanism of antigen activation	Archive location (in protozoa)
Protozoa				
T. brucei	VSG	1000	Transcriptional and recombinational	Subtelomeric and telomeric
P. falciparum	PfEMP1 (var)	59	Transcriptional	Subtelomeric and interstitial
	Rifin	200		Subtelomeric
P. carinii	MSG	100	Recombinational	Subtelomeric and telomeric
G. lamblia	VSP	150	Transcriptional	Interstitial
Bacteria				
Borrelia hermsii	VMP (Vlp + Vsp)	40	Recombinational	
Borrelia burgdorferi	VLS	15	Recombinational	
Anaplasma marginale	MSP2	7-9	Recombinational	
	MSP3	7-9		
Neisseria gonorrheae	PilE + PilS	20	Recombinational	

1.6 Specific aims of project

This project fits into the efforts of Prof. Dave Barry's group to provide a model that combines trypanosome growth with antigenic variation rate and sequential *VSG* expression and then to test the hypothesis that these parameters are driven by the DNA recombination mechanisms responsible for individual *VSG* switches.

The specific aims are to analyse the *T. brucei* strain 927/4 silent *VSG* archive in order to make predictions of its mode of use, exploring in detail the internal composition of the archive, looking for patterns of relatedness amongst variants and building a model of how these sequences are likely to contribute to antigenic variation (Chapter 3). The model was tested *in vivo* by conducting a chronic mouse infection with strain 927/4 (Chapter 4), providing further insights into the kinetics and features of antigenic variation, with the ability to trace expressed *VSG*s back to the genome sequence. Bioinformatic findings provided the scope for undertaking a bioinformatics and preliminary experimental analysis of a novel *VSG*-related gene family (Chapter 5).

CHAPTER 2

MATERIALS AND METHODS

2 Materials and Methods

2.1 Culturing trypanosomes

Two different trypanosome strains were handled, TREU 927 both *in vivo* and *in vitro*, Lister 427 only *in vitro*. gDNA for the other two strains (EATRO 795 and STIB 247) used in PCR experiments in Chapter 5 (see Figure 5.9), was kindly provided by Rebecca Barnes.

In vitro growth of strain Lister 427 was performed solely for the PCR experiment in Chapter 5 (see Figure 5.11). It was achieved by using HMI-9 medium (Hirumi and Hirumi, 1989), supplemented with hygromycin (10 μg/ml), at 37°C in a humidified 5% CO_2 incubator.

The bloodstream GUTat 10.1 derivative of strain TREU 927 was used for setting up a chronic infection. This trypanosome strain is pleomorphic and cannot be grown easily in culture. Bloodstream stage trypanosomes from the infection were differentiated to the procyclic form, by the protocol given below. The procyclic GUTat 10.1 line was grown at 27°C to obtain RNA, for the PCR experiment shown in Figure 5.10, using SDM79 medium (Brun and Schonenberger, 1979).

2.1.1 Stabilate preparation

Bloodstream form trypanosome culture stabilates were prepared by addition of 10% sterile glycerol to 900 μl of culture at a density of ~2 x 10^6 trypanosomes/ml, in cryotubes (Nunc). In the case of procyclic trypanosomes, the culture (at a density of 6 x 10^6 trypanosomes/ml) was concentrated five-fold by spinning at 360 g for 10 mins in a falcon tube, removing four fifths of the supernatant, then gently resuspending the cells prior to glycerol addition. Stabilate tubes were left overnight at -80°C in a cotton-padded box to allow the freezing process to occur gradually. They were then transported in dry ice and transferred to liquid nitrogen.

2.1.2 Differentiation from bloodstream to procyclic stage

In differentiating bloodstream forms to procyclic forms, 50-100 µl of blood with trypanosome density at around 10^7 were added to 3 mls SDM79, in the presence or absence of 6 mM cis aconitate (Sigma), and trypanosomes were left to differentiate at 27°C.

2.2 Procedures related to growing trypanosomes *in vivo*

2.2.1 Host immunosuppression

Immunosuppression of ICR or BALB/c mice to allow trypanosome growth from stabilate was performed by cyclophosphamide treatment (250 mg/kg body weight, Sigma), 24 hr prior to trypanosome injection.

2.2.2 Trypanosome growth and collection

For the chronic mouse infection, 11 mice were each injected with ~10^6 trypanosomes (80 µl blood volume with parasitaemia of 1.3 x 10^7 trypanosomes/ml), derived from an immunosuppressed mouse infected four days previously with a TREU 927/4 GUTat 10.1 stabilate, kindly provided by Prof. M. Turner. Mouse blood was harvested by cardiac puncture into CBSS (Carter's Balanced Salt Solution (1 x): 0.023 M HEPES, 0.12 M NaCl, 5.41 mM KCl, 0.4 mM $MgSO_4$, 5.6 mM Na_2HPO_4, 0.035 M glucose, 0.05 mM phenol red, pH 7.4), containing 5% sodium citrate anticoagulant (0.15 ml CBSS per 0.85 ml blood).

2.2.3 Stabilate preparation

Stabilates were prepared by mixing blood 2:1 with 22.5% v/v DMSO (Dimethyl sulphoxide) in CBSS. Blood was injected into fine bore polythene tubing (Portex), which was then cut into "straws" and fitted into cryotubes, perforated to provide contact between straws and liquid nitrogen. The freezing of the stabilate then followed the procedure outlined in section 2.2.1.

2.3 Basic laboratory procedures

2.3.1 Parasite isolation and lysis

2.3.1.1 gDNA isolation

Trypanosomes were isolated from mouse blood (normally a volume of ~1ml, with a trypanosome density of ~10^7 cells/ml) by centrifugation at 1000 g for 10 mins, resulting in trypanosomes forming a white layer ("buffy coat" on top of lysed blood cells). The buffy coat was transferred to a clean eppendorf tube and trypanosomes were lysed by addition of 500 µl lysis buffer (50mM Tris pH 8, 1 mM EDTA, 100mM NaCl). 50 µl 10% SDS (Sodium dodecyl sulphate) and 2.5 µl proteinase K (at 20 mg/ml) were also added in order to denature the proteins present in the sample, which was incubated overnight at 37°C. In the case of trypanosomes in culture, cultures were spun at 1000 g for 10 mins, resulting in formation of a trypanosome pellet, which was resuspended into lysis buffer after removal of most of the supernatant. A maximum number of 10^8 cells were used for each individual lysis reaction.

2.3.1.2 Phenol/chloroform extraction of gDNA

In order to separate gDNA from other lysed cellular components (see section 2.4.1.1), the trypanosome lysate was mixed with an equal volume (normally ~600 µl) of 1:1 phenol/chlorophorm (Sigma). Mixing was followed by centrifugation at maximum speed for 10 mins resulting in separation of the acqueous and organic phases. The acqueous (top) phase was transferred to a clean eppendorf tube and two volumes of absolute ethanol (100%) were added to result in gDNA precipitation, aided by incubation of the sample at -20°C for at least one hour. The precipitate was then pelleted by centrifugation at maximum speed for 10 mins. The pellet was washed by addition of 1 ml 70% ethanol accompanied by a 1 min centrifugation; it was then allowed to air-dry and resuspended in 50-200 µl water. gDNA was then stored at –20°C.

2.3.2 RNA and cDNA preparation

2.3.2.1 RNA isolation

In the case of RNA isolation from mouse blood, 1 ml erythrocyte lysis buffer (ELB, Qiagen) was added to 200 µl of blood and the sample was then incubated on ice for 5 mins and then vortexed. The incubation and vortexing steps were repeated twice more, before centrifugation for 10 mins at 360 g. The supernatant was discarded and the pellet resuspended in 400 µl ELB, after which the centrifugation step was repeated. The pellet obtained was then resuspended in 350 µl RLT buffer (Qiagen RNeasy kit) plus 1% β-mercaptoethanol and RNA was isolated using the Qiagen RNeasy mini kit, following the protocol provided by the manufacturer. In the case of RNA isolation from culture, ~10^7 cells were centrifuged at 360 g for 10 mins and then resuspended in RLT buffer as above.

2.3.2.2 Reverse transcription (RT-PCR)

RNA (maximum 1 µg) was first treated with 1 µl DNaseI (Invitrogen), in a 10 µl reaction containing 1 µl 10 x DNase I reaction buffer and made up with water. The reaction was performed at room temperature for 15 mins, and stopped by addition of 1 µl 25 mM EDTA and incubation at 65°C for 10 mins. cDNA was prepared by reverse transcription using the Superscript II First-Strand Synthesis System for RT-PCR kit (Invitrogen). An RNA/primer mix was then prepared by addition of 2µl oligo[dT] primer, 2 µl 10 mM dNTPs and 5 µl DEPC (Diethyl pyrocarbonate) water, and incubation at 65°C for 5 mins. The 20 µl samples were then transferred on ice and split in two 10 µl aliquots, one used for reverse transcription and one as a control for gDNA contamination (no reverse transcriptase added). To each aliquot were added 9 µl of a reaction mixture containing 2 µl 10 RT buffer, 4µl 25mM $MgCl_2$, 2 µl 0.1 M DTT and 1 µl RNaseOUT Recombinant Ribonuclease inhibitor. The reactions were left at room temperature for 2 mins and then 1 µl of reverse transcriptase (SuperScript II RT) was added to the samples (with exclusion of the negative controls). The samples were then incubated at 25°C for 10 mins, 42°C for 50 mins and 70°C for 15 mins, then chilled on ice. Prior to PCR, 1 µl RNase H was added to each sample in order to remove RNA, with incubation at 37°C for 20 mins.

2.3.3 Gel electrophoresis

Separation of DNA by electrophoresis was achieved by loading on 1% agarose gels (Seakem LE, BMA) run at 100V in 1 x TAE buffer (40 mM Tris, 19 mM acetic acid, 1mM EDTA). Gels contained 0.2 μg/ml ethidium bromide (EtBr) to allow DNA visualisation under UV light. DNA size was estimated using a 1 kb ladder (Invitrogen) as a size standard.

2.3.4 Polymerase Chain Reaction (PCR)

In order to amplify *VSG* sequences from cDNA for cloning and sequencing, PCR products were generated by Herculase proofreading polymerase (Stratagene) to maximise sequence accuracy, using a forward primer annealing to the spliced leader sequence and a reverse primer recognising a conserved 16-mer (see Appendix, Tables 7.4 and 7.10, for primer sequences). Reactions were composed of 1-3 μl cDNA, 2 μl of each primer (5mM), 1 μl of 10mM dNTPs, 1 μl Herculase, 5 μl 10X Herculase buffer (Stratagene) and made up to 50 μl with water. Reaction conditions were: 5 mins at 95°C, followed by 30 cycles of 95°C for 1 min., 38°C for 2 mins, 72°C for 2 mins and a final extension of 5 mins at 72°C, using a Robocycler machine (Stratagene).

In the case of the other PCRs performed in Chapters 4 and 5, the polymerase used was Taq (*Thermus aquaticus*) DNA polymerase (ABGene). Reactions were composed of 1-3 μl cDNA/gDNA, 2 μl of each primer (1 μM final concentration), 0.2 μl Taq, 2 μl custom buffer mix (provided by Annette MacLeod, resulting in a PCRs containing 45 mM Tris-HCl pH 8.8, 11 mM $(NH_4)_2SO_4$, 4.5 mM $MgCl_2$, 6.7 mM 2-mercaptoethanol, 4.4 μM EDTA pH 8.0, 113 μg/ml BSA, 1 mM each of the four dNTPs) and made up to 20 μl with water. Reaction conditions were 30 cycles of 95°C for 50 s, 55-63°C for 50 s, and 65°C for 1 min per kb of expected product.

2.3.5 Cloning of PCR products and recombinant plasmid isolation

2.3.5.1 Cloning of PCR products using TOPO vector (Invitrogen)

Prior to cloning, PCR products generated by Herculase were subjected to a 20 minute Taq polymerase extension at 72°C, to add A residues to the ends, in order to provide cohesive ends for the T overhangs present in the vector. PCR products were TOPO-cloned directly

without any prior purification step (TOPO TA vector and cloning procedure developed by Invitrogen). 1-4 µl of PCR reaction was incubated with 1µl TOPO vector and 1 µl of salt solution (1.2 M NaCl, 0.06 M $MgCl_2$), and the reaction was made up to 6 µl with water. The length of incubation was of 10-15 minutes, following the manufacturer's advice for cloning a PCR product containing a mixture of different sequence, in order to maximise the diversity of cloned PCR products. 5 µl of the reaction were added to 50 µl of TOP 10 F' competent cells (Invitrogen), and these were then kept on ice for 10-25 minutes, prior to a 40 s heat shock at 42° C. After the heat shock, cells were again placed on ice and plated after 2 mins on pre-warmed L-agar plates containing 0.27 M ampicillin (Sigma), and incubated overnight at 37°C. Single colonies were then picked and used to inoculate 3 mls of L-broth (with added ampicillin at 10 mg/ml) and grown overnight at 37°C, for plasmid isolation.

2.3.5.2 Single colony lysis for screening colonies for recombinant plasmids

In order to screen for colonies containing plasmids with the correct insert size, 100 µl of each overnight culture were added to 35 µl of single colony lysis buffer (SCL; 1% SDS, 1.5% Ficoll, 1 x TAE, 1 x loading buffer, in water). This reaction was left to incubate at room temperature for 20 mins and then centrifuged at 14000 rpm at 4°C, using a benchtop centrifuge. 15 µl of supernatant were loaded directly onto a 1% agarose gel, allowing detection of clones containing a TOPO plasmid with a shift in size due to presence of insert. 1.5 ml of the overnight cultures testing positive by the procedure described above were then used to isolate recombinant plasmids, using a small-scale plasmid DNA purification kit ("miniprep", Qiagen).

2.3.5.3 Plasmid DNA digestion

Restriction enzyme digestion of recombinant TOPO vector to test for PCR product presence was conducted by using the *Eco*RI enzyme, present on either side of the site of insert integration. 1 µl miniprep plasmid DNA was routinely used, with 1 µl enzyme and 5 µl 10 x *Eco*RI reaction buffer, and the reaction was made up with water in a total volume of 50 µl. This was incubated at 37°C for 1-2 hrs.

2.4 Bioinformatics

2.4.1 VSG annotation

VSG arrays were identified by GeneDB automated blast searches. Further annotation was conducted using the Artemis software (Rutherford *et al.*, 2000). The software allowed identification of 70-bp repeats (low GC content) and then to run blastX (NCBI, www.ncbi.nlm.nih.gov/BLAST, low complexity filter disabled, *Trypanosoma* entrez query limit) of DNA sequences between 70-bp repeats (3kb average). This procedure allowed capturing the full *VSG* reading frame within the *VSG* cassette. The reading frames of *VSG* pseudogenes were established by analysing blast hits and then manually assemblying the full degenerate gene. The start and end of the putative *VSG* genes were checked for presence of signal peptide (SignalP 3.0, http://www.cbs.dtu.dk/services/SignalP/) and GPI anchor (by similarity with conserved VSG GPI signal sequences, then by big-PI predictor, http://mendel.imp.univie.ac.at/sat/gpi/gpi_server.html and DGPI http://129.194.185.165/dgpi/DGPI_demo_en.html). The protein sequence was then analysed for its N- and C- terminal domain types by looking at cysteine residue presence and spacing and performing further alignments against VSGs of known domain types where necessary. Where applicable, the dividing line between the two domains was set at 50 amino acids upstream of the first conserved cysteine in the C terminal domain. VSGs were classed as functional if they had a recognisable cysteine pattern and potentially functional signal peptide and GPI anchor, as atypical if any of the above requirements were not met, but the gene was not interrupted by frameshifts or stop codons, or as pseudogene if the sequence was either incomplete or degenerate.

2.4.2 Phylogenetic analysis

Multiple sequence alignments were conducted using ClustalX, or ClustalW at EBI (http://www.ebi.ac.uk/clustalw/) or at the Pasteur Institute (http://bioweb.pasteur.fr/seqanal/interfaces/clustalw.html), the alignment was then edited with Bioedit (http://www.mbio.ncsu.edu/BioEdit/bioedit.html) and then a second alignment with the edited sequences was performed. The tree produced by ClustalX was then visualised and coloured using the HyperTree software (Bingham and Sudarsanam, 2000).

CHAPTER 3

A BIOINFORMATIC ANALYSIS OF THE *VSG* ARCHIVE

3 A Bioinformatic Analysis of the *VSG* Archive

3.1 Introduction

The Variant Surface Glycoprotein (VSG) constitutes the major surface antigen of African trypanosomes, forming a dense coat on their surface. It is encoded in the genome as a very large gene family (~1000 copies), of which only one allele is expressed at any given time in bloodstream stage trypanosomes, the others remaining silent. VSG N-terminal domains are hypervariable (they normally share 20-25% identity) and part of their sequence forms the outer exposed coat surface, while the C-terminal domain is more conserved (40-50% identity) and forms the innermost part of the coat, adjacent to the plasma membrane. An N-terminal signal peptide and a C-terminal GPI anchor signal allow respectively targeting and attachment to the cell surface. Given the generally low homology, cysteine pattern conservation is the most useful criterion for identifying and classifying *VSG*s; based on it, three types of N-terminal domain (A-C) and four types of C-terminal domain (1-4) have been described (See Chapter 1, sections 1.2 and 1.3 for further details).

A bioinformatic analysis of the silent *VSG* copies in the genome of *Trypanosoma brucei* strain TREU 927/4 GUTat10.1 was conducted, covering all megabase chromosomes. Firstly, an overview will be given of the basic findings regarding the *VSG* archive in terms of its size and coding potential. In this context, the database created for the purpose of storing and retrieving the archive (VSGdb) will briefly be described, as it formed the basis for subsequent analyses. Secondly, a detailed analysis of *VSG* N-terminal and C-terminal domain types will be presented. This will then converge to provide a general picture of how domains relate to each other as encoded in full-length genes. Thirdly, the sequence environment in which *VSG* genes are found (subtelomeric arrays) will be described.

The results presented in this chapter have been achieved owing to the availability of the genome sequence for *Trypanosoma brucei* strain TREU 927/4 GUTat10.1, due to the joint sequencing initiative of The Institute for Genomic Research (TIGR) and The Wellcome Trust Sanger Institute. All data were merged and analysed through the *Trypanosoma brucei* GeneDB (http://www.genedb.org/genedb/tryp/index.jsp), a database created by the Sanger Institute Pathogen Sequencing Unit (PSU) with the aim to provide reliable access to the latest sequence data and their annotation and curation. Basic *VSG* annotation and curation were undertaken using Artemis, a software package developed at the Sanger Institute that enables visualisation and characterisation of large (chromosome-size) DNA

molecules (Rutherford *et al.*, 2000). Initial guidelines in *VSG* annotation were given by Dr. Mark Carrington (University of Cambridge), who had constructed a small, independent *VSG* database. Further developments were discussed and validated with him and with Drs Hubert Renauld and Christiane Hertz-Fowler, genome curators at the Sanger Institute. Some of the basic findings derived from work presented in this chapter were included in the antigenic variation section of the trypanosome genome paper (Berriman *et al.*, 2005).

3.2 Analysis of the *T. brucei* strain 927 *VSG* archive - overview

3.2.1 Estimation of VSG *archive size*

*VSG*s present in the larger assemblies of megabase chromosomes 1-11 as found in the GeneDB v4 release (July 2005) were analysed, and additional *VSG*s were annotated from the putative chr 8 homologue array and unordered contigs of chr 9, 10, 11, yielding a total of 940 genes. A contig (short for contiguation) represents the outcome of using bioinformatics to join sequencing reads, based on overlap between them. The ordering of contigs with respect to one another can be achieved either through sequence overlap, providing a physical map, or positioning of contigs in a genetic map by the use of specific markers, to provide a scaffold or assembly. Both approaches are problematic for *VSG* arrays, due to the repetitive nature of the arrays and the lack of genetic markers. Therefore, there is still considerable work left in order to complete sequencing and annotating *VSG* arrays and to distinguish between genes present on the two homologues of each chromosome.

In order to put the current analysis into perspective and be able to draw meaningful conclusions, an estimate of the total size of the *VSG* archive will first be attempted. Knowledge of chromosome sizes from pulsed field gels (Melville, 1997; Melville *et al.*, 1998) and of the extent of the core region of each chromosome from genome sequencing (C. Hertz-Fowler, pers. comm.) allows estimation of the core haploid genome size at 22 Mb; this leaves the value of the combined subtelomeric region for each chromosome and its homologue at around 9.5 Mb (Table 3.1).

Table 3.1: Estimate of subtelomeric chromosome size for *Trypanosoma brucei* strain TREU 927/4 (data analysis by Prof. J.D. Barry, pers. comm.).

PFG stands for Pulsed Field Gel (see text).

Chromosome number	927 chromosome size (PFG) (Mb)	core Mb (genome sequencing)	927 subtelomere size (Mb)
1A	1.10	0.80	0.30
1B	1.20	0.80	0.40
2A	1.30	0.90	0.40
2B	1.34	0.90	0.44
3A	1.63	1.46	0.18
3B	1.76	1.46	0.30
4A	1.65	1.39	0.26
4B	1.78	1.39	0.40
5A	1.70	1.29	0.41
5B	1.72	1.29	0.43
6A	1.75	1.22	0.53
6B	2.00	1.22	0.78
7A	2.17	2.15	0.02
7B	2.17	2.15	0.02
8A	2.10	2.35	-0.25
8B	2.17	2.35	-0.18
9A	3.33	2.16	1.18
9B	3.49	2.16	1.33
10A	4.39	3.85	0.54
10B	4.41	3.85	0.56
11A	5.22	4.47	0.75
11B	5.22	4.47	0.75
diploid	-	44.05	9.54
haploid	-	22.00	-
total	53.59	-	-

Around 2 Mb are likely to be taken up by 15-20 bloodstream expression sites (average length 50 kb) with a barren region upstream of up to 80 kb[20] (LaCount *et al.*, 2001), and allowing 0.5 Mb of telomeres (~10 kb at each end), 7.5 Mb of subtelomeres are left, likely to be occupied almost exclusively by *VSG* arrays. Assuming that all arrays are haploid (which appears to be the case, as discussed towards the end of the chapter), and given an average gene density of one *VSG* every 5 kb (observed in the current study), yields a value of around 1500 genes. Taking a less conservative approach and assuming that telomere tracts and expression sites combined contribute 1 Mb of subtelomeres, a value of 1700 is obtained (8.5 Mb of arrays). All these computations are subject to error, but it can be said it is highly unlikely that the silent archive would exceed 2000 genes. GeneDB currently contains around 4 Mb of subtelomeres in the main chromosome molecules and another 2.8

[20] Assuming that all BES are located on megabase chromosomes, which possibly is not the case. 100kb is the "arbitrary" average value taken for each BES plus barren region upstream, in making the 2Mb estimate.

Mb of *VSG*-containing unordered contigs (C. Hertz-Fowler, pers. comm.), giving around 7 Mb of subtelomeres. If the above estimate is correct, GeneDB houses currently most of the *VSG* arrays (with ~1250 *VSG*s annotated and more *VSG*s likely to be as yet unannotated). Taking a value of 1600 as a likely estimate, it therefore emerges that *VSG*s analysed in this study (940) represent between half and two thirds of the total *VSG* archive. The 927 genome strain appears to have the shortest chromosomes amongst *T. brucei* strains, and this is thought to be due to changes in *VSG* array length: if this is correct, the 427 strain has 2.5 times the array length of 927, corresponding to around 3000 *VSG* genes. Whether an "optimal" archive size is selected for in the wild or whether archive size fluctuations are very common, this is currently not known.

3.2.2 Annotation

In most cases annotation of *VSG*s has been reliant on a first pass automated gene "detection" conducted by the sequencing centres. This was then refined to encompass the full extent of genes or fragments thereof. Specific features were used to capture the characteristics of each encoded protein, in terms of signal peptide, N-terminal and C-terminal domain and GPI signal. Whereas the signal peptide was always based on prediction by algorithm (SignalP 3.0, http://www.cbs.dtu.dk/services/SignalP/), the other three features were curated manually on the basis of sequence homology by using blastX and blastP at NCBI. In the case of putative functional GPI signal anchors, the prediction "by homology" was additionally verified by specific prediction programs (big-PI, http://mendel.imp.ac.at/sat/gpi/gpi_server.html, and DGPI, http://129.194.185.165/dgpi/DGPI_demo_en.html). N-glycosylation prediction was not performed routinely, being conducted only on putative functional VSGs: whereas the amino acid motif allowing N-glycosylation is conserved, there is less of a consensus with regards to the stringency of N-glycosylation requirements, which seem to be based on more complex structural observations, implying knowledge of tertiary structure (see Section 1.2.2 of Introduction). There will therefore be no further mention of this aspect of VSG modification: suffice it to say that *VSG*s classed as functional in this study seem to have the same range of variation in N-glycosylation as that found in expressed VSGs characterised to date (data not shown). Study of detailed tertiary structure of proteins encoded by candidate functional genes could enable predictions to be made in terms of the likelihood of them being expressed and could possibly question the "intactness" of some sequences, but such analysis is beyond the scope of the work presented here.

A total of 940 *VSG*s were analysed and they can be divided into four categories: functional, atypical, pseudogene and incomplete. Whereas "functional" *VSG*s meet all parameters for the features described above and are likely to be expressed, what have been termed "atypical" *VSG*s encode a complete open reading frame that has potential problems associated with folding (cysteine pattern incomplete) and targeting or anchoring (alterations in GPI signal or N-terminal signal peptide sequence); this makes the prediction that these *VSG*s could be expressed uncertain. As for the final two categories, "pseudogenes" have been defined as full-length coding sequences interrupted by frameshifts and/or stop codons and "fragments" are shorter than a complete *VSG* coding sequence.

The striking observation made upon completing the analysis of the *VSG* archive was that the great majority of the 940 silent *VSG* copies analysed are pseudogenes (611, 65%) or fragments (197, 21%), only a small percentage being functional (43, 4.5%) or atypical (89, 9.5%) (see Figure 3.1). This raises questions as to how this "defective" archive contributes to antigenic variation, an issue that will be discussed once other important aspects of the data have been presented.

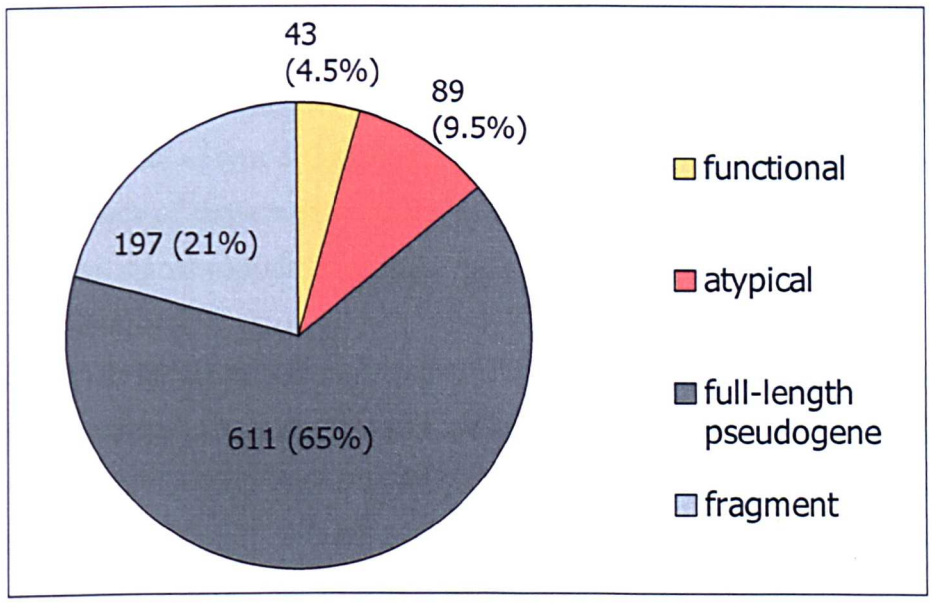


Figure 3.1. Basic features of the *T. brucei* strain TREU 927 *VSG* archive (940 genes).

The dividing line between "functional" and "atypical" *VSG*s is very fine, especially as in most cases (67 "atypicals" out of 89, 75%) it comes down to minute differences in GPI anchor signal sequence (Table 7.3 of the Appendix gives all atypical *VSG*s and their associated features). These differences can be broken down into changes at the cleavage site (38 instances), presence of polar residues in the hydrophobic extension (22), or the hydrophobic extension being longer or shorter than the consensus sequence (28). The latter can often be explained by the presence of a frameshift within the GPI signal sequence, extending the hydrophobic region or making it unusually short. Whereas the predictions for longer anchor sequences or sequences with alterations at the cleavage site seem to be unfavourable, there are borderline cases in which there are only a few substitutions from the GPI signal consensus obtained from expressed *VSG*s, making a neat classification into either functional or atypical very difficult. It is expected that an increased knowledge of GPI signal sequence requirements and a larger sample of expressed *VSG*s will shed light on this currently grey area. This will mean that some of the currently functional *VSG*s might be classed as atypical or *vice versa*, depending on how stringent are the requirements for stable *in vivo* expression. Other *VSG*s (9) were classed as atypical due solely to missing cysteines in the C-terminal domain[21], and another 9 because they encode a novel C-terminal domain (type 5, described later in this chapter). Lack of experimental evidence for processing of this domain type within a mature polypeptide makes it difficult to define with confidence whether or not the anchor signal sequences of these silent copies are functional or not (see Figure 7.1 of the Appendix for an alignment of type 5 C-terminal domains). Only four *VSG*s have been classed as atypical due to lack of conserved cysteines in the N-terminal domain or because of weak N-terminal signal peptide prediction.

Just as there is a grey area between functional and atypical *VSG*s, it could be argued that many atypical *VSG*s are "not fit for expression" and therefore could follow under a broader definition of pseudogene, namely as a gene that is not able to form a functional expressed protein. *VSG* fragments will be discussed in more detail towards the end of this chapter, as the architecture of *VSG* arrays is described.

[21] A total of 30 atypical VSGs have a non-conserved cysteine pattern in the C-terminal domain.

3.2.3 VSGdb: a tool for VSG domain analysis

The overview results presented above were obtained in the first-pass analysis of the data: as chromosomes were scanned for *VSG* arrays and *VSG*s were analysed using Artemis, a table was compiled giving details of each *VSG* in terms of coding sequence features (sequence intactness, domain types), gene location, associated non-*VSG* genes and repeats (see section 3.7 for further details on *VSG* arrays). From this initial "catalogue" (presented in Table 7.2 of the Appendix) the analysis was greatly aided by the development of *VSG*db (available online at www.*VSG*db.org). This flat-file based database was developed using BioPerl in collaboration with bioinformaticians Pauline Ward and Suraj Menon and is currently maintained by Jon Wilkes, Wellcome Centre bioinformatician. At the core of the database are EMBL files containing the sequence and *VSG* annotation for each individual chromosome or contig; it is likely that *VSG*db would need to be converted to a relational (MySQL) database if more data are to be added, as the current setup, while handling the data effectively, has limitations in terms of speed and complexity of queries.
The database allows access to all annotated *VSG* sequences and retrieval of sequences for particular groups of *VSG*s (by domain, domain combination, chromosome, etc.), blast-searching a *VSG* sequence against known *VSG*s, separate retrieval of N-terminal and C-terminal domains. Each *VSG* also has its own page with details of all associated features, and all *VSG*s have GeneDB identifiers, allowing cross-talk with GeneDB.
Apart from providing open access to the community, making the database a useful tool for researchers wanting to explore the *VSG* family with a specifically designed and flexible tool, *VSG*db allowed cross-checking the data, amending the annotation and also analysing more closely each N-terminal and C-terminal domain type and how they relate to each other. In short, domain types were retrieved separately, and the assignment to functional, atypical and pseudogene categories and to specific domain types was verified. This was followed by several rounds of multiple sequence alignment (ClustalX), together with alignment editing (Bioedit, http://www.mbio.ncsu.edu/BioEdit/bioedit.html) in order to remove indels (large insertions and deletions) and sequences that were too diverged (as judged from tree branch lengths and cysteine conservation). The data could then be compared at different levels using domain type subgroups or different domain combinations.

The following paragraphs present the results obtained, starting from N-terminal domains, moving then to C-terminal domains, then to integrate the data and look at full length coding sequences: as each component was dissected, a clearer appreciation of the

sequences available for antigenic variation was gained. *VSG* sequence diversity, divergence and evolution could be closely assessed, based on intradomain type diversification and interdomain type interaction.

3.3 N-terminal domains

In the 771 full-length N-terminal domains, types A and B are the most abundant, each amounting to about half the total number, with type C forming a small cluster within type A (see global N-terminal domain alignment below, Figure 3.2). No novel N-terminal domain types were found. For each N-terminal domain type, about one in three domains are functional (see Figure 3.3) suggesting, as will become more apparent in section 3.4, that most of the degeneracy of the *VSG* archive is due to defective C-terminal domains.

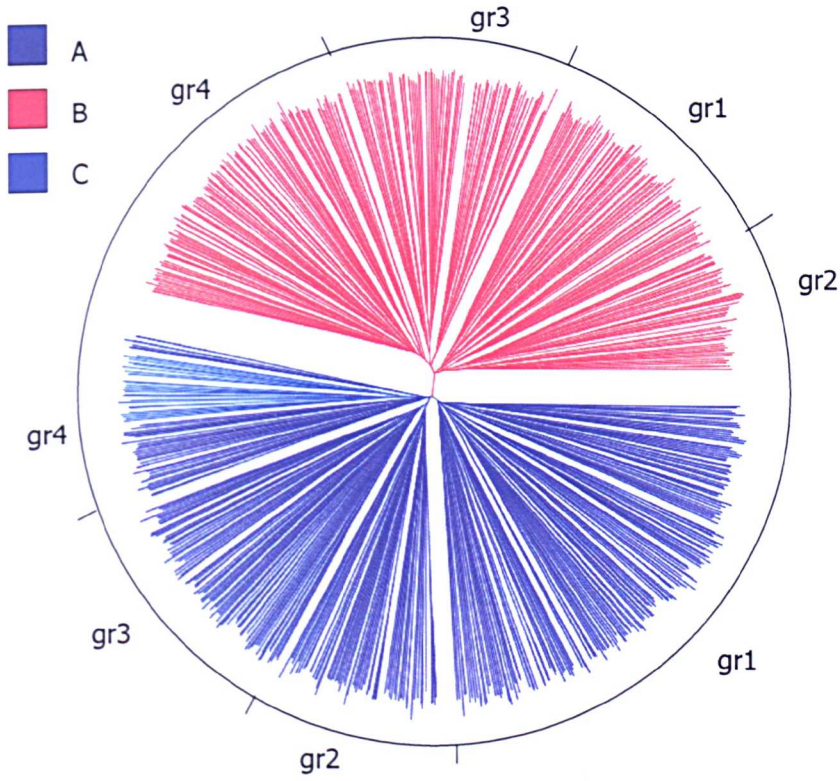


Figure 3.2: Tree derived from amino acid alignment of 725 *VSG* N-terminal domains.

The tree and multiple sequence alignment were both generated by ClustalX (Thompson *et al.*, 1997). The tree was visualised and the branches coloured using HyperTree (Bingham and Sudarsanam, 2000). The three N-terminal domain types (A, B and C) were coloured in dark blue, red and light blue, respectively. Gr1 to gr4 in both type A and type B domains represent groups 1 to 4, the significance of which is explained in Figures 3.4 and 3.5.

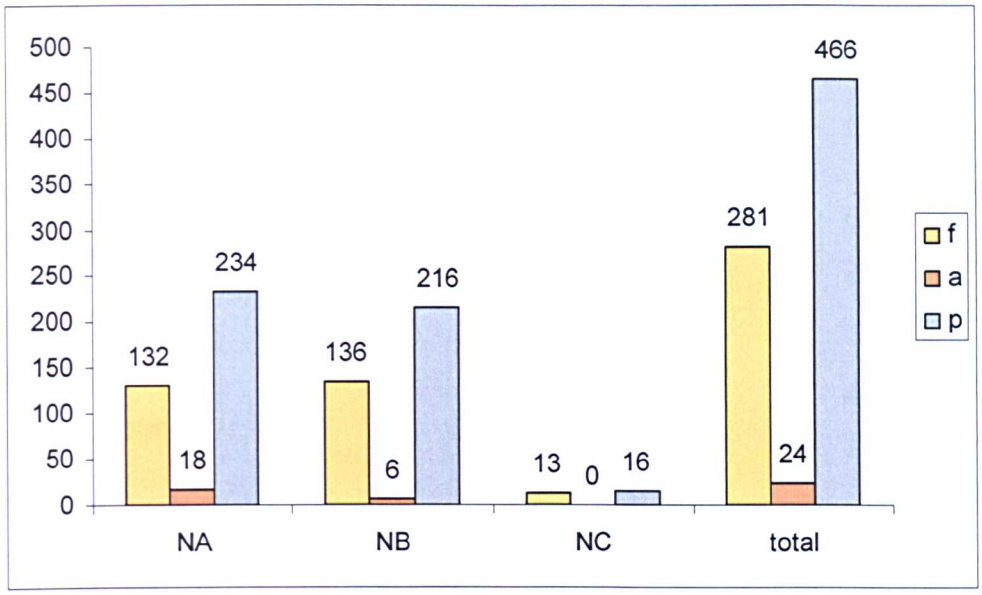


Figure 3.3: Abundance and basic characteristics of the different *VSG* N-terminal domain types (A-C).

The y axis indicates number of domains; NA, NB, NC indicate N-terminal domains A, B and C, respectively; the three columns indicate functional (f), atypical (a) and pseudogene (p) domains. The number of domains in each category is given above the columns.

3.3.1 Types A and C

A more detailed analysis of N-terminal domain types A and C involved analysis of 338 type A and 29 type C, totalling 377 sequences. Alignment of these N-terminal domains indicated several distinct clusters (see Figure 3.2, giving a summary tree for all N-terminal domains and their subgroups), so the data were analysed further and broken down into smaller subsets, which were then edited to reduce uninformative indels and optimise the alignment. The editing allowed a closer look at potential differences between groups, to explain their distinctness, and this appears to be due to significant variations in the cysteine pattern. Four main groups (corresponding to four main cysteine patterns) can be identified (see Figure 3.4): group 1 represents the most common pattern type (44%), the one for which tertiary structure has been solved experimentally (Freymann *et al.*, 1990). Groups 2, 3 and 4 reveal the "hidden" diversity of type A domains, of which type C can be considered part, its cysteine pattern being closely related to that of type A domain group 4.

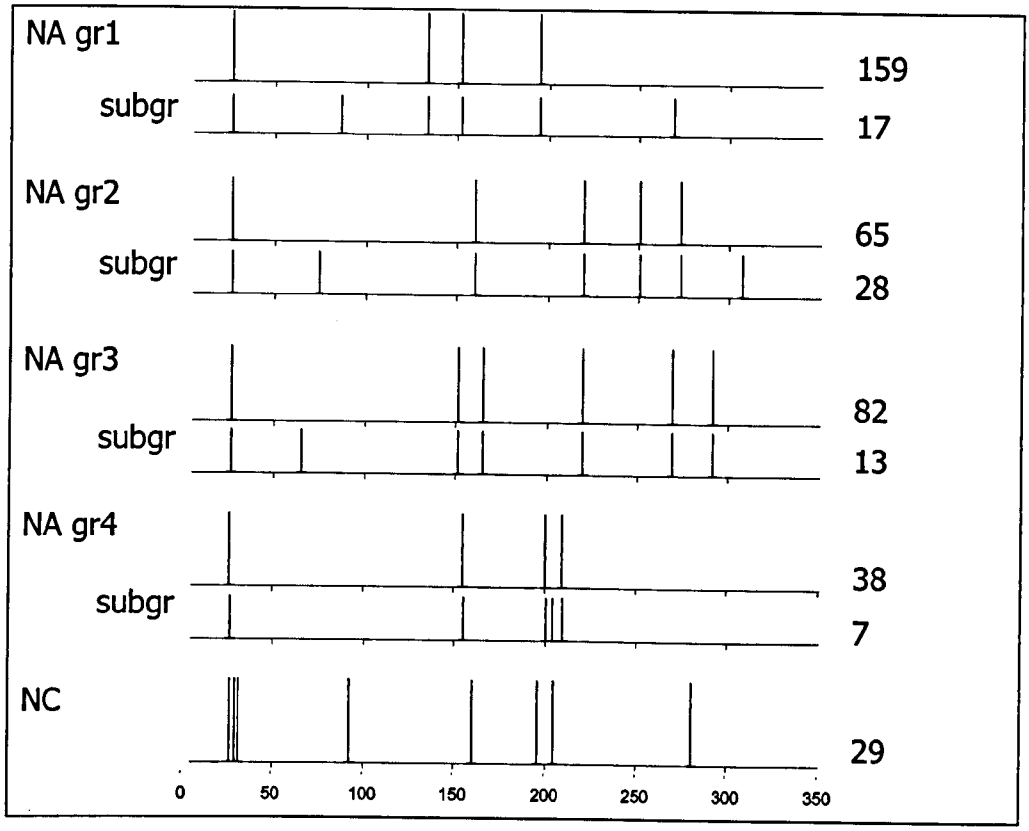


Figure 3.4: Cysteine pattern of type A N-terminal domain groups and subgroups. The number to the right indicates the number of domains in each group. Gr strands for group and subgr for subgroup.

It was of interest to look at secondary structure prediction for these four groups, to appreciate whether they were marked also by variation in alpha helical arrangement. While it was noted that the length of helices A and B (at the beginning of the domain, see Chapter 1, Figures 1.2 and 1.3) was greater in groups 2-4 than in group 1, possibly correlating with the wider spacing of cysteines, there did not appear to be any other significant group-specific difference (data not shown).

3.3.2 Type B

As for N-terminal domain type B, 335 sequences were analysed and groups could be identified from the tree and multiple sequence alignment (see Figure 3.2); however, unlike in the type A analysis, the differences between the groupings are not striking in terms of cysteine pattern. The main difference between groups appears to be the slight variation in distance between conserved cysteines, although all groups have the same cysteine pattern (see Figure 3.5).

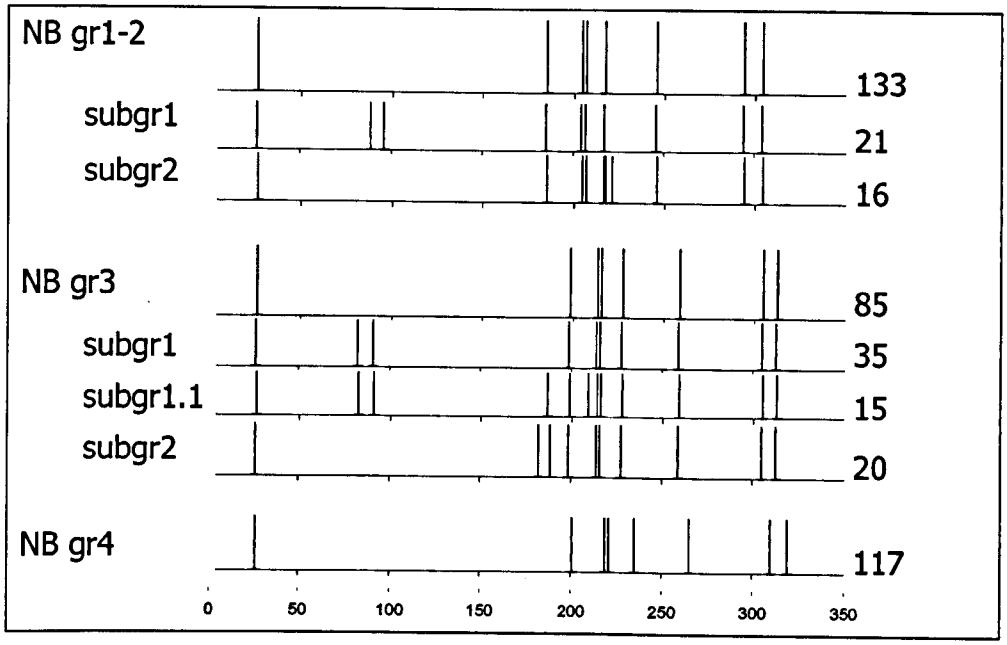


Figure 3.5: Cysteine pattern of type B N-terminal domain groups and subgroups. The number to the right indicates the number of domains in each group. Gr strands for group and subgr for subgroup. cysteine pattern of N-terminal domain type B subgroups.

When producing a tree with either the first or last 100 amino acids of all N-terminal domains (these regions were chosen because of smaller differences in cysteine pattern between types A and B), clear separation between sequences encoding the two domain types was maintained, suggesting a lack of interdomain short-range recombination, likely to be due to excessive sequence divergence (data not shown).

3.4 C-terminal domains

In the analysis of *VSG* C-terminal domains, more diversity and complexity was encountered than for the N-terminal domains. A C-terminal domain can be described by a combination of features: the presence of one or two 4-cysteine subdomains, the spacing of cysteines within the subdomains, and the type of GPI signal. The table below proposes a classification of C-terminal domains based upon these features. In addition to domain types 1 to 4, two other domain types were found to be encoded in the *VSG* archive, and were named 5 and 6.

Table 3.2: C-terminal domain type proposed classification.

domain type based on last four cysteines and GPI signal	1 cysteine subdomain (4-cys)		2 cysteine subdomains (8-cys)	
	Domain type	Number of domains	Domain type	Number of domains
1	-	0	C1	235
2	C2	186	-	0
3	-	0	C3	152
4	C4	10	C6	38
5	C5a	28	C5b	2

C-terminal domain type 1 is the most abundant (235), followed by type 2 (186) then type 3 (152). Types 4, 5 and 6 are less common (10, 30 and 38 respectively). In contrast to the N-terminal domain, there are only 13% (85/651) functional domains. A somewhat less degenerate domain is type 2, with 23% (43/186) being functional, whereas only 7% of type 1 are predicted to be functional, the values for the others lying somewhere between (see Fig 3.6 for details). The difference between the level of degeneracy of type 1 and type 2 could be due to the fact that type 2 is shorter (only four cysteines) and therefore less likely to accumulate deleterious mutations. An additional explanation would be that the abundance of type 1 domains makes them statistically more prone to recombination than other domains, leading to the spreading and amplification of errors due to inaccuracies in the recombination reactions.

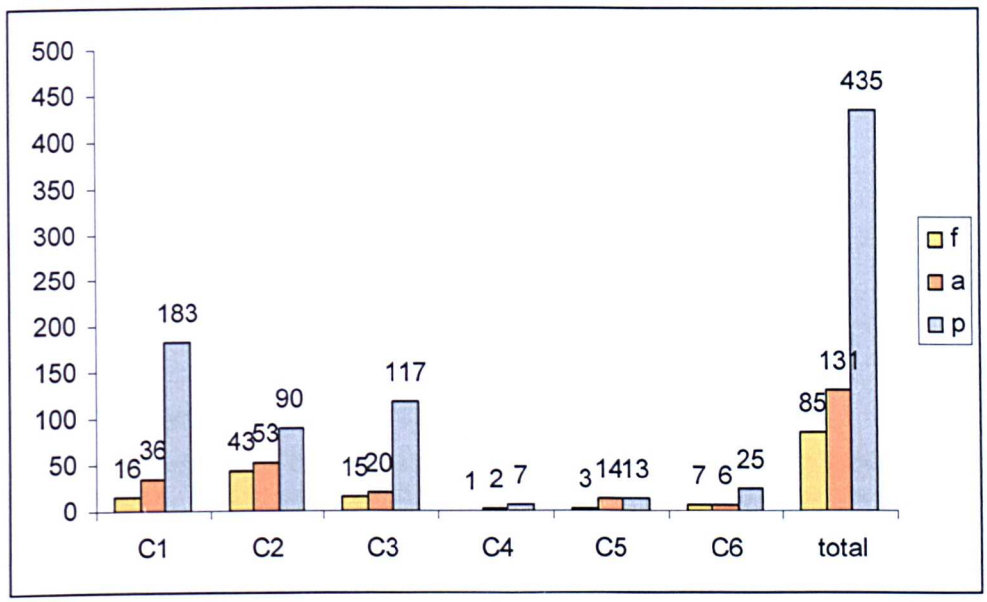


Figure 3.6: Abundance and basic characteristics of the different *VSG* C-terminal domain types (1-6).

The y axis indicates number of domains. C1 through to C6 indicate C-terminal domains 1 to 6. The three columns indicate functional (f), atypical (a) and pseudogene (p) domains. The number of domains in each category is given above the columns.

Analysis of internal variability of cysteine pattern in each individual domain failed to reveal discrete subgroups, in contrast to the case of N-terminal domains. The subgroups presented in Figure 3.7 are a small subset of similar sequences taken from a very heterogenous pool rather than representing neat and comprehensive subdivisions, as found in N-terminal domain subgroups. It was of interest to note that some of the first 4-cysteine subdomains were shared amongst different domain types (subdomains underlined in blue in Figure 3.7), whereas the last 4-cysteine subdomain tended to be the unique signature for the domain type (apart from C4 and C6, which share one, as underlined in red in Figure 3.7).

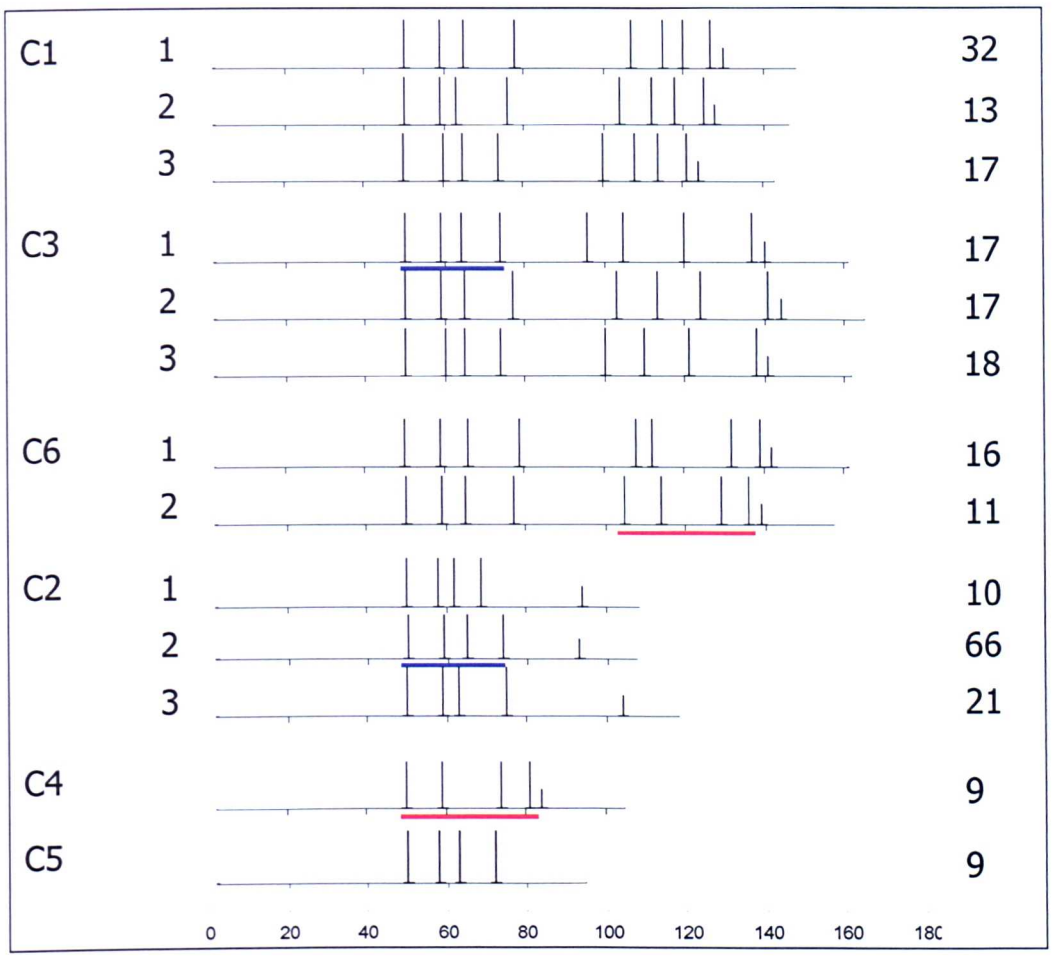


Figure 3.7: Cysteine pattern of C-terminal domains, types 1-6.

Cysteine subdomains that are shared between different domain types are highlighted in the same colour. Two to three examples of domain type subgroups are given for the main domain types, with the number of domains within each subgroup given on the left. The half height bar towards the end of the domain indicates the start of the GPI anchor signal (not determined for type 5).

It then became of interest to investigate the relationship between the 4- and 8-cysteine domain types and how different 4-cysteine subdomains relate to each other. This was done by producing and analysing C-terminal domain trees. Of the 651 domains that were initially used for the trees (many others were omitted because their excessive degeneracy rendered them unassignable to a domain type), only 541 were used in the analysis presented below (Figure 3-8). The trees include only domains in which allocation to a type was unambiguous and the cysteine subdomain pattern was either intact or, if degenerate, at least similar to that of other intact ones. Two additional trees were produced, in which the first and second 4-cysteine subdomain of 8-cysteine domains (types 1, 3, 6) were taken separately and each compared with the core cysteine-containing part of 4-cysteine domains (types 2, 4 and 5).

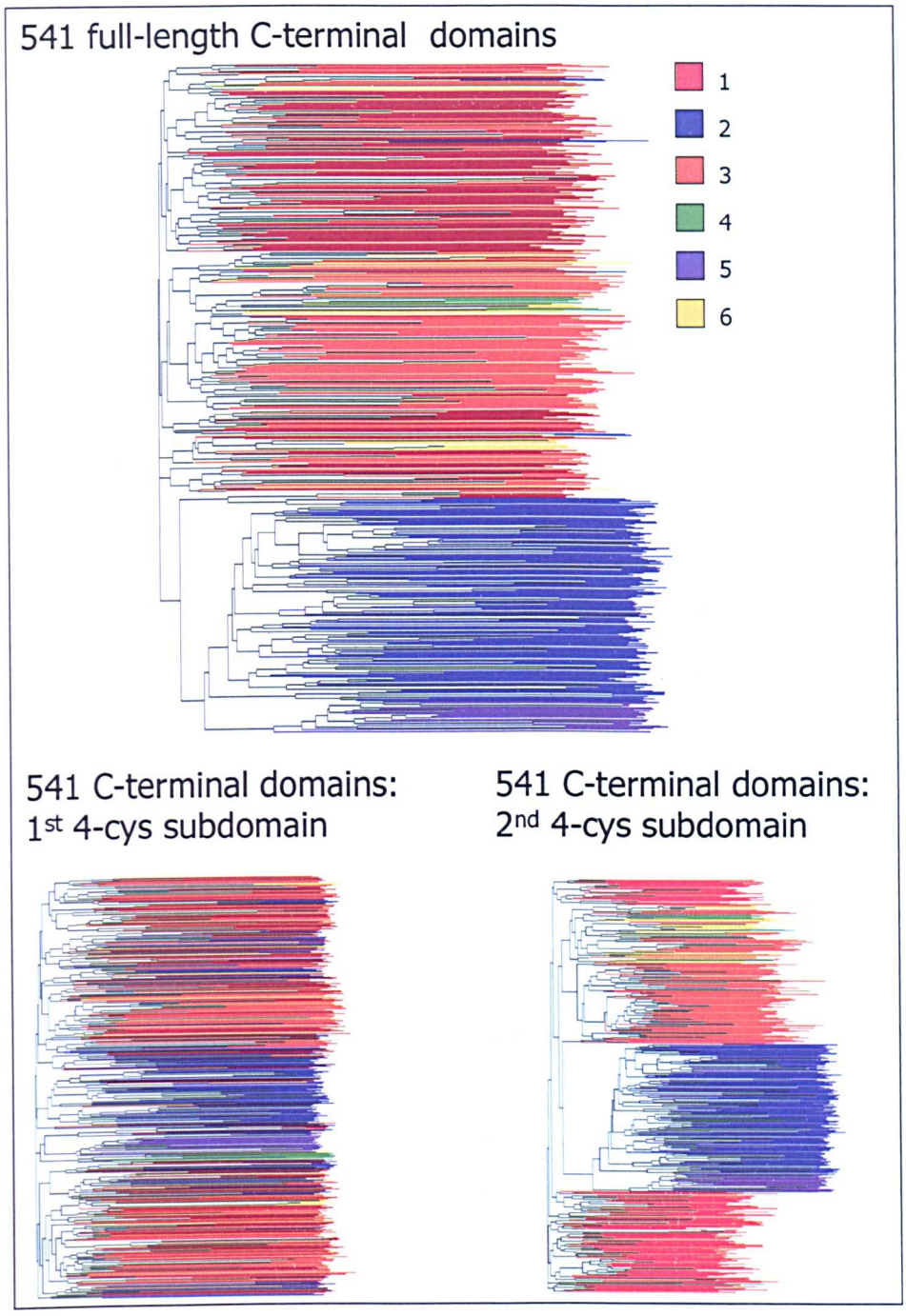


Figure 3.8: Trees derived from amino acid alignment of 541 *VSG* C-terminal domains.

The tree and multiple sequence alignment were both generated by ClustalX (Thompson *et al.*, 1997). The tree was visualised and the branches coloured using HyperTree (Bingham and Sudarsanam, 2000). Above: global tree comparing full-length sequence of all domains. Below to the left: tree comparing the first 4-cysteine subdomain of types 1, 3, 6 to the 4-cysteine subdomains of type 2, 4 and 5. Below to the right: tree comparing the second 4-cysteine subdomain of types 1, 3, 6 to the four cysteines of type 2, 4 and 5.

Figure 3.8 suggests that there is promiscuous mixing of different C-terminal domain types in the N-terminal domain-proximal region, corresponding to the first four cysteines of the 8-cysteine domain types. Domains seem not to mix so extensively towards the C-terminal end of the gene, where they show more specific features (last four cysteines and GPI signal). This confirms what was suggested from analysis of cysteine pattern of domain type subgroups and provides the context in which the documented cases of recombination between sequences encoding the first four cysteines of a C-terminal domain type 1 and a type 2 C-terminal domain have been shown to occur (Hutchinson *et al.*, 2003). A model of recombination between C-terminal domain types can be proposed based on these analyses (see Figure 3.9).

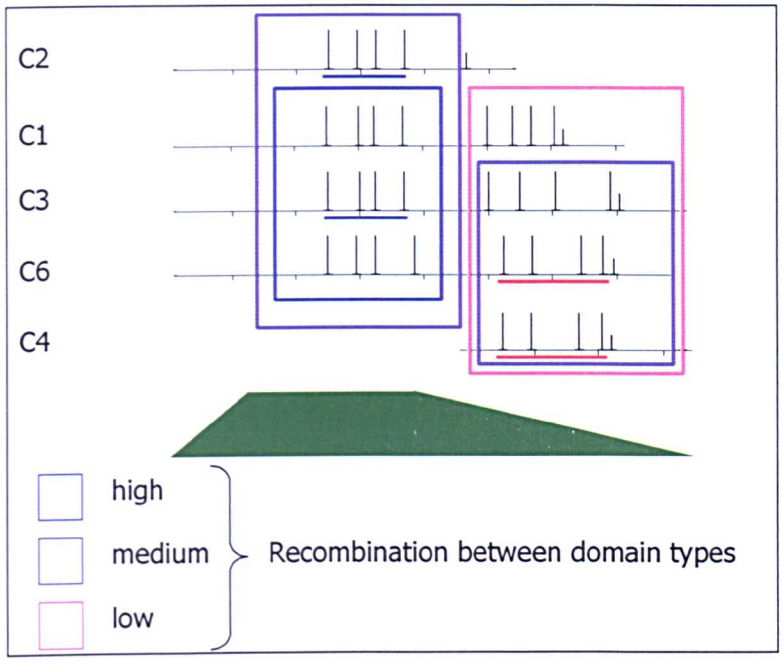


Figure 3.9: Model of recombination between C-terminal domain types.

The height of the green area below the domain depictions is proportional to the proposed level of interdomain recombination, increasing as it approaches the first set of four cysteines, and decreasing in the second part of the domain. Groups of four cysteines underlined in blue or red indicate identical spacing of cysteines.

3.5 Full-length *VSGs*

Now that both N- and C-terminal domains have been described and analysed, it is of interest to observe how they relate to each other. In order to do this, all domain combinations in full-length *VSG* genes were considered and the expected and observed values of these combinations were calculated. These are presented in Table 3.3; they match an older set of data (produced in July 2005), but are nevertheless informative, as the set of data was similar in size and composition to the final one presented at the beginning of the chapter, so essentially the same results would be obtained by repeating the analysis.

Table 3.3: *VSG* N- and C- terminal domain combinations: expected and observed values. Underlined in green are the main preferential domain combinations (more than 10% difference between observed and expected valued).

		N domain					
		A		B		C	
		Obs[1]	Exp[2]	Obs	Exp	Obs	Exp
C domain	1	101 (28)	130	149 (49)	112	3 (12)	9.5
	2	143 (40)	91	19 (6.2)	78	11 (42)	6.6
	3	43 (12)	76	101 (33)	65	3 (12)	5.5
	4	8 (2.3)	5.1	1 (0.3)	4.4	1 (3.8)	0.4
	5	27 (7.6)	16	0 (0)	14	4 (15)	1.2
	6	9 (2.5)	17	24 (7.8)	15	1 (3.8)	1.3
	other	24 (6.8)	20	12 (3.9)	17	3 (12)	1.5
	total		355		306		26
		355 (100)		306 (100)		26 (100)	

[1]Each cell shows: frequency and (percentage of total for that domain type)
[2]Expected values were calculated as: (observed frequency proportion of each N-terminal domain type)/(observed frequency of individual C-terminal domain type)

There appears to be a bias in domain combination, with A2, B1, B3 and B6 being the most common. This could suggest that a single surface molecule diverged into separate families (A and B), possibly acquiring different functions, and that later the formation of a large repertoire of protective coat proteins was achieved by expansion and mixing of these families, which still show traces of their long separation. This mixing of N-terminal domains of one type with C-terminal domains associated with a different N-terminal domain type is likely to have occurred because of the shared four-cysteine motif in the C-terminal domain (see Figure 3.9), acting as a bridge between diverged hypervariable N-terminal domains. How both N-terminal domain families acquired related C-terminal domains is as yet beyond explanation, although comparison with *T. congolense* might shed some light, because of the lack of a C-terminal domain in the VSGs of this parasite. A

preliminary analysis of *T. congolense VSG* emerging from the *T. congolense* genome project will be given in Chapter 5, section 5.4.1.

As might be expected, a global tree for full-length *VSG*s (Figure 3.10, obtained from the July 05 data, see beginning of section 3.5) discriminates primarily between N-terminal domains, whereas, within the N-terminal domain groupings, the specific C-terminal domain does not form a neat subgroup, again mirroring the above-mentioned overlap between C-terminal domains.

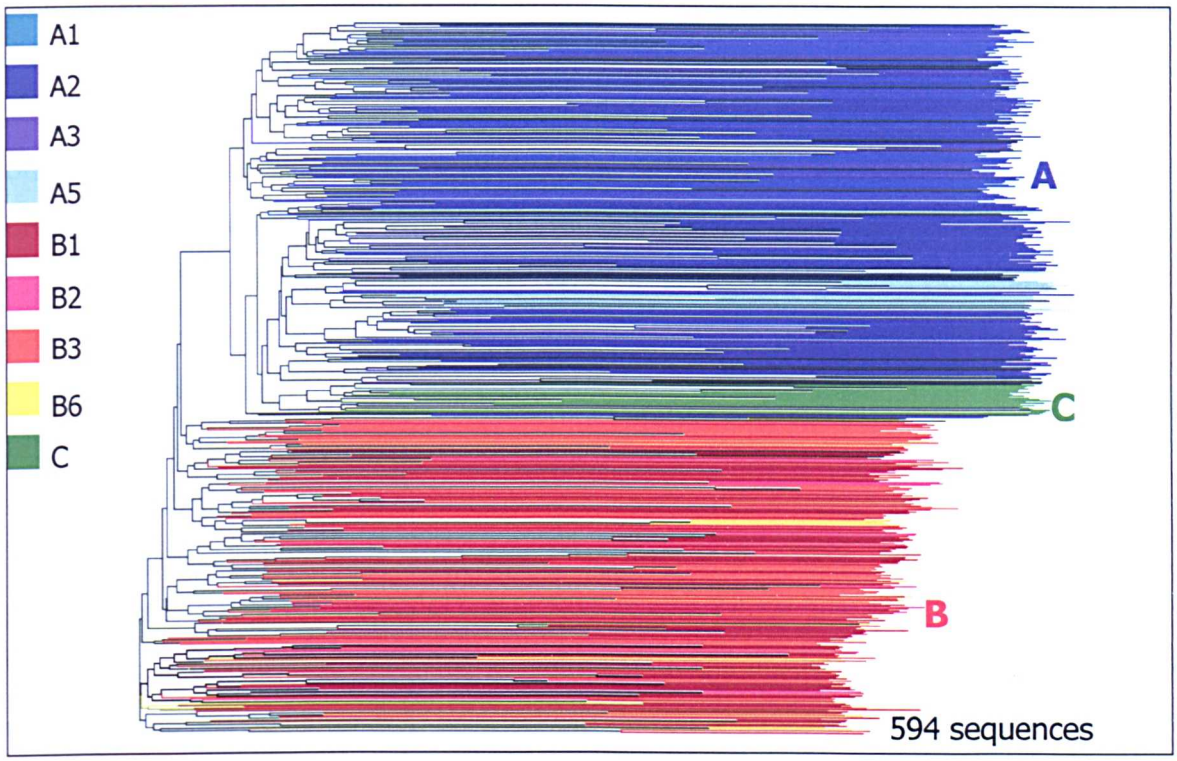


Figure 3.10: Tree derived from amino acid alignment of 594 full-length *VSG*s.

The tree and multiple sequence alignment were both generated by ClustalX (Thompson *et al.*, 1997). The tree was visualised and the branches coloured using HyperTree (Bingham and Sudarsanam, 2000). The July 2005 dataset was used, see text for more details. Domain combinations are highlighted in different colours, apart from VSGs with N-terminal domain type C, which were grouped together because of their small number.

3.6 *VSG* families?

At this point it is important to address the question of whether, within the broad *VSG* domain types and subgroups, there is further substructuring that could allow definition of smaller *VSG* families, or whether there is a continuum of variation. Although the long branches in the trees suggest the latter is correct, a quantitative estimate of this divergence will be attempted, by analysing the pairwise score generated by the multiple sequence alignment of N-terminal domains (http://www.ebi.ac.uk/clustalw/). This was done for N-terminal domain type A[22], with the scores being converted to percentage identity and the proportion of different identity ranges being plotted (Figure 3.11). Between the 361 A type N-terminal domains considered, 64980 pairwise comparisons were generated. Out of these, only 90 show an amino acid sequence identity above 60%, that is to say that only one in two of these domains would have a close orthologue[23]. Sequence identity drops dramatically between *VSGs,* with a 95% probability of any alignment between *VSGs* to result in less than 25% sequence identity, down to 18% (see Figure 3.11). Below 25% identity, errors in aligning sequences increase dramatically (at around 20% sequence identity, misalignment has been estimated to occur with a frequency of one in two). In turn, the presence of alignment errors increases the probability of inferring mistaken relationships between protein sequences, with a maximum accuracy of 75% when the identity between sequences is lower than 25% (Vogt *et al.*, 1995). When looking at low scores below 18 (corresponding to below 25% identity) there was not a linear decrease in identity between domains, but rather a fluctuation in alignment values between 18 and 25% identity. The fact that most *VSG* comparisons lie below the 25% "twilight zone" (Vogt *et al.*, 1995) results in a very high level of noise in the data. The *VSG* groups defined earlier in this chapter are held together by pairwise alignments in the 25-35% range, which help to order the "noise" below 25%. This means that within each *VSG* group there will be *VSGs* with as low as 18% sequence identity, which implies that groups cannot be defined simply in terms of sequence homology, but rely on more complex relationships between multiple pairs of *VSGs*. The implications of this highly diverged *VSG* archive will be discussed in Chapter 6.

[22] The same analysis was conducted with type B N-terminal domains, giving similar results (data not shown).

[23] Assuming that each of the 90 pairwise alignments corresponds to 180 unique domains, which is reasonably accurate, as there are very few domains sharing more than one close orthologue (data not shown).

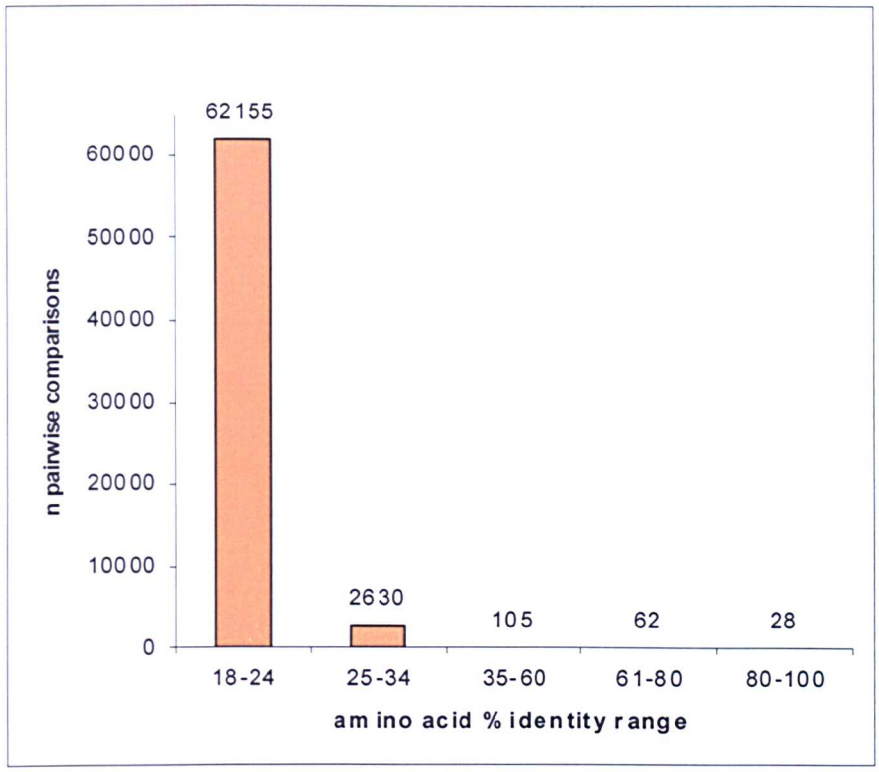


Figure 3.11: Pairwise scores from a multiple sequence alignment of 361 type A *VSG* N-terminal domains (64980 comparisons).

Each range of % amino acid sequence identity (given on legend to the right) has been associated with the number of pairwise alignments giving score values within this range. The absolute number of each group of alignments is given above each bar.

3.7 *VSG* arrays: structure and features

A major finding arising from the genome sequencing and annotation project is that the great majority of *VSG*s appears to be located in distinct subtelomeric arrays rather than "internally" in the genome, as was previously thought (Van der Ploeg *et al.*, 1982b) (see Figure 3.12). The arrays appear to be the only context in which *VSG*s are found in the genome, only one *VSG*-like gene out of 940 being located outside them (and it has unusual features, see Chapter 5, section 5.2.7). A separate class of genes, related to type B N-terminal domain-encoding genes but with many similarities to *T. congolense VSG*s, which are shorter than those of *T. brucei*, has been found. They reside outside the arrays, and have been termed "*VSG*-related" genes. Chapter 5 will be devoted to the analysis of this gene family, drawing comparisons and showing differences to *T. brucei* and *T. congolense VSG*s.

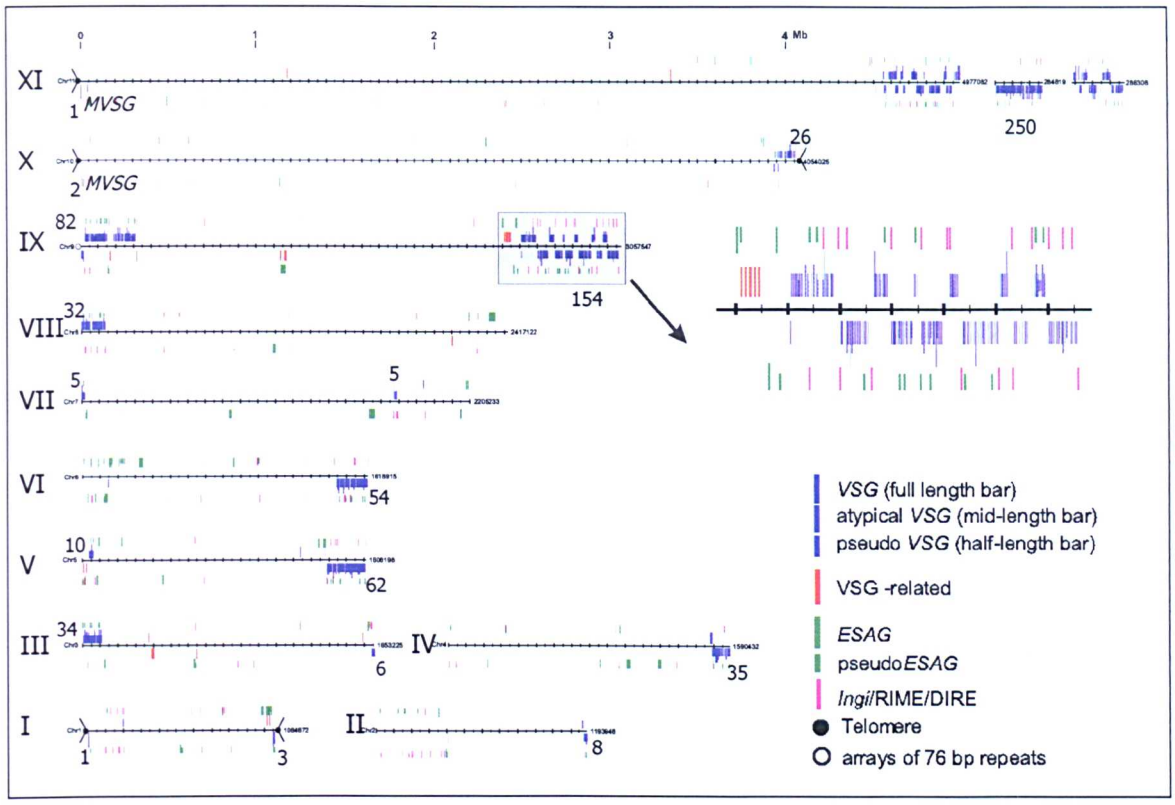


Figure 3.12: Chromosome maps for *T. brucei* 927 with *VSG* arrays highlighted, modified from Berriman (2005).

Roman numerals indicate the megabase chromosomes 1-11 in order not to confuse them with arabic numerals indicating the number of *VSG*s present in each array (correct as of July 2005). For up to date maps see www.genedb.org (arabic numerals normally indicate chromosome numbers in this context).

The basic unit in a *VSG* array has been termed the *VSG* cassette: it is delimited upstream by 70-bp repeats and downstream by the 3' end of the *VSG*, extending from within the end of the coding sequence to conserved elements immediately downstream, and is around 3-4 kb long (see Chapter 1 section 1.3.4 for more details and references, and Figure 3.13). Some 92% of full-length *VSG*s (687/743) are associated with at least one 70-bp repeat upstream. By analysing variation in 70-bp repeat numbers upstream of 5 array *VSG* genes it was suggested that differences amongs 70-bp might impose an order of expression amongst array *VSG*s (Aline *et al.*, 1985). The present bioinformatic analysis reveals a less extensive diversity amongst 70-bp repeats: 80% of array genes with 70-bp repeats have only one copy of the repeat, 15% have two copies and only 5% have between 3 and 15 copies, with no association between large number of repeats and putative functional *VSG*s. In addition, it appears that the level of degeneracy amongst 70-bp repeats is greater than what has been shown so far, with frequent departures from the consensus sequence, especially in the case of section 3 of the repeat (see Introduction, section 1.3.4), which sometimes is absent or incomplete (data not shown).

A novel feature found to be associated to *VSG* cassettes is the presence of one or two genes between the 70-bp repeat and the *VSG* sequence: the most commonly found is a pseudogene copy of *ESAG3* (11% of *VSG* cassettes). The second, annotated as an *UDP-Gal* or UDP-GlcNAc-dependent glycosyltransferase by Prof. M. Ferguson, was found to be present in 4% of *VSG* cassettes, mostly as a pseudogene. Whether there was any significance in coexpression of these genes with *VSG*s at some time in the past, or even at present, following cassette duplication into *VSG* expression sites, it is likely that currently the array copies of these genes act primarily, if not exclusively, as blocks of sequence homology facilitating recombination between silent *VSG*s.

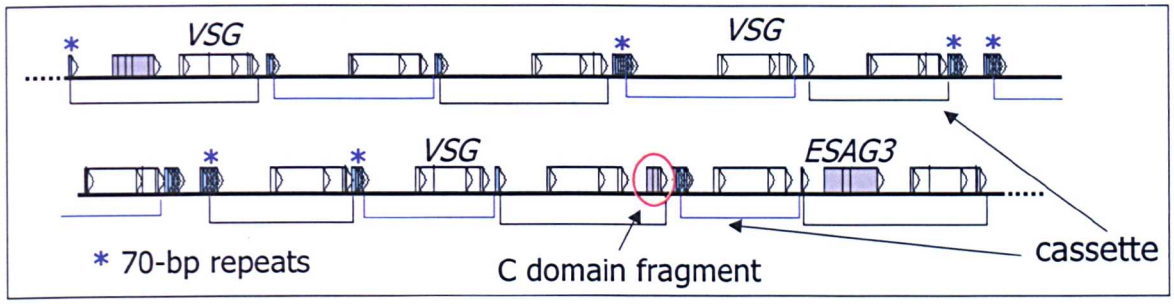


Figure 3.13: Example of *VSG* array with cassette structure highlighted.
Cassette boundaries (indicated with alternating blue and black brackets) are approximate and given only as rough guide to *VSG* array structure.

Most arrays contain several copies of the retrotransposon *INGI,* which appears to cluster in these subtelomeric regions and largely to be absent from the core region of the chromosome (see Figure 3.12) (Kimmel *et al.*, 1987; Bringaud *et al.*, 2004). All *VSG* arrays are directed away from the telomere[24], apart from some on chromosomes 9 and 11, where there are many strand-switches associated with *INGI* insertion. *INGI* is partly associated to truncated *VSG*s in the genome, as depicted in Figure 3.14, showing that one fourth of incomplete *VSG*s are flanked by an *INGI* element. A significant proportion (around half) of these incomplete genes are C-terminal domains, or fragments thereof, which are found just downstream of a full-length *VSG* in a cassette. An example of these short C-terminal fragments is highlighted with a red circle in Figure 3.13: it is likely that also these sequences (as *ESAG3* and *UDP-Gal*) are functioning as an extra homology

[24] This will be discussed in Chapter 6, section 6.1.1.3.

flanking region, as the vast majority are degenerate, containing stop codons and/or frameshifts.

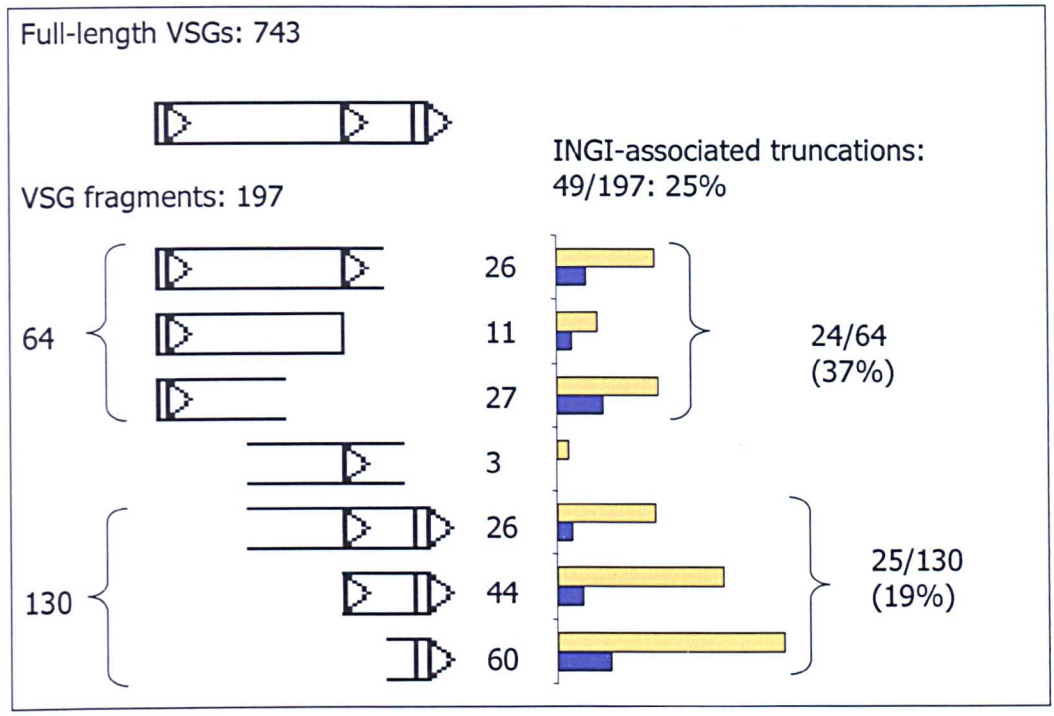


Figure 3.14: *VSG* fragments, proportion of different truncated versions.

The four *VSG* features are highlighted in the full-length gene at the top of the figure (signal peptide, N-terminal domain, C-terminal domain, GPI signal). Below are *VSG* fragments grouped into N-and C-terminal domain fragments and divided into subgroups depending on the extent of the truncation. To the right of each subgroup is the number of *VSG* fragments found to exhibit the particular truncation type. This number is then represented as a yellow bar further to the right, and below it is a blue bar indicating the proportion of truncations associated with *INGI*.

Looking for patterns of *VSG* gene distribution and relatedness, there appears not to be any obvious linkage of related genes within arrays, as shown in the tree below (Figure 3.15), which is the same alignment of full-length *VSG* given in Figure 3.10, but coloured instead according to array location. This is in keeping with earlier research showing that a family of related *VSG*s was dispersed in the genome (Beals and Boothroyd, 1992) and suggests that, if gene duplication is acting in tandem[25], genes are then rapidly reshuffled amongst the arrays. The great divergence of the archive, mentioned when discussing *VSG* domains and *VSG* families, significantly complicates gene duplication analysis, as only a small group of sequences can meaningfully be compared (data not shown) when tracing locations of related genes.

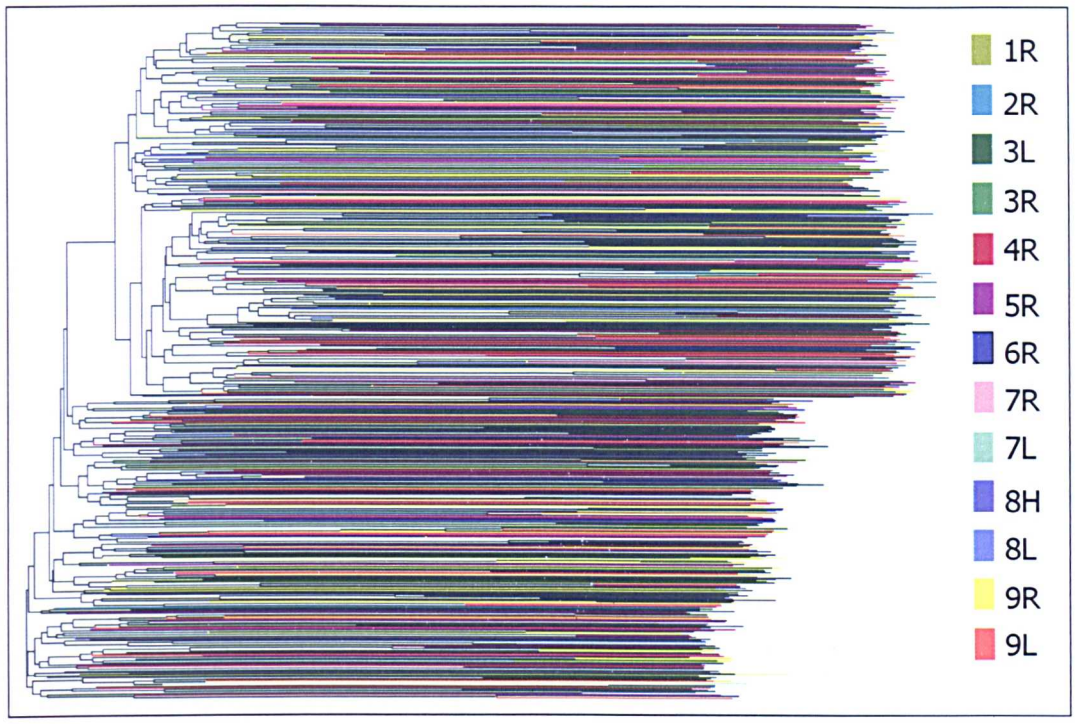


Figure 3.15: Tree derived from amino acid alignment of 594 full-length *VSGs* (same as Figure 3.10, but coloured by array).

To the right are colours associated to each complete array available (July 2005). Numbers 1 to 9 correspond to chromosomes 1 to 9, L and R correspond to the left and right subtelomeric arm of the chromosome, respectively. 8H is the 8L array homologue.

It was mentioned in the Introduction (section 1.1.3) that genome sequencing has brought to the fore the finding that at least some, if not all, subtelomeric regions containing *VSG* arrays are haploid. This was explored in more detail for the chr8 array and the array on the homologous chromosome, the first pair of homologues to be unambiguously determined, both subtelomeric contigs being linked to a common core housekeeping region. It appears that synteny is restricted to merely three of the >30 genes (Figure 3.16), suggesting that there might not be any preferential interaction of arrays present on homologous chromosomes.

[25] Further considerations on the issue of gene duplications will be made in Chapter 6 (Discussion), section 6.3.1.

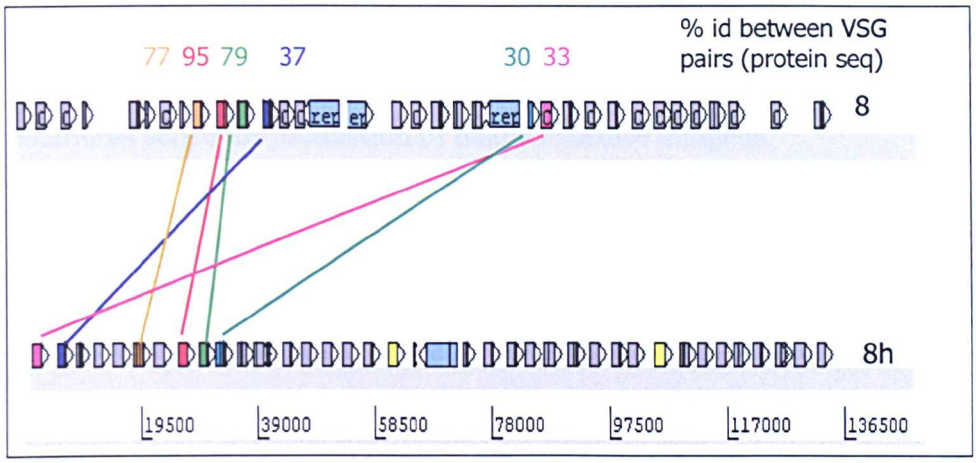


Figure 3.16: Comparison of the *VSG* array of chr 8 and the array on the homologue.

Elements in light blue correspond to *INGI* sequences; the telomere is to the left; numbers below indicate bp from the beginning of the two arrays. The five highlighted VSG pairs share at least 30% amino acid sequence identity. All VSGs that are not highlighted share less than 30% sequence identity.

3.8 Conclusions

The bioinformatic analysis conducted in this chapter has provided many clues as to how silent genes forming the *VSG* archive can contribute to antigenic variation. First and foremost, it was shown that the *VSG* archive is composed largely of pseudogenes and sequences that have diverged from the expressed consensus (~95% of silent *VSG*s). Having made an estimate of archive size at around 1600 genes, it is likely that about 80 functional genes are present, a number that is significant enough to envisage an important contribution of these genes in the course of a chronic infection. Whether during infection functional array genes are expressed with a specific timing, then to be replaced by hybrid genes exploiting the ~500 functional N-terminal domains (one third of total full-length *VSG*s), or whether the latter appear concurrently with their most closely related functional counterpart, remains to be elucidated.

The C-terminal domains are more degenerate than N-terminal domains, but the necessity for a large pool of the former is probably not as strong as that for the latter, as C-terminal domains do not contribute to antigenic variation, not containing any epitopes exposed to the surface. The *VSG* cassette structure might support the idea that most C-terminal domains have become mere flanks of homology for recombination reactions involving the N-terminal part of the protein, as outlined in Figure 3.9.

It therefore appears that, despite its degeneracy, the silent *VSG* archive contains a huge potential for creating new variants and that most of its potential would come to fruition through the formation of hybrid genes, utilising intact N-terminal domains and allowing

them to recombine with functional C-terminal domains resident at telomeres. Further recombination reactions creating mosaicism within the N-terminal domain would provide additional scope for production of distinct surface antigens.

Another important observation was made with regards to *VSG* divergence within the archive. The large number of *VSG* analysed allowed a more precise definition of the concept of "*VSG* families". It seems that such a definition is in general an overstatement, as any *VSG* would not have more than four other genes sharing above 60% amino acid sequence identity (data not shown), the average being around one close orthologue per gene. This level of divergence within a gene family is highly unusual, even amongst surface antigens, which normally share around 90% sequence identity (see section 6.3.2 for further discussion on this matter).

As for the sequence context of *VSG*s, it appears that subtelomeres (as in the case for other parasites, see Introduction, section 1.5) have been exploited for the expansion and evolution of the *VSG* archive. Several repetitive elements within the arrays are candidates for mediating *VSG* archive expansion and rearrangement (apart from *VSG* themselves). These include primarily the retrotransposon *INGI* and 70-bp repeats, with ESAG3 and UDP gal glucosyltransferase having a putative ancillary role. The analysis presented suggests that arrays are haploid and are highly prone to interacting with each other, given the current jumbling of related sequences.

The three key features mentioned above, *VSG* degeneracy, divergence and organisation/evolution, formed the three main directions of research conducted in the rest of this project. Considerations on these aspects of antigenic variation will emerge in the following two chapters, the first exploring the use of the *VSG* archive during a chronic infection (with the aim of understanding the contribution of array genes and pseudogenes to the infection) and the second analysing *VSG*-related genes, highlighting the main features of their divergence from *VSG*s and their distinct organisational pattern in the genome.

CHAPTER 4

HIERARCHICAL EXPRESSION OF *VSGS* IN SUBTELOMERIC ARRAYS: EXPERIMENTAL ANALYSIS

4 Hierarchical Expression of *VSGs* in Subtelomeric Arrays: Experimental Analysis

4.1 Introduction

The finding that most *VSGs* in arrays are pseudogenes leads to the question of how the *VSG* archive contributes to antigenic variation in the course of a chronic infection. It seems that most genes could be activated only through the prior formation of a hybrid or mosaic gene, by one or more gene conversion events. This phenomenon has been described qualitatively and is believed to occur with such a low probability that its switch products are observed only late in infection (Thon *et al.*, 1990), but its frequency has not been measured. A new tool to investigate the dynamics of *VSG* switching and activation is represented by the availability of the genome sequence for *T. brucei* strain TREU 927/4 GUTat 10.1. A chronic infection with this strain was expected to allow tracing, using bioinformatics analysis, a number of archive donor sequences (see estimate of *VSG* archive coverage in Chapter 3, section 3.2.1), and to distinguish simple duplicative activations of functional silent genes from more complex events involving pseudogenes. Published information on parasitaemia in strain 927 in mice confirmed the experimental feasibility of this approach (van Deursen *et al.*, 2001). The hypothesis is that the products of mosaic gene formation might be detected earlier in the course of an infection than is currently thought and then gradually come to represent the predominant source of novel VSG coats, once all other intact *VSGs* have been expressed and the host has become immune to their products. Mosaic genes may become more complex as infection progresses, originating from more than 2-3 donor sequences, and possibly resulting from the interaction of more distantly related donors.

4.2 Chronic mouse infection experiment: overview

4.2.1 Infection details

Initially, 11 mice were each injected with 10^6 trypanosomes (80 µl blood volume with parasitaemia of 1.3×10^7 trypanomes/ml), derived from an immunosuppressed mouse infected four days previously with a TREU 927/4 GUTat 10.1 stabilate, kindly provided by Prof. M. Turner. Before the experiment began, the identity of the trypanosomes was confirmed with genome markers specific to TREU 927 (data not shown).

As repeated sampling of individual animals would yield only a small sample volume, the protocol adopted involved batch infection and terminal sampling of groups of mice on days 3, 9, 14, 21-24 and 28 of infection. The mice to be sacrificed were selected on the basis of their parasitaemia, those with the highest being chosen, but the increasing effect of anaemia on mouse health was also taken into account as infection progressed (see Figure 4.1 for infection profile). Blood was collected from each sample and RNA, gDNA and stabilates were prepared. Infection became patent more rapidly than in chronic mouse infections previously carried out with pleomorphic EATRO 795 trypanosomes (Morrison *et al.*, 2005): the initial peak was reached at day 3 (rather than day 7), whereas the first relapse peak occurred at around day 10 (rather than day 14). Nevertheless, the distance between the first two peaks is the same in both 795 and 927 experiments, suggesting a similar course of infection. As noticed in the above-mentioned study, also here the onset of a second relapse peak was much more variable between mice.

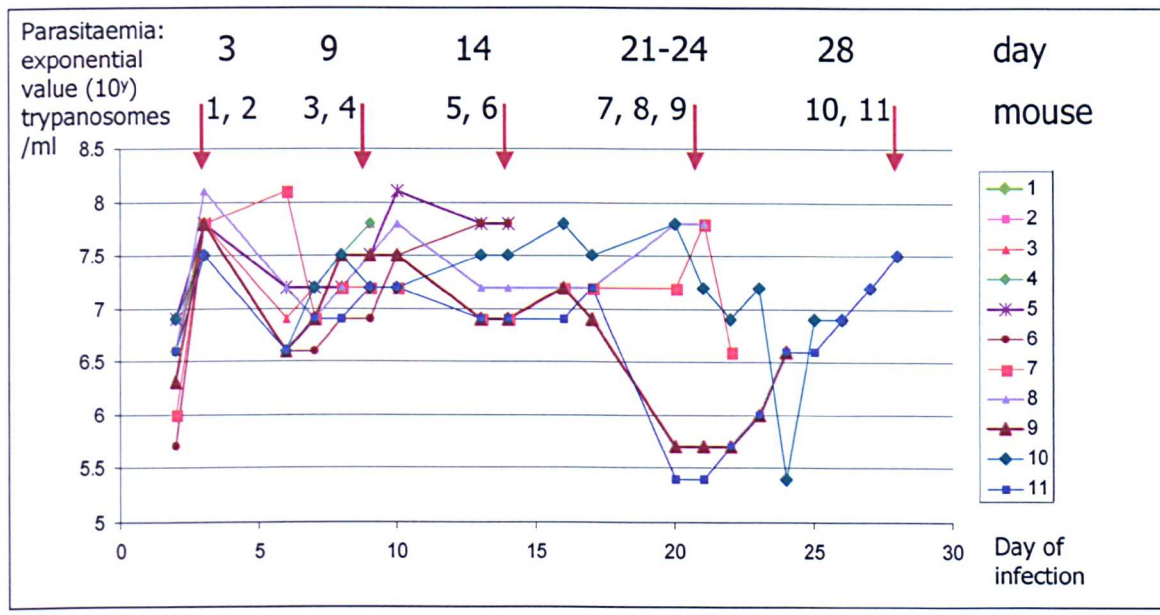


Figure 4.1: Infection profile of eleven chronic mouse infections carried out with *T. brucei* TREU 927/4 GUTat 10.1.

x axis = day of infection, y axis = parasitaemia (parasitaemia given as exponential value (10^y) trypanosomes/ml). Numbers 1 to 11 in legend and above chart represent the 11 mice used in the experiment. The five red arrows indicate the five timepoints.

4.2.2 Amplification of VSG *cDNA*

Reverse transcription was performed on the RNA obtained, using as primer oligo[dT] to maximise the production of full-length transcripts. PCR products were generated by Herculase proofreading polymerase (Stratagene) to maximise sequence accuracy, using a

forward primer annealing to the spliced leader sequence (found at the 5' end of all trypanosome mRNA molecules) and a reverse primer recognising a conserved 16-mer sequence specific to the 3' end of all *VSG* mRNAs (see Table 7.4 for primer sequences). Reaction conditions were: 5 mins at 95°C, followed by 30 cycles of 95°C for 1 min., 38°C for 2 mins, 72°C for 2 mins and a final extension of 5 mins at 72°C. Amplified products of the expected size (1.6 kb, see Figure 4.2) were cloned using the TOPO vector (see section 2.3.5 of Materials and Methods).

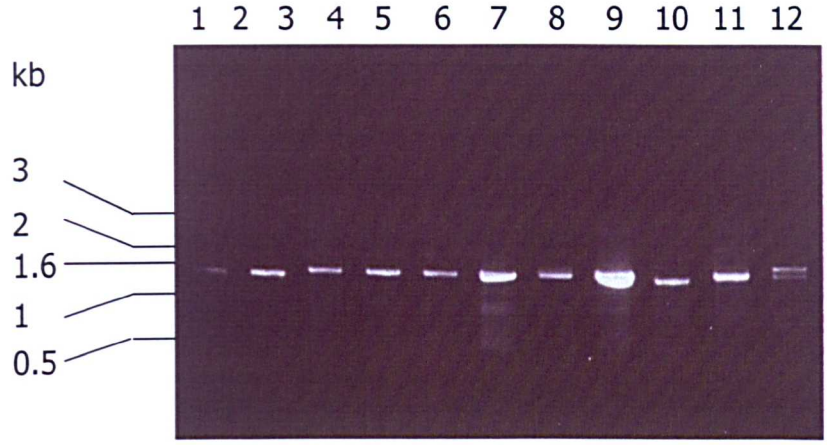


Figure 4.2: Ethidium bromide stained gel with *VSG* PCR reactions on cDNA collected from the GUTat 10.1 clone and its derivatives from eleven separate mouse infections, used for TOPO cloning.

Lane 1: GUTat 10.1 cDNA pre-infection; lanes 2-12: cDNA from mouse 1 to 11. Primers used annealed to spliced leader sequence (5') and 16-mer in 3' UTR.

4.2.3 VSG *cloning and sequencing*

Each positive clone obtained from the TOPO cloning reaction was given a three-part name (*e.g.* 03-01-01), indicating the day of harvest (3 to 28), the mouse (1 to11), and the specific clone number. A two part name with day of harvest and mouse will be used to indicate each separate infection (*e.g.* 03-01). Cloning efficiency varied between cloning attempts, but was generally low (average of 1 positive in 10 clones, with a minimum of 1 positive in 50 clones for the PCR products of 14-06); initial colony screening was conducted using the single colony lysis method (see Materials and Methods), which allows assessment of insert presence by alteration in plasmid size. Candidate plasmids were isolated using a small-scale plasmid DNA purification kit ("miniprep", Qiagen) and insert length checked by *Eco*RI digest (data not shown). Table 4.1 summarises the raw data, indicating the number of positive clones analysed for each PCR cloning and specifying the nature of the analysis (sequencing or PCR based on previously sequenced clones).

Table 4.1: *VSG*-containing clones obtained from RNA by RT-PCR.

Day and mouse	Sequenced clones	Clones analysed by PCR only	Total number of clones
03-01	2	1	3
03-02	2	1	3
09-03	4	4	8
09-04	2 (+ 1 partial)	7	10
14-05	2	3	5
14-06	1	1	2
22-07	4	0	4
21-08	2	2	4
24-09	1	1	2
28-10	5	6	11
28-11	11	4	15
Total	37	30	67

The full-length VSG coding sequence was assembled by overlapping sequencing reads, produced from M13 forward and reverse primers present at either sides of the TOPO vector cloning site, and internal primers designed specifically to cover the central region of the gene (see Figure 4.3 for primer design details and Table 7.4 of Appendix for a full list of primers used in this experiment). At least two rounds of sequencing were performed to confirm independently the sequence of the insert across its whole length, unless the first result gave 100% sequence identity to a *VSG* already present in the database.

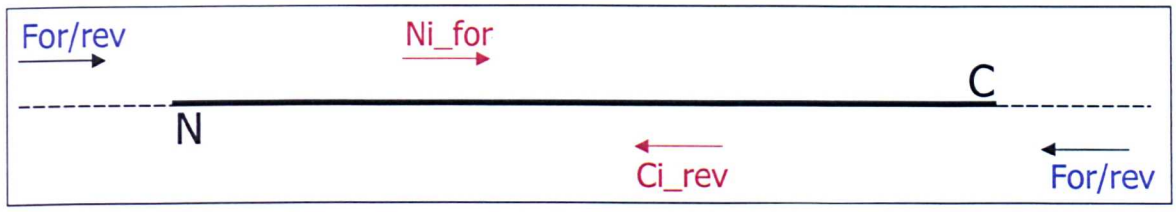


Figure 4.3: Primer nomenclature for *VSG* sequencing and analysis. N and C refer respectively to *VSG* N-terminal and C-terminal domains.

In blue are the M13 forward and reverse primers used for the first round of sequencing (for and rev amplified either the N- or the C-terminal domain depending on the orientation of the insert in the vector). In red are the primers used to cover the central region (Ni= N internal, Ci = C internal), designed on the basis of the M13 sequencing results.

The two Figures below (Figure 4.4 and Figure 4.5) give details of the PCR analysis of unsequenced clones based on primers developed for sequenced clones; it is shown for mice 3, 4, 5, 10 and 11, for which different inserts were present in the clones obtained. All unsequenced clones derived from mice 1, 2, 6, 8 and 9 corresponded to the single variant sequenced (PCR data not shown). It is worth noting that localised sequence differences could be present between a given sequenced clone and unsequenced clones identified solely by PCR analysis. Such "positive" clones were nevertheless not considered further, as the main purpose of the analysis was to capture the full diversity of each infection, and

then to sequence clones that appeared not to match any of the known characterised *VSG* sequences. The number of different *VSG*s obtained ultimately derives from the diversity of antigen types present in the blood at the time of harvest, which is expected to be very low in the first peak and to increase gradually with time (Capbern *et al.*, 1977), but the PCR reaction may preferentially amplify the most abundant mRNAs, overshadowing the actual sample composition. Four different *VSG*s was the maximum number obtained per infection, and was observed at day 9, 22 and 28, so the only increase in the number of variants isolated was observed between day 3 (one variant) and the rest of the timepoints (four variants).

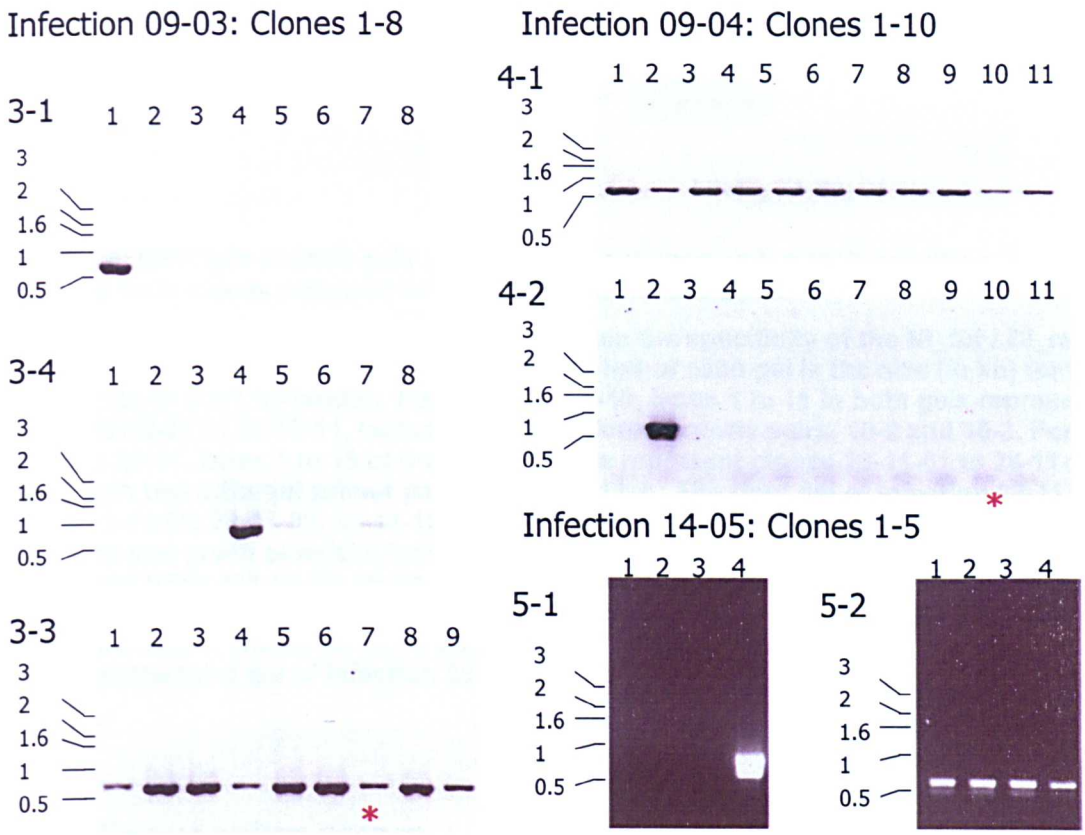


Figure 4.4: Ethidium stained gels showing PCR products using *VSG*-specific primers, amplified from clones obtained from infection 09-03, 09-04 and 14-05.

The number to the upper left of each gel indicates the specificity of the Ni_for / Ci_rev primer pair used for PCR analysis of the clones (e.g. 3-1 indicates mouse 3 clone 1; primers 3-3 and 4-1 are the same, as the *VSG* is shared in both infections). Negative reversal of the gel image is used in gels of infection 09-03 and 09-04, whereas the gel image for 14-05 was not reversed. To the left of each gel is the size (in kb) for the main bands of the 1 kb ladder. For infection 09-03, lanes 1 to 8 represent clones 09-03-01 to 09-03-08, tested with three different primers (3-1, 3-4 and 3-3). PCR with 3-3 primers includes a negative control (lane 9, clone 28-10-02), needed due to likely primer contamination. For infection 09-04, lanes 1 to 10 represent clones 09-04-01 to 09-04-10, tested with two different primers (4-1 and 4-2). Lane 11 in both gels represents a negative control (clone 28-10-02). For infection 14-05, lanes 1 to 3 represent clones 14-05-03 to 14-05-05, tested with two different primers (5-1 and 5-2). Lane 4 on gel 5-1 represents clone 14-05-01, lane 4 on gel 5-2 represents clone 14-05-02. The asterisk indicates clones that are not amplified by any of the primer pairs (clones 09-03-07 and 09-04-10).

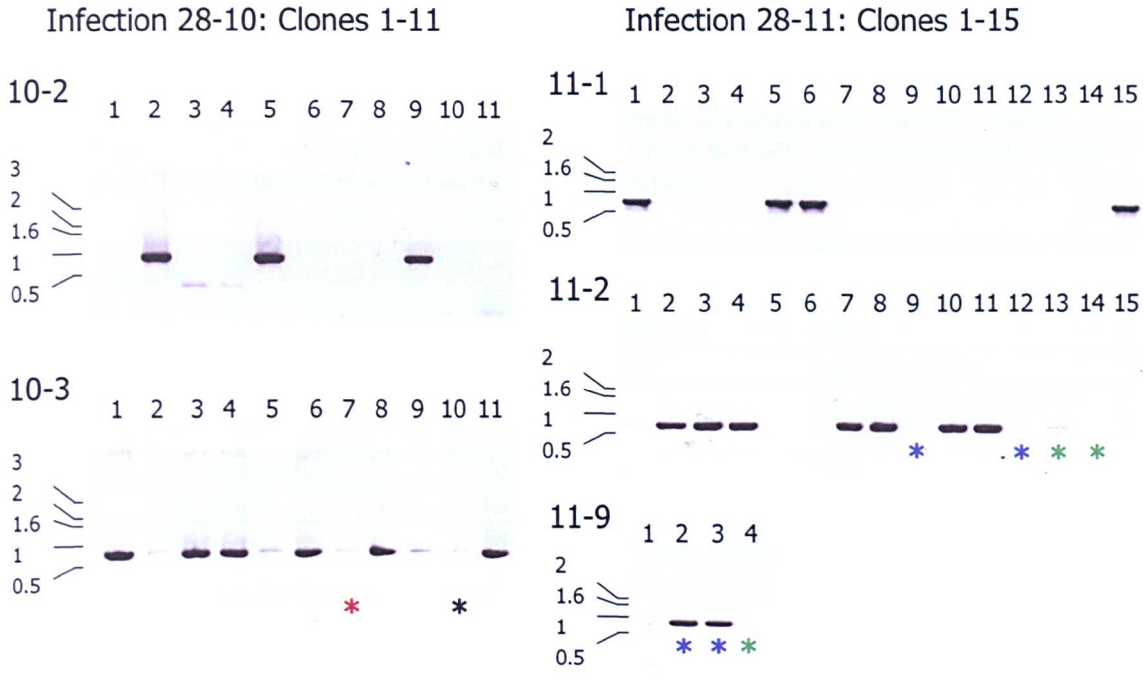


Figure 4.5: Ethidium stained gels showing PCR products using *VSG*-specific primers, amplified from clones obtained from infection 28-10 and 28-11.

The number to the upper left of each gel indicates the specificity of the Ni_for / Ci_rev primer pair used for PCR on the clones. To the left of each gel is the size (in kb) for the main bands of the 1 kb ladder. For infection 28-10, lanes 1 to 11 in both gels represent clones 28-10-01 to 28-10-11, tested with two different primer pairs, 10-2 and 10-3. For infection 28-11, lanes 1 to 15 of the first two gels represent clones 28-11-01 to 28-11-15, tested with two different primer pairs, 11-1 and 11-2. The third gel of infection 28-11 has clones 28-11-02, 28-11-09, 28-11-12, 28-11-14 (lanes 1-4) tested with primer pair 11-9. The red, black and green asterisks indicate clones that are not amplified by any primer pair of sequenced *VSG*s (28-10-07, 28-10-10 and 28-11-14, respectively). Clone 28-11-13 (faint band in 11-2 gel) was found to have the same *VSG* as clone 28-11-14, see Figure 4.6. Blue asterisks indicate clones 28-11-09 and 28-11-12, corresponding to the same *VSG*, as can be seen from the third gel of infection 28-11.

4.2.4 Sequencing errors

Errors were present in the cDNA sequence of 09-03-01, 09-03-04, 09-04-01, 14-05-02, 22-07-02 and 28-11-14. These were all observed as differences between independent sequencing reads and could be accounted for by gDNA contamination (unlikely, as PCR was performed using mRNA-specific primers), or errors in reverse transcription, PCR, or cloning and replication in *E. coli* (*VSG* sequences have sometimes resulted in unstable clones with poor growth). The approach taken to correct these errors was to amplify the corresponding *VSG* with specific primers, from gDNA and cDNA, when a putative donor sequence was not present to identify the error. Table 4.2 gives details of the errors detected.

Table 4.2: Details of errors present in sequenced clones.

The fourth and fifth column refer to the method by which the presence of the error was detected and corrected. Column four gives details of the sequence used as comparison with the error-containing sequence. Column five gives the details of the new clones obtained by reamplification of the initial clone by using specific internal primers; these clones were then sequenced and used as a comparison

Clone	Error type	bp	Corrected by sequence comparison	Corrected by reamplification
09-03-01	Deletion	1	donor read (98% id)	No
09-04-01	Deletion	1	Clone 09-03-03	gDNA clone
09-03-04	Insertion	39	No	2 cDNA and 2 gDNA clones
14-05-02	Point mutation	1	Clone 14-06-01	No
22-07-02	Insertion	1	Donor	No
28-11-14	Insertion	1	Clone 28-11-13	No

In short, five indels (of 1 bp in four cases out of five) and one mismatch were found in 36 fully sequenced clones (totalling around 53 kb). The fidelity of the RT polymerase supplied by Invitrogen has been shown to be around 10^{-6} errors. The higher error rate detected (10^{-4}) could therefore be due to PCR amplification. This is further suggested by the fact that, in the case of 09-03-04, the reamplified cDNA did not contain the insertion found in the clone. The magnitude of errors detected is also above that commonly found with the Herculase polymerase (10^{-6} errors) used in PCR, so inherent features of the *VSG* sequence itself impeding accurate copying in the PCR reaction should not be discounted as an explanation. The possibility of rearrangments being generated by template copy choice, a process by which, during PCR, the polymerase would switch to another closely related template (Zaphiropoulos, 1998) has also been considered, and deemed unlikely: there is no evidence as yet that any of the *VSG*s containing errors are present in multiple copies and the estimate of divergence of the *VSG* archive given in Chapter 3, section 3.6, makes the chance that six *VSG*s might have ~99% identical copies in the genome very small. In addition, Herculase polymerase is a blend of *Pfu* and *Taq* polymerase and *Pfu* has been shown to be less prone to copy choice (Shafikhani, 2002).

4.3 Overview of *VSG*s isolated in current study

Figure 4.6 summarises the basic information with regards to all *VSG* clones isolated in this study, divided by day, mouse and clone number[26]. Table 4.3 gives a more concise list of all *VSG*s identified by sequencing and relates each *VSG* to the clones in which it was found.

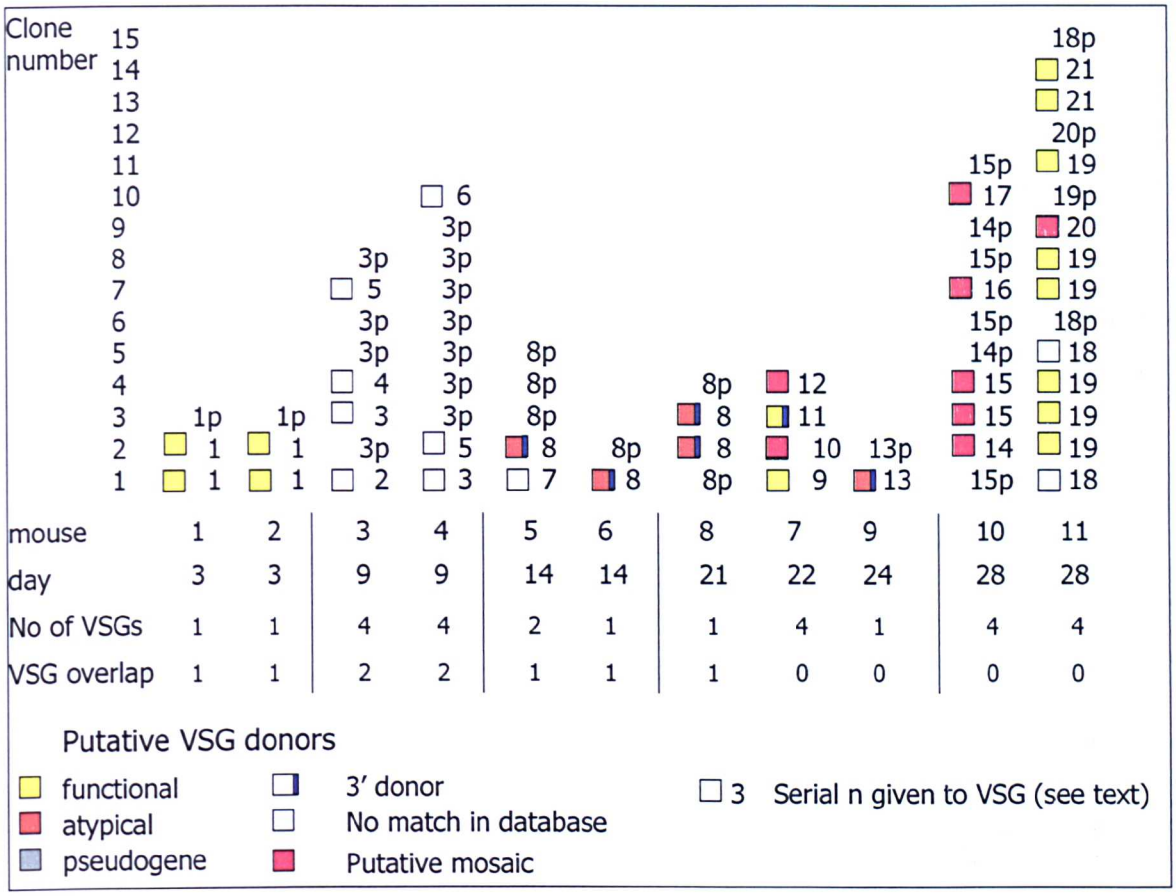


Figure 4.6. Summary of cDNA clones isolated from each of the eleven mouse infections.

Each vertical column of squares and numbers represents the clones isolated in each mouse. Squares show the nature of each sequence, based on its source(s) in the *VSG* archive, and associated numbers are the number of clones isolated for each sequence; p =PCR analysis only. Below each column is a summary giving the mouse number, the day, the number of VSGs isolated in each infection, and the number of *VSG*s isolated that were common to different infections (*VSG* overlap).

[26] This is present in tabulated form with additional details (identifier and location of putative donors, domain annotation) in Table 7.5 of the Appendix.

Table 4.3. *VSG*s identified, with day of appearance and prevalence amongst clones.

The presence of a donor for the 3' end of the gene is specified in the third column (see text for details). Novel *VSG*s are given the clone identifier preceded by CMI (chronic mouse infection), whereas *VSG*s corresponding to known archive genes are given the GeneDB identifier of the donor.

VSG n	VSG id	Donor type	Day	mouse	Clones (sequenced)		Clones (PCR)	
					n	id	n	Id
1	Tb10.v4.0001	functional	03	01, 02	4	03-01-01, 03-01-02, 03-02-01, 03-02-02	2	03-01-03, 03-02-03
2	CMI_09-03-01	?	09	03	1	09-03-01	0	
3	CMI_09-03-03	?	09	03, 04	2	09-04-01, 09-03-03	11	09-04-03 to 09-04-09; 09-03-02, 09-03-05, 09-03-06, 09-03-08
4	CMI_09-03-04	?	09	03	1	09-03-04	0	
5	CMI_09-04-02	?	09	03, 04	2	09-04-02, 09-03-07	0	
6	CMI_09-04-10	?	09	04	1	09-04-10	0	
7	CMI_14-05-01	?	14	05	1	14-05-01	0	
8	Tb09.v4.0077	Atypical + 3' donor	14	05, 06, 08	4	14-05-02, 14-06-01, 21-08-02, 21-08-03	6	14-05-03, 14-05-04, 14-05-05, 14-06-02, 21-08-01, 21-08-04
9	Tb09.244.1580	functional	22	07	1	22-07-01	0	
10	CMI_22-07-02	Mosaic	22	07	1	22-07-02	0	
11	Tb927.3.190	Functional +3' donor	22	07	1	22-07-03	0	
12	CMI_22-07-04	Mosaic	22	07	1	22-07-04	0	
13	Tb09.v4.0102	Atypical + 3' donor	24	09	1	24-09-01	1	24-09-02
14	CMI_28-10-02	Mosaic	28	10	1	28-10-02	2	28-10-05, 28-10-09
15	CMI_28-10-03	Mosaic	28	10	2	28-10-03, 28-10-04	4	28-10-01, 28-10-06, 28-10-08, 28-10-11
16	CMI_28-10-07	Mosaic	28	10	1	28-10-07	0	
17	CMI_28-10-10	Mosaic	28	10	1	28-10-10	0	
18	CMI_28-11-01	Functional/ hybrid	28	11	2	28-11-01, 28-11-05	2	28-11-06, 28-11-15
19	Tb927.5.5080	Functional	28	11	6	28-11-02, 28-11-03, 28-11-04, 28-11-07, 28-11-08, 28-11-11	1	28-11-10
20	CMI_28-11-09 (Tb927.7.6530)	Mosaic	28	11	1	28-11-09	1	28-11-12
21	Tb10.v4.0088	Functional	28	11	2	28-11-13, 28-11-14	0	

The 37 sequenced clones were found to correspond to 21 different *VSG*s, as some clones encoded the same *VSG*. A summary of characteristics of these 21 *VSGs* is given in Figure 4.7: they have been divided into different groups depending on the features of the donors identified in GeneDB. Both "functional" and "atypical" *VSG* donors[27] were found, although activation of the two atypical donors was probably associated with separate donation of the 3' end of the gene, at which point mismatches were present. Because of the small extent of this hypothetical 3' donor sequence these genes have not been defined as "hybrids", a term reserved for genes in which N-terminal and C-terminal domains are contributed by separate genes. Seven cDNA sequences could be best explained as putative hybrids or mosaics arising from the combination of different donor genes (which included functional genes and pseudogenes). For other genes no donors could account for the full extent of their sequence, and only partial matches were found. Figure 4.7 divides them

into two groups: one group is likely to correspond to telomeric minichromosomal genes (which have not been included in the sequencing project but which are known to be activated early in infection, Robinson *et al.*, 1999) and the other to be array *VSG*s (or mosaics thereof) that have not yet been fully sequenced or assembled. If only the 14 genes with putative array donors are considered, it appears that ten have full-length matches to either contigs or main chromosome assemblies, whereas, for the remaining four, partial donors were found in short reads. This confirms indirectly the prediction that the contigs and main chromosome molecules in GeneDB include approximately three quarters of the *VSG* archive, and suggests that most of the archive is captured when genome sequence survey reads are also included.

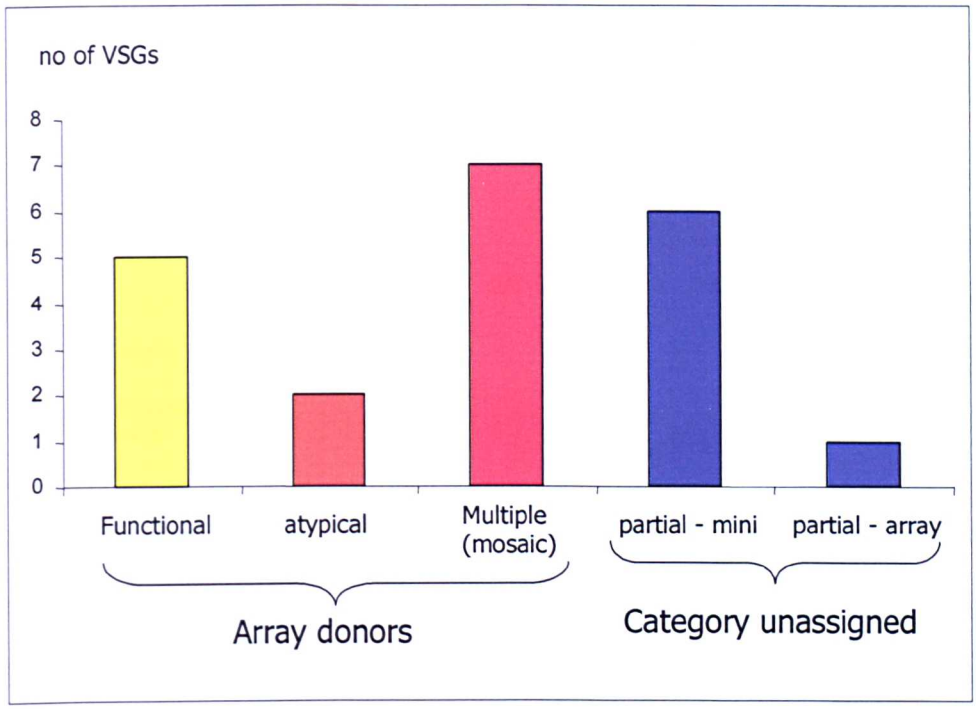


Figure 4.7: 21 unique *VSG*s isolated from eleven chronic mouse infections, grouped by donor features.

Putative array donors are present in yellow (functional), orange (atypical), red (mosaic) bars. *VSG*s for which only partial donors were found are divided into putative minichromosomal genes (Partial-mini) and putative array genes (partial-array).

[27] See Chapter 3, section 3.2.2 for an explanation of these terms.

4.4 Summary of basic findings

A condensed view of Figure 4.6 is given in Figure 4.8, highlighting the relative contribution of *VSG*s with unknown, single or multiple donors at the different timepoints (day 9, 14, 21-24, 28), in order to provide a dynamic view of the data, mirroring the course of a single infection.

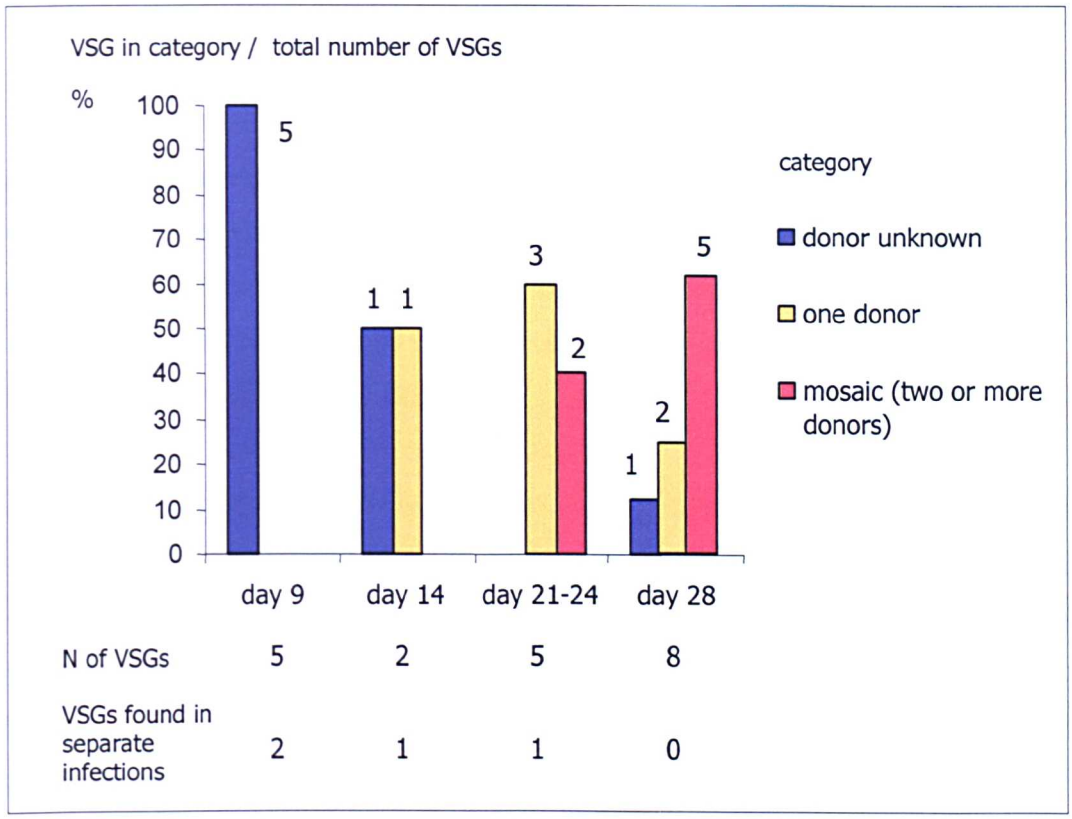


Figure 4.8: Relative contribution of *VSG*s with unknown, single or multiple donors at different timepoints (day 9, 14, 21-24, 28).

Data in bars are expressed as percentage of total number of *VSG*s per timepoint; above each bar is the absolute number of *VSG*s in each category. Below the chart are highlighted the number of *VSG*s per timepoint and the number of *VSG*s common to separate infections.

The most striking observation is that, although based solely on bioinformatic detection of putative donors, the data suggest that by day 28 mosaic gene formation contributes significantly to variants arising, giving for the first time a quantitative appreciation of the importance of this process *in vivo*. Putative mosaic genes were first detectable at day 22, although no data are available between day 14 and 22, so their first appearance might be significantly earlier. Appearance at day 22 fits with previous studies in rabbits (Capbern *et al.*, 1977), in which day 23-25 saw the emergence of the first mosaic. The timing corresponds, at least in a subset of mice, to the emergence of a second relapse peak. Interestingly, the fact that the infection profile starts to vary significantly between mice around this time might suggest a possible correlation with the onset of silent array genes

and their less predictable use to generate mosaics that are potentially unique to each infection. A more detailed analysis of mosaic genes will follow in section 4.7.

Assuming that unknown *VSG*s isolated on day 9 correspond to minichromosomal telomeric genes, the individual clones examined seem to confirm the hierarchy observed in previous experiments (telomere-proximal *VSG*s preceding array *VSG*s, Robinson *et al.*, 1999). There appears to be an overlap between telomere-proximal and array genes on day 14, possibly mirroring an overlap in activation probabilities between these two loci; whether this is due to coding sequence homology of an array gene for the expressed *VSG* taking precedence over homology at telomere sequence, this remains to be established, as a clear path for activation of these genes cannot be determined in this experiment. The appearance of array genes seems not to correlate with the number of 70-bp repeats present upstream of the donors, confirming more directly the hypothesis made in section 3.7 that 70-bp repeats might not contribute greatly to the hierarchy of array genes, as most of them have only one or two repeats, with few exceptions.

As found in previous studies (Robinson *et al.*, 1999; Morrison *et al.*, 2005), it appears that at the beginning of infection a set of predominantly, if not exclusively, telomeric genes is activated as "default" in separate hosts (see Figure 4.8, below the chart). This is more prominent at day 9 and gradually becomes less common as infection progresses and the large archive of silent array genes is accessed. The two infections analysed at day 28 did not share any expressed *VSG*s, suggesting that by this time the infections take distinct routes.

All but one of the genes isolated after day 9 could be identified as having either functional, atypical or multiple donors. The only exception is *VSG* 28-11-01, for which an N- and a C-terminal donor have been found, however N and C-terminal reads do not overlap, making the gene either a putative functional or a hybrid. Although donors for *VSG* 28-10-10 do not span the whole length of the open reading frame, the overlap between putative N-terminal domain donors suggests the gene to be a mosaic.

An important link with the genome project analysis presented in Chapter 3 was the finding of expressed *VSG*s with novel C-terminal domain types (types 5 and 6), first observed bioinformatically in the course of the *VSG* array annotation and never previously isolated *in vivo*. In addition, mosaic 22-07-02 showed expression of a type A N- terminal domain with only two cysteines (similar to MVAT5, (Carrington *et al.*, 1991)), and a type 1 C-terminal domain (as judged by the GPI signal sequence) with type 6-like cysteine spacing.

The assembling of the latter is likely to represent a rare event in the gradient of recombination between C-terminal domain types proposed in Figure 3.9 (in which the beginning of the domain is more prone to intertype recombination): in the present case it involves the normally domain-specific C-terminal end of the gene. These examples add more information on the structural constraints (cysteine pattern and GPI signal) and complexity (domain types) of *VSG*s and could possibly prompt further biochemical studies, investigating the folding of N-terminal domains with only two cysteines and the cleavage features of novel GPI signals.

Three aspects of the work outlined so far will be considered in the remaining part of this chapter. Examples of expressed *VSG*s with 3' donor sequence (section 4.6) will be presented, followed by a more detailed analysis of the *VSG* mosaics detected in this study, through a bioinformatic comparison with and between their donors (section 4.7) and by experimental confirmation of the mosaic nature of one of these genes (section 4.8). Before this, the putative presence of several expressed *VSG*s containing "real" mismatches with their donors will be discussed (section 4.5).

4.5 Point mutation in expressed *VSG*s?

After initial interest was sparked by a few clones having a significant number of mismatches with their donors, a systematic analysis of presence of point mutation in all fully sequenced clones was undertaken (36 clones, with 09-04-10 excluded, as it is only partially sequenced) (see Table 7.6 of Appendix for full list of unique mutations). It was possible to observe that, in the case of clones 09-03-03 and 09-04-01, both expressing the same *VSG*, there were no mismatches when the two sequences were compared, whereas three mismatches were found when a comparison was made with the putative donor read. This, together with the identical point mutation found between the database GUTat 10.1 *VSG* sequence and that of four clones from day 3, and a similar comparison made between clones 09-03-07 and 09-04-02, suggests that there might be a slight baseline divergence between the clone used for the genome sequencing and the clone used to start the infections. If point mutation occurs relatively rapidly this might have occurred just by passaging. This finding gives a good "internal control" to assess the validity of point mutations detected in other clones, in the absence of experimental confirmation. Moreover, if point mutations were generated by artefactual amplification during RT-PCR, the chance of detecting two independent clones originating from separate infections sharing the same base pair change, as just described, would be very remote. Figure 4.9

gives an overview of the number of point mutations found across the 19 *VSG*s for which complete or sufficient donor sequence was present to enable the comparison, and it also highlights the number of clones analysed per individual *VSG*. Only five out of 19 do not show any point mutation differences.

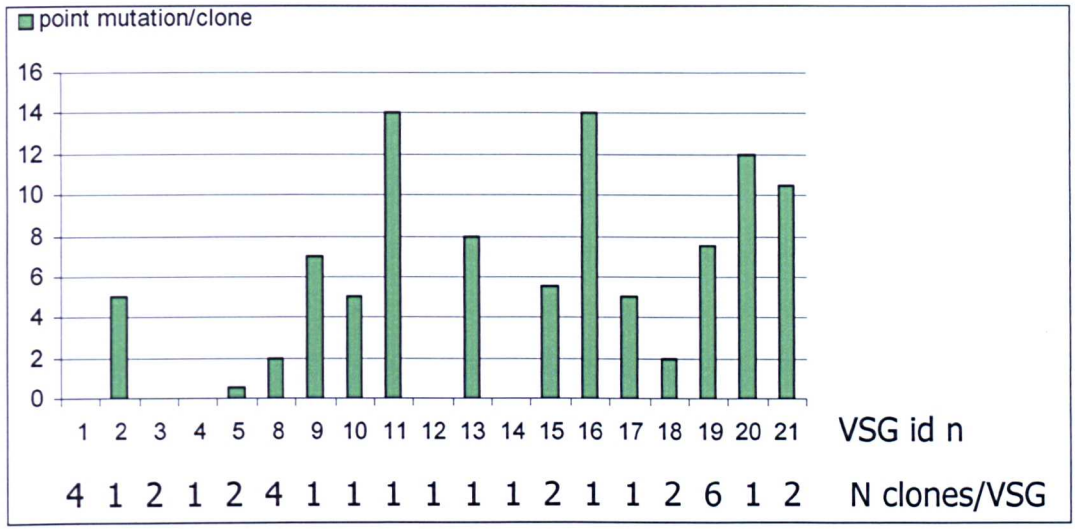


Figure 4.9. Point mutation in 19 unique expressed *VSG*s (1-21, except *VSG*s 6 and 7).

For each *VSG*, the number of sequenced cDNA clones is shown below the *VSG* id number, and the average number of point mutations in each *VSG*, in comparison with the silent gene in GeneDB, is recorded in the green bars.

The data, rearranged "chronologically" into the same groupings as in Figure 4.8, are displayed in Fig 4.10 below, again to see whether a pattern can be found.

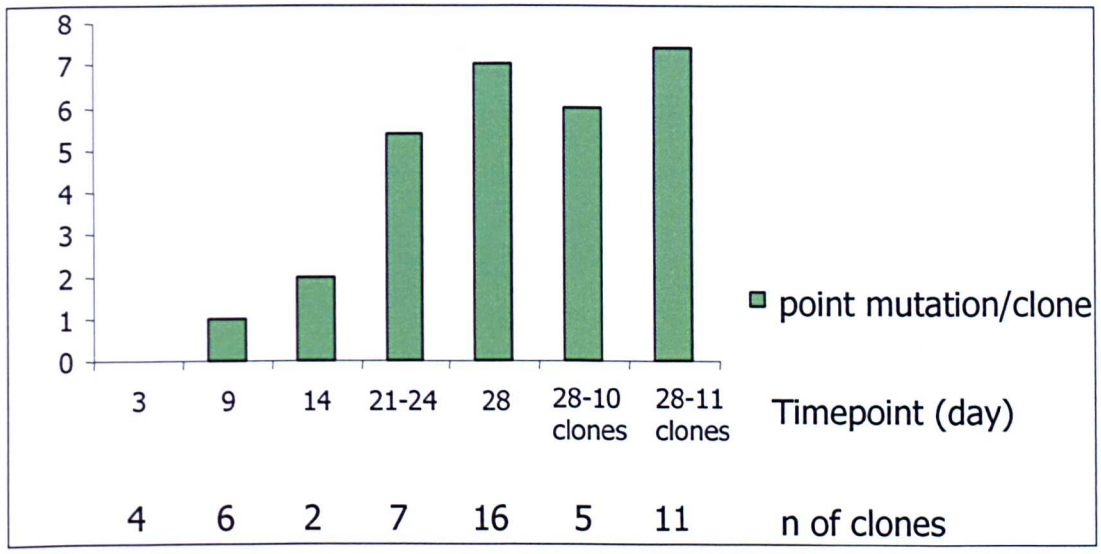


Figure 4.10: Number of point mutations per clone at the different timepoints (day 3, 9, 14, 21-24, 28).

The number of clones considered is given below the table. Day 28 has also been split into infections 28-10 and 28-11, to allow for the fact that at this advanced stage the infection profile might differ significantly between mice.

The number of point mutations per clone seems to increase dramatically when "early" and "late" timepoints are compared. Interpreting these data is problematic, as the detection of these mutations cannot be immediately correlated with antigenic variation and the production of a non-crossreactive coat: this will be discussed at the end of the chapter. The rest of this section will be devoted to further validating the point mutation data.

4.5.1 Location of point mutations

The location of differences between expressed and putative donor genes strongly suggests a non-random distribution of the mutations detected, as 173/184 (94%) unique codon changes are located in the N-terminal domain. Figure 4.11 gives the combined location of all mutations across the 18 *VSG*s analysed, highlighting mutations in N-terminal domains type A and B: it is apparent that many of the point mutations are clustered. This was statistically validated by a Chi-squared test, in which point mutations in the N-terminal domain (between bp 1 and 1000) were grouped into ten 100 bp intervals. Types A and B were considered separately, and in both cases the Chi-squared value supported a non-random distribution of mutations at the 0.01 significance level[28], with concentration of mutations at bp 400-500 and 600-700 for type A, and at bp 600-800 for type B.

[28] The expected chi square value at 0.01 significance level is 21.67 for nine degrees of freedom, and the observed values were of 29.05 and 24.06 for mutation distribution in type A and B domains respectively (see Table 7.7 of Appendix for full data analysis).

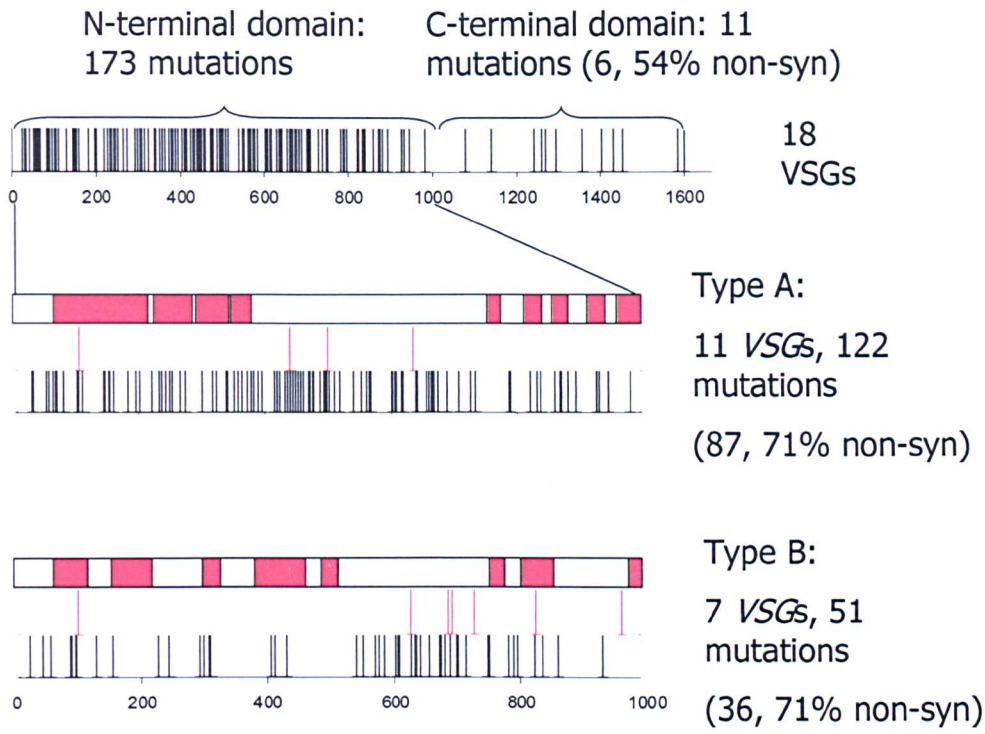


Figure 4.11: Physical location of point mutations along the length of 18 expressed *VSG*s.

All 181 unique point mutations have been pooled together along the length of an "ideal" *VSG*. N-terminal domain point mutations are then considered separately in type A and type B domains. Above each set of point mutations are vertical red lines indicating the location of cysteine residues and a depiction of the putative alpha helical content of each type (helical regions indicated in red). The exact correspondence of a point mutation with cysteine residues is due to the fact that the average of the cysteine positions across all *VSG*s analysed was taken; no mutation of cysteines was observed. The 11 type A *VSG*s are 1, 2, 3, 5, 9, 10, 11, 12, 16, 19 and 21. The 7 type B *VSG*s are 4, 8, 13, 15, 17, 18, 20. In brackets are indicated the number and percentage of non-synonymous point mutations.

A crude prediction of secondary structure potential (using Jpred, http://www.compbio.dundee.ac.uk/~www-jpred/) was also undertaken, to see whether a correlation between putative surface loops (in white) and location of point mutations could be found. Two "global" secondary structure predictions encompassing respectively all type A and all type B VSGs present in this study were attempted, using a multiple sequence alignment of these two sets of protein sequences. The secondary structure prediction suggests that the statistically significant concentration of mutations detected occurs in non-alpha helical regions. As none of the genes are very closely related to each other, the accuracy of the prediction is expected to be lower than if querying with a single sequence and allowing the program to align this with its close orthologues. In order to refine the analysis and overcome the above-mentioned limitations of "global" secondary structure prediction, three individual *VSG*s in which there are many point mutations are presented in Figure 4.12. A clearer correlation between non-helical regions and point

mutations was found: this was again statistically validated with a Chi-squared test (significance at the 0.05 confidence level was observed for all three *VSGs*[29]), strengthening the hypothesis that a process of positive selection might be operating, concentrating mutations in regions with reduced secondary structure potential, possibly corresponding or overlapping with surface epitope regions. Further analysis considering the number of synonymous versus non-synonymous mutations is conducted in section 4.9.1, where the measurement and the implications of positive selection operating on VSGs are discussed in more detail.

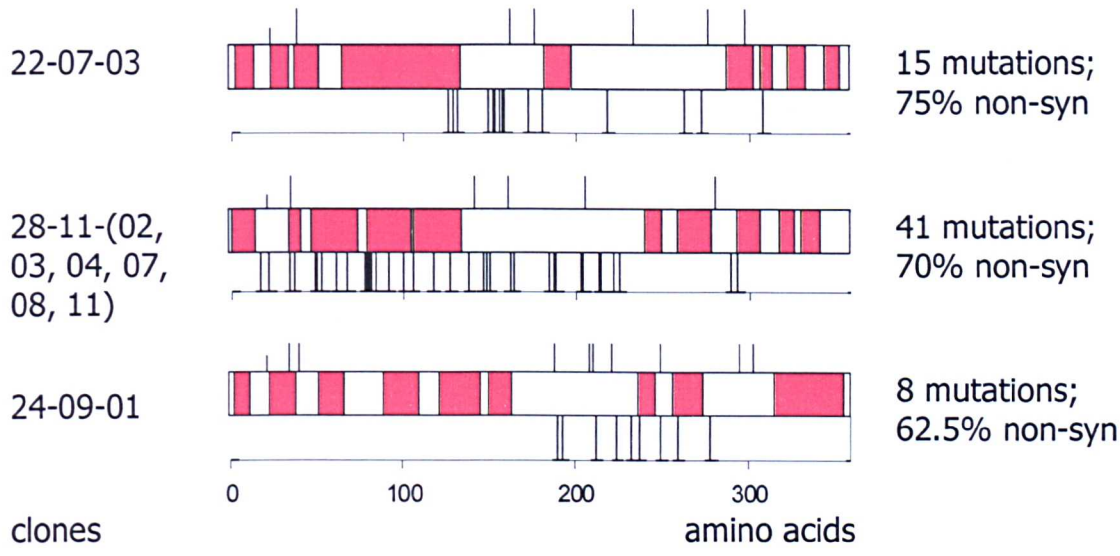


Figure 4.12: Correlation between point mutation and predicted helical content in a subset of N-terminal domains.

The first two examples are type A domains, whereas the third is type B. The short vertical lines above each graphic show positions of cysteines, and those below each graphic show the positions of point mutations. Predicted alpha helical regions are shown in red. The half-height line at the beginning of each cysteine pattern indicates the location of the signal peptide cleavage site.

[29] See Table 7.8 of Appendix for the full data analysis.

4.6 3' donor analysis

The hypothesis that the very end of expressed *VSG*s derived from a gene conversion of a silent array gene is commonly contributed by a 3' donor due to the degeneracy of C-terminal domains in the archive, as brought to light by the annotation of subtelomeric arrays (see Chapter 3, section 3.4), receives support from the common occurrence of conversions ending within the C-terminal domain (eight instances out of 21 genes, two atypical donors, two functional donors and four mosaics). This fits with previous studies in which the C-terminal domain end was found to be retained in expressed *VSG*s with unrelated N-terminal domains, such as *VSG*20, 20* and 20bis (Thon *et al.*, 1990). Nevertheless, termination of *VSG* gene conversion within the coding sequence happens also simply because it is a region of high identity between *VSG*s (Rice-Ficht *et al.*, 1981). This is the case for duplications involving *VSG* 117 and 118 (strain 427): both are functional array *VSG*s and most clones that activated these *VSG*s showed end of conversion around or within the GPI signal sequence (Bernards *et al.*, 1981; Michels *et al.*, 1983). Therefore, in addition to the previously detected "spontaneous" termination of conversion within coding sequence, there appears to be selection to retain the C-terminal end of the previously expressed *VSG*, as most incoming donors would have a faulty GPI signal. The identity of a 3' donor is difficult to ascertain, but the fact that only in one case there was a 3' donor match to a known array gene lends strength to the possibility that these donor C-terminal domains reside at telomeres and originated from *VSG*s previously occupying the expression site. Three instances of 3' donor detection will be considered, relating to three main donors: Tb09.v4.0077, appearing in four clones derived from three different mice (see Table 4.4), Tb927.3.190 (one clone) and Tb09.v4.0102 (one clone).

Table 4.4: List of clones with 3' donor.
Highlighted in red are the three cases analysed in section 4.6

Clone	Id of main VSG donor	Main VSG donor features	Domain type	VSG n
14-05-02, 14-06-01, 21-08-02, 21-08-03	Tb09.v4.0077	Atypical	B2	8
22-07-03	Tb927.3.190	Functional	A2	11
22-07-04	mosaic	Mosaic	A4	12
24-09-01	Tb09.v4.0102	Atypical	B2	13
28-10-03	mosaic	Mosaic	B3	15
28-10-07	mosaic	Mosaic	A3	16
28-11-09	mosaic	Mosaic	B3	20
28-11-14	Tb10.v4.0088	Functional	A2	21

VSG Tb09.v4.0077 appears to have been activated frequently between days 14 and 21, in a way reminiscent to *VSG* 118 in strain 427 (Timmers *et al.*, 1987). It seems that this partial duplication must have occurred at least once in the parent clone, at a time prior to the start of the infection, as three of four duplications have the same breakpoint between the two donors (see Figure 4.13). Such a coincidence could suggest that a telomere-proximal copy was present at the time of infection, and this in turn could explain frequent activation of this gene in separate infections. Nevertheless, the possibility that an array gene is activated by identical conversion reactions in different infections cannot be formally ruled out without further experimental analysis.

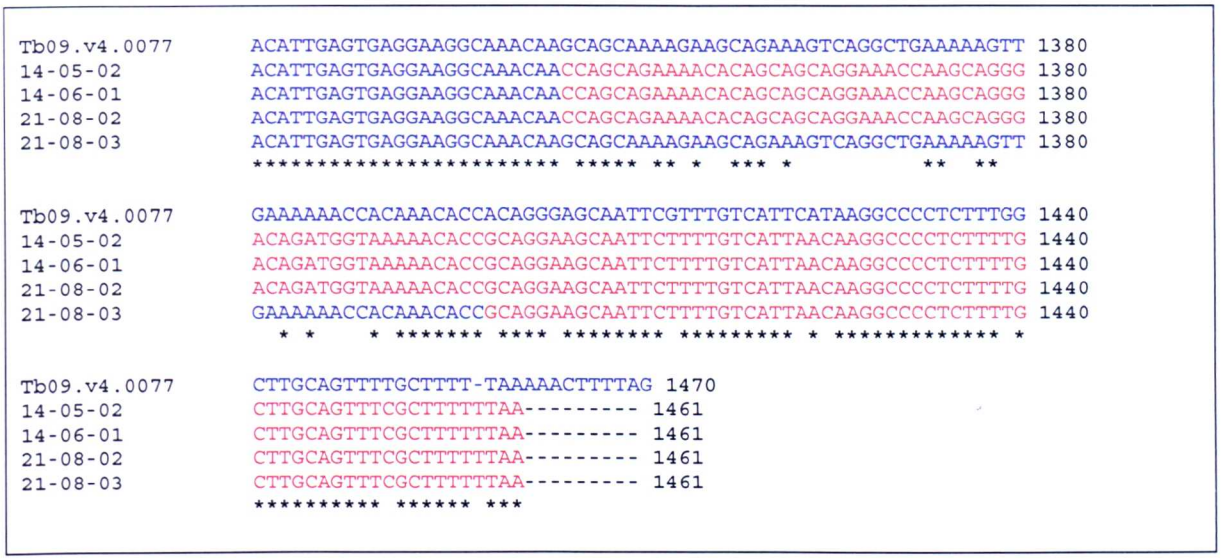


Figure 4.13: Nucleotide alignment of the C-terminal domain end of Tb09.v4.0077 (donor), with related expression-linked copies (ELC).

Donor sequence is indicated in blue, as are the ELC regions with 100% identity to the donor. All ELC regions in red are 100% identical to each other.

Regardless of whether they occurred during or prior to the infection, there were at least two separate activations of the same array gene, both using the same 3' donor. The use of the 3' donor sequence appears to be mandatory for the expression of this *VSG*, and this could be due to the fact that its GPI signal is "atypical" and possibly non-functional (see Figure 4.14).

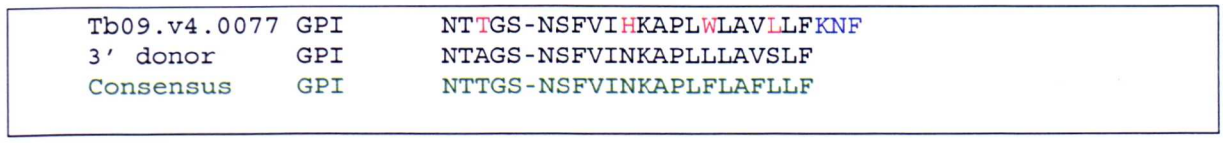


Figure 4.14: Putative GPI anchor signal of Tb09.v4.0077 (main donor), compared with that of 3' donor.

A consensus example is given for an expressed *VSG*, accession code AF335472. Residues in blue indicate the non-conserved additional amino acids at the end of the GPI signal. Residues in red indicate amino acid mismatches between Tb09.v4.0077 and 3' donor.

This type 2 GPI signal is three amino acids longer than the consensus sequence and has one polar lysine residue, normally not found in the hydrophobic extension. The nucleotide alignment in Figure 4.13 gives a possible explanation for this: the donor *VSG* appears to have a deletion of one T, producing a frameshift, resulting in the incorporation of an extra three amino acids.

The second *VSG*, Tb927.3.190, appears not to have undergone such a prominent alteration through gene conversion (see Figure 4.15). From this example it could be concluded that either the gene conversion fortuitously included the very end of the gene (as in the case of *VSG*s 117 and 118 mentioned above) or that it was a requirement for expression. Subtle departures from the consensus sequence are present, although the GPI signal has been classified as functional in the bioinformatic analysis, as the signal is still conserved in length and has a hydrophobic profile very similar to the consensus sequence.

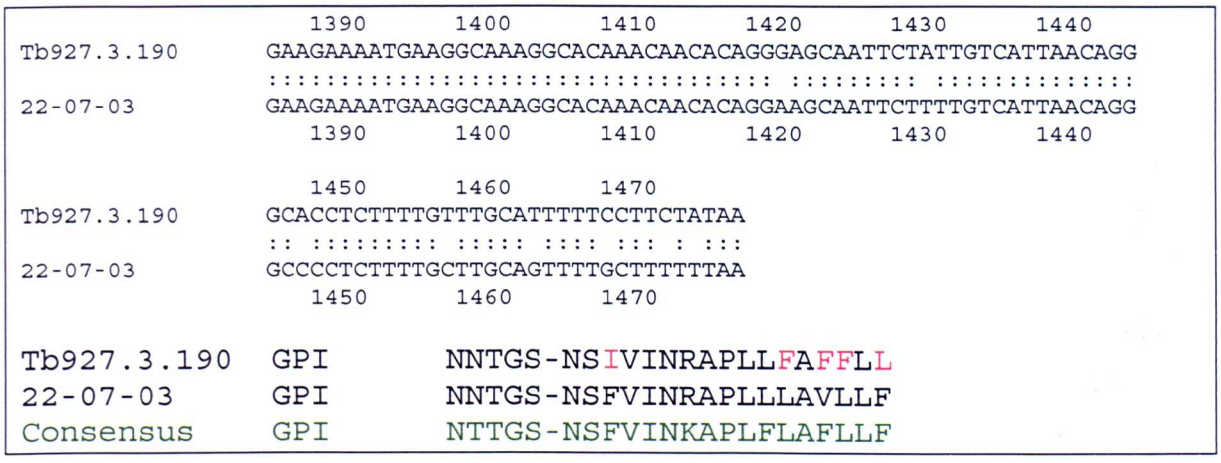


Figure 4.15: Nucleotide and amino acid alignment of the GPI signal region of Tb927.3.190 (donor) with its matching expressed clone 22-07-03.

Mismatches are suggested to be due to the presence of a 3' donor in clone 22-07-03.

The third in this subset of *VSG*s, Tb09.v4.0102, is a striking example of the requirements of a C-terminal gene conversion for the expression of intact *VSG*s with faulty GPI signals (classified as atypical or pseudogene depending on the nature of the fault). The donor in this case was classified as atypical with a type 3 C-terminal domain[30], and the gene conversion has "utilised" the first four cysteines in this domain and spliced them together with a type 2 GPI signal (See Figure 4.16).

```
                  370        380        390        400        410        420
Tb09.v4.0102 NTQVLEALLKAATNQNQGPEAVAGKAQQEKIEEECNKQDKDTDCTANPKCAWNEKAADPK
             :::::::::::::::::::::::::::::::::::::::::::::::::::::.:::::::
24-09-01     NTQVLEALLKAATNQNQGPEAVAGKAQQEKIEEECNKQDKDTDCTANPKCAWNDKAADPK
                    370        380        390        400        410        420

                    430        440        450        460        470        480
Tb09.v4.0102 KKCSLSEEAKQAVEKANQETGGKDGKPDCSKLTTQTECEAVNKDGKKYSGLRSGKDNEEE
             ::::::::::::
24-09-01     KKCSLSEEGKQPAENTAAGNQAGTDGKNTAGSNSFVINKAPLLLAVSLF
               430          440                          450

                    490        500
Tb09.v4.0102 KDKVKCRSASFLVNYKLSLSIAAGF
```

Figure 4.16: Amino acid alignment of the C-terminal domain of Tb09.v4.0102 (donor) with its matching expressed clone 24-09-01.

Highlighted in yellow are the cysteine residues. Letters in red indicate the GPI anchor signal sequence from the first position after the cleavage site.

This *VSG* activation, together with that mentioned at the end of section 4.4, accords with the model suggesting a gradient of recombination between C-terminal domain types, as proposed in Figure 3.9: it is a clear example that the beginning of the domain is more prone to intertype recombination and that interactions between different domain types occur predominantly at the site of the first four cysteines.

[30] CURATION = sequence has diverged from the expressed VSG consensus sequence; cysteine pattern not conserved in C-terminal domain, one cysteine missing in the second four- cysteine subdomain; GPI prediction uncertain.

4.7 Mosaic analysis

Seven putative mosaics were detected in three separate infections, at day 22 and 28 (see Fig 4.17). Only the first four (22-07-02, 22-07-04, 28-10-02 and 28-10-03) will be considered here, as the other three putative mosaics identified (28-10-07, 28-10-10, 28-11-09) are based on incomplete donor reads, rather than full-length donors, preventing as complete an interpretation of the data.

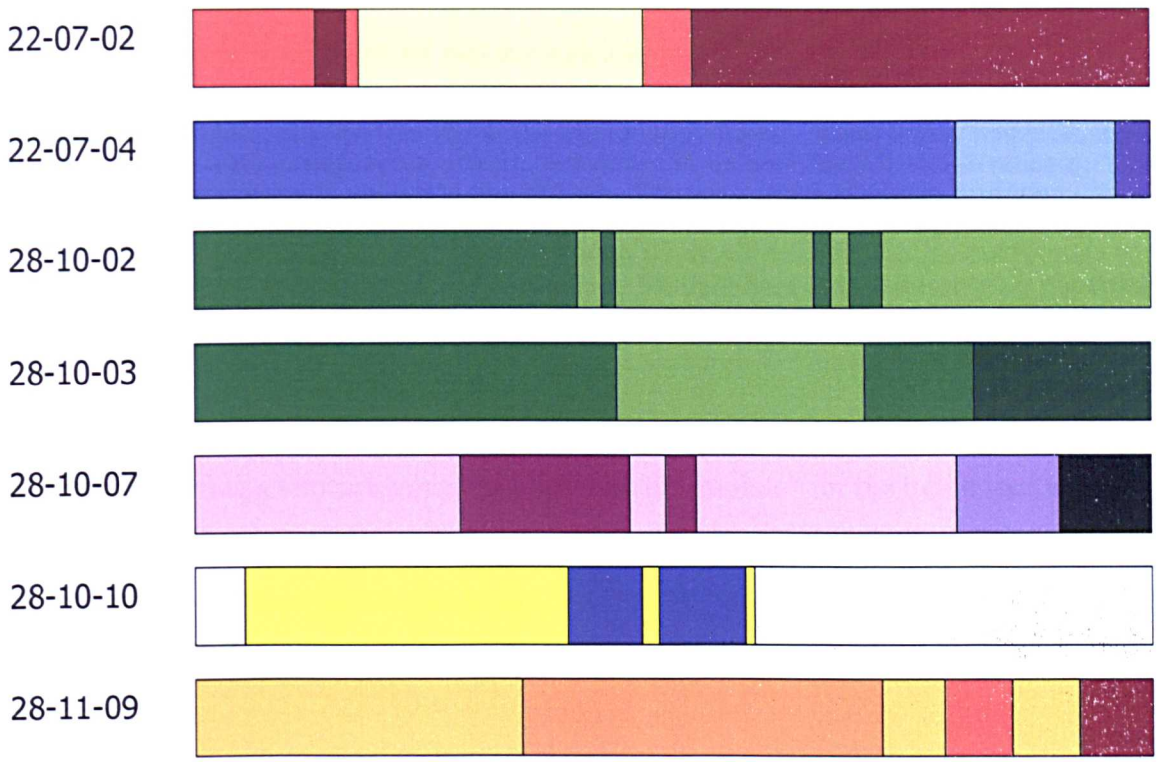


Figure 4.17: Depiction of the seven putative mosaics detected in the experiment. The different colours indicate the different donors contributing to the mosaic. The clone number is given to the left. The white regions in mosaic 28-10-10 lack identified donors.

Below (Figure 4.18) is an overview of putative mosaic 22-07-02, which is proposed to have been assembled from three donors. All three donors are pseudogenes: A is a type A6 *VSG* with a frameshift after bp 444 and a stop codon at bp 46 and it contributed the region spanning the C-terminal domain and a small segment of the N-terminal domain. B and C are also type A6 *VSG* pseudogenes with a single stop codon at bp 46 and bp 607 respectively. In addition, B has some irregularities before the cysteine pattern in the C-terminal domain, including CVC in the hinge region that could cause folding problems and has never been observed in any expressed *VSG*.

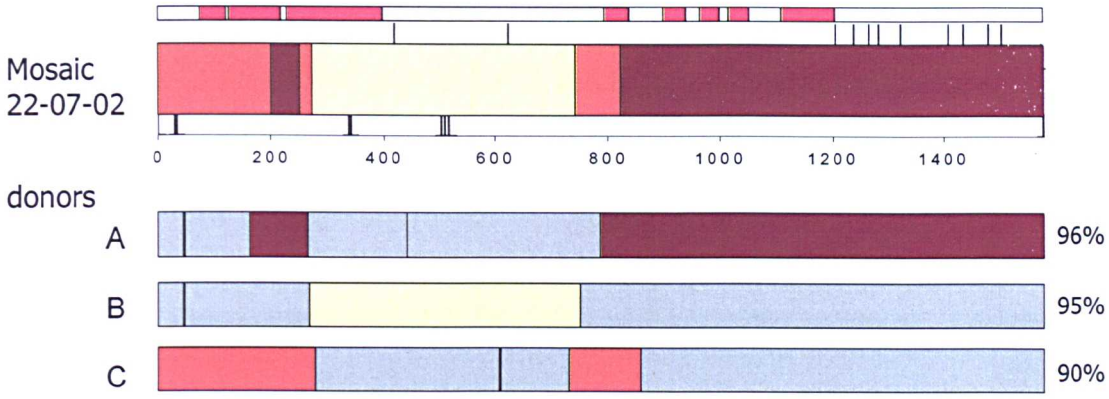


Figure 4.18: Mosaic gene 22-07-02 and its likely donors.

Donor A is Tb11.v4.0074, B is Tb10.v4.0161, C is Tb11.14.0001; beside each donor is the percent nucleotide sequence identity to the expressed mosaic. Each donor has indicated, in brown, yellow and orange respectively, the extent of its contribution to the mosaic. The overlap between donors is shown in the individual donor gene depictions, and the breakpoint between two donors in the mosaic is the average of the overlap. Below the mosaic are bars indicating sites of mismatch with the three donors. Above the mosaic is the cysteine pattern and alpha helical prediction. Vertical bars in the donor gene depictions represent frameshifts or stop codons, with the exception of those separating blocs of different colour.

The fact that this mosaic gene appears to have been constructed entirely from pseudogenes suggests that either a telomeric gene has acted as a "template" for the initial recruitment of one of the pseudogenes and/or that the assembly of the mosaic might have taken place stepwise in a silent expression site, as any direct gene conversion of these donor genes into the active expression site would have resulted in a trypanosome with a lethally defective coat.

Mosaic 22-07-04 is the most simple mosaic and is best described as a hybrid: two putative donors, Tb11.30.0005 (A) and Tb11.v4.0021 (B), contribute respectively to the N-terminal domain and to the C-terminal domain. In addition, a 3' donor sequence is postulated for the very end of the gene (see Figure 4.19). A is a pseudogene due to a single frameshift in the GPI signal, whereas B has been annotated as atypical due to the fact that the GPI signal is not well conserved in sequence. As noted previously in section 4.6, it appears also in this case that irregularities in the GPI signal of donor gene B result in expression being mediated by an additional GPI signal donor. A putative 3' donor has been identified in Tb09.354.0090, covering the last 60 bp, but the stretch is so short that other unknown *VSG*s resident at telomeres could also have acted as donors. Formation of hybrid genes appears to accommodate donors sharing a lower level of sequence identity, as localised identity at the C-terminal domain would be sufficient to catalyse the gene conversion reaction. Conversely, in the case of genes with mosaicism within the N-terminal domain,

the level of identity required between donors appears to be higher (see mosaic 22-07-02 above and mosaics 28-10-02 and 28-10-03, discussed in the next paragraph).

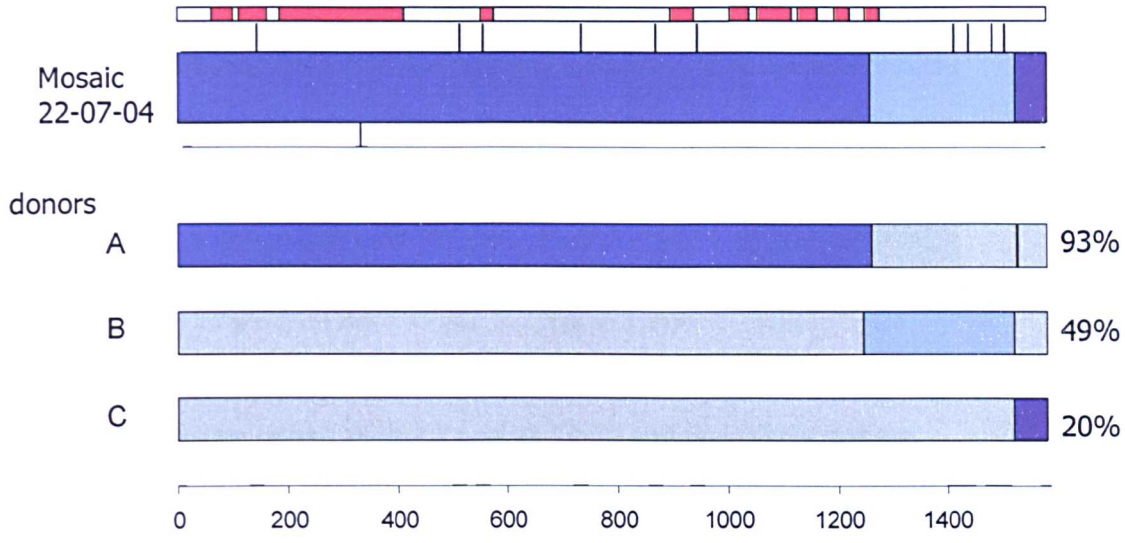


Figure 4.19: Mosaic gene 22-07-04 and its likely donors.

"A" is Tb11.30.0005, "B" is Tb11.v4.0021, "C" is Tb09.354.0090. Beside each donor is the percent protein sequence identity to the expressed mosaic. Below the mosaic is a bar indicating sites of mismatch with the three donors. Above the mosaic is the cysteine pattern and alpha helical prediction. Vertical bars in the donor gene depictions represent frameshifts or stop codons, with the exception of those separating blocs of different colour. Other general information is given in Figure 4.18.

Finally, mosaics 28-10-02 and 28-10-03, originating from the same donors, will be considered. It is interesting to observe the relationship between mosaics 28-10-02 and 28-10-03, as the presence of one shared gene conversion boundary (highlighted with black arrows in Figure 4.20) suggests that these mosaics emerged from the same predecessor. The dynamics of formation of this mosaic seem to be complex: gene A as been classed as functional (although it has a polar amino acid, E, in the hydrophobic extension, and could therefore have been classed as atypical: a borderline case), gene B is a pseudogene due to a single stop codon in the N-terminal domain and of gene C only the C-terminal part is known. The C-terminal domain is provided by either B in the case of 28-10-02 or a combination of B and C in the case of 28-10-03.

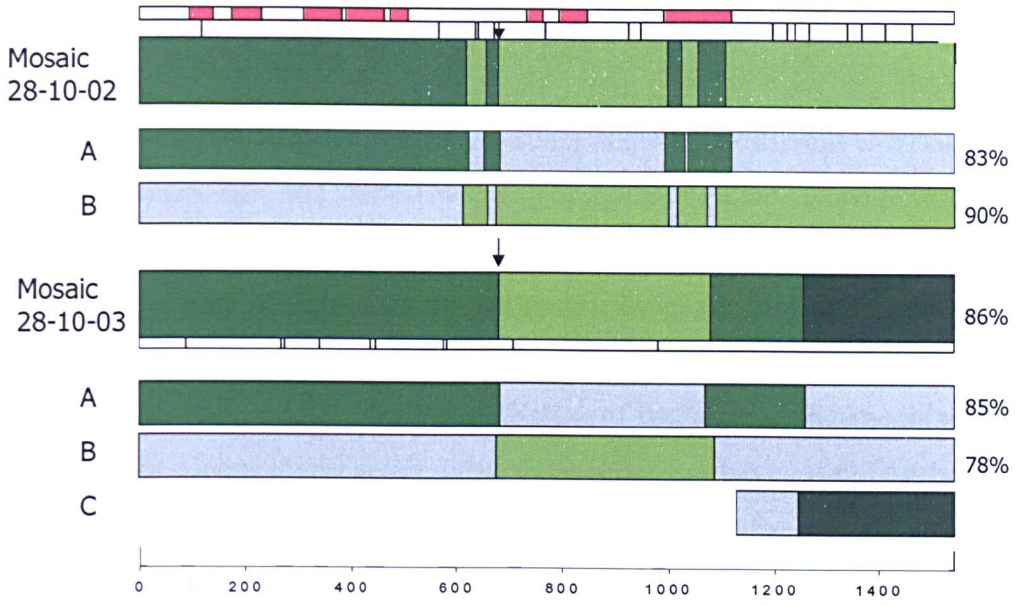


Figure 4.20: Mosaic genes 28-10-02 and 28-10-03 and their likely donors.

A is Tb11.09.0005, B is Tb11.0260, C is read AZ217061. Beside each donor is the percent nucleotide sequence identity to the expressed mosaic (beside mosaic 28-10-03 is the identity to 28-10-02). Below the mosaic are bars indicating sites of mismatch with the three donors (no mismatches are present in 28-10-02). Above mosaic 28-10-02 is the cysteine pattern and alpha helical prediction; these apply also to mosaic 28-10-03. Vertical bars in the donor gene depictions represent frameshifts or stop codons, with the exception of those separating blocs of different colour. Black arrows indicate the shared gene conversion event between the two mosaics. Other general information is given in Figure 4.18.

A diagram highlighting the possible development of this mosaic is shown below (Figure 4.21).

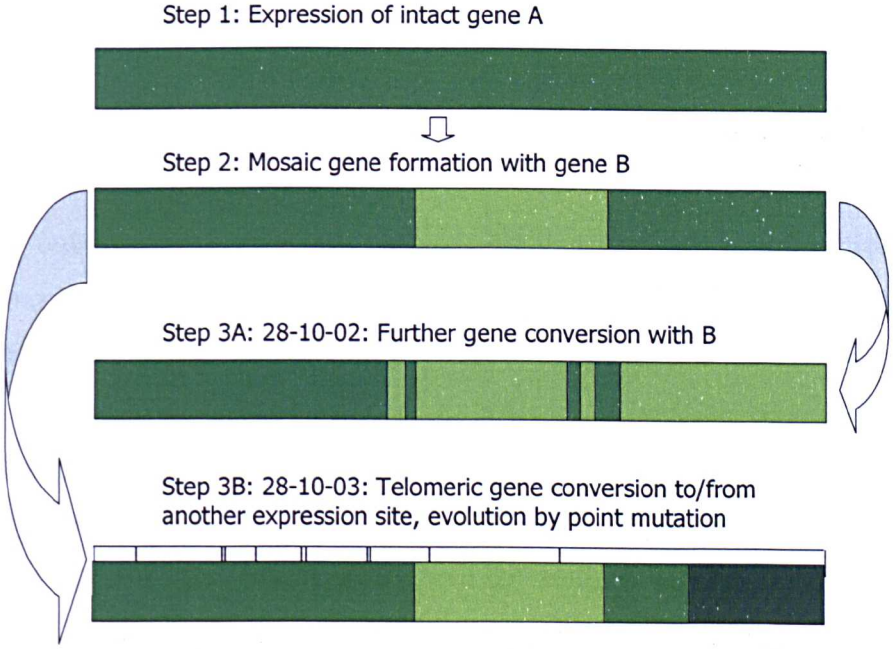


Figure 4.21. Diagram of possible development of mosaics 28-10-02 and 28-10-03.

Possibly at an earlier stage A was expressed and then it gave rise to a simple mosaic, as in step 2 in Figure 4.21, with sequence coming from gene B. After this, the two mosaics seem to have taken different routes: 28-10-03 acquired a different C-terminal end, possibly from a telomeric site, and started evolving by point mutation (10 point mutations are present in the N-terminal domain). 28-10-02 had further interaction with gene B and the array copy of gene A to produce a more complex mosaic: gene B replaced gene A at the C-terminal domain and a total of seven different transitions between the donors A and B are present in the end product. The colocalisation of further rearrangements with putative non-helical surface loops might suggest that these processes have altered (or be in the process of altering) the exposed epitope. This example highlights the possible interaction between two diversification processes, namely mosaic gene formation and point mutation. It still leaves unanswered the question of whether this diversification is producing new non-crossreactive antigens: this will be discussed further in section 4.9.4.

When comparing mosaics from day 21 with those from day 28 there appears not to be an increase in mosaic complexity. If anything, the earlier mosaics appear to be more complex, but when looking at relatedness between donors there is a marked drop in sequence identity: from 87-95% at day 21 to 73% at day 28. A longer infection would have to be run and more mosaics would need to be analysed in order to test whether this is likely to be a general trend; it is nevertheless possible to envisage this as a convincing possibility, as there are few closely related *VSG*s in the archive and interactions producing viable mosaics may eventually have to rely on more fortuitous conversions, using only limited and therefore localised sequence similarity.

4.8 Mosaic analysis: experimental confirmation by PCR

In order to confirm the bioinformatic predictions made in the previous section, it was important to determine experimentally whether putative mosaics had been assembled in the course of an infection or were actually intact genes that had been assembled as mosaics at a time before the infection started. In order to differentiate between these two possibilities, the approach taken was to design a PCR-based experiment in which primer pairs would distinguish between donors and would amplify from only the mosaics, only the donors, or both mosaics and donors. Results of this analysis are presented only for mosaic 28-10-02, as PCRs for the remaining mosaics were not completed due to lack of time. Figure 4.22 gives a graphical overview of the experiment, and Table 4.5 gives more details for each PCR, including the expected result.

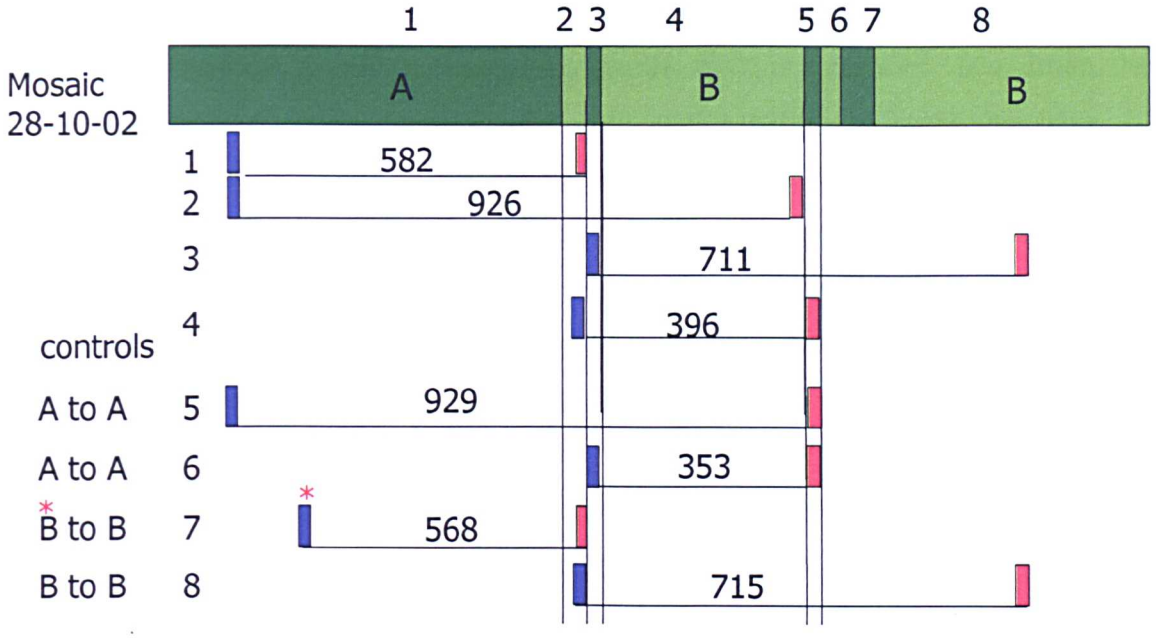


Figure 4.22: Experimental confirmation by PCR of mosaic 28-10-02.

Blue blocks indicate forward primers, red blocks reverse primers. Vertical lines delineate the recombined tracts to which primers anneal. Numbers above the lines connecting the sets of primers indicate the size (bp) of each expected PCR product. The first four reactions (1-4) are specific to the mosaic, reactions 5-8 amplify either donor A or donor B. The forward primer of reaction 7 is indicated by an asterisk, as it anneals specifically to the non-converted N-terminal section of donor B and therefore does not relate to the mosaic depiction given above. Numbers above the mosaic indicate the different gene conversion tracts, for ease of discussion.

Table 4.5: Details of PCRs to confirm mosaic 28-10-02, with PCR result predictions.

Pr1 and pr2 give the names of the two primers used in each PCR. "Don pr1" and "Don pr2" tell which donor the primer anneals to (A or B, followed by gene conversion tract number). "Length product" is the length of the expected PCR product. "Pre" stands for preinfection 927 GUTat 10.1 gDNA, "28-11" for gDNA from the other infection at day 28; "28-10" indicates gDNA of the infection from which the mosaic was isolated. "28-10-02" indicates the plasmid DNA of the clone corresponding to the mosaic.

Pcr n	Pr1	Pr2	Don pr 1	Don pr 2	Length product (bp)	Pre/ 28-11 gDNA	28-10 gDNA	28-10-02 clone	28-10 CDNA +
1	10-2dA_95for	10-2dB_657rev	A1	B2	582	n	y	y	y
2	10-2dA_95for	10-2dB_1001rev	A1	B4	926	n	y	y	y
3	10-2dA_671for	10-2dB_1362rev	A3	B8	711	n	y	y	y
4	10-2dB_648for	10-2dA_1024rev	B2	A5	396	n	y	y	y
	controls								
5	10-2dA_95for	10-2dA_1024rev	A1	A5	929	y	y	y	y
6	10-2dA_671for	10-2dA_1024rev	A3	A5	353	y	y	y	y
7	10-2_Db_89for	10-2dB_657rev	B	B2	568	y	y	n	n
8	10-2dB_648for	10-2dB_1362rev	B2	B8	715	y	y	y	y

The PCR results given in Figure 4.23 confirm the expected results and give backing to the initial suggestion from bioinformatic analysis that 28-10-02 is a mosaic[31]. In addition, they validate against the phenomenon of template copy choice that might fortuitously have yielded some "*in vitro*" mosaic genes (see section 4.2.4 for a discussion of template copy choice). The variable signal given by different primers is likely to be due to difference in efficiency between primers, as the fluctuations are visible consistently throughout the eight set of reactions. It is of interest that amongst the gene conversion tracts validated are also three very short regions, B2, A3 and A5 (average value 38, 23 and 28 bp respectively), suggesting that these short gene conversions are not artefacts or explainable with the presence of unsequenced *VSG* genes. Short gene conversions are the most likely to produce changes in exposed epitopes, as longer gene conversion events tend to reduce rather than promote diversity.

[31] PCR reaction 2 is applicable also to mosaic 28-10-03, which is therefore validated as well.

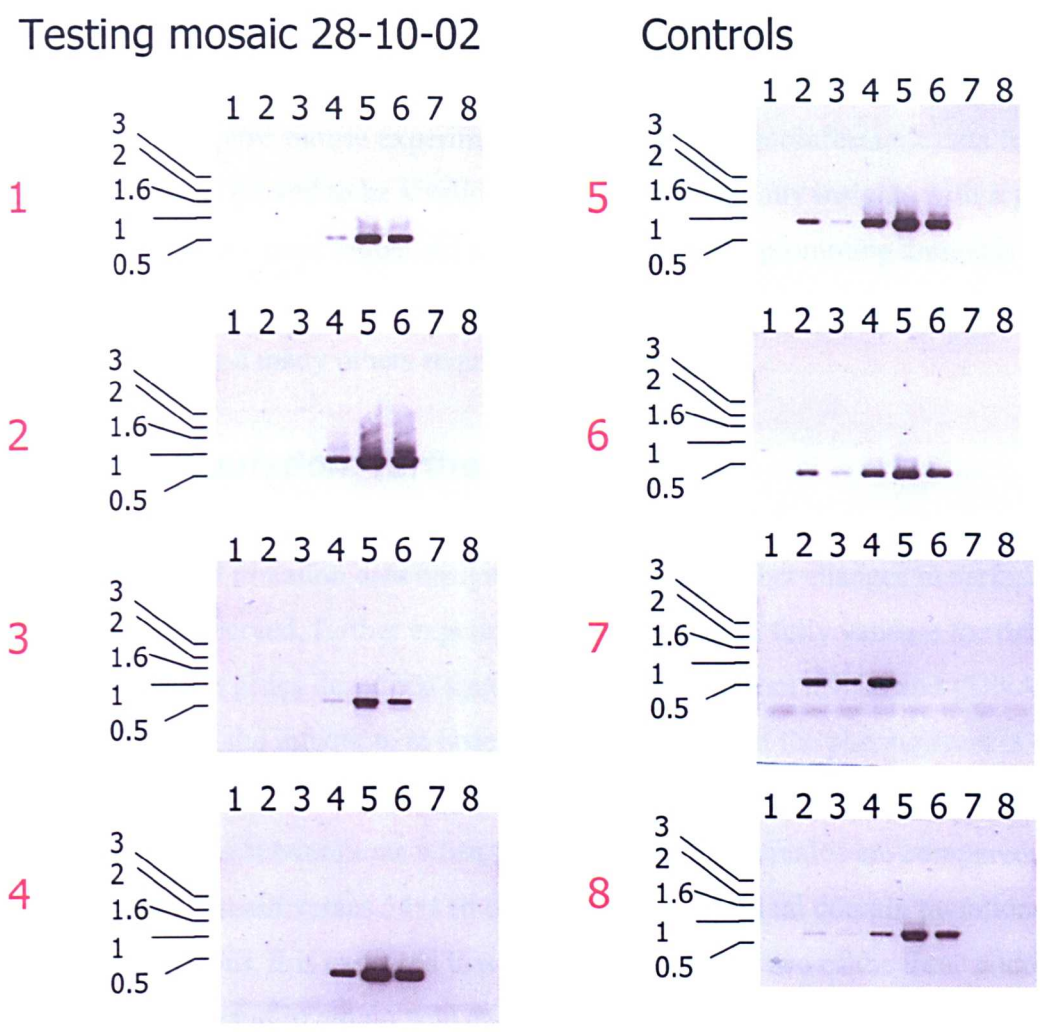
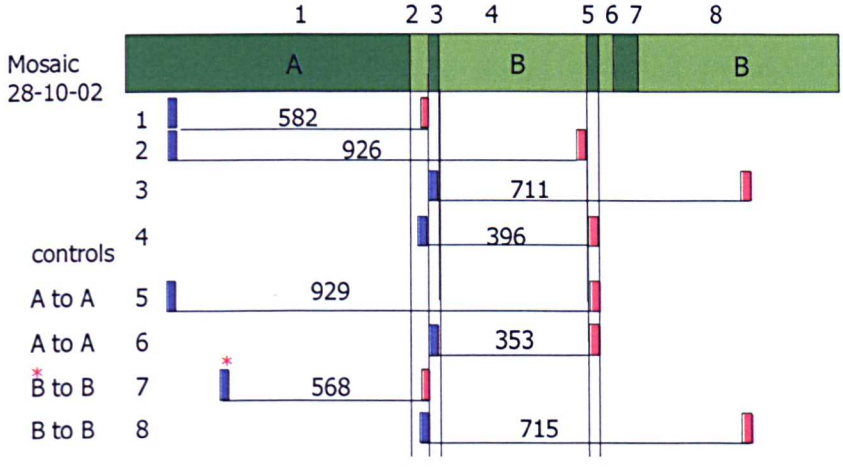


Figure 4.23: Ethidium stained gels showing PCRs to test whether 28-10-02 is a mosaic sequence.

In each gel the following samples are loaded: lane 1, strain 427 gDNA (negative control); lane 2, strain 927 gDNA pre-infection (negative control); lane 3, gDNA from infection 28-11 (negative control); lane 4, gDNA from infection 28-10 (positive control); lane 5, plasmid DNA of clone 28-10-02; lane 6, cDNA + from infection 28-10; lane 7, cDNA - from infection 28-10; lane 8, no DNA control. Primer pairs for the 8 reactions (numbers in red) are given in Table 4.5. To the left of each gel is the size (in kb) for the main bands of the 1 kb ladder. Below is a smaller version of Figure 4.22, for ease of consultation of the PCR results of the gels above.



4.9 Conclusions

Combining the *in vivo* mouse experiment with the *in silico* bioinformatic data from the 927 genome strain has proved to be a valid approach yielding many insights with regards to our understanding of the gene sequences and underlying forces promoting antigenic variation in *T. brucei*. It should nevertheless be considered a pilot study in which many points are left unanswered and many others require validation.

4.9.1 Point mutation: further experiments

Whereas the point mutation data analysed thus far suggest that changes in surface epitopes of VSGs were observed, further experiments are required to fully validate the data. First of all, a subset of the genes described should be reamplified from cDNA and gDNA from before and during the infection, in order to add certainty that the phenomenon is occurring during the infection, as the data suggest. There appears to be a difference in the percentage of nonsynonymous substitutions when N- and C-terminal domains are compared: 70% in the N-terminal domain versus 54% (6 out of 11) for C-terminal domain mutations. Due to the nature of codons, it is expected that random mutation at two of the three codon positions would result in amino acid change, making the value of 70% substitution not significant *per se*. Whereas the clustered location of mutations and their statistically significant association with putative surface loop regions suggests a process of positive selection, no evidence could be gathered, by running a dN-dS test[32] using a recently developed online software for detection of positive and negative selection (http://www.datamonkey.org/) (Pond and Frost, 2005). A possible explanation is the low number of point mutations per clone due to the putative recent mutation of *VSG* 19 (see next paragraph) and/or the small number of clones analysed (six clones of *VSG* 19, with 49 mutations).

Two indirect lines of evidence suggest that mutations are accumulating in expressed *VSG*s: firstly, as shown above in section 4.7, the onset of point mutations in the case of mosaic 28-10-03 can be tied to the prior expression of an earlier form of mosaic 28-10-02 (see Figure 4.21). Secondly, multiple clones corresponding to *VSG* 19 were shown to have different point mutations, suggesting a recent acquisition of mutations, subsequent to *VSG* expression. The presence of only eight shared mutations out of 49 further suggests that the

[32] The dN-dS test gives a statistical validation for the likelihood that nonsynonymous substitutions are positively selected for, allowing to establish the significance of mutations found in different alleles of the same gene.

process of accumulation must be rapid, leading to divergence at the population level. Further sequencing of cDNA clones might provide a clearer estimate of the extent of this diversification process, and possibly (if more diversity is uncovered) enable application of the dN-dS test to a statistically valid number of data, overcoming the limitations of the currently small dataset. Further considerations on the possible significance of point mutations will be made in the Discussion (Chapter 6).

4.9.2 VSGs with a 3' donor: further experiments

The 3' donor analysis also requires experimental validation, with a similar PCR approach as that used for mosaic genes in section 4.7, in order to define whether the duplication occurred during the infection. Evidence gathered so far points to the requirements of a C-terminal end donor for most array *VSG*s, but further experiments generating a larger sample size could allow more precise definition of the *in vivo* requirements of GPI signal sequences, by observing whether a pattern emerges with regards to the GPI signals that are not "allowed" into the expression site. This in turn could shed light on the grey area constituted by "atypical" array *VSG*s (See Chapter 3).

4.9.3 Mosaic genes: further experiments

As for mosaic genes, support has been gained for the tested hypothesis that they contribute significantly to antigenic variation when the "degenerate" set of array genes starts to be utilised. Half of the 14 expressed, putative array genes were shown, at least bioinformatically, to be mosaic genes, and for two of them this was confirmed experimentally. As information on donor genes was insufficient in three out of seven cases and two mosaics were derived from the same donor, the actual sample size from which to draw meaningful conclusions was reduced to three, or two if the hybrid gene is discounted. A rabbit infection allowing collection of daily samples would enable more accurate definition of the onset of mosaic genes, together with their development and changes in donor composition, expanding from the tentative dynamic framework suggested in this current study.

CHAPTER 5

VSG-RELATED GENES: BIOINFORMATICS AND MOLECULAR ANALYSIS

5 *VSG*-Related Genes: Bioinformatics And Molecular Analysis

5.1 Introduction

The bioinformatic analysis of the *VSG* archive, as outlined in Chapter 3, unveiled the presence of a small gene family that shares features with *VSG*s but is nevertheless distinct enough to warrant the use of the term *VSG*-related, as will be argued in this chapter. The chapter is divided into three parts, the first presenting the bioinformatic evidence for *VSG*-related as a separate gene family (referred to in the remainder of the chapter as the *VR* gene family) and the second giving a preliminary experimental characterisation of *VSG*-related expression. Due to the similarities between *VR* genes and *T. congolense* VSGs (see Section 3.2.3), the third part will give a brief description of what is known of *T. congolense VSG* arrays, looking for shared themes between *T. congolense VSG*s, *T. brucei VSG*s and *VR* genes.

5.2 *VSG*-related genes: bioinformatic analysis

5.2.1 Full-length VRs *vs* VSGs

An initial distinction between *VSG* and *VR* genes is clearly made by aligning sequences of these two gene families. Figure 5.1 shows a full-length alignment of 35 functional *VSG*s with 29 *VR* genes. Three clusters are formed, *VSG*s with a type A N-terminal domain, *VSG*s with a type B N-terminal domain and *VR* genes.

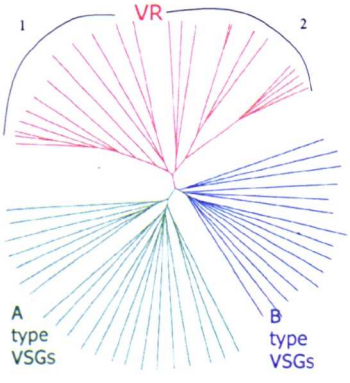


Figure 5.1: Tree based on amino acid alignment of 35 full-length functional *VSG*s with 29 VR genes.

Numbers 1 and 2 indicate the two main VR gene clusters, further illustrated in Figure 5.2.

After this preliminary distinction was made with *VSG*s, the next step taken was to identify clusters within *VR* genes: the 29 *VR* genes analysed fall into two main clusters, 1 and 2, with cluster 3 forming a smaller cluster and being related to cluster 1 (see Figure 5.2). Cluster 1 contains *VR* genes with short N-terminal domains while cluster 2 genes have longer N-terminal domains, very similar to those of type B *VSGs*. This similarity is further investigated in the next section.

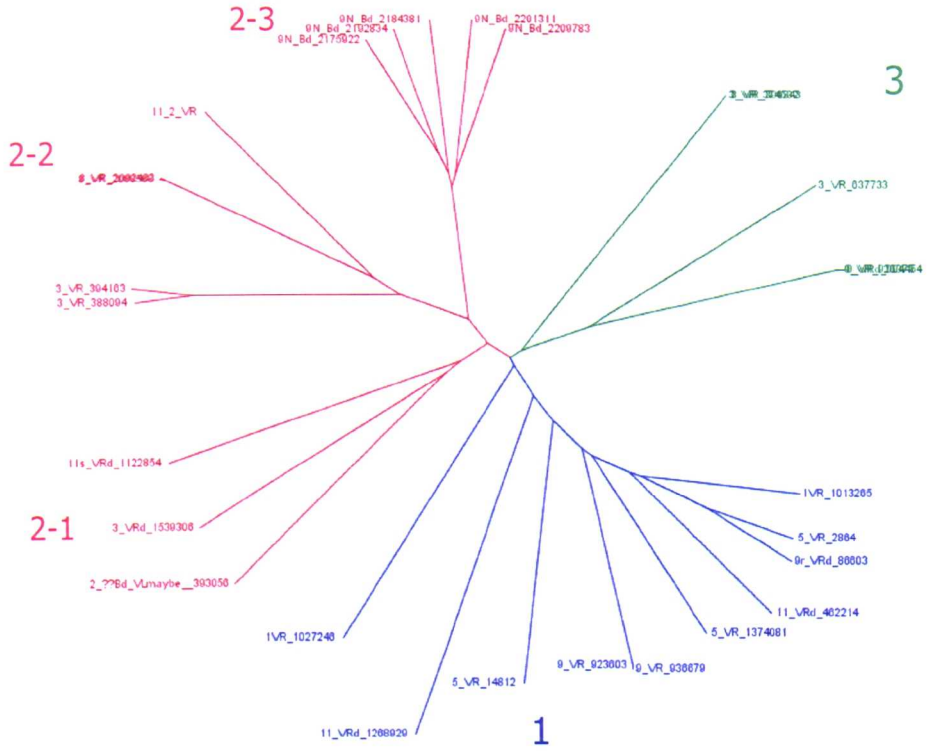


Figure 5.2: Tree based on amino acid alignment of full-length *VR* genes.

Three main clusters were defined: cluster 1 (blue), 2 (red) and 3 (green). Cluster 2 was further partitioned in three subclusters, 2.1, 2.2. and 2.3, cluster 2.1 including *VR* genes encoding cysteines at their C-terminal end. The two longest branches of cluster 1 have been termed 1o (outgroup within cluster 1), see Table 7.9 of Appendix for details.

5.2.2 VSG *Type B and* VSG*-related N-terminal domains*

Due to the close similarity of *VR* genes to type B *VSG* N-terminal domains, it was of interest to see whether it was possible to characterise the *VR* N-terminal domain region and identify potential motifs or signature sequences that could enable its differentiation from *VSG*s. The first step was to align the 29 *VR* N-terminal domains with 332 type B N-terminal domains (see Figure 5.3): *VSG*-related N-terminal domains clustered more closely with what was defined in Chapter 3 as *VSG* N-terminal domain type B group 2.

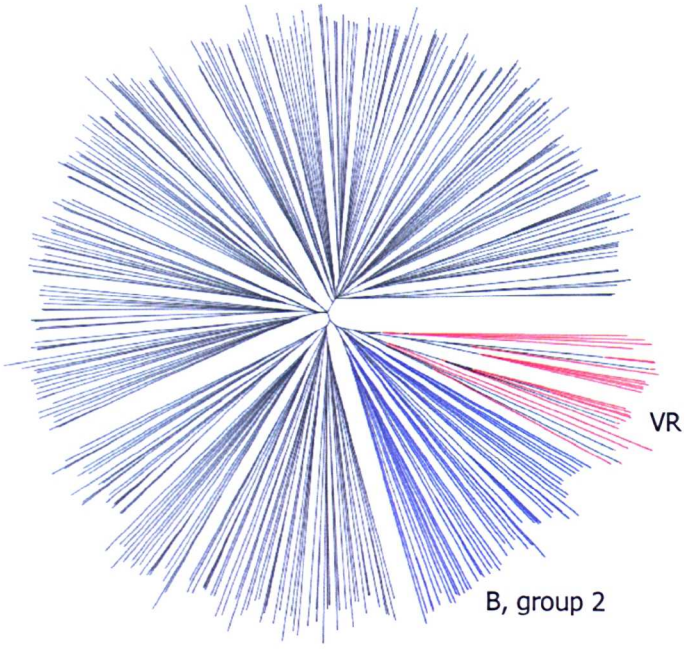


Figure 5.3: *VR* N-terminal domains aligned with 332 *VSG* type B N-terminal domains.

In red are *VR* N-terminal domains (VR), in blue type B group 2 N-terminal domains (B, group2). The three other main groups (in grey), are type B groups 1, 3 and 4 (clockwise from type B group 2).

Focusing on a more detailed comparison with group 2, and looking for differences from *VR* genes, it appears that there is no localised difference in sequence that can enable distinction between the *VSG* type B domains and *VR* N-terminal domains; the cysteine pattern is remarkably conserved (as is generally the case across all type B N-terminal domains, see Chapter 3) and so are other conserved residues (data not shown). The only significant departure in terms of structure is in *VR* N-terminal domain cluster 1 (see Figure 5.4), which have a shorter N-terminal domain, with a deletion of ~100 amino acids after the first conserved cysteine, spanning the long domain stretch with no cysteines (see Figure 3.5, Chapter 3, for cysteine spacing of type B VSG domains). So, with this sole exception, it appears that there is a continuum of variation throughout the domain, rather than different clusters with more or less significant structural differences. Within this continuum there is no overlap, suggesting that *VR* genes are genetically isolated from the recombination reactions that lead to shuffling of sequences in the case of *VSG* genes.

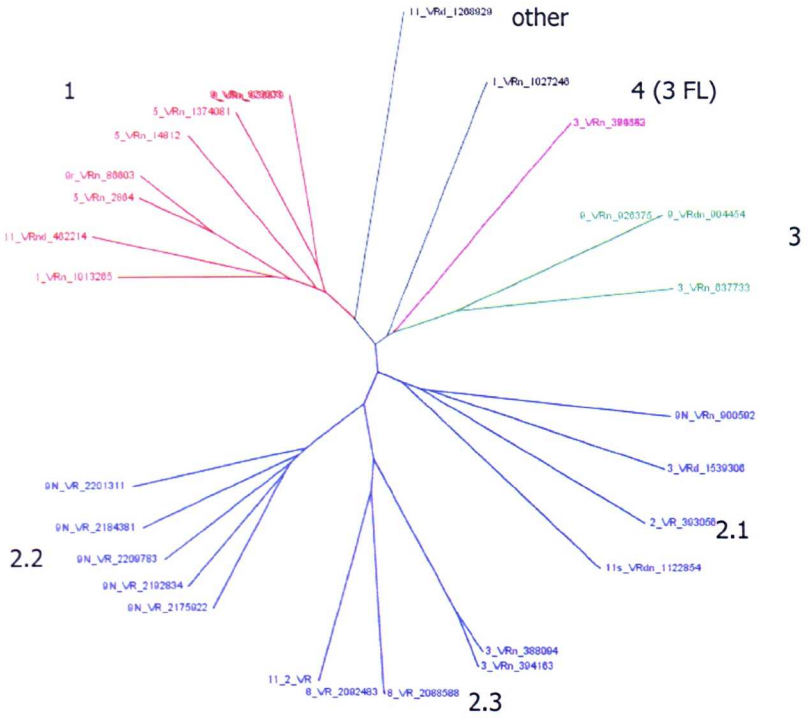


Figure 5.4: Tree based on amino acid alignment of *VR* N-terminal domains.

Cluster 1 and 2 are the main clusters. Cluster 2 can be divided into three subclusters, 2.1, 2.2 and 2.3. Cluster 3 of full-length VR genes is here divided into clusters 3 and 4 (the latter is further qualified in the graph by "3 FL" (= full-length cluster 3), with two additional sequences not clustering well (labelled as "other", corresponding to the full-length cluster 1 outgroup, see Figure 5.2).

The tree for *VR* gene N-terminal domains is remarkably similar to the tree for full-length sequences, as occurs for *VSG*s. Domains were divided into four clusters: cluster 1 and 2 correspond to full-length clusters 1 and 2, respectively, while full-length cluster 3 is here split into N-terminal domain clusters 3 and 4; two *VR*s (the most diverged of cluster 1 full-length, referred to 1o, cluster 1 outgroup) do not group well with any of the four clusters.

5.2.3 VR *C-terminal domains*

A comparison between all 85 functional *VSG* C-terminal domains (and respectively two and 13 atypical type 4 and 5 C-terminal domains) and 29 *VR* C-terminal domains was made (see Figure 5.5). The tree shows clear separation of *VR* C-terminal domains.

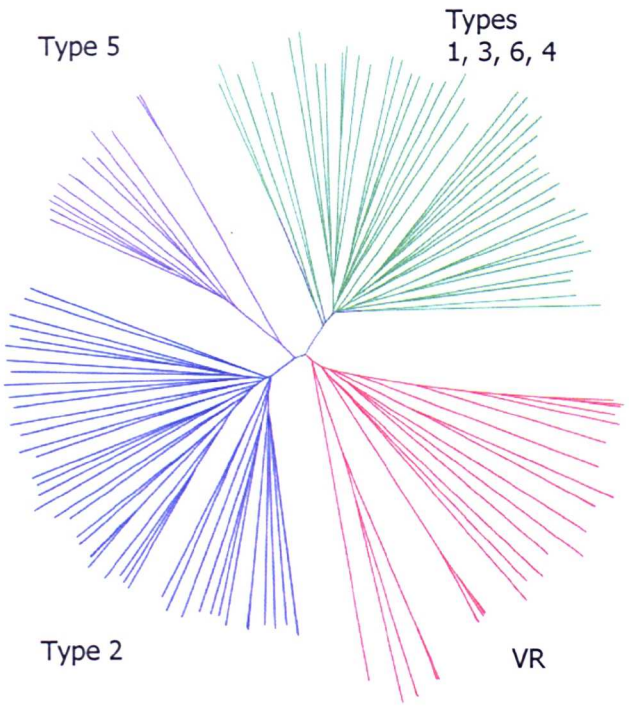


Figure 5.5: Tree based on amino acid alignment of *VR* C-terminal domains (indicated as *VR*) with all 85 functional *VSG* C-terminal domain types (indicated as types 1 to 6), plus two atypical type 4, and 13 atypical type 5 C-terminal domains.

The separate analysis of the C-terminal domain of *VR*s was prompted by the presence of a subset of genes with cysteine residues at their C-terminal end, reminiscent of *VSG* C-terminal domains. All the sequence downstream of the conserved N-terminal cysteine pattern was considered (90-130 amino acids). Three C-terminal domain clusters were identified, 1 and 3 being more related to each other and cluster 2 being composed predominantly of cysteine-containing C-terminal domains (See Figure 5.6). It appears that there is a potential for a GPI anchor in clusters 1 and 2, while cluster 3 appears not to have a hydrophobic extension (C-terminal domain alignments are given in the Appendix, section 7.4.2). A significant presence of proline residues was detected in domains with no cysteines, a similarity with *T. congolense* C-terminal ends (see Introduction, section 1.2.2). This was further explored by comparing VR and *T. congolense* VSG C-terminal domains, it appears that although both are rich in proline, the sequence motifs are distinct, and clear separation between the two families is present (see Figure 7.5 of Appendix for a multiple sequence alignment)

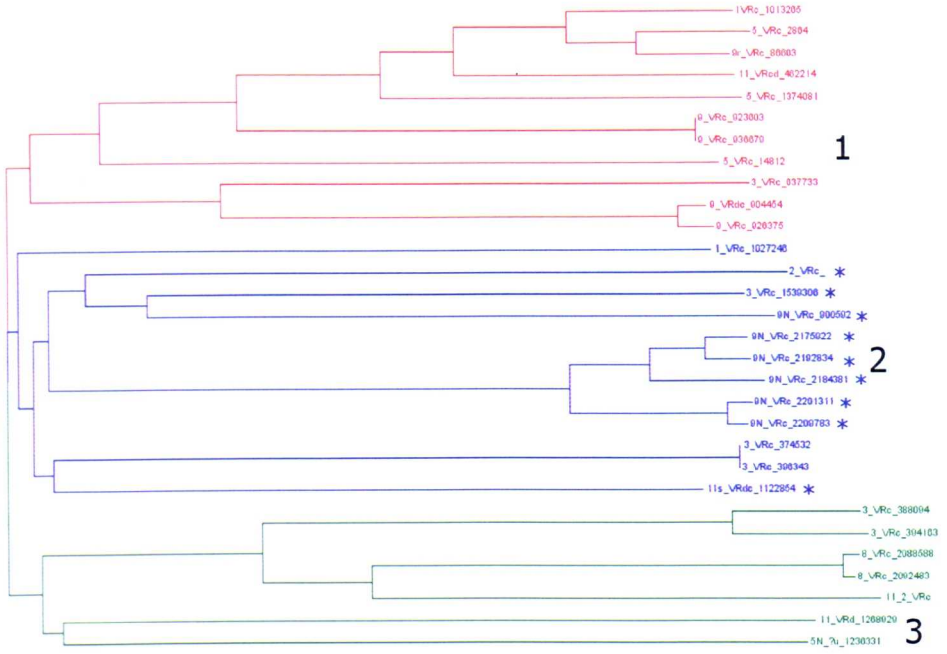


Figure 5.6: Tree based on amino acid alignment of *VR* C-terminal domains.

Three main clusters (1-3, colour-coded) are present. Cluster 2 is mainly composed of domains with cysteine residues (highlighted with asterisk).

5.2.4 N- and C-terminal domain analysis

When looking at how the C-terminal domain clusters relate to the N-terminal domain ones (Table 5.1), it is apparent that, unlike for *VSG*s, there is a tight association between the two domains, as no domain combinations appear to be present. This suggests a different pattern of evolution from that of *VSG*s, in which the C-terminal end of *VR* genes does not provide a 3' region of homology to allow N-terminal domains to recombine with each other.

Table 5.1. Relation between full-length (FL), N and C clusters.

FL CLUSTER TYPE	N CLUSTER TYPE	C CLUSTER TYPE	OCCURRENCE (NUMBER OF VR)
1	1	1	8
2.1	2.1	2.1	4
2.2	2.2	3	5
2.3	2.3	2.1	5
3	3	1	3
3	4	2	2
Other[33]			
1o	N[34]	2	1
1o	N[2]	3	1

[33] FL cluster 1o = 1 outgroup, more diverged.

[34] Does not cluster.

5.2.5 *Flanking regions of* VR *and* VSG *genes*

It was of interest to compare the *VSG* and *VR* flanking regions to see whether some common traits were found. For the downstream region, ~300 bp were used in the comparison, corresponding to the length of sequence between the end of a *VSG* and the beginning of the next set of 70-bp repeats in the array (see Chapter 3, section 3.7 for *VSG* array structure). In the case of the upstream region, 300 bp immediately upstream of the start codon were used in the alignment. Figure 5.7 shows the tree generated from the upstream and downstream regions. It can be seen that, whereas the 3' region separates well, with short branch lengths, the 5' region, especially in the case of *VSG*s, is very diverged, leading to a degree of overlap between *VSG* and *VR* regions that is probably due to the weakness of the alignment. In addition, 3' and 5' UTRs tend to group in a similar way to the coding sequence, when the different trees are compared (data not shown). An observation made with regards to the *VSG* 3' region is that the two separate clusters correlate with the presence or absence of the 16-mer motif present in the 3' UTR of all expressed *VSG*s (see section 1.3.4 of Introduction). This could suggest recombination events between expression sites and *VSG* arrays.

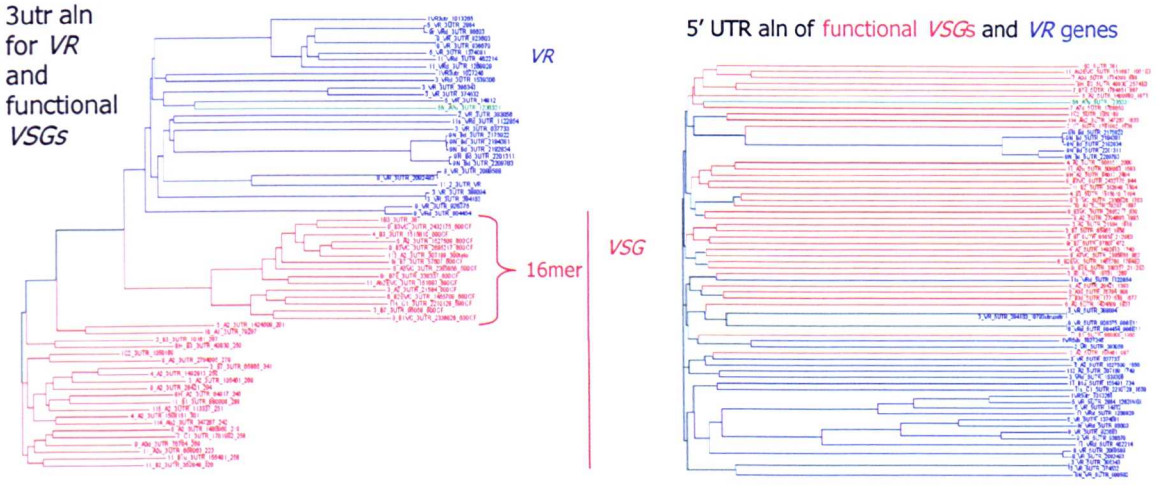


Figure 5.7: Alignment of 5' and 3' flanks of functional *VSG*s and *VR* genes.

Sequence in green represents Tb927.5.3990, the only gene with intermediate features between *VSG* and *VR* genes. *VSG* genes with 16 mer in 3' UTR are highlighted.

5.2.6 VR *gene location*

VSG-related genes were allocated a number from 1 to 29 (see Figure 5.8 and Table 7.9 of Appendix for correspondence between this identification and the GeneDB ids) according to their location in the genome, from chr 1 to chr 11. Twenty *VR* genes out of 29 are organised in arrays composed of between two and five genes, and *VR* genes occupy a total of 15 different loci. Of these, 13 are internal and two (*VR* 10-11 and *VR* 26) are telomeric[35]. Twelve of the 15 loci are located just upstream of a strand switch, and 9 of them are associated with *ESAG* genes (*ESAG1*, *2*, *3*, *4*, *5*, *6*, *9*, *11*, but in particular *1*, *2*, *9* and *11*). A tight association between a dispersed subset of 8 *VR* genes and *ESAG9*[36] 2-3 kb downstream of these genes was found (all *VR* with this association belonged to cluster 1, see Figure 5.8), whereas other *ESAG*s appear not to form a precise pattern of association with *VR*s. None of the regions upstream of *VR*s were found to contain 70-bp repeats. Although no particular DNA features have been ascribed to strand switch regions, they appear to match closely with the breaks of syntheny found when comparing *T. brucei* with *Leishmania* and *T. cruzi* (H. Renauld, pers. comm.). This suggests that these regions might be more prone to recombination and might have been telomeric, before chromosome fusions occurred in *T. brucei,* as mentioned in section 1.1.3 of the Introduction (Ghedin *et al.*, 2004). The current association of a subset of *ESAG* genes with telomeres at the Bloodstream expression site (BES) further suggests an ancestral telomeric location for all *ESAG* genes and *VR* genes.

[35] More recently, another *VSG*-related gene has been found upstream of a telomeric M*VSG* expression site on chromosome 10 and *VR* 25 has been relocated in a discontinuity within the subtelomeric *VSG* array in the left arm of chromosome 9.

[36] Of these eight *ESAG9* genes, six are intact and two are pseudogenes.

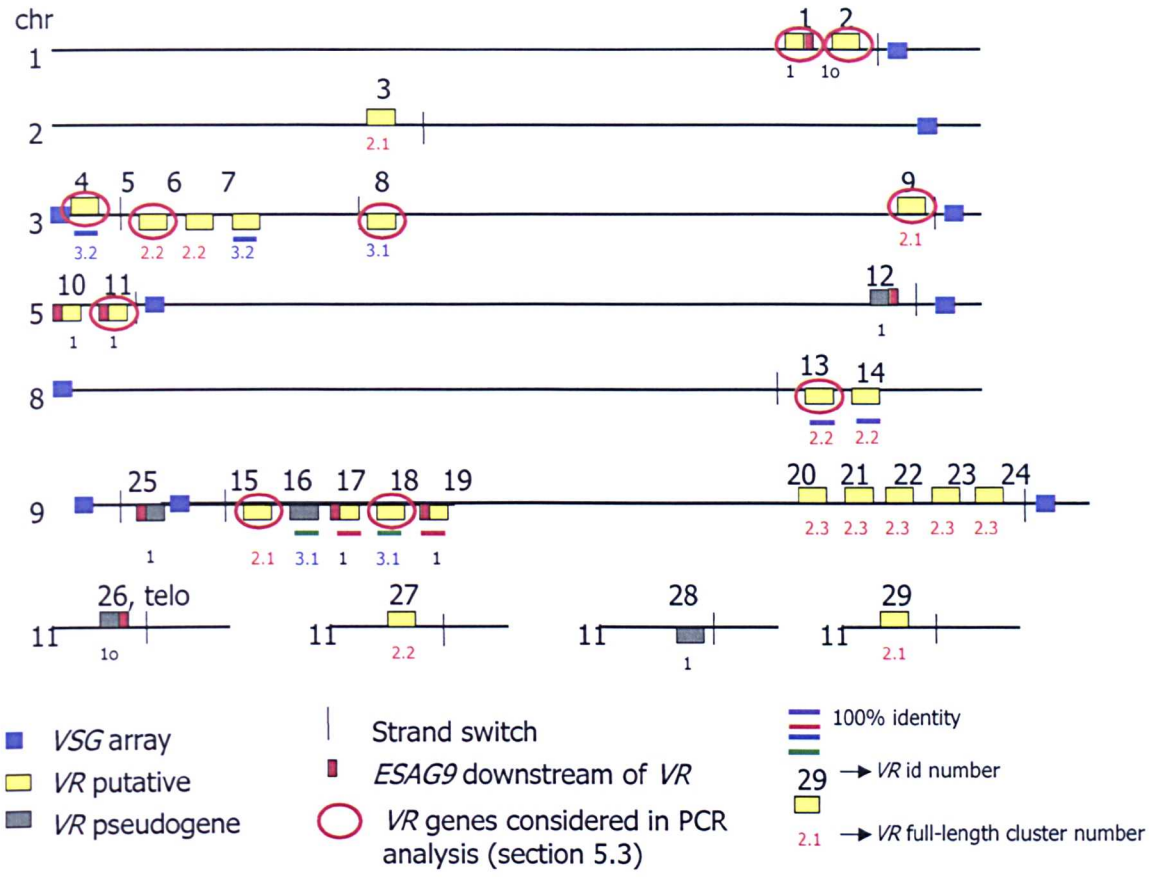


Figure 5.8: Chromosome location and basic features of *VR* genes.

To the left is the chromosome number. The *VR* id number is given above each *VR* gene, whereas full-length *VR* cluster types are highlighted below. *VR* 25 to 29 are located in small contigs that have been assigned to chr11. pairs of *VR* genes underlined with the same colour share 100% sequence identity. *VR* 25 was previously located in a contig, but it now maps to a discontinuity in the array at the left arm of chr9. *VR* genes circled in red were selected for the PCR analysis presented in section 5.3.

5.2.7 Overview of features of VSG-related genes

In summary, 31 *VSG*-related genes have been found to form a separate group, distinct from 70-bp associated array *VSGs*[37]. Twenty-nine *VR* genes were analysed and a full-length alignment shows clear separation from 35 functional *VSG* sequences (Figure 5.1). Their N-terminal domain is type B, but forms a separate group when compared with all other type B *VSG* N-terminal domains. The C-terminal domain is not closely related to any of the conserved *VSG* C-terminal domains and forms a separate cluster. With regards to the "integrity" of *VR* sequences, there are few pseudogenes, with only seven out of 31 (22%), compared with a minimum estimate of 86% for the *VSG* archive. These estimates are very

[37] This study considers 29 of the 31, the other two *VR* genes have more recently been found: they are not included in the analysis, but are included Table 7.9 of the Appendix.

tentative: as no consensus for expressed *VR* genes is present, it is possible that the proportion of pseudogenes might be slightly higher.

Together with these differences at the gene level, when looking at the location of these genes, it is striking that they are not found to be interspersed amongst *VSG*s. They are predominantly located internally, as single genes or in small arrays. On the contrary, *VSG*s are located within 70-bp repeat-associated arrays that are almost exclusively subtelomeric (with the exception of a short internal *VSG* array on chromosome 7). *VR* 5' and 3' UTRs cluster separately from *VSG* UTRs. A common feature 5' of *VSG*s in subtelomeric arrays is the presence of 70-bp repeats; these are absent from the region upstream of *VR* genes. Only one *VSG,* Tb927.5.3990, appears to have features intermediate between these two groups. Its 3' UTR groups with *VSG*-related, its 5' UTR tentatively groups with *VSG*s, but does not have 70-bp repeats, the N-terminal domain is type A, and the C-terminal domain has 7 cysteines and does not cluster well with either *VSG*s or *VR* C-terminal domains. Because of its unusual features, it is possible that Tb927.5.3990 has also diverged from *VSG*s in a way similar to that of *ESAG6*, *ESAG7* and *SRA*: all these surface molecules have a common origin with type A N-terminal domains of *VSG*s (see section 1.2.3 of introduction for a discussion of these genes). Table 5.2 gives a general overview of the comparison between *VSG* and *VR* genes.

Table 5.2: A comparison of *VSG*s and *VSG*-related proteins.

Features	*VSG*	*VSG*- related
Number	940	31
Pseudogenes	808	7
70-bp repeat upstream	Most	None
Present in arrays with 70-bp repeats	All but 1 (Tb927.5.3990)	None
Length	474-540	340-490
N-terminal domain types	A, B, C	B type clusters 1, 2, 3
C-terminal domain types	1, 2, 3, 4, 5, 6	3 clusters, cluster 2 with cysteines
GPI signal	Present	Absent or not conserved; possibly present in clusters 1 and 2
3' utr	Two main subtypes clustering closely	Cluster together, separately from *VSG*s
5' utr	Diverged, slight overlap with *VR*	Diverged, slight overlap with *VSG*s

5.3 Expression of *VR* genes

The discovery of *VSG*-related genes with distinct genomic locations and separate clustering from array *VSG*s prompted investigation of whether these genes are functional and whether their expression patterns match those of *VSG*s. Because of their distinct features, the hypothesis is that *VR* genes encode surface molecules with a role separate from that of *VSG*s. Here, a subset of 10 *VR* genes are subjected to preliminary characterisation by PCR.

5.3.1 Comparison of VSG*-related archive across* T. brucei *strains*

Genomic DNA from strains TREU 927, Lister 427, EATRO 795 and STIB 247 of *T. brucei* was tested by PCR for presence of 10 different *VR* genes. Half of these genes was found to be ubiquitous, showing a conservation that is possibly higher than that present amongst *VSG* archives of different strains (see Figure 5.9 and Appendix for correspondence of *VR* numbers and GeneDB identifiers). A good way of verifying the preceding statement regarding *VSG* archive conservation across strains would be to repeat the same analysis with ~ 20 putative functional 927 array *VSG*s and see what proportion give a positive signal in the three other strains.

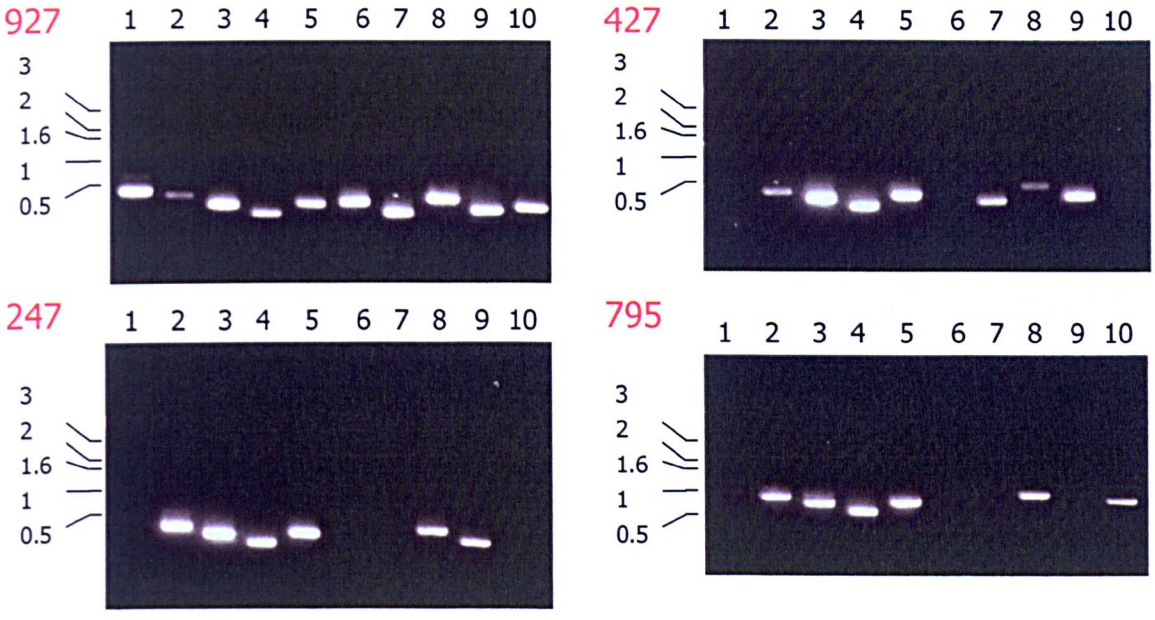


Figure 5.9. Ethidium stained gels showing *VR* presence detected by PCR on gDNA in 927, 247, 427, 795 *T. brucei* strains.

Strain numbers are indicated in red to the left of each gel. For each gel, lanes one to ten correspond to PCRs with primers for *VR* 1, 2, 4, 5, 8, 9, 11, 13, 15 and 18, respectively. To the left of each gel is the size (in kb) for the main bands of the 1 kb ladder.

The next approach taken was to question whether the genes are expressed. This was undertaken by reverse transcriptase PCR (RT-PCR). cDNA was prepared, with the primer oligo[dT], from 927 bloodstream and procyclic trypanosomes and was then analysed further with the same set of primers for the 10 genes as used above (see Figure 5.10). Results show expression of at least eight *VSG*-related genes within the trypanosome bloodstream population of strain 927, including four of the five ubiquitous genes. *VSG* related genes appear to be expressed also in procyclic trypanosomes, with an identical pattern to that seen in bloodstream forms, although in five instances (*VR 2*, *5*, *8*, *15*, *18*) there appears to be a lower level of transcript abundance, when comparing the two gels for the apparent ratio of VR RT-PCR product to tubulin. These results would need to be confirmed by Northern blot, which was not possible due to lack of time.

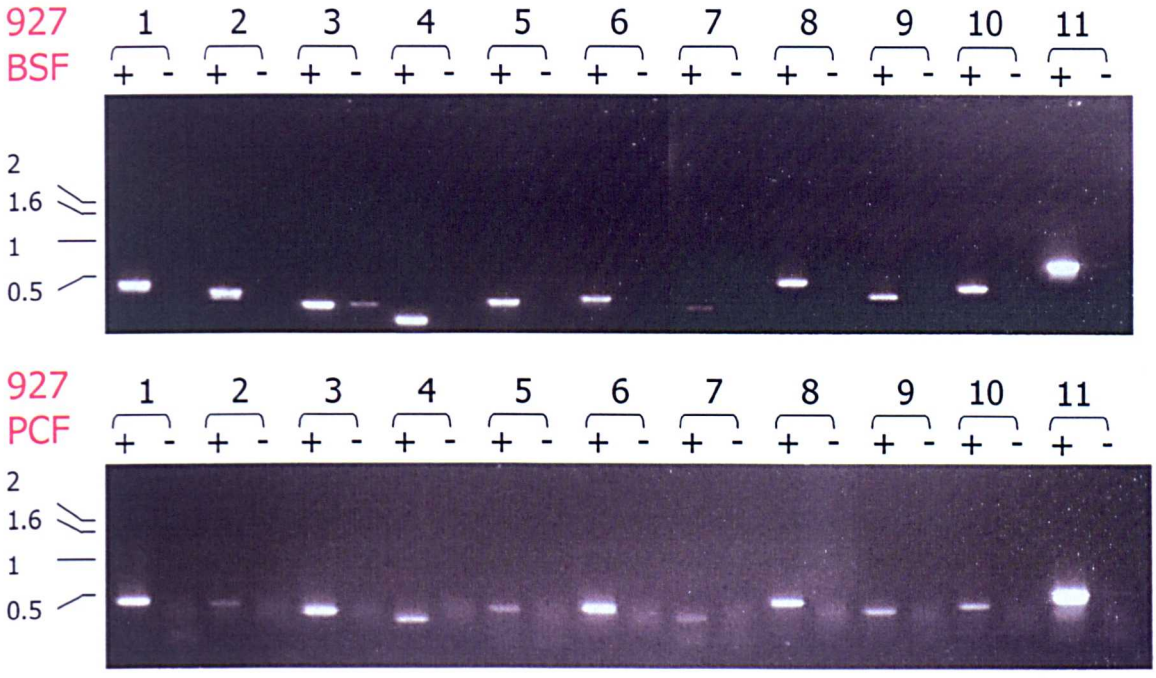


Figure 5.10: Ethidium stained gel showing *VR* expression, as detected by RT-PCR analysis of cDNA from 927 bloodstream and procyclic trypanosomes.

Above the gel, + and – indicate whether RT+ or RT- cDNA from 927 was used as template for the eleven different sets of PCR (numbers 1 to 11 above gel). Reactions one to ten correspond in both gels to primers specific to 10 *VR* genes, VR 1, 2, 4, 5, 8, 9, 11, 13, 15 and 18 respectively. Reaction 11, in both gels, corresponds to a positive control using tubulin-specific primers.

5.3.2 *Are* VR *genes used as* VSG *coats?*

Having gathered tentative evidence that *VSG*-related genes are expressed in bloodstream and possibly also procyclic trypanosomes, it was important to test whether *VSG*-related expression is concomitant with that of canonical *VSG*s or whether it occurs mutually exclusively of that of *VSG*s. To distinguish between these two possibilities, expression was analysed in the 427 strain that has been genetically modified by insertion of the *HYG* gene in tight linkage with the 221 *VSG*, such that 221 expression correlates with hygromycin resistance (McCulloch *et al.*, 1997). This strain was grown under drug selection and then RNA was prepared. RT-PCR results show expression of three *VSG*-related genes (VR2, 4 and 15), so it appears that they are coexpressed with the 221 *VSG*, rather than being expressed mutually exclusively, as would occur for authentic *VSG*s. This has been further confirmed by showing that 221 is the only *VSG* expressed, as there were no RT-PCR products for five other *VSG*s that are commonly expressed as variants within the 221 strain (see Figure 5.11).

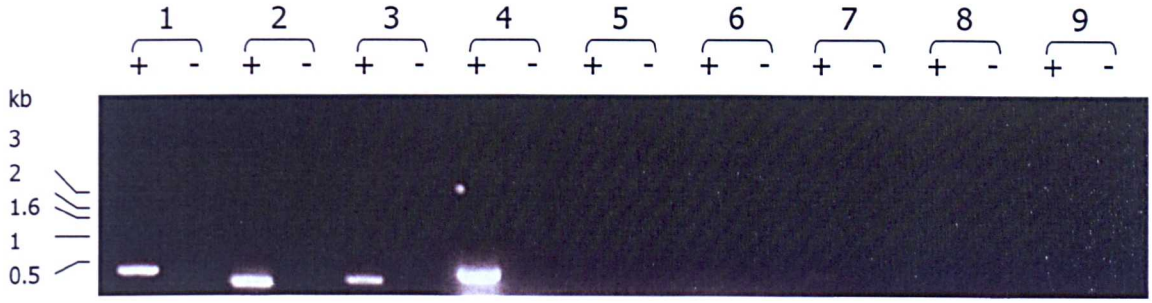


Figure 5.11: Ethidium stained gel showing RT-PCRs to test whether *VR* genes are coexpressed with *VSG*s in *T. brucei* strain 427 expressing the 221 *VSG* under hygromycin selection.

Above the gel, + and – indicate whether RT+ or RT- cDNA from 427 was used as template for the nine different sets of PCR (numbers 1 to 9 above gel). Reactions 1 to 3 used primers specific to *VR* 2, *VR* 5 and *VR* 15 respectively. Reaction 4 amplified part of the 221 VSG. Reactions 5 to 9 correspond to other *VSG*s known to be expressed in 427 (*VSG* 118, Vo2, 121, G4 and S8 respectively), as a control for homogeneity of *VSG* expression in the trypanosome population used to harvest RNA from. Primer sequences can be found in the Appendix, in Table 7.10. To the left is the size (in kb) for the main bands of the 1 kb ladder.

5.4 *T. brucei VSG*s, *VSG*-related genes and *T. congolense VSG*s: a comparative analysis

The striking differences between *VSG* and *VR* gene sequences and organisation, and the partial similarity between *VR* genes and *T. congolense VSG*s, suggest that the *T. congolense VSG* system might shed some light onto the *T. brucei* system.

5.4.1 *Analysis of* T. congolense VSGs

The sequence of twenty-three *T. congolense VSG*s is analysed here, including the ten full-length *VSG* sequences that have been published (Strickler *et al.*, 1987; Eshita *et al.*, 1992; Rausch *et al.*, 1994; Urakawa *et al.*, 1997). The 13 novel sequences were obtained by blasting a published sequence against the *T. congolense* database (http://www.genedb.org/genedb/tcongolense/blast.jsp) (see Table 7.11). These *VSG*s appear to be present in arrays of tandemly duplicated genes, and one of the contigs analysed (chr 10) suggests the array to be immediately adjacent to a telomere tract (see Figure 5.12). At least four out of the 13 *VSG*s are pseudogenes, possibly hinting to a lower degeneracy of the *VSG* archive when compared to *T. brucei,* however the sample size is too small to draw definite conclusions. *VSG*s in the arrays are not evenly spaced, do not have any repeats associated with their flanks, and seem not to have a cassette structure (see Figure 5.12)[38]. The range of divergence between genes seems to be comparable with that found among *T. brucei VSG*s (data not shown).

[38] A comparison with *VR* gene arrays is not included as the sequence context is too different. *VR*s are either present in small numbers (two-three genes), or interspersed with *ESAG* genes, or in arrays of closely related tandemly duplicated genes (one instance, *VR*s 20 to 24).

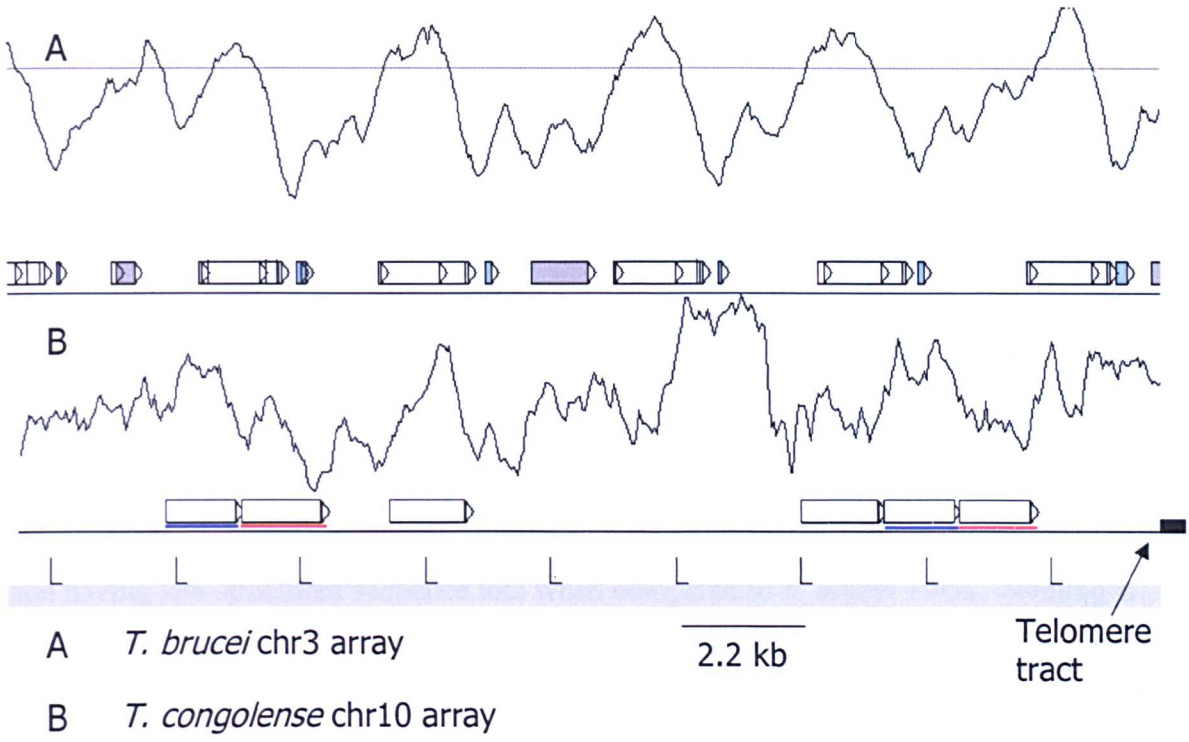


Figure 5.12: Comparison of *T. brucei* and *T. congolense* array structure.

Image derived from the Artemis annotation software. In section A, *T. brucei VSGs* are in white, narrow blue boxes indicate 70bp repeats, gray boxes indicate ESAG3 pseudogenes. In sectionB, *T. congolense VSGs* are highlighted in pairs of different colours. The *VSG* pair underlined in red share 60% amino acid identity, the pair underlined in blue shares 50% sequence identity. The telomere tract found at the end of the Chr 10 assembly is highlighted as a black box. Above each gene prediction is a graph highlighting the GC content of the DNA sequence. *T. congolense VSG* arrays do not share the periodical pattern of GC content found in *T. brucei* arrays.

Searching the *T. congolense* GeneDB database with a type A N-terminal domain of *T. brucei* (ILTat 1.24) yielded only one ORF with significant match, which when blasted in turn against the GeneDB *T. brucei* protein database yielded *ESAG6*. Blasting the *congolense ESAG6*-like gene against the *congolense* database, more than 60 contigs appear to contain good hits to the sequence, suggesting the presence of a multigene family of *ESAG6*-like molecules in *T. congolense*. The relationship between this multigene family and *T. congolense VSGs* needs to be further investigated bioinformatically. So far, a single contig (congo502c05.p1k_12) was found to display both gene types in the same array, indicating that they could be part of the *T. congolense VSG* archive, or at least share the same genomic locus.

5.5 Conclusions

Yet unanswered questions with regards to the *T. congolense* array system include whether domain combinations are present (like in the case of *VSG*s) or whether they are not (as in the case of VSG related genes). It would be also interesting to identify patterns of gene duplication to see whether any flanking sequences promoting archive expansion (and possibly copying into an expression site) can be found. For both the above questions to be answered, many more sequences would need to be analysed, and more extensive sequencing coverage and annotation of *VSG* arrays would be required. So far it looks like *T. congolense VSG*s might be more similar to *VSG*-related genes, possibly lacking domains and having less structured sequence loci when compared to *T. brucei VSG*s. Nothing is known about the mode of expression of *T. congolense* VSGs, whether expression sites similar to those of *T. brucei* are present: the striking differences in array structure between the two trypanosome species suggests that there might be many more divergent traits in the mode of antigenic variation.

A more extensive coverage of *T. congolense VSG*s in the genome strain might reveal a *T. congolense* equivalent of the *VSG*-related family presented here for *T. brucei.* With additional experimental data characterising the role of *VSG*-related genes, it would then be possible to undertake a large comparative study in which surface receptors (such as *ESAG6* and possibly *VSG*-related genes) and their putatively derived antigens could be analysed across species. This would allow common themes and differences to emerge in the antigenic variation strategies of African trypanosomes and would produce a clearer picture of the evolution of the *VSG* superfamily, together with a better understanding of its past and present biological roles.

CHAPTER 6

DISCUSSION

6 Discussion

This discussion will focus on the dynamics of antigenic variation, in terms of whole *VSG* archive evolution in the course of a chronic infection and its relationship to the sequential expression of the evolving archive. A holistic approach is called for, as post-genome studies will be most effective if thought is given to working simultaneously on bioinformatics, molecular and *in vivo* studies, requiring new ideas to be introduced and tested in trypanosome antigenic variation research. Key to this holistic approach is the concept of adaptive mutation, in which interaction with the host environment establishes a pattern of evolution on the whole antigenic variation system (and organism), allowing a more tailored (less random) response to immune selection (adaptive mutation is reviewed in Caporale (2003)). As selection in a specific environment is maintained, the pattern of evolution is reinforced, producing or selecting sequence loci and molecular mechanisms that are better tailored to, and more focused in, the generation of the "required" diversity, sometimes resulting in the evolution of a mutator phenotype. The clear aim of elucidating antigenic variation as a mutator phenotype is crucial in overcoming, or at least understanding, the limitations of experiments *in vitro*, where key pathways might not be operational; it also furthers the scope of bioinformatics studies, allowing the genome to be queried not solely for its coding, but also for its evolving, potential.

The immune system of mammals has selected for convergent evolution of antigenic variation strategies (Deitsch *et al.*, 1997). The common vocabulary of antigenic variation includes the presence of at least a multigene family of surface antigens, a mechanism for single antigen allele expression, often mediated by a dedicated expression site, the role of repeats in gene conversion of a silent copy to the expression site, the frequently telomeric, or clustered, nature of loci encoding the antigens (with often unique sequence features when compared with the housekeeping region of the genome), the use of partial gene conversion and point mutation to further enhance antigen diversity, and the presence of pseudogenes. *Trypanosoma brucei* has outstanding features, namely a huge antigen gene archive size (at least 10-fold more than that of all other known parasites), a very high level of divergence and a high number of pseudogenes amongst its antigen gene copies. It also seems to rely upon several distinct strategies: transcriptional switching, full-length gene conversion, mosaic gene conversion and possibly point mutation, as has emerged from this study.

6.1 *VSG* Sequence mutation and evolution

The aim of this first section is to explore some of the DNA features of *VSG*s and *VSG* arrays, outlining the sequence basis upon which point mutations arise and antigenic variation takes place. First of all the concept of strand bias will be introduced, the importance of which will become clear in section 6.1.1.2.

6.1.1 Strand bias, leading and lagging strands

6.1.1.1 Background on strand bias and its significance

It has been shown that, in the absence of asymmetric substitution occurring between leading and lagging strands during replication, the relative amounts of the four nucleotide bases found in DNA are roughly equal. In contrast, the presence of asymmetric substitution might lead to a higher incidence of a particular base over another: in the case of many bacterial species, the leading strand was shown to be more abundant in G over C and T over A (Frank and Lobry, 1999). Strand bias has been studied extensively for bacteria and has been found in archaea, mitochondria and bacteriophage (Rocha, 2004) and recently also in eukaryotes (Niu *et al.*, 2003; Hou *et al.*, 2006). No mutational mechanism has been described that satisfactorily explains this G+T bias: cytosine deamination, resulting in a C to T transition, has been suggested to play a role, but would have to act on the leading strand in order to generate the bias, which is in conflict with data suggesting the lagging strand as being more prone to mutation (Frank and Lobry, 1999). Furthermore, AID, the enzyme involved in cytosine deamination, has been implicated in mutations occurring during transcription in somatic hypermutation of the immunoglobulin genes, but the spectrum of mutations generated is not directly linked to the enzymatic specificities of this protein and the observed mutations do not result in a strand bias (Samaranayake *et al.*, 2006). There appear therefore to be big gaps in the current understanding of strand bias and mutational mechanisms, so what is presented below is just a brief overview of some of the main points related to this area of study, which are of relevance for discussing *VSG* arrays and *VSG* point mutations. In bacteria, the lagging strand is suggested to be maintained for longer in single-stranded conformation (ssDNA), compared with the leading strand, and cytosine deamination occurs more than one hundred-fold faster in ssDNA than in double-stranded DNA (Rocha, 2004). Indeed, it has been shown that genes on the lagging strand mutate more rapidly than those on the leading strand (Szczepanik *et al.*, 2001). This could also be due to head-on collisions between DNA and RNA polymerases, causing stalling

and the potential for errors being introduced during replication fork restart. It is in fact the case in many bacteria that replication seems to affect the orientation of operons, which tend to be located on the leading strand (Rocha, 2004).

As for eukaryotes, the okazaki fragments generated during lagging strand replication are much shorter than those in prokaryotic DNA, suggesting that ssDNA-dependent mutation would exert a weaker effect. Replicons also tend to be shorter in eukaryotes, and are generally not identifiable bioinformatically; this, together with the lack of such a rigid gene organisation as the bacterial operon, makes eukaryotic strand bias more difficult to detect (Niu *et al.*, 2003). Strand biases have nevertheless been detected, although they have been ascribed also to factors other than asymmetric substitution during replication, such as transcription-associated mutation bias and selective codon usage bias (Niu *et al.*, 2003).

6.1.1.2 Strand bias in *VSG* point mutations and in *VSG* arrays

In the case of *T. brucei*, a single report of strand bias was made with reference to point mutations in three *VSGs* (Lu *et al.*, 1994). The type of mutations in expressed *VSG*s detected in this study (see Table 6.1 below) are comparable with those published data (comparison not shown). If both strands were mutating at random in the course of replication, the number of G to A mutations on the leading and lagging strands should match; those on the lagging strand would result in the leading strand mutating from C to T, so comparing G:A (G to A) with C:T mutations on the same strand provides indirect evidence for preferential mutation of one strand versus the other. Table 6.1 shows 171 mutations ordered in pairs corresponding to the same mutation on the two strands. While it clearly appears that mutations are not generated on either strand at random, there being asymmetric substitution, what remains unclear is whether, at least for this dataset, this would correlate with a general compositional strand bias at *VSG* loci: although skewed, these mutations are not predicted to generate any skew, as, overall, the number of mutations giving rise to A is roughly the same as the number of A residues mutating to the other three bases, and this is true also for C, G and T.

Table 6.1: 171 point mutations in expressed *VSG*s isolated in current study, ordered in six pairs, corresponding to identical mutations arising on the *VSG* coding strand and the opposite strand.

VSG strand mutation	Number of mutations	VSG strand mutation (equivalent mutation on opposite strand)	Number of mutations
A:C	15	T:G (A:C)	5
C:A	17	G:T (C:A)	2
A:G	38	T:C (A:G)	11
G:A	41	C:T (G:A)	9
A:T	1	T:A (A:T)	4
C:G	16	G:C (C:G)	12

A recent publication analysing strand asymmetry in kinetoplastids (Nilsson, 2005) detected asymmetric substitution at strand switch regions of *T. brucei* chromosome 1, adding to the similarities in gene organisation between trypanosomes and bacteria (presence of polycistrons) and possibly implicating the switch regions as origins of replication. The scarcity of *VSG* sequences on this chromosome (only four *VSG*s) has meant that *VSG* arrays were not included in the analysis. To address more specifically whether a bias is present within the *VSG* archive, a comparison of DNA composition between the 154 *VSG*s of the array on the right arm of chr 9 and 99 housekeeping genes on chr 8 was made, and is presented in Table 6.2. This suggests that genes present in the polycistrons of megabase chromosomes do not show any significant strand bias (see Table 6.2), whereas *VSG*s show a marked bias in favour of A over T (35% A content), with no obvious bias in G versus C. Interestingly, the one clear example of strand bias to be detected in eukaryotes has been reported for yeast subtelomeres (Niu *et al.*, 2003): whether this is a wider feature of subtelomeric contingency loci (Barry *et al.*, 2003) remains to be investigated.

Table 6.2: Percentage of the four DNA bases at different genomic loci in *T. brucei*.

	Chr 8 99 internal genes	Chr 9 *VSG* N-terminal domains	Expressed *VSG* N-terminal domains	Chr 9 *VSG* C-terminal domains	Expressed *VSG* C-terminal domains	Chr9 GPI signal
A	23.89	34.64	32.73	39.96	38.31	24.06
C	23.60	25.75	27.41	18.77	21.69	16.12
G	28.07	24.46	25.54	22.37	23.47	16.81
T	24.44	15.15	14.32	18.89	16.54	43.00

To summarise, *VSG* mutations are not generated at random between leading and lagging strand and *VSG* arrays appear to have a DNA composition (and bias) different from that of the core region of megabase chromosomes (see Table 6.2). Interestingly, the prevalence of mutations in the VSG N-terminal domain (see Chapter 4) seems to correlate with a localised skew in favour of A and C in this domain (see Table 6.2), which corresponds in *E. coli* with the lagging strand of replication. Additionally, such localised bias fits with the concept of adaptive mutations as enunciated by Caporale (2003): "When we observe different rates of mutation at different positions in a gene, we should consider the possibility this may be due to evolution of mechanisms that modulate the rate of variation, rather than selection for and against mutations one by one". Another way mutations could be targeted to the N-terminal domain is by repeated error-prone gene conversion reactions, in which C-terminal domains would act as homology flanks and therefore might not be affected, at least not in their entirety. There is evidence against such error-prone gene conversion reactions, as a study monitoring repeated gene conversions between two genes (AnTat 1.1 and AnTat 1.10) did not find any evidence of point mutation (Pays *et al.*, 1985). This was reported in a monomorphic strain, so an error-prone gene conversion might still be part of the recombination events that are characteristic of pleomorphic trypanosomes.

6.1.1.3 Significance of *VSG* array strand bias

If bacteria and trypanosomes have similar strand bias, then it could be suggested that *VSG*s are, or have been, preferentially located on the lagging strand of replication. Nothing is known about origins of replication in *T. brucei* and their directionality, but the above data prompt at least a speculative comparison with bacteria. As mentioned above, in bacteria orthologues present on the lagging strand have been shown to diverge at a faster rate than those in the leading strand (Szczepanik *et al.*, 2001). There might therefore be significance in the orientation of arrays away from the telomere and in expression sites being oriented towards the telomere: different mutation rates might be established between expression site and donor regions, either to enhance or to reduce mutation at expression sites. It is also possible that a replication origin is placed between the array and the expression site, and in this case the mutation rate of these two loci might be the same. An additional reason for this opposite orientation could be for ease of aligning telomeric and array *VSG*s during gene conversion; this could occur by the telomere-proximal region folding back and pairing with more internal sites of the same DNA molecule or ectopically with other arrays. A similar head to head orientation of silent antigen copies and expression site is observed in *Borrelia burgdorferi* linear plasmids, suggesting a common architectural

theme (Zhang *et al.*, 1997). Another remarkable finding with regards to mutation in bacteria is that genes that have recently changed strand exhibit a faster mutation rate (Szczepanik *et al.*, 2001): this might suggest an advantage in the mixed orientation of *VSG*s in the long arrays of chr 9 and chr 11. An alternative explanation for this could be that the strand bias in *VSG* arrays produces some instability in DNA conformation, requiring it to be balanced by strand switching, when the length of the array surpasses a certain threshold. A further effect imputed to residency within the lagging strand, deletions between direct repeats are observed frequently and preferentially on the lagging strand in *E. coli* (Sinden *et al.*, 1999), so again, to stretch parallels to their limits, the preferential location of *VSG* arrays on one strand might facilitate deletions (and rearrangements in general) involving 70-bp repeats. It is difficult to conceive how this could be tested in *T. brucei* in any other way apart from sequencing at least some *VSG* arrays for the genome strain after prolonged growth *in vivo*, as a large number of events would need to be detected for statistical evaluation of the different changes. Prior to this, an experimental study to provide an estimate of the frequency of such events could be conducted, involving genomic DNA digestion and probing of selected array fragments, with the aim to observe changes in their size over time (across different clones). This would also allow to select specific clones that have diverged sufficiently and to then use them for array sequencing.

In parallel to an analysis of large sequence rearrangements, a more thorough analysis of strand (and codon) bias and a wider knowledge of mutations at *VSG* loci could reveal the relative contributions to mutation rates in the *VSG* archive provided by DNA sequence, sequence location, chromatin conformation and possibly transcription[39].

6.1.2 *Understanding* VSG *mutations*

6.1.2.1 Mutational mechanisms for *VSG*s

It is currently unknown whether there is a dedicated mechanism for mutation at subtelomeres, and how it might operate. Hypermutation in the variable region of immunoglobulin genes begins with regulated, targeted, enzymatic cytosine deamination events on either strand at the sequence RGYW, followed by action of a mutator polymerase (Caporale, 2003): the basic use of cytosine deamination to promote mutation in both bacteria and humans, together with the preliminary evidence provided above of a

[39] Somatic hypermutation has been shown to be strongly dependant on transcription, as mutation frequencies correlate with transcriptional activity (Aguilera, 2002).

mutation bias in trypanosomes, might suggest an involvement of this process in *VSG* mutation. An alternative mechanism could involve base J, which replaces a fraction of thymine at subtelomeres (van Leeuwen *et al.*, 1997), and for which a function has not yet been fully ascribed: its association with subtelomeres might increase mutation rate, possibly by causing the DNA polymerase to stall. Base J, or other factors, could also result in promotion of double strand breaks (DSBs), and lead to a phenomenon termed recombination-dependent adaptive point mutation; in this pathway, the DSB leads to mutations being produced in its proximity, by a process involving (in *E. coli*) the error prone polymerase polIV and downregulation of mismatch repair (Rosenberg, 2001). So far this process has been described only in *E. coli*, although some analogies with somatic hypermutation in humans have been drawn[40] (Rosenberg, 2001). The potential for it to operate in trypanosomes has been suggested, based on the presence in the genome of all the required factors, such as error-prone polymerases (McKenzie and Rosenberg, 2001). While looking for a specificmutation mechanism, it should not be forgotten that a mutation rate is achieved through "integration of a wide range of cellular activities, including the level of and balance between distinct repair, polymerase and proofreading activities and their interaction with different sites in the genome" (Caporale, 2003), so further explanations could be found by increased knowledge in processes such as DNA replication, mismatch repair and genome architecture dynamics.

6.1.2.2 Where do gene conversion and point mutation occur?

Where mosaic formation and point mutation occur remains to be established: in the case of mosaic genes it has been suggested that silent expression sites might be the loci at which they are assembled (see section 1.3.1.5), then to be switched on or converted into the active expression site. Telomeric 3' donor *VSG*s have been found in two instances to be pseudogenes (Thon *et al.*, 1989; Roth *et al.*, 1989), lending strength to the idea that not all telomeres harbour intact *VSG*s. Although there are reasons to think that point mutations detected in the experiment originate from the *VSG* Expression-Linked Copy (ELC) (see Chapter 4, section 4.9.1), an analysis of mutation across the archive as a whole during infection would be highly informative, especially to observe whether the donor copies also mutate and whether they would be transferring (or eliminating) mutations to the ELC by gene conversion.

[40] An error prone polymerase, polμ, has been found to be involved in somatic hypermutation, see also section 6.3.2.

6.1.3 Is VSG gene evolution fast enough to contribute to antigenic variation?

Both point mutation and mosaic gene data produced in this work lead to a major and yet unanswered question, namely whether there is evolution of expressed *VSG* sequences contributing to antigenic variation alongside *VSG* switching (be it by full-length or segmental gene conversion).

Modelling based on current data suggests that within 5 to 6 days of appearance of a VAT beyond a certain threshold, polyclonal antibodies would be produced by the host and the VAT would be eliminated (Morrison *et al.*, 1982). Within this time window, point mutation or further mosaic formation would need to affect the initial expressed *VSG* to an extent that the outcome of these diversification mechanisms renders it non-cross reactive with the starting VAT. The time frame is further reduced by the fact that once switching to a non-cross reactive variant has occurred, current knowledge of mechanisms suggests it may take up to two days for the coat to be completely replaced (exact timing unknown, see Introduction, section 1.3.2), extending the time in which the switching trypanosome is vulnerable to existing antibodies. Point mutation is also restricted by the limitations imposed by VSG structure and function and possibly also by the "homogenising" effect of repeated gene conversions from the original silent donor. What therefore remains unanswered is not so much whether point mutations occur, but whether they have biological significance in terms of helping prolong a chronic infection. Their role could be in further diversification of the *VSG* archive on a longer time scale, rather than in short term immune evasion. Alternatively, point mutated or mosaic versions could be retained in a silent expression site after an unrelated *VSG* has been switched on, allowing them more time for divergence to occur, possibly to be reactivated later on in the infection, once they have achieved non-cross reactivity.

If indeed point mutation and "real-time" evolution of mosaic genes are promoting antigenic variation, this could explain the observation made when comparing the two infections at day 28, that the point mutations per clone in infection 28-10 were lower than in 28-11. Whereas all *VSG*s isolated in 28-10 are suggested to be mosaics, only one putative mosaic is present in 28-11, perhaps hinting that in this infection the diversification strategy (mutator phenotype) used was biased towards point mutation. This in turn could be due to the presence or absence of closely related orthologues for a *VSG* that has been duplicated into the expression site: lack of orthologues would invoke point mutations,

whereas presence of orthologues could result in the preferential formation of more complex mosaic genes.

6.1.3.1 Point mutation in other systems

Point mutations have been reported alongside gene conversion in both *Borrelia hermsii* (Restrepo and Barbour, 1994) and *Borrelia burgdorferi* (Sung *et al.*, 2001); in both cases they are thought to occur after gene conversion , at the expression site. Interestingly, in the case of *Borrelia hermsii*, increase in point mutation incidence was observed as the infection progressed. On closer inspection, this latter example, unlike that of *Borrelia burgdoferi*, does not constitute an accurate analogy, as mutations in *Borrelia hermsii* were templated, actually resulting from short-range gene conversion events from very closely related silent pseudogene copies of the antigen.

To formally demonstrate that point mutated antigen variants contribute to the generation of escape mutants is not trivial and is currently being addressed in the field of virology, for example in the case of hepatitis C virus (HCV) (Guglietta *et al.*, 2005); in this study, previously characterised epitopes were followed in their evolution over a period of a year in a single chronic infection, and by *ex vivo* analysis of synthetic peptides representing the altered epitopes, were shown to be recognised with diminished efficiency by the wild-type specific cytotoxic T cell response.

6.1.3.2 Do trypanosomes rely also on "escape mutants"?

A way to find out whether point mutated genes result in immune evasion would be to force their expression (and also that of its non-mutated predecessor) into the BES of a monomorphic trypanosome strain; this would allow us to test these variants *in vitro* for reactivity with infection serum. Putative non-cross reactive variants could then be tested *in vivo* by initial infection with one variant, followed by curing the host and superinfecting with the second variant. By pursuing this line of research it would be possible to appreciate the upper limit of non-crossreactivity between variants and the interplay between mosaic gene formation and point mutation.

6.2 Modelling infection: infection evolution

6.2.1 Host parasite interactions

Another interesting aspect of antigenic variation is the relationship between infection and mutation, which could be better characterised by comparing mutation rates for the *VSG* archive *in vitro* and *in vivo* (the latter both in the absence and presence of immune responses). If, as it does, the *VSG* switching rate rapidly decreases upon prolonged passaging in the absence of immune responses (Turner, 1997), it might be that the whole process of *VSG* archive diversification is affected. Different cues for the mutator phenotype to be switched on might be present, and not all simply connected to selection by the immune system. A good example of (indirect) second order selection is that of segmental gene conversion of the *vls* antigen in *Borrelia burgdorferi.* Compared with the high levels of gene conversion detected in chronic infections, parasites grown *in vitro* do not exhibit any gene conversion and tend to lose plasmids and infectivity; on the other hand, *in vivo* growth in the absence of immune selection results in an intermediate level of gene conversion (Zhang and Norris, 1998), suggesting that there might be several host-parasite interactions influencing the mutator phenotype. Drawing analogies with bacteria, this phenotype might also be transient during infection and expressed only in a subpopulation, allowing short term rather than long term survival (McKenzie and Rosenberg, 2001), in a balance between survival and general organismal viability. Drawing analogies with the process of differentiation from long slender to short stumpy bloodstream forms, a stress-mediated response based on accumulation of unknown factors in the bloodstream as the infection progresses might also be operating.

6.2.2 Stage 2 antigenic variation?

As pointed out in Chapter 4 (section 4.2.1), the timing of the onset of array *VSG*s and mosaic genes and the putative upregulation or build up in the number of point mutations seem to coincide, at least in mice, with a dramatic change in infection profile, in which distinct peaks are absent. The change in infection profile suggests that point mutation of *VSG*s might have an immunomodulatory role, reminiscent of the difference between antigenic shift (large changes, exchange of RNA segments between viral strains) and drift (point mutations) in the neuroaminidase gene of the influenza virus. While shift causes severe pathogenesis, drift causes only a mild disease (Vossen *et al.*, 2002). Immunomodulation by point mutation could be due to the host needing to select

continually B cells through the process of affinity maturation. Additionally, the divergence in expression of different surface coats late in infection might contribute to spreading the resources of the host to engage in a more subtle "hide and seek" game. High affinity antibodies can be derived from low affinity ones by selection of six to eight point mutations (Beale and Iber, 2006), a number comparable with that of mutations found in *VSG*s, which therefore might be sufficient to lower the affinity of antibodies, rather than be completely non cross reactive. The fact that a polyclonal antibody response is mounted against each VAT, possibly limiting the efficacy of immunomodulation by point mutation, might be countered by the observed rapid and divergent mutation pattern within the VAT population, likely to result in a broad spectrum of altered epitopes. As an important *caveat* to the above statements, the effectiveness of antibodies during infection has been shown to be long-lasting (Morrison *et al.*, 1982), and is a requirement for a hierarchy of expression to be maintained: in the absence of specific antibodies, the order is immediately reset to "early" variants. Therefore, the proposed modulation would be operating "at equilibrium" with the immune response, rather than subverting it completely. On the other hand, it has been suggested that the process of antibody affinity maturation is impeded late in infection by a inhibition of T helper cells, and this would correlate to the prevalence of the IgM antibody subclass in response to infection (Turner, 1989). More immunological data on the course of prolonged trypanosome infections are needed to provide a framework for more robust models of chronic stage antigenic variation to be built.

6.2.3 Hierarchy late in infection

At the level of hierarchy of *VSG* expression, stage 2 antigenic variation could represent a "genome-encoded" phenotypic shift from order to chaos, based on progressive lowering of gene activation probabilities to levels at which new variants seem to appear at random (i.e. non reproducibly between separate infections). Reanalysis of rabbit infections originally conducted by Capbern *et al.*, (1977) suggests that variants appear as a continuum rather than in distinct groups of early, middle and late (J. D. Barry, pers. comm.). This correlates well with the lack of distinct *VSG* classes within the subtelomeric archive reported in this work, likely to result in only small differences in activation probabilities between genes. These differences appear nonetheless significant, as the study mentioned above reported a very consistent pattern of expression across eleven rabbit infections; they are expected to derive from the additive effect of small variations in locus (proximity to expression site), sequence environment and sequence homology with telomeric *VSG*s. In this continuum, the "dramatic" shift from telomeric to array *VSG*s is liable to remain the only exception to

a less notable progression across the variant archive, complicating the prospects of producing an accurate model of *VSG* hierarchical expression. As the infection progresses, the variants arising appear to be increasingly unique, found only in single infections, and are likely to be formed entirely by mosaic genes, through low probability "random" events. Mosaic genes are likely to be a feature in both "ordered" and "chaotic" stages of antigenic variation: one testable hypothesis would be that some mosaics would initially appear alongside functional array genes, due to further recombination of the latter with related pseudogenes, and then later in infection would manifest as the outcome of rarer recombination events between more diverged and / or degenerate donors. This could be tested by monitoring the onset of 10 functional genes across a statistically significant number of infections. Doing this could define whether timing of appearance is indeed present and reproducible, rather than stochastic. A "generic" *VSG* PCR could then be performed on cDNA from the days in which these *VSG*s are present, to see whether related *VSG*s are consistently being coexpressed in the same time window (or immediately after). Sampling of variants appearing after 2-3 months of infection could also be carried out, to probe into the currently unknown depths of antigenic variation.

6.3 *VSG* array evolution

6.3.1 VSG *array structure*

The present analysis, as part of the genome sequencing and annotation efforts, has unveiled much complexity in the organisation and features of the *VSG* archive. It is now apparent that *VSG*s are organised in subtelomeric arrays, in a unique sequence environment in which *VSG* cassettes (70-bp to VSG coding sequence) are enriched with *ESAG3* and *UDP Gal* pseudogenes and *VSG* genes are interrupted and rearranged around *INGI* and *RHS* sequences (see Chapter 3, section 3.7). A hint uncovering the plasticity of *VSG* arrays was provided by early studies comparing the sequence environment of the donor copy of *VSG* 118 in strain 427 and related strains (Bernards *et al.*, 1986), showing lack of gene order conservation and breaks of syntheny, leading to a model involving deletion and duplication of *VSG* cassettes. This outline has been confirmed: tracing recent duplications in the genome and looking for putative conversion tracts shows that most gene duplications in *VSG* arrays involve 70-bp repeats and the end of the C-terminal domain (data not shown). A more up to date model of array evolution can now be suggested: shorter rearrangements (including deletions) and conversions utilising the *VSG* cassette promote at the same time gene duplication and divergence; these are accompanied by larger scale contractions and

expansions due to transposition events involving *INGI*. A role for transposons in reshaping genomes has been shown in other systems (Kazazian, 2004) and, in the case of trypanosomes, it could explain large differences in archive size between strains and between arrays and homologues; on the other hand, cassette duplication might explain the original expansion of the arrays and the shuffling of sequences covering up larger rearrangements caused by *INGI*. That the latter are rarer than 70-bp mediated gene conversions can be inferred from the low incidence of detection of such event in the genome (data not shown).

70-bp repeats have been suggested to be related to transposons, as the conserved TGTTG sequence in the GT-rich tract has been found in eukaryotic transposable elements (Aline *et al.*, 1985). An alternative scenario would be that they originated from within the *VSG* coding sequence: TAATGATGATAATAACAATAATAATGATAATGATAATATTAA[41] was found within the coding sequence of *VR* 4 (Tb03.1J15.190, a *VSG*-related gene), and shorter runs can be found in several *VSG* C-terminal domains (data not shown). Whichever might be the explanation, the highly recombinogenic properties of 70-bp repeats are evident in the absolute predominance of gene conversion events involving these repeats, both within arrays (current study, data not shown) and between arrays and expression sites (Peter Burton, PhD thesis, 2003, University of Glasgow).

Interestingly, gene duplication has recently been proposed to be a form of adaptive mutation: "a link between gene duplication and mutation is explicit in the finding that decreased activity of the mismatch repair proteins MLH1 or MSH6 increases the rate of gene amplification in eukaryotic cells" (Caporale, 2003).

6.3.1.1 Antigen archive architecture in other systems

The types of organisation and expansion seen for the *VSG* archive are not without precedent. In a recent study of the silent archive of the antigen (vlp / vsp) genes in *Borrelia hermsii*, the spirochete responsible for relapsing fever, an organisation remarkably similar to that of *VSG* genes has been found, including a cassette structure and repeats providing one of the flanks, although downstream, rather than upstream, (Dai *et al.*, 2006). Most gene conversions in this system appear to use flanks rather than create mosaics, at least in the first relapse studied (only one mosaic gene detected), and this might suggest a

[41] Triplets highlighted with alternating blue and black colour.

similar hierarchy in switching mechanisms, allowing more complex mosaics to appear later in infection.

Another example of subtelomeric antigen archive with analogies to the *VSG* system is that of *Pneumocystis carinii*. The ~75 copies of the *MSG* antigen genes are arranged in tandem arrays and, like *VSGs*, close relatives are jumbled in the arrays (Keely *et al.*, 2005). Sequence shuffling seems to be due to intergenic spacer sequences, which appear to be more conserved than the coding sequence itself, as is the case for *VSG* cassette flanks.

6.3.2 VSG *divergence and* VSG *pseudogenes*

When considering that at least one half to two thirds of the 927 *VSG* archive has been analysed, it appears that the level of divergence between *VSG*s is generally very great. A conservative definition of VSG family is proposed here, encompassing genes that are at least 60% identical at the protein level: these would be defined as orthologues (See Chapter 3, section 3.6). It might be that the high level of divergence (and therefore mutation) amongst *VSG*s is required for long-term survival in the field, as this might maximise the chances of being able to spread in a host population that is already at least partially immune against the products of a given archive. This need for high levels of mutation could have in turn pushed for the expansion of the archive, possibly to compensate for the number of deleterious mutations accumulating; alternatively, mutation and repertoire expansion could have been (and possibly are) simultaneous, concerted processes, reaching a steady state based on optimal transmission. The high number of pseudogenes (~90%)[42] might be a side effect of the high mutation rate (in the case of the N-terminal domain) together with the inaccurate resolution of gene conversion reactions (in the case of the C-terminal domain). The unusual DNA content (50% of T residues) and likely instability of the *VSG* GPI signal might lead to a heightened level of recombination involving this sequence; this correlates well with what was observed with regards to cassette duplication within arrays (see section 6.3.1), where the C-terminal domain is preferentially used as one of the boundaries of conversion (data not shown). This heightened recombination of a putatively unstable sequence would lead to errors accumulating, and this is indeed what is observed in the archive (see Chapter 3, section 3.4), where most C-terminal ends are defective, due to stop codons or frameshifts. The generation of errors through recombination could be explained by analogy to the role of polμ in somatic hypermutation:

[42] Only limited numbers of pseudogenes are present in *Borrelia burgdorferi* (3 out of 15 silent genes) (Zhang *et al.*, 1997) and in *Pneumocystis carinii* ("few" out ~75) (Keely *et al.*, 2005).

polμ is involved in processing microhomology junctions (similar junctions might be used in *VSG* cassette recombination) and has been shown to generate more frameshifts than the conventional polβ (Tippin *et al.*, 2004). It might be that a similar error prone polymerase dedicated to processing microhomology junctions is involved in *VSG* gene conversion. Alternatively, the DNA conformation in which VSGs are found might prevent access of repair factors.

As mentioned above, the extent of divergence needed to produce non-cross reactive coats is unknown. The detection of point mutations in this study and speculations made so far on their role should not detract from the possibility that a significant amount of divergence (*e.g.* 10%) is required for non cross-reactivity. On the other hand, the level of divergence might be purely fortuitous, due to the plasticity of the alpha helical conserved regions, which impose only limited constraint on mutation. This in turn might be the reason for the lack, within the N-terminal domain-encoding region, of conserved sequences that might anchor segmental gene conversion, which therefore seems to have to rely on more extensive homology between donors (see Chapter 4, section 4.7). Again a comparison with *Borrelia burgdorferi* is informative: unlike VSGs, the vls surface protein is suggested to have retained function in infection, and displays six variable and six conserved regions, resulting in the possibility of utilising the conserved regions as anchors for segmental duplication, focusing divergence in restricted regions and minimising overall divergence (Zhang *et al.*, 1997). A comparison of the tertiary structure constraints of other antigens, together with knowledge of level of divergence within each antigen archive, might provide an answer as to whether the extensive *VSG* divergence is due to the lack of strong restriction in sequence variation of its helical backbone.

6.4 Perspectives: archive evolution

The high level of divergence might suggest that the whole archive is constantly mutating, and that mutation occurs at a higher rate than gene conversion and repertoire expansion. Using the reference genome strain, both gene duplication and mutation can now be followed over time, allowing analysis of the speed of these events. A new, dynamic era in trypanosome antigenic variation research has begun.

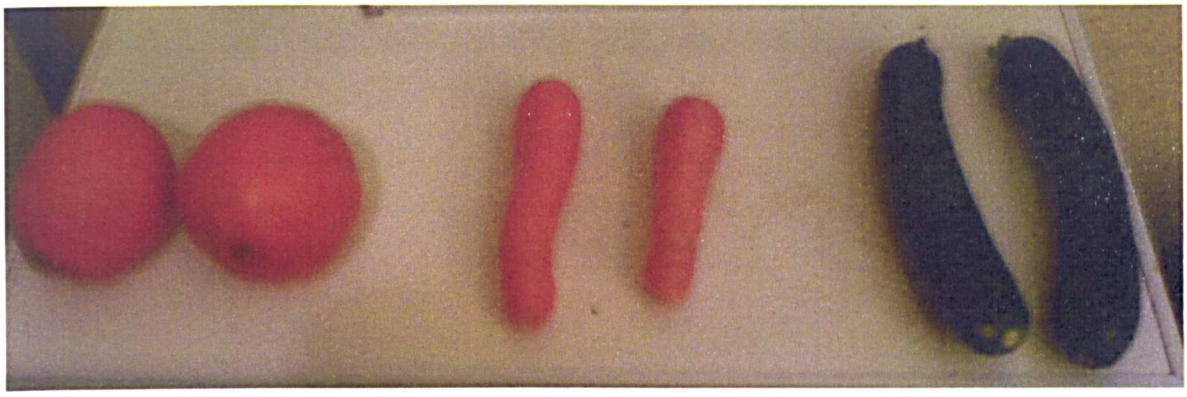


Figure 6.1: A model of archive evolution.

CHAPTER 7

APPENDIX

7 Appendix

7.1 Supplementary material to Introduction

7.1.1 Published VSG *table*

Table 7.1: Summary of all published full-length *VSG* sequences.

Species are T.b.b. (*T. brucei brucei*), T.b.r. (*T. brucei rhodesiense*), T.b.g. (*T. brucei gambiense*), T.ev. (*T. evansi*), T.eq. (*T. equiperdum*), T.c. (*T. congolense*), T.v. (*T.vivax*). Domain combinations with an asterisk beside indicate that the type 2 C-terminal domain is devoid of cysteines; *VSG* sequences without a reference were submitted to the database and are not connected to any publication (unpublished); the four *VSG* sequences with "pers. comm." have not been submitted yet.

	species	Strain	*VSG* name	Accession n	Domain combination	reference
1	T.b.b.	AnTat	1.1ELC	P06015	A1	(Matthyssens *et al.*, 1981)
2	T.b.b.	AnTat	1.1(B)	K00398	A1	(Pays *et al.*, 1983b)
3	T.b.b.	AnTat	1.10	K00397	A1	(Pays *et al.*, 1983b)
4	T.b.b.	ILTat	1.21/VAT-E	X56766	B2	(Carrington *et al.*, 1991)
5	T.b.b.	ILTat	1.22/VAT-J	X56765	A2	(Carrington *et al.*, 1991)
6	T.b.b.	ILTat	1.23/VAT-G	X56768	B3	(Carrington *et al.*, 1991)
7	T.b.b.	ILTat	1.24	X56767	A1	(Carrington *et al.*, 1991)
8	T.b.b.	ILTat	1.25/VAT-A	X56769	B1	(Carrington *et al.*, 1991)
9	T.b.b.	ILTat	1.1BC	AF317934	C2	Urakawa, T., 2000
10	T.b.b.	ILTat	1.1B	AF317935	A1	Urakawa, T., 2000
11	T.b.b.	ILTat	1.61(MVSG)	CAA09956	B1	(Matthews *et al.*, 1990)
12	T.b.b.	ILTat	1.67/VAT-B		A1	J.D. Barry, pers. comm.
13	T.b.b.	ILTat	1.68/VAT-C		A2*	J.D. Barry, pers. comm.
14	T.b.b.	ILTat	1.69/VAT-D		A2	J.D. Barry, pers. comm.
15	T.b.b.	ILTat	1.73		A2	J.D. Barry, pers. comm.
16	T.b.b.	MITat	1.1	X56761	A2	(Carrington *et al.*, 1991)
17	T.b.b.	MITat	1.2 (221)	X56762	A2	(Carrington *et al.*, 1991)
18	T.b.b.	MITat	1.4A (117)	P02896	A1	(Boothroyd *et al.*, 1982) (Allen and Gurnett, 1983) (Bohme and Cross, 2002).
19	T.b.b.	MITat	1.5 (118)	X56763	A3	(Carrington *et al.*, 1991)
20	T.b.b.	MITat	1.6	X56764	A1	(Carrington *et al.*, 1991)
21	T.b.b.	MITat	1.11	AY935571	B1	Cross, G.A., 2005
22	T.b.b.	MITat	1.12	AY935577	B1	Cross, G.A., 2005
23	T.b.b.	MITat	1.13	AY935576	C1	Cross, G.A., 2005
24	T.b.b.	MITat	1.21	AY935572	A2	Cross, G.A., 2005
25	T.b.b.	MITat	1.3	AY935575	B1	Cross, G.A., 2005
26	T.b.b.	MITat	1.8	AY935574	A1	Cross, G.A., 2005
27	T.b.b.	MITat	1.9	AY935573	B3	Cross, G.A., 2005
28	T.b.b.	427	G4	AAG03079	A2	(Alsford *et al.*, 2001)
29	T.b.b.		222	AJ007019	B2	(Ansorge *et al.*, 1999)
30	T.b.b.		Buteba 1	AJ549081	B2	(Hutchinson *et al.*, 2003)
31	T.b.b.		Bugosa 1	AJ560648	A3	(Hutchinson *et al.*, 2003)
32	T.b.b.			AAN78184	C2*	Beg, O.U., 2002
33	T.b.b.	GUTAT	10.1	AF335471	A2*	(LaCount *et al.*, 2001)
34	T.b.b.	GUTAT	10.3	AF335472	C2*	(LaCount *et al.*, 2001)
35	T.b.b.		MVAT5-RX2	A45175	A1	(Lu *et al.*, 1993; Lu *et al.*, 1994)
36	T.b.b.	ILTat	1.3	J01221	A1	(Rice-Ficht *et al.*, 1981)
37	T.b.b.		MSA1.13	1909283C	A1	(Kamper and Barbet, 1992)
38	T.b.g.			M62629	A1	(Dai Do *et al.*, 1991)

39	T.b.g.	LiTat 1.3		AJ304413	B1	(Berberof *et al.*, 2001)
40	T.b.r.			AF097332	B2	(Milner and Hajduk, 1999)
41	T.b.r.			AAA21275	A1	(Alarcon *et al.*, 1994)
42	T.b.r.	LouTat	1	X56643	A1	(Reinitz *et al.*, 1992)
43	T.b.r.	WRATat	A	M33823	B2	(Reddy *et al.*, 1990)
44	T.b.r.	WRATat	B	M33824	B2	(Reddy *et al.*, 1990)
45	T.b.r.		MVAT5-like	M33825	A1	(Reddy *et al.*, 1990)
46	T.b.r.		MVAT4	O76421	A1	(Pedram and Donelson, 1999)
47	T.b.r.		MVAT7	U83435	C4	(Kim and Donelson, 1997)
48	T.b.r.	LouTat	1.5	M89932	B2	(Schopf and Mansfield, 1998)
49	T.b.r.	ETat	1.2RC	AJ010198	A2	Pays, E., 1998
50	T.b.r.	ETat	1.2SC	AJ010199	A2	Pays, E., 1998
51	T.b.r.	AnTat1	MVAT5	AF259553	A1	(Bringaud *et al.*, 2001)
52	T.b.r.	WaTat	1	M83694	A1	(Barbet *et al.*, 1982)
53	T.b.r.	WATat	1.2	M86646	A2	Clarke, M.W, 1992
54	T.b.r.	WATat	1.12	M83695	A1	(Barbet *et al.*, 1982)
55	T.b.r.	WATat	1.13	M83696	A1	(Barbet *et al.*, 1982)
56	T.b.r.	WATat	1.14	M83697	A1	(Barbet *et al.*, 1982)
57	T.b.r.			A45175	A1	(Lu *et al.*, 1993)
58	T.b.r.			AAA30249	A1	(Kamper and Barbet, 1992)
59	T.eq.	BoTat	VSG78-20	M29498	B2	(Roth *et al.*, 1989)
60	T.eq.	BoTat	VSG20bis	X55534	A2	(Thon *et al.*, 1990)
61	T.eq.	BoTat	VSG78	M29497	B2	(Roth *et al.*, 1989a)
62	T.eq.	BoTat	1	X60228	A2	(Baltz *et al.*, 1991)
63	T.eq.	BoTat	VSG20	X16723	C2	(Thon *et al.*, 1989)
64	T.ev.	AnTat	3.3	AF317915	A1	Urakawa, T., 2000
65	T.ev.			AF317914	A2	Urakawa, T., 2000
66	T.ev.			AF317916	A2	Urakawa, T., 2000
67	T.ev.			AF317917	A2	Urakawa, T., 2000
68	T.ev.			AF317918	B1	Urakawa, T., 2000
69	T.ev.			AF317919	B1	Urakawa, T., 2000
70	T.ev.			AF317920	A1	Urakawa, T., 2000
71	T.ev.			AF317921	B2	Urakawa, T., 2000
72	T.ev.			AF317922	B1	Urakawa, T., 2000
73	T.ev.			AF317923	A2	Urakawa, T., 2000
74	T.ev.			AF317924	A1	Urakawa, T., 2000
75	T.ev.			AF317925	A2	Urakawa, T., 2000
76	T.ev.			AF317926	A1	Urakawa, T., 2000
77	T.ev.			AF317927	B2	Urakawa, T., 2000
78	T.ev.			AF317928	A2	Urakawa, T., 2000
79	T.ev.			AF317929	A2	Urakawa, T., 2000
80	T.ev.			AF317930	A2	Urakawa, T., 2000
81	T.ev.			AF317931	A2	Urakawa, T., 2000
82	T.ev.			AF317932	A2	Urakawa, T., 2000
83	T.ev.			AF317933	B1	Urakawa, T., 2000
84	T.ev.	ShTat	1.3	AF418693	A1	Zhou, J., 2001
85	T.ev.	ShTat	1.1	AAS66653	A2	Xiang, F., 2004
86	T.ev.	ShTat	1.2	AAS66654	B1	Xiang, F., 2004
87	T.ev.	ShTat	1.5	AY216527	A2	Zhou, J., 2003
88	T.c.	ILNAR2	IL3000 bVSG	U07140		Urakawa, T., 1994
89	T.c.	ILNAR2	ILNat2.1	U07141		Urakawa, T., 1994
90	T.c.	ILNAR2	ILNat3.2	U07142		Urakawa, T., 1994
91	T.c.	ILNAR2	mVSG1	M74802		(Eshita *et al.*, 1992)
92	T.c.	ILNAR2	MVSG2	M74803		(Eshita *et al.*, 1992)
93	T.c.	BeNAR1	BeNat1	X79399		(Rausch *et al.*, 1994)
94	T.c.	BeNAR1	BeNat1.2	X79400		(Rausch *et al.*, 1994)
95	T.c.	BeNAR1	BeNat1.3	X79401		(Rausch *et al.*, 1994)
96	T.c.	YNAR1	YNat1.1	M15112		(Strickler *et al.*, 1987)
97	T.c.	YNAR1	YNat1.3	M15113		(Strickler *et al.*, 1987)
98	T.v.		ILDat2.1	Z48228		(Gardiner *et al.*, 1996)

7.2 Supplementary material to Chapter 3

7.2.1 VSG *table*

Table 7.2: Sequence, cassette and locus features of all 940 *T. brucei* 927 silent array *VSG*s analysed.

Locus features include three columns, chr (chromosome number, 1-11, plus 8h for the homologue array of chr 8), array (l = left subtelomere, r = right subtelomere, n = not at subtelomere), orient (orientation, b = backwards, t = towards, with respect to general chromosome orientation). Sequence features include seven columns, FL (full-length, y or n), IN (intact, no stop codons, y or n), F (putative functional, y or n), N type (N-terminal domain type, a, b or c; d = degenerate / pseudogene; u = uncertain / atypical; d = too degenerate to assign to type; ? = no similarity to any domain types; n = not present), C type (C-terminal domain type, 1 to 6, with type 5 with 8 cysteines present at 5b; for additional notation see N type), N (N-terminal domain presence, y or n; f = fragment), C (C-terminal domain, y or n; f = fragment). Cassette features include four columns, 70 (number of 70-bp repeats upstream of VSGs). Alternate shading of rows matches separate chromosomes / contigs to which the VSGs belong. No information is available for array and orientation features of short contigs in chr 9, 10 and 11 (names not provided, indicated only as 9, 10 and 11).

GENEdb identifier	Locus features			Sequence features							Cassette features			
	Chr	Array	Orient	FL	IN	F	N type	C type	N	C	70	ingi	E3	UDP GAL
Tb927.1.05	1	n	b	y	y	y	b	3	y	y	31	n	n	n
Tb927.1.520	1	n	t	n	n	n	n	1d	f	y	0	y	n	n
Tb927.1.5300	1	r	b	y	y	y	c	2	y	y	1	n	n	n
Tb927.1.5320	1	r	b	y	n	n	ad	3d	y	y	1	n	n	n
Tb927.1.5330	1	r	b	y	y	n	b	3u	y	y	0	n	n	n
Tb927.2.6350	2	r	t	y	n	n	ad	5d	y	y	0	n	n	n
Tb927.2.6372	2	r	b	n	n	n	bd	3d	f	y	0	n	n	n
Tb927.2.6375	2	r	b	n	n	n	bd	n	f	n	1	n	n	n
Tb927.2.6380	2	r	b	y	n	n	b	1d	y	y	1	n	n	n
Tb927.2.6390	2	r	b	y	n	n	bd	3d	y	y	2	n	y	y
Tb927.2.6410	2	r	b	y	y	n	au	2u	y	y	1	n	n	n
Tb927.2.6430	2	r	b	y	n	n	b	1d	y	y	1	n	n	n
Tb927.3.120	3	l	t	y	n	n	b	3d	y	y	?	n	y	n
Tb927.3.130	3	l	t	n	n	n	d	n	f	n	1	n	n	n
Tb927.3.140	3	l	t	y	n	n	ad	2d	y	y	0	n	n	n
Tb927.3.150	3	l	t	y	y	y	b	3	y	y	2	n	n	n
Tb927.3.170	3	l	t	y	n	n	b	3d	y	y	1	n	y	n
Tb927.3.180	3	l	t	y	y	n	a	1u	y	y	1	n	n	n
Tb927.3.190	3	l	t	y	y	y	a	2	y	y	1	n	n	n
Tb927.3.200	3	l	t	n	n	n	n	d	n	f	0	y	n	n
Tb927.3.210	3	l	t	y	y	n	b	3u	y	y	1	y	n	n
Tb927.3.220	3	l	t	n	n	n	d	1d	f	y	1	n	n	n
Tb927.3.230	3	l	t	y	n	n	ad	2u	y	y	2	n	n	n
Tb927.3.240	3	l	t	y	n	n	b	1d	y	y	1	n	n	n
Tb927.3.260	3	l	t	n	n	n	n	1d	n	y	1	n	n	n
Tb927.3.270	3	l	t	y	n	n	ad	2	y	y	1	n	n	n
Tb927.3.280	3	l	t	y	n	n	a	1d	y	y	1	n	n	n
Tb927.3.300	3	l	t	y	n	n	b	1d	y	y	1	n	n	n
Tb927.3.310	3	l	t	y	n	n	ad	3d	y	y	1	n	n	n
Tb927.3.320	3	l	t	y	n	n	ad	2	y	y	1	n	n	n
Tb927.3.330	3	l	t	y	n	n	d	2d	y	y	1	n	n	n
Tb927.3.340	3	l	t	y	y	y	b	6	y	y	1	n	n	n
Tb927.3.350	3	l	t	y	n	n	ad	4d	y	y	3	n	n	n
Tb927.3.360	3	l	t	y	n	n	bd	1d	y	y	1	n	n	n
Tb927.3.370	3	l	t	y	y	n	au	1u	y	y	1	n	n	n
Tb927.3.380	3	l	t	n	n	n	n	3d	n	y	0	n	n	n
Tb927.3.390	3	l	t	y	n	n	ad	1d	y	y	1	n	n	n
Tb927.3.400	3	l	t	y	n	n	bd	3d	y	y	1	n	n	n
Tb927.3.410	3	l	t	y	n	n	b	3d	y	y	5	n	n	n
Tb927.3.420	3	l	t	y	n	n	ad	1d	y	y	1	n	n	n

Tb927.3.440	3	l	t	y	y	y	b	6	y	ya	2	n	n	n
Tb927.3.450	3	l	t	n	n	n	n	3d	n	y	0	n	n	n
Tb927.3.470	3	l	t	y	n	n	b	1d	y	y	1	n	y	n
Tb927.3.480	3	l	t	y	n	n	ad	1d	y	y	1	n	n	n
Tb927.3.490	3	l	t	y	y	n	a	2u	y	y	1	n	n	n
Tb927.3.500	3	l	t	y	n	n	bd	3d	y	y	1	n	n	n
Tb927.3.5850	3	r	b	y	n	n	ad	2d	y	y	1	y	n	n
Tb927.3.5860	3	r	b	y	n	n	a	1d	y	y	1	n	n	n
Tb927.3.5870	3	r	b	n	n	n	n	3d	n	y	0	n	n	n
Tb927.3.5880	3	r	b	y	n	n	a	2d	y	y	1	n	n	n
Tb927.3.5890	3	r	b	y	n	n	au	1d	y	y	1	n	n	n
Tb927.4.5400	4	r	t	y	y	n	a	5u	y	y	0	n	n	n
Tb927.4.5410	4	r	t	y	y	n	a	5u	y	y	1	n	n	n
Tb927.4.5420	4	r	t	y	y	n	a	5u	y	y	1	n	n	n
Tb927.4.5430	4	r	t	y	y	n	a	5u	y	y	1	n	n	n
Tb927.4.5450	4	r	b	y	n	n	ad	?d	y	y	1	3'	n	n
Tb927.4.5460	4	r	b	y	y	y	a	2	y	y	1	n	n	n
Tb927.4.5470	4	r	b	y	n	n	bd	3d	y	y	1	n	y	n
Tb927.4.5490	4	r	b	n	n	n	n	3d	n	f	0	n	n	n
Tb927.4.5500	4	r	b	y	n	n	au	3d	y	y	1	n	n	n
Tb927.4.5510	4	r	b	y	n	n	bd	3d	y	y	1	n	n	y
Tb927.4.5530	4	r	b	y	y	y	a	2	y	y	1	n	n	n
Tb927.4.5540	4	r	b	y	n	n	ad	d	y	y	1	n	n	n
Tb927.4.5550	4	r	b	n	n	n	n	2d	n	y	0	n	n	n
Tb927.4.5560	4	r	b	y	y	y	b	3	y	y	2	n	n	n
Tb927.4.5570	4	r	b	y	n	n	ad	2	y	y	1	n	n	n
Tb927.4.5580	4	r	b	y	y	n	a	2u	y	y	1	n	n	n
Tb927.4.5590	4	r	b	y	n	n	ad	d	y	y	1	n	n	n
Tb927.4.5600	4	r	b	y	n	n	ad	2d	y	y	1	n	n	n
Tb927.4.5610	4	r	b	y	n	n	bd	1d	y	y	1	n	n	n
Tb927.4.5620	4	r	b	y	n	n	ad	d	y	y	1	n	n	n
Tb927.4.5630	4	r	b	n	n	n	n	3d	n	y	0	n	n	n
Tb927.4.5640	4	r	b	y	n	n	bd	1	y	y	1	n	n	n
Tb927.4.5650	4	r	b	y	n	n	bd	3d	y	y	0	n	y	n
Tb927.4.5670	4	r	b	n	n	n	n	1d	n	y	1	y	n	n
Tb927.4.5680	4	r	b	y	n	n	bd	3d	y	y	1	n	n	n
Tb927.4.5690	4	r	b	y	n	n	b	6d	y	y	1	n	n	n
Tb927.4.5700	4	r	b	y	y	n	a	2u	y	y	1	n	n	n
Tb927.4.5710	4	r	b	y	n	n	bd	3d	y	y	1	n	n	n
Tb927.4.5720	4	r	b	y	n	n	ad	2d	y	y	1	n	n	n
Tb927.4.5730	4	r	b	y	n	n	a	3d	y	y	1	n	n	n
Tb927.4.5740	4	r	b	y	n	n	ad	1d	y	y	1	n	n	n
Tb927.4.5750	4	r	b	n	n	n	n	d	n	y	0	n	n	n
Tb927.4.5760	4	r	b	y	n	n	bd	3d	y	y	1	n	y	n
Tb927.4.5780	4	r	b	y	n	n	bd	3d	y	y	1	n	n	y
Tb927.4.5800	4	r	b	n	n	n	n	d	n	f	0	n	n	n
Tb927.5.180	5	l	t	y	n	n	ad	1d	y	y	1	n	n	n
Tb927.5.190	5	l	t	y	n	n	a	2d	y	y	1	n	n	n
Tb927.5.200	5	l	t	y	n	n	ad	2d	y	y	1	n	n	n
Tb927.5.220	5	l	t	y	n	n	bd	3d	y	y	1	n	y	n
Tb927.5.230	5	l	t	y	y	n	a	1u	y	y	2	n	n	n
Tb927.5.240	5	l	t	y	n	n	bd	1	y	y	1	n	n	n
Tb927.5.250	5	l	t	n	n	n	n	4d	n	f	0	n	n	n
Tb927.5.270	5	l	t	n	n	n	b	n	f	n	2	y	n	n
Tb927.5.3990	5	n	t	y	y	n	a	?u	y	y	0	n	n	n
Tb927.5.4650	5	r	b	y	n	n	bd	3d	y	y	0	n	y	n
Tb927.5.4670	5	r	b	y	y	n	a	5u	y	y	1	n	n	n
Tb927.5.4680	5	r	b	y	n	n	ad	2d	y	y	1	n	n	n
Tb927.5.4690	5	r	b	y	y	n	a	5u	y	y	15	n	n	n
Tb927.5.4700	5	r	b	y	n	n	b	2d	y	y	3	n	n	n
Tb927.5.4710	5	r	b	n	n	n	bd	1d	f	y	1	y	y	n
Tb927.5.4730	5	r	b	y	y	n	c	u	y	y	1	n	n	n
Tb927.5.4740	5	r	b	y	n	n	ad	1d	y	y	3	n	n	n
Tb927.5.4750	5	r	b	y	n	n	b	3d	y	y	3	n	y	n
Tb927.5.4770	5	r	b	y	y	n	a	2u	y	y	1	n	n	n
Tb927.5.4780	5	r	b	y	n	n	ad	1d	y	y	1	y	n	n
Tb927.5.4790	5	r	b	y	n	n	ad	2u	y	y	0	n	n	n
Tb927.5.4800	5	r	b	n	n	n	ad	n	f	n	1	n	n	n
Tb927.5.4810	5	r	b	y	y	n	a	4u	y	y	1	n	n	n
Tb927.5.4820	5	r	b	y	n	n	ad	1d	y	y	1	n	n	n
Tb927.5.4830	5	r	b	y	n	n	ad	4d	y	y	2	n	n	n
Tb927.5.4840	5	r	b	y	y	n	a	2u	y	y	1	n	n	n
Tb927.5.4850	5	r	b	y	n	n	bd	1u	y	y	1	n	y	n
Tb927.5.4870	5	r	b	y	n	n	bd	3d	y	y	1	n	n	n

Tb927.5.4880	5	r	b	y	n	n	ad	1d	y	y	1	n	n	n
Tb927.5.4890	5	r	b	n	n	n	n	3d	n	f	0	n	n	n
Tb927.5.4900	5	r	b	y	n	n	ad	2u	y	y	11	n	n	n
Tb927.5.4910	5	r	b	y	n	n	bd	1d	y	y	1	n	n	n
Tb927.5.4920	5	r	b	y	n	n	ad	5u	y	y	1	n	n	n
Tb927.5.4930	5	r	b	y	n	n	bd	2	y	y	1	n	n	n
Tb927.5.4940	5	r	b	y	n	n	ad	2	y	y	1	n	n	n
Tb927.5.4950	5	r	b	y	y	n	b	6u	y	y	1	n	n	n
Tb927.5.4960	5	r	b	y	n	n	ad	2d	y	y	1	n	n	n
Tb927.5.4970	5	r	b	y	n	n	bd	1d	y	y	1	n	n	n
Tb927.5.4980	5	r	b	y	n	n	a	2d	y	y	1	n	n	n
Tb927.5.4990	5	r	b	y	n	n	bd	1d	y	y	1	n	n	n
Tb927.5.5000	5	r	b	y	n	n	ad	2d	y	y	1	n	n	n
Tb927.5.5010	5	r	b	y	n	n	bd	3d	y	y	1	n	n	y
Tb927.5.5030	5	r	b	n	n	n	n	1d	n	f	0	n	n	n
Tb927.5.5040	5	r	b	y	n	n	a	3d	y	y	2	n	n	n
Tb927.5.5050	5	r	b	y	y	n	b	3u	y	y	0	n	n	n
Tb927.5.5060	5	r	b	n	n	n	bd	n	f	n	2	n	n	n
Tb927.5.5070	5	r	b	n	n	n	n	2d	n	f	0	n	n	n
Tb927.5.5080	5	r	b	y	y	y	a	2	y	y	2	n	n	n
Tb927.5.5090	5	r	b	y	n	n	b	1d	y	y	1	n	n	n
Tb927.5.5100	5	r	b	y	n	n	cd	5d	y	y	1	n	n	n
Tb927.5.5110	5	r	b	y	n	n	ad	1d	y	y	2	n	n	n
Tb927.5.5120	5	r	b	y	n	n	ad	1d	y	y	1	n	n	n
Tb927.5.5130	5	r	b	y	n	n	b	3d	y	y	1	n	n	y
Tb927.5.5150	5	r	b	y	n	n	ad	1d	y	y	1	n	n	n
Tb927.5.5160	5	r	b	y	n	n	ad	5d	y	y	1	n	n	n
Tb927.5.5170	5	r	b	y	n	n	bd	1d	y	y	1	n	n	y
Tb927.5.5190	5	r	b	y	n	n	bd	1d	y	y	0	n	n	n
Tb927.5.5200	5	r	b	n	n	n	bd	n	f	n	1	n	n	n
Tb927.5.5210	5	r	b	y	y	n	b	2u	y	y	2	n	n	n
Tb927.5.5220	5	r	b	y	n	n	bd	1d	y	y	1	n	y	n
Tb927.5.5240	5	r	b	y	n	n	b	d	y	y	4	n	n	n
Tb927.5.5250	5	r	b	n	n	n	n	3d	n	f	0	n	n	n
Tb927.5.5260	5	r	b	y	n	n	b	6d	y	y	1	n	n	n
Tb927.5.5270	5	r	b	y	n	n	cd	2	y	y	3	n	n	n
Tb927.5.5280	5	r	b	y	n	n	b	3d	y	y	9	n	n	n
Tb927.5.5290	5	r	b	y	n	n	bd	3d	y	y	1	n	n	n
Tb927.5.5300	5	r	b	y	n	n	a	1d	y	y	6	n	n	n
Tb927.5.5310	5	r	b	y	n	n	b	1d	y	y	2	n	n	n
Tb927.5.5320	5	r	b	y	n	n	ad	1d	y	y	2	n	n	n
Tb927.5.5330	5	r	b	y	n	n	bd	1d	y	y	1	n	y	n
Tb927.5.5350	5	r	b	n	n	n	n	d	n	y	2	y	n	n
Tb927.6.5210	6	r	b	y	n	n	ad	3d	y	y	1	n	n	n
Tb927.6.5220	6	r	b	n	n	n	n	2d	n	f	0	n	n	n
Tb927.6.5230	6	r	b	y	n	n	bd	1d	y	y	1	n	n	n
Tb927.6.5240	6	r	b	y	y	n	a	2u	y	y	2	n	n	n
Tb927.6.5250	6	r	b	n	n	n	n	2d	n	f	0	n	n	n
Tb927.6.5260	6	r	b	y	y	y	b	2	y	y	1	n	y	n
Tb927.6.5280	6	r	b	y	n	n	ad	3d	y	y	1	n	n	n
Tb927.6.5290	6	r	b	y	n	n	ad	2d	y	y	0	n	n	n
Tb927.6.5300	6	r	b	y	n	n	bd	3d	y	y	2	n	n	y
Tb927.6.5320	6	r	b	n	n	n	ad	d	y	f	1	n	n	n
Tb927.6.5330	6	r	b	y	n	n	bd	1d	y	y	1	n	n	y
Tb927.6.5350	6	r	b	y	n	n	ad	2	y	y	1	n	n	n
Tb927.6.5360	6	r	b	y	n	n	ad	2d	y	y	1	n	n	n
Tb927.6.5370	6	r	b	y	y	y	a	2	y	y	0	y	n	n
Tb927.6.5380	6	r	b	n	n	n	d	2d	f	f	1	n	n	n
Tb927.6.5390	6	r	b	y	n	n	bd	1d	y	y	1	n	y	n
Tb927.6.5410	6	r	b	y	n	n	ad	1d	y	y	1	n	n	n
Tb927.6.5420	6	r	b	y	n	n	ad	4d	y	y	1	n	n	n
Tb927.6.5430	6	r	b	n	n	n	ad	2d	y	f	1	n	n	n
Tb927.6.5440	6	r	b	y	n	n	ad	2u	y	y	2	n	n	n
Tb927.6.5450	6	r	b	y	y	n	au	2	y	y	0	n	n	n
Tb927.6.5460	6	r	b	n	n	n	bd	n	y	n	1	n	n	n
Tb927.6.5470	6	r	b	y	n	n	bd	3d	y	y	0	y	n	n
Tb927.6.5480	6	r	b	y	n	n	bd	d	y	y	1	n	y	n
Tb927.6.5500	6	r	b	y	n	n	bd	1d	y	y	1	n	y	n
Tb927.6.5520	6	r	b	y	n	n	ad	2d	y	y	1	n	n	n
Tb927.6.5530	6	r	b	y	n	n	bd	1d	y	y	1	n	n	n
Tb927.6.5540	6	r	b	y	n	n	ad	3d	y	y	1	n	n	n
Tb927.6.5550	6	r	b	y	y	n	a	3u	y	y	1	n	n	n
Tb927.6.5560	6	r	b	y	n	n	ad	2	y	y	1	n	n	n
Tb927.6.5570	6	r	b	n	n	n	n	2d	n	y	1	y	n	n

Tb927.6.5590	6	r	b	y	n	n	bd	3d	y	y	1	n	n	n
Tb927.6.5600	6	r	b	y	n	n	b	1d	y	y	2	n	n	n
Tb927.6.5610	6	r	b	y	n	n	ad	1d	y	y	1	n	n	n
Tb927.6.5620	6	r	b	y	n	n	bd	1d	y	y	1	n	n	n
Tb927.6.5630	6	r	b	y	n	n	au	1d	y	y	1	n	n	n
Tb927.6.5640	6	r	b	y	n	n	b	2d	y	y	1	n	n	n
Tb927.6.5650	6	r	b	y	n	n	bu	3d	y	y	1	n	n	n
Tb927.6.5660	6	r	b	y	n	n	ad	1d	y	y	1	n	n	n
Tb927.6.5670	6	r	b	y	n	n	ad	d	y	y	0	n	n	n
Tb927.6.5680	6	r	b	n	n	n	bd	n	f	n	0	n	n	n
Tb927.6.5690	6	r	b	n	n	n	d	n	f	n	1	n	n	n
Tb927.6.5700	6	r	b	n	n	n	n	3d	n	f	0	n	n	n
Tb927.6.5710	6	r	b	y	n	n	bd	3d	y	y	2	n	y	n
Tb927.6.5730	6	r	b	y	n	n	ad	2d	y	y	1	n	n	n
Tb927.6.5740	6	r	b	y	y	n	b	1u	y	y	0	n	n	n
Tb927.6.5750	6	r	b	n	n	n	bd	n	f	n	2	n	n	n
Tb927.6.5760	6	r	b	y	n	n	bd	3d	y	y	1	n	y	n
Tb927.6.5780	6	r	b	y	n	n	cd	3u	y	y	1	n	n	n
Tb927.6.5790	6	r	b	y	n	n	au	2d	y	y	1	n	n	n
Tb927.6.5800	6	r	b	n	n	n	n	3d	n	f	0	n	n	n
Tb927.6.5820	6	r	b	n	n	n	ad	n	y	n	1	n	n	n
Tb927.6.5830	6	r	b	y	n	n	ad	2d	y	y	0	y	n	n
Tb927.7.110	7	l	t	y	n	n	bd	1d	y	y	1	n	n	n
Tb927.7.120	7	l	t	y	n	n	ad	3d	y	y	1	n	n	n
Tb927.7.140	7	l	t	y	n	n	bd	1d	y	y	0	n	y	n
Tb927.7.150	7	l	t	y	n	n	bd	6	y	y	1	n	n	n
Tb927.7.6500	7	n	t	y	y	y	c	1	y	y	2	y	n	n
Tb927.7.6510	7	n	t	y	n	n	bd	6d	y	y	1	n	n	n
Tb927.7.6520	7	n	t	y	n	n	a	6d	y	y	1	n	n	n
Tb927.7.6530	7	n	t	y	n	n	b	3d	y	y	1	n	n	n
Tb927.7.6540	7	n	t	y	y	n	a	3u	y	y	2	n	n	n
Tb927.8.100	8	l	t	n	n	n	ad	2	f	y	0	n	n	n
Tb927.8.110	8	l	t	y	n	n	ad	2d	y	y	2	n	n	n
Tb927.8.130	8	l	t	y	y	n	b	3u	y	y	1	n	y	n
Tb927.8.140	8	l	t	n	n	n	d	n	f	n	1	y	n	n
Tb927.8.150	8	l	t	y	n	n	a	3d	y	y	0	y	n	n
Tb927.8.160	8	l	t	n	n	n	n	3d	n	y	0	n	n	n
Tb927.8.170	8	l	t	y	y	n	c	1u	y	y	1	n	n	y
Tb927.8.180	8	l	t	n	n	n	n	3d	n	y	0	n	n	n
Tb927.8.190	8	l	t	y	n	n	bd	6d	y	y	1	n	n	n
Tb927.8.210	8	l	t	y	n	n	bd	1d	y	y	2	n	n	y
Tb927.8.220	8	l	t	y	n	n	bd	1d	y	y	1	n	n	n
Tb927.8.240	8	l	t	y	n	n	bd	1d	y	y	2	n	y	n
Tb927.8.250	8	l	t	y	n	n	bd	1d	y	y	1	n	n	n
Tb927.8.260	8	l	t	y	n	n	a	2d	y	y	2	n	n	n
Tb927.8.270	8	l	t	y	n	n	ad	1	y	y	1	n	n	n
Tb927.8.280	8	l	t	y	y	n	c	2u	y	y	3	n	n	n
Tb927.8.290	8	l	t	y	n	n	bd	1d	y	y	1	n	n	n
Tb927.8.300	8	l	t	y	n	n	ad	2d	y	y	1	n	n	n
Tb927.8.310	8	l	t	y	n	n	bd	1d	y	y	1	n	n	n
Tb927.8.320	8	l	t	y	n	n	ad	2d	y	y	0	y	n	n
Tb927.8.330	8	l	t	y	n	n	ad	2	y	y	2	n	n	n
Tb927.8.340	8	l	t	y	n	n	?d	d	y	y	3	n	n	n
Tb927.8.350	8	l	t	y	n	n	bd	6	y	y	1	n	n	n
Tb927.8.360	8	l	t	y	n	n	ad	2d	y	y	3	n	n	n
Tb927.8.380	8	l	t	y	n	n	bd	1d	y	y	2	n	y	n
Tb927.8.400	8	l	t	y	n	n	b	3d	y	y	1	n	y	n
Tb927.8.410	8	l	t	y	n	n	ad	3d	y	y	1	n	n	n
Tb927.8.420	8	l	t	y	n	n	au	1d	y	y	1	n	n	n
Tb927.8.430	8	l	t	y	n	n	ad	d	y	y	1	n	n	n
Tb927.8.440	8	l	t	y	n	n	b	d	y	y	2	n	n	n
Tb927.8.460	8	l	t	y	y	n	b	3u	y	y	0	n	y	n
Tb927.8.470	8	l	t	n	n	n	bd	3d	f	y	y	y	n	n
Tb08.27P2.120	8h		b	n	n	n	bd	3d	f	y	0	y	n	n
Tb08.27P2.140	8h		b	y	y	n	b	6u	y	y	1	n	n	n
Tb08.27P2.150	8h		b	y	n	n	bd	1d	y	y	1	n	y	n
Tb08.27P2.165	8h		b	y	n	n	?d	3d	y	y	1	n	n	n
Tb08.27P2.190	8h		b	n	n	n	ad	d	y	f	1	n	n	n
Tb08.27P2.210	8h		b	y	n	n	ad	2	y	y	1	n	n	n
Tb08.27P2.220	8h		b	y	y	n	au	3	y	y	1	n	n	n
Tb08.27P2.230	8h		b	y	n	n	bd	1d	y	y	9	n	n	n
Tb08.27P2.240	8h		b	y	y	y	b	3	y	y	1	n	y	n
Tb08.27P2.260	8h		b	y	y	n	a	1u	y	y	1	n	n	n
Tb08.27P2.270	8h		b	y	n	n	a	2d	y	y	1	n	n	n

Tb08.27P2.290	8h		b	y	y	n	b	1u	y	y	0	n	n	n
Tb08.27P2.310	8h		b	y	n	n	bd	d	y	y	1	n	y	n
Tb08.27P2.340	8h		b	y	n	n	ad	3d	y	y	1	n	n	n
Tb08.27P2.370	8h		b	y	n	n	a	2d	y	y	1	n	n	n
Tb08.27P2.380	8h		b	y	n	n	a	3d	y	y	1	n	n	n
Tb08.27P2.390	8h		b	y	n	n	b	1d	y	y	1	n	y	n
Tb08.27P2.410	8h		b	n	n	n	n	d	n	f	1	y	n	n
Tb08.27P2.435	8h		b	n	n	n	n	3d	n	f	1	n	n	n
Tb08.27P2.460	8h		b	y	y	y	a	2	y	y	1	n	n	n
Tb08.27P2.480	8h		b	y	n	n	bu	3d	y	y	1	n	n	n
Tb08.27P2.490	8h		b	y	n	n	au	1d	y	y	2	n	n	n
Tb08.27P2.510	8h		b	y	n	n	ad	2	y	y	2	n	y	n
Tb08.27P2.530	8h		b	y	n	n	ad	5d	y	y	1	n	n	n
Tb08.27P2.550	8h		b	y	n	n	ad	3d	y	y	1	n	n	n
Tb08.27P2.565	8h		b	n	n	n	n	6d	n	f	0	n	n	n
Tb08.27P2.580	8h		b	y	n	n	bd	3	y	y	1	n	n	n
Tb08.27P2.610	8h		b	y	n	n	ad	1d	y	y	1	n	n	n
Tb08.27P2.620	8h		b	y	n	n	ad	3d	y	y	1	n	n	n
Tb08.27P2.630	8h		b	y	n	n	bd	1d	y	y	1	n	n	n
Tb08.27P2.640	8h		b	y	n	n	bd	1d	y	y	1	n	n	y
Tb08.27P2.650	8h		b	y	n	n	a	1d	y	y	1	n	n	n
Tb08.27P2.660	8h		b	y	n	n	bd	6d	y	y	1	n	n	n
Tb08.27P2.670	8h		b	y	y	n	b	1u	y	y	1	n	n	n
Tb08.27P2.680	8h		b	y	n	n	bd	3	y	y	1	n	n	n
Tb08.27P2.690	8h		b	n	n	n	bd	1d	f	y	1	n	n	n
Tb08.27P2.710	8h		b	y	n	n	bd	1d	y	y	1	n	n	n
Tb08.27P2.720	8h		b	y	n	n	a	2d	y	y	?	?	?	?
Tb09.354.0010	9	I	b	y	n	n	ad	2d	y	y	2	n	n	n
Tb09.354.0030	9	I	b	y	n	n	ad	2d	y	y	1	n	n	n
Tb09.354.0040	9	I	b	y	n	n	ad	2d	y	y	0	n	n	n
Tb09.354.0055	9	I	b	n	n	n	d	n	f	n	1	n	n	n
Tb09.354.0060	9	I	b	y	y	n	a	2u	y	y	0	y	n	n
Tb09.354.0070	9	I	t	n	n	n	ad	1d	f	y	0	y	n	n
Tb09.354.0090	9	I	t	y	y	y	b	6	y	y	1	n	y	n
Tb09.354.0100	9	I	t	n	n	n	n	6	n	y	0	n	n	n
Tb09.354.0120	9	I	t	y	n	n	b	1d	y	y	2	n	y	n
Tb09.354.0130	9	I	t	y	n	n	bd	3d	y	y	4	n	n	n
Tb09.354.0140	9	I	t	y	n	n	ad	3d	y	y	1	n	n	n
Tb09.354.0150	9	I	t	n	n	n	n	2d	n	y	1	n	n	n
Tb09.354.0160	9	I	t	y	n	n	b	3d	y	y	1	n	n	n
Tb09.354.0180	9	I	t	y	y	n	b	3u	y	y	1	y	y	n
Tb09.354.0200	9	I	t	y	n	n	cd	2	y	y	0	n	n	n
Tb09.354.0210	9	I	t	y	n	n	ad	1d	y	y	1	n	n	n
Tb09.354.0220	9	I	t	y	n	n	b	1d	y	y	1	n	n	n
Tb09.354.0230	9	I	t	y	n	n	ad	2d	y	y	1	n	n	n
Tb09.354.0240	9	I	t	y	n	n	bd	3d	y	y	1	n	n	n
Tb09.354.0250	9	I	t	y	n	n	bd	1d	y	y	1	n	n	n
Tb09.354.0260	9	I	t	y	n	n	ad	3d	y	y	1	n	n	n
Tb09.142.0010	9	I	t	y	n	n	a	1d	y	y	0	n	n	n
Tb09.142.0020	9	I	t	n	n	n	ad	d	y	f	1	n	n	n
Tb09.142.0040	9	I	t	y	n	n	bd	1d	y	y	1	n	y	n
Tb09.142.0050	9	I	t	y	n	n	ad	3d	y	y	1	n	n	n
Tb09.142.0060	9	I	t	y	n	n	ad	2d	y	y	1	n	n	n
Tb09.142.0070	9	I	t	y	n	n	ad	3d	y	y	1	n	n	n
Tb09.142.0090	9	I	t	y	n	n	ad	2d	y	y	1	n	n	n
Tb09.142.0100	9	I	t	y	y	n	a	5u	y	y	1	n	n	n
Tb09.142.0120	9	I	t	y	n	n	bd	3d	y	y	1	n	y	n
Tb09.142.0130	9	I	t	n	n	n	bd	1d	y	f	1	n	n	n
Tb09.142.0140	9	I	t	n	n	n	n	?	n	y	0	n	y	n
Tb09.142.0160	9	I	t	y	n	n	ad	1d	y	y	1	n	n	n
Tb09.142.0170	9	I	t	y	n	n	bd	3d	y	y	1	n	n	n
Tb09.142.0180	9	I	t	y	n	n	a	d	y	y	2	n	n	n
Tb09.142.0200	9	I	t	y	n	n	cd	d	y	y	1	n	n	n
Tb09.142.0210	9	I	t	n	n	n	n	6d	n	y	0	n	n	n
Tb09.142.0230	9	I	t	y	y	n	b	1u	y	y	1	n	y	n
Tb09.142.0240	9	I	t	y	y	y	b	6	y	y	2	n	n	n
Tb09.142.0245	9	I	t	n	n	n	n	4d	n	f	0	n	n	n
Tb09.142.0250	9	I	t	y	n	n	ad	1d	y	y	1	n	n	n
Tb09.142.0280	9	I	t	y	n	n	bd	3d	y	y	1	n	y	y
Tb09.142.0290	9	I	t	y	n	n	bd	3d	y	y	1	n	n	n
Tb09.142.0460	9	I	t	y	n	n	bd	3d	y	y	1	n	n	y
Tb09.142.0470	9	I	t	y	n	n	b	3d	y	y	1	n	n	n
Tb09.142.0480	9	I	t	y	n	n	bd	3d	y	y	1	n	n	n
Tb09.142.0490	9	I	t	y	n	n	ad	2d	y	y	1	n	n	n

Tb09.142.0500	9	l	t	y	n	n	ad	5d	y	y	1	n	n	n
Tb09.v4.0001	9	l	t	y	n	n	au	2d	y	y	0	y	n	n
Tb09.v4.0002	9	l	t	y	n	n	bd	3	y	y	1	n	y	n
No id	9	l	t	n	n	n	n	3d	n	f	0	n	n	n
Tb09.v4.0003	9	l	t	y	n	n	bd	3d	y	y	1	n	n	n
Tb09.v4.0004	9	l	t	y	n	n	au	1d	y	y	2	n	n	n
Tb09.v4.0005	9	l	t	y	y	n	a	2u	y	y	2	n	n	n
No id	9	l	t	y	n	n	ad	d	y	y		n	n	n
Tb09.160.0010	9	l	t	y	n	n	a	2d	y	y	1	n	n	n
Tb09.160.0030	9	l	t	y	n	n	ad	1u	y	y	1	n	n	n
Tb09.160.0040	9	l	t	n	n	n	a	n	f	n	1	y	n	n
Tb09.160.0045	9	l	t	n	n	n	n	4d	n	f	0	y	n	n
Tb09.160.0060	9	l	t	y	n	n	ad	2d	y	y	1	n	n	n
Tb09.160.0070	9	l	t	y	n	n	bd	1d	y	y	3	n	n	n
Tb09.160.0080	9	l	t	y	n	n	c	2d	y	y	1	n	n	n
Tb09.160.0100	9	l	t	y	n	n	bd	1d	y	y	1	n	n	n
Tb09.v2.0100	9	l	t	n	n	n	n	3d	n	f	0	n	n	n
Tb09.160.0110	9	l	t	y	y	y	a	2	y	y	2	n	n	n
Tb09.160.0120	9	l	t	y	n	n	b	1d	y	y	2	n	n	n
Tb09.160.0150	9	l	t	y	n	n	ad	2u	y	y	1	n	y	n
Tb09.v2.0110	9	l	t	n	n	n	n	3d	n	y	0	n	n	n
Tb09.160.0160	9	l	t	y	y	n	a	5u	y	y	2	n	n	n
Tb09.160.0170	9	l	t	n	n	n	b	d	y	f	1	n	n	y
Tb09.160.0180	9	l	t	y	n	n	ad	d	y	y	1	n	n	n
Tb09.160.0185	9	l	t	n	n	n	n	d	n	f	0	n	n	n
Tb09.v2.0120	9	l	t	n	n	n	n	1d	n	f	0	n	n	n
Tb09.160.0200	9	l	t	y	n	n	bd	1d	y	y	1	n	n	n
Tb09.160.0220	9	l	t	n	n	n	ad	d	y	f	1	y	n	n
Tb09.v2.0130	9	l	t	n	n	n	n	2d	n	f	0	y	n	n
Tb09.160.0230	9	l	t	y	n	n	ad	1d	y	y	1	n	n	n
Tb09.v2.0140	9	l	t	n	n	n	n	3d	y	f	0	n	n	n
Tb09.160.0250	9	l	t	y	n	n	b	6d	y	y	4	n	y	n
Tb09.v2.0150	9	l	t	n	n	n	n	1d	n	y	0	n	n	n
Tb09.160.0260	9	l	t	y	y	n	au	3	y	y	2	n	n	n
Tb09.160.0280	9	l	b	y	y	n	a	5b_u	y	y	0	y	n	n
Tb09.244.1940	9	r	b	n	n	n	n	6d	n	y	0	n	n	n
Tb09.244.1930	9	r	b	y	n	n	ad	3d	y	f	1	n	n	n
Tb09.244.1920	9	r	t	y	n	n	a	2d	y	y	1	n	n	n
Tb09.v2.0160	9	r	t	n	n	n	n	1d	n	f	0	n	n	n
Tb09.v2.0170	9	r	t	y	n	n	ad	2d	y	y	1	n	n	n
Tb09.244.1890	9	r	t	n	n	n	n	3d	n	y	0	n	n	n
Tb09.244.1880	9	r	t	y	n	n	ad	4d	y	y	1	n	n	n
Tb09.244.1870	9	r	t	y	n	n	a	5u	y	y	1	n	n	n
Tb09.244.1860	9	r	t	y	n	n	b	1d	y	y	1	n	n	n
Tb09.244.1850	9	r	t	y	n	n	ad	3d	y	y	1	n	n	n
Tb09.244.1830	9	r	t	y	y	n	b	1u	y	y	1	n	n	y
Tb09.v2.0200	9	r	t	y	n	n	bd	3d	y	y	1	n	y	n
Tb09.244.1800	9	r	t	n	n	n	ad	1d	f	y	1	y	n	n
Tb09.244.1790	9	r	t	y	y	n	au	5u	y	y	1	n	n	n
Tb09.244.1780	9	r	t	y	n	n	b	1d	y	y	3	n	y	n
Tb09.v2.0220	9	r	t	y	n	n	ad	2d	y	y	1	n	n	n
Tb09.v2.0230	9	r	t	n	n	n	d	n	f	n	1	y	n	n
Tb09.244.1750	9	r	t	y	n	n	c	5d	y	y	1	n	n	n
Tb09.244.1740	9	r	t	y	y	y	b	1	y	y	3	n	n	n
Tb09.244.1730	9	r	t	n	n	n	n	1d	n	y	0	n	n	n
Tb09.244.1720	9	r	t	y	n	n	a	6d	y	y	1	n	n	n
Tb09.244.1710	9	r	t	y	n	n	a	1d	y	y	1	n	n	n
Tb09.244.1700	9	r	t	y	n	n	a	2	y	y	1	n	n	n
Tb09.244.1690	9	r	t	y	n	n	bd	1	y	y	0	n	n	n
Tb09.244.1680	9	r	t	n	n	n	ad	2d	y	f	1	y	n	n
Tb09.v2.0240	9	r	b	y	n	n	bd	1d	y	y	1	n	n	n
Tb09.244.1620	9	r	b	y	n	n	ad	2d	y	y	1	y	n	n
Tb09.244.1600	9	r	b	y	y	n	b	1u	y	y	1	n	n	n
Tb09.244.1590	9	r	b	y	n	n	bd	3d	y	y	1	n	n	y
Tb09.v2.0250	9	r	b	n	n	n	n	1d	n	y	0	n	n	n
Tb09.244.1580	9	r	b	y	y	y	a	3	y	y	1	n	n	n
Tb09.244.1570	9	r	b	y	y	n	a	1u	y	y	1	n	n	n
Tb09.244.1560	9	r	b	y	n	n	bd	2d	y	y	1	n	n	n
Tb09.244.1550	9	r	b	y	n	n	ad	2	y	y	1	n	n	n
Tb09.244.1540	9	r	b	y	n	n	ad	2d	y	y	1	n	n	n
Tb09.244.1530	9	r	b	y	n	n	ad	1d	y	y	1	n	n	n
Tb09.244.1510	9	r	b	y	n	n	bd	1d	y	y	1	n	n	n
Tb09.244.1490	9	r	b	y	n	n	ad	3d	y	y	3	n	n	n
Tb09.244.1480	9	r	b	y	n	n	bd	3d	y	y	1	n	y	n

Tb09.244.1450	9	r	b	y	n	n	cd	2d	y	y	1	n	n	n
Tb09.244.1440	9	r	b	y	n	n	c	2d	y	y	2	n	n	n
Tb09.v2.0260	9	r	b	n	n	n	n	1d	n	f	0	y	n	n
Tb09.v2.0270	9	r	t	n	n	n	n	1d	n	f	0	y	n	n
Tb09.244.1410	9	r	t	y	y	y	b	3	y	y	1	n	n	n
Tb09.v2.0280	9	r	t	n	n	n	n	3d	n	f	0	n	n	n
Tb09.244.1390	9	r	t	y	n	n	a	1d	y	y	2	n	n	n
Tb09.244.1380	9	r	t	y	n	n	ad	3d	y	y	1	n	n	n
Tb09.244.1360	9	r	t	n	n	n	ad	d	y	n	1	n	n	n
Tb09.244.1350	9	r	t	y	n	n	cd	6d	y	y	1	n	n	n
Tb09.244.1330	9	r	t	n	n	n	d	1d	f	y	1	n	y	n
Tb09.244.1310	9	r	t	y	n	n	ad	2d	y	y	1	n	n	n
Tb09.244.1300	9	r	t	y	n	n	ad	d	y	y	1	y	n	n
Tb09.244.1280	9	r	b	y	n	n	ad	1d	y	y	1	y	n	n
Tb09.244.1260	9	r	b	y	n	n	ad	5d	y	y	1	n	n	n
Tb09.v2.0290	9	r	b	n	n	n	n	3d	n	f	0	n	n	n
Tb09.244.1250	9	r	b	y	n	n	bd	3d	y	y	1	n	n	y
Tb09.244.1230	9	r	b	y	n	n	ad	2d	y	y	1	n	n	n
Tb09.v2.0300	9	r	b	n	n	n	n	3d	n	f	0	n	n	n
Tb09.244.1220	9	r	b	y	n	n	bd	d	y	y	0	n	y	n
Tb09.244.1200	9	r	b	n	n	n	bd	n	f	n	2	n	n	n
Tb09.244.1190	9	r	b	y	n	n	bd	1d	y	y	1	n	n	n
Tb09.244.1180	9	r	b	y	n	n	bd	1d	y	y	1	n	y	n
Tb09.v2.0310	9	r	b	y	n	n	ad	1d	y	y	1	n	n	n
Tb09.244.1140	9	r	b	y	n	n	ad	2d	y	y	3	n	n	n
Tb09.v2.0320	9	r	b	n	n	n	n	1d	n	f	0	n	n	n
Tb09.244.1130	9	r	b	y	n	n	ad	2	y	y	1	n	n	n
Tb09.244.1110	9	r	b	y	n	n	bd	1d	y	y	1	n	n	n
Tb09.244.1090	9	r	t	y	n	n	bd	2d	y	y	0	n	y	n
Tb09.244.1070	9	r	t	y	n	n	b	2d	y	y	3	n	n	n
Tb09.244.1050	9	r	t	n	n	n	bd	n	f	n	1	y	n	y
Tb09.v2.0330	9	r	b	n	n	n	n	d	n	f	0	n	n	n
Tb09.244.1020	9	r	b	y	n	n	bd	3d	y	y	0	y	n	n
Tb09.v2.0340	9	r	b	n	n	n	d	n	f	n	1	n	y	n
Tb09.244.1000	9	r	b	y	y	n	a	1u	y	y	1	n	n	n
Tb09.244.0990	9	r	b	y	n	n	bd	3d	y	y	2	n	n	n
Tb09.244.0980	9	r	b	y	n	n	a	2d	y	y	1	n	n	n
Tb09.244.0970	9	r	b	y	n	n	ad	1d	y	y	1	n	n	n
Tb09.244.0960	9	r	b	y	n	n	bd	1d	y	y	1	n	y	n
Tb09.v2.0350	9	r	b	n	n	n	n	1d	n	f	0	n	n	n
Tb09.244.0930	9	r	b	y	n	n	a	3d	y	y	1	n	n	n
Tb09.244.0920	9	r	b	y	n	n	ad	1d	y	y	1	n	n	n
Tb09.244.0910	9	r	b	y	y	y	a	2	y	y	1	n	n	n
Tb09.244.0900	9	r	b	y	n	n	bd	3d	y	y	1	n	n	n
Tb09.244.0880	9	r	b	n	n	n	bd	1d	f	y	1	n	n	n
Tb09.244.0870	9	r	b	y	n	n	b	1d	y	y	1	n	n	n
Tb09.244.0860	9	r	b	y	n	n	ad	1d	y	y	0	n	n	n
Tb09.244.0815	9	r	b	n	n	n	n	2d	n	f	0	y	n	n
Tb09.v2.0360	9	r	t	n	n	n	n	1d	n	y	0	n	n	n
Tb09.244.0800	9	r	t	y	n	n	ad	1u	y	y	1	n	n	n
Tb09.244.0790	9	r	t	y	n	n	ad	1d	y	y	1	n	n	n
Tb09.244.0780	9	r	t	y	n	n	a	1d	y	y	1	n	n	n
Tb09.244.0770	9	r	t	y	n	n	bd	1d	y	y	2	n	n	n
Tb09.244.0750	9	r	t	y	n	n	ad	5d	y	y	2	n	n	n
Tb09.v2.0370	9	r	b	y	n	n	ad	2d	y	y	0	n	n	n
Tb09.v2.0380	9	r	b	n	n	n	ad	n	f	n	1	n	n	n
Tb09.244.0710	9	r	b	y	n	n	bd	3d	y	y	1	n	y	y
Tb09.244.0690	9	r	b	y	n	n	ad	2d	y	y	1	n	n	n
Tb09.244.0680	9	r	b	n	n	n	bd	d	y	f	1	n	n	n
Tb09.244.0670	9	r	b	y	n	n	ad	1d	y	y	1	n	n	n
Tb09.244.0660	9	r	b	y	n	n	a	1d	y	y	1	n	n	n
Tb09.244.0650	9	r	b	y	n	n	ad	2d	y	y	1	n	n	n
Tb09.v2.0390	9	r	b	n	n	n	n	1d	n	f	0	n	n	n
Tb09.244.0640	9	r	b	y	y	n	a	u	y	y	1	n	n	n
Tb09.244.0630	9	r	b	y	n	n	ad	2d	y	y	1	n	n	n
Tb927.2.6480	9	r	b	y	n	n	ad	d	y	y	1	n	n	n
Tb927.2.6490	9	r	b	y	n	n	ad	1d	y	y	1	n	n	n
Tb09.v2.0400	9	r	b	n	n	n	ad	d	y	f	1	n	n	n
Tb927.2.6520	9	r	b	y	n	n	ad	1d	y	y	1	n	n	n
Tb927.2.6530	9	r	b	y	n	n	bd	1d	y	y	1	n	n	n
Tb927.2.6550	9	r	b	y	n	n	ad	3d	y	y	2	n	n	n
Tb927.2.6560	9	r	b	y	n	n	bd	2d	y	y	1	n	y	n
Tb09.v1.0950	9	r	b	y	n	n	cd	2d	y	y	1	n	n	n
Tb927.2.6590	9	r	b	y	n	n	c	2d	y	y	2	n	n	n

Tb09.v2.0410	9	r	b	n	n	n	n	1d	n	f	0	y	n	n
Tb09.v2.0420	9	r	t	n	n	n	n	6d	n	f	0	y	n	n
Tb927.2.6620	9	r	t	y	n	n	bd	1d	y	y	1	n	n	n
Tb927.2.6630	9	r	t	n	n	n	d	2	f	y	1	n	n	n
Tb927.2.6640	9	r	t	n	n	n	ad	d	y	f	1	n	n	n
Tb927.2.6650	9	r	t	y	y	y	b	3	y	y	1	n	n	n
Tb09.244.0475	9	r	t	n	n	n	n	2d	n	f	0	n	n	n
Tb09.v2.0430	9	r	t	y	n	n	ad	6d	y	y	1	n	n	n
Tb09.v2.0440	9	r	t	n	n	n	n	3d	n	f	0	y	n	n
Tb927.2.6690	9	r	b	n	n	n	bd	n	f	n	1	y	n	y
Tb927.2.6710	9	r	b	y	n	n	bd	2d	y	y	2	n	n	n
Tb927.2.6730	9	r	b	y	n	n	b	6d	y	y	1	n	n	n
Tb09.244.0375	9	r	b	y	n	n	ad	d	y	y	1	n	n	n
Tb927.2.6760	9	r	b	y	n	n	ad	1d	y	y	1	n	n	n
Tb927.2.6770	9	r	b	y	n	n	ad	d	y	y	3	n	n	n
Tb927.2.6790	9	r	b	y	n	n	b	1d	y	y	1	n	n	n
Tb927.2.6800	9	r	b	n	n	n	d	2	f	y	0	y	n	n
Tb927.2.6830	9	r	t	y	n	n	bd	1d	y	y	1	n	n	n
Tb927.2.6850	9	r	t	y	n	n	b	1d	y	y	1	n	y	n
Tb09.244.0250	9	r	t	y	y	n	u	2u	y	y	1	n	n	n
Tb927.2.6880	9	r	t	n	n	n	d	n	f	n	0	y	n	n
Tb927.2.6890	9	r	t	y	n	n	cd	3d	y	y	1	n	n	n
Tb927.2.6910	9	r	t	y	n	n	ad	1d	y	y	1	n	n	n
Tb09.244.0200	9	r	t	y	y	n	b	3u	y	y	1	n	y	n
Tb927.2.6970	9	r	t	n	n	n	bd	n	y	n	1	n	n	n
Tb09.244.0150	9	r	b	y	n	n	b	3d	y	y	1	n	n	n
Tb09.244.0140	9	r	b	y	n	n	ad	3d	y	y	1	n	n	n
Tb09.244.0130	9	r	b	y	n	n	b	1d	y	y	1	n	n	n
Tb09.244.0120	9	r	b	y	n	n	ad	5d	y	y	1	n	n	n
Tb09.244.0110	9	r	b	y	n	n	ad	1d	y	y	1	n	n	n
Tb09.244.0090	9	r	b	y	n	n	b	1d	y	y	2	n	n	n
Tb09.244.0080	9	r	b	n	n	n	ad	1d	f	y	1	y	n	n
Tb09.244.0060	9	r	b	y	n	n	ad	2	y	y	1	n	n	n
Tb09.244.0050	9	r	b	y	y	n	a	2u	y	y	1	n	n	n
Tb09.244.0030	9	r	b	y	n	n	bd	1d	y	y	1	n	n	n
Tb09.244.0020	9	r	b	y	n	n	bd	d	y	y	0	n	n	n
Tb09.v2.0060	9	r	b	n	n	n	ad	n	f	n	1	y	n	n
Tb09.v2.0070	9	r	b	y	n	n	bd	1d	y	y	1	n	n	n
Tb09.v2.0080	9	r	b	y	n	n	ad	2d	y	y	1	n	n	n
Tb09.v2.0090	9	r	b	y	n	n	b	6d	y	y	2	n	n	n
Tb09.244.0005	9	r	b	n	n	n	n	d	n	f	0	y	n	n
Tb09.218.0010	9		t	n	n	n	d	1d	f	y	0	n	n	n
Tb09.218.0020	9		t	y	y	y	c	?	y	y	2	n	n	n
Tb09.218.0030	9		t	y	n	n	bd	6d	y	y	0	n	n	n
Tb09.218.0040	9		t	y	n	n	cd	5d	y	y	0	n	n	n
Tb09.218.0050	9		t	y	n	n	bd	1d	y	y	1	n	n	n
Tb09.218.0060	9		t	y	n	n	bd	3d	y	y	2	n	n	n
Tb09.218.0070	9		t	n	n	n	n	3d	n	y	0	n	n	n
Tb09.218.0071	9		t	n	n	n	d	n	f	n	1	n	n	n
Tb09.218.0073	9		t	y	n	n	ad	2d	y	y	0	n	n	n
Tb09.218.0090	9		t	n	n	n	bd	n	f	n	1	y	n	y
Tb09.218.0110	9		b	n	n	n	bd	1d	y	f	1	n	n	n
Tb09.v4.0102	9		t	y	y	n	b	3u	y	y	1	n	n	y
Tb09.218.0150	9		t	y	n	n	ad	1u	y	y	2	n	n	n
Tb09.218.0260	9		t	n	n	n	a	1d	y	f	1	y	n	n
Tb09.218.0270	9		t	y	n	n	a	2d	y	y	1	n	n	n
Tb09.218.0280	9		t	y	y	n	b	1u	y	y	14	n	n	n
Tb09.218.0290	9		t	y	y	n	a	u	y	y	1	n	n	n
Tb09.218.0310	9		t	n	n	n	b	d	y	f	1	n	y	n
No id	9		t	n	n	n	d	1	f	y	0	n	n	n
Tb09.218.0330	9		t	n	n	n	bd	d	y	f	0	n	n	n
Tb09.218.0340	9		t	y	n	n	b	3d	y	y	1	n	n	n
Tb09.218.0350	9		t	y	n	n	b	1d	y	y	1	y	n	n
Tb09.218.0360	9		b	y	n	n	bd	3d	y	y	2	n	y	n
Tb09.218.0370	9		b	y	n	n	b	1d	y	y	1	n	n	n
Tb09.218.0390	9		b	y	n	n	bd	1d	y	y	1	y	n	n
Tb09.218.0400	9		b	y	n	n	ad	1d	y	y	1	n	n	n
Tb09.218.0410	9		b	y	n	n	b	6d	y	y	1	n	n	n
Tb09.218.0420	9		b	n	n	n	n	d	n	f	0	n	n	n
Tb09.218.0430	9		b	y	n	n	bd	4	y	y	2	n	y	n
Tb09.218.0450	9		b	y	n	n	bd	1d	y	y	1	n	y	n
Tb09.218.0470	9		b	y	n	n	a	1d	y	y	1	n	n	n
Tb09.v4.0134	9		b	y	n	n	ad	2	y	y	1	n	n	n
Tb09.218.0485	9		b	n	n	n	n	d	n	f	0	n	n	n

Tb09.218.0490	9		b	y	y	n	a	1u	y	y	1	n	n	n
Tb09.218.0500	9		b	y	n	n	bd	1d	y	y	0	y	n	n
Tb09.v4.0073	9		t	y	n	n	b	1d	y	y	1	n	n	n
Tb09.v4.0077	9		t	y	y	y	b	2	y	y	1	n	n	n
Tb09.354.0270	9		t	n	n	n	n	5d	n	y	1	y	n	n
Tb09.354.0290	9		t	y	n	n	bd	1d	y	y	1	n	n	y
Tb09.354.0300	9		t	y	n	n	a	d	y	y	1	n	n	n
Tb09.354.0320	9		t	y	n	n	bd	1d	y	y	1	n	y	n
Tb09.354.0330	9		t	y	n	n	ad	3d	y	y	1	n	n	n
Tb09.354.0350	9		t	y	n	n	bd	1d	y	y	1	n	n	n
No id	9		t	n	n	n	ad	n	f	n	1	y	y	n
Tb09.354.0370	9		t	y	n	n	ad	?u	y	y	1	n	n	n
Tb09.354.0390	9		t	y	n	n	bd	1d	y	y	1	n	y	n
Tb09.354.0400	9		t	y	n	n	b	3d	y	y	1	n	n	n
Tb09.354.0410	9		t	y	n	n	ad	1d	y	y	3	n	n	n
Tb09.354.0420	9		t	n	n	n	ad	1d	f	f	1	n	n	n
Tb09.354.0430	9		b	n	n	n	au	n	y	n	1	y	n	n
Tb09.354.0440	9		b	y	n	n	ad	?u	y	y	4	n	n	n
Tb09.v4.0168	9		b	y	n	n	b	1d	y	y	1	n	n	y
Tb09.v4.0166	9		b	y	y	n	a	1u	y	y	1	n	n	n
Tb09.354.0475	9		b	n	n	n	n	1	n	f	0	y	n	n
Tb09.354.0480	9		b	n	n	n	bd	d	y	f	1	y	n	n
Tb09.354.0490	9		b	y	n	n	bd	3d	y	y	2	n	n	n
Tb09.354.0500	9		b	n	n	n	n	1d	n	f	0	n	n	n
No id	9			n	n	n	n	2d	n	f	0	n	n	n
Tb09.v4.0177	9			y	y	y	a	1	y	y	3	n	n	n
No id	9			n	n	n	n	3d	n	f	0	n	n	n
Tb09.v4.0178	9			y	n	n	ad	3d	y	y	1	n	n	n
Tb10.v4.0001	10		b	y	y	y	a	2	y	y	10	n	n	n
Tb10.v4.0004	10		b	y	n	n	ad	6d	y	y	1	n	n	n
Tb10.v4.0005	10		b	y	n	n	ad	2u	y	y	1	n	n	n
Tb10.v4.0006	10		b	y	n	n	bd	1d	y	y	1	n	n	n
Tb10.v4.0007	10		b	y	n	n	b	1d	y	y	1	n	n	n
No id	10		t	n	n	n	n	6d	n	y	0	y	n	n
Tb10.v4.0008	10		t	y	n	n	ad	2d	y	y	1	n	n	n
Tb10.v4.0009	10		t	y	n	n	bd	2d	y	y	1	n	n	n
Tb10.v4.0010	10		t	y	n	n	b	3d	y	y	1	n	n	n
Tb10.v4.0012	10		t	y	n	n	bd	1d	y	y	1	n	n	y
No id	10		t	n	n	n	n	3d	n	y	0	n	n	n
Tb10.v4.0013	10		t	y	n	n	bd	6d	y	y	1	n	n	n
Tb10.v4.0014	10		t	y	n	n	b	1d	y	y	1	n	n	y
Tb10.v4.0015	10		t	y	n	n	bd	1d	y	y	1	n	n	n
Tb10.v4.0016	10		t	n	n	n	bd	n	y	n	1	n	n	n
No id	10		t	n	n	n	bd	1d	f	y	0	n	n	n
Tb10.v4.0017	10		t	y	n	n	ad	1d	y	y	1	n	n	n
Tb10.v4.0018	10		t	y	n	n	ad	2d	y	y	1	n	n	n
Tb10.v4.0019	10		t	y	n	n	cd	1d	y	y	1	n	n	n
No id	10			y	n	n	bd	3d	y	y	1	n	n	n
No id	10			n	n	n	n	5u	f	y	1	n	n	n
Tb10.v4.0020	10			y	n	n	bd	1	y	y	1	n	n	n
Tb10.v4.0021	10			y	n	n	au	d	y	y	1	n	n	n
No id	10			n	n	n	n	2d	n	f	0	n	n	n
Tb10.v4.0022	10			y	n	n	ad	5u	y	y	1	n	n	n
Tb10.v4.0023	10			y	y	n	a	5u	y	y	1	n	n	n
Tb10.v4.0024	10			y	n	n	a	3d	y	y	2	n	n	n
Tb10.v4.0025	10			y	n	n	a	3d	y	y	2	n	n	n
Tb10.v4.0026	10			y	n	n	ad	1u	y	y	1	n	n	n
Tb10.v4.0027	10			y	n	n	a	d	y	y	1	n	n	n
Tb10.v4.0028	10			y	n	n	cd	1u	y	y	1	n	n	n
No id	10			n	n	n	d	3d	f	y	1	n	n	n
Tb10.v4.0029	10			n	n	n	d	d	f	f	1	n	n	n
Tb10.v4.0030	10			y	n	n	bd	2d	y	y	1	n	n	n
Tb10.v4.0031	10			y	y	y	b	2	y	y	45	n	n	n
Tb10.v4.0160	10		t	y	n	n	a	2d	y	y	1	n	n	n
Tb10.v4.0161	10		t	y	n	n	ad	6	y	y	1	n	n	n
Tb10.06.0030	10		t	n	n	n	n	3d	n	y	0	n	n	n
Tb10.06.0040	10		t	y	y	y	a	1	y	y	1	n	n	n
Tb10.06.0050	10		t	y	n	n	d	2d	y	y	1	n	n	n
Tb10.06.0070	10		t	n	n	n	bd	3d	f	y	1	n	n	n
Tb10.06.0080	10		t	y	n	n	bd	6d	y	y	2	n	n	n
Tb10.06.0090	10		t	y	n	n	ad	2u	y	y	1	n	n	n
Tb10.06.0100	10		t	y	n	n	ad	3d	y	y	1	n	n	n
Tb10.06.0105	10		t	n	n	n	n	3d	n	y	0	n	n	n
Tb10.06.0108	10		t	n	n	n	a	n	y	n	1	n	n	n

Tb10.06.0110	10		t	y	n	n	bd	1d	y	y	0	n	n	n
Tb10.06.0120	10		t	n	n	n	bd	2d	y	f	1	n	n	n
Tb10.06.0125	10		t	n	n	n	n	3d	n	y	0	n	n	n
Tb10.06.0130	10		t	y	n	n	bd	1d	y	y	2	n	n	n
Tb10.06.0140	10		t	y	n	n	ad	5d	y	y	1	n	n	n
Tb10.06.0155	10		t	n	n	n	n	3d	n	y	0	n	n	n
Tb10.06.0160	10		t	y	y	n	b	3u	y	y	1	n	n	n
Tb10.06.0170	10		t	y	n	n	ad	5d	y	y	2	n	n	n
Tb10.06.0180	10		t	n	n	n	bd	n	y	n	2	n	n	n
Tb10.06.0190	10		t	y	n	n	b	1d	y	y	0	n	y	n
Tb10.06.0200	10		t	y	n	n	bd	3d	y	y	2	n	n	n
Tb10.06.0220	10		t	y	n	n	bd	1d	y	y	1	n	n	n
Tb10.06.0230	10		t	y	n	n	a	1d	y	y	2	n	n	n
Tb10.06.0240	10		t	y	n	n	a	1d	y	y	2	n	n	n
Tb10.06.0250	10		t	y	n	n	b	3d	y	y	1	n	n	n
Tb10.06.0270	10		t	y	y	n	b	3u	y	y	4	n	n	y
Tb10.06.0280	10		t	y	n	n	ad	2d	y	y	1	n	n	n
Tb10.06.0290	10		t	y	n	n	ad	2d	y	y	1	n	n	n
Tb10.06.0300	10		t	y	n	n	a	1d	y	y	1	n	n	n
Tb10.v4.0193	10		t	y	n	n	a	2d	y	y	2	n	n	n
Tb10.v4.0105	10		b	n	n	n	b	d	y	f	1	y	n	y
Tb10.07.0020	10		b	y	y	n	au	2	y	y	1	n	n	n
Tb10.07.0030	10		b	y	n	n	bd	1d	y	y	1	n	n	n
Tb10.07.0040	10		b	y	n	n	bd	3d	y	y	1	n	n	n
Tb10.07.0050	10		b	n	n	n	n	6d	n	f	0	y	n	n
Tb10.07.0060	10		b	n	n	n	bd	n	f	n	1	y	n	n
Tb10.07.0070	10		b	y	n	n	ad	2d	y	y	1	n	n	n
Tb10.07.0080	10		b	y	n	n	c	2d	y	y	1	n	n	n
Tb10.07.0090	10		b	y	n	n	bd	1d	y	y	1	n	n	n
Tb10.07.0095	10		b	n	n	n	n	2d	n	y	0	n	n	n
Tb10.07.0100	10		b	y	n	n	a	3d	y	y	1	n	n	n
Tb10.07.0110	10		b	n	n	n	au	d	y	f	1	n	n	n
Tb10.07.0120	10		b	y	n	n	?d	2d	y	y	1	y	n	n
Tb10.07.0130	10		b	y	n	n	ad	1d	y	y	1	n	n	n
Tb10.v4.0120	10		b	y	n	n	ad	1d	y	y	1	n	n	n
Tb10.07.0150	10		b	y	n	n	ad	1d	y	y	1	n	n	n
Tb10.07.160	10		b	y	n	n	a	5d	y	y	1	n	n	n
Tb10.07.0170	10		b	y	n	n	bd	1d	y	y	1	n	n	n
Tb10.v4.0124	10		b	y	n	n	ad	1d	y	y	5	n	n	n
Tb10.v4.0144	10		b	n	n	n	n	d	n	y	0	n	n	n
Tb10.08.0020	10		b	y	y	n	bu	2	y	y	1	n	n	n
Tb10.08.0030	10		b	y	n	n	bd	2d	y	y	1	n	y	n
Tb10.08.0040	10		b	y	n	n	ad	2d	y	y	1	n	n	n
Tb10.08.0050	10		b	y	n	n	bd	1d	y	y	1	n	n	n
Tb10.08.0060	10		b	y	n	n	a	1d	y	y	1	n	n	n
Tb10.08.0070	10		b	y	n	n	bd	1d	y	y	1	n	n	n
Tb10.08.0080	10		b	y	y	n	bu	3u	y	y	1	n	n	n
Tb10.08.0090	10		b	y	n	n	b	1d	y	y	1	n	n	n
Tb10.08.0100	10		b	y	n	n	a	1d	y	y	1	n	n	n
Tb10.08.0110	10		b	n	n	n	n	3d	n	y	0	n	n	n
Tb10.08.0120	10		b	n	n	n	n	2d	n	y	0	n	n	n
Tb11.01.8850	11		b	y	n	n	bu	d	y	y	1	n	n	n
Tb11.05.0001	11		t	n	n	n	ad	1d	f	y	0	y	n	n
Tb11.06.0002	11		t	y	n	n	bd	1d	y	y	2	n	n	n
Tb11.06.0003	11		t	y	n	n	b	1d	y	y	1	n	n	n
Tb11.09.0001	11		b	y	n	n	ad	2d	y	y	1	n	n	n
Tb11.09.0002	11		b	y	n	n	ad	2	y	y	1	n	n	n
Tb11.09.0003	11		b	y	y	y	a	2	y	y	2	n	y	n
Tb11.09.0005	11		b	y	y	y	b	1u	y	y	2	n	n	n
Tb11.09.0006	11		b	y	n	n	bd	1d	y	y	2	n	n	n
Tb11.09.0007	11		b	y	n	n	a	1d	y	y	1	n	n	n
Tb11.09.0008	11		b	y	n	n	ad	2d	y	y	1	n	n	n
Tb11.09.0010	11		b	y	n	n	au	3d	y	y	1	n	n	n
Tb11.09.0011	11		b	y	n	n	bd	3d	y	y	0	y	n	n
Tb11.09.0012	11		b	n	n	n	n	2d	n	f	0	n	n	n
Tb11.09.0013	11		b	n	n	n	n	d	n	f	0	n	n	n
Tb11.13.0001	11		b	y	n	n	ad	2d	y	y	1	y	n	n
Tb11.13.0002	11		b	y	n	n	bd	6d	y	y	2	n	n	n
Tb11.13.0003	11		t	y	n	n	bd	3	y	y	0	n	n	n
Tb11.13.0004	11		b	y	y	n	b	2u	y	y	1	n	n	n
Tb11.13.0005	11		b	y	n	n	ad	3d	y	y	0	n	n	n
Tb11.13.0006	11		b	y	n	n	bd	1d	y	y	1	n	n	n
Tb11.13.0007	11		b	n	n	n	d	n	f	n	2	n	n	n
Tb11.13.0008	11		b	y	n	n	bd	1d	y	y	0	y	n	n

Tb11.14.0001	11		b	y	n	n	ad	6	y	y	1	n	n	n
Tb11.14.0002	11		b	y	n	n	ad	2d	y	y	2	n	n	n
Tb11.14.0003	11		b	y	n	n	ad	1d	y	y	2	n	n	n
Tb11.14.0004	11		b	y	n	n	cd	3u	y	y	3	n	y	n
Tb11.14.0006	11		b	y	n	n	bd	3d	y	y	1	n	n	n
Tb11.14.0007	11		b	y	y	n	a	2u	y	y	1	n	n	n
Tb11.14.0008	11		b	y	y	n	a	2u	y	y	2	n	n	n
Tb11.14.0009	11		b	y	n	n	bd	1d	y	y	1	n	n	n
Tb11.14.0010	11		b	y	n	n	bd	3d	y	y	0	n	n	n
Tb11.14.0011	11		b	y	n	n	ad	d	y	y	1	n	n	n
Tb11.14.0012	11		b	y	n	n	bd	6d	y	y	1	n	y	n
Tb11.14.0013	11		b	y	n	n	bd	2d	y	y	1	n	n	n
Tb11.14.0014	11		b	y	n	n	a	d	y	y	1	n	n	n
Tb11.14.0015	11		b	y	y	y	b	2	y	y	1	n	n	n
Tb11.14.0016	11		b	y	n	n	ad	3d	y	y	1	n	n	n
Tb11.14.0017	11		b	y	n	n	ad	2d	y	y	1	n	n	n
Tb11.14.0018	11		b	y	n	n	ad	1d	y	y	1	n	n	n
Tb11.14.0019	11		b	y	y	n	a	2u	y	y	1	n	n	n
Tb11.14.0021	11		b	y	n	n	ad	6d	y	y	1	n	n	n
Tb11.14.0022	11		b	y	n	n	bd	1d	y	y	1	n	n	n
Tb11.14.0023	11		b	y	n	n	ad	2d	y	y	1	n	n	n
Tb11.14.0024	11		b	y	n	n	ad	5d	y	y	1	n	n	n
Tb11.14.0025	11		b	y	n	n	ad	5d	y	y	1	n	n	n
Tb11.14.0026	11		b	y	n	n	bd	3d	y	y	1	n	n	n
Tb11.14.0027	11		b	n	n	n	bd	3d	f	y	0	y	n	n
Tb11.14.0029	11		b	n	n	n	n	3d	n	f	0	n	n	n
Tb11.14.0031	11		b	n	n	n	n	1d	n	f	0	n	n	n
Tb11.14.0032	11		b	y	n	n	ad	1d	y	y	1	n	n	n
Tb11.14.0033	11		b	n	n	n	n	3d	n	f	0	n	n	n
Tb11.14.0034	11		b	n	n	n	n	3d	n	y	0	n	n	n
Tb11.15.0001	11		b	y	n	n	ad	5d	y	y	2	n	n	n
Tb11.15.0002	11		b	y	n	n	bd	1d	y	y	1	y	n	n
Tb11.15.0003	11		t	y	n	n	ad	1d	y	y	1	y	n	n
Tb11.15.0004	11		t	y	n	n	bd	3d	y	y	1	n	n	y
Tb11.15.0006	11		t	y	n	n	bd	1d	y	y	1	n	n	n
Tb11.15.0007	11		t	y	n	n	b	3u	y	y	0	n	y	n
Tb11.16.0001	11		t	y	n	n	a	2d	y	y	1	n	n	n
Tb11.16.0002	11		t	y	y	n	a	2u	y	y	1	n	n	n
Tb11.16.0003	11		t	y	y	n	a	2u	y	y	1	n	n	n
Tb11.16.0004	11		t	y	n	n	bd	1d	y	y	1	n	n	n
Tb11.20.0001	11		t	n	n	n	n	2u	n	y	0	y	n	n
Tb11.20.0003	11		b	y	n	n	ad	1d	y	y	1	y	n	n
Tb11.20.0004	11		b	y	n	n	ad	2	y	y	1	n	n	n
Tb11.21.0001	11		t	y	n	n	ad	1d	y	y	2	n	n	n
Tb11.21.0002	11		t	y	n	n	ad	1d	y	y	1	n	n	n
Tb11.21.0003	11		t	y	n	n	ad	1d	y	y	2	n	n	n
Tb11.21.0004	11		t	y	y	n	a	2u	y	y	1	n	n	n
Tb11.21.0005	11		t	y	n	n	b	3d	y	y	1	n	n	n
Tb11.21.0006	11		t	y	n	n	bd	1d	y	y	1	n	n	n
Tb11.21.0007	11		t	y	n	n	ad	2d	y	y	1	n	n	n
Tb11.21.0008	11		t	n	n	n	d	n	f	n	1	n	n	n
Tb11.21.0009	11		t	y	n	n	bd	1d	y	y	0	n	n	n
Tb11.21.0011	11		t	n	n	n	bd	n	y	n	1	y	y	n
Tb11.24.0002	11		b	y	n	n	a	2d	y	y	1	n	n	n
Tb11.24.0003	11		b	y	n	n	bd	1u	y	y	1	n	y	n
Tb11.24.0005	11		b	y	n	n	bd	1d	y	y	2	n	y	n
Tb11.24.0007	11		b	y	y	n	b	1u	y	y	1	n	n	n
Tb11.24.0008	11		b	y	n	n	ad	d	y	y	1	n	n	n
Tb11.24.0009	11		b	n	n	n	d	1d	f	y	1	n	n	n
Tb11.24.0010	11		b	n	n	n	n	3d	n	y	0	y	n	n
Tb11.24.0011	11		b	n	n	n	b	d	y	f	1	y	n	n
Tb11.24.0012	11		b	y	y	n	bu	1	y	y	1	n	n	n
Tb11.24.0013	11		b	y	n	n	ad	2d	y	y	?	n	?	n
Tb11.24.0014	11		b	n	n	n	ad	n	y	n	2	y	n	n
Tb11.29.0001	11		t	y	n	n	bd	1d	y	y	?	n	n	n
Tb11.30.0001	11		b	y	n	n	a	3d	y	y	9	n	n	n
Tb11.30.0003	11		b	y	n	n	bd	1d	y	y	1	n	y	n
Tb11.30.0005	11		b	y	n	n	a	4d	y	y	2	n	n	n
Tb11.30.0007	11		t	y	n	n	ad	d	y	y	0	y	y	n
Tb11.30.0008	11		t	y	n	n	ad	1d	y	y	2	n	n	n
Tb11.30.0009	11		t	y	n	n	bd	6d	y	y	2	n	n	y
Tb11.30.0010	11		t	y	n	n	a	2d	y	y	1	n	n	n
Tb11.30.0012	11		t	y	n	n	b	1d	y	y	1	n	y	n
Tb11.30.0013	11		t	y	n	n	ad	2d	y	y	1	n	n	n

Tb11.30.0014	11		t	y	n	n	bd	2d	y	y	1	n	n	n
Tb11.30.0015	11		t	y	n	n	a	2d	y	y	1	n	n	n
Tb11.30.0016	11		t	y	n	n	ad	1u	y	y	1	n	n	n
Tb11.32.0001	11		t	y	n	n	ad	2u	y	y	1	n	n	n
Tb11.32.0002	11		t	y	n	n	a	1d	y	y	1	n	n	n
Tb11.34.0001	11		t	y	n	n	bd	3d	y	y	1	n	n	n
Tb11.35.0001	11		b	y	y	n	a	1u	y	y	2	n	n	n
Tb11.35.0002	11		b	y	n	n	b	3d	y	y	2	n	n	n
Tb11.35.0003	11		b	y	n	n	bd	3d	y	y	2	n	y	n
Tb11.35.0005	11		b	y	n	n	bd	3d	y	y	1	n	n	n
Tb11.35.0006	11		b	y	n	n	a	3d	y	y	1	n	n	n
Tb11.38.0001	11		b	n	n	n	b	n	y	n	1	y	n	n
Tb11.38.0002	11		b	y	n	n	a	1d	y	y	1	n	n	n
Tb11.38.0003	11		b	y	y	y	b	1	y	y	2	n	n	n
Tb11.38.0004	11		b	y	n	n	bd	1d	y	y	1	n	n	n
Tb11.38.0005	11		b	y	y	n	b	1u	y	y	1	n	n	n
Tb11.38.0006	11		b	y	n	n	c	2d	y	y	2	n	n	n
Tb11.38.0007	11		b	y	y	y	a	2	y	y	1	n	n	n
Tb11.38.0008	11		b	n	n	n	n	d	n	f	0	n	n	n
Tb11.38.0009	11		b	n	n	n	n	3d	n	f	0	n	n	n
Tb11.41.0001	11		t	y	n	n	b	1d	y	y	1	n	n	n
Tb11.43.0001	11		b	y	y	n	bu	1u	y	y	1	n	n	n
Tb11.43.0002	11		b	y	n	n	ad	2d	y	y	1	n	n	n
Tb11.43.0003	11		b	y	n	n	bd	1d	y	y	1	n	n	n
Tb11.43.0004	11		b	y	n	n	ad	2d	y	y	1	n	n	n
Tb11.44.0002	11		t	y	n	n	bu	1d	y	y	0	y	y	n
Tb11.44.0004	11		t	y	n	n	bd	1d	y	y	1	n	n	y
Tb11.44.0006	11		t	y	n	n	bd	3d	y	y	1	n	y	n
Tb11.44.0007	11		t	y	n	n	ad	4d	y	y	2	n	n	n
Tb11.44.0008	11		t	n	n	n	n	3u	n	f	0	n	n	n
Tb11.44.0009	11		t	y	n	n	cd	5u	y	y	1	n	n	n
Tb11.44.0010	11		t	y	n	n	b	6d	y	y	2	n	n	n
Tb11.45.0001	11		b	y	n	n	a	1d	y	y	1	n	n	n
Tb11.45.0002	11		b	y	n	n	b	3d	y	y	1	n	n	n
Tb11.45.0003	11		b	y	n	n	b	3d	y	y	1	n	y	n
Tb11.48.0002	11		t	y	n	n	bd	3d	y	y	2	n	y	n
Tb11.48.0003	11		t	y	n	n	a	3d	y	y	1	y	n	n
Tb11.49.0002	11		t	y	n	n	b	2d	y	y	1	n	y	n
Tb11.49.0003	11		t	y	n	n	ad	2u	y	y	1	n	y	n
Tb11.49.0004	11		t	n	n	n	bu	d	y	f	2	n	y	n
Tb11.49.0005	11		t	y	n	n	a	d	y	y	0	n	n	n
Tb11.49.0007	11		t	y	n	n	bd	3d	y	y	1	n	n	y
Tb11.49.0008	11		t	y	n	n	cd	1d	y	y	1	n	n	n
Tb11.49.0009	11		t	y	n	n	ad	1d	y	y	1	n	n	n
Tb11.49.0010	11		t	y	n	n	ad	2u	y	y	1	n	n	n
Tb11.49.0012	11		t	y	n	n	bd	3d	y	y	1	n	y	n
Tb11.51.0005	11		t	y	y	y	a	2	y	y	7	n	n	n
Tb11.54.0001	11		t	y	n	n	ad	2u	y	y	1	n	n	n
Tb11.57.0011	11		t	y	n	n	cd	d	y	y	1	n	n	n
Tb11.57.0012	11		t	y	n	n	b	3d	y	y	1	n	n	n
Tb11.57.0013	11		t	n	n	n	n	3d	n	f	0	n	n	n
Tb11.57.0015	11		t	y	n	n	b	6d	y	y	0	n	y	n
Tb11.57.0016	11		t	n	n	n	n	d	n	y	0	n	n	n
Tb11.57.0018	11		t	n	n	n	d	n	f	n	1	y	y	n
Tb11.57.0019	11		b	y	y	n	a	2u	y	y	1	n	n	n
Tb11.57.0020	11		b	y	n	n	ad	2d	y	y	1	n	n	n
Tb11.57.0021	11		b	y	n	n	bd	1d	y	y	1	n	y	n
Tb11.57.0023	11		b	y	n	n	ad	1d	y	y	2	n	n	n
Tb11.57.0024	11		b	y	y	n	b	1u	y	y	2	n	n	y
Tb11.57.0026	11		b	y	n	n	ad	3u	y	y	1	n	n	n
Tb11.57.0027	11		b	y	n	n	bu	1d	y	y	1	n	n	n
Tb11.57.0028	11		b	y	n	n	ad	5	y	y	1	n	n	n
Tb11.57.0029	11		b	y	n	n	ad	4d	y	y	1	n	n	n
Tb11.57.0030	11		b	y	n	n	ad	2d	y	y	1	n	n	n
Tb11.57.0031	11		b	y	n	n	ad	2d	y	y	1	n	n	n
Tb11.57.0032	11		b	y	n	n	bd	d	y	y	2	n	n	n
Tb11.57.0033	11		b	y	n	n	bd	1d	y	y	0	n	y	n
Tb11.57.0035	11		t	y	y	n	a	2u	y	y	0	n	n	n
Tb11.57.0037	11		t	y	n	n	b	3d	y	y	1	n	y	n
Tb11.57.0038	11		t	y	n	n	ad	2d	y	y	1	n	n	n
Tb11.57.0039	11		t	y	n	n	ad	2d	y	y	1	n	n	n
Tb11.57.0040	11		t	y	n	n	ad	1d	y	y	1	n	n	n
Tb11.57.0041	11		b	y	n	n	bd	3d	y	y	?	n	n	n
Tb11.57.0043	11		t	y	n	n	bd	2d	y	y	2	n	n	n

Tb11.v4.0072	11		t	y	n	n	ad	2	y	y	1	n	n	n
Tb11.57.0045	11		t	y	n	n	b	1d	y	y	0	n	n	n
Tb11.57.0049	11		b	y	n	n	bd	3d	y	y	1	n	y	n
Tb11.57.0051	11		b	y	n	n	ad	1d	y	y	1	n	n	n
Tb11.57.0052	11		b	y	y	n	b	1u	y	y	1	n	n	n
Tb11.57.0053	11		b	y	n	n	bd	1u	y	y	1	n	n	n
Tb11.57.0055	11		b	y	n	n	bd	1d	y	y	1	n	n	n
Tb11.57.0076	11		t	n	n	n	d	d	f	y	0	y	n	n
Tb11.57.0082	11		t	y	n	n	bd	2	y	y	2	n	y	n
Tb11.57.0083	11		t	y	n	n	ad	2d	y	y	1	n	n	n
Tb11.57.0084	11		b	y	y	y	a	2	y	y	?	n	?	?
Tb11.57.0085	11		b	n	n	n	n	d	n	f	0	n	n	n
Tb11.57.0086	11		b	n	n	n	n	d	n	f	0	n	n	n
Tb11.57.0087	11		t	n	n	n	n	1d	n	y	1	n	n	n
Tb11.57.0088	11		t	n	n	n	n	4d	n	f	0	y	n	n
Tb11.v4.0010	11		t	n	n	n	a	d	y	f	1	y	n	n
Tb11.v4.0015	11		t	y	y	y	a	2	y	y	6	n	n	n
Tb11.v4.0016	11		t	y	n	n	a	1d	y	y	1	n	n	n
Tb11.v4.0021	11			y	y	n	a	4u	y	y	1	n	n	n
Tb11.v4.0027	11			n	n	n	a	3d	y	f	1	n	n	n
Tb11.v4.0028	11			y	n	n	a	1d	y	y	1	n	n	n
Tb11.v4.0029	11			y	y	y	a	5	y	y	1	n	n	n
Tb11.v4.0030	11			y	y	n	a	2u	y	y	1	n	n	n
Tb11.v4.0033	11			y	n	n	b	1d	y	y	2	n	y	n
Tb11.v4.0034	11			y	y	y	b	3	y	y	1	n	n	y
Tb11.v4.0035	11			y	n	n	b	1d	y	y	1	n	n	n
Tb11.v4.0036	11			y	y	n	b	5b_u	y	y	2	n	n	y
Tb11.v4.0038	11			y	y	y	b	6	y	y	1	n	y	n
Tb11.v4.0040	11			y	y	n	a	2u	y	y	1	n	n	n
Tb11.v4.0050	11			y	y	y	b	1	y	y	11	n	n	n
Tb11.v4.0058	11			y	y	y	b	6	y	y	2	n	y	n
Tb11.v4.0063	11			y	n	n	a	1d	y	y	1	n	n	n
Tb11.v4.0065	11			y	y	n	a	2u	y	y	2	n	n	n
Tb11.v4.0067	11			y	n	n	b	1d	y	y	1	n	n	n
Tb11.v4.0068	11			y	n	n	b	1d	y	y	1	n	n	n
Tb11.v4.0070	11			y	n	n	a	1d	y	y	1	n	n	n
Tb11.v4.0074	11			y	n	n	ad	6	y	y	1	n	n	n
Tb11.56.0001	11			y	n	n	ad	2d	y	y	2	n	n	n
Tb11.56.0002	11			y	n	n	b	1d	y	y	1	n	n	n
Tb11.56.0003	11			y	n	n	ad	3d	y	y	1	n	n	n
No id	11			n	n	n	n	2d	n	f	0	n	n	n
Tb11.56.0004	11			y	n	n	ad	1d	y	y	1	n	n	n
Tb11.56.0006	11			y	n	n	bd	1d	y	y	1	n	n	n
Tb11.01.8840	11			n	n	n	d	2d	y	f	3	y	n	n
Tb11.1085	11			y	n	n	ad	1d	y	y	1	y	n	n
Tb11.0905	11			n	n	n	n	3d	n	y	0	n	n	n
Tb11.0910	11			y	y	y	a	1	y	y	1	n	n	n
Tb11.01.8870	11			y	n	n	a	1d	y	y	1	n	n	n
Tb11.01.8880	11			y	n	n	ad	5d	y	y	3	n	n	n
Tb11.0935	11			n	n	n	n	3d	n	y	0	n	n	n
Tb11.01.8890	11			y	n	n	bd	3d	y	y	2	n	n	n
Tb11.01.8910	11			y	n	n	bd	1d	y	y	0	n	n	n
Tb11.01.8920	11			y	n	n	a	1d	y	y	1	n	n	n
Tb11.1145	11			n	n	n	n	6d	n	y	0	n	n	n
Tb11.01.8930	11			n	n	n	b	2d	y	f	2	n	y	n
Tb11.0495	11			y	n	n	ad	1d	y	y	0	n	n	n
Tb11.0500	11			y	n	n	ad	6d	y	y	1	n	n	n
Tb11.0510	11			y	n	n	ad	1d	y	y	1	n	n	n
Tb11.0610	11			y	y	n	b	1u	y	n	2	n	n	n
Tb11.0620	11			y	n	n	ad	2d	y	n	1	n	n	n
Tb11.57.0058	?			y	n	n	ad	2d	y	y	1	n	n	n
Tb11.57.0059	?			y	n	n	ad	d	y	y	1	n	n	n
Tb11.57.0060	?			y	n	n	bd	1u	f	n	1	n	n	n
Tb11.57.0061	?			n	n	n	d	1u	f	y	1	n	n	n
Tb11.57.0063	?			y	n	n	ad	3d	y	y	2	n	n	n
Tb11.57.0064	?			y	n	n	bd	1d	y	y	2	n	n	n
Tb11.57.0065	?			y	n	n	bd	1d	y	y	0	y	n	n
Tb11.28.0001	?			y	y	n	a	2u	y	y	1	n	n	n
Tb11.28.0002	?			n	n	n	n	3d	n	y	0	n	n	n
Tb11.40.0001	?			y	n	n	ad	d	y	y	1	n	n	n
Tb11.40.0002	?			y	n	n	a	2d	y	y	1	n	n	n
Tb11.40.0004	?			y	n	n	ad	2d	y	y	1	n	y	n
Tb11.40.0005	?			y	n	n	b	1d	y	y	1	y	n	n
Tb11.40.0006	?			y	n	n	b	1d	y	y	1	y	n	n

Tb11.40.0007	?			y	y	n	a	1u	y	y	1	n	n	n
Tb11.40.0008	?			y	n	n	b	6d	y	y	1	n	n	y
Tb11.57.0066	?			y	n	n	b	1d	y	y	1	n	n	n
Tb11.57.0067	?			y	n	n	ad	2u	y	y	1	n	n	n
Tb11.57.0070	?			y	n	n	bd	6d	y	y	1	n	y	y
Tb11.57.0071	?			y	n	n	ad	d	y	y	1	n	n	n
Tb11.57.0072	?			y	n	n	bd	3d	y	y	1	n	n	n
Tb11.57.0073	?			y	n	n	bd	2d	y	y	1	n	n	n
Tb11.57.0074	?			y	n	n	a	1d	y	y	1	n	n	n
Tb11.57.0075	?			y	n	n	bd	1d	y	y	1	n	n	n
Tb11.31.0001	?			y	n	n	ad	6	y	y	1	n	n	n
Tb11.57.9999	?			n	n	n	n	d	n	f	0	n	n	n

7.2.2 *Atypical* VSG *table*

Table 7.3: Atypical *VSG*s and associated departures from expressed VSG consensus sequence.

The first column gives the GeneDB identifier, the second and third give the N and the C-terminal domain type, respectively. Domain types followed by u (uncertain), indicate they include "atypical" features. The remaining eight columns each highlight different problems associated with these VSGs. "GPI cl" indicates mutations at the cleavage site. "GPI hyd polar/not cons" indicates that the GPI signal hydrophobic extension contains polar or non conserved amino acids. "GPI hyd short" indicates that the GPI signal hydrophobic extension is shorter than the consensus, whereas "GPI hyd long" indicates that it is longer than the consensus. "cys C t" indicates non-conserved cysteine pattern in the C-terminal domain. "type 5" refers to C-terminal type 5, for which no consensus or cleavage site requirements are known. "Sig pep" indicates weak signal peptide prediction. "cys N t" indicates non-conserved cysteine pattern in the N-terminal domain. A description for each individual gene is available at VSGdb.

Atypical VSGs	N	C	GPI cl	GPI Hyd Polar/ not cons	GPI hyd short	GPI hyd long	Cys C t	Type 5	Sig pep	Cys N t
summary			38	22	9	19	30	12	5	4
Tb927.1.5330	b	3u	1	0	0	0	0	0	0	0
Tb927.2.6410	au	2u	0	0	0	1	0	0	0	0
Tb927.3.210	b	3u	0	1	0	0	0	0	0	0
Tb927.3.370	au	1u	1	0	0	1	0	0	0	0
Tb927.3.180	a	1u	1	0	0	0	0	0	0	0
Tb927.3.490	A	2u	0	0	0	1	0	0	0	0
Tb927.4.5400	A	5u	0	0	0	0	0	1	0	0
Tb927.4.5410	A	5u	0	0	0	0	0	1	0	0
Tb927.4.5420	A	5u	0	0	0	0	0	1	0	0
Tb927.4.5430	A	5u	0	0	0	0	0	1	0	0
Tb927.4.5580	A	2u	1	0	0	0	0	0	0	0
Tb927.4.5700	a	2u	1	1	0	0	0	0	0	0
Tb927.5.230	a	1u	1	0	0	0	1	0	0	0
Tb927.5.4950	B	6d	0	0	0	0	1	0	0	0
Tb927.5.5050	B	3d	0	0	0	1	0	0	0	0
Tb927.5.4670	a	5u	0	0	0	0	1	1	0	0
Tb927.5.4690	a	5u	0	0	0	0	0	1	0	0
Tb927.5.4730	c	4u	1	1	0	0	1	0	0	0
Tb927.5.4810	a	4d	0	0	0	1	1	0	0	0
Tb927.5.4840	A	2u	0	0	0	1	0	0	0	0
Tb927.5.3990	a	?u	1	1	0	0	0	0	0	0
Tb927.5.5210	b	2u	1	0	0	1	0	0	0	0
Tb927.5.4770	A	2u	0	1	0	0	0	0	0	0
Tb927.6.5240	a	2u	0	0	0	0	1	0	0	0

Tb927.6.5450	au	2	0	1	0	0	0	0	1	0
Tb927.6.5550	a	3d	0	1	0	0	0	0	0	0
Tb927.6.5740	b	1d	0	1	1	0	0	0	0	0
Tb927.7.6540	a	3d	0	0	0	1	0	0	0	0
Tb927.8.170	C	1d	1	0	0	0	0	0	0	0
Tb927.8.280	C	2d	0	0	0	1	0	0	0	0
Tb927.8.130	B	3u	1	1	0	0	0	0	0	0
Tb927.8.460	B	3u	1	0	0	0	0	0	0	0
Tb08.27P2.140	b	6d	1	0	0	0	0	0	0	0
Tb08.27P2.220	au	3	0	0	1	0	0	0	1	0
Tb08.27P2.260	a	1d	0	0	0	0	1	0	0	0
Tb08.27P2.290	b	1d	1	0	0	0	0	0	0	0
Tb08.27P2.670	b	1d	0	0	0	1	1	0	0	0
Tb09.142.0100	a	5d	0	0	0	1	0	1	0	0
Tb09.142.0230	b	1d	1	0	1	0	1	0	0	0
Tb09.160.0160	A	5u	0	0	0	0	0	1	0	0
Tb09.160.0260	ad	3	0	0	0	0	0	0	0	1
Tb09.160.0280	a	5u	0	0	0	0	0	1	0	0
Tb09.244.0050	a	2u	0	1	0	0	0	0	0	0
Tb09.244.1000	a	1u	0	1	0	0	0	0	0	0
Tb09.244.1570	a	1d	1	0	0	1	1	0	0	0
Tb09.244.1600	b	1	0	0	1	0	0	0	0	0
Tb09.244.1790	Au	5d	0	0	0	0	0	1	0	0
Tb09.244.1830	b	1d	1	0	0	1	1	0	0	0
Tb09.354.0060	a	2u	0	1	0	0	0	0	0	0
Tb09.354.0180	b	3u	1	1	0	0	0	0	0	0
Tb09.244.0640	a	d	1	0	0	0	0	0	0	1
Tb09.244.0250	a	2u	0	1	0	0	0	0	0	0
Tb09.244.0200	b	3d	1	0	0	1	1	0	0	0
Tb09.218.0280	b	1u	1	0	0	0	1	0	0	0
Tb09.218.0290	a	u	1	0	0	0	1	0	1	0
Tb09.218.0490	a	1u	0	0	0	0	1	0	0	0
Tb09.v4.0102	B	3u	1	0	0	0	1	0	0	0
Tb09.v4.0166	A	1u	0	0	0	0	1	0	0	0
Tb10.06.0160	b	3d	1	0	0	1	1	0	0	0
Tb10.06.0270	B	3u	1	1	0	0	1	0	0	0
Tb10.07.0020	Au	2	0	0	0	0	0	0	0	1
Tb10.08.0020	Bu	2	0	0	0	0	0	0	1	0
Tb10.08.0080	B	3u	1	0	0	0	0	0	0	0
Tb10.v4.0023	A	5u	1	0	0	0	0	1	0	0
Tb11.0610	B	1u	0	0	1	0	1	0	0	0
Tb11.13.0004	B	2d	1	0	0	0	1	0	0	0
Tb11.14.0007	A	2u	0	1	0	1	0	0	0	0
Tb11.14.0008	a	2u	0	1	0	0	0	0	0	0
Tb11.14.0019	a	2u	1	0	0	0	0	0	0	0
Tb11.16.0002	a	2d	0	0	1	0	0	0	0	0
Tb11.16.0003	a	2d	1	0	0	0	1	0	0	0
Tb11.21.0004	a	2u	0	0	0	0	1	0	0	0
Tb11.24.0007	b	1d	0	0	0	1	0	0	0	0
Tb11.24.0012	bu	1	0	0	0	0	0	0	0	1
Tb11.35.0001	a	1d	1	0	1	0	0	0	0	0
Tb11.38.0005	b	1d	0	0	0	0	1	0	0	0
Tb11.43.0001	bu	d	1	0	1	0	1	0	1	0
Tb11.57.0019	a	2u	1	1	0	0	0	0	0	0
Tb11.57.0024	b	1u	1	1	0	0	0	0	0	0
Tb11.57.0035	a	2u	1	0	0	0	0	0	0	0
Tb11.57.0052	b	1u	0	1	1	0	1	0	0	0
Tb09.v4.0005	a	2u	1	0	0	0	1	0	0	0
Tb11.v4.0021	a	4u	0	1	0	0	0	0	0	0
Tb11.v4.0030	a	2u	1	0	0	0	1	0	0	0
Tb11.v4.0036	b	5u	0	0	0	0	0	1	0	0
Tb11.v4.0040	a	2u	1	1	0	0	0	0	0	0
Tb11.v4.0065	A	2u	0	0	0	1	0	0	0	0
Tb11.28.0001	A	2u	0	0	0	0	1	0	0	0
Tb11.40.0007	A	1u	0	0	0	1	1	0	0	0

7.2.3 C-terminal type 5 domain alignment

Alignments were performed with clustalX, then edited with Bioedit. The edited sequences were realigned using ClustalW at Poly bioinformatique Lyonnais (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_clustalw.html), for the graphical output style.

```
5ex_CMI09-04-02   GDKK--GEVCTGTEEKDCDKTKCDWNAEKKECKVKEGAAVISAVIKAPLLLAFLLF-
5f_Tb11570028     GDNKTNAAACKGTEEGKCDTKKCAWDKDNKECKVKEGAAVISYLMKVPLLLAFLLL-
5a_Tb092441870    GDNKTNADACKGTEEGKCDTKKCAWDKDNKECKAKEGAAVISYLMKVPLLLAFLLLQ
5p_Tb092440120    GDNKTNAAACKGTEEGKCDTKKCTWDKDKNECKVKEGTFFISAVMKAPLLLAFLLL-
5a_Tb091600160    GDNKTTAADCKASSETNCDKTKCDWNAEKKQCKVKEGAAVISAVIKSPLLLEFLLLA
5p_Tb09v40079     DTVA--TTDCAATEADKCDTTKCTWNKEKEKCKVKDGAFIISAVIKSPFALFLLLP-
5p_Tb110930       HNKA--ATECLATEEKDCDKTKCNWNKEKSECKVKEGVFIISAVIKTTLLLAILFLY
5f_Tb11v40029     GDKK--DEECKGKLETNCDTTKCTWNKEENECKVKKGAAVISAVIKAPLLPAFLLF-
5p_Tb092441750    VDKK--DEVCTGAEEGKCDKTKCDWNAEKKQCKVKEGAVTISAVIKAPLLLAFLVLA
5p_Tb091420100    GDKK--DEVCKATDEKDCDKNKCEWNKEKNECKVKESAFIISVSIKAILLLKYFGFS
5a_Tb92754920     EQNK--AEVCTATEEKDCDTKKCDWNAENKQCKVKEGTVIISALMKAHLLLAFLIL-
5a_Tb092441790    GDNKTTAADFTGTEEDKCDKTKCDWNKEKNECKVKEGAVVISAVINAPLLLAFLILA
5a_Tb10v40022     GETK---EECTGTVETDCDKTKCTWNKEKNECKVKES-AYISSVTKVPLLLAFLLL-
5a_Tb10v40023     GEKK---EECTGTVETDCDKTKCTWNKEKNECKVKES-AYISSVTKVPLLLAFLLL-
5p_Tb10v40179     DDKT--GAECTATEEGKCDKNKCTWDKEKNQCKVKEGA-VISAVIKVLLVLAVLLF-
5a_Tb92754670     GDNKTTAAECKATEEGKCDKTKCTWNAEKKQFKVKEGAAVISAVIKAPLLLAFLLI-
5a_Tb92754690     GDNKTTTADCTGTEEGKCDKTKCDWNAEKKQCKVKEGAAVISAVIKAPLLLAVLLP-
5p_Tb92755100     GDNKTTAAECKATEEGKCDKNKCDWDKEKNECKVKIA-VVISAVTKVPLLFLQLFLE
5p_Tb092441260    GSKE--KNECTATEEGKYDKNKCTWDKEKNECKVKEGAVVISAVTKAPLLLAFLLF-
5a_Tb0827P2530    GDNKAIATDCKATEEGKCDKEKWTWDKEKKECKVKEG-AVLLAVIKVPLLFAFFLLA
                                .     . *. *  *: :::.: *.* .    :     :  :     :
Prim.cons.        GDNKTTA2ECTGTEEGKCDKTKCTWNKEKKECKVKEGAAVISAVIKAPLLLAFLLLA
```

Figure 7.1: Alignment of novel type 5 C-terminal domain silent copies with the C-terminal domain of expressed VSG CMI 09-04-02.

Cysteines are highlighted in yellow. The cleavage site is unknown, but could occur after the last glycine, as glycine is a known acceptor of GPI anchors. Each VSG is given the GeneDB identifier, preceded by 5ex, 5f, 5a, 5p (expressed, putative functional, atypical and pseudogene, respectively).

7.3 Supplementary material to Chapter 4

7.3.1 Primers

Table 7.4: Primers used in Chapter 4 experiments.

Primer name	sequence
Cloning of expressed VSGs	
Spliced leader	GTTTCTGTACTATATTG
16 mer	GTGTTAAAATATATCA
Internal primers for VSG sequencing	
2-1_Nifor	CGCAAACGCAACTATTAGCC
2-1_Cirev	CAAGTGCAACAGCCTTCTTG
3-1_Nifor	CAACAGCCAAATAAACGCCC
3-1_Cirev	TCAAGCTTATTGCACTCGGC
3-4_Nifor	AAGCTGCACCTACTAAGCGC
3-4_Cirev	GCAGTTAGCGTCTGTGCAAG
4-1_Nifor	GAGCTAGCAGGCTGCTTAAC
4-1_Cirev	GCATCCCTTACCCTCAAGGT
4-2_Nifor	AATGGCTGAATGCTGCATCA
4-2_Cirev	TTGCCATGTCGCTAGAAGAG
5-1_Nifor	AAAAGCGGACACGGCCGCAG
5-1_Cirev	GATGGTTGTGCGCAGAAGAA
5-2_Nifor	TAAATTTCTAACGGCGATCG
5-2_Cirev	CGCTAAATCAGCCTTGCCTC
7-1_Nifor	AGCAGACTCGCCGACTTAAA
7-1_Cirev	GCAGTATCAGTGCGCCTTCT
7-2_Nifor	CTATCAACAAACAACGCGGC
7-2_Cirev	CGGGCTAGCAGTACATGTCA
7-3_Nifor	CGAGCACAAAAACGGCATAA
7-3_Cirev2	CGTGGCTAGGGGCTCTAGAT
7-4_Nifor	AGTAACAGACAGGGCCAACG
7-4_Cirev2	AAGCTGCTTAATGAAGTCATCGT
8-3_Nifor	GCTCTCAGCCTAGGAAGCAA
8-3_Cirev	TCAGTGTTGGCGCTATTGTT
10-2_Nifor	GCATAGCAGCAGCACAGAAA
10-2_Cirev	GTTCCTGTTGCTGGGTTTGT
10-3_Cirev	GTTGGTTCGCCTTTCACTGT
10-7_Nifor	CGCTGAAAAATTTACTCGGC
10-7_Cirev	TTGTGTTATCTGCGTCGCA
10-10_Nifor	TGAAGTCGAACGGACAATTG
10-10_Cirev	TTGTGTTATCTGCGTCGCA
11-1_Nifor	CACTCAATGGACAAGTGCGG
11-1_Cirev	TTGGAGCTGTAGTGCCTCCT
11-2_Nifor	AGCAAGCCTCCAGGTATTCA
11-2_Cirev	GTATTGGTAGTGGCGCTTGG
11-9_Nifor	AAAACTCAAAAGGGTCGTGA
11-9_Cirev	GTTCCTGTTTTGGATTTGCA
11-13_Nifor	CGGAGTTACTCAGGAAAGCC
11-13_Cirev	AGCCGCTGTTTTGCGTATAG
Testing whether clone 28-10-02 is a mosaic	
10-2dA_95for	GACAGGCAGCGGATGACCAG
10-2dA_671for	ACGGCCAAACCGGGATGAAA
10-2dB_648for	CGACTTAGTCTGTCTTTGTG
10-2_Db_89for	CCATTGTCGACGCAACCGTT
10-2dB_657rev	GGCCGTCGACACAAAGACAG
10-2dB_1001rev	CCAGCGGATCTCGCTGTTCC
10-2dB_1362rev	TGCTTTCTATGCCAGTACAT
10-2dA_1024rev	ATCGCTTCGCGGACTTCCTC

7.3.2 Table with extended details of all sequenced clones

Table 7.5: Donor and location of each independent sequenced VSG clone.

Clone	VSG donor id	Donor location	Donor features and dom type	Dom type	VSG n
03-01-01	Tb10.v4.0001	BES chr 10	Functional	A2	1
03-01-02	Tb10.v4.0001	BES chr 10	Functional	A2	1
03-02-01	Tb10.v4.0001	BES chr 10	Functional	A2	1
03-02-02	Tb10.v4.0001	BES chr 10	Functional	A2	1
09-03-01	reads	Minichromosome?	?	A6	2
09-03-03	reads	Minichromosome?	?	A6	3
09-03-04	reads	Minichromosome?	?	B3	4
09-03-07	reads	Minichromosome?	?	A5	5
09-04-01	reads	Minichromosome?	?	A6	3
09-04-02	reads	Minichromosome?	?	A5	5
09-04-10	reads	Minichromosome?	?	C1	6
14-05-01	reads	Minichromosome?	?	A2	7
14-05-02	Tb09.v4.0077	tryp_IXb-217g08.q1c	Atypical	B2	8
14-06-01	Tb09.v4.0077	tryp_IXb-217g08.q1c	Atypical	B2	8
21-08-02[43]	Tb09.v4.0077	tryp_IXb-217g08.q1c	Atypical	B2	8
21-08-03	Tb09.v4.0077	tryp_IXb-217g08.q1c	Atypical	B2	8
22-07-01	Tb09.244.1580	Chr9	Functional	A3	9
22-07-02	Tb11.v4.0074	Chr11	Pseudogene A6	A6	10
	Tb10.v4.0161	tryp_X-324h11.p1k	Pseudogene A6		
	Tb11.14.0001	Chr11	Pseudogene A6		
22-07-03	Tb927.3.190	Chr3	Functional	A2	11
22-07-04	Tb11.30.0005	Chr11	Pseudogene A4	A4	12
	Tb11.v4.0021	Chr11	Atypical A4		
	Tb09.354.0090	Chr9	functional B6		
24-09-01	Tb09.v4.0102	tryp_IXb-218d07.p1c	Atypical	B2	13
28-10-02	Tb11.09.0005	Chr11	Functional	B3	14
	Tb11.13.0003	Chr11	Pseudogene		
28-10-03	Tb11.09.0005	Chr11	Functional	B3	15
	Tb11.13.0003	Chr11	Pseudogene		
	Read (3'donor)	AZ217061	?		
28-10-04	Same as 28-10-03	Same as 28-10-03	Same as 28-10-03	B3	15
28-10-07	Tb11.v4.0027	tryp_XI-339b02.q2k2024	Pseudogene	A3	16
	read	tryp_XI-934d12.q1k_rev	?		
	read	tryp_X-334d06.p1c	?		
28-10-10	Read	tryp_XI-1058d01.q1k	?	B2	17
	Tb11.57.0032	Chr11	Pseudogene Bd		
	Tb09.244.0090	Chr9	Pseudogene B1		
28-11-01	read	CONTIG10023	1100 bp of VSG	B1	18
28-11-02	Tb927.5.5080	Chr5	Functional	A2	19
28-11-03	Tb927.5.5080	Chr5	Functional	A2	19
28-11-04	Tb927.5.5080	Chr5	Functional	A2	19
28-11-05	Short contig	CONTIG10023	1100 bp of VSG	B1	18
28-11-07	Tb927.5.5080	Chr5	Functional	A2	19
28-11-08	Tb927.5.5080	Chr5	functional	A2	19
28-11-09	Tb927.7.6530	Chr7	Pseudogene B3	B3	20
	read	1AL455630	?		
	read	2AQ660659	?		
	Read (3' donor)	3AL474336	?		
28-11-11	Tb927.5.5080	Chr5	functional	A2	19
28-11-13	Tb10.v4.0088	tryp_X-206a03.p1c	functional	A2	21
28-11-14	Tb10.v4.0088	tryp_X-206a03.p1c	functional	A2	21

[43] The day 21 infection corresponds to mouse 8 instead of 7 because blood was not harvested on day 21, but it was passaged in an immunocompromised mouse and harvested on day 23, after mouse 7 (day 22).

7.3.3 Table listing all point mutations found in VSG cDNAs

Table 7.6: All unique point mutations detected in sequenced VSG clones derived from chronic infection study, when compared with putative donors.

VSG (clone identifier)	VSG type	Codon position	Parent Codon	Aa	Codon	aa	Syn (N)/ non syn (Y)	Base change
03-01-02	A2	619	GCG	A	GAA	E	Y	2,3
03-03-01	A6	64	ATA	I	CTA	L	Y	1
03-03-01	A6	70	AAA	K	GAA	E	Y	1
03-03-01	A6	82	GGT	G	GGG	G	N	3
03-03-01	A6	196	AAG	K	AAC	N	Y	3
03-03-01	A6	220	ATC	I	AAC	N	Y	2
03-03-01	A6	436	GCC	A	GGC	G	Y	2
03-03-01	A6	496	GCA	A	GCC	A	N	3
09-03-03	A6	670	CTG	L	CTA	L	N	3
09-03-03	A6	820	GCA	A	GAA	E	Y	2
09-03-03	A6	832	ATT	I	ATC	I	N	3
09-03-04	B3	22	GTG	V	GGG	G	Y	2
09-03-04	B3	127	AGT	S	AGC	S	N	3
09-03-07	A5	373	GCA	A	ACA	T	Y	1
14-05-01	A2	943	AAG	K	AGG	R	Y	2
14-05-02	B2	1267	TTA	L	CTA	L	N	1
14-06-01	B2	127	AAA	k	AAC	N	Y	3
14-06-01	B2	310	AGC	S	AAC	N	Y	2
14-06-01	B2	430	AGC	S	AAC	N	Y	2
21-08-02	B2	583	GGA	G	AGA	R	Y	1
21-08-02	B2	643	GCG	A	ACG	T	Y	1
21-08-02	B2	688	AGC	S	GGC	G	Y	1
21-08-03	B2	607	GCC	A	GAC	D	Y	2
22-07-01	A3	58	GAA	E	CAA	Q	Y	1
22-07-01	A3	472	CAC	H	AAC	N	Y	1
22-07-01	A3	499	CCT	P	GCT	A	Y	1
22-07-01	A3	568	AAC	N	GAC	D	Y	1
22-07-01	A3	607	GAT	D	AAT	N	Y	1
22-07-01	A3	610	CTG	L	CTA	L	N	3
22-07-01	A3	688	GCC	A	GGC	G	Y	2
22-07-01	A3	736	AAA	K	AAG	K	N	3
22-07-01	A3	871	GGC	G	AAC	N	Y	1,2
22-07-02	A1	31	ATA	I	CTA	L	Y	1
22-07-02	A1	34	ACG	T	GCG	A	Y	1
22-07-02	A1	340	CGC	R	AGG	R	N	1,3
22-07-02	A1	343	AGC	S	GGC	G	Y	1
22-07-02	A1	502	GAA/GAT	E/D	GAG	E	N	3
22-07-02	A1	511	ACC	T	ACT	T	N	3
22-07-02	A1	517	ACC/GAT	T/D	GCC	A	Y	1 /2,3
22-07-03	A2	379	CAG	Q	CAA	Q	N	3
22-07-03	A2	388	CTT	L	CTA	L	N	3
22-07-03	A2	394	CAC	H	CAT	H	N	3
22-07-03	A2	448	GGC	G	AGC	S	Y	1
22-07-03	A2	457	CCC	P	GCC	A	Y	1
22-07-03	A2	460	TTA	L	TTG	L	N	3
22-07-03	A2	469	GCA	A	ACA	T	Y	1
22-07-03	A2	475	AGC	S	ACT	T	Y	2,3
22-07-03	A2	478	TCA	S	CCA	P	Y	1
22-07-03	A2	517	AAC	N	ACC	T	Y	2
22-07-03	A2	541	GTT	V	CTT	L	Y	1
22-07-03	A2	655	GCA	A	GCG	A	N	3
22-07-03	A2	790	GGC	G	GCC	A	Y	2
22-07-03	A2	820	AGC	S	AAC	N	Y	2
22-07-03	A2	925	AAA	K	CAA	Q	Y	1

22-07-03	A2	1138	GCA	A	GTA	V	Y	2
22-07-04	A4	316	ACC	T	ACG	T	N	3
24-09-01	B2	571	AAC	N	AGC	S	Y	2
24-09-01	B2	577	GGG	G	GGC	G	N	3
24-09-01	B2	637	GGA	G	GGC	G	N	3
24-09-01	B2	673	ACG	T	GCG	A	Y	1
24-09-01	B2	700	TCA	S	AAA	K	Y	1,2
24-09-01	B2	712	AAC	N	GAC	D	Y	1
24-09-01	B2	751	CCA	P	CCG	P	N	3
24-09-01	B2	781	GCG	A	ACG	T	Y	1
24-09-01	B2	835	GGG	G	GGC	G	N	3
24-09-01	B2	1240	GAA	E	GAC	D	Y	3
28-10-03	B3	55	AAC	N	AGC	S	Y	2
28-10-03	B3	226	GAC	D	GAG	E	Y	3
28-10-03	B3	244	GAA	E	GAC	D	Y	3
28-10-03	B3	310	GAT	D	AAT	N	Y	1
28-10-03	B3	406	GAC	D	GAA	E	Y	3
28-10-03	B3	412	GCC	A	GCA	A	N	3
28-10-03	B3	541	ACA	T	ACG	T	N	3
28-10-03	B3	550	ACC	T	AGC	S	Y	2
28-10-03	B3	676	CAC	H	CAT	H	N	3
28-10-03	B3	859	ACC	T	ACA	T	N	3
28-10-07	A2	106	AAC	N	AGC	S	Y	2
28-10-07	A2	145	AGT	S	ACT	T	Y	2
28-10-07	A2	154	ACC	T	ACA	T	N	3
28-10-07	A2	196	ATG	M	ATT	I	Y	3
28-10-07	A2	202	CAG	Q	CAA	Q	N	3
28-10-07	A2	244	AGC	S	AAC	N	Y	2
28-10-07	A2	373	GCC	A	GTC	V	Y	2
28-10-07	A2	433	GCA	A	AAA	K	Y	1,2
28-10-07	A2	472	GAA	E	GAG	E	N	3
28-10-07	A2	601	GGA	G	GGG	G	N	3
28-10-07	A2	604	AGC	S	ATC	T	Y	2
28-10-07	A2	616	GGC	G	GCC	A	Y	2
28-10-07	A2	676	ACG	T	ACA	T	N	3
28-10-07	A2	709	GGC	G	GAC	D	Y	2
28-10-07	A2	979	AAA	K	ACA	T	Y	2
28-10-07	A2	1075	ACA	T	GCA	A	Y	1
28-10-10	B2	154	ACG	T	GCG	A	Y	1
28-10-10	B2	604	AAC	N	AGC	S	Y	2
28-10-10	B2	610	GAA	E	GCA	A	Y	2
28-10-10	B2	634	ACC	T	AGC	S	Y	2
28-10-10	B2	703	AAA	K	AGA	R	Y	2
28-10-10	B2	748	CCC	P	GCC	A	Y	1
28-10-10	B2	790	GGA	G	GGC	G	N	3
28-11-01	B1	583	CCG	P	ACG	T	Y	1
28-11-01	B1	681	AAG	K	AAA	K	N	3
28-11-01	B1	700	CCA	P	GCA	A	Y	1
28-11-01	B1	823	AAC	N	GAC	D	Y	1
28-11-01	B1	1261	GTG	V	GCG	A	Y	2
28-11-01	B1	1582	AGA	R	AAA	K	Y	2
28-11-02	A2	202	GCG	A	ACG	T	Y	1
28-11-02	A2	235	TCC	S	GCC	A	y	1
28-11-02	A2	238	GAC	D	GAA	E	Y	3
28-11-02	A2	352	AGC	S	AGA	R	Y	3
28-11-02	A2	415	AGC	S	AAC	N	Y	2
28-11-02	A2	451	GAG	E	GAC	D	Y	3
28-11-02	A2	562	AAG	K	AAA	K	N	3
28-11-03	A2	103	CCG	P	CCC	P	N	3
28-11-03	A2	112	ATA	I	CTA	L	Y	1
28-11-03	A2	151	ACA	T	ACG	T	N	3
28-11-03	A2	241	GTC	V	GCC	A	Y	2
28-11-03	A2	301	AGC	S	AGA	R	Y	3
28-11-03	A2	439	ACC	T	AGC	S	Y	2
28-11-03	A2	553	ACT	T	GCT	A	Y	1
28-11-03	A2	868	GCA	A	GTA	V	Y	2

28-11-03	A2	880	ACG	T	GCG	A	Y	1
28-11-03	A2	1450	TTG	L	TTA	L	N	3
28-11-04	A2	52	AAC	N	GAC	D	Y	1
28-11-04	A2	148	TAC	Y	TAT	Y	N	3
28-11-04	A2	274	GCA	A	GCT	A	N	3
28-11-04	A2	316	AGC	S	AAC	N	Y	2
28-11-04	A2	487	GAG	E	GAA	E	N	3
28-11-04	A2	610	TCA	S	ACA	T	Y	1
28-11-04	A2	640	GGT	G	GCT	A	Y	2
28-11-04	A2	1402	GGA	G	GGG	G	N	3
28-11-04	A2	1429	GCC	A	GCA	A	N	3
28-11-07	A2	67	ACC	T	AGC	S	Y	2
28-11-07	A2	943	AAA	K	AAG	K	N	3
28-11-08	A2	244	GAA	E	GAC	D	Y	3
28-11-08	A2	382	GCA	A	ACA	T	Y	1
28-11-08	A2	493	AGA	R	AAA	K	Y	2
28-11-08	A2	565	ACT	T	AAT	N	Y	2
28-11-08	A2	610	TCA	S	TCG	S	N	3
28-11-08	A2	640	GGT	G	AAT	N	Y	1,2
28-11-08	A2	664	ACC	T	AAC	N	Y	2
28-11-08	A2	676	GCC	A	GGC	G	Y	2
28-11-09	B3	43	GCG	A	GCA	A	N	3
28-11-09	B3	85	GTG	V	GCG	A	Y	2
28-11-09	B3	88	CGT	R	GAT	D	Y	1,2
28-11-09	B3	94	CAG	Q	CAA	Q	N	3
28-11-09	B3	97	CCG	P	CTG	L	Y	2
28-11-09	B3	292	TAC	Y	CAC	H	Y	1
28-11-09	B3	298	ACC	T	AAC	N	Y	2
28-11-09	B3	307	TCG	S	GCG	A	Y	1
28-11-09	B3	628	GCC	A	GAA	E	Y	2,3
28-11-09	B3	655	GAT	D	GAC	D	N	3
28-11-09	B3	700	GTC	V	GTG	V	N	3
28-11-09	B3	796	ACG	T	GCG	A	Y	1
28-11-09	B3	931	GCG	A	GAC	D	Y	2
28-11-09	B3	1357	GGG	G	GGA	G	N	3
28-11-11	A2	160	AGC	S	AAC	N	Y	2
28-11-11	A2	181	AGA	T	AAA	K	Y	2
28-11-11	A2	232	CAG	Q	CGG	R	Y	2
28-11-11	A2	253	GCA	A	GCG	A	N	3
28-11-11	A2	445	ACC	T	AAC	N	Y	2
28-11-11	A2	640	GGT	G	GGC	G	N	3
28-11-11	A2	643	GCA	A	ACA	T	Y	1
28-11-13	A2	112	GCA	A	ACA	T	Y	1
28-11-13	A2	322	CCT	P	ACT	T	Y	1
28-11-13	A2	358	ACA	T	GCA	A	Y	1
28-11-13	A2	364	GGC	G	GGG	G	N	3
28-11-13	A2	373	GCC	A	GGC	G	Y	2
28-11-13	A2	424	GCT	A	AGT	T	Y	1
28-11-13	A2	661	GCT	A	GCG	A	N	3
28-11-13	A2	667	ACA	T	GCA	A	Y	1
28-11-13	A2	787	AGC	S	AAA	D	Y	2,3
28-11-13	A2	838	ACA	T	GCA	A	Y	1
28-11-13	A2	859	GGA	G	GGG	G	N	3
28-11-13	A2	1294	AAA	K	AGA	R	Y	2
28-11-14	A2	247	GCA	A	ACA	T	Y	1
28-11-14	A2	265	GCA	A	GRA	V	Y	2
28-11-14	A2	415	AGC	S	AAC	N	Y	2
28-11-14	A2	418	GGC	G	GAC	D	Y	2
28-11-14	A2	442	TCG	S	ACG	T	Y	1
28-11-14	A2	601	GGG	G	GGA	G	N	3
28-11-14	A2	676	GCC	A	GGG	G	Y	2,3
28-11-14	A2	727	AAC	N	CAC	H	Y	1
28-11-14	A2	892	GCA	A	CCA	P	Y	1
28-11-14	A2	931	CTA	L	TTA	L	N	1

7.3.4 Chi squared tests

Table 7.7: Chi squared test for non-random distribution of mutations in type A and type B N-terminal domains of expressed *VSGs*.

Point mutations found in the first 1000 bp of expressed *VSGs* were grouped according to domain type (type A and type B). The occurrence of mutations was recorded in 100 bp intervals and compared to the expected random distribution. The Chi squared value for nine degrees of freedom is 21.67 at the 0.01 significance level.

N-terminal domain 1000 bp, divided into ten groups	Mutation distribution (o)	Mutation distribution (e)	o-e	$(o-e)^2$	$(o-e)^2/e$
Type A N-terminal domains					
1-100	8	12	-4	16	1.33
101-200	12	12	0	0	0
201-300	13	12	1	1	0.08
301-400	16	12	4	16	1.33
401-500	22	13	9	81	6.23
501-600	9	12	-3	9	0.75
601-700	23	12	11	121	10.08
701-800	5	12	-7	49	4.08
801-900	9	12	-3	9	0.75
901-1000	5	13	-8	64	4.92
sum	122	122	0	366	29.57
Type B N-terminal domains					
1-100	7	5	2	4	0.8
101-200	3	5	-2	4	0.8
201-300	4	5	-1	1	0.2
301-400	3	5	-2	4	0.8
401-500	3	6	-3	9	1.5
501-600	6	5	1	1	0.2
601-700	12	5	7	49	9.8
701-800	10	5	5	25	5
801-900	3	5	-2	4	0.8
901-1000	1	6	-5	25	4.16
sum	52	52	0	126	24.06

Table 7.8: Chi squared test for preferential association of point mutation with predicted non-helical regions in the N-terminal domain of three expressed VSGs.

The Chi squared value for 1 degree of freedom is 3.84 at the 0.05 confidence level.

N-terminal domain secondary structure prediction	Mutation distribution (o)	Mutation distribution (e)	o-e	$(o-e)^2$	$(o-e)^2/e$
22-07-03, helical content 167/360 amino acids					
helical	3	7	-4	16	2.28
non helical	12	8	4	16	2
sum	15	15	0	32	4.28
24-09-01, helical content 164/360 amino acids					
helical	0	4	-4	16	4
non helical	8	4	4	16	4
sum	8	8	0	32	8
28-11-02 and related clones (see Chapter 4): 166/360 amino acids					
helical	12	19	-7	49	2.57
non helical	29	22	7	49	2.22
sum	41	41	0	98	4.80

7.4 Supplementary material to Chapter 5

7.4.1 VSG-related table

Table 7.9: *VSG*-related (*VR*) gene identifiers.

The first column gives the identifier used in Chapter 5, the second column gives the identifiers present in the sequence alignments given below (Figures 7.2 to 7.5), while the third column gives the GeneDB identifiers. Columns four to six give the cluster number to which *VR*s belong, based on full-length, N-terminal domain and C-terminal domain alignment respectively. The last column specifies whether *VR*s are pseudogenes (ps) (Y or N).

VR id	Old id	GeneDB id	FL cl	N cl	C cl	ps
1	1_VR_1013265	Tb927.1.5060	1	1	1	N
2	1_VR_1027246	Tb927.1.5170	1o	n	2	N
3	2_VR_393056	Tb927.2.2060	2.1	2.1	2.1	N
4	3_VR_374532	Tb03.1J15.190	3	4	2	N
5	3_VR_388094	Tb03.1J15.250	2.2	2.2	3	N
6	3_VR_394163	Tb03.1J15.200	2.2	2.2	3	N
7	3_VR_396343	Tb03.1J15.350	3	4	2	N
8	3_VR_637733	Tb03.48O8.310	3	3	1	N
9	3_VR_1539306	Tb03.2H15.460	2.1	2.1	2.1	N
10	5_VR_2864	Tb05.25N21.420	1	1	1	N
11	5_VR_14812	Tb05.25N21.360	1	1	1	N
12	5_VRd_1374081	Tb05.26C7.250	1	1	1	Y
13	8_VR_2088588	Tb08.10K10.350	2.2	2.2	3	N
14	8_VR_2092483	Tb08.10K10.330	2.2	2.2	3	N
15	9N_VR_900592	Tb09.160.5350	2.1	2.1	2.1	N
16	9_VRd_904454	Tb09.160.5370	3	3	1	Y
17	9_VR_923603	Tb09.160.5440	1	1	1	N
18	9_VR_926375	Tb09.v1.0300	3	3	1	N
19	9_VR_936679	Tb09.v1.0290	1	1	1	N
20	9N_VR_2175922	Tb09.244.2330	2.3	2.3	2.1	N
21	9N_VR_2184381	Tb09.244.2310	2.3	2.3	2.1	N
22	9N_VR_2192834	Tb09.244.2280	2.3	2.3	2.1	N
23	9N_VR_2201311	Tb09.244.2240	2.3	2.3	2.1	N
24	9N_VR_2209783	Tb09.244.2200	2.3	2.3	2.1	N
25	9r_VRd_86603	Tb09.142.0340	1	1	1	Y
26	11_VRd_1268929	Tb11.51.0003	1o	n	3	Y
27	11_2_VR	Tb11.02.1566	2.2	2.2	3	N
28	11_VRd_462214	Tb11.17.0003	1	1	1	Y
29	11s_VR_1122854	Tb11.01.4560	2.1	2.1	2.1	n
30	Chr 10_30829	No id				
31	Chr 10_20376	Tb10.v4.0003				

7.4.2 VR C-terminal domain alignments

Alignments in Figures 7.2 to 7.5 were performed with ClustalX, then edited with Bioedit. The edited sequences were realigned using ClustalW at Poly bioinformatique Lyonnais (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_clustalw.html), for the graphical output style.

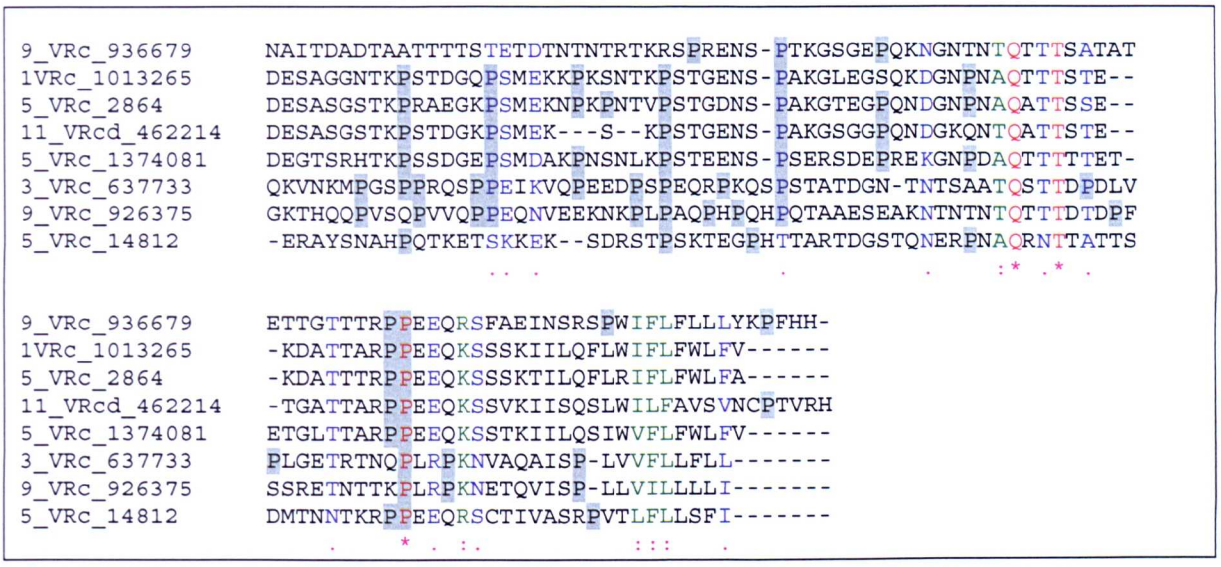


Figure 7.2: VR C-terminal domain Cluster 1 alignment.

Proline residues are highlighted in gray. Conversion to current GeneDB id can be made by use of Table 7.9.

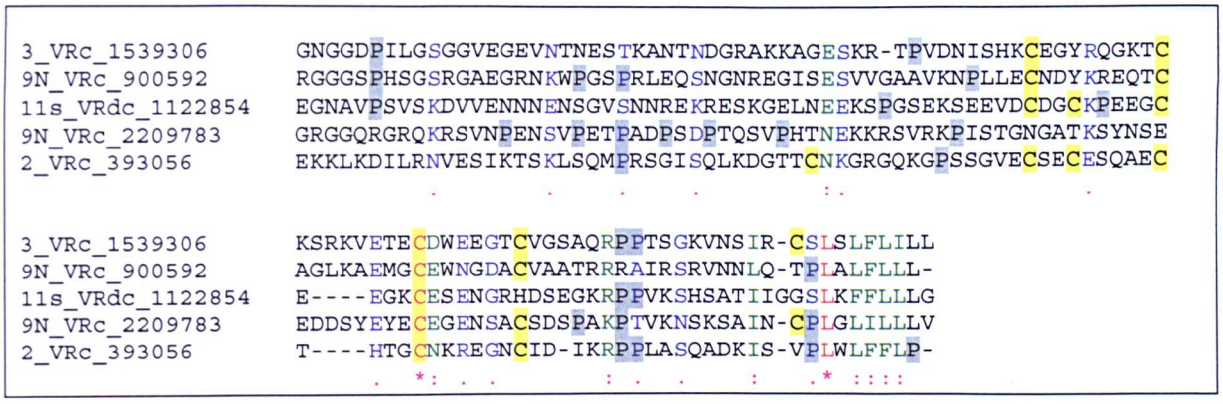


Figure 7.3: VR C-terminal domain Cluster 2.1 alignment (C-terminal domain cluster with cysteines).

Proline residues are highlighted in gray, cysteine residues in yellow. Conversion to current GeneDB id can be made by use of Table 7.9.

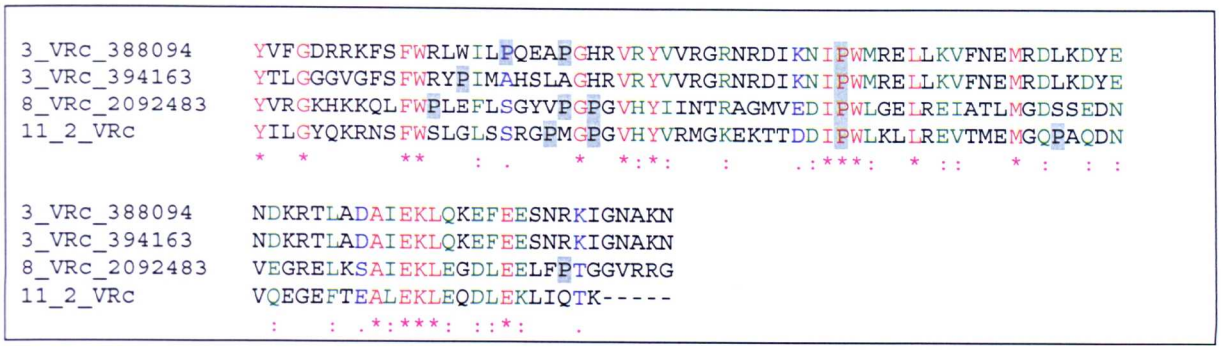


Figure 7.4: VR C-terminal domain Cluster 3 alignment.

Proline residues are highlighted in gray. Conversion to current GeneDB id can be made by use of Table 7.9.

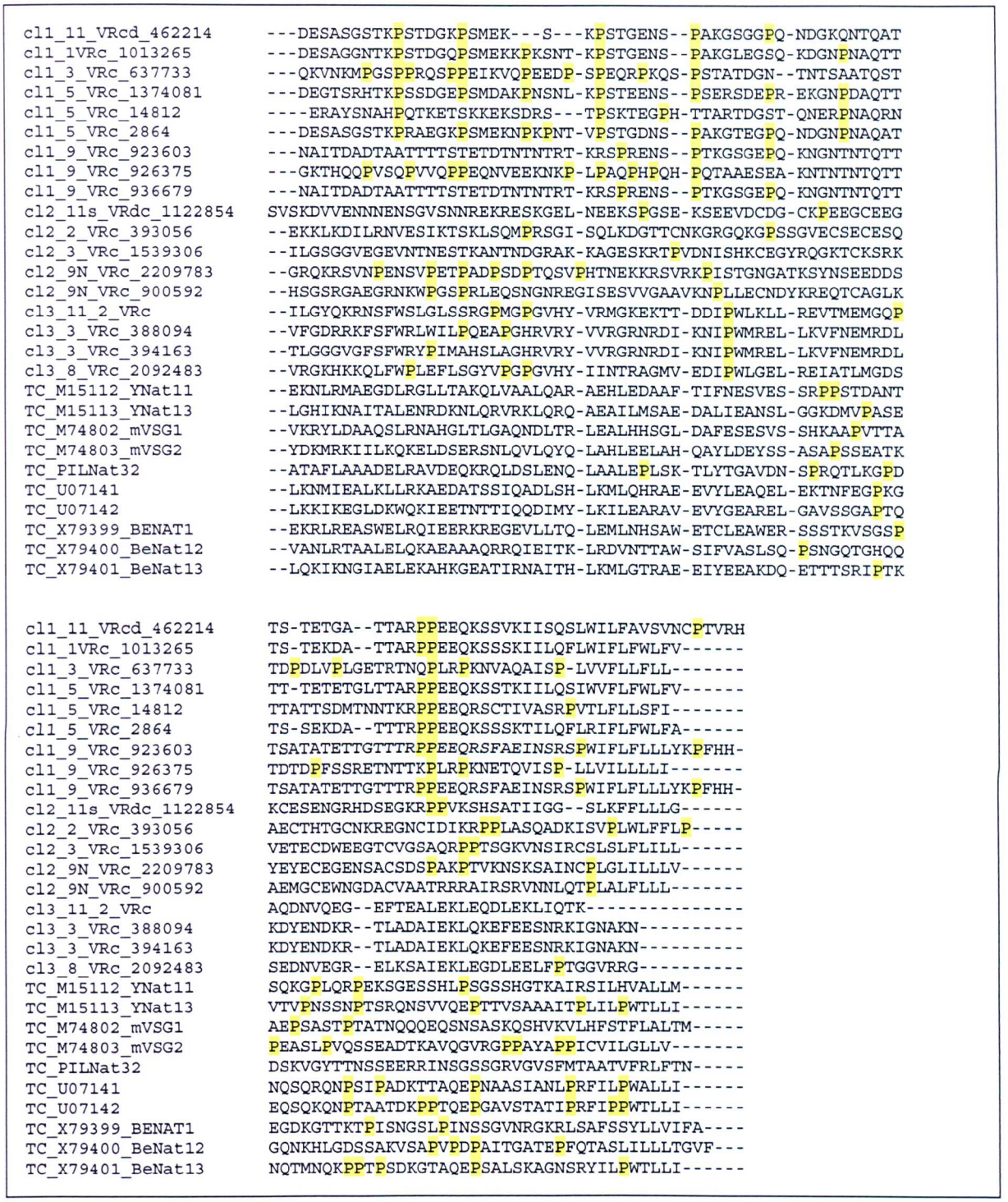


Figure 7.5: Multiple sequence alignment of C-terminal domain of 10 *Trypanosoma congolense* cDNAs with 18 *T. Brucei VR* C-terminal domains, from clusters 1, 2 and 3.

The first part of each sequence id specifies the nature of the domain: cl1, cl2 and cl3 correspond to the three *VR* C-terminal domain clusters, TC corresponds to *Trypanosoma congolense*.

7.4.3 Primers

Table 7.10: Primers used in Chapter 5 experiments.

Primer name	sequence
VR primers used in experiments described in section 5.3.1 and 5.3.2	
VR1_f	GCTACTGGAGCGAATGAAGG
VR1_r	TGCAATCCATTGCTTGTGTT
VR2_f	GGCAACGTCCAAAATATGCT
VR2_r	ATTCTTGCGGTTTGGATTTG
VR4_f	AGCTAAGCAGCATTGGGAGA
VR4_r	TCATTCGGTATTTTCTCGGC
VR5_f	GCCTCGTATGAATTTGGGAA
VR5_r	TTCCGGTACATCCATCAACA
VR8_f	ACCAAAGGGAGTTTGACGTG
VR8_r	TCTTCGGCTCTGTCCTTGTT
VR9_f	AACATCAGGAAGACCAACGC
VR9_r	TCTTTTGCTTTCACCAGCCT
VR11_f	TGGCCAAGATAGTGGAAACC
VR11_r	CGCATCCTGAATTGAACCTT
VR13_f	ACAACCCGAAGTGAATCAGG
VR13_r	TCAGCCCTCTTCCCTCTACA
VR15_f	GGAAATTTTGGACAGGAGCA
VR15_r	CGATGCACCACATAGACAC
VR18_f	GGAACACAAGCAGAACAGCA
VR18_r	CCTAGGTTGTTGGTGCGTTT
VSG primers used in experiment described in section 5.3.2	
221f	ATGCCTTCCAATCAGGAGGC
221r	TGTATGGGCGACAACTGCAG
121f	TAACCTTTACAACAGAGCGCACAAACTTAA
121r	CGCTGGCTGTGGTGCTCAGAATCATGCAGA
118f	CAGGTTCAGTCGAAGTATCAAGCAACAGGC
118r	TTCGTCTAGGACCCCGGCGGCCCTACCGGC
G4f	CTAACAGCAGCAGCGATGCAAAGTAGGATG
G4r	TCATCAAGGTAGTCCGTTGTGCGTGGCGTT
S8f	TCCAGCAAACGAGCGGATGCGGCGCTGGTG
S8r	CACGGCCTTGTCTTGTGCTGGCCCTGTTGT
VO2f	AGCCGCCTCGCCTGACGCAGCAACAGCGTG
VO2r	CGGTTCCGGCGCTGCAAAGGCAGAGCAAGT

7.4.4 T. congolense VSGs *analysed*

Table 7.11: Brief description of the 13 *T. congolense VSGs* analysed in current study.

The contig name and coordinate of start codon are given, in the absence of GeneDB identifiers. As for *T. brucei VSGs*, the sequences are characterised by whether they are "full-length", "intact" (no stop codons) and "functional" (putative functional). The question marks in the last column indicate that not enough data are available for expressed *T. congolense* VSGs to make accurate predictions based on a known consensus.

	coordinate	contig	Full length	Intact	functional
1	8988	congo502c05.p1k	y	y	?
2	7537	congo502c05.p1k	y	y	?
3	5320	congo502c05.p1k	y	y	?
4	2072	congo385g06.q1k	y	n	n
5	6118	congo385g06.q1k	y	n	n
6	27950	945999.c000216108	y	n	n
7	735	945999.c000541607	y	y	?
8	2860	Chr10	y	y	?
9	4230	Chr10	y	y	?
10	5540	Chr10	y	y	?
11	12812	Chr10	y	y	?
12	15138	Chr10	y	y	?
13	16858	Chr10	y	y	?

8 Reference list

Aguilera, A. (2002). The connection between transcription and genomic instability. *The EMBO Journal* **21**, 195-201.

Alarcon, C. M., Son, H. J., Hall, T. and Donelson, J. E. (1994). A monocistronic transcript for a trypanosome variant surface glycoprotein. *Mol. Cell. Biol.* **14**, 5579-5591.

Aline, R., Macdonald, G., Brown, E., Allison, J., Myler, P., Rothwell, V. and Stuart, K. (1985). $(TAA)_n$ within sequences flanking several intrachromosomal variant surface glycoprotein genes in *Typanosoma brucei*. *Nucleic Acids Research* **13**, 3161-3177.

Allen, G. and Gurnett, L. P. (1983). Locations of the six disulfide bonds in a variant surface glycoprotein (VSG 117) from *Trypanosoma brucei*. *Biochemical Journal* **209**, 481-487.

Allen, G., Gurnett, L. P. and Cross, G. A. M. (1982). Complete amino-acid sequence of a variant surface glycoprotein (vsg 117) from *Trypanosoma brucei*. *Journal of Molecular Biology* **157**, 527-546.

Alsford, S., Wickstead, B., Ersfeld, K. and Gull, K. (2001). Diversity and dynamics of the minichromosomal karyotype in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* **113**, 79-88.

Amiguet-Vercher, A., Perez-Morga, D., Pays, A., Poelvoorde, P., Van Xong, H., Tebabi, P., Vanhamme, L. and Pays, E. (2004). Loss of the mono-allelic control of the *VSG* expression sites during the development of *Trypanosoma brucei* in the bloodstream. *Mol. Microbiol.* **51**, 1577-1588.

Ansorge, I, Steverding, D., Melville, S., Hartmann, C. and Clayton, C. (1999). Transcription of 'inactive' expression sites in African trypanosomes leads to expression of multiple transferrin receptor RNAs in bloodstream forms. *Mol. Biochem. Parasitol.* **101**, 81-94.

Auffret, C. A. and Turner, M. J. (1981). Variant specific antigens of *Trypanosoma brucei* exist in solution as glycoprotein dimers. *Biochemical Journal* **193**, 647-650.

Baltz, T., Giroud, C., Bringaud, F., Eisen, H., Jacquemot, C. and Roth, C. W. (1991). Exposed epitopes on a *Trypanosoma equiperdum* variant surface glycoprotein altered by point mutations. *The EMBO Journal* **10**, 1653-1659.

Barbet, A. F., Davis, W. C. and McGuire, T. C. (1982). Cross-neutralization of two different trypanosome populations derived from a single organism. *Nature* **300**, 453-456.

Barrett, M. P., Bringaud, F., Doua, F., Melville, S. E. and Baltz, T. (1996). Hypervariability in gene copy number for the glucose-transporter genes in trypanosomes. *Journal of Eukaryotic Microbiology* **43**, 244-249.

Barry, J. D. (1986). Antigenic variation during *Trypanosoma vivax* infections of different host species. *Parasitology* **92**, 51-65.

Barry, J. D. (1997). The relative significance of mechanisms of antigenic variation in African trypanosomes. *Parasitology Today* **13**, 212-218.

Barry, J. D. (2006). *Implicit Information in Eukaryotic Pathogens as the Basis of Antigenic Variation.* In 'The Implicit Genome'. (Ed. L. H. Caporale.) pp. 91-106 (Oxford University Press: Oxford.)

Barry, J. D. and Carrington, M. (2004). *Antigenic Variation.* In 'The Trypanosomiases'. (I. Maudlin, P. H. Holmes and M. A. Miles Eds.) pp. 25-37. (CABI: Wallingford.)

Barry, J. D., Ginger, M. L., Burton, P. and McCulloch, R. (2003). Why are parasite contingency genes often associated with telomeres? *Int. J. Parasitol* **33**, 29-45.

Barry, J. D., Graham, S. V., Fotheringham, M., Graham, V. S., Kobryn, K. and Wymer, B. (1998). *VSG* gene control and infectivity strategy of metacyclic stage *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* **91**, 93-105.

Barry, J. D. and McCulloch, R. (2001). Antigenic variation in trypanosomes: Enhanced phenotypic variation in a eukaryotic parasite. *Advances In Parasitology* **49**, 1-70.

Beale, R. and Iber, D. (2006). *Somatic evolution of antibody genes.* In 'The Implicit Genome'. (Ed.L. H. Caporale.) pp. 177-190 (Oxford University Press: Oxford.)

Beals, T. P. and Boothroyd, J. C. (1992). Genomic organization and context of a trypanosome variant surface glycoprotein gene family. *J.Mol. Biol.* **225**, 961-971.

Becker, M., Aitcheson, N., Byles, E., Wickstead, B., Louis, E. and Rudenko, G. (2004). Isolation of the repertoire of *VSG* expression site containing telomeres of *Trypanosoma brucei* 427 using transformation-associated recombination in yeast. *Genome Research* **14**, 2319-2329.

Bell, J. S., Harvey, T. I., Sims, A. M. and McCulloch, R. (2004). Characterization of components of the mismatch repair machinery in *Trypanosoma brucei. Mol. Microbiol.* **51**, 159-173.

Belli, S. I. (2000). Chromatin remodelling during the life cycle of trypanosomatids. *International Journal for Parasitology* **30**, 679-687.

Benz, C., Nilsson, D., Andersson, B., Clayton, C. and Guilbride, D. L. (2005). Messenger RNA processing sites in *Trypanosoma brucei. Mol. Biochem. Parasitol.* **143**, 125-134.

Berberof, M., Perez-Morga, D. and Pays, E. (2001). A receptor-like flagellar pocket glycoprotein specific to *Trypanosoma brucei gambiense. Mol. Biochem. Parasitol.* **113**, 127-138.

Berberof, M., Vanhamme, L., Tebabi, P., Pays, A., Jefferies, D., Welburn, S. and Pays, E. (1995). The 3'-terminal region of the mRNAs for *VSG* and procyclin can confer stage specificity to gene expression in *Trypanosoma brucei. EMBO J.* **14**, 2925-2934.

Bernards, A., Van der Ploeg, L. H. T., Frasch, A. C. C., Borst, P., Boothroyd, J. C., Coleman, S. and Cross, G. A. M. (1981). Activation of trypanosome surface glycoprotein genes involves a duplication-transposition leading to an altered 3' end. *Cell* **27**, 497-505.

Bernards, A., Van der Ploeg, L. H. T., Gibson, W. C., Leegwater, P., Eijgenraam, F., de Lange, T., Weijers, P., Calafat, J. and Borst, P. (1986). Rapid change of the repertoire of variant surface glycoprotein genes in trypanosomes by gene duplication and deletion. *Journal of Molecular Biology* **190**, 1-10.

Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renauld, H., Bartholomeu, D. C., Lennard, N. J., Caler, E., Hamlin, N. E., Haas, B., Bohme, U., Hannick, L., Aslett, M. A., Shallom, J., Marcello, L., Hou, L., Wickstead, B., Alsmark, U. C., Arrowsmith, C., Atkin, R. J., Barron, A. J., Bringaud, F., Brooks, K., Carrington, M., Cherevach, I., Chillingworth, T. J., Churcher, C., Clark, L. N., Corton, C. H., Cronin, A., Davies, R. M., Doggett, J., Djikeng, A., Feldblyum, T., Field, M. C., Fraser, A., Goodhead, I., Hance, Z., Harper, D., Harris, B. R., Hauser, H., Hostetler, J., Ivens, A., Jagels, K., Johnson, D., Johnson, J., Jones, K., Kerhornou, A. X., Koo, H., Larke, N., Landfear, S., Larkin, C., Leech, V., Line, A., Lord, A., MacLeod, A., Mooney, P. J., Moule, S., Martin, D. M., Morgan, G. W., Mungall, K., Norbertczak, H., Ormond, D., Pai, G., Peacock, C. S., Peterson, J., Quail, M. A., Rabbinowitsch, E., Rajandream, M. A., Reitter, C., Salzberg, S. L., Sanders, M., Schobel, S., Sharp, S., Simmonds, M., Simpson, A. J., Tallon, L., Turner, C. M., Tait, A., Tivey, A. R., Van Aken, S., Walker, D., Wanless, D., Wang, S., White, B., White, O., Whitehead, S., Woodward, J., Wortman, J., Adams, M. D., Embley, T. M., Gull, K., Ullu, E., Barry, J. D., Fairlamb, A. H., Opperdoes, F., Barrell, B. G., Donelson, J. E., Hall, N., Fraser, C. M., Melville, S. E. and El Sayed, N. M. (2005). The genome of the African trypanosome *Trypanosoma brucei. Science* **309**, 416-422.

Berriman, M., Hall, N., Sheader, K., Bringaud, F., Tiwari, B., Isobe, T., Bowman, S., Corton, C., Clark, L., Cross, G. A., Hoek, M., Zanders, T., Berberof, M., Borst, P. and Rudenko, G. (2002). The architecture of variant surface glycoprotein gene expression sites in *Trypanosoma brucei. Mol. Biochem. Parasitol.* **122**, 131-140.

Bingham, J. and Sudarsanam, S. (2000). Visualizing large hierarchical clusters in hyperbolic space. *Bioinformatics.* **16**, 660-661.

Bitter, W., Gerrits, H., Kieft, R. and Borst, P. (1998). The role of transferrin-receptor variation in the host range of *Trypanosoma brucei. Nature* **391**, 499-502.

Blum, M. L., Down, J. A., Gurnett, A. M., Carrington, M., Turner, M. J. and Wiley, D. C. (1993). A structural motif in the variant surface glycoproteins of *Trypanosoma brucei. Nature* **362**, 603-609.

Bohme, U. and Cross, G. A. M. (2002). Mutational analysis of the variant surface glycoprotein GPI- anchor signal sequence in *Trypanosoma brucei*. *Journal of Cell Science* **115**, 805-816.

Boothroyd, J. C., Paynter, C. A., Coleman, S. L. and Cross, G. A. M. (1982). Complete nucleotide-sequence of complementary-DNA coding for a variant surface glycoprotein from *Trypanosoma brucei*. *Journal of Molecular Biology* **157**, 547-556.

Borst, P. and Fairlamb, A. H. (1998). Surface receptors and transporters of *Trypanosoma brucei*. *Annual Review of Microbiology* **52**, 745-778.

Borst, P. and Ulbert, S. (2001). Control of *VSG* gene expression sites. *Mol. Biochem. Parasitol.* **114**, 17-27.

Bringaud, F., Biteau, N., Donelson, J. E. and Baltz, T. (2001). Conservation of metacyclic variant surface glycoprotein expression sites among different trypanosome isolates. *Mol Biochem.Parasitol.* **113**, 67-78.

Bringaud, F., Biteau, N., Zuiderwijk, E., Berriman, M., El Sayed, N. M., Ghedin, E., Melville, S. E., Hall, N. and Baltz, T. (2004). The *Ingi* and RIME non-LTR retrotransposons are not randomly distributed in the genome of *Trypanosoma brucei*. *Mol Biol.Evol.* **21**, 520-528.

Brun, R. and Schonenberger, M. (1979). Cultivation and *in vitro* cloning of procyclic culture forms of *Trypanosoma brucei* in a semi-defined medium. *Acta Tropica* **36**, 289-292.

Campbell, D. A., Vanbree, M. P. and Boothroyd, J. C. (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 Research* **12**, 2759-2774.

Capbern, A., Giroud, C., Baltz, T. and Mattern, P. (1977). *Trypanosoma equiperdum*: étude des variations antigéniques au cours de la trypanosomose expérimentale du lapin. *Experimental Parasitology* **42**, 6-13.

Caporale, L. H. (2003). Natural selection and the emergence of a mutation phenotype: an update of the evolutionary synthesis considering mechanisms that affect genome variation. *Annual Review Of Microbiology* **57**, 467-485.

Carrington, M. and Boothroyd, J. (1996). Implications of conserved structural motifs in disparate trypanosome surface proteins. *Mol. Biochem. Parasitol.* **81**, 119-126.

Carrington, M., Miller, N., Blum, M., Roditi, I., Wiley, D. and Turner, M. (1991). Variant specific glycoprotein of *Trypanosoma brucei* consists of 2 domains each having an independently conserved pattern of cysteine residues. *Journal of Molecular Biology* **221**, 823-835.

Carruthers, V. B., Navarro, M. and Cross, G. A. M. (1996). Targeted disruption of Expression Site Associated Gene 1 in bloodstream form *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* **81**, 65-79.

Chattopadhyay, A., Jones, N. G., Nietlispach, D., Nielsen, P. R., Voorheis, H. P., Mott, H. R. and Carrington, M. (2005). Structure of the C-terminal domain from *Trypanosoma brucei* variant surface glycoprotein MITat1.2. *Journal of Biological Chemistry* **280**, 7228-7235.

Chaves, I, Rudenko, G., Dirks-Mulder, A., Cross, M. and Borst, P. (1999). Control of variant surface glycoprotein gene expression sites in *Trypanosoma brucei*. *The EMBO Journal* **18**, 4846-4855.

Chung, W. L., Carrington, M. and Field, M. C. (2004). Cytoplasmic targeting signals in transmembrane invariant surface glycoproteins of trypanosomes. *Journal of Biological Chemistry* **279**, 54887-95.

Conway, C., Proudfoot, C., Burton, P., Barry, J. D. and McCulloch, R. (2002). Two pathways of homologous recombination in *Trypanosoma brucei*. *Mol. Microbiol.* **45**, 1687-1700.

Cross, G. A. M., Wirtz, L. E. and Navarro, M. (1998). Regulation of *VSG* expression site transcription and switching in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* **91**, 77-91.

Cully, D. F., Gibbs, C. P. and Cross, G. A. M. (1986). Identification of proteins encoded by variant surface glycoprotein expression site-associated genes in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* **21**, 189-197.

Cully, D. F., Ip, H. S. and Cross, G. A. M. (1985). Coordinate transcription of variant surface glycoprotein genes and an expression site associated gene family in *Trypanosoma brucei*. *Cell* **42**, 173-182.

D'Orso, I., De Gaudenzi, J. G. and Frasch, A. C. (2003). RNA-binding proteins and mRNA turnover in trypanosomes. *Trends in Parasitology* **19**, 151-155.

Dai Do, Thi C., Aerts, D., Steinert, M. and Pays, E. (1991). High homology between variant surface glycoprotein gene expression sites of *Trypanosoma brucei* and *Trypanosoma gambiense*. *Mol.Biochem.Parasitol.* **48**, 199-210.

Dai, Q., Restrepo, B. I., Porcella, S. F., Raffel, S. J., Schwan, T. G. and Barbour, A. G. (2006). Antigenic variation by *Borrelia hermsii* occurs through recombination between extragenic repetitive elements on linear plasmids. *Molecular Microbiology* **60**, 1329-43.

Das, A. and Bellofatto, V. (2003). RNA polymerase II-dependent transcription in trypanosomes is associated with a SNAP complex-like transcription factor. *Proc.Natl.Acad.Sci.U.S.A* **100**, 80-85.

de Lange, T. and Borst, P. (1982). Genomic environment of the expression-linked extra copies of genes for surface antigens of *Trypanosoma brucei* resembles the end of a chromosome. *Nature* **299**, 451-453.

Deitsch, K. W., Moxon, E. R. and Wellems, T. E. (1997). Shared themes of antigenic variation and virulence in bacterial, protozoal, and fungal infections. *Microbiology and Molecular Biology Reviews* **61**, 281-294.

Devaux, S., Lecordier, L., Uzureau, P., Walgraffe, D., Dierick, J. F., Poelvoorde, P., Pays, E. and Vanhamme, L. (2006). Characterization of RNA polymerase II subunits of *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* **148**, 60-68

DiPaolo, C., Kieft, R., Cross, M. and Sabatini, R. (2005). Regulation of trypanosome DNA glycosylation by a SWI2/SNF2-like protein. *Mol. Cell.* **17**, 441-451.

Duggan, A. J. (1970). *An Historical Perspective.* In 'The African Trypanosomiases'. (Ed.H. W. Mulligan.) (George Allen & Unwin ltd: London.)

El Sayed, N. M., Ghedin, E., Song, J., MacLeod, A., Bringaud, F., Larkin, C., Wanless, D., Peterson, J., Hou, L., Taylor, S., Tweedie, A., Biteau, N., Khalak, H. G., Lin, X., Mason, T., Hannick, L., Caler, E., Blandin, G., Bartholomeu, D., Simpson, A. J., Kaul, S., Zhao, H., Pai, G., Van Aken, S., Utterback, T., Haas, B., Koo, H. L., Umayam, L., Suh, B., Gerrard, C., Leech, V., Qi, R., Zhou, S., Schwartz, D., Feldblyum, T., Salzberg, S., Tait, A., Turner, C. M., Ullu, E., White, O., Melville, S., Adams, M. D., Fraser, C. M. and Donelson, J. E. (2003). The sequence and analysis of *Trypanosoma brucei* chromosome II. *Nucleic Acids Research* **31**, 4856-4863.

El Sayed, N. M., Hegde, P., Quackenbush, J., Melville, S. E. and Donelson, J. E. (2000). The African trypanosome genome. *International Journal for Parasitology* **30**, 329-345.

Engstler, M., Thilo, L., Weise, F., Grunfelder, C. G., Schwarz, H., Boshart, M. and Overath, P. (2004). Kinetics of endocytosis and recycling of the GPI-anchored variant surface glycoprotein in *Trypanosoma brucei*. *Journal of Cell Science* **117**, 1105-1115.

Ersfeld, K., Melville, S. E. and Gull, K. (1999). Nuclear and genome organization of *Trypanosoma brucei*. *Parasitology Today* **15**, 58-63.

Eshita, Y., Urakawa, T., Hirumi, H., Fish, W. R. and Majiwa, P. A. (1992). Metacyclic form-specific variable surface glycoprotein-encoding genes of *Trypanosoma (Nannomonas) congolense*. *Gene* **113**, 139-148.

Esser, K. M. and Schoenbechler, M. J. (1985). Expression of two variant surface glycoproteins on individual African trypanosomes during antigen switching. *Science* **229**, 190-193.

Ferguson, M. A. J., Duszenko, M., Lamont, G. S., Overath, P. and Cross, G. A. M. (1986). Biosynthesis of *Trypanosoma brucei* variant surface glycoproteins n- glycosylation and addition of a phosphatidylinositol membrane anchor. *Journal of Biological Chemistry* **261**, 356-362.

Field, M. C. and Boothroyd, J. C. (1996). Sequence divergence in a family of variant surface glycoprotein genes from trypanosomes - coding region hypervariability and downstream recombinogenic repeats. *Journal of Molecular Evolution* **42**, 500-511.

Florent, I., Baltz, T., Raibaud, A. and Eisen, H. (1987). On the role of repeated sequences 5' to variant surface glycoprotein genes in African trypanosomes. *Gene* **53**, 55-62.

Frank, A. C. and Lobry, J. R. (1999). Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene* **238**, 65-77.

Frank, S. A. (1999). A model for the sequential dominance of antigenic variants in African trypanosome infections. *Proceedings Of The Royal Society Of London Series B-Biological Sciences* **266**, 1397-1401.

Frank, S. A. and Barbour, A. G. (2006). Within-host dynamics of antigenic variation. *Infect.Genet.Evol.* **6**, 141-146.

Freymann, D., Down, J., Carrington, M., Roditi, I., Turner, M. and Wiley, D. (1990). 2.9 Å resolution structure of the N-terminal domain of a variant surface glycoprotein from *Trypanosoma brucei*. *Journal of Molecular Biology* **216**, 141-160.

Futse, J. E., Brayton, K. A., Knowles, D. P. and Palmer, G. H. (2005). Structural basis for segmental gene conversion in generation of *Anaplasma marginale* outer membrane protein variants. *Mol. Microbiol.* **57**, 212-221.

Gardiner, P. R., Nene, V., Barry, M. M., Thatthi, R., Burleigh, B. and Clarke, M. W. (1996). Characterization of a small variable surface glycoprotein from *Trypanosoma vivax*. *Mol. Biochem. Parasitol.* **82**, 1-11.

Ghedin, E., Bringaud, F., Peterson, J., Myler, P., Berriman, M., Ivens, A., Andersson, B., Bontempi, E., Eisen, J., Angiuoli, S., Wanless, D., Von Arx, A., Murphy, L., Lennard, N., Salzberg, S., Adams, M. D., White, O., Hall, N., Stuart, K., Fraser, C. M. and El Sayed, N. M. (2004). Gene synteny and evolution of genome architecture in trypanosomatids. *Mol. Biochem.Parasitol.* **134**, 183-191.

Gibson, W. and Bailey, M. (2003). The development of *Trypanosoma brucei* within the tsetse fly midgut observed using green fluorescent trypanosomes. *Kinetoplastid.Biol.Dis.* **2**, 1.

Gibson, W. and Stevens, J. (1999). Genetic exchange in the trypanosomatidae. *Advances In Parasitology* **43**, 1-45.

Gottesdiener, K., Garciaanoveros, J., Lee, M. G. and Van der Ploeg, L. H. T. (1990). Chromosome organization of the protozoan *Trypanosoma brucei*. *Molecular and Cellular Biology* **10**, 6079-6083.

Graham, V. S. and Barry, J. D. (1996). Is point mutagenesis a mechanism for antigenic variation in *Trypanosoma brucei? Mol. Biochem. Parasitol.* **79**, 35-45.

Gray, A. R. (1965). Antigenic variation in a strain of *Trypanosoma brucei* transmitted by *Glossina morsitans* and *G.palpalis*. *Journal of General Microbiology* **41**, 195-214.

Gruszynski, A. E., van Deursen, F. J., Albareda, M. C., Best, A., Chaudhary, K., Cliffe, L. J., Del Rio, L., Dunn, J. D., Ellis, L., Evans, K. J., Figueiredo, J. M., Malmquist, N. A., Omosun, Y., Palenchar, J. B., Prickett, S., Punkosdy, G. A., van Dooren, G., Wang, Q., Menon, A. K., Matthews, K. R. and Bangs, J. D. (2006). Regulation of surface coat exchange by differentiating African trypanosomes. *Mol. Biochem. Parasitol.* **147**, 211-223.

Guglietta, S., Garbuglia, A. R., Pacciani, V., Scotta, C., Perrone, M. P., Laurenti, L., Spada, E., Mele, A., Capobianchi, M. R., Taliani, G., Folgori, A., Vitelli, A., Ruggeri, L., Nicosia, A., Piccolella, E. and Del Porto, P. (2005). Positive selection of cytotoxic T lymphocyte escape variants during acute hepatitis C virus infection. *European Journal of Immunology* **35**, 2627-2637.

Gunzl, A., Bruderer, T., Laufer, G., Schimanski, B., Tu, L. C., Chung, H. M., Lee, P. T. and Lee, M. G. (2003). RNA polymerase I transcribes procyclin genes and variant surface glycoprotein gene expression sites in *Trypanosoma brucei*. *Eukaryot.Cell* **2**, 542-551.

Guyaux, M., Cornelissen, A. W. C. A., Pays, E., Steinert, M. and Borst, P. (1985). *Trypanosoma brucei* - a surface antigen messenger RNA is discontinuously transcribed from 2 distinct chromosomes. *The EMBO Journal* **4**, 995-998.

Hajduk, S. L., Cameron, C. R., Barry, J. D. and Vickerman, K. (1981). Antigenic variation in cyclically transmitted *Trypanosoma brucei*. Variable antigen type composition of metacyclic trypanosome populations from the salivary glands of *Glossina morsitans*. *Parasitology* **83**, 595-607.

Hirumi, H. and Hirumi, K. (1989). Continuous cultivation of *Trypanosoma brucei* bloodstream forms in a medium containing a low concentration of serum protein without feeder cell layers. *Journal of Parasitology* **75**, 985-989.

Hoare, C. A. (1970). The Mammalian Trypanosomes of Africa. In 'The African Trypanosomiases'. (Ed.H. W. Mulligan.) (George Allen & Unwin ltd: London.)

Hoek, M., Zanders, T. and Cross, G. A. M. (2002). *Trypanosoma brucei* expression-site-associated-gene-8 protein interacts with a Pumilio family protein. *Mol. Biochem. Parasitol.* **120**, 269-283.

Hope, M., McLeod, A., Leech, V., Melville, S., Sasse, J., Tait, A. and Turner, C. M. R. (1999). Analysis of ploidy (in megabse chromosomes) in *Trypanosoma brucei* after genetic exchange. *Mol. Biochem. Parasitol.* **104**, 1-9.

Horn, D. (2001). Nuclear gene transcription and chromatin in *Trypanosoma brucei. Int.J.Parasitol* **31**, 1157-1165.

Horn, D. and Barry, J. D. (2005). The central roles of telomeres and subtelomeres in antigenic variation in African trypanosomes. *Chromosome Res.* **13**, 525-533.

Horn, D. and Cross, G. A. M. (1997). Position-dependent and promoter-specific regulation of gene expression in *Trypanosoma brucei. The EMBO Journal* **16**, 7422-7431.

Hou, W. R., Wang, H. F. and Niu, D. K. (2006). Replication-associated strand asymmetries in vertebrate genomes and implications for replicon size, DNA replication origin, and termination. *Biochemical and Biophysical Research Communications* **344**, 1258-1262.

Hutchinson, O. C., Smith, W., Jones, N. G., Chattopadhyay, A., Welburn, S. C. and Carrington, M. (2003). VSG structure: similar N-terminal domains can form functional VSGs with different types of C-terminal domain. *Mol. Biochem. Parasitol.* **130**, 127-131.

Johnson, J. G. and Cross, G. A. M. (1979). Selective cleavage of variant surface glycoproteins from *Trypanosoma brucei. Biochemical Journal* **178**, 689-697.

Johnson, P. J., Kooter, J. M. and Borst, P. (1987). Inactivation of transcription by UV-irradiation of *T. brucei* provides evidence for a multicistronic transcription unit including a VSG gene. *Cell* **51**, 273-281.

Kamper, S. M. and Barbet, A. F. (1992). Surface epitope variation via mosaic gene formation is potential key to long-term survival of *Trypanosoma brucei. Mol. Biochem.Parasitol.* **53**, 33-44.

Kazazian, H. H. (2004). Mobile elements: drivers of genome evolution. *Science* **303**, 1626-1632.

Keely, S. P., Renauld, H., Wakefield, A. E., Cushion, M. T., Smulian, A. G., Fosker, N., Fraser, A., Harris, D., Murphy, L., Price, C., Quail, M. A., Seeger, K., Sharp, S., Tindal, C. J., Warren, T., Zuiderwijk, E., Barrell, B. G., Stringer, J. R. and Hall, N. (2005). Gene arrays at *Pneumocystis carinii* telomeres. *Genetics* **170**, 1589-1600.

Kim, K. S. and Donelson, J. E. (1997). Co-duplication of a variant surface glycoprotein gene and its promoter to an expression site in African trypanosomes. *J.Biol.Chem.* **272**, 24637-24645.

Kimmel, B. E., Ole Moi Yoi, O. and Young, J. R. (1987). *Ingi*, a 5.2-kb dispersed sequence element from *Trypanosoma brucei* that carries half of a smaller mobile element at either end and has homology with mammalian LINES. *Molecular and Cellular Biology* **7**, 1465-1475.

Koenig-Martin, E., Yamage, M. and Roditi, I. (1992). A procyclin-associated gene in *Trypanosoma brucei* encodes a polypeptide related to ESAG 6 and 7 proteins. *Mol. Biochem. Parasitol.* **55**, 135-145.

Kooter, J. M. and Borst, P. (1984). Alpha amanitin insensitive transcription of variant surface glycoprotein genes provides further evidence for discontinuous transcription in trypanosomes. *Nucleic Acids Research* **12**, 9457-9472.

Kooter, J. M., Van der Spek, H. J., Wagter, R., d'Oliveira, C. E., Van der Hoeven, F., Johnson, P. J. and Borst, P. (1987). The anatomy and transcription of a telomeric expression site for variant specific surface antigens in *T. brucei. Cell* **51**, 261-272.

Kooter, J. M., Winter, A. J., Doliveira, C., Wagter, R. and Borst, P. (1988). Boundaries of telomere conversion in *Trypanosoma brucei. Gene* **69**, 1-11.

Kraemer, S. M. and Smith, J. D. (2006). A family affair: *var* genes, PfEMP1 binding, and malaria disease. *Curr.Opin.Microbiol.* **9**, 374-380.

LaCount, D. J., El Sayed, N. M., Kaul, S., Wanless, D., Turner, C. M. and Donelson, J. E. (2001). Analysis of a donor gene region for a variant surface glycoprotein and its expression site in African trypanosomes. *Nucleic Acids Res.* **29**, 2012-2019.

Laurent, M., Pays, E., Van der Werf, A., Aerts, D., Magnus, E., Van Meirvenne, N. and Steinert, M. (1984). Translocation alters the activation rate of a trypanosome surface antigen gene. *Nucleic Acids Research* **12**, 8319-8328.

Lee, M. G. and Van der Ploeg, L. H. T. (1987). Frequent independent duplicative transpositions activate a single *VSG* gene. *Molecular and Cellular Biology* **7**, 357-364.

Liniger, M., Urwyler, S., Studer, E., Oberle, M., Renggli, C. K. and Roditi, I. (2004). Role of the N-terminal domains of EP and GPEET procyclins in membrane targeting and the establishment of midgut infections by *Trypanosoma brucei. Mol. Biochem.Parasitol.* **137**, 247-251.

Liu, A. Y. C., Michels, P. A. M., Bernards, A. and Borst, P. (1985). Trypanosome variant surface glycoprotein genes expressed early in infection. *Journal of Molecular Biology* **182**, 383-396.

Liu, A. Y. C., Van der Ploeg, L. H. T., Rijsewijk, F. A. M. and Borst, P. (1983). The transposition unit of variant surface glycoprotein gene 118 of *Trypanosoma brucei* - presence of repeated elements at its border and absence of promoter-associated sequences. *Journal of Molecular Biology* **167**, 57-75.

Lowell, J. E. and Cross, G. A. (2004). A variant histone H3 is enriched at telomeres in *Trypanosoma brucei. Journal of Cell Science* **117**, 5937-5947.

Lu, Y., Alarcon, C. M., Hall, T., Reddy, L. V. and Donelson, J. E. (1994). A strand bias occurs in point mutations associated with variant surface glycoprotein gene conversion in *Trypanosoma rhodesiense. Mol. Cell. Biol.* **14**, 3971-3980.

Lu, Y., Hall, T., Gay, L. S. and Donelson, J. E. (1993). Point mutations are associated with a gene duplication leading to the bloodstream reexpression of a trypanosome metacyclic *VSG. Cell* **72**, 397-406.

Lukes, J., Hashimi, H. and Zikova, A. (2005). Unexplained complexity of the mitochondrial genome and transcriptome in kinetoplastid flagellates. *Current Genetics* **48**, 277-299.

Lythgoe, K., Morrison, L. J., Read, A. F. and Barry, J. D. (2006). A mathematical model based on empirical data suggests that order in trypanosome antigenic variation is determined by the parasite. *submitted.*

Majumder, H. K., Boothroyd, J. C. and Weber, H. (1981). Homologous 3'-terminal regions of mRNAs for surface antigens of different antigenic variants of *Trypanosoma brucei. Nucleic Acids Research* **9**, 4745-4753.

Masterson, W. J., Taylor, D. and Turner, M. J. (1988). Topologic analysis of the epitopes of a variant surface glycoprotein of *Trypanosoma brucei. Journal Of Immunology* **140**, 3194-3199.

Matthews, K. R., Ellis, J. R. and Paterou, A. (2004). Molecular regulation of the life cycle of African trypanosomes. *Trends in Parasitology* **20**, 40-47.

Matthews, K. R., Shiels, P. G., Graham, S. V., Cowan, C. and Barry, J. D. (1990). Duplicative activation mechanisms of two trypanosome telomeric *VSG* genes with structurally simple 5' flanks. *Nucleic Acids Research* **18**, 7219-7227.

Matthyssens, G., Michiels, F., Hamers, R., Pays, E. and Steinert, M. (1981). Two variant surface glycoproteins of *Trypanosoma brucei* have a conserved C-terminus. *Nature* **293**, 230-233.

McCulloch, R. and Barry, J. D. (1999). A role for RAD51 and homologous recombination in *Trypanosoma brucei* antigenic variation. *Genes & Development* **13**, 2875-2888.

McCulloch, R., Rudenko, G. and Borst, P. (1997). Gene conversions mediating antigenic variation in *Trypanosoma brucei* can occur in variant surface glycoprotein expression sites lacking 70 base-pair repeat sequences. *Molecular and Cellular Biology* **17**, 833-843.

McKenzie, G. J. and Rosenberg, S. M. (2001). Adaptive mutations, mutator DNA polymerases and genetic change strategies of pathogens. *Curr.Opin.Microbiol.* **4**, 586-594.

Mehlert, A., Bond, C. S. and Ferguson, M. A. (2002). The glycoforms of a *Trypanosoma brucei* variant surface glycoprotein and molecular modeling of a glycosylated surface coat. *Glycobiology* **12**, 607-612.

Melville, S. E. (1997). Genome research in *Trypanosoma brucei*: Chromosome size polymorphism and its relevance to genome mapping and analysis. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **91**, 116-120.

Melville, S. E., Gerrard, C. S. and Blackwell, J. M. (1999). Multiple causes of size variation in the diploid megabase chromosomes of African trypanosomes. *Chromosome Res.* **7**, 191-203.

Melville, S. E., Leech, V., Gerrard, C. S., Tait, A. and Blackwell, J. M. (1998). The molecular karyotype of the megabase chromosomes of *Trypanosoma brucei* and the assignment of chromosome markers. *Mol. Biochem. Parasitol.* **94**, 155-173.

Michels, P. A. M., Liu, A. Y. C., Bernards, A., Sloof, P., Vanderbijl, M. M. W., Schinkel, A. H., Menke, H. H., Borst, P., Veeneman, G. H., Tromp, M. C. and Vanboom, J. H. (1983). Activation of the genes for variant surface glycoprotein-117 and glycoprotein-118 in *Trypanosoma brucei*. *Journal of Molecular Biology* **166**, 537-556.

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

Miller, E. N., Allan, L. M. and Turner, M. J. (1984a). Mapping of antigenic determinants within peptides of a variant surface glycoprotein of *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* **13**, 309-322.

Miller, E. N., Allan, L. M. and Turner, M. J. (1984b). Topological analysis of antigenic determinants on a variant surface glycoprotein of *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* **13**, 67-81.

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

Milner, J. D. and Hajduk, S. L. (1999). Expression and localization of serum resistance associated protein in *Trypanosoma brucei rhodesiense*. *Mol. Biochem. Parasitol.* **104**, 271-283.

Morgan, R. W., Elsayed, N. M. A., Kepa, J. K., Pedram, M. and Donelson, J. E. (1996). Differential expression of the Expression Site Associated Gene I family in African trypanosomes. *Journal of Biological Chemistry* **271**, 9771-9777.

Morrison, L. J., Majiwa, P., Read, A. F. and Barry, J. D. (2005). Probabilistic order in antigenic variation of *Trypanosoma brucei*. *Int. J. Parasitol.* **35**, 961-972.

Morrison, W. I., Black, S. J., Paris, J., Hinson, C. A. and Wells, P. W. (1982). Protective immunity and specificity of antibody responses elicited in cattle by irradiated *Trypanosoma brucei*. *Parasite Immunology* **4**, 395-407.

Nash, T. E. (2002). Surface antigenic variation in *Giardia lamblia*. *Mo.l Microbiol.* **45**, 585-590.

Navarro, M., Cross, G. A. M. and Wirtz, E. (1999). *Trypanosoma brucei* variant surface glycoprotein regulation involves coupled activation/inactivation and chromatin remodeling of expression sites. *The EMBO Journal* **18**, 2265-2272.

Navarro, M. and Gull, K. (2001). A pol I transcriptional body associated with *VSG* monoallelic expression in *Trypanosoma brucei*. *Nature* **414**, 759-763.

Nilsson, D., Andersson, B. (2005). Strand asymmetry in trypanosomatid parasites. *Exp. Parasitol.* **109**, 143-149

Niu, D. K., Lin, K. and Zhang, D. Y. (2003). Strand compositional asymmetries of nuclear DNA in eukaryotes. *Journal of Molecular Evolution* **57**, 325-334.

OBeirne, C., Lowry, C. M. and Voorheis, H. P. (1998). Both IgM and IgG anti-VSG antibodies initiate a cycle of aggregation- disaggregation of bloodstream forms of *Trypanosoma brucei* without damage to the parasite. *Mol. Biochem. Parasitol.* **91**, 165-193.

Ohshima, K., Kang, S., Larson, J. E. and Wells, R. D. (1996). TTA.TAA triplet repeats in plasmids form a non-H bonded structure. *Journal of Biological Chemistry* **271**, 16784-16791.

Oli, M. W., Cotlin, L. F., Shiflett, A. M. and Hajduk, S. L. (2006). Serum resistance-associated protein blocks lysosomal targeting of trypanosome lytic factor in *Trypanosoma brucei*. *Eukaryot.Cell* **5**, 132-139.

Pal, A., Hall, B. S., Jeffries, T. R. and Field, M. C. (2003). Rab5 and Rab11 mediate transferrin and anti-variant surface glycoprotein antibody recycling in *Trypanosoma brucei*. *Biochemical Journal* **374**, 443-451.

Palenchar, J. B., Liu, W., Palenchar, P. M. and Bellofatto, V. (2006). A divergent transcription factor TFIIB in trypanosomes is required for RNA polymerase II-dependent spliced leader RNA transcription and cell viability. *Eukaryot.Cell* **5**, 293-300.

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

Pays, E., Houard, S., Pays, A., Van Assel, S., Dupont, F., Aerts, D., Huetduvillier, G., Gomes, V., Richet, C., Degand, P., Van Meirvenne, N. and Steinert, M. (1985). *Trypanosoma brucei* - the extent of conversion in antigen genes may be related to the DNA coding specificity. *Cell* **42**, 821-829.

Pays, E., Van Assel, S., Laurent, M., Darville, M., Vervoort, T., Van Meirvenne, N. and Steinert, M. (1983b). Gene conversion as a mechanism for antigenic variation in trypanosomes. *Cell* **34**, 371-381.

Pedram, M. and Donelson, J. E. (1999). The anatomy and transcription of a monocistronic expression site for a metacyclic variant surface glycoprotein gene in *Trypanosoma brucei*. *Journal of Biological Chemistry* **274**, 16876-16883.

Penchenier, L., Alhadji, D., Bahebegue, S., Simo, G., Laveissiere, C. and Cuny, G. (2005). Spontaneous cure of domestic pigs experimentally infected by *Trypanosoma brucei gambiense*. Implications for the control of sleeping sickness. *Veterinary Parasitology* **133**, 7-11.

Perez-Morga, D., Amiguet-Vercher, A., Vermijlen, D. and Pays, E. (2001). Organization of telomeres during the cell and life cycles of *Trypanosoma brucei*. *Journal of Eukaryotic Microbiology* **48**, 221-226.

Pond, S. L. and Frost, S. D. (2005). Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics.* **21**, 2531-2533.

Raibaud, A., Gaillard, C., Longacre, S., Hibner, U., Buck, G., Bernardi, G. and Eisen, H. (1983). Genomic environment of variant surface antigen genes of *Trypanosoma equiperdum*. *Proc.Natl.Acad.Sci.U.S.A* **80**, 4306-4310.

Rausch, S., Shayan, P., Salnikoff, J. and Reinwald, E. (1994). Sequence determination of three variable surface glycoproteins from *Trypanosoma congolense*. Conserved sequence and structural motifs. *European Journal of Biochemistry* **223**, 813-821.

Reddy, L. V., Hall, T. and Donelson, J. E. (1990). Sequences of 3 *VSG* messenger-RNAs expressed in a mixed population of *Trypanosoma brucei-rhodesiense*. *Biochemical and Biophysical Research Communications* **169**, 730-736.

Redpath, M. B., Windle, H., Nolan, D., Pays, E., Voorheis, H. P. and Carrington, M. (2000). *ESAG11*, a new *VSG* expression site-associated gene from *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* **111**, 223-228.

Reinitz, D. M., Aizenstein, B. D. and Mansfield, J. M. (1992). Variable and conserved structural elements of trypanosome variant surface glycoproteins. *Mol. Biochem. Parasitol.* **51**, 119-132.

Restrepo, B. I. and Barbour, A. G. (1994). Antigen diversity in the bacterium *B. hermsii* through "somatic" mutations in rearranged *vmp* genes. *Cell* **78**, 867-876.

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

Robinson, N. P., Burman, N., Melville, S. E. and Barry, J. D. (1999). Predominance of duplicative *VSG* gene conversion in antigenic variation in African trypanosomes. *Molecular And Cellular Biology* **19**, 5839-5846.

Rocha, E. P. (2004). The replication-related organization of bacterial genomes. *Microbiology* **150**, 1609-1627.

Roditi, I. (1996). The VSG-procyclin switch. *Parasitology Today* **12**, 47-49.

Rosenberg, S. M. (2001). Evolving responsively: adaptive mutation. *Nature Reviews Genetics* **2**, 504-515.

Roth, C., Bringaud, F., Layden, R. E., Baltz, T. and Eisen, H. (1989). Active late-appearing variable surface antigen genes in *Trypanosoma equiperdum* are constructed entirely from pseudogenes. *Proc.Natl.Acad.Sci.U.S.A* **86**, 9375-9379.

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

Rudenko, G., Blundell, P. A., Dirksmulder, A., Kieft, R. and Borst, P. (1995). A ribosomal DNA promoter replacing the promoter of a telomeric *VSG* gene expression site can be efficiently switched on and off in *Trypanosoma brucei*. *Cell* **83**, 547-553.

Rudenko, G., McCulloch, R., Dirksmulder, A. and Borst, P. (1996). Telomere exchange can be an important mechanism of variant surface glycoprotein gene switching in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* **80**, 65-75.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A. and Barrell, B. (2000). Artemis: sequence visualization and annotation. *Bioinformatics.* **16**, 944-945.

Salmon, D., Hanocq-Quertier, J., Paturiaux-Hanocq, F., Pays, A., Tebabi, P., Nolan, D. P., Michel, A. and Pays, E. (1997). Characterization of the ligand-binding site of the transferrin receptor in *Trypanosoma brucei* demonstrates a structural relationship with the N-terminal domain of the variant surface glycoprotein. *The EMBO Journal* **16**, 7272-7278.

Salmon, D., Paturiaux-Hanocq, F., Poelvoorde, P., Vanhamme, L. and Pays, E. (2005). *Trypanosoma brucei*: growth differences in different mammalian sera are not due to the species-specificity of transferrin. *Exp.Parasitol.* **109**, 188-194.

Samaranayake, M., Bujnicki, J.M., Carpenter, M., Bhagwat, A.S. (2006) Evaluation of molecular models for the affinity maturation of antibodies: roles of cytosine deamination by AID and DNA repair. *Chem. Rev.* **106**, 700-719.

Schopf, L. R. and Mansfield, J. M. (1998). Characterization of a relatively rare class B, type 2 trypanosome variant surface glycoprotein gene. *Journal of Parasitology* **84**, 284-292.

Schweizer, J., Tait, A. and Jenni, L. (1988). The timing and frequency of hybrid formation in African trypanosomes during cyclical transmission. *Parasitology Research* **75**, 98-101.

Sechman, E. V., Rohrer, M. S. and Seifert, H. S. (2005). A genetic screen identifies genes and sites involved in pilin antigenic variation in *Neisseria gonorrhoeae*. *Molecular Microbiology* **57**, 468-483.

Seed, J. R. (1978). Competition among serologically different clones of *Trypanosoma brucei gambiense in vivo*. *J. Protozool.* **25**, 526-529.

Seyfang, A., Mecke, D. and Duszenko, M. (1990). Degradation, recycling, and shedding of *Trypanosoma brucei* variant surface glycoprotein. *J.Protozool.* **37**, 546-552.

Shafikhani, S. (2002). Factors affecting PCR-mediated recombination. *Environ.Microbiol.* **4**, 482-486.

Shah, J. S., Young, J. R., Kimmel, B. E., Iams, K. P. and Williams, R. O. (1987). The 5' flanking sequence of a *Trypanosoma brucei* variable surface glycoprotein gene. *Mol. Biochem. Parasitol.* **24**, 163-174.

Simpson, A. G., Stevens, J. R. and Lukes, J. (2006). The evolution and diversity of kinetoplastid flagellates. *Trends in Parasitology* **22**, 168-174.

Sinden, R. R., Hashem, V. I. and Rosche, W. A. (1999). DNA-directed mutations. Leading and lagging strand specificity. *Ann.N.Y.Acad.Sci.* **870**, 173-189.

Steverding, D., Stierhof, Y. D., Chaudhri, M., Ligtenberg, M., Schell, D., Becksickinger, A. G. and Overath, P. (1994). ESAG-6 AND ESAG-7 products of *Trypanosoma brucei* form a transferrin-binding protein complex. *European Journal of Cell Biology* **64**, 78-87.

Steverding, D., Stierhof, Y. D., Fuchs, H., Tauber, R. and Overath, P. (1995). Transferrin-binding protein complex is the receptor for transferrin uptake in *Trypanosoma brucei*. *Journal of Cell Biology* **131**, 1173-1182.

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 (1987). *Trypanosoma congolense*: structure and molecular organization of the surface glycoproteins of two early bloodstream variants. *Biochemistry* **26**, 796-805.

Strickler, J. E. and Patton, C. L. (1982). *Trypanosoma brucei*: nearest neighbor analysis on the major variable surface coat glycoprotein--crosslinking patterns with intact cells. *Experimental Parasitology* **53**, 117-132.

Sung, S. Y., McDowell, J. V. and Marconi, R. T. (2001). Evidence for the contribution of point mutations to vlsE variation and for apparent constraints on the net - Accumulation of sequence changes in vlsE during infection with Lyme disease spirochetes. *Journal of Bacteriology* **183**, 5855-5861.

Szczepanik, D., Mackiewicz, P., Kowalczuk, M., Gierlik, A., Nowicka, A., Dudek, M. R. and Cebrat, S. (2001). Evolution rates of genes on leading and lagging DNA strands. *Journal of Molecular Evolution* **52**, 426-433.

Tetley, L., Turner, C. M. R., Barry, J. D., Crowe, J. S. and Vickerman, K. (1987). Onset of expression of the variant surface glycoproteins of *Trypanosoma brucei* in the tsetse fly studied using immunoelectron microscopy. *Journal of Cell Science* **87**, 363-372.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. and Higgins, D. G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* **25**, 4876-4882.

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

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

Timmers, H. T. M., de Lange, T., Kooter, J. M. and Borst, P. (1987). Coincident multiple activations of the same surface antigen gene in *Trypanosoma brucei*. *Journal of Molecular Biology* **194**, 81-90.

Tippin, B., Kobayashi, S., Bertram, J. G. and Goodman, M. F. (2004). To slip or skip, visualizing frameshift mutation dynamics for error-prone DNA polymerases. *Journal of Biological Chemistry* **279**, 45360-45368.

Turner, C. M., McLellan, S., Lindergard, L. A., Bisoni, L., Tait, A. and MacLeod, A. (2004). Human infectivity trait in *Trypanosoma brucei*: stability, heritability and relationship to SRA expression. *Parasitology* **129**, 445-454.

Turner, C. M. R. (1997). The rate of antigenic variation in fly-transmitted and syringe- passaged infections of *Trypanosoma brucei*. *Fems Microbiology Letters* **153**, 227-231.

Turner, C. M. R. (1999). Antigenic variation in *Trypanosoma brucei* infections: an holistic view. *Journal of Cell Science* **112**, 3187-3192.

Turner, C. M. R., Aslam, N. and Dye, C. (1995). Replication, differentiation, growth and the virulence of *Trypanosoma brucei* infections. *Parasitology* **111**, 289-300.

Turner, C. M. R., Barry, J. D., Maudlin, I. and Vickerman, K. (1988). An estimate of the size of the metacyclic variable antigen repertoire of *Trypanosoma brucei rhodesiense*. *Parasitology* **97**, 269-276.

Ulbert, S., Chaves, I. and Borst, P. (2002). Expression site activation in *Trypanosoma brucei* with three marked variant surface glycoprotein gene expression sites. *Mol. Biochem. Parasitol.* **120**, 225-235.

Urakawa, T., Eshita, Y. and Majiwa, P. A. O. (1997). The primary structure of *Trypanosoma (Nannomonas) congolense* variant surface glycoproteins. *Experimental Parasitology* **85**, 215-224.

Urwyler, S., Vassella, E., Van Den Abbeele, J., Renggli, C. K., Blundell, P. A., Barry, J. D. and Roditi, I (2005). Expression of procyclin mRNAs during cyclical transmission of *Trypanosoma brucei. PLOS Pathogens* **1**, e22

Van der Ploeg, L. H. T., Bernards, A., Rijsewijk, F. A. M. and Borst, P. (1982a). Characterization of the DNA duplication-transposition that controls the expression of 2 genes for variant surface glycoproteins in *Trypanosoma brucei. Nucleic Acids Research* **10**, 593-609.

Van der Ploeg, L. H. T., Valerio, D., de Lange, T., Bernards, A., Borst, P. and Grosveld, F. G. (1982b). 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 Research* **10**, 5905-5923.

van Deursen, F. J., Shahi, S. K., Turner, C. M. R., Hartmann, C., Guerra-Giraldez, C., Matthews, K. R. and Clayton, C. E. (2001). Characterisation of the growth and differentiation *in vivo* and *in vitro* of bloodstream form *Trypanosoma brucei* strain TREU 927. *Mol. Biochem. Parasitol.* **112**, 163-171.

van Leeuwen, F., Wijsman, E. R., Kieft, R., vanderMarel, G. A., Vanboom, J. H. and Borst, P. (1997). Localization of the modified base J in telomeric VSG gene expression sites of *Trypanosoma brucei. Genes Dev.* **11**, 3232-3241.

Vanhamme, L., Poelvoorde, P., Pays, A., Tebabi, P., Xong, H. V. and Pays, E. (2000). Differential RNA elongation controls the variant surface glycoprotein gene expression sites of *Trypanosoma brucei. Molecular Microbiology* **36**, 328-340.

Vanhamme, L., Postiaux, S., Poelvoorde, P. and Pays, E. (1999). Differential regulation of *ESAG* transcripts in *Trypanosoma brucei. Mol. Biochem. Parasitol.* **102**, 35-42.

Vickerman, K. (1969). On the surface coat and flagellar adhesion in trypanosomes. *Journal of Cell Science* **5**, 163-194.

Vogt, G., Etzold, T. and Argos, P. (1995). An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. *Journal of Molecular Biology* **249**, 816-831.

Vossen, M. T., Westerhout, E. M., Soderberg-Naucler, C. and Wiertz, E. J. (2002). Viral immune evasion: a masterpiece of evolution. *Immunogenetics* **54**, 527-542.

Wang, J., Bohme, U. and Cross, G. A. (2003). Structural features affecting variant surface glycoprotein expression in *Trypanosoma brucei. Mol Biochem.Parasitol.* **128**, 135-145.

Wickstead, B., Ersfeld, K. and Gull, K. (2004). The small chromosomes of *Trypanosoma brucei* involved in antigenic variation are constructed around repetitive palindromes. *Genome Research* **14**, 1014-1024.

Xong, H. V., Vanhamme, L., Chamekh, M., Chimfwembe, C. E., Van Den Abbeele, J., Pays, A., Van Meirvenne, N., Hamers, R., Debaetselier, P. and Pays, E. (1998). A *VSG* expression site-associated gene confers resistance to human serum in *Trypanosoma rhodesiense. Cell* **95**, 839-846.

Zaphiropoulos, P. G. (1998). Non-homologous recombination mediated by *Thermus aquaticus* DNA polymerase I. Evidence supporting a copy choice mechanism. *Nucleic Acids Research* **26**, 2843-2848.

Zhang, J. R., Hardham, J. M., Barbour, A. G. and Norris, S. J. (1997). Antigenic variation in Lyme disease borreliae by promiscuous recombination of VMP-like sequence cassettes. *Cell* **89**, 275-285.

Zhang, J. R. and Norris, S. J. (1998). Kinetics and in vivo induction of genetic variation of vlsE in *Borrelia burgdorferi. Infection and Immunity* **66**, 3689-3697.

Zhang, Q. Y., DeRyckere, D., Lauer, P. and Koomey, M. (1992). Gene conversion in *Neisseria gonorrhoeae*: evidence for its role in pilus antigenic variation. *Proc.Natl.Acad.Sci.U.S.A* **89**, 5366-5370.

Ziegelbauer, K. and Overath, P. (1992). Identification of invariant surface glycoproteins in the bloodstream stage of *Trypanosoma brucei*. *Journal of Biological Chemistry* **267**, 10791-10796.

Ziegelbauer, K. and Overath, P. (1993). Organization of two invariant surface glycoproteins in the surface coat of *Trypanosoma brucei*. *Infection and Immunity* **61**, 4540-4545.

Zitzmann, N., Mehlert, A., Carroue, S., Rudd, P. M. and Ferguson, M. A. J. (2000). Protein structure controls the processing of the N-linked oligosaccharides and glycosylphosphatidylinositol glycans of variant surface glycoproteins expressed in bloodstream form *Trypanosoma brucei*. *Glycobiology* **10**, 243-249.

Zomerdijk, J. C. B. M., Ouellette, M., ten Asbroek, A. L. M. A., Kieft, R., Bommer, A. M. M., Clayton, C. E. and Borst, P. (1990). The promoter for a variant surface glycoprotein gene expression site in *Trypanosoma brucei*. *The EMBO Journal* **9**, 2791-2801.