Dissertation

submitted to the

Combined Faculty of Natural Sciences and Mathematics

Of the Ruperto Carola University Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

Presented by

M.Sc. Monica Chandra

Born in Surabaya, Indonesia

Oral examination: 15.06.2021

# Sampling the structure-space of metacyclic Variant Surface Glycoprotein (mVSG) of *Trypanosoma brucei* coat

Referees : Prof. Dr. Christine Clayton
Prof. Dr. Nina Papavasiliou

**Table of contents**

**Summary**

*Trypanosoma brucei (T. brucei)* is an extracellular unicellular pathogen, which is the causative agent of African sleeping sickness in sub-Saharan Africa. It has a two-host life cycle in its vector tsetse fly (*Glossina sp*.) and its mammalian hosts. The work of this thesis focuses on the metacyclic stage of the parasite, which is the mammalian infectious form injected by the fly, and the metacyclic Variant Surface Glycoprotein (mVSG). VSG is expressed in the metacyclic and bloodstream forms. The latter differentiate from metacyclic in the mammalian host and are capable of evading the host immune system by constantly changing the expression of its surface proteome makeup. Despite a vast number of VSG genes present in the genome repertoire (~2000 genes and pseudogenes), only a specific subset of five VSGs, called metacyclic VSGs (mVSGs), are expressed in the metacyclic stage of the Lister 427 strain. In this study, we are comparing mVSGs and bloodstream VSGs at the protein structure level. Using structure-based prediction by protein threading, we showed that VSGs in *T. brucei* are divided into two different structure classes, A and B. We have determined the structure of mVSG 531 and mVSG1954, classified in Class A and B, respectively, by X-ray crystallography. The homodimer structure of mVSG531's N-terminal domain shows the conservation of the three-helix bundle core and the dumbbell-like shape observed in the previously determined VSGs structures in Class A, VSG2 (MITat 1.2) and VSG1 (MITat 1.1), which are expressed in the bloodstream form. The structure of mVSG1954 also shows similarities to VSG3 (MITat 1.3), classified in Class B and expressed in the bloodstream form. Like VSG3, it exists as a monomer in the crystal asymmetric unit, but also and forms non-crystallographic trimers. While the VSG3 is *O*-glycosylated, there is no evidence of *O*-glycosylation in the mVSG1954 structure. The results address a long-standing question of whether the mVSGs were similar to or different from the bloodstream VSG coat proteins, showing that structurally and biochemically, the two life-stages classes of VSGs are very similar.

**Zusammenfassung**

*Trypanosoma brucei* (*T. brucei*) ist ein einzelliger Parasit und der Hauptverursacher der afrikanischen Schlafkrankheit. *T. brucei* wird von seinem Vektor, der Tsetsefliege (*Glossina sp*.) auf den Säugetierwirt übertragen. Während des metazyklischen Stadiums des Parasiten, welches die infektiöse Form für Säugetiere darstellt und von der Fliege durch die Blutmahlzeit injiziert wird, exprimiert *T. brucei* die metazyklische Form des Variant Surface Glycoprotein (VSG). VSG wird vom Parasiten sowohl in der metazyklischen als auch in der Blutstromform exprimiert, jedoch unterscheiden sich die beiden Formen. Die Blutsform VSG tragen, durch Veränderung der Oberflächenproteomzusammensetzung, maßgeblich zur Resistenz von *T. brucei* gegenüber dem Wirtsimmunsystems bei. Trotz einer großen Anzahl von VSG-Genen im Genom von *T. brucei* (~2000 Gene und Pseudogene) wird nur eine spezifische Untergruppe von fünf VSGs, genannt metazyklische VSGs (mVSGs), in der metazyklischen Form im Lister 427 Stamm exprimiert. In dieser Studie vergleichen wir mVSGs und Blutstrom-VSGs auf der Ebene der Proteinstruktur. Strukturbasierte Vorhersagen mit Hilfe von Protein-Threading zeigen, dass sich VSGs in *T. brucei* in zwei verschiedene Strukturklassen, A und B, einteilen lassen. Mit Hilfe von Röntgenkristallographie haben wir die Struktur von mVSG531 und mVSG1954, die in Klasse A bzw. B eingestuft sind, entschlüsselt. Ein Vergleich der Homodimerstruktur der N-terminalen Domäne von mVSG531 mit den bereits existierenden Proteinstrukturen von VSG2 und VSG1 (beide Blutsform Klasse A) zeigt, dass der charakteristische Drei-Helix-Bündelkern und die Hantelform von VSG2/1 in mVSG531 enthalten sind. Die Struktur von mVSG1954 ähnelt VSG3 (Blutsform Klasse B) insofern, dass es als Monomer in der kristallinen Asymmetrieeinheit vorliegt und kristallographische Trimere bildet. Während VSG3 O-glykosyliert ist, gibt es in der mVSG1954-Struktur keine Hinweise auf O-Glykosylierung. Diese Arbeit geht auf die seit langem bestehende Frage ein, ob die mVSGs den VSG-Hüllproteinen des Blutstroms ähnlich sind oder sich von ihnen unterscheiden. Die Ergebnisse zeigen, dass die beiden VSG-Klassen der Lebensstadien strukturell und biochemisch sehr ähnlich sind.

**List of abbreviations**

| | |
|---|---|
| BES | : Bloodstream expression sites |
| BSF | : Bloodstream form |
| BSF mVSG1954 | : BSF cells expressing mVSG1954 |
| BSF mVSG531 | : BSF cells expressing mVSG531 |
| BSF VSG2 | : BSF cells expressing VSG2 |
| CTD | : C-terminal domain |
| *E. coli* | : *Escherichia coli* |
| ESAG | : Expression Sites Associated Genes |
| gDNA | : genomic DNA |
| GDT | : global distance test |
| GPI | : Glycophosphatidylinositol |
| HAT | : Human African Trypanosomiasis |
| i.p. | : intraperitoneal |
| *L. donovani* | : *Leishmania donovani* |
| MAD | : Multi-angle anomalous dispersion |
| MCF | : Metacyclic form |
| MES | : Metacyclic expression sites |
| MIR | : Multiple Isomorphous Replacement |
| MIRAS | : Multiple Isomorphous Replacement with anomalous scattering |
| mVSGs | : VSGs expressed in metacyclic life stages of *T. brucei* |
| NTD | : N-terminal domain |
| PCF | : Procyclic form |
| PEG | : Polyethylene glycol |
| PMSF | : Phenylmethylsulfonylfluoride |
| RBP6 | : RNA Binding Protein 6 |
| RMSD / r.m.s.d. | : root mean square deviation |
| RNAP | : RNA Polymerase |
| SAD | : Single-wavelength anomalous dispersion |
| SIR | : Single Isomorphous Replacement |
| SIRAS | : Single Isomorphous Replacement with anomalous scattering |
| *T. brucei* | : Trypanosoma brucei |
| TM-score | : template modeling score |
| uGDT | : unnormalized global distance test |

VSG : Variant Surface Glycoprotein

WHO : World Health Organization

**List of Figures**

## List of Tables

## Chapter I: Introduction

### 1. *T. brucei* as the causative pathogen of African trypanosomiasis

*Trypanosoma brucei* (*T. brucei*) is a unicellular pathogen, the causal agent of Human African Trypanosomiasis (HAT) in humans and Nagana in cattle (Büscher et al., 2017). It is found mainly in sub-Saharan African countries (Figure 1). The parasite is transmitted to the mammalian host by tsetse flies (*Glossina sp.*), endemic to the sub-Saharan Africa region. The disease has been a significant economic burden in these relatively under-developed countries. Nagana causes both death and early termination of the pregnancy in cattle. Moreover, the surviving cattle have a decrease in milk and meat production, which causes an even more economic burden to the developing countries in the area (Swallow, 1999).

Two major subspecies of *T. brucei* are known to infect humans, which are *T. brucei gambiense* and *T. brucei rhodesiense* (Figure 1) The former is known to cause chronic trypanosomiasis, while the latter causes acute infection in humans (Büscher et al., 2017). Another subspecies, *T. brucei brucei*, is known to infect cattle and not humans under normal circumstances (and is the primary laboratory strain cultivated for biomedical research) (Pays et al., 2006). The first stage of Trypanosoma infection occurs in the subcutaneous tissues, blood, and lymph. As the disease progresses, the parasites will begin to migrate to the brain, passing the blood-brain barrier. In the later stage of infection, neurological symptoms start to appear, including deregulation of the sleep cycle, from which the disease takes its common name ("Trypanosomiasis, human African (sleeping sickness)," WHO.). Several drugs are available for trypanosomiasis treatment, which include suramin, pentamidine, eflornithine, nifirtimox, fexinidazole, and melarsoprol (Lindner et al., 2020; Steverding, 2010). However, these drugs cause adverse side effects, which makes these drugs unsuitable for prophylactic consumption (Barrett et al., 2007). The prognosis of the disease is better when detected as early as possible. No vaccine against HAT or Nagana is available to this date, making disease prevention dependent on other methods such as vector control. The continuous joint-measures between the affected countries and the World Health Organization have significantly decreased the number of trypanosomiasis cases (Franco et al., 2020). However, continuous research for better

prevention of the disease is needed to avoid future outbreaks and improve the people's life quality in the affected areas.

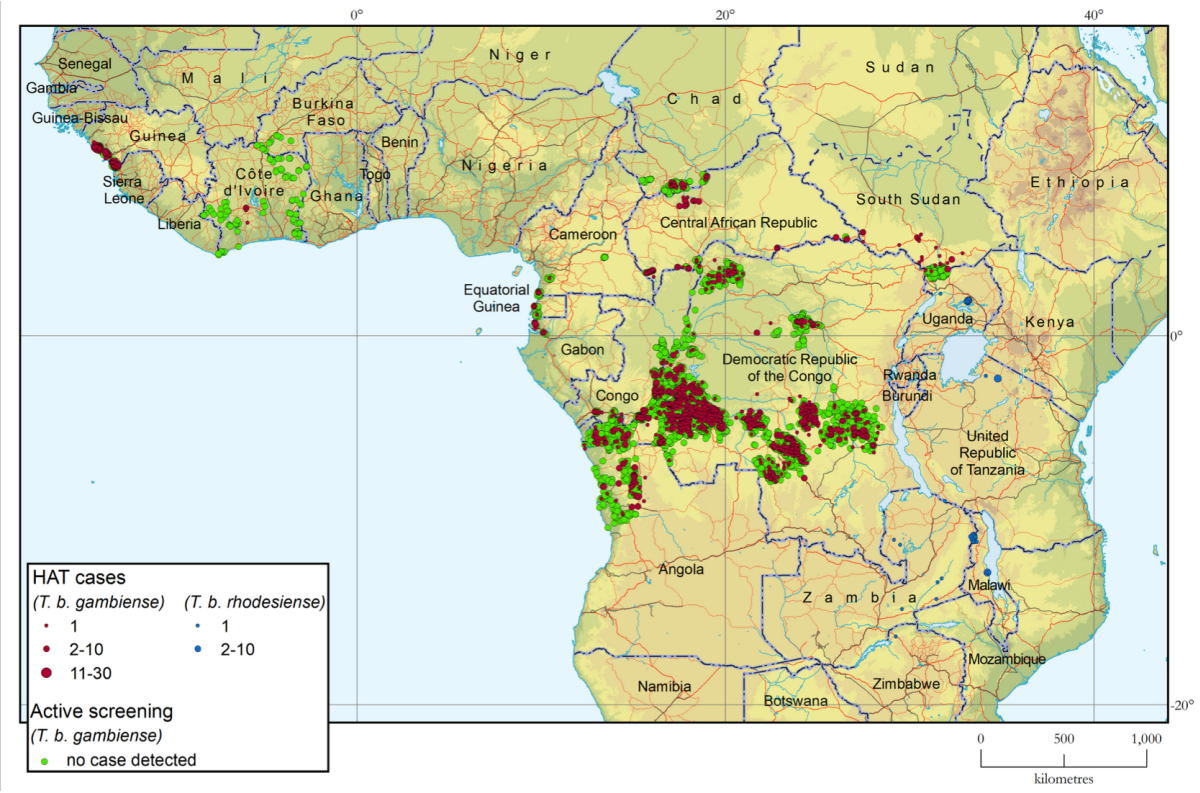


***Figure 1.* Geographic distribution of Human African Trypanosomiasis (HAT) in Sub-Saharan Africa in 2017 to 2018.**

Most of the HAT cases in the western part of Sub-Saharan Africa are caused by *T. brucei gambiense*, while most of the cases in the eastern part are dominated by *T. brucei rhodesiense*. The most affected areas are the Democratic Republic of the Congo and the Central African Republic. (Franco et al., 2020)

## 2. Life cycle of *T. brucei*

*T. brucei* has a complex life cycle in both its mammalian host and its insect vector, the tsetse fly (Figure 2). The procyclic form (PCF) of *T. brucei* inhabits the tsetse fly's midgut (Büscher et al., 2017). The PCF move toward the proventriculus and develop further into elongated epimastigotes. The elongated epimastigotes undergo asymmetric cell division, resulting in longer and shorter daughter cells. The smaller epimastigotes eventually attach themselves to the fly's salivary gland. These cells will eventually mature into the mammalian-infective metacyclic form (MCF) (Van Den Abbeele et al., 1999). MCF is cell-cycle arrested and have reduced transcription and

protein synthesis levels (Christiano et al., 2017). MCF colonizes the salivary gland of the tsetse fly and are injected into the mammalian host's bloodstream when the fly takes a blood meal (Figure 2) (Büscher et al., 2017).

Inside the mammalian host, MCF differentiates into the long, slender bloodstream form (BSF) (Figure 2). The BSF proliferates by binary fission and swims freely inside their mammalian host's bloodstream. To ensure the life cycle's continuity, BSF will eventually progress into the fly infective form, called stumpy form (Figure 2) (Büscher et al., 2017). The progression from BSF to stumpy form is in a cell density-dependent manner by quorum sensing (Zimmermann et al., 2017). The stumpy form is further taken up by the fly when it takes another blood meal (Figure 2) (Büscher et al., 2017). Inside the fly's midgut, the stumpy form undergoes many changes in their cell morphology, protein expression profiles, as well as the remodeling of the cell surface's proteome to become PCF once again (Dean et al., 2009; Gruszynski et al., 2006; Rico et al., 2013).

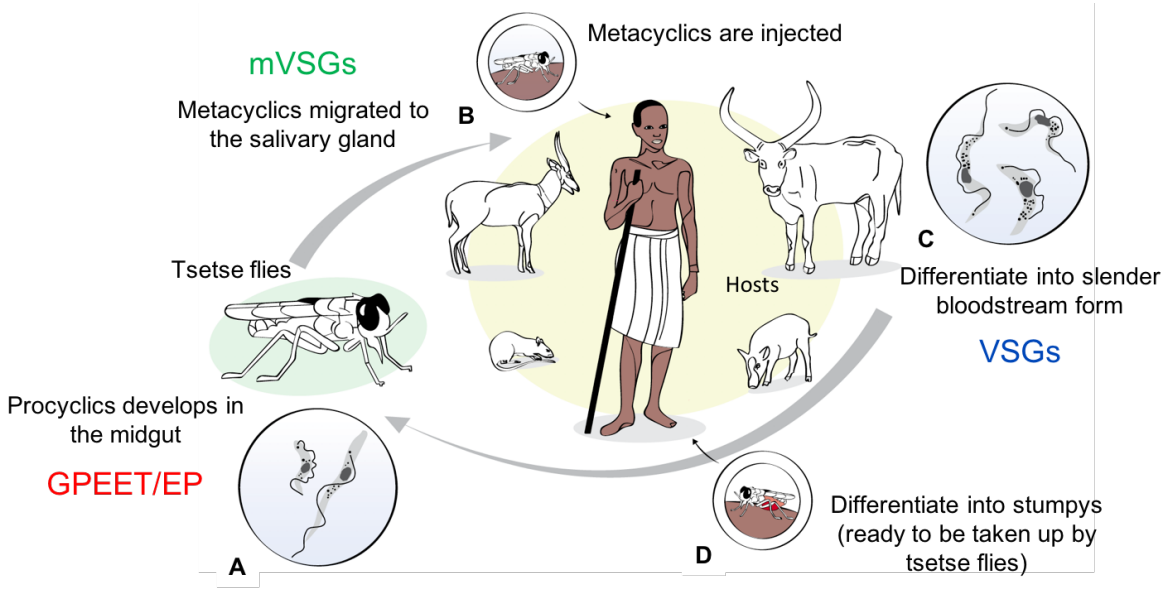


***Figure 2.*** ***T. brucei life cycle***

**A)** PCF cells, covered with GPEET and EP repeat proteins on their surfaces, inhabit and develop in the tsetse fly's midgut. **B)** PCF cells eventually migrate to the salivary gland and develop into MCF cells, which are covered by mVSGs. MCF cells are injected into the mammalian host when the fly takes a blood meal. **C)** MCF cells develop into long slender BSF cells, which are covered by VSGs. BSF cells are able to start the infection in the mammalian host and swim freely in the bloodstream. **D)** Some of the BSF cells differentiate further into stumpy form, which will be taken up by the tsetse fly to complete the life cycle and start the colonization of the midgut by PCF cells. (image adapted from: Büscher et al., 2017)

## 3. Surface proteome of *T. brucei* at different life stages

One of the landmarks that distinguishes one life stage from another is the surface proteome's remodeling. The infectivity of *T. brucei* in a mammalian host is marked by the expression of Variant Surface Glycoprotein (VSG) on its surface. VSGs are expressed in the infective forms MCF and BSF and are still retained in the stumpy form (Cross et al., 2014; Rico et al., 2013). VSG covers the surface of MCF and BSF densely (approximately ten million copies of the VSG protein on the surface of each cell) (Jackson et al., 1985). When the fly takes up the stumpy form in a bloodmeal, the stumpy form develops into PCF inside the fly's midgut. The stumpy form's progression to PCF is marked by the loss of VSG on their surface (Rico et al., 2013). Instead, procyclins cover the PCF (Richardson et al., 1988). Both VSGs and procyclins are inserted into the membrane of the *T. brucei* by a glycophosphoinositol anchor (GPI-

anchor). The GPI-anchor is located at the end of the polypeptide chain. The GPI-anchor that harbors the VSGs are susceptible to the cleavage by phospholipase C, while the one that harbors the procyclins are not. The susceptibility is caused by the acetylation of the inositol moiety in the GPI-anchor attached to the procyclins (Field et al., 1991).

BSF cells' surfaces are carpeted by VSGs (Cross, 1975; Jackson et al., 1985). Almost 7-11% of the total mRNA in *T. brucei* consists of VSG's transcripts, and greater than 95% of the proteins on the surface are VSGs (Grünfelder et al., 2002; Nilsson et al., 2010). There are about 2,000 different VSG genes and pseudogenes present in the *T. brucei* genome repertoire (Cross et al., 2014). However, only about 350 genes are predicted to be functional (complete reading frame, no frameshifts or early termination codon present - (Cross et al., 2014)). On the amino acid sequence level, these VSGs are highly diverse, sharing about 10-35% sequence identity. In BSF, VSGs are transcribed by RNA Polymerase I (RNAP I) from their approximately 20 telomeric bloodstream expression sites (BES) (Figure 3B) (Günzl et al., 2015; Hertz-Fowler et al., 2008). In trypanosome, the trans splicing spliced the splice leader sequence to the 5' end of each mRNA. This process represents a post-transcriptional mode of capping that lets trypanosomes to use other RNA Polymerases other than RNAP II (Günzl et al., 2015). Not only VSGs, but there are also other genes that are co-expressed with VSGs from BES, namely Expression Sites Associated Genes (ESAGs). Fourteen to 25 different ESAGs have been found in BSF T. brucei and encode an array of different proteins (Cully et al., 1985). ESAG4, for example, encodes a transmembrane protein, which has a adenylate cyclase activity (Ross et al., 1991). Both ESAG6 and 7 encode a heterodimer transferrin receptor (Salmon et al., 1997; Schell et al., 1991) . ESAG 10 encodes a transmembrane protein that is found to be homologous to biopterin receptor in *Leishmania donovani* (*L. donovani*) (Lemley et al., 1999). Despite the vast number of VSG genes and pseudogenes in the *T. brucei* genome repertoire, only one VSG is expressed at a given time from one of the BES (monoallelic expression) (Borst and Ulbert, 2001). *T. brucei* continuously evades the mammalian host immune response by antigenic variation, which is achieved by changing the currently expressed VSG into an antigenically distinct VSG. Because the immune system only recognizes the previously expressed VSG (as the VSGs are antigenically distinct), those parasites that switch their coat can no longer be

recognized and can proliferate (Hovel-Miner et al., 2015). Thus, it results in a continuous infection of the mammalian host. This behavior is reflected in the parasitemia waves that have been observed and is characteristic of trypanosome infection (Figure 3A).

Procyclins are repeat proteins, which harbor a set of tandem sequence repeats. There are two forms of procyclins: GPEET and EP (Roditi et al., 1998). GPEET consists of six recurring sequences of Gly-Pro-Glu-Glu-Thr residues and transcribed from a single gene. Meanwhile, EP is transcribed from a set of three genes and consists of up to thirty Glu-Pro repeats (Clayton and Mowatt, 1989; Mowatt and Clayton, 1987; Richardson et al., 1988; Roditi and Clayton, 1999). EP was identified and purified from PCF cells by combinations of Concanavalin A affinity chromatography in tandem with other chromatography techniques (Clayton and Mowatt, 1989; Ferguson et al., 1993; Richardson et al., 1988). Alike VSGs, procyclins genes are also transcribed by RNA Polymerase I in PCF (Günzl et al., 2003). The promoter region of procyclins are not telomeric like VSG promoters in BSF and MCF (Vanhamme and Pays, 1995). Procyclins are expressed in high density at the surface of PCF. Approximately, six million procyclins covered the PCF cell's surface (Clayton and Mowatt, 1989). Such repeat proteins are proposed to form an amyloid mesh (Tsiolaki et al., 2018). The expression of GPEET and EP depends on how long the cells have been cultured and the culturing condition (Bütikofer et al., 1999, 1997; Morris et al., 2002; Vassella et al., n.d.). However, GPEET is primarily expressed in the early stage of PCF. As the PCF differentiates further, the surface proteome transitions to EP dominance (Acosta-Serrano et al., 2001). When the PCF cells differentiate further into elongated epimastigotes and undergo asymmetric division, the smaller epimastigotes, which is attached to the fly's salivary gland, express BARP (Brucei Alanine-rich Proteins) before it matures into the MCF cells (Urwyler et al., 2007).

Like BSF cells, the surface of MCF cells is also covered by mVSG (Ramey-Butler et al., 2015), although the exact number of mVSG molecules present on the surface is still unknown. In MCF, VSGs are also transcribed by RNA polymerase I from dedicated, also telomeric, metacyclic expression sites (MES) (Figure 3B) (Günzl et al., 2003). While the VSGs transcripts in BSF cells are polycistronic, the VSGs transcripts in MCF cells are monocistronic and there is no co-expression of ESAGs (Figure 3B)

(Alarcon et al., 1994; Graham and Barry, 1995). The shorter promoter region and VSGs transcripts in MCF cells have been presumed to conserve the energy and resources used in MCF cells due to the quiescent nature of the MCF cells (Christiano et al., 2017). While there is a vast number of VSGs genes present in genome repertoire of *T. brucei*, there are only five VSGs expressed in *T. brucei* Lister 427 strain MCF cells (mVSGs), namely mVSG397, 531, 639, 653, and 1954 (Christiano et al., 2017; Cross et al., 2014). Alike bloodstream VSGs, individual cell only expresses one mVSG at a given time (Ramey-Butler et al., 2015).

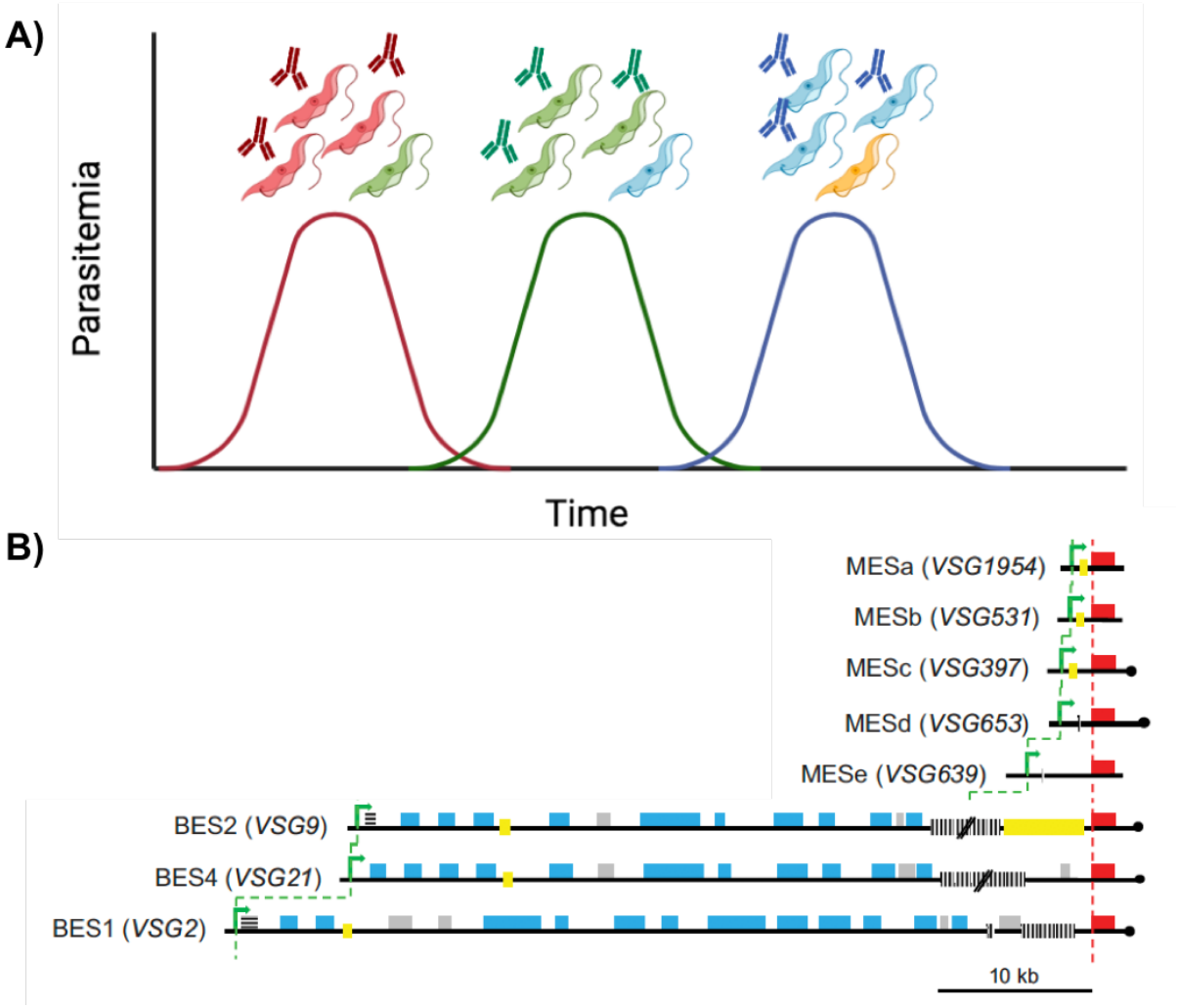


***Figure 3.*** ***Antigenic Variation and VSG expression sites in T. brucei BSF and MCF.***

**A)** *T. brucei* evades the mammalian host immune system by antigenic variation. When an immune response is mounted against one type of VSGs, a small population of parasites changes its VSG expression into an antigenically different VSG. The antigenic variation results in the continual infection of the mammalian host, characterized by waves of parasitemia. **B)** mVSGs transcripts are shorter and monocistronic than the VSGs expressed in the BSF cells. (green arrow: VSG promoter, blue: Expression Sites Associated Genes (ESAGs), red: VSGs) (Christiano et al., 2017)

## 4. Variant Surface Glycoprotein (VSG) structural diversity

The VSG polypeptide chain is divided into four regions. The first 20 to 30 amino acids at the N-terminus of the polypeptide chain is the signal peptide (Al-Qahtani et al., 1998). The signal peptide is eventually cleaved during the VSG's maturation process. The next and largest part of the polypeptide chain is the VSG N-terminal domain (Blum and Turner, 1993; Freymann et al., 1990). The N-terminal domain consists of about 350 amino acids (Figure 4A). The VSG's N-terminal domain is highly antigenic and serves as the primary host-parasite interaction site (Figure 4B, 4C) (Hempelmann et al., 2021; Mugnier et al., 2016). The VSG C-terminal domain is smaller than the N-terminal domain, consisting of around 80 amino acids (Chattopadhyay et al., 2005). The GPI-anchor that inserts the VSG into the membrane is attached at the C-terminal domain, and a short far-C-terminal peptide is removed during this process (Figure 4A).

Most VSG classification schemes are based on the similarity of their amino acid sequence and the number and location of cysteine residues likely to form disulfide bridges. Weirather et al. (2012) suggest five VSG N-terminal domain classes and six VSG C-terminal domain classes on these and other criteria (Weirather et al., 2012). .

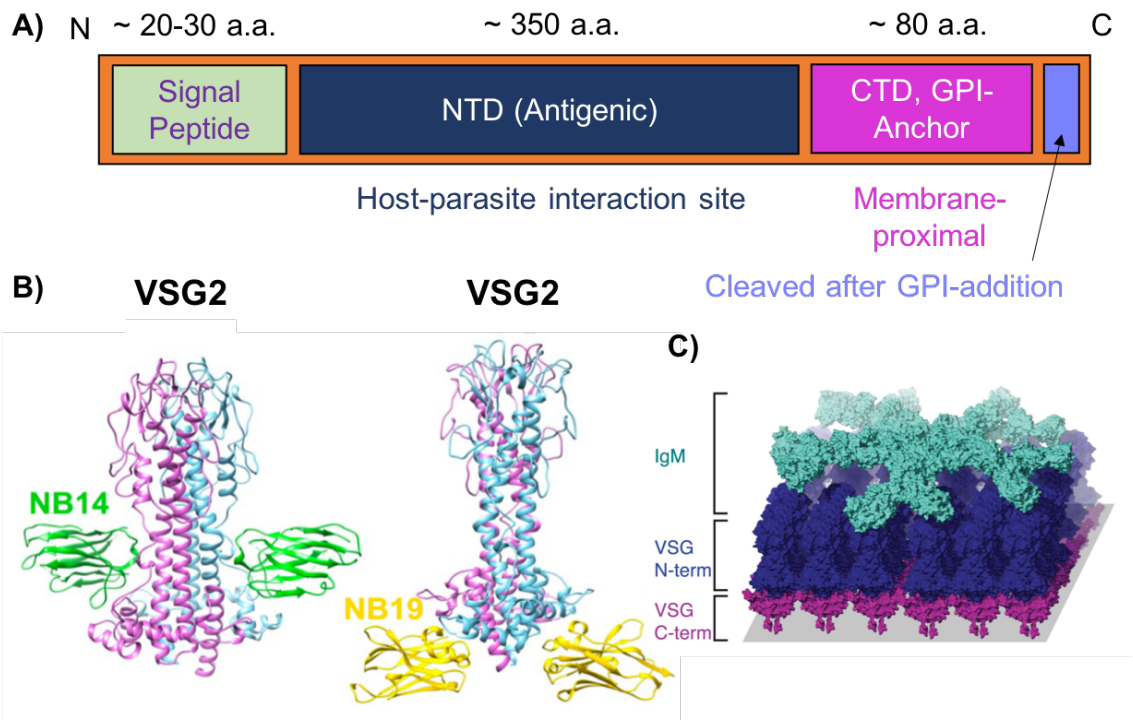


***Figure 4.*** ***VSG polypeptide and VSG-antibody interactions.***

**A)** VSG polypeptide is divided into four parts, the signal peptide, the N-terminal domain (NTD), the C-terminal domain (CTD), and a short peptide at the C-terminus removed upon GPI addition. The signal peptide is 20 to 30 amino acids long and cleaved during the protein's maturation. The NTD is the most significant part of the polypeptide chain. The NTD is antigenic and the host-parasite interaction site. The CTD is the smaller domain closer to the membrane and bearing the GPI-anchor. **B)** Interaction between the VSG2's NTD (blue and magenta) and several nanobodies (green and yellow) produced from camelids (image adapted from: Hempelmann et al., 2021). **C)** Rough schematic illustrating how VSGs (purple: NTD, magenta: CTD) are organized on the parasite's surface and recognized by the pentameric IgM molecule (teal) (image adapted from: (Mugnier et al., 2016).

The first VSG N-terminal domain structure, namely MITat 1.2 (also known as VSG221 and lately as VSG2), was determined by X-ray crystallography nearly three decades ago. The atomic-resolution structure reveals that VSG2 possesses a "dumbbell-like" architecture made up of a core a three-helix bundle capped at either end by an upper and lower "lobe" of polypeptide. The structure also shows that VSG2 forms a homodimer and that an N-linked glycan is attached at the structure's bottom lobe (Freymann et al., 1990). Next, the structure of VSG ILTat 1.24 was also determined and showed an overall similarity with the VSG2 structure but with variations. Both structures share the dumbbell-like structure and the three-helix bundle

core (Blum and Turner, 1993). These findings have supported the assumption that VSGs' distinct antigenicity relies solely on their amino acid sequence differences (particularly on the surface of the structures). Three decades later, the structure of VSG1 (MITat 1.1) was published. Like VSG2 and IITat 1.24 structures, the VSG1 structure also reveals the dumbbell-like architecture with the three-helix bundle core (Figure 5) (Bartossek et al., 2017). It's close similarity to VSG2 seemed to support the idea that the VSGs were fairly related to each other in overall protein fold.

In 2018, the publication of a new VSG structure, BSF VSG3 (MITat 1.3), challenged many of these long-held assumptions. VSG3 does possess the three-helix bundle motif like the VSG2, IITat 1.24, and VSG1 structures. However, the VSG3 structure reveals significant topological differences compared to the previous structures. In the structures of VSG1, VSG2, and ILTat1.24, the bottom lobe is composed of sequence from the C-terminal region of the N-terminal domain's polypeptide chain. In contrast, the bottom lobe of VSG3 consists of polypeptide sequence from the beginning of the N-terminal domain, thus showing the connectivity in the 3-dimensional structure differs between VSG3 and the previously determined protein structures. VSG3 is further distinguished from the previous work as it possesses an *O*-linked glycan at the N-terminal domain's top lobe attached to a serine residue (S317) (Pinger et al., 2018). *O*-linked glycosylation had not been previously described in *T. brucei*. In a mouse infection study, it was shown that the *O*-linked glycan is essential for *T. brucei's* survival in mice and that its presence dampens the immune response. Mice infected with *T. brucei* expressing the wildtype VSG3 succumbed to death in the first parasitemia wave. In contrast, the mice infected with the *O*-glycan mutant of VSG3 survived the first parasitemia wave (other data revealed clear distinctions in antibody responses to these different coats). The data shows that the *O*-linked glycan is essential in the immune evasion of *T. brucei* in a mouse infection. Like the previously determined structures, an N-linked glycan is also present in the bottom lobe of VSG3. However, the VSG3 N-terminal domain structure exists as a monomer instead of a dimer in the crystal asymmetric unit and in solution, and the monomers form a crystallographic trimer (Pinger et al., 2018). The physiological relevance of the monomeric and trimeric structures has not been shown.

The recently published structures of VSGsur and VSG13 have also shown that VSG poteins are even more diverse than previously presumed from the small subset of structures known to the field for thirty years (Figure 5) (Zeelen et al., 2021). Interestingly, VSGsur has also been shown to confer resistance against suramin (Wiedemar et al., 2018), the first-line drug for trypanosome infection, by binding to suramin at the cavity between the dimer. While maintaining the conserved three-helix bundle core, the structures have revealed a highly organized top-lobe, a large β-sheet that combines in the VSG dimer to form an intermolecular β-sandwich. Compared to the previously determined structures, the N-linked glycan is located at the middle of the core stalk instead of at the structure's bottom lobe. The multiple disulfide bonds are not clustered to the top lobe but are dispersed throughout the proteins (Figure 5) (Zeelen et al., 2021). All these features show that previous assumptions about the conservation of VSG structural elements must be reconsidered, including the idea that the VSGs are passive conduits of antigenic variation and not active in the biology of the organism (as VSGsur is in conferring heightened drug resistance (Wiedemar et al., 2018)).

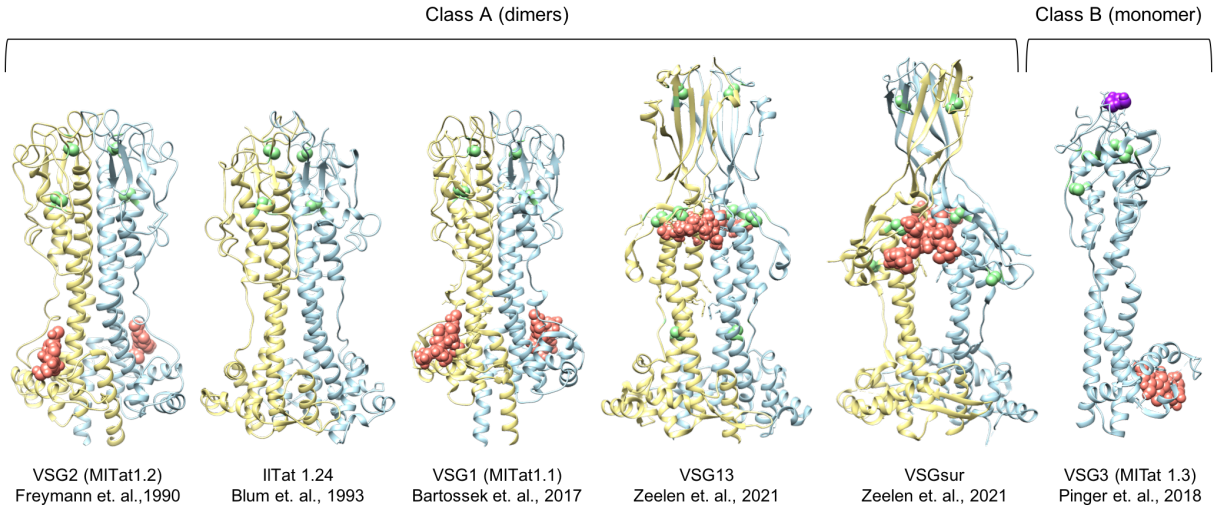


***Figure 5.*** ***Bloodstream VSGs' structures.***

The structures of VSG2, IITat 1.24, VSG1, VSG13, and VSGsur in "Class A" form homodimers. The VSG3's structure, included in "Class B" exists as monomer and *O*-glycosylated at the molecule's top lobe (salmon pink: N-linked glycan, green: disulfide bonds, purple: O-linked glycan) (image adapted from: Bartossek et al., 2017; Blum and Turner, 1993; Freymann et al., 1990; Pinger et al., 2018; Zeelen et al., 2021) Image was created using CHIMERA (Pettersen et al., 2004)

## 5. X-ray crystallography

(The content of this section is mostly adapted from the book Outline of Crystallography for Biologist, Blow, D. (2010) and Protein Crystallization, Bergfors, T. M. (2009), otherwise stated)

X-ray crystallography is one of the primary techniques in the field of structural biology. One can determine a biomolecule structure to atomic resolution by this method, determining the precise locations of all amino acids, co-factors, and solvating ions around the protein. This technique has aided researchers in determining enzyme function and mechanism, small-molecule drug design, protein-protein interaction, and many other aspects of biology at the molecular level. However, X-ray crystallography relies on the existence of protein crystals that are sometimes very challenging to obtain. A general and simplified X-ray crystallography workflow from protein to structure can be viewed in Figure 6A (Büttner et al., 2015).

The first step of the technique starts with obtaining a large amount of high purity soluble protein of interest. Many proteins exist in a minuscule amount in the cells under normal conditions; therefore, many proteins have to be overexpressed to get enough protein to do a crystallization trial. It is crucial to determine the suitable protein expression system and optimal conditions for protein expression and methods of purification.

After obtaining the purified protein, the next step is determining suitable crystallization conditions to obtain crystals that diffract to high resolution (showing ordered scattering beyond ~3.5Å). The most common technique used to obtain protein crystals is vapor diffusion (Figure 6B). In this technique, the purified protein is mixed with a mother liquor (or reservoir solution) in a specific ratio. The mother liquor consists of precipitating agents (e.g., Polyethylene glycol (PEG), ammonium sulfate, and many more) and other chemicals (salts, buffer, organic compounds). The reagent concentration in the mixture is lower than that in the reservoir as the two volumes are combined. The difference in concentration between the drop and the well will cause the crystallization drop to shrink as water diffuses to equilibrate. This step remains the technique's bottleneck due to the numerous variables influencing the crystallization process. Some of the variables are the nature and concentration of precipitating agents, added salts, buffer, pH, temperature, and many more. The initial crystallization

condition's initial search is usually done by a sparse matrix screen. The sparse matrix screen is a collection of very diverse crystallization conditions. It is usually available commercially but can be formulated in the laboratory as well. The initial conditions are often based on crystallization conditions in the available literature that have successfully produced crystals. Should crystalline material be obtained in some combination of reagents, this initial condition is further optimized by making concentration grids of the different chemicals in the crystallization mix to obtain better crystals for X-ray diffraction. Diffraction data collection is usually performed with laboratory sources or the much more powerful and wavelength-tunable X-ray sources at synchrotrons.

The X-ray diffraction data is recorded by a detector as thin-slices of two-dimensional image and converted into electron density maps using a Fourier transforms mathematical equation. While the intensity and amplitude information can be extracted from the diffraction data, the phase information is missing, known as the "phase problem" in X-ray crystallography. There are two ways to solve the phase problem: molecular replacement and experimental phasing. Molecular replacement uses the phase information from a related structure with high sequence identity to generate the electron density map. Most of the structures deposited in the Protein Data Bank (PDB) have been determined using molecular replacement (Abergel, 2013). There are several different methods to experimentally phase diffraction data. Among those are multiple isomorphous replacement (MIR), single isomorphous replacement (SIR), single and multiple isomorphous replacement with anomalous scattering (SIRAS and MIRAS), single-wavelength anomalous dispersion (SAD), and multiwavelength anomalous dispersion (MAD). Heavy atom derivative crystals are needed to perform experimental phasing, whereby derivative means that a crystal (the native) has been soaked or otherwise exposed to electron rich atoms (sometimes achieved recombinantly as with the replacement of sulfur in methionine with selenium). A series of different isomorphous heavy atoms derivative crystals are needed to perform MIR to determine the phase angle. The derivative crystals need to have the same unit cells as the native crystal to be called isomorphous. The difference in the X-ray diffraction patterns between the native and the derivative crystals can be used to solve the structure. Obtaining a series of different isomorphous derivative crystals is hard due to the disruption of the structure, which comes from the heavy atom binding.

Therefore, another method, SIR, was developed to enable the structure determination with only one type of isomorphous derivative crystal. Nowadays, similar methods, which are based on the difference in the diffraction patterns by anomalous scattering, are used more widely compared to the isomorphous replacement methods. Anomalous scattering occurs when the incident X-ray photon has an energy close to an electron absorption edge and the scattering is no longer elastic. One important effect of anomalous scattering is that some of the scattered energy is no longer in phase with the incident photon, information that can be used to solve the phase problem. Anomalous data collection is performed by collecting X-ray diffraction near an accessible absorption edge of the heavy atom (one that the synchrotron can reach). Such diffraction contains an anomalous structural factor component and the difference between the normal and the anomalous structure factors can determine the phase. While SAD uses one wavelength on one crystal, MAD uses several different wavelengths, each producing different anomalous components. Two other techniques, SIRAS and MIRAS, combine both isomorphous replacement and anomalous scattering. These two methods can be achieved when the anomalous scatterer is the same at the heavy atoms used in the isomorphous replacement. Following the generation of the electron density map, the structure's molecular modeling can be performed, as the polypeptide chemical groups are built into the maps. The initial model undergoes several refinement cycles.

After each refinement cycle, the model is assessed by several parameters. One of the essential parameters in assessing the model quality is R-value. The R-values is defined by the equation below:

$$R = \frac{\Sigma\big||F_{obs}| - |F_{calc}|\big|}{\Sigma|F_{obs}|}$$

$F_{obs}$ is the structure amplitude observed in the experimental data, while $F_{calc}$ is the structure amplitudes calculated from the model. In other words, the R-value measure how well the structure generated from the refinement agrees with the experimental data. Typical, well-refined structures have R-values ($R_{work}$) between 15 to 25% (or can be written as 0.15 – 0.25), depending on the resolution and quality of the model and data. However, during the refinement process, the model is continuously fitted to the

experimental data, which can result in a biased (lower) R-value in which changes in model parameters (x, y, z position and temperature factor) do not reflect improvements to the model. This is called *overfitting* and is more likely to occur when the ratio of variables to experimental data is high. To overcome this bias and obtain an R-value that reflects the quality of the model, 5-10% of the experimental data is removed when performing refinement (a specific implementation of the statistical principle of *cross-validation*). In the end, the model is assessed against the remaining 5-10% unused observations, creating an unbiased parameter called $R_{free}$. In non-problematic structures, the $R_{work}$ and $R_{free}$ are usually similar (<5% difference is considered acceptable), with $R_{free}$ to be commonly higher than the $R_{work}$. A difference of more than 5% between the two usually indicates that the existing model may be over-refined. Other parameters such as clash score, Ramachandran outliers, and sidechain outliers have also been used to assess the model's quality. These parameters consider each atom and amino acid residue's geometric and chemical constraints in the model.

Despite the efforts to make the workflow more autonomous, X-ray crystallography is still a lengthy and laborious process. Therefore, choosing a suitable protein of interest is needed at the beginning of the project. Several software packages, such as PHENIX and CCP4i2 (Adams et al., 2010; Potterton et al., 2018), make the structure determination process more accessible to general users.

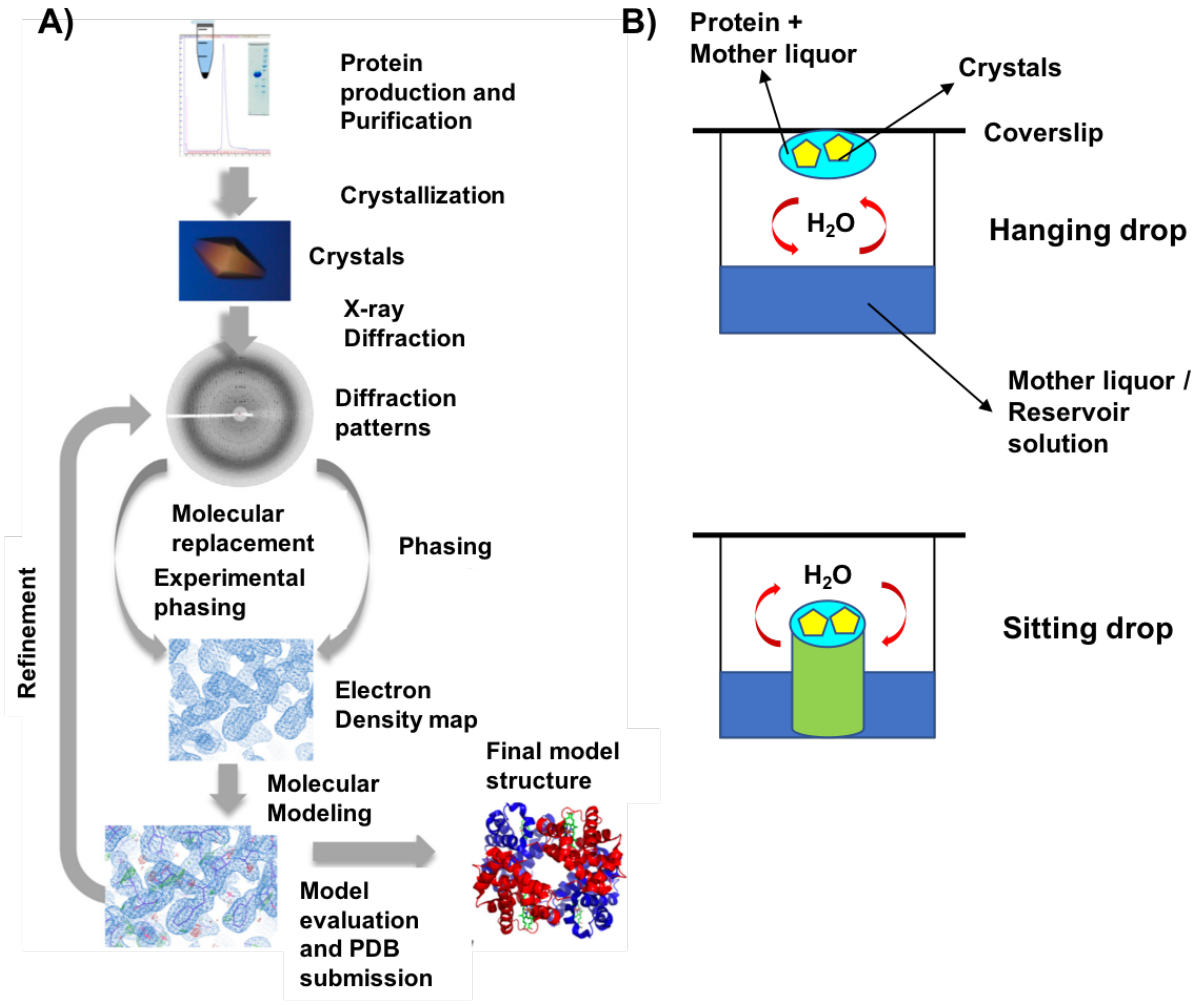


***Figure 6.*** ***X-ray crystallography at a glance.***

**A)** General x-ray crystallography workflow (Büttner et al., 2015) **B)** Diagram of hanging and sitting drop vapor diffusion technique to obtain protein crystals.

## 6. Thesis scope

When this study was started, there was little information available on the mVSGs. Particularly, no information was available on the structure of mVSGs; and so, it was not known whether they could, for instance, fold differently, or have distinct PTMs – or whether they were in any way different than "standard" VSGs expressed in bloodstream form parasites. This is particularly important as mVSGs are the first antigenic surface presented by the parasites to the mammalian host. Therefore, this doctoral thesis is focused on determining the structures of various mVSGs and comparing them to the BSF VSGs. This study's primary technique is X-ray

crystallography because of its capability to determine a molecule's structure down to atomic resolution to get a detailed picture.

A large amount of high purity protein is needed to perform X-ray crystallography. Because the VSGs are highly expressed endogenously in the cells of the African trypanosome and 95% of the cell surface proteome consists of VSGs (Grünfelder et al., 2002), mVSGs can be isolated is significant quantities and at high purity directly from the native organism using a modification of the method described in Cross, 1984 (Cross, 1984, p. 19). However, MCF cells are cell-cycle arrested and cannot be grown in culture (Christiano et al., 2017). Therefore, to get many MCF cells, large amounts of flies would need to be dissected to obtain enough protein, which is not practical for this study. Besides, MCF cells randomly express one of the five annotated mVSGs (Christiano et al., 2017), and the presence of other mVSGs will complicate the downstream purification process.

To circumvent this challenge, we therefore expressed two of the mVSGs, mVSG531 and mVSG1954, in BSF cells. BSF cells can proliferate in culture; thus, enough cells can be obtained in a short time. BSF cells were transfected with a linearized plasmid that encodes the coding sequence of mVSG531 or mVSG1954 with the hygromycin drug-resistance gene (for details, see Chapter II). Upon transfection, the plasmid will integrate into the active BES, and the mVSG of interest will be continuously transcribed under the hygromycin selection drug included in the growth media. The surviving cells that express mVSG531 or mVSG1954 were grown to high density to obtain enough protein. The subsequent purification steps from BSF cells' membranes were performed according to Cross, 1984 with modifications as described in Pinger et al., 2018.

The purified proteins were used in crystallization trials, the initial crystallization conditions determined by a sparse matrix screen. The conditions yielding imperfect crystals were further optimized until high-resolution diffracting crystals were obtained. The X-ray diffraction data were collected at the synchrotrons. Molecular replacement (mVSG531 using the model from VSG1) or experimental phasing using SAD (mVSG1954 using soaks with sodium bromide) determined the structures.

Because the mVSGs were not expressed in the proper life stage, it is possible that there could be differences in the antigenicity of these VSGs (e.g., due to folding or post-translational modifications). Therefore, comparisons between the mVSGs expressed in the BSF cells and the mVSGs expressed in the MCF cells were made. Antisera against mVSG531 ($\alpha$-mVSG531) or mVSG1954 ($\alpha$-mVSG1954) were raised in mice by infecting the mice with the BSF cells expressing mVSG531 (BSF mVSG531) or BSF cells expressing mVSG1954 (BSF mVSG1954). Meanwhile, by overexpressing RNA Binding Protein 6 (RBP6) in PCF cells, the PCF cells are differentiated into MCF cells. The generated antisera were used in flow cytometry to detect the MCF cells that express mVSG531 or mVSG1954 on their surfaces.

With this study, we expected to gain more insight on the structures of mVSGs and how those compare to their bloodstream counterparts. The structures of mVSGs might also reveal probable new PTMs of other additional attributes on VSGs. The proteomic data in Christiano et. al., 2017 has hinted a possible phosphorylation on the mVSG653, a type of modification that has not been observed in any of the bloodstream VSGs structures to date. As mVSGs are the first antigenic surface that is presented to the mammalian host, the study on mVSGs will also be beneficial in designing a potential vaccine for the prevention of trypanosome infection.

## Chapter II: Materials and Methods

### 1. Reagents List

| Reagents | Manufacturer | Catalog Number |
|---|---|---|
| 2-Mercaptoethanol | Sigma | M3148 |
| 2-Propanol | Sigma | 33539 |
| Acetic acid, 100% | Sigma | 3738.4 |
| Agarose | Roth | 3810.3 |
| Ampicillin sodium salt | Sigma | A9518 |
| Bathocuproinedisulfonic acid | Santa Cruz Biotechnology | SC-217698A |
| Bradford reagent | Serva | 39222 |
| Bromophenol Blue Na-salt | Serva | 15375 |
| Calcium chloride dihydrate | Carl Roth | 5239.1 |
| cOmplete Mini | Roche | 11836153001 |
| Coomassie Brilliant Blue G 250 | Serva | 17524 |
| DL-Dithiothreitol | Sigma-Aldrich | D0632 |
| DNAzol R reagent | VWR International | J60637.AD |
| Ethanol absolute | Honeywell | 32205 |
| Fetal Bovine Serum (FBS), Qualified | Gibco | 10270-106 |
| Geneticin (G418 sulfate) | Gibco | 11811023 |
| Glycerol | Roth | 56-81-5 |
| Glycerol for cell culture | Sigma-Aldrich | G2025 |
| Hemin, Bovine | Sigma-Aldrich | H9039 |
| HEPES | Carl Roth | 9105.4 |
| Hydrochloric acid fuming, 37% | Carl Roth | 4625.1 |
| Hygromycin B | Invitrogen | 10687010 |
| L-Cysteine-HCl-$H_2O$ | Serva | 17769 |
| LB Broth | Sigma | L3022 |
| LB Broth with agar | Sigma | L2897 |
| Midori green advance | Biozym scientific | 617004 |

| | | |
|---|---|---|
| MMT pH 4.0 -1M | Jena Biosciences | CS-398 |
| MMT pH 9.0 -1M | Jena Biosciences | CS-399 |
| N-acetyl-D-glucosamine | Sigma | A3286 |
| Nuclease free water | Invitrogen | AM9937 |
| Penicillin Streptomycin (Pen-Strep) | Sigma | P4333 |
| Polyethylene glycol 1500 | Sigma | 86101 |
| Polyethylene glycol 3350 | Sigma | 88276 |
| Puromycin | Gibco | A11138-03 |
| Q-Sepharose FastFlow | GE Healthcare | GE17-0510-01 |
| SDS solution 10% | PanReac Applichem | A3950 |
| Sodium bicarbonate | Sigma-Aldrich | S5761 |
| Sodium bromide | Sigma-Aldrich | 71329 |
| Sodium dodecyl sulfate (SDS) | Sigma | 75746 |
| Sodium hydroxide | Sigma-Aldrich | 30620 |
| Tetracycline | Gibco | A39246 |
| Tris-(2-carboxyethyl) phosphine hydrochloride (Trizma Base) | Sigma-Aldrich | 75259 |
| Trypsin | Sigma | T1426 |
| Zinc Chloride | Merck | 1.8816 0250 |

## 2. Antibody List

| Antigen | Host | Isotype | Manufacturer | Clone | Catalog Number |
|---|---|---|---|---|---|
| *T. brucei* procyclins (GPEET) | Mouse | IgG3 | Cedarlane Labs | 9G4 | CLP009A |
| *T. brucei* procyclins (EP) | Mouse | IgG1 | Cedarlane Labs | TBRP1/ 247 | CLP001A |
| α-Mouse IgM-FITC | Goat | IgG | Southern Biotech | - | 1021-02 |
| α-mVSG531 | Mouse C57BL6/J | Polyclonal | Papavasilliou Lab (DKFZ) | - | - |
| α-mVSG1954 | Mouse C57BL6/J | Polyclonal | Papavasilliou Lab (DKFZ) | - | - |

## 3. Buffers list

| Buffer Name | Buffer composition |
|---|---|
| SEC-Buffer | 20 mM HEPES pH 7.5, 150 mM NaCl |
| TGS-Buffer | 25 mM Tris-HCl, 192 mM Glycine, 0.1% SDS |
| 4x SDS Loading Buffer | 0.2M Tris-HCl pH 6.8, 0.4M DTT, 8% SDS, 40% v/v glycerol, 0.4% w/v bromophenol blue. |

## 4. Plasmid list

| Plasmid Name | Description |
|---|---|
| pmVSG531 | pFAB6 with insert encoding mVSG531 CDS and Ampicillin resistance gene |
| pmVSG1954 *wildtype* | pFAB6 with insert encoding mVSG1954 *wildtype* CDS and Ampicillin resistance gene |
| pmVSG1954 S321A | pFAB6 with insert encoding mVSG1954 S321A CDS and Ampicillin resistance gene |

## 5. Cell lines list

| Cells name | Organism | Description and Usage |
|---|---|---|
| DH5α | *E. coli* | Cloning, plasmid propagation |
| BSF mVSG531 | *T. brucei* | Bloodstream cells expressing mVSG531 used in mVSG531 protein purification and generation of α-mVSG531 in mice |
| BSF mVSG1954 wildtype | *T. brucei* | Bloodstream expressing mVSG1954 wildtype used in mVSG1954 wildtype protein purification and generation of α-mVSG531 in mice |
| BSF mVSG1954 S321A | *T. brucei* | Bloodstream cells expressing mVSG1954 S321A used in mVSG1954 S321A protein purification |
| PCF 29-13 RBP6 | *T. brucei* | Procyclic cells with tetracycline-inducible RBP6 overexpression system used in the procyclic differentiation to metacyclic |

## 6. T. brucei's growth media

### HMI-9 *T. brucei* growth media

| Components | Concentration (g/L) |
|---|---|
| IMDM (without HEPES, Phenol red, L-glutamine, sodium bicarbonate, with 4.5050 g/L NaCl) | 11. 0965 |
| L-glutamine | 0.649 |
| HEPES | 6.620 |
| Phenol red | 0.017 |
| Sodium bicarbonate | 3.360 |
| Hypoxanthine | 0.151 |
| Thymidine | 43.3 |
| Sodium pyruvate | 0.122 |
| Bathocuproinedisulfonic acid disodium salt | 0.202 |

The HMI-9 media in this study is manufactured by PAN Biosciences, Germany.

Additional components to be added into 900 ml of HMI-9 to make 1 L of supplemented HMI-9 media:

| Component | 100x stock | Comments |
|---|---|---|
| L-cysteine-HCl-$H_2O$ | 18.2 mg/ml | The mixture is made fresh prior to use and sterile-filtered, add 10 ml into the media |
| 2-Mercaptoethanol | 1.4 ul/ml | |
| Penicillin-Streptomycin | | Add 10 ml into the media |
| Fetal Bovine Serum | | Add 100 ml into the media, final concentration: 10% |

**SDM-80 *T. brucei* PCF growth media**

| Salts | CAS Number | Molecular weight (g/mol) | Concentration (mg/L) |
|---|---|---|---|
| Calcium chloride x $2H_2O$ | 10035-04-8 | 147.01 | 53.77 |
| Iron (III) nitrate x $9H_2O$ | 7782-61-8 | 404.00 | 0.1462 |
| Magnesium sulfate dried | 7487-88-9 | 120.37 | 126.173 |
| Potassium chloride | 7447-40-7 | 74.55 | 345.68 |
| Sodium chloride | 7647-14-5 | 58.44 | 5876.56 |
| Sodium dihydrogen phosphate x 2 $H_2O$ | 13472-35-0 | 156.00 | 1078.63 |
| **Amino Acids** | **CAS Number** | **Molecular weight (g/mol)** | **Concentration (mg/L)** |
| L-Alanine | 56-41-7 | 89.09 | 210.415 |
| L-Arginine free base | 74-79-3 | 174.20 | 100 |
| L-Arginine HCl | 1119-34-2 | 210.66 | 148.08 |
| L-Asparagine x $H_2O$ | 5794-13-8 | 150.13 | 9 |
| L-Aspartic acid | 56-84-8 | 133.10 | 14.07 |
| L-Cysteine x HCl x $H_2O$ | 7048-04-6 | 175.63 | 0.02233 |
| L-Cystine milled | 56-89-3 | 240.30 | 9.6 |
| L-Cystine x 2 HCl | 30925-07-6 | 313.22 | 25.968 |
| Glycine | 56-40-6 | 75.07 | 14.65 |
| L-Histidine x HCl x $H_2O$ | 5934-29-2 | 209.63 | 49.012 |
| L-Hydroxyproline | 51-35-4 | 131.10 | 2.03 |
| L-Isoleucine | 73-32-5 | 131.17 | 59.44 |
| L-Leucine x HCl | 61-90-5 | 131.18 | 67.52 |
| L-Lysine x HCl | 657-27-2 | 182.65 | 91.15 |
| L-Methionine | 63-68-3 | 149.21 | 89.005 |
| L-Phenylalanine | 63-91-2 | 165.19 | 119.435 |
| L-Serine | 56-45-1 | 105.09 | 71.375 |
| Taurine | 107-35-7 | 125.15 | 160 |

| | | | |
|---|---|---|---|
| L-Tryptophan | 73-22-3 | 204.23 | 12.722 |
| L-Tyrosine | 60-18-4 | 181.19 | 114.4 |
| L-Tyrosine x 2 Na x $2H_2O$ | 122666-87-9 | 261.19 | 46.02 |
| L-Valine | 72-18-4 | 117.15 | 54.21 |
| **Vitamins** | **CAS Number** | **Molecular weight (g/mol)** | **Concentration (mg/L)** |
| 4-Aminobenzoic acid | 150-13-0 | 137,14 | 2,01015 |
| L(+)-Ascorbic acid | 50-81-7 | 176,13 | 0,01015 |
| D(+)-Biotin | 58-85-5 | 244,31 | 0,20203 |
| D-Calcium-Pantothenate | 137-08-6 | 476,54 | 0,66323 |
| Choline chloride | 67-48-1 | 139,62 | 0,7627 |
| Folic acid | 59-30-3 | 441,41 | 4,6632 |
| myo-Inositol | 87-89-8 | 180,16 | 1,33255 |
| Menadione | 58-27-5 | 172,19 | 0,00203 |
| Nicotinamide | 98-92-0 | 122,13 | 0,666275 |
| Nicotinic Acid | 59-67-6 | 123,11 | 0,005075 |
| Pyridoxal x HCl | 65-22-5 | 203,62 | 0,666275 |
| Pyridoxine x HCl | 58-56-0 | 205,64 | 0,005075 |
| **Other components** | **CAS Number** | **Molecular weight (g/mol)** | **Concentration (mg/L)** |
| Adenine sulfate | 321-30-2 | 368,34 | 2,03 |
| Adenosine | 58-61-7 | 267.25 | 10 |
| Adenosine-5'-monophosphate x 2Na x $2H_2O$ | 4578-31-8 | 427.22 | 0,04506 |
| Adenosine -5'-triphosphate x 2Na | 987-65-5 | 551.10 | 0,223 |
| Cholesterol | 57-88-5 | 386.66 | 0,0406 |
| Deoxyribose | 533-67-5 | 134.13 | 0,1015 |
| L-Glutathione reduced | 70-18-8 | 307.32 | 0,01015 |
| Guanine x HCl | 635-39-2 | 187.59 | 0,05556 |

| | | | |
|---|---|---|---|
| Guanosine | 118-00-3 | 283/24 | 10 |
| HEPES | 7365-45-9 | 238/31 | 8000 |
| Hypoxanthine | 68-94-0 | 136.11 | 0,0619 |
| MOPS | 1132-61-2 | 209.26 | 5000 |
| Phenol Red sodium salt | 34487-61-1 | 376.36 | 10,924 |
| D-Ribose | 50-69-1 | 150.13 | 0,1015 |
| Thymine | 65-71-4 | 126.11 | 0,0609 |
| Tween 80 | Not applicable | Not applicable | 4,06 |
| Uracil | 66-22-8 | 112.09 | 0,0609 |
| Xanthine | 69-89-6 | 152.11 | 0,061 |

## 7. Extraction of T. brucei genomic DNA (gDNA) isolation

*T. brucei* gDNA was extracted from BSF cells expressing VSG2. Ten million cells were used in the genomic DNA extraction. The cells were harvested by centrifugation at 1,800 rpm for 10 minutes at 22$^{o}$C. The supernatant was removed, while leaving 500 μl of the media to resuspend the cells pellet. The cells mixture was transferred into a 1.5 ml Eppendorf tube, and subsequently centrifugated at 2,500 xg for 4 minutes at 22$^{o}$C. The supernatant was removed completely, and the cells pellet was resuspended and lysed by mixing it with 1 ml of DNAzol R reagents to isolate the genomic DNA. The mixture was incubated for 5 minutes at room temperature. Subsequently, 500 μl 100% Ethanol was added into the mixture. The tube was inverted multiple times to homogenized the mixture. The mixture was again incubated at room temperature for 5 minutes. The genomic DNA was pelleted by centrifugation at 12,000 xg for 15 minutes at 4$^{o}$C. The supernatant was discarded, and the precipitated genomic DNA was washed 1 ml of 70% Ethanol. The mixture was centrifugated at 12,000 xg for 5 minutes. The ethanol was discarded, and the pellet was centrifugated again at 20,000 xg for 1 minute. The remaining ethanol was removed by pipetting, and the genomic DNA pellet was subsequently resuspended in 50 μl 5 mM Tris HCl pH 8.5 and stored in 4$^{o}$C.

## 8. Cloning of pmVSG531, pmVSG1954 wildtype, and pmVSG1954 S321A

The inserts encoding mVSG531 and mVSG1954 were amplified from *T. brucei* gDNA (mVSG531 forward primer: 5'-CTACGCGACACGTACGATGATTCATAGGCTCACATTC-3', mVSG531 reverse primer: 5'-TTGAGGGGGGAAGTTAACTTAAAACAGCAAAAGTGCG-3', mVSG1954 forward primer: 5'-CTACGCGACACGTACGATGCTACAAAAGGTATTAACGTTG-3', mVSG1954 reverse primer: 5'-TTGAGGGGGGAAGTTAACCTAAAAAAGCAAGGCCGC-3') (Eurofins), using Phusion High Fidelity DNA Polymerase (New England Biosciences, NEB). PCR product was analyzed in 1% agarose DNA gel for purity. The band corresponding to mVSG531 and mVSG1954 insert was cut out and purified using Gel and PCR Clean-up Kit (Macherey Nagel). The pFAB6 backbone vector was generated by modifying the previously established pSY37F1D-CTR-BSD vector (Pinger et al., 2018) by replacing the blasticidin-resistance gene (BSD) by a hygromycin resistance gene (HYG) and the addition of an HpaI restriction site immediately downstream of the VSG2 CDS. Prior to ligation, the backbone pFAB6 was digested by restriction enzymes HpaI and BsiWI (NEB). The digestion reaction was analyzed in 0.8% agarose DNA gel and purified from the gel. The purified insert was ligated to the digested pFAB6 vector using In-Fusion Cloning Kit (Takara Bio). *E. coli* DH5$\alpha$ was transformed with ligation mix, plated on LB-Ampicilin plate, and incubated overnight at 37$^{o}$C. The surviving colonies were used to inoculate LB-Amp and grown with shaking at 37$^{o}$C overnight. The plasmids from overnight cultures were purified by NucleoSpin (No Lid) Kit for Plasmid DNA (Macherey Nagel). The purified plasmids were further confirmed by Sanger sequencing. pmVSG1954 S321A plasmid was generated by site directed mutagenesis using QuikChange Lightning Site directed Mutagenesis Kit (Agilent). The mutagenesis primers used to introduce S321A mutation are: forward primer: 5'-CTTGCCGCCATTAGCCCCGCCGATGCCG-3', reverse primer: 5'-CGGCATCGGCGGGGCTAATGGCGGCAAG-3'.

## 9. Generation of T. brucei cell line expressing mVSG531

pmVSG531 was linearized using BgIII restriction and purified by DNA ethanol precipitation. 10 μg of linearized pmVSG531 was used to transfect 3 x $10^7$ Lister427 BSF *T. brucei* cells using Amaxa Nucleofector (Lonza). Transfected cells were allowed to recover in 24-well cell culture dish at 37°C, 5% $CO_2$ for 6 hours before adding the selection drugs Hygromycin (25 μg/ml). Surviving clones were observed after 5 days. RNA from surviving clones were collected to produce cDNA using ProtoScript First Strand cDNA synthesis kit (NEB). To confirm the expression of mVSG531, PCR was performed on the cDNA (forward primer (splice leader): 5'-ACAGTTTCTGTACTATATT-3', reverse primer (3'-UTR): 5'-GATTTAGGTGACACTATAGTGTTAAAATATATC-3') (Eurofins). The PCR product was analyzed for purity on 1% agarose DNA gel and purified from the gel. The purified PCR products was further confirmed by Sanger sequencing.

## 10. Generation of T. brucei cell line expressing mVSG1954 wildtype and mVSG1954 S321A

pmVSG1954 wildtype and pmVSG1954 S321A were linearized using BgIII restriction and purified by DNA ethanol precipitation. 10 μg of linearized plasmids was used to transfect 3 x $10^7$ Lister427 2T1 BSF *T. brucei* cells using Amaxa Nucleofector (Lonza). Transfected cells were allowed to recover in 24-well cell culture dish at 37°C, 5% $CO_2$ for 6 hours before adding the selection drugs Hygromycin (25 μg/ml). Surviving clones were observed after 5 days. RNA from surviving clones were collected to produce cDNA using ProtoScript First Strand cDNA synthesis kit (NEB). To confirm the expression of mVSG1954 wildtype and mVSG1954 S321A, PCR was performed on the cDNA (forward primer (splice leader): 5'-ACAGTTTCTGTACTATATT-3', reverse primer (3'-UTR): 5'-GATTTAGGTGACACTATAGTGTTAAAATATATC-3') (Eurofins). The PCR product was analyzed for purity on 1% agarose DNA gel and purified from the gel. The purified PCR products was further confirmed by Sanger sequencing.

## 11. Protein purification of full-length mVSG531 and mVSG1954 (wildtype and S321A)

BSF *T. brucei* expressing mVSG531 or mVSG1954 was cultivated to 3.5 x $10^6$ cells/ml in HMI-9 media (supplemented with 10% FBS) at 37°C, 5% $CO_2$. The cells were harvested by centrifugation at 4,000 rpm for 20 minutes (TX-750 rotor, Thermo Scientific). The cells were subsequently homogenized and lysed in 0.2 mM $ZnCl_2$. The membrane fraction containing VSG was separated from the supernatant by centrifugation at 10,000xg for 10 minutes. VSG was solubilized by activating endogenous GPI-PLC in pre-warmed (42°C) SEC-Buffer at 37°C for 5 minutes. The soluble fraction containing VSG was separated from the membrane fraction by centrifugation at 10,000xg for 10 minutes. The soluble fractions were subjected for further purification by anion exchange chromatography using Q-Sepharose Fast Flow beads (GE Healthcare). The soluble fraction was incubated with the beads on rotating device for 30 minutes at 4°C. Flow through fraction containing VSG was collected upon separation of the soluble fraction from the beads. The beads were washed with SEC-Buffer. The flow through and wash fractions were concentrated using Amicon Stirred Cell protein concentrator (Millipore) with 10,000 Da MWCO membrane to 1 mg/ml. The proteins were further purified using gel filtration chromatography with HiLoad 16/600 Superdex 200 column equilibrated with SEC-Buffer in ÄKTA Pure FPLC system (GE Healthcare).

## 12. Limited proteolysis to produce N-terminal domain of mVSG1954

To generate N-terminal domain (NTD) of mVSG1954, the concentrated protein was subjected to limited proteolytic digestion using trypsin. The VSG at the concentration of 1 mg/ml was mixed with trypsin (5 mg/ml) (Sigma Aldrich) at 1:50 trypsin:VSG ratio and incubated for 3 hours on ice. The reaction was terminated by adding PMSF to 1 mM final concentration. The NTD is further purified using gel filtration chromatography with HiLoad 16/600 Superdex 200 column equilibrated with SEC-Buffer in ÄKTA Pure FPLC system (GE Healthcare).

## 13. Intact protein mass spectrometry for mVSG1954 wildtype and S321A

Protein samples were loaded on a self-packed reversed-phase column (0.8 × 2 mm, Poros R1) for desalting and concentration using 0.3% formic acid (0.3 ml/min). After 3 minutes proteins were eluted with 40% isopropanol, 5% acetonitrile, 0.3% formic acid (0.04ml/min) and analyzed with a QTof mass spectrometer (maXis, Bruker Daltonics) after electrospray ionization. Data were deconvoluted using the ESI Compass 1.3 Maximum Entropy Deconvolution Option (Bruker) to determine the molecular weight of the proteins.

## 14. Size exclusion chromatography with Multi Angle Light Scattering (SEC-MALS) for mVSG 1954

50 ul of full-length mVSG1954 or N-terminal domain of mVSG1954, at 1.0 mg/ml concentration (determined by nanodrop, using the predicted molecular weight and extinction coefficient of the protein) was injected onto a Superdex 200 10/300 GL gel-filtration column (GE Healthcare) in SEC-Buffer at room temperature. The column was coupled to a MALS system (MiniDAWN and Optilab, Wyatt Technology). Data were analysed using the Astra 7 software (Wyatt Technology). Measurements were performed in triplicates.

## 15. Generation of antisera in C57BL/6J against mVSG531 or mVSG1954

The infection procedure was adapted from Aresta-Branco et al., 2019. Antisera against BSF expressing mVSG531 or mVSG1954 was generated by infecting two C57BL/6J female mice. On day 0, two C57BL/6J female mice were injected (i.p.) by 1000 parasites (BSF expressing either mVSG531 or mVSG1954) suspended in 200 µl HMI-9 media. Berenil (1.25 μg/ml) was administered to both mice on day 4 and 5 post infection (i.p.). Both mice were sacrificed on day 8 by $CO_2$ asphyxiation. Blood were collected from both mice by cardiac puncture. The antisera were separated from the other cellular component by centrifugation in BD Microtainer blood collection tubes. The antisera were further confirmed by flow cytometry to stain the original cell line used for the infection.

## 16. Differentiation of PCF T. brucei into MCF

*T. brucei* PCF 29-13 cells were transfected by plasmid that encodes for tetracycline-inducible RBP6 overexpression (courtesy of Lucy Glover, Institute Pasteur, Paris, France). The cells were cultured in SDM-80 media (SDM-79 without glucose). SDM-80 media was supplemented with 7.5 μg/ml Hemin, 10% heat-inactivated FBS, and 50 mM N-Acetylglucosamine. On day 0, RBP6 overexpression was induced with 10 μg/μl tetracycline at 2 x $10^6$ cell density. The growth of both induced and uninduced cultures were examined and diluted to 2 x $10^6$ cells per ml. Tetracycline was added every day during differentiation. The occurrence of MCF was examined using flow cytometry on day 6 by detecting the loss of GPEET and EP expression.

## 17. Separation of undifferentiated PCF from the induced culture

Undifferentiated PCF cells were separated from the non-PCF cells by negative selection using magnetic beads. On day 6 post-induction of RBP6 overexpression. 1 x $10^8$ cells in induced culture were collected by centrifugation at 1800 xg for 10 minutes at 4°C (TX-750 rotor, Thermo Scientific). Remaining growth media was aspirated, and the cells were resuspended in 1 ml SDM-80 media. Monoclonal α-GPEET and α-EP IgG were used in 1:50 antibody to cells volume as the primary staining. The cells were stained for 15 minutes on ice. Following staining, 10 ml of SDM-80 was added to wash the primary α-GPEET and α-EP. The cells were further harvested by centrifugation at 1800 xg for 10 minutes at 4°C (TX-750 rotor, Thermo Scientific). The cells were further resuspended in 800 μl SDM-80 media. 20 μl of goat α-mouse IgG magnetic beads was added into the cell suspension and left to stain for 15 minutes on ice. 10 ml of SDM-80 was added to wash the magnetic beads. The cells were further harvested by centrifugation at 1800 xg for 10 minutes at 4°C (TX-750 rotor, Thermo Scientific). After the removal of the supernatant, the cells are suspended in 1 ml of SDM-80 media. To separate the undifferentiated PCF cells, the cell suspension was pass through LS-column placed in magnetic field, and flow through that contain the non-PCF cells was collected. The column was subsequently washed with 3 x 3 ml of SDM-80 media to wash off all the non-PCF cells. The bound PCF cells were eluted by pushing 5 ml of

SDM-80 through the LS-column after the column was removed from the magnetic field. Workflow of the magnetic separation process is depicted in Figure 7 below.

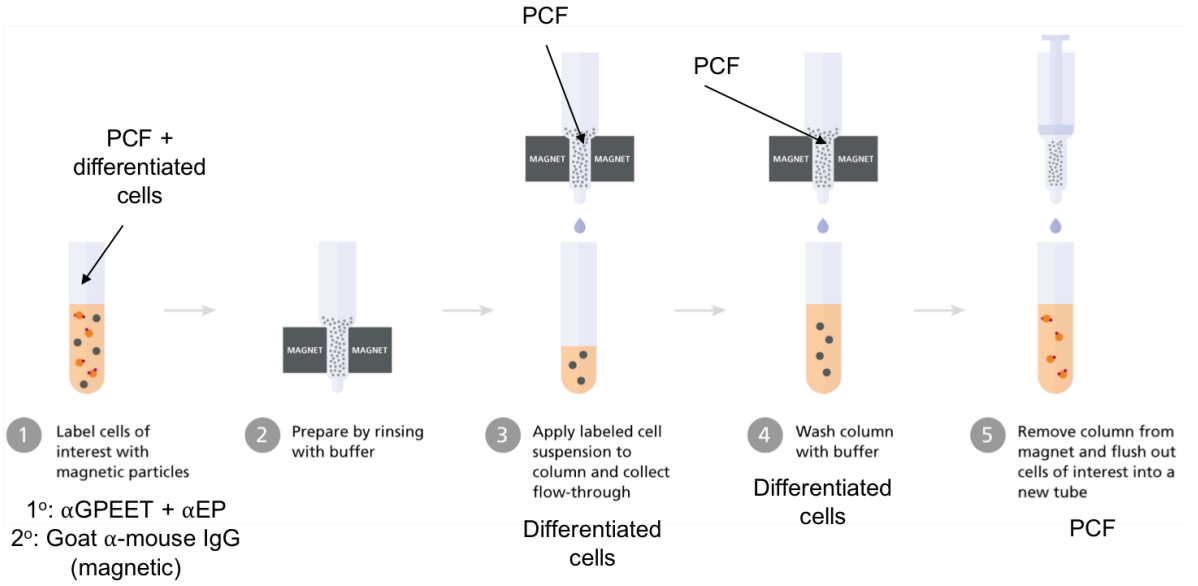


***Figure 7. Magnet-based cells separation used in separating GPEET and EP expressing cells from non-GPEET and EP expressing cells.***

### 18. Flow cytometry on T. brucei MCF

Flow cytometry was performed to detect MCF cells that express mVSG531 or mVSG1954 within the differentiated cells. Undifferentiated PCF cells and the non-purified differentiated were used as control to observed the number of differentiated cells. The cell density from both uninduced and tetracycline-induced culture were counted. Two million cells were used in each flow cytometry sample. The cells were harvested by centrifugation at 5,400 xg for 4 minutes at 4ºC. Monoclonal α-GPEET and α-EP IgG were conjugated with Allophycocyanin (APC) fluorescence molecule by APC Conjugation Kit – Lightning Link (Abcam). Cells were stained with α-GPEET or α-EP or both α-GPEET and α-EP to detect the loss of GPEET and EP as the indicator of the presence of non-PCF cells in the induced culture. The cells were subsequently resuspended in the SDM-80 medium with the respective antibody in 1:100 antibody to medium ratio (α-GPEET-APC, α-EP-APC, or both α-GPEET-APC and α-EP-APC). The samples were incubated on ice for 10 minutes. Washing step was done by adding 1 ml SDM-80 into each sample. The samples were subsequently centrifugated at 5,400

xg for 4 minutes at 4°C to remove the remaining media and staining solution. The cells were subsequently the resuspended in 250 μl SDM-80 and kept on ice prior to analysis. The cells were analyzed by BD-FACSCalibur Flow Cytometry. 50,000 cells were analyzed from each sample.

Two million undifferentiated PCF cells were mixed with 500,000 BSF cells expressing either mVSG531 or mVSG1954. These samples were used as a positive control that α-mVSG531 and α-mVSG1954 are able to recognize a small population of cells expressing mVSG531 or mVSG1954 in a mixture with other cells. Two million purified differentiated cells were also used in each flow cytometry sample. The differentiated cells were stained by α-mVSG531, α-mVSG1954, or both α-mVSG531 and α-mVSG1954. The cells were harvested by centrifugation at 5,400 xg for 4 minutes at 4°C. The cells were subsequently resuspended in the SDM-80 medium with the respective antibody in 1:100 antibody to medium ratio (α-mVSG531, α-mVSG1954, or both α-mVSG531 and α-mVSG1954). The samples were incubated on ice for 10 minutes. The samples were subsequently centrifugated at 5,400 xg for 4 minutes at 4°C to remove the remaining media and staining solution. The cells were resuspended in 200 μl SDM-80 medium with 1:1000 α-Mouse IgM-FITC as secondary antibody stain. The cells were incubated for 10 minutes on ice. The samples were subsequently centrifugated at 5,400 xg for 4 minutes at 4°C to remove the remaining media and staining solution. The cells were subsequently the resuspended in 250 μl SDM-80 and kept on ice prior to analysis. The cells were analyzed by BD-FACSCalibur Flow Cytometry. 50,000 cells were analyzed from each sample.

## Chapter III: Classification of T. brucei's VSGs based on their structural properties

### 1. Introduction

More diversity within VSG structures has been revealed as more structures were determined (Pinger et al., 2018; Zeelen et al., 2021). Therefore, the initial presumption that the VSGs were all very similar, differing mostly by surface properties due to amino acid divergence, is no longer accurate, especially with the discovery of critical immune-modulatory post-translational modifications like the *O*-linked glycan at the top of VSG3 (Pinger et al., 2018). Structure determination of more VSGs will give a better understanding of how diverse the VSG structures are and help better understand the mechanism of immune evasion in *T. brucei*. However, structure determination is a laborious process and therefore, a classification on VSG using bioinformatics analysis using structure-based prediction by protein threading was performed to decide which VSGs will be used in the structural study.

Structure-based prediction by protein threading online server called PHYRE2 was used in this section. In contrast to homology-based structure prediction, protein threading aligns the given sequence into different structure templates that are available in the PDB database. Subsequently, based on the alignment with the suitable template, a predicted model of the given protein sequence will be built. Protein threading is a very useful method to predict protein structures that only have distant homologs with low sequence identity. Not only predicting the tertiary structure, PHYRE2 also does secondary structure prediction and domain analysis of the submitted protein sequence. (Kelley et al., 2015)

### 2. Results

#### a. Structure-based prediction protein threading

According to Cross, 2014, the VSGs are categorized grossly into two classes, class A and B. Combining their classification system with the classification of the N-terminal domain from Marcello and Barry, class A is divided into three subclasses (A1, A2, and A3), while class B is divided into two subclasses (B1 and B2) (Cross et al.,

2014; Marcello and Barry, 2007). When the study was started, there were only four VSGs structures available, VSG2, IlTat 1.24, VSG1, and VSG3 (Bartossek et al., 2017; Blum and Turner, 1993; Freymann et al., 1990; Pinger et al., 2018). Prior to the VSG3 structure determination, VSG3 did not thread against VSG2 in PHYRE2, showing a clear structure divergence between the two proteins. The VSG3 structure eventually showed a topological difference in arrangement of the polypeptide chain in the tertiary structure (Figure 8).

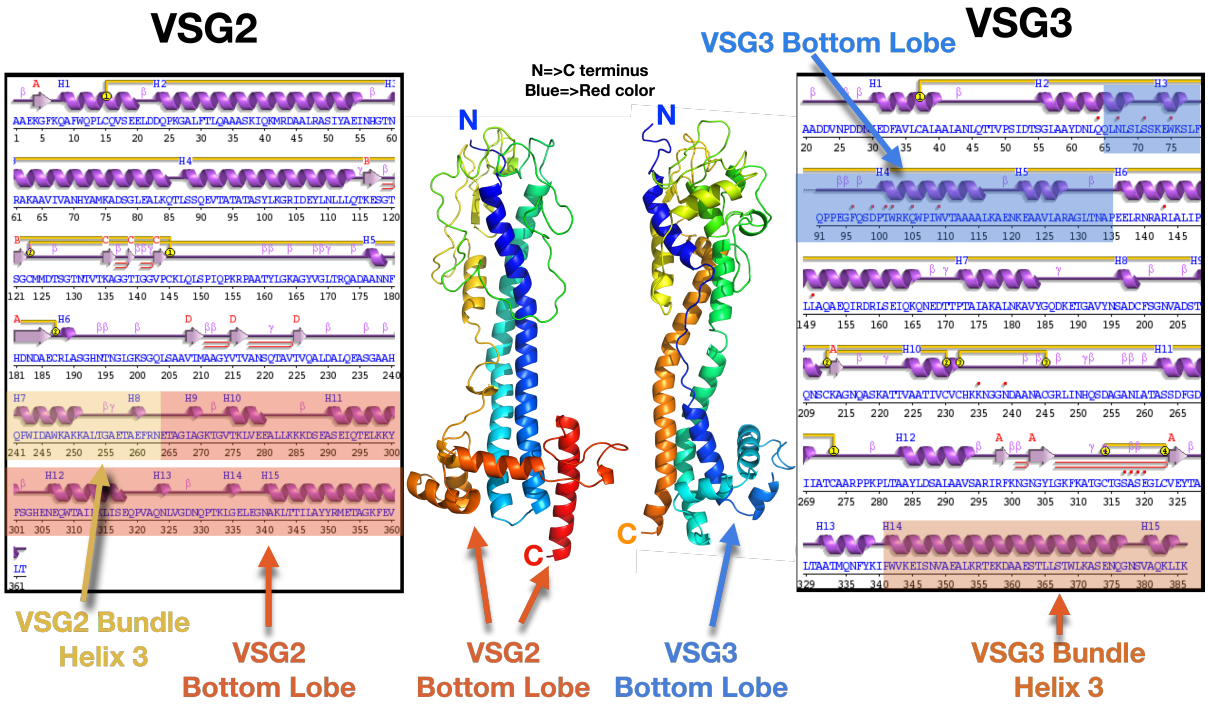


***Figure 8.*** ***Atomic resolution structures of VSG2 and VSG3 reveals difference in the organization of amino acid sequences in the tertiary structure.***

Shown above is the monomers of both VSG2 (PDB code: 1VSG) and VSG3 structures (PDB code: 6ELC). Both structures reveal that VSG2's bottom lobe structure comprises the later amino acids in the primary sequence (left) whereas those of VSG3 are comprised of those from the N-terminus. (Pinger et al., 2018)

386 functional VSGs from Lister427 described in Cross, 2014 were analyzed by PHYRE2 (Kelley et al., 2015). 201 VSGs scored highly only against VSG2 (PDB code: 1VSG) and VSG1 structures (which scored high against VSG2, PDB code: 5LY9), while the other 153 VSGs scored highly only against VSG3 structure (PDB code; 6ELC). Only 32 VSGs were shown to score highly against both VSG2 and VSG3 structures (Table 1). The result from PHYRE2 suggests a clear structural divergence

between the two groups, supporting the broad characterization of the VSGs into two primary classes. Within the functional VSGs that were analyzed, five of them were annotated as VSGs that are expressed in MCF cells (mVSGs), namely mVSG397, 531, 639, 653, and 1954 (Cross et al., 2014). These mVSGs are expressed from a dedicated expression site called the metacyclic expression site (MES) by RNAP I (Günzl et al., 2003). In contrast to VSG transcripts in BSF, mVSG transcript is monocistronic (Alarcon et al., 1994; Graham and Barry, 1995). mVSG397, 531, 639, and 653 only scored highly against VSG2 and VSG1 (Table 1). Based on the classification described in Cross, 2014, mVSG 397 is classified in subclass A1, and mVSG531, 639, and 653 are in subclass A2 (Cross et al., 2014). mVSG1954 is the only mVSG that scored highly against the VSG3 structure (Table 1). According to the classification in Cross, 2014, mVSG1954 is classified in subclass B2 (Cross et al., 2014). To date, there is no structural information available on mVSGs. Thus, mVSG531 and mVSG1954 were chosen as members of each VSG subclass to be used in structural study using X-ray crystallography to assess the similarities and differences between VSGs expressed in BSF and MCF cells.

| Types of VSGs | Subclasses described in Cross, 2014 | | | | |
|---|---|---|---|---|---|
| | **A1** | **A2** | **A3** | **B1** | **B2** |
| **Bloodstream VSGs** | 92 | 105 | 32 | 77 | 75 |
| **Metacyclic VSGs** | 1 | 3 | | | 1 |
| mVSG397 | √ | | | | |
| mVSG531 | | √ | | | |
| mVSG639 | | √ | | | |
| mVSG653 | | √ | | | |
| mVSG1954 | | | | | √ |

***Table 1.*** ***Classifications of functional VSGs (bloodstream VSGs and mVSGs) according to Cross, 2014.***

386 VSGs sequences were analyzed by PHYRE2 (Kelley et al., 2015). 201 VSG sequences scored high only against VSG2 and VSG structures. Four of the 201 are mVSGs. These VSGs are classified in subclass A1 and A2 according to the classification in Cross, 2014. 32 VSG sequences scored high against both VSG2 and VSG1 family and VSG3 structures. These VSGs are classified in subclass A3. The remaining VSGs sequences, which scored high against VSG3 structure only, are classified in Class B according to the classification in Cross, 2014. mVSG397, 531, 639, and 653 are classified in Class A. mVSG1954 is the only mVSG that is classified in Class B. (Cross et al., 2014)

## Chapter IV: mVSG531 structure determination

### 1. Introduction

Among the four mVSGs included in the Class-A VSGs, mVSG531 was chosen to be used in the structural study by X-ray crystallography as a likely member of the VSG1/2 subclass (Class A). A large amount of pure mVSG531 soluble protein was needed to perform an X-ray crystallography project. However, there are some challenges to obtain purified mVSG531 from MCF cells. MCF cells can be obtained by infection of the tsetse fly with the PCF cells and the protein eventually harvested from the MCF cells from the infected fly's salivary gland. However, a large amount of tsetse fly is needed to get a sufficient protein amount because MCF is cell-cycle arrested and cannot be multiplied in culture (Christiano et al., 2017). Also, the type of expressed mVSG is not homogeneous, which will complicate the downstream purification process (metacyclic VSGs co-express up to five different VSGs in the Lister427 strain (Christiano et al., 2017)). Recent literature shows that MCF cells can also be obtained by overexpressing RNA binding protein 6 (RBP6) (Kolev et al., 2012). Nevertheless, the differentiation process is not efficient to obtain the large amount of MCF cells needed to produce a large amount of protein required for an x-ray crystallography experiment.

To overcome these challenges, mVSG531 was expressed in BSF cells because it is quickly grown in culture and has rapid doubling time. BSF cells were transfected with a linearized plasmid that encoded the mVSG531 and hygromycin resistance gene sequences (see Chapter II for details). Therefore, the coding sequence of mVSG531 was integrated into the active bloodstream expression site (BES), and the clones that express mVSG531 were selected by supplying the growth media with hygromycin. Five days after the transfection, the surviving clones were collected and grown to obtain sufficient RNA. The expression of mVSG531 was further confirmed by RT-PCR using primers that recognize the splice leader and 3’-UTR regions. The PCR products were subsequently sequenced by Sanger sequencing.

## 2. Results

### a. Purification of mVSG531

BSF cells expressing mVSG531 (BSF mVSG531) were grown to high density (2.5 – 3.5 million cells per ml) in a large amount of *T. brucei* growth media HMI-9, supplemented with 10% FBS and L-cysteine. The cells were subsequently harvested to purify mVSG531. Samples from each purification step were collected to be analyzed in SDS-PAGE (Figure 9A). The first step of the purification is cell lysis by osmotic shock using low concentration of $ZnCl_2$. The presence of $ZnCl_2$ inhibits phospholipase C's activity, which cleaves VSG from the membrance (Cross, 1984). The insoluble fraction containing the membrane with the membrane proteins was separated from the soluble fraction containing the intracellular protein (Figure 9A, SN1). The release of membrane-bound VSG proteins was done by activating phospholipase C that cuts on the GPI-moiety anchoring VSG to the membrane (Figure 9A, SN2, and SN3). Anion exchange chromatography was further done by incubating the fractions that contain the soluble VSG proteins with Q-Sepharose beads to remove the remaining contaminants. Due to VSGs' anionic property, the soluble mVSG531 does not bind to the beads and is retained in the flow-through and washes fractions (Figure 9A, FT and W). The predicted molecular weight from the full-length protein is 51.5 kDa (without counting in the N-linked glycans' molecular weight) using Expasy ProtParam (Walker, 2005). However, as we see on the Coomassie-stained gel, the full-length mVSG531 migrated with electrophoretic mobility in between the 55 and 70 kDa molecular weight marker protein, which is bigger than the theoretical molecular weight for mVSG531 (Figure 9A, FT and W). The difference in the molecular weight could be due to N-linked glycans observed in most VSG structures up to this date. The protein was further subjected to an additional purification step using size exclusion chromatography. Full-length mVSG531 eluted approximately at the retention volume of 67.8 ml from HiLoad 16/600 Superdex 200 column. The retention volume for mVSG531 was used to calculate the molecular weight of mVSG531 based on the linear equation obtained from the linear regression of the logarithmic values of the standard proteins' molecular weight their retention volume on HiLoad 16/600 Superdex 200 column (Figure 9B). The estimated molecular weight of 135 kDa corresponds to either 2 molecules of full-length mVSG531. The existence of mVSG531 as a dimer is consistent with the

previous data on VSG2, which also belongs to Class-A VSG, it is more probable that mVSG531 eluted as a dimer.

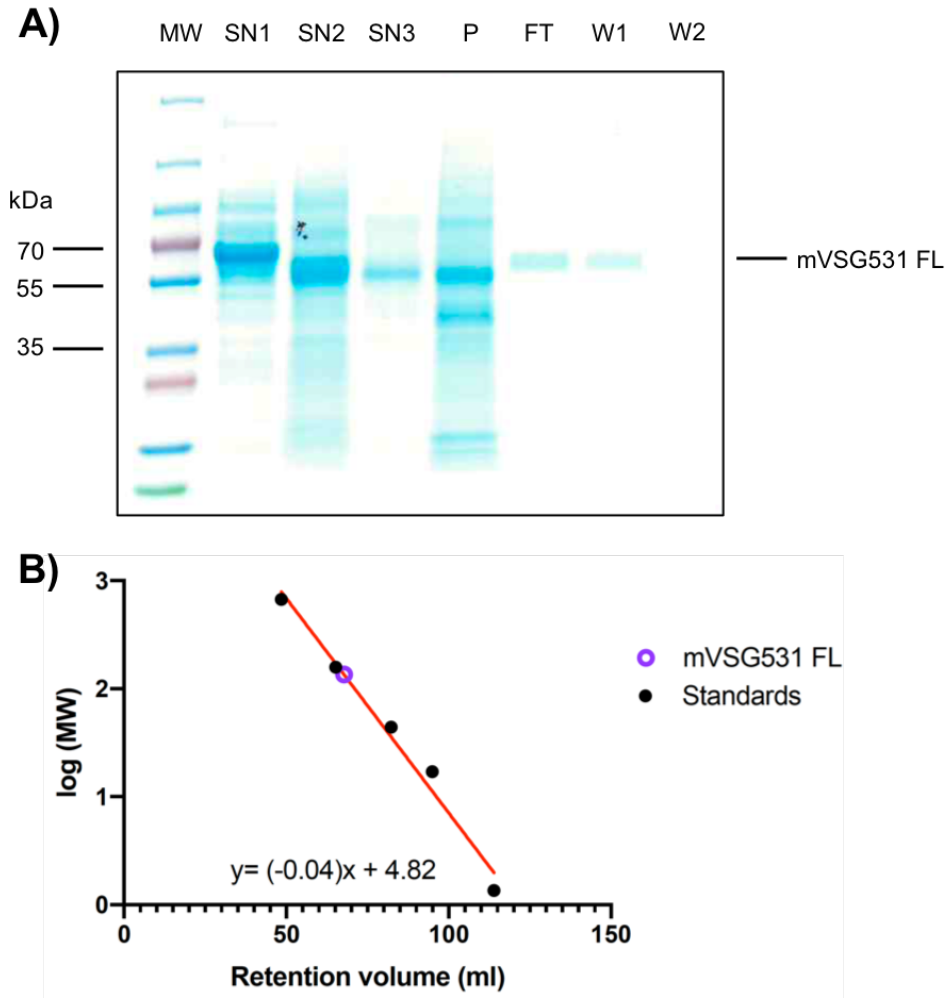


***Figure 9.* mVSG531 was purified endogenously as soluble protein from T. brucei and forms a dimer in solution.**

**A)** Aliquots from each purification step of mVSG531 were collected and analyzed in 4-20% gradient SDS-PAGE. The gel was stained with Coomassie Blue stain. BSF mVSG531 was lysed by osmotic shock in 0.2 mM $ZnCl_2$. The soluble fraction containing the intracellular protein (SN1) was separated from the membrane fraction. The GPI-bound mVSGs on the membrane was release by activating phospholipase C at 37°C (SN2 and SN3). The remaining membrane fraction without mVSGs bound was discarded (P). The fractions containing the soluble mVSGs were incubated with Q-sepharose anion exchange beads to remove the contaminant proteins. The soluble mVSG is retained in the flow through (FT) and wash fractions (W1, W2). Purified soluble full-length mVSG531 migrated with the electrophoretic mobility in between the 70 and 55 kDa molecular weight marker (MW). **B)** The retention volume of mVSG531 was used to estimate the molecular weight of the protein, based on the linear equation obtained from the linear regression of molecular weights of standard proteins and their retention volumes on HiLoad 16/600 Superdex 200 FPLC column. The calculated molecular weight of mVSG531 is 135 kDa.

### b. Crystallization of purified mVSG531 protein

The purified soluble mVSG531 was crystallized at a concentration of 7 mg/ml. Firstly, the initial crystallization condition was determined by performing a sparse matrix screen in vapor diffusion sitting drop 96-well plate format. 384 different conditions were tried in the sparse matrix screen using commercial screens Wizard I, II, III, IV (Jena Biosciences), ProComplex (QIAGEN), and Morpheus (Molecular Dynamics). The sparse matrix screens were incubated in a 22$^{o}$C incubator. Several initial crystallization conditions that enabled crystal formation were found from the sparse matrix screen. One of the most promising crystals was grown in 25% PEG 1500 and 0.1 M MMT Buffer pH 6.5. MMT buffer is a combination between three different compounds, which was malic acid, MES, and Tris, in a 1:2:2 ratio, respectively. The combination of three different compounds with different buffering capacities ensured pH stability over a large range, enabling exploring the effect of pH in the crystallization experiment. The crystals found in the initial condition consisted of thin plates and small thin rod-like crystals overlapping with other crystals (Figure 10A). It was challenging to manipulate these crystals without the risk of destroying the crystals and challenging to harvest a single crystal needed in X-ray diffraction data collection.

Therefore, the initial condition was further optimized. The optimization was done by replicating the initial condition and creating a gradient of different concentrations of the precipitating agent (PEG 1500) and different pH of the MMT buffer in vapor diffusion hanging drop 24-well plate format to get a harvestable-sized single crystal. After three weeks of incubation at 22$^{o}$C, single harvestable-sized crystals appeared in the initial crystallization condition replicated in the hanging drop 24-well plate format (Figure 10B). However, the crystals' surface has some imperfections; thus, more optimization was performed by incorporating several different additive compounds that can alter crystal growth. The final optimized crystallization condition for mVSG531 was 25% PEG 1,500, 0.1M MMT buffer pH 6.5, with 3% glucose as additive (Figure 10C).

Due to the long time needed for crystal formation, the crystals were further analyzed in SDS-PAGE to observe whether the crystallized species is still the full-length mVSG531 or if degradation had taken place. After three weeks of incubation in the drops, it was found that the full-length mVSG531 had been proteolyzed during the

crystallization process, as shown in the Coomassie-stained SDS-PAGE (Figure 10D). The crystallized species migrated with the electrophoretic mobility between the 35 and 55 kDa molecular weight markers. The crystallized species were smaller in molecular weight when compared to the full-length mVSG531, which is approximately 55 kDa in size (Figure 9A, FT and W). It is suspected that the more disordered C-terminal domain was cleaved during the crystallization process by remaining proteases, leaving the more structured N-terminal domain (as has been seen previously with other VSGs) (Pinger et al., 2018).

The crystals were preserved by freezing it in liquid nitrogen to transport for X-ray data collection in a synchrotron radiation facility. For freezing crystals were harvested using a nylon cryo-loop (Crystal Cap SPINE HT, Hampton Research) and immediately transferred into premixed stabilization buffer (28% PEG 1500, 0.1 MMT Buffer pH 6.5) that contained 15% glycerol as a cryoprotectant to avoid the formation of ice crystals during crystal freezing in liquid nitrogen. After soaking the crystals in the stabilization buffer for 30 seconds, the crystals were transferred into a new nylon cryo-loop and immediately flash-freeze in liquid nitrogen. Such crystals are referred to as native crystals in the subsequent text. The crystals were shipped to the destination in a special metal container prechilled by liquid nitrogen.

X-ray data collection was performed in European Synchrotron Radiation Facility (ESRF), Grenoble, France. The diffraction data on the native crystal were collected at 1.0 Å wavelength. Initial crystal indexing and data processing by the synchrotron processing software EDNA reveals the crystals formed in space group P1211 with a= 50.1 Å, b= 170.9 Å, c= 163.7 Å and $\alpha$= 90$^{o}$, $\beta$= 90.7$^{o}$, $\gamma$= 90$^{o}$. The best crystal diffracted to 1.96 Å resolution.

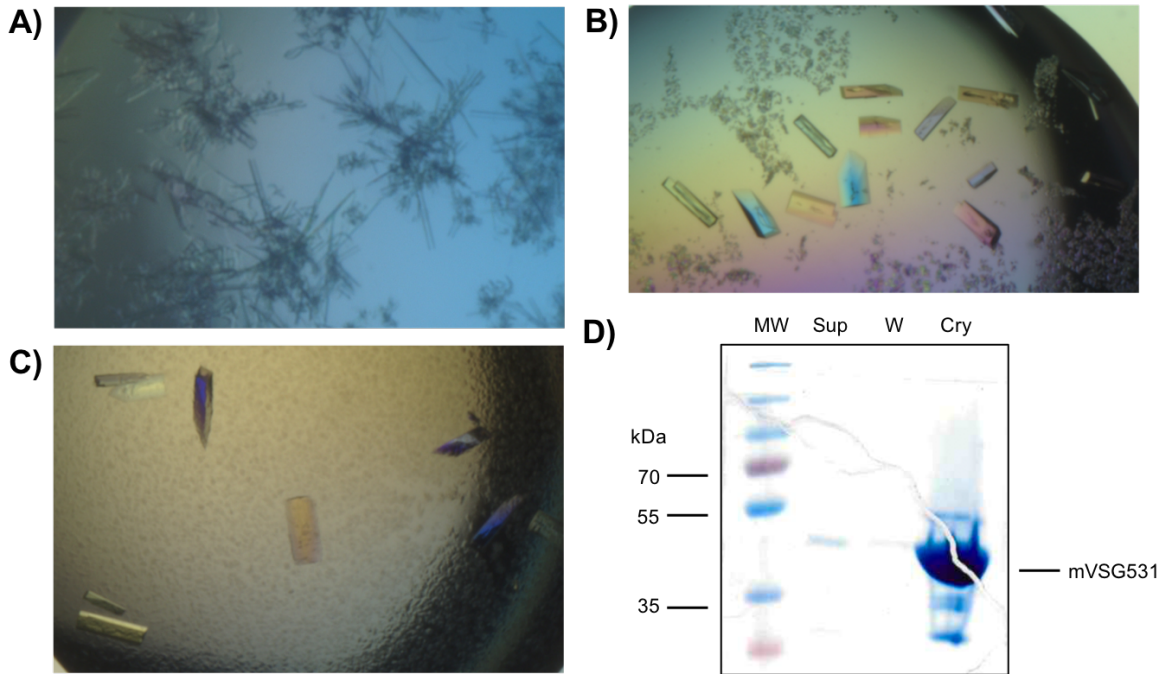


***Figure 10.*** ***Crystallization of mVSG531.***

**A)** Initial crystals obtained from sparse matrix screen in vapor diffusion sitting drop 96-well plate format. The crystallization condition was 25% PEG 1500, 0.1 M MMT Buffer pH 6.5. **B)** Crystals obtained in the initial condition which was replicated in vapor diffusion hanging drop 24-well plate format. **C)** Crystals obtained after optimization by including 3% glucose in 25% PEG 1500, 0.1M MMT Buffer pH 6.5. **D)** The crystallized protein in the drop was analyzed in SDS-PAGE. The crystallized mVSG531 migrated with the electrophoretic mobility in between the 35 and 55 kDa molecular weight marker (Cry). The same protein species was also found in the drop solution (Sup) and the crystal washing step (W).

### c. Molecular replacement to solve mVSG531 atomic resolution structure

The mVSG531 structure was determined by molecular replacement, using the structure of VSG1 (PDB code: 5LY9) as the search model (Bartossek et al., 2017). VSG1 shares 35% sequence identity with mVSG531. The space group of mVSG531 crystal is $P12_11$ (see Appendix 1). Calculations based on average protein density in macromolecular crystals estimated that there were between six and eight molecules in the crystal asymmetric unit (Table 2) (Kantardjieff and Rupp, 2003; Matthews, 1968; Weichenberger and Rupp, 2014). Phaser-MR performed molecular replacement in the PHENIX crystallography software suite, and eight molecules of mVSG531 were

identified in the crystal asymmetric unit, forming four classical VSG dimers (analogous to those seen in VSG1 and VSG2) (Figure 11A). The dimeric structure shown in the crystal structure corroborates the molecular weight estimation using size exclusion chromatography that mVSG531 exists as a dimer in solution. Model building was performed with AutoBuild, an automatic model building and initial structure refinement software integrated into PHENIX. Subsequent refinement was performed in phenix.refine, also included in PHENIX (Adams et al., 2010), followed by repeated cycles of manual model examination and improvement (using COOT, (Emsley and Cowtan, 2004)) and automated refinement.

| N (mol) | Prob (N) for resolution | Prob (N) for overall | Vm A**3/Da | Vs % solvent | Mw (Da) |
|---|---|---|---|---|---|
| 1 | 0.0000 | 0.0000 | 17.52 | 92.98 | 40000 |
| 2 | 0.0004 | 0.0002 | 8.76 | 85.96 | 80000 |
| 3 | 0.0015 | 0.0122 | 5.84 | 78.94 | 120000 |
| 4 | 0.0112 | 0.0427 | 4.38 | 71.92 | 160000 |
| 5 | 0.0362 | 0.0912 | 3.50 | 64.89 | 200000 |
| 6 | 0.1391 | 0.1963 | 2.92 | 57.87 | 240000 |
| 7 | 0.2935 | 0.2948 | 2.50 | 50.85 | 280000 |
| 8 | 0.3550 | 0.2606 | 2.19 | 43.83 | 320000 |
| 9 | 0.1378 | 0.0847 | 1.95 | 36.81 | 360000 |
| 10 | 0.0216 | 0.0134 | 1.75 | 29.79 | 400000 |
| 11 | 0.0038 | 0.0022 | 1.59 | 22.77 | 440000 |
| 12 | 0.0000 | 0.0000 | 1.46 | 15.75 | 480000 |

***Table 2.*** ***Prediction of number of molecules in the crystal asymmetric unit using Matthew's coefficient.***

N: number of molecules, Prob: probability, Vm: Matthew's coefficient, Vs: solvent percentage, Mw: molecular weight. (Kantardjieff and Rupp, 2003; Matthews, 1968; Weichenberger and Rupp, 2014)

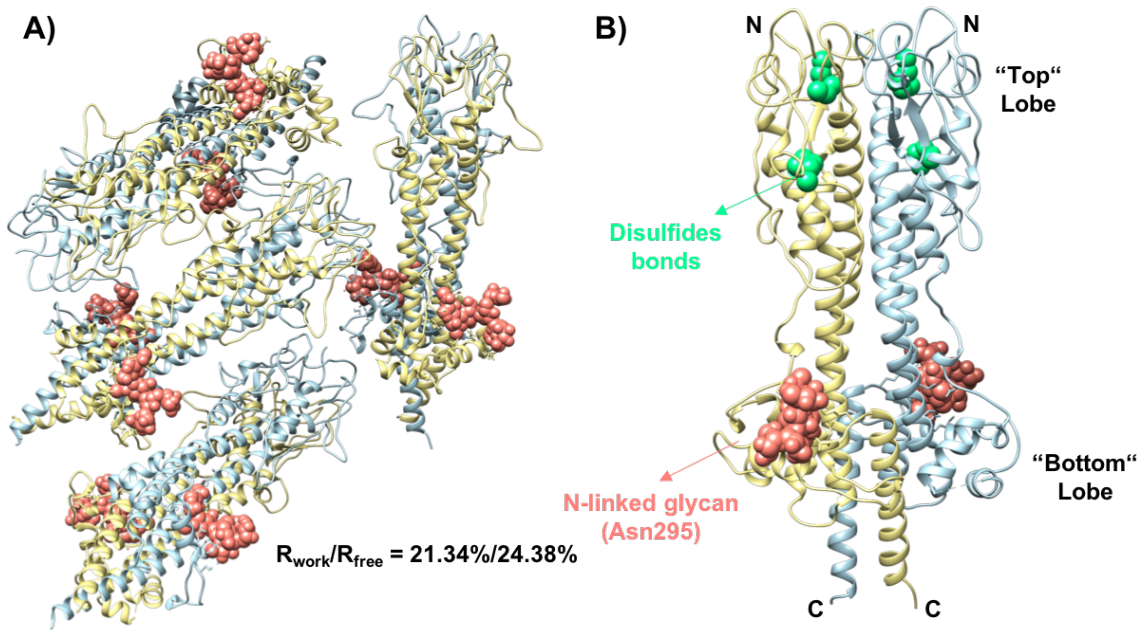


***Figure 11.*** ***Atomic resolution structures of mVSG531.***

Molecular replacement was performed by PHASER-MR on the native crystal dataset using BSF VSG1 as search model to determine the structure of mVSG531. **A)** Eight molecules of mVSG531 was found in the crystal asymmetric unit. Model building was performed by AutoBuild, an automated model building and initial structure refinement software. Subsequent structure refinement was performed by phenix.refine. PHASER-MR, AutoBuild, and phenix.refine are integrated in the PHENIX crystallography software suite (Adams et al., 2010). The final resolution for mVSG531 structure is 1.95 Å. Final $R_{work}/R_{free}$ for mVSG531 structure is 21.34%/24.38% (definition on R-values, see Chapter I, Part 5). **B)** N-linked glycan is attached at Asn295 at the dimer's bottom lobe (salmon pink spheres). Two disulfide bonds are observed at the top lobe of each molecule (green spheres). For a more detailed data collection and final model refinement statistics, see Appendix 1. Images of protein structures were generated and edited using CHIMERA (Pettersen et al., 2004).

### d. Analysis of mVSG531 structure

mVSG531 scored the highest against VSG2 and VSG1 based on the structure-based prediction algorithm PHYRE2 (Kelley et al., 2015). N-linked glycan is observed at the dimer's bottom lobe, attached to the residue Asn298 (Figure 11B). Like VSG1 and VSG2, the lower lobe of the mVSG531 structure comprises the later part of the amino acids in the polypeptide chain. Four cysteines forming two disulfide bonds were observed at the structure's top lobe (Figure 11B), alike VSG2 and VSG1 (Figure 5). To further confirm the similarities between the three structure, structure alignment was

performed on the overall structures and only the three-helix bundle core (Figure 12A) and the bottom lobe (Figure 12B). Structure alignment was performed by DeepAlign in RaptorX structure alignment server (Wang et al., 2013, 2011). Different parameters are used to assess how well the structures align with each other, which are root mean square difference (RMSD), unnormalized and normalized global distance test (uGDT/GDT), and template modeling score (TM-score). RMSD is the most used parameter to assess structure alignment, and it measures the deviation between the equivalent atom pairs. uGDT/GDT measures the number of C$\alpha$ set that have distance cutoffs (Xu and Zhang, 2010). The normalized GDT is uGDT score divided by the length of the protein and multiplied by 100. Related proteins usually have more than 50 in the values uGDT or GDT (Zemla, 2003). TM-score measures the topological similarity between the structures. A TM-score more than 0.6 indicates similarity in protein fold between the structures (Xu and Zhang, 2010). Considering the values of the above-mentioned parameters in the overall structures' alignment, it is concluded that the mVSG531, VSG2 and VSG1 structures aligned well to each other and share a high degree of similarity (Figure 12A). The similarity was further confirmed by the parameters' scores from the individual alignment of different parts of the structures (three-helix bundle core and bottom lobe) (Figure 12B).

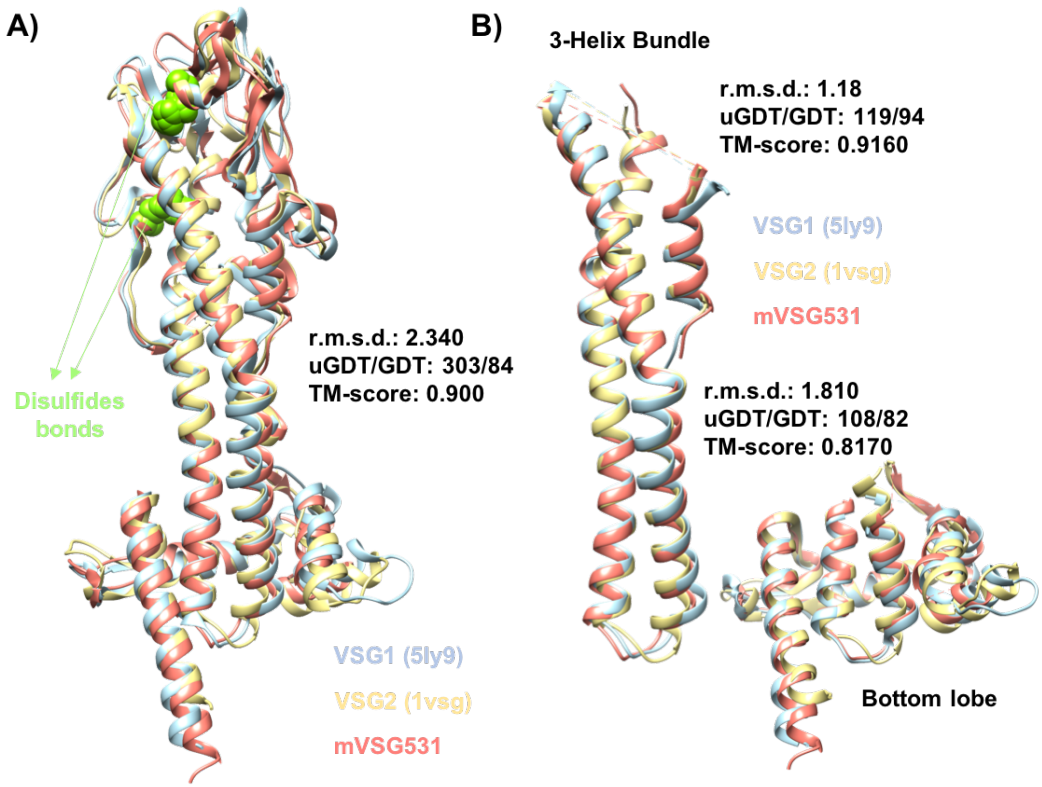


***Figure 12.*** ***Structure alignment between mVSG531, VSG2, and VSG1 reveals structure similarity between mVSG531 and bloodstream VSGs in Class A.***

**A)** Overall structure alignment of mVSG531, VSG2 (PDB code: 1VSG), and VSG1 (PDB code: 5LY9) monomers (yellow: VSG2, blue: VSG1, salmon pink: mVSG531, green: disulfide bonds) **B)** Three-helix bundle core **(left)** and the bottom lobe **(right)** structures alignment. Structure alignment was performed by DeepAlign in RaptorX structure alignment server. (Wang et al., 2013, 2011) (r.m.s.d.: root mean square difference, uGDT/GDT: (unnormalized) global distance test, TM-score: template modeling score). Images of protein structures were generated and edited using CHIMERA (Pettersen et al., 2004).

Despite the identified similarities between mVSG531 and VSG2, and VSG1, some differences were observed. When the surface of mVSG531, VSG2, and VSG1 structures was visualized based on their electrostatic potential (Figure 13 second row) and hydrophobicity (Figure 13, third row), stark differences were observed between the molecular surfaces of three different proteins. The top surface was thought as the primary host-parasite interaction site. However, recent literature has challenged the idea by showing that small nanobodies produced against *T. brucei* coat with VSG2 are able to bind closer to the VSG2's bottom lobe (Hempelmann et al., 2021). Besides, calcium metal binding was also observed in VSG2. The removal of calcium-binding by mutation of the residues that coordinate the metal binding has led to differences in host

antibody response in mice (Gkeka and Aresta-Branco, 2021, in preparation). No metal binding has been observed in the mVSG531 structure. The data above suggests that the three proteins are recognized differently by the host immune system, despite their overall tertiary structure similarities.

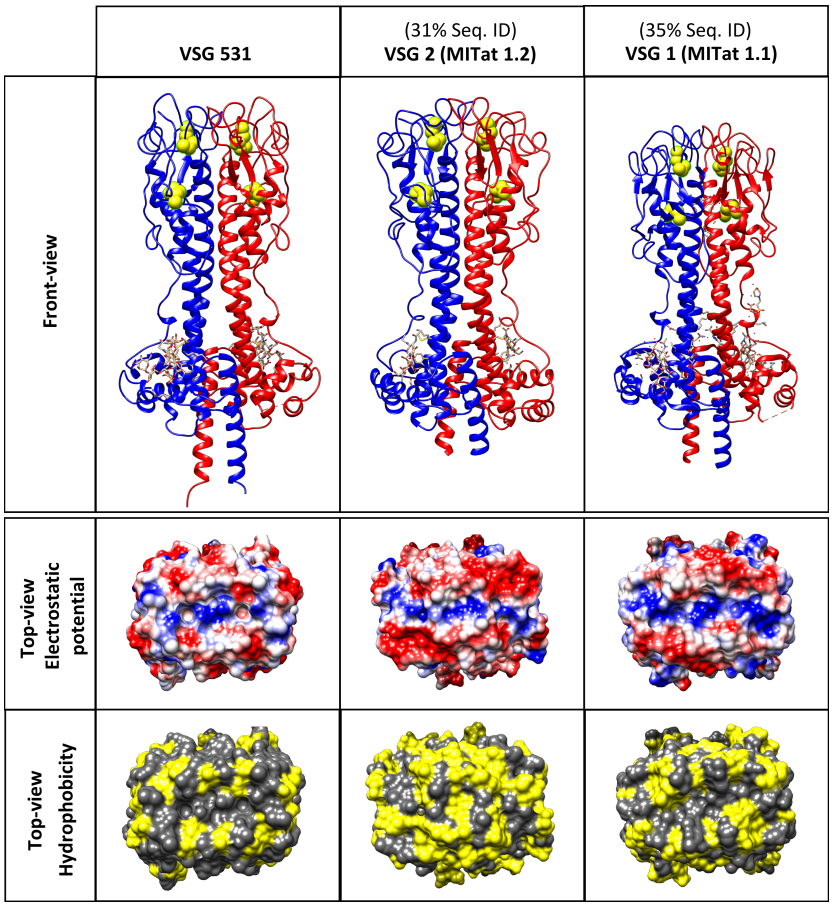


***Figure 13.*** ***The mVSG531's surface properties are different compared to VSG2 and VSG1 amidst their tertiary structure similarity.***

mVSG531 is compared to VSG 2 (PDB code: 1VSG) and VSG 1 (PDB code: 5LY9), which share above 30% sequence identity. From the front-view (top-row), all VSGs forms a dimer, has two disulfides bonds at the dimer's top lobe, which are positioned at the same location, and N-glycans towards the bottom of the molecule. The top surfaces of different VSGs are colored based on the electrostatic potential and the hydrophobicity of the surface amino acids. Images of protein structures were generated and edited using CHIMERA (Pettersen et al., 2004).

## Chapter V: mVSG1954 structure determination

### 1. Introduction

From the five mVSGs that have been annotated in *T. brucei* strain Lister427, mVSG1954 is the only mVSG that scored highly against VSG3 in PHYRE2 (Table 1) (Kelley et al., 2015). Therefore, mVSG1954 was chosen to be used in the structural study using X-ray crystallography to get the structure representative from the second VSG class (Class B). mVSG1954 was also expressed by the same technique as mVSG531 full-length expression in BSF cells to get enough protein to perform X-ray crystallography.

### 2. Results

#### a. Purification of mVSG1954

BSF mVSG1954 were grown to high density in a large volume of *T. brucei* growth media HMI-9, supplemented with 10% FBS and L-cysteine to get a sufficient amount of protein. The cells were subsequently harvested and lysed by osmotic shock. The insoluble fraction that contains the membrane with the attached membrane protein was separated from the soluble fraction that contains the intracellular protein (Figure 14A, SN1). The release of membrane-bound VSG proteins was done by activating phospholipase C that cuts on the GPI-moiety anchoring VSG to the membrane (Figure 14A, SN2 and SN3). Anion exchange chromatography was further done by incubating the fractions that contain the soluble VSG proteins with Q-sepharose beads to remove the remaining contaminants. Due to VSGs' anionic property, the soluble mVSG1954 does not bind to the beads and is retained in the flow-through and washes fractions (Figure 14A, FT and W). Expasy ProtParam was used to calculate the theoretical molecular weight of full-length mVSG1954 based on its primary amino acid sequence (Walker, 2005). The calculated theoretical molecular weight is 53.9 kDa.

However, as we see on the Coomassie-stained gel, the full-length mVSG1954 migrated with electrophoretic mobility between the 55 and 70 kDa molecular weight marker proteins is larger than the theoretical molecular weight (Figure 14A). The difference in the molecular weight could be caused by N-linked glycans that have been

observed in most VSG structures up to this date. The full-length mVSG1954 was further subjected to an additional purification step using size exclusion chromatography. Full-length mVSG1954 eluted approximately at the retention volume of 72.94 ml from HiLoad 16/600 Superdex 200 column. The retention volume for mVSG1954 was used to calculate the molecular weight of mVSG1954 based on the linear equation obtained from the linear regression of the logarithmic values of the standard's molecular weight proteins and their retention volume on the HiLoad 16/600 Superdex 200 column. The estimated molecular weight of 84.4 kDa (Figure 14B). However, the predicted molecular weight is ambiguous, as it is smaller than the molecular weight of two molecules and bigger than one molecule of full-length mVSG1954. This discrepancy in the molecular weight can be caused by the non-globular shape of VSG proteins that have been observed in the previously determined VSG structures. Protein size determination by size exclusion chromatography relies on assuming that proteins with the same hydrodynamic radius elute at the same retention volume. All the standard proteins used to establish the linear equation to estimate the molecular weight are globular proteins. Therefore, the equation cannot be applied to properly determine the molecular weight of a non-globular protein, such as VSG proteins.

Additional size exclusion chromatography coupled with multi-angle light scattering (SEC-MALS) experiment was performed to determine mVSG1954's molecular weight more accurately (Table 3). Dr. Karine Lapouge performed the SEC-MALS experiments at the Protein Expression and Purification at European Molecular Biology Laboratory in Heidelberg, Germany. The SEC-MALS data shows that the molecular weight of full-length mVSG1954 is 55.1 kDa, similar to the theoretical molecular weight of full-length mVSG1954. Therefore, it is concluded that full-length mVSG1954 exists as a monomer in solution.

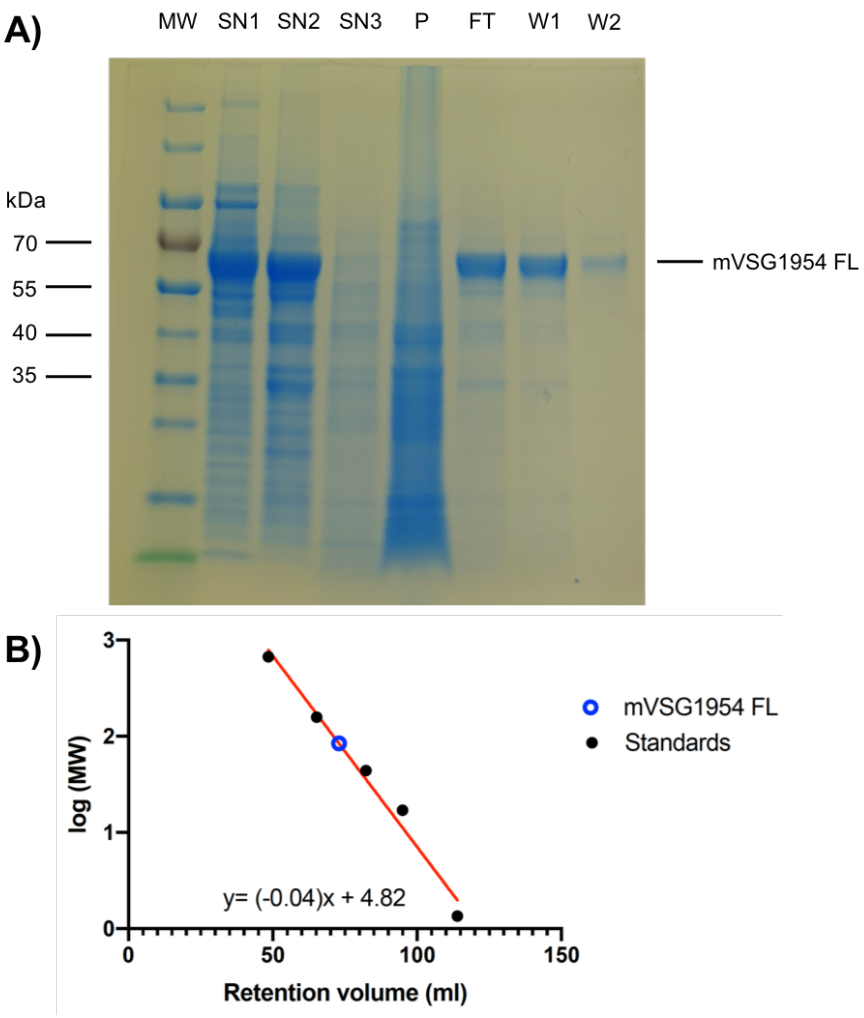


***Figure 14.*** ***mVSG1954 was purified endogenously as soluble protein from T. brucei and exists as a monomer in solution.***

**A)** Aliquots from each purification step of mVSG1954 were collected and analyzed in 4-20% SDS-PAGE. The gel was stained with Coomassie Blue stain. BSF cells expressing mVSG1954 was lysed by osmotic shock in 0.2 mM $ZnCl_2$. The soluble fraction containing the intracellular protein (SN1) was separated from the membrane fraction. The GPI-bound mVSGs on the membrane was release by activating phospholipase C at 37°C (SN2 and SN3). The remaining membrane fraction without mVSGs bound was discarded (P). The fractions containing the soluble mVSGs were incubated with Q-sepharose anion exchange beads to remove the contaminant proteins. The soluble mVSG is retained in the flow through (FT) and washes fractions (W1, W2). Purified soluble full-length mVSG1954 migrated with electrophoretic mobility in between the 55 and 70 kDa molecular weight marker (MW). **B)** The retention volume of mVSG1954 was used to estimate the molecular weight of the protein, based on the linear equation obtained from the linear regression of molecular weights of standard proteins and their retention volumes on HiLoad 16/600 Superdex 200 FPLC column. The calculated molecular weight of mVSG1954 is 84.4 kDa.

| Protein | Concentration | Average Molecular Weight (kDa) |
|---|---|---|
| mVSG1954 FL | 1 mg/ml | 55,1 |
| mVSG1954 NT | 1 mg/ml | 43,2 |

***Table 3.*** ***Full-length and N-terminal domain of mVSG1954 molecular weight determination by SEC-MALS.***

### b. Crystallization of purified mVSG1954 protein

Crystallization trials with full-length mVSG1954 did not produced suitable crystals for X-ray data collection. Because the C-terminal domain is tethered to the N-terminal domain through a flexible linker which poisons crystallization, limited proteolysis was performed to remove the protein's flexible regions. mVSG1954 was incubated with a low concentration of different proteases at low temperatures to allow the protease to remove only the more accessible and flexible parts of the protein. Different proteases, different protease concentrations, and different time points were tested. A 1:50 trypsin to protein ratio with three hours incubation time on ice was found to produce the optimal cleavage. The reaction was stopped by adding phenylmethyl sulfonyl fluoride (PMSF), a serine protease inhibitor that binds covalently to the serine residue at the protease's active site.

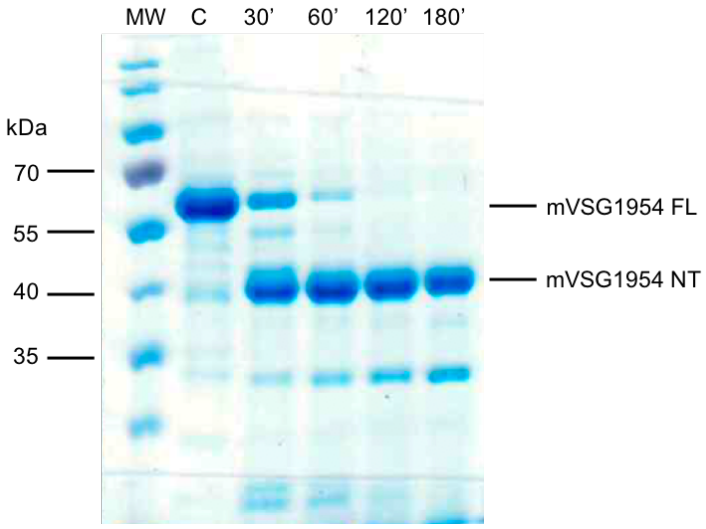


***Figure 15.*** ***Limited proteolysis on full-length mVSG1954 using trypsin to produce the mVSG1954's N-terminal.***

Limited proteolysis using 1:50 trypsin to protein ratio on ice at different timepoints (30, 60, 120, and 180 minutes). (MW: molecular weight, C: full-length mVSG1954)

Following limited proteolysis, the N- and C-terminal domains of mVSG1954 were separated by size exclusion chromatography. The mixture separated into two distinct peaks in the HiLoad 16/600 Superdex 200 column (Figure 16A). The first peak has a retention volume of 80.64 ml, and the second peak has a retention volume of 99.98 ml. The species that eluted in each peak were concentrated individually and analyzed by SDS-PAGE. The full-length mVSG1954 was used as a control. The species that eluted in the first peak, the N-terminal domain of mVSG1954, migrated with the similar electrophoretic mobility as the 40 kDa protein standard (Figure 16B). The species that eluted in the second peak migrated with similar electrophoretic mobility as the 15 kDa protein standard (Figure 16B). The retention volume for the species eluted in the first peak is used to estimate each species' molecular weight based on the linear equation obtained from the linear regression of the logarithmic value of the standard proteins' molecular weight and their retention volumes on the HiLoad 16/600 Superdex 200. The calculated molecular weight from the species that eluted in the first peak is 41.8 kDa (Figure 16C).

Additional SEC-MALS experiment was performed on the species eluted in the first peak by Dr. Karine Lapouge at the European Molecular Biology Laboratory, Heidelberg, Germany. The molecular weight of the N-terminal domain of mVSG1954

is 43.2 kDa (Table 3), which corroborates the result from the size exclusion chromatography. The data concludes that the N-terminal domain of mVSG1954 exists as a monomer in solution.

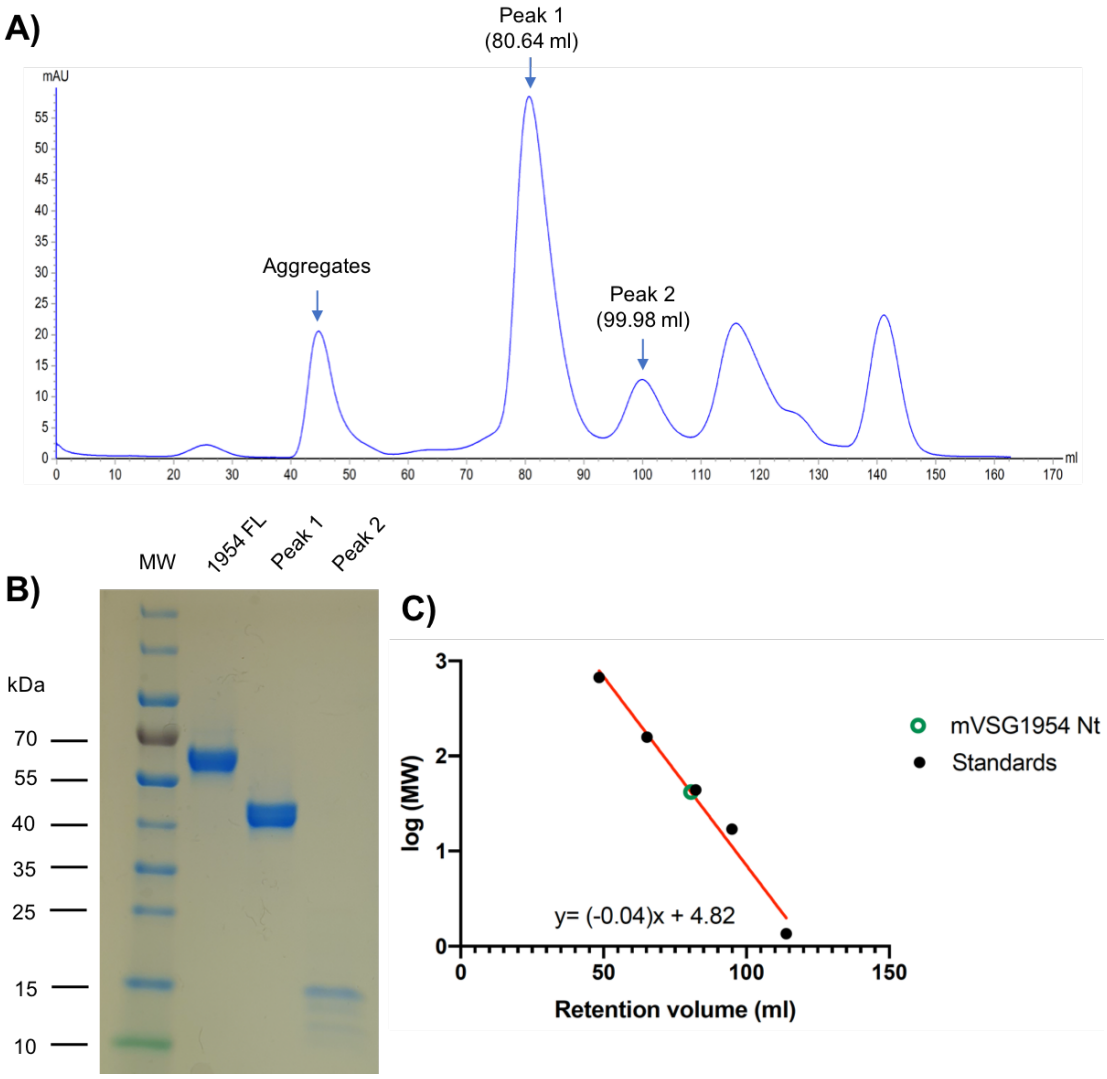


***Figure 16.*** ***Limited proteolysis on full-length mVSG1954 to produce N-terminal domain of mVSG1954.***

**A)** Separation of species generated during limited proteolysis on full-length mVSG1954 by trypsin. The peak at the beginning at the chromatogram contains protein aggregates. The main peak (peak 1) has retention volume of 80.64 ml, and the second peak has retention volume of 99.98 ml. **B)** The fractions that eluted from the first and second peaks were concentrated separately. The concentrated protein from each peak is analyzed by SDS-PAGE. The species that eluted in the first peak migrated with similar electrophoretic mobility as the 40 kDa molecular weight marker, and the species that eluted in the second peak migrated with similar electrophoretic mobility as the 15 kDa molecular weight. Full-length mVSG1954 protein **C)** The retention volumes were used to calculate the molecular weight of the species that eluted in each peak using the linear equation obtained from the linear regression of the logarithmic value of the molecular weight of the standard proteins and their retention volumes on HiLoad 16/600 Superdex 200 column. The calculated molecular weight from the species that eluted in the first and second peak are 41.8 and 7.1 kDa, respectively.

The initial crystallization condition was determined by sparse matrix screen. The sparse matrix screen was performed in-house and also in collaboration with Dr. Jürgen Kopp and Claudia Siegmann from the Protein Crystallization Core Facility at the University of Heidelberg, Heidelberg, Germany. In-house screening was performed by setting up three 96-well plate with Wizard 1&2 (Jena Biosciences), Wizard 3&4 (Jena Biosciences), and ProComplex (QIAGEN) Commercial Screens at both 22$^{o}$C and 4$^{o}$C. At the core facility, five 96-well plate were set up at 4$^{o}$C and 18$^{o}$C to find the initial crystallization condition. In addition, two different protein to crystallization solution ratios (1:1 and 2:1) were done for each crystallization condition. Several initial crystallization conditions with PEG 3350 at 18$^{o}$C were found, and the initial condition was further optimized by setting up four 96-well plate optimization screen with different sodium salts and different PEG 3350 concentrations. The most promising initial crystallization condition was 20% PEG 3350 and 0.4M NaCl at 18$^{o}$C. The crystals appeared 19 hours after the protein has been mixed with the crystallization solution. The crystals produced resemble hexagonal plates and needle-like crystals. Further optimization on the initial crystallization condition was done to improve crystal's quality.

Further optimization to obtain larger and harvestable crystals was performed by the vapor diffusion hanging drop method in 24-well plate format. The concentration of the main crystallization agent, PEG 3350, was varied in small increments. Different molecular weight PEGs were also used in the optimization. Different concentrations of NaCl and other sodium salts (sodium bromide and sodium iodide) were also used in the optimization experiment. Different protein to crystallization solution ratio was also tested in the experiment. The final optimized crystallization condition for mVSG1954 was 25% PEG 1,500, 0.4M NaBr with a 1:1 protein to crystallization solution ratio. The crystals' shape are three-dimensional plates. The crystals appeared after three days of incubation at 22$^{o}$C.

Before freezing, the crystals were transferred into a stabilization buffer containing 28% PEG1500, 0.75M NaBr, and 20% glycerol. A higher concentration of sodium bromide was added to 0.75M to incorporate more bromine atoms into the crystals to obtain anomalous datasets (see Chapter I part 5). The crystals were immersed in the stabilization buffer for 30 seconds and subsequently transferred and frozen in liquid nitrogen. X-ray diffraction data was collected at the Swiss Light Souce

at the Paul Scherrer Institute, Villigen, Switzerland, at 0.9198 Å wavelength. This is an X-ray absorption edge for bromine, the corresponding anomalous scattering measured used to obtain single-wavelength anomalous dispersion (SAD) dataset (See Chapter I). The most promising crystal diffracted to 1.68 Å resolution (Figure 17A). The space group is P321 with a= 69.614 Å, b= 69.614 Å, c= 115.36 Å, $\alpha$= 90$^o$, $\beta$= 90$^o$, $\gamma$= 120$^o$.

### c. Experimental phasing to determine the structure of mVSG1954

Molecular replacement cannot be performed to determine mVSG1954's structure due to the absence of a suitable model that shares high sequence identity with mVSG1954. Therefore, an anomalous diffraction experiment was performed to determine the structure of mVSG1954. The crystals were soaked in a high sodium bromide concentration to incorporate the bromine into the crystal, and anomalous X-ray diffraction data were collected at 0.9198 Å wavelength to obtain single-wavelength anomalous dispersion (SAD) dataset.

The substructure atoms were found using SHELX in the CCP4i2 crystallography software suite. Initial model building was performed in Arp/wArp and BUCCANEER in the same suite (Potterton et al., 2018). Subsequently, the model was further refined in phenix.refine, included in the PHENIX crystallography software suite. However, after data quality analysis, the last 300 frames in the dataset were opted out from the processing in xia2/DIALS to avoid incorporating data with radiation damage. The refined structure was used to perform molecular replacement with PHASER-MR on the new reprocessed data. Subsequent refinement was performed in phenix.refine, also included in PHENIX crystallography software suite (Adams et al., 2010). Cycles of manual model examination using COOT (Emsley and Cowtan, 2004) and improvement with automated refinement completed the model. One molecule of mVSG1954 presents in the crystal asymmetric unit (Figure 17B), and the molecules form trimer by crystallographic symmetry, identical to the crystallographic trimer of VSG3 (Figure 18B).

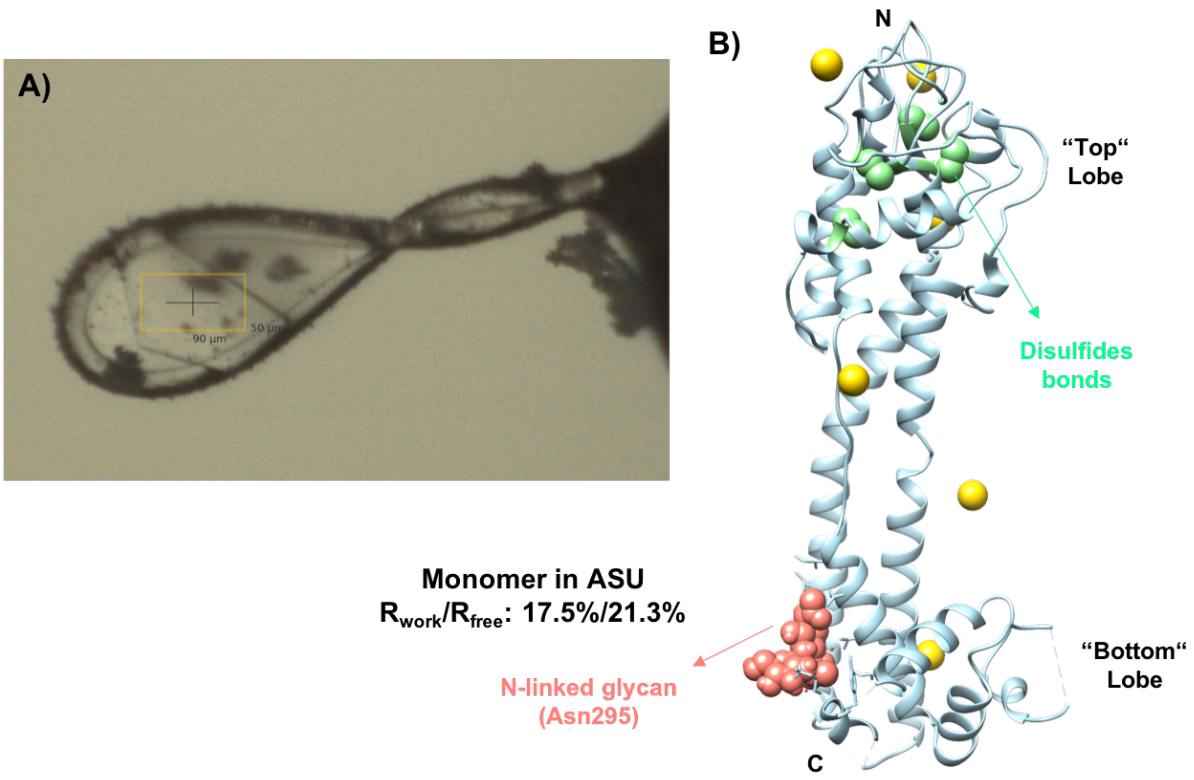


***Figure 17.*** ***Atomic resolution structure of the N-terminal domain of mVSG1954.***

The crystal of mVSG1954 was grown in 25% PEG 1500, 0.4M NaBr. The crystal was transferred to stabilization buffer with 28% PEG1500, 0.75M NaBr, and 20% glycerol prior to freezing in liquid nitrogen. The crystal was incubated for 30 seconds in the stabilization buffer before harvesting it with a nylon loop and flash-freezing **(A)**. Data collection was performed at Bromine K-Edge 0.9198 Å wavelength to obtain single-wavelength anomalous dispersion (SAD) dataset to solve the structure by experimental phasing. **B)** One monomer was found in the crystal asymmetric unit (light blue). The final resolution for mVSG1954 structure is 1.68 Å. Final $R_{work}/R_{free}$ for mVSG1954 structure is 17.5 %/21.3% (definition on R-values, see Chapter I, Part 5). N-linked glycan is attached at Asn295 at the structure's bottom lobe (pink salmon). Four disulfides bonds were observed at the structure's top lobe (light green). Bromine atoms used in experimental phasing is colored in yellow. For a more detailed data collection and final model refinement statistics, see Appendix 2. Images of protein structures were generated and edited using CHIMERA (Pettersen et al., 2004).

### d. Analysis of mVSG1954 structure

mVSG1954 scored highly only against BSF VSG3 based on the structure-based prediction algorithm PHYRE2 (Kelley et al., 2015). The N-linked glycan is attached to the residue Asn374 at the monomer's bottom lobe (Figure 17B). The N-linked glycan

position is different when compared to the VSG3 structure. In the VSG3 structure, the N-linked glycan is attached at the Asn67, rather at the N-terminal part of the VSG3 N-terminal polypeptide chain (Pinger et al., 2018). When the VSG3 and mVSG1954 trimers are superposed with each other, it is shown that the positions of the N-linked glycans are opposing each other (Figure 18A). The N-linked glycan at the VSG3 structure is embedded in the pocket between the structure' bottom lobe ear and the core three-helix bundle core. Meanwhile, the N-linked glycan at the mVSG1954 structure is placed at the interface between the monomers (Figure 18B). Eight cysteines forming four disulfide bonds were observed at the monomers' top lobe, with one free cysteine residue located at the structure's bottom lobe. Similar to VSG3's structures, the disulfide bonds are concentrated at the structure's top lobe (Figure 5B, 17B) (Pinger et al., 2018). The similarity was confirmed further by performing structure alignment between mVSG1954 and VSG3 monomer structure. The structure alignment is performed by DeepAlign in RaptorX structure alignment server (Wang et al., 2013, 2011). The RMSD value obtained from this alignment is considerably high, compared to, for instance the alignment between mVSG531 and VSG2 and VSG1 structures (Figure 12). The RMSD. measure is sensitive towards outlier regions, and in this case, it might be caused by the high variability at the top lobe structure (Figure 18C). The value of uGDT is still above 50, and the TM-value is still above 0.5, which indicate, the proteins aligned well and share the same fold (Xu and Zhang, 2010; Zemla, 2003; Zhang, 2005). By sequence alignment, mVSG1954 was shown to have a similar loop anchored by two cysteine residues, which contain the *O*-glycosylated serine residue. However, the crystal structure of mVSG1954 does not show any evidence for the presence of an *O*-linked glycan on the only serine residue on the loop (S321) (Figure 17B, 18B left).

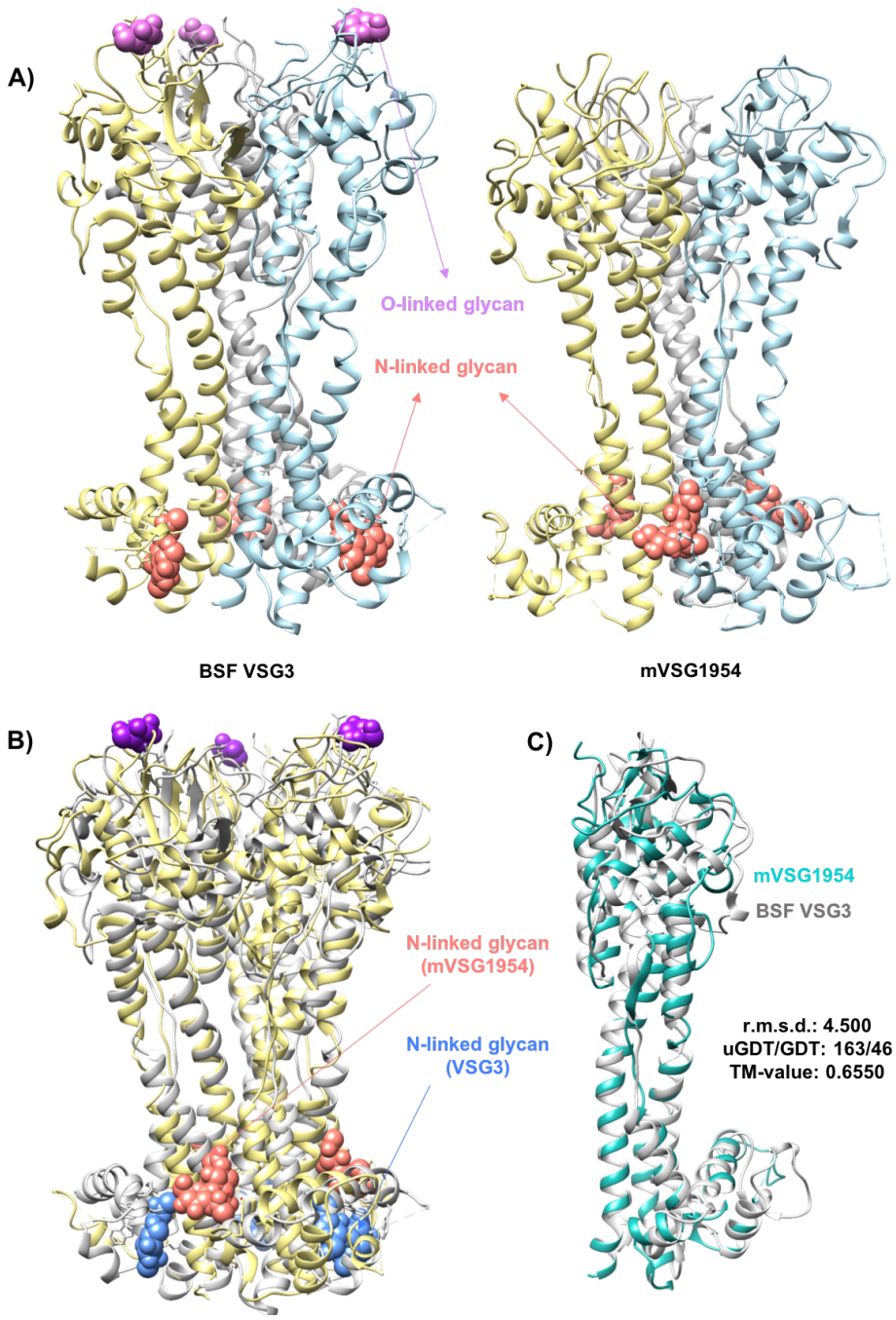


***Figure 18.* Comparison between mVSG1954 and BSF VSG3.**

**A)** Crystallographic trimer of VSG3 (left) and mVSG1954 (right). N-linked glycan is attached at the structures' bottom lobes (salmon pink). An *O*-linked glycan is attached at the VSG3 structure's top lobe (lilac). No evidence of an *O*-linked glycan is observed in mVSG1954 structure. **B)** Structure superposition between trimers of mVSG1954 (yellow) and VSG3 (grey) performed by SSM Superpose function in COOT (Krissinel and Henrick, 2004). **C)** Superposition of mVSG1954 (aquamarine) and BSF VSG3 (grey) monomers by DeepAlign in RaptorX structure alignment server (Wang et al., 2013, 2011) Images of protein structures were generated and edited using CHIMERA (Pettersen et al., 2004).

The absence of an *O*-linked glycan was also corroborated by intact mass spectrometry analysis. The intact mass spectrometry measurements were performed and analyzed by Nicole Lübbehüsen, at the mass spectrometry core facility at Zentrum für Molekular Biologie Heidelberg, Germany. The samples measured were the full-length mVSG1954 and mVSG1954 S321A, which are non-treated and treated with PNGaseF (NEB). PNGaseF is a bacterial enzyme that cleaves the bond between the asparagine residue and the innermost N-acetylglucosamine molecule in high mannose (thereby removing N-linked glycans but not *O*-linked) (Tarentino and Plummer, 1994). In the full-length mVSG1954 wildtype sample, a mass difference of 162 Da between the peaks corresponds to an attached hexose mass (Figure 19A). When treated with PNGaseF, a very small peak with a mass difference of 162 Da persists, suggesting that mVSG1954 could be *O*-glycosylated (Figure 19B), although the small size of the peak raises doubts to the interpretation. To test this idea, the serine residue that was presumed to be *O*-glycosylated was mutated to alanine (S321A). In the non-treated sample, mass differences of 203 Da and 162 Da were observed, which correspond to N-acetylglucosamine and hexose molecules. Besides, a mass difference of 16 Da was also observed compared to the full length mVSG1954 wildtype, indicating the serine to alanine mutation (Figure 19C). However, looking at the PNGaseF-treated S321A sample, the same mass difference of 162 Da was observed (Figure 19D). The data clearly indicate that S321 is not *O-glycosylated.* Several interpretations are possible for the 162Da mass differential: (1) the *O*-linked sugar is present in mVSG1954 but not on residue 321, (2) the *O*-linked sugar is present in mVSG1954 but not on residue 321 but only when produced in the metacyclic form (our protein is produced in the BSF), (3) the small peak is in fact “noise” in the spectrum and not indicative of a biochemical post-translational modification, (4) the mass is due to a non-covalently bound hexose group. Nevertheless, no evidence of *O*-glycosylation is observed anywhere else on the crystal structure's molecule. Together with the very weak to non-existent evidence for *O*-linked glycosylation by mass spectrometry, it is likely that mVSG1954 is not so modified (at least when expressed in the BSF).

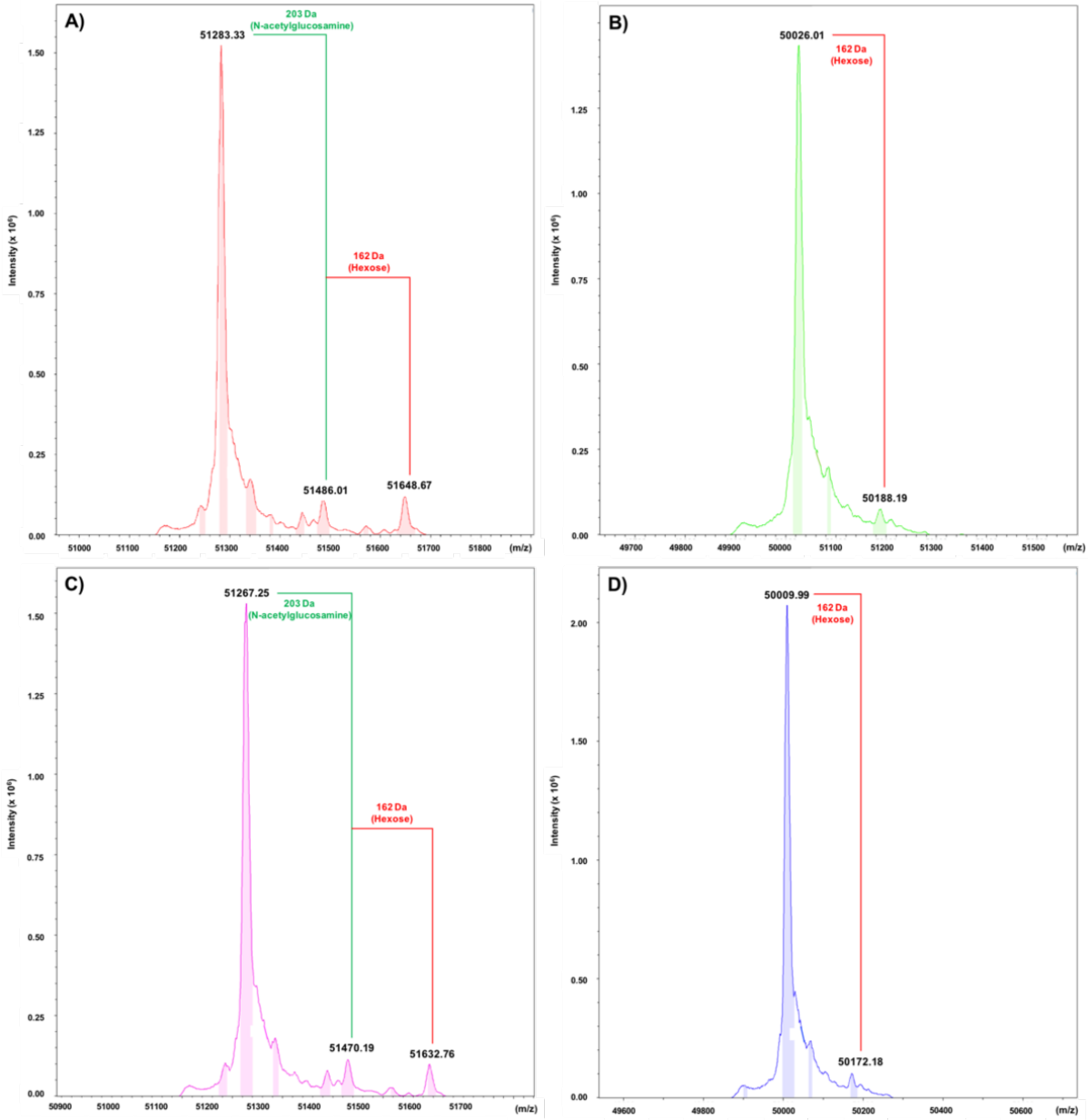


***Figure 19.* *mVSG1954 is not likely O-glycosylated.***

Intact mass spectrometry was performed on full-length mVSG1954 wildtype and S321A **(A and C)**. Both protein constructs are also treated with PNGaseF to cleave the N-linked glycans **(B and D)**. Mass differences of 203 Da and 162 Da was observed in the full-length mVSG1954 wildtype **(A)**, which correspond to N-acetylglucosamine and hexose molecules, respectively. When the sample was treated with PNGaseF **(B)**, only the 162 Da mass difference was observed, indicating the loss of N-linked glycan and the possible presence of *O*-linked glycan. This was examined by mutating the putative *O*-glycosylation site at S321 to alanine, which cannot be *O*-glycosylated. A 16 Da mass difference was observed between the full-length mVSG1954 wildtype **(A)** and S321A **(C)**, indicating the serine to alanine mutation. However, when treated with PNGaseF **(D)**, the 162 Da mass difference still presents, indicating that the protein is not *O*-glycosylated on that residue.

## Chapter VI: Antigenic Similarity between mVSGs expressed in BSF and MCF cells

### 1. Introduction

The structures of the two metacyclic VSGs, mVSG531 and mVSG1954, have revealed overall similarities between these mVSGs and the structures of VSGs present in bloodstream form trypanosomes (BSFs). However, the structures, though representing in sequence metacyclic VSGs, were determined after expression and purification of the relevant polypeptides in BSF cells. That is because MCF cells are cell cycle-arrested and cannot be cultured in vitro. One question that arises then, is whether antigenic differences between the two are due to the expression of mVSGs in BSF, not in the biologically relevant MCF, where for example expressed VSGs could incur MCF-specific PTMs. Proteomic data in Christiano et. al., 2017, has suggest the phosphorylation on two serine and two threonine residues in mVSG653. Phosphorylation has not been observed to date in the bloodstream VSGs (Christiano et al., 2017).

Recent experiments in the lab have demonstrated that even tiny differences in protein sequence (e.g. differences due to the presence or absence of a PTM) have substantial antigenic consequences (Pinger et al., 2018, Gkeka and Aresta-Branco, 2021, in preparation). To attempt to answer the question of whether antigenic differences exist between the mVSGs expressed in BSF cells compared to their native MCF context, we took the following route: we decided to generate antisera against these mVSGs after infection of animals with the relevant BSF cells, and to then test whether those antisera recognized equally the mVSGs expressed in MCF vs BSF. This chapter details results of these experiments.

## 2. Results

### a. Antisera raised against mVSG531 and mVSG1954 (α-mVSG531 and α-mVSG1954) are specific and exclusive

Two forms of *T. brucei* can be stably cultured *in vitro*, both deriving from the Lister427 strain: PCFs and BSFs. In 2012, Kolev and colleagues made the seminal if serendipitous discovery that overexpression of RNA binding protein 6 (RBP6) in PCFs "pushed" their differentiation all the way through to MCFs (Kolev et al., 2012). This differentiation protocol produces all intermediate forms as well, so that the RBP6 overexpressing culture will consist of undifferentiated PCF cells, epimastigotes, and MCF cells (which express all possible mVSGs present in the 427 strain with each individual cell expressing one type of mVSG at a time) (Christiano et al., 2017; Kolev et al., 2012).

PCF cells with tetracycline-inducible RBP6 overexpression plasmid were first differentiated by culturing the PCF cells with tetracycline for six days in low concentration glucose media. The differentiated cells (epimastigotes and MCF cells) were purified by selective depletion using magnetic labeling (see Chapter II), to arrive at a population of cells that almost entirely MCFs. These are still expected to be heterogeneous however, with regard to their expressed mVSG coat (Kolev et al., 2012).

Meanwhile, antisera raised against mVSG531 (α-mVSG531) and mVSG1954 (α-mVSG1954) were produced by infecting female C57BL6/J mice with BSF cells expressing mVSG531 or mVSG1954 respectively (see Chapter II). Specifically, on days 4 and 5 post-infection, berenil was administered to the mice to clear the infection, and then on day eight post-infection, the mice were sacrificed by $CO_2$ asphyxiation, and total blood was collected by cardiac puncture. The antisera were subsequently separated from the other blood components through centrifugation.

Both α-mVSG531 and α-mVSG1954 primary antisera, were subsequently used in flow cytometry (in conjunction with a FITC-labeled anti-mouse IgM secondary) to

detect mVSG531 and mVSG1954 expressing cells within the purified differentiated MCF population.

Both α-mVSG531 and α-mVSG1954 antisera were first tested to determine whether they can recognize BSF mVSG531 or BSF mVSG1954 (used to infect the mice). For this, BSF mVSG531 or mVSG1954 were stained with α-mVSG531 or α-mVSG1954 and subsequently counterstained using Goat α-Mouse IgM conjugated to FITC. The cells were examined using flow cytometry. Clearly, both α-mVSG531 and α-mVSG1954 can recognize the BSF cells expressing their corresponding cognate mVSGs, as indicated by the shift in fluorescence intensity compared to the negative controls (unstained cells and BSF cells stained with antisera from naïve mice) (Figure 20).

Additionally, both antisera were tested for cross-reactivity: α-mVSG531 was tested against BSF mVSG1954 and BSF cells expressing VSG2 (BSF VSG2), while α-mVSG1954 was tested against BSF mVSG531 and BSF VSG2 as well. No cross-reactivity was detected, and each antibody was only able to recognize the cell type against which each was raised (Figure 20).

Therefore, both α-mVSG531 and α-mVSG1954 can specifically and exclusively recognize the BSF cells expressing the corresponding mVSGs used in the infection.

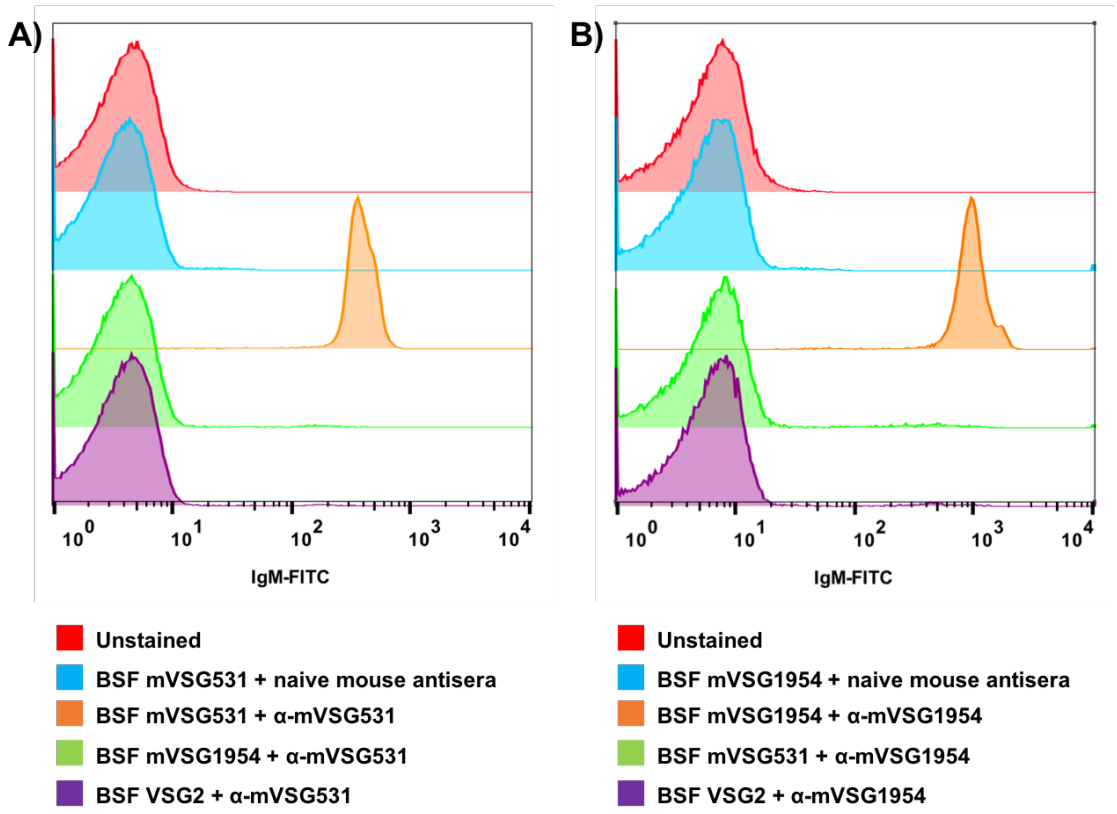


***Figure 20. Antisera raised against mVSG531 or mVSG1954 in C57BL6/J mice specifically recognizes the BSF cells expressing respective mVSGs.***

**A)** α-mVSG531 was tested to stain BSF mVSG531 (orange), BSF mVSG1954 (green) and BSF VSG2 (purple). After staining with α-mVSG531, the cells were counterstained with secondary Goat anti-Mouse IgM coupled to FITC. Unstained cells and BSF cells expressing mVSG531 stained with antisera from naïve mouse were used as negative controls. **B)** α-mVSG1954 was tested to stain BSF mVSG1954 (orange), BSF mVSG531 (green) and BSF VSG2 (purple). After staining with α-mVSG1954, the cells were counterstained with secondary Goat anti-Mouse IgM coupled to FITC. Unstained cells and BSF cells expressing mVSG1954 stained with antisera from naïve mouse were used as negative controls. Shift in fluorescence intensity was observed in the sample in which the antisera were used to stain the BSF cells expressing the respective mVSG (orange), but not in the sample in which the antisera were used to stain BSF cells expressing the opposite mVSG (green) or BSF VSG2 (purple).

### b. mVSGs expressed in BSF and in differentiated MCF are antigenically similar

As mentioned, PCF cells can be differentiated by inducing the overexpression of RBP6 to obtain MCF cells. As differentiation proceeds in the fly, early procyclic cells, whose surface is covered by a protein made of GPEET repeat proteins, differentiate into late procyclic cells (covered by an EP repeat protein coat). The Lister 427 PCF cells that we propagate *in vitro* most likely represent a transition state between these two procyclic cell subtypes, in that they express both GPEET and EP proteins on their coat. As these then differentiate further, they also lose the EP repeat coat. Therefore, we used antibodies against these two coat types to follow their loss, as a marker for how well differentiation proceeded.

Differentiation was examined on days 5 and 6 post-induction by detecting GPEET and EP's loss on the parasite surface by flow cytometry. In the starting population, which is the uninduced culture, 91.4% of the cells expressed GPEET on their surface (Figure 21A), and 96.6% of the cells express EP on the surface (Figure 21B). 98% of the cells are positive for GPEET and EP together, when stained with both $\alpha$-GPEET and $\alpha$-EP, indicated by the shift in fluorescence intensity (Figure 21C). It is concluded that most of the starting population cells are PCF cells. In the induced culture, 89.9% of the cells in the culture had lost their GPEET protein from the surface on day six post-infection compared to the starting population (Figure 21D). However, when stained with $\alpha$-EP, 86.1% of the cells were positive for EP, indicated by the shift in the fluorescence intensity (Figure 21E). Only 9.01% of the mixture's cells do not express GPEET or EP on the surface (Figure 21F), and these likely consist of epimastigotes (which express BARP), and MCF cells expressing all five possible Lister427 mVSGs.

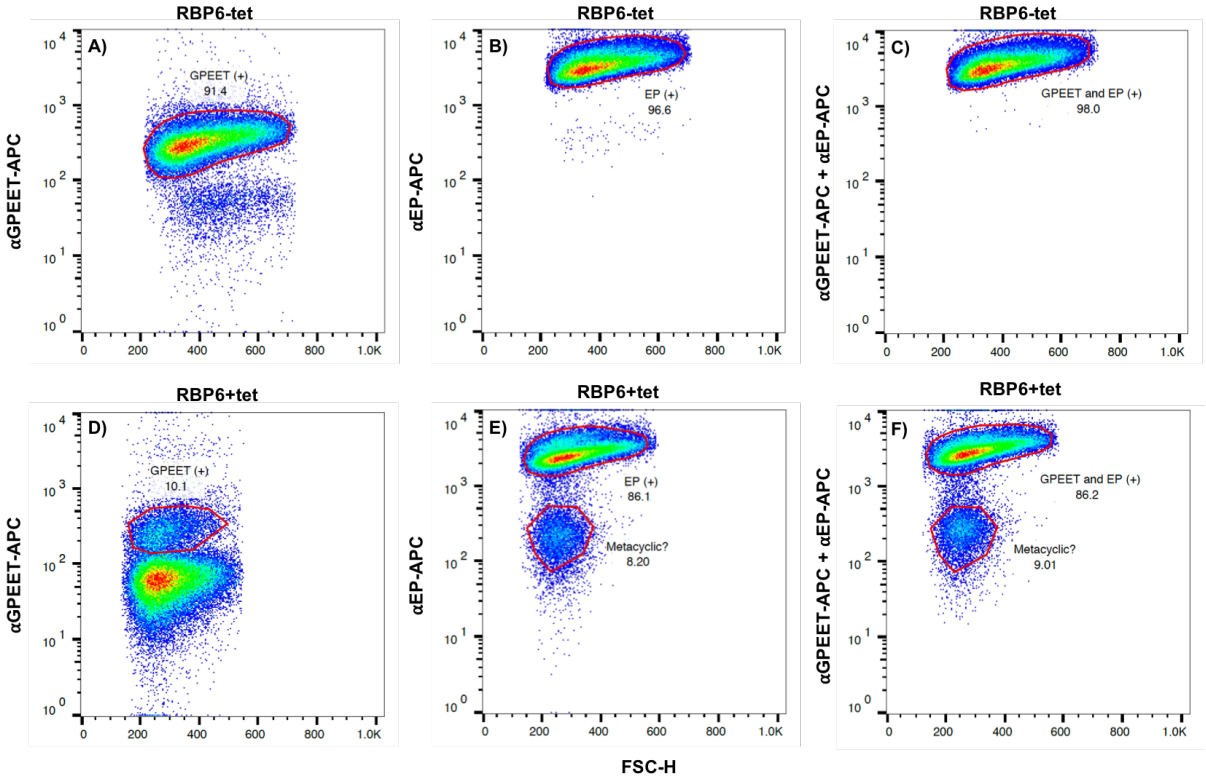


***Figure 21.*** ***Overexpression of RBP6 in PCF cells initiate differentiation of PCF cells into later life stages of T. brucei.***

Wildtype 29-13 PCF cells tetracycline-inducible RBP6 overexpression plasmid was cultured in SDM-80 media supplemented with 7.5 μg/ml Hemin, 50 mM N-Acetylglucosamine, and 10% $^{v}/_{v}$ heat-inactivated FBS at 27°C and 5% CO2. The induction was started at 2 million cells/ml cell density with 10 μg/ml tetracycline for 6 days. On day 6, both uninduced and induced culture were analyzed by flow cytometry. Both uninduced and induced cultures were stained with α-GPEET **(A, D)**, α-EP **(B, E)**, or both α-GPEET and α-EP **(C, F)**. Both α-GPEET and α-EP were coupled to fluorescence molecule APC. 91.4% of the cells in starting population (uninduced culture) is positive for GPEET on their cell surface (A), and 96.6% of the cells is positive for EP (B), indicated by the shift in the fluorescence intensity. When stained with both α-GPEET and α-EP, 98% of the starting population has shifted in fluorescence intensity **(C)**. In the induced culture, only 10.1% from the cells is positive for GPEET and 86.1% of the cells is positive for EP. When the induced culture is stained with both α-GPEET and α-EP 86.2% is positive for both GPEET and EP. 9.01% of the cells in the induced culture is negative for either GPEET and EP, indicating that this cell population is no longer PCF cells.

To enrich for differentiated MCFs, cells which still expressed GPEET and EP, were separated from the cell mixture by magnetic separation (see Chapter II). The mixture was first stained with both α-GPEET and α-EP and counterstained with Goat anti-mouse IgG coupled to magnetic beads. The stained PCF cells were subjected to a magnetic field to retain them in the column and separate it from the GPEET and EP – negative differentiated cells, which contain the MCF cells of interest.

Having purified a reasonable number of MCF cells ($7 \times 10^6$ cells), we then aimed to stain them using the previously collected and validated α-mVSG531 and α-mVSG1954 antisera, and compare that to the ability of the same antisera to recognize these mVSGs when expressed in BSFs.

Firstly, the possibility that these antisera recognize any marker on the surface of PCFs, which did not cross react with non-cognate BSF cells, also need to be excluded. Therefore, BSF mVSG531 or BSF mVSG1954 were mixed with undifferentiated PCF cells in 1:20 BSF cells to PCF cells ratio. The mixtures were stained with α-mVSG531 or α-mVSG1954 and counterstained with α-mouse IgM conjugated to FITC. The stained cells were subsequently analyzed by flow cytometry. From flow cytometry data, both α-mVSG531 and α-mVSG1954 were able to recognize and bind only the small population of BSF cells expressing mVSG531 mVSG1954 in the mixture with PCF cells, indicated by 2-3% of cells that has an increased shift in fluorescence intensity (Figure 22). This not only demonstrates the lack of cross-reactivity with PCFs, but it also shows that small percentages of cells can be recognized easily by a shift in fluorescence intensity (here, the “spiked” BSF material, comprising only 2-3% of the mix can be identified with ease).

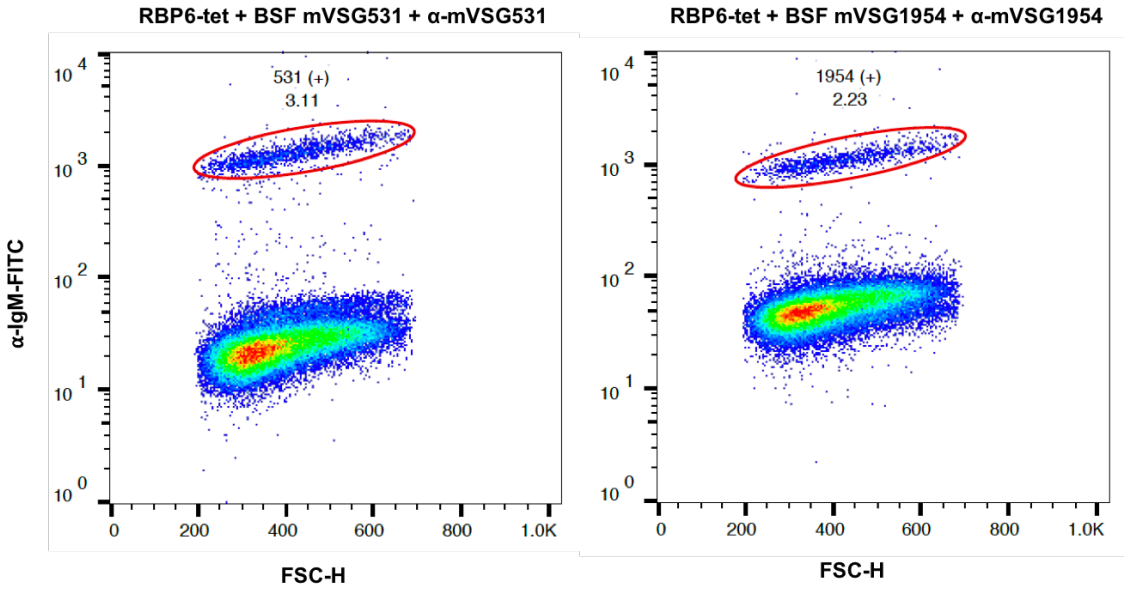


***Figure 22.*** ***Antisera α-mVSG531 and α-mVSG1954 elicited in C57BL6/J mice is able to recognize small positive cell population from a mixture with undifferentiated PCF cells.***

5% of BSF cells expressing mVSG531 (left) or mVSG1954 (right) was mixed with 95% undifferentiated PCF cells. The mixtures were stained with either α-mVSG531 or α-mVSG1954 and counterstained with Goat anti-mouse IgM coupled to FITC. 3.11% of the cells in the mixture of undifferentiated PCF cells and BSF cells expressing mVSG531 has shifted in fluorescence intensity (left). 2.33% of the cells in the mixture of undifferentiated PCF cells and BSF cells expressing mVSG1954 has shifted in fluorescence intensity (right).

Successively, the purified differentiated (MCF) cells were stained with α-mVSG531 or α-mVSG1954 or both α-mVSG531 and α-mVSG1954. The cells were counter stained with α-mouse IgM conjugated to FITC and analyzed by flow cytometry. Approximately 3.83% of the cells in the cell mixture were recognized by the α-mVSG531 antisera, and the level of fluorescence intensity was identical between these MCFs and mVSG531 expressing BSFs (Figure 23A, D). Similar results were collected for the 1.90% of cells that were recognized by the α-mVSG1954 antisera (Figure 23 B, E). Together, the lack of fluorescence intensity shift results strongly suggest that the antisera raised against these VSGs recognize them regardless of whether they are expressed natively in MCFs, or in BSFs through genetic manipulation. It is concluded that lack of an obvious difference in antigenicity likely means that these VSGs are identical at least as far as the immune system is concerned. By extension, it is concluded that there is nothing special about mVSGs – with the exception of their genomic location they function just like “standard” VSGs.

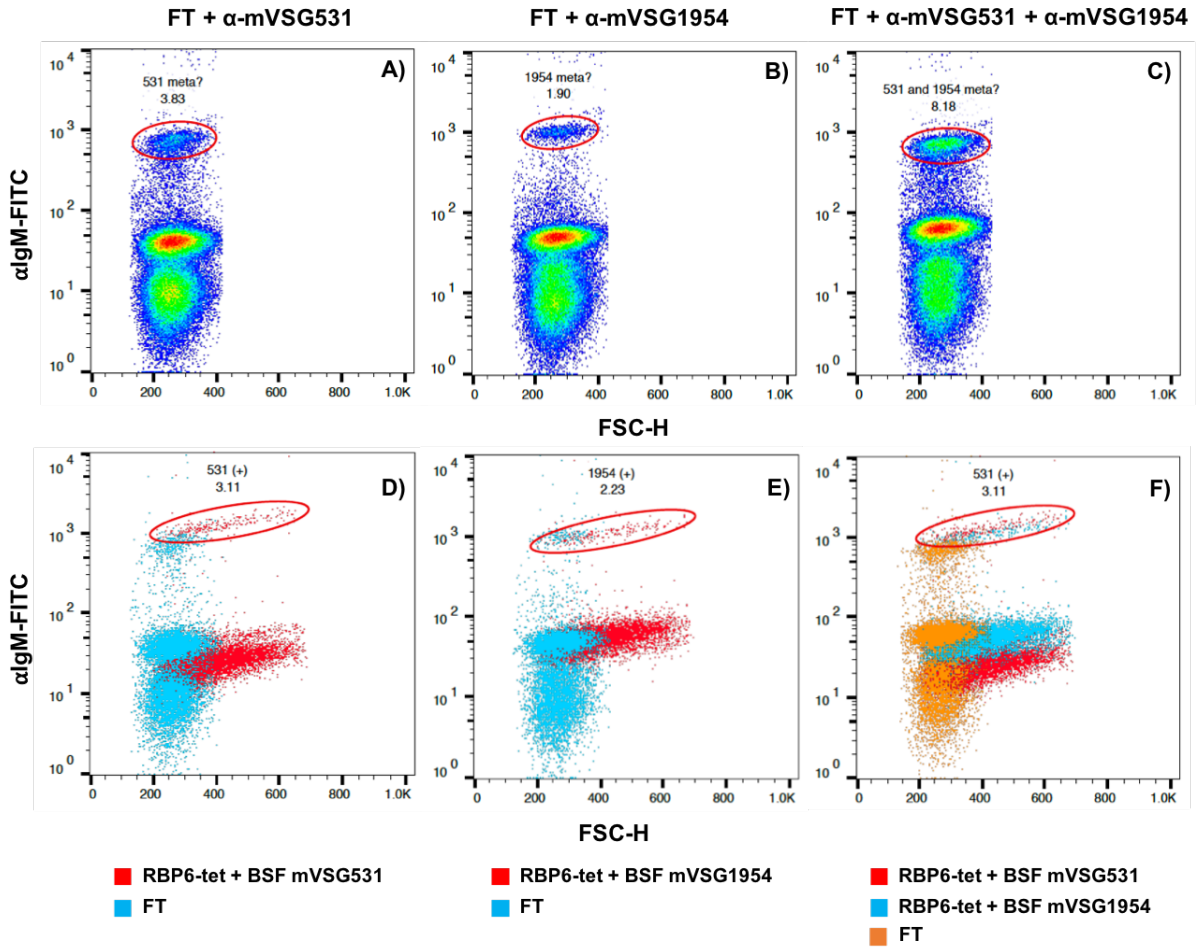


***Figure 23.*** ***Antisera α-mVSG531 or α-mVSG1954 elicited in C57BL6/J mice can recognize mVSG531 or mVSG1954 expressed in MCF cells.***

The undifferentiated PCF cells in the induced culture was separated from the differentiated cells by magnetic labelling. The flow through fraction that contains the differentiated cells was stained with α-mVSG531 **(A)**, α-mVSG1954 **(B)**, or both α-mVSG531 and α-mVSG1954 **(C)**. The cells were further counterstained with Goat α-mouse IgM coupled with FITC. 3.83% and 1.90% cells were positive (A and B, red circle) for either mVSG531 or mVSG1954 on the cells surface, indicated by the shift in the fluorescence intensity. When the cells were stained with both α-mVSG531 and α-mVSG1954, 8.18% of the cells were positive **(C, red circle)**. The flow cytometry dot plots from the differentiated cells were overlaid with the dot plot from figure 13 **(D, E, F)** to compare the binding of α-mVSG531 or α-mVSG1954 to the cells expressing mVSG531 or mVSG1954.

## Chapter VII: Discussion

The evolution of host antibody response and parasite immune evasion mechanisms is a continuous arms race. *T. brucei* overcomes the host immune system response by switching its current expressed VSG into an antigenically distinct VSG. The presence of the large archive of VSG genes available in *T. brucei's* genome enables it to evade the host immune response continuously by antigenic variation (Hovel-Miner et al., 2015; Mugnier et al., 2016). Aside from amino acid sequence diversity, recent literature has shown that VSGs are also diverse in their tertiary structure organization, oligomerization states, and post-translational modification, all factors with significant influence over the host's immune response (Pinger et al., 2018; Zeelen et al., 2021).

This study focuses on the metacyclic life stage of *T. brucei*, which is the life stage that inhabits the tsetse fly's salivary gland and is subsequently transmitted into the mammalian host when the fly takes a blood meal (refer to Chapter I part 2). Similarly, to BSF cells, MCF cells also express VSGs on their surface (Ramey-Butler et al., 2015). These are not expressed from standard Bloodstream Expression Sites but rather from metacyclic expression sites (MESs - which are very similar in BESs in that they are telomeric, but rather dissimilar in that they do not co-express ESAGs and do not include repeat regions that characterize BESs) (refer to Chapter I part 3). When injected into the mammalian host, the mVSGs expressed on the parasites' surfaces are still retained for up to 7 days before these first-to-be-injected MCFs differentiating into BSF cells (which now express "proper" VSGs from one of 15-20 telomeric BESs) (Graham et al., 1990; Helm et al., 2009). These MES/mVSG and BES/VSG differences have been thought to conserve energy and resource in the cell cycle-arrested MCF cells (Christiano et al., 2017).

In the Lister427 *T. brucei* lab strain used in our lab, there have been five VSGs that reside in MESs and which are therefore expressed in MCF cells, namely mVSG397, 531, 639, 653, and 1954 (Christiano et al., 2017; Kolev et al., 2012). When this study was started, several VSG structures were already available. Among those were the bloodstream VSG2, llTat 1.24, VSG1, and VSG3 (Bartossek et al., 2017; Blum and Turner, 1993; Freymann et al., 1990; Pinger et al., 2018). The comparison between these had shown unexpected but clear divergence between class A VSGs

(VSG2, IlTat1.24, VSG1) and class B VSGs (VSG3) in most structural aspects (oligomeric state, topology, PTM presence). However, information on mVSGs was missing, and my PhD studies centered upon identifying differences, if any, between mVSGs and other VSGs.

Firstly, bioinformatics tool structure-based protein structure prediction PHYRE2 (Kelley et al., 2015) was used to decide which mVSGs would be used in the study. The prediction results show that mVSG397, 531, 639, and 653 are structurally related to the bloodstream VSG structures in Class A. Only mVSG1954 was predicted to be structurally related to VSG3 structure in class B. Thus, mVSG531 and mVSG1954 were chosen to be used in the structural study using x-ray crystallography to compare mVSG structures of Class A and B.

The N-terminal mVSG531's structure indeed revealed similarities with VSGs from the Class A family. Size exclusion chromatography has shown that purified full-length mVSG531 expressed in the BSF cells elutes with a dimer's molecular weight. The determined structure by x-ray crystallography indeed corroborates the result. Eight molecules of mVSG531, which form four dimers, were found in the crystal asymmetric unit by molecular replacement. The mVSG531 structure also reveals striking similarities in the overall tertiary structure compared with VSG2, VSG1, and IlTat 1.24 structures. The mVSG531 structure has the characteristic dumbbell-like structure with a three-helix core. Two disulfide bonds were concentrated at the molecule's top lobe, alike with the VSG2, VSG1, and IlTat 1.24's structures (Bartossek et al., 2017; Blum and Turner, 1993; Freymann et al., 1990). N-linked glycan is also observed in mVSG531's structure, attached at the structures' bottom lobe, similar to those in VSG2 and VSG1 structures (Bartossek et al., 2017; Freymann et al., 1990). Structure alignment between mVSG531, VSG2, and VSG1 structure also resulted in low RMSD value and high TM-value, indicating that the structures are highly related to each other and belong to the same protein fold. Despite the similarities, the electrostatic potential and hydrophobicity properties at the surface differ significantly due to the variations on the amino acid sequence level. Thus, mVSG531, though overall highly similar to the other VSGs mentioned, is still recognized differently by the mammalian host immune system.

Similarly, as predicted, the N-terminal domain mVSG1954 structure has revealed similarities with VSG3. The result from the size exclusion chromatography and SEC-MALS has revealed that the full-length and the N-terminal domain of mVSG1954 exist as a monomer and the crystal structure validates the result. A monomer of mVSG1954 was found in the crystal asymmetric unit, similar to the VSG3 structure. Four disulfide bonds were observed and concentrated at the structure's top lobe, like the VSG3 structure. The N-linked glycan is attached at the structure's bottom lobe, similar to VSG3. However, VSG3's N-linked glycan is attached at the asparagine residue at the N-terminal part of the polypeptide sequence of the N-terminal domain of mVSG1954. Meanwhile, the N-linked glycan in mVSG1954 is attached at Asn374, at the C-terminal part of the N-terminal domain's polypeptide chain. The monomers also form a crystallographic trimer, as shown in the VSG3 structure. In another crystal form, we also observed that the monomers form a non-crystallographic trimer, which further signifying the relevance of the trimer structure. Different oligomerization states have challenged the assumption that VSGs always exist as dimers on the parasites' surfaces; therefore, the discovery that VSG3 and now also mVSG1954 are likely trimers is biologically relevant. The relevance of the trimer structure seen in mVSG1954 was also observed in the study by Umaer et. al, 2021 (submitted). Analysis of whole lysate of BSF mVSG1954 on the native gel showed the existence of high molecular weight species that resembles a potential trimer.

Intact protein mass spectrometry also shows that mVSG1954 is not O-glycosylated, despite having the same loop region and a serine residue that was O-glycosylated in other Class B's VSGs (Pinger et al., 2018). This difference in glycosylation pattern might come due to the difference in the life stages in which the protein should be expressed. It is possible that the MCF cells are using a different subset of O-glycosyltransferase compared to the BSF cells.

The structures of mVSGs have shown that MES resident VSGs are not different than VSGs in BESs that are normally expressed in BSF cells. However, we wondered if mVSGs expressed in MCF cells, even if not structurally distinct, were in any way antigenically distinct. We therefore used flow cytometry together with $\alpha$-mVSG531 and $\alpha$-mVSG1954 sera raised against the BSF expressed forms, to detect any difference in binding to mVSG531 and mVG1954 expressed in the MCF cells differentiated in

vitro from PCF cells. The data clearly demonstrate that the mVSG531 and mVSG1954 expressed in MCF cells are antigenically indistinguishable from those on the surface of BSF cells. Thus, there is likely nothing structurally or antigenically “special” to MES resident VSGs.

To date, mVSG’s function in MCF cells is unknown. One possibility is that the similarities of mVSGs to their bloodstream VSG counterpart might function as the MCF cells' defense system as they enter the mammalian host. The alternative complement system lyses the PCF cells and potentially other pre-MCF cell types (Ferrante and Allison, 1983). By having an intact VSG coat, the MCF cells can survive in the mammalian host and eventually differentiate into infectious, rapid-growing BSF cells. Of course, this hypothesis must also account for the fact that mVSGs are expressed from MES – why not express these in MCFs already within BESs? Unlike MESs, which is relatively shorter, there are many genes that are co-expressed with VSG within BESs (ESAGs) (Cully et al., 1985). This raises the possibility that other BES-co-expressed genes (such as ESAGs for example) are toxic to MCFs (and are therefore specifically avoided). This can be tested in the future by inserting one of the BES-co-expressed genes in a specific MES, which expresses specific mVSG, in PCF cells with the RBP6 overexpression system. The PCF cells can be further differentiated into MCF cells, and the inserted gene will be expressed from the MES. If the gene is indeed toxic to the cells, the cells expressing the gene will die and we will observe the loss of the particular mVSGs from the pool of the MCF cells.

Finally, as MCF cells are the first antigenic surface presented to the mammalian host, and as their mVSG repertoire is limited (5 mVSGs in Lister427), it has been suggested that mVSGs might be an attractive target in the context of prophylaxis treatment (e.g. a vaccine). However, the findings herein that an mVSG itself is “just” another VSG both structurally and antigenically, together with previous studies that have identified “mVSGs” expressed within BESs in populations of VSG switchers (Mugnier et al., 2015), strongly suggest that the repertoire of mVSGs (or rather, of MES-resident VSGs expressed within MCFs) in nature can be (or can become) just as diverse as that of BES-expressed variants. It is therefore unlikely that mVSG vaccines can be a prophylactic proposition in the context of trypanosomiasis.

## Chapter VIII: Conclusion and Future directions

The work performed in this thesis has shed light on mVSGs structures for the first time. Specifically, this study determined the structures of mVSG531 and mVSG1954. Extensive similarities were uncovered between the mVSGs structures when compared to the bloodstream VSGs. We also have determined that the similarities were not likely due to the expression of mVSGs in the BSF cells: specifically, we note that the mVSGs expressed in the MCF cells are antigenically similar as compared to mVSGs expressed in the BSF cells.

Of course, there are still numerous things that need to be confirmed and investigated. Proteomic experiments need to be performed to further confirm the similarity between the mVSGs expressed in the BSF form and the mVSGs expressed in the MCF cells. Such a proteomic experiment will also be helpful in detecting potentially new PTMs (such as phosphorylation, first detected on mVSG653 – (Christiano et al., 2017)). Determining the structures of additional mVSGs will also be helpful to see if there are any differences between Class A mVSGs. Lastly, co-crystallization experiments of mVSGs with anti-mVSG cognate antibody will be very crucial if we are to identify the interaction site between the mVSGs and the host immune system.

**Acknowlegdment**

The completion of the work in this thesis would not be possible without the extensive support of the people in my support system. To them I would like to extend my forever deepest gratitude:

To my mentors and scientific advisors, Prof. Dr. F. Nina Papavasiliou and Prof. Dr. C. Erec Stebbins, thank you for believing in me despite all my shortcomings and giving me the chance to pursue my doctoral degree in your laboratories. I thank you both for the fruitful scientific discussions and numerous advices. I also thank you both for being the people I can look up to and being the role models for my scientific career.

To my thesis committee advisors, Prof. Dr. Christine Clayton and Prof. Dr. Hedda Wardemann, thank you for the mentorship and scientific and career advices during my doctoral studies, and for the time and commitment to discuss the progress of my project in the progress meeting. I also thank Christine for the time and commitment to join my defense committee.

To my collaborators, Prof. Dr. Lucy Glover, Nicole Lübbehüssen and Dr. Thomas Ruppert from the ZMBH Mass Spectrometry facility, Dr. Karine Lapouge from EMBL, Heidelberg, Claudia Siegmann and Dr. Jürgen Kopp from the University of Heidelberg Protein Crystallization Core Facility, for sharing their resources and expertise. The project will not be completed if not because of all of you.

To my defense committee members, Prof. Dr. Christian Janzen and Dr. Guoliang Cui, thank you for the time and commitment to join my defense committee.

To my family, my father and mother, thank you for the emotional and financial support that have brought me to this point of my life. Thank you for believing in and supporting your daughter who is doing stuffs that still puzzle you most of the time.

To my partner in everything, Dr. Thomas Plum, thank you for the love and your never-ending support in my ups and downs during my study, for always believing in me at times when I even did not believe in myself, and for scientific discussions at the most odd-times of the day.

To ones who have welcomed me to their family, Jutta and Georg Plum, thank you for the warm welcome whenever I am there. Thank you for believing in me and being proud of me.

To my direct supervisor, Dr. Francisco Aresta-Branco, thank you for sharing your expertise on trypanosome biology and all the advices I need to jump start my project when I first join the lab. Thank you for all the fruitful scientific discussion and all the help you gave during my study.

To my foodie partners and the coolest girls in the lab, Anastasia Gkeka, Dr. Gianna Triller, Beatrice Casati, Dimitra Stamkopoulou, and Taga Lerner, thank you for the delicious home-made dinners, coolest girls' nights, and the chats about life. I will forever cherish it.

To my D160 colleagues, Alexander Hempelmann, Monique van Straaten, and Johan Zeelen, thank you for all the scientific discussions and advices, especially on protein crystallography, you all are really helpful and this project will not be completed without your help.

To my sweet dealer, Dr. Riccardo Pecori, thank you for always willing to be my victim of my impulse for sharing sweets.

To the one who left the lab too soon, Dr. Esteban Erben, thank you for the scientific discussion and advices, especially on the trypanosome biology.

To our sushi master, Dr. Joey Verdi, thank you for the numerous delicious sushi dinners and of course your scientific discussion and advices.

To the ones who always made everything interesting, Rafail Tasakis and Jose Paulo Lorenzo, thank you for your positive energy to light up every situation.

To the ones who always keep the lab running, Sandra Ruf and Annette Arnold, thank you for keeping the things in the lab running and working despite all our shortcomings in being tidy.

To my sister from another parents, Pearlyn Ler, thank you for all the delicious home-made food, being a couch potato with me, and sending each other memes. Thank you for always being there for me.

To another foodie buddy, Marius Felix, thank you for always supplying me with good food, criticizing my knife skill, and the chats about life over glasses of wine.

To all my other friends whom I cannot mention one by one, everybody whom I have met along the way, old and new friends, I will be forever grateful of all of you for making me the person that I am today.

## Bibliography


Abergel, C., 2013. Molecular replacement: tricks and treats. Acta Crystallogr. D Biol. Crystallogr. 69, 2167–2173. https://doi.org/10.1107/S0907444913015291

Acosta-Serrano, A., Vassella, E., Liniger, M., Kunz Renggli, C., Brun, R., Roditi, I., Englund, P.T., 2001. The surface coat of procyclic Trypanosoma brucei: programmed expression and proteolytic cleavage of procyclin in the tsetse fly. Proc. Natl. Acad. Sci. U. S. A. 98, 1513–1518. https://doi.org/10.1073/pnas.041611698

Adams, P.D., Afonine, P.V., Bunkóczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.-W., Kapral, G.J., Grosse-Kunstleve, R.W., McCoy, A.J., Moriarty, N.W., Oeffner, R., Read, R.J., Richardson, D.C., Richardson, J.S., Terwilliger, T.C., Zwart, P.H., 2010. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr. D Biol. Crystallogr. 66, 213–221. https://doi.org/10.1107/S0907444909052925

Alarcon, C.M., Son, H.J., Hall, T., Donelson, J.E., 1994. A monocistronic transcript for a trypanosome variant surface glycoprotein. Mol. Cell. Biol. 14, 5579–5591. https://doi.org/10.1128/MCB.14.8.5579

Al-Qahtani, A., Teilhet, M., Mensa-Wilmot, K., 1998. Species-specificity in endoplasmic reticulum signal peptide utilization revealed by proteins from Trypanosoma brucei and Leishmania. Biochem. J. 331 ( Pt 2), 521–529. https://doi.org/10.1042/bj3310521

Aresta-Branco, F., Sanches-Vaz, M., Bento, F., Rodrigues, J.A., Figueiredo, L.M., 2019. African trypanosomes expressing multiple VSGs are rapidly eliminated by the host immune system. Proc. Natl. Acad. Sci. 116, 20725–20735. https://doi.org/10.1073/pnas.1905120116

Barrett, M.P., Boykin, D.W., Brun, R., Tidwell, R.R., 2007. Human African trypanosomiasis: pharmacological re-engagement with a neglected disease. Br. J. Pharmacol. 152, 1155–1171. https://doi.org/10.1038/sj.bjp.0707354

Bartossek, T., Jones, N.G., Schäfer, C., Cvitković, M., Glogger, M., Mott, H.R., Kuper, J., Brennich, M., Carrington, M., Smith, A.-S., Fenz, S., Kisker, C., Engstler, M., 2017. Structural basis for the shielding function of the dynamic trypanosome variant surface glycoprotein coat. Nat. Microbiol. 2, 1523–1532. https://doi.org/10.1038/s41564-017-0013-6

Bergfors, T. M. (2009) Protein Crystallization. International University Line. ISBN 978-0-0720774-4-6

Blow, D. (2010) Outline of Crystallography for Biologists. Oxford University Press. ISBN 978-0-19-851051-2

Blum, M.L., Turner, M.J., 1993. A structural motif in the variant surface glycoproteins of Trypanosoma brucei 362, 7.

Borst, P., Ulbert, S., 2001. Control of VSG gene expression sites. Mol. Biochem. Parasitol. 114, 17–27. https://doi.org/10.1016/S0166-6851(01)00243-2

Büscher, P., Cecchi, G., Jamonneau, V., Priotto, G., 2017. Human African trypanosomiasis. The Lancet 390, 2397–2409. https://doi.org/10.1016/S0140-6736(17)31510-6

Bütikofer, P., Ruepp, S., Boschung, M., Roditi, I., 1997. 'GPEET' procyclin is the major surface protein of procyclic culture forms of Trypanosoma brucei brucei strain 427. Biochem. J. 326, 415–423. https://doi.org/10.1042/bj3260415

Bütikofer, P., Vassella, E., Ruepp, S., Boschung, M., Civenni, G., Seebeck, T., Hemphill, A., Mookherjee, N., Pearson, T.W., Roditi, I., 1999. Phosphorylation of a major GPI-anchored surface protein of Trypanosoma brucei during transport to the plasma membrane. J. Cell Sci. 112 ( Pt 11), 1785–1795.

Büttner, F.M., Renner-Schneck, M., Stehle, T., 2015. X-ray crystallography and its impact on understanding bacterial cell wall remodeling processes. Int. J. Med. Microbiol. 305, 209–216. https://doi.org/10.1016/j.ijmm.2014.12.018

Chattopadhyay, A., Jones, N.G., Nietlispach, D., Nielsen, P.R., Voorheis, H.P., Mott, H.R., Carrington, M., 2005. Structure of the C-terminal Domain from Trypanosoma brucei Variant Surface Glycoprotein MITat1.2 *. J. Biol. Chem. 280, 7228–7235. https://doi.org/10.1074/jbc.M410787200

Christiano, R., Kolev, N.G., Shi, H., Ullu, E., Walther, T.C., Tschudi, C., 2017. The proteome and transcriptome of the infectious metacyclic form of *Trypanosoma brucei* define quiescent cells primed for mammalian invasion: Metacyclic *Trypanosoma brucei* cells are quiescent. Mol. Microbiol. 106, 74–92. https://doi.org/10.1111/mmi.13754

Clayton, C.E., Mowatt, M.R., 1989. The Procyclic Acidic Repetitive Proteins of Trypanosoma brucei. J. Biol. Chem. 264, 15088–15093. https://doi.org/10.1016/S0021-9258(18)63815-3

Cross, G.A., 1975. Identification, purification and properties of clone-specific glycoprotein antigens constituting the surface coat of Trypanosoma brucei. Parasitology 71, 393–417. https://doi.org/10.1017/s003118200004717x

Cross, G.A.M., 1984. Release and purification of trypanosoma brucei variant surface glycoprotein. J. Cell. Biochem. 24, 79–90. https://doi.org/10.1002/jcb.240240107

Cross, G.A.M., Kim, H.-S., Wickstead, B., 2014. Capturing the variant surface glycoprotein repertoire (the VSGnome) of Trypanosoma brucei Lister 427. Mol. Biochem. Parasitol. 195, 59–73. https://doi.org/10.1016/j.molbiopara.2014.06.004

Cully, D.F., Ip, H.S., 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. https://doi.org/10.1016/S0092-8674(85)80113-6

Dean, S., Marchetti, R., Kirk, K., Matthews, K.R., 2009. A surface transporter family conveys the trypanosome differentiation signal. Nature 459, 213–217. https://doi.org/10.1038/nature07997

Emsley, P., Cowtan, K., 2004. Coot: model-building tools for molecular graphics. Acta Crystallogr. D Biol. Crystallogr. 60, 2126–2132. https://doi.org/10.1107/S0907444904019158

Ferguson, M.A., Murray, P., Rutherford, H., McConville, M.J., 1993. A simple purification of procyclic acidic repetitive protein and demonstration of a sialylated glycosyl-phosphatidylinositol membrane anchor. Biochem. J. 291, 51–55.

Ferrante, A., Allison, A.C., 1983. Alternative pathway activation of complement by African trypanosomes lacking a glycoprotein coat. Parasite Immunol. 5, 491–498. https://doi.org/10.1111/j.1365-3024.1983.tb00763.x

Field, M.C., Menon, A.K., Cross, G.A., 1991. Developmental variation of glycosylphosphatidylinositol membrane anchors in Trypanosoma brucei. Identification of a candidate biosynthetic precursor of the glycosylphosphatidylinositol anchor of the major procyclic stage surface glycoprotein. J. Biol. Chem. 266, 8392–8400. https://doi.org/10.1016/S0021-9258(18)92988-1

Franco, J.R., Cecchi, G., Priotto, G., Paone, M., Diarra, A., Grout, L., Simarro, P.P., Zhao, W., Argaw, D., 2020. Monitoring the elimination of human African trypanosomiasis at continental and country level: Update to 2018. PLoS Negl. Trop. Dis. 14, e0008261. https://doi.org/10.1371/journal.pntd.0008261

Freymann, D., Down, J., Carrington, M., Roditi, I., Turner, M., Wiley, D., 1990. 2·9 Å resolution structure of the N-terminal domain of a variant surface glycoprotein from Trypanosoma brucei. J. Mol. Biol. 216, 141–160. https://doi.org/10.1016/S0022-2836(05)80066-X

Gkeka, A., Aresta-Branco, F., Triller, G., Stebbins, C. E., 2021. Immunodomminant surface epitopes power immune evasion in the African trypanosome. (in preparation).

Graham, S.V., Barry, J.D., 1995. Transcriptional regulation of metacyclic variant surface glycoprotein gene expression during the life cycle of Trypanosoma brucei. Mol. Cell. Biol. 15, 5945–5956. https://doi.org/10.1128/MCB.15.11.5945

Graham, S.V., Matthews, K.R., Shiels, P.G., Barry, J.D., 1990. Distinct, developmental stage-specific activation mechanisms of trypanosome VSG genes. Parasitology 101 Pt 3, 361–367. https://doi.org/10.1017/s0031182000060558

Grünfelder, C.G., Engstler, M., Weise, F., Schwarz, H., Stierhof, Y.-D., Boshart, M., Overath, P., 2002. Accumulation of a GPI-anchored protein at the cell surface requires sorting at multiple intracellular levels. Traffic Cph. Den. 3, 547–559. https://doi.org/10.1034/j.1600-0854.2002.30805.x

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., Bangs, J.D., 2006. Regulation of surface coat exchange by differentiating African trypanosomes. Mol. Biochem. Parasitol. 147, 211–223. https://doi.org/10.1016/j.molbiopara.2006.02.013

Günzl, A., Bruderer, T., Laufer, G., Schimanski, B., Tu, L.-C., Chung, H.-M., Lee, P.-T., Lee, M.G.-S., 2003. RNA polymerase I transcribes procyclin genes and variant surface glycoprotein gene expression sites in Trypanosoma brucei. Eukaryot. Cell 2, 542–551. https://doi.org/10.1128/ec.2.3.542-551.2003

Günzl, A., Kirkham, J.K., Nguyen, T.N., Badjatia, N., Park, S.H., 2015. Mono-allelic VSG expression by RNA polymerase I in Trypanosoma brucei: expression site control from both ends? Gene 556, 68–73. https://doi.org/10.1016/j.gene.2014.09.047

Helm, J.R., Hertz-Fowler, C., Aslett, M., Berriman, M., Sanders, M., Quail, M.A., Soares, M.B., Bonaldo, M.F., Sakurai, T., Inoue, N., Donelson, J.E., 2009. Analysis of expressed sequence tags from the four main developmental stages of Trypanosoma congolense. Mol. Biochem. Parasitol. 168, 34–42. https://doi.org/10.1016/j.molbiopara.2009.06.004

Hempelmann, A., Hartleb, L., van Straaten, M., Hashemi, H., Zeelen, J.P., Papavasiliou, F.N., Engstler, M., Stebbins, C.E., Jones, N.G., 2021. Nanobody Mediated Macromolecular Crowding Induces Membrane Fission and Remodeling in the African Trypanosome (preprint). Microbiology. https://doi.org/10.1101/2021.01.13.426364

Hertz-Fowler, C., Figueiredo, L.M., Quail, M.A., Becker, M., Jackson, A., Bason, N., Brooks, K., Churcher, C., Fahkro, S., Goodhead, I., Heath, P., Kartvelishvili, M., Mungall, K., Harris, D., Hauser, H., Sanders, M., Saunders, D., Seeger, K., Sharp, S., Taylor, J.E., Walker, D., White, B., Young, R., Cross, G.A.M., Rudenko, G., Barry, J.D., Louis, E.J., Berriman, M., 2008. Telomeric expression sites are highly conserved in Trypanosoma brucei. PloS One 3, e3527. https://doi.org/10.1371/journal.pone.0003527

Hovel-Miner, G., Mugnier, M., Papavasiliou, F.N., Pinger, J., Schulz, D., 2015. A Host-Pathogen Interaction Reduced to First Principles: Antigenic Variation in T. brucei. Results Probl. Cell Differ. 57, 23–46. https://doi.org/10.1007/978-3-319-20819-0_2

Jackson, D.G., Owen, M.J., Voorheis, H.P., 1985. A new method for the rapid purification of both the membrane-bound and released forms of the variant surface glycoprotein from Trypanosoma brucei. Biochem. J. 230, 195–202. https://doi.org/10.1042/bj2300195

Kantardjieff, K.A., Rupp, B., 2003. Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein–nucleic acid complex crystals. Protein Sci. 12, 1865–1871. https://doi.org/10.1110/ps.0350503

Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., Sternberg, M.J.E., 2015. The Phyre2 web portal for protein modeling, prediction and analysis. Nat. Protoc. 10, 845–858. https://doi.org/10.1038/nprot.2015.053

Kolev, N.G., Ramey-Butler, K., Cross, G.A.M., Ullu, E., Tschudi, C., 2012. Developmental Progression to Infectivity in Trypanosoma Brucei Triggered by an Rna-Binding Protein. Science 338, 1352–1353. https://doi.org/10.1126/science.1229641

Krissinel, E., Henrick, K., 2004. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr. D Biol. Crystallogr. 60, 2256–2268. https://doi.org/10.1107/S0907444904026460

Lemley, C., Yan, S., Dole, V.S., Madhubala, R., Cunningham, M.L., Beverley, S.M., Myler, P.J., Stuart, K.D., 1999. The Leishmania donovani LD1 locus gene ORFG encodes a biopterin transporter (BT1). Mol. Biochem. Parasitol. 104, 93–105. https://doi.org/10.1016/S0166-6851(99)00132-2

Lindner, A.K., Lejon, V., Chappuis, F., Seixas, J., Kazumba, L., Barrett, M.P., Mwamba, E., Erphas, O., Akl, E.A., Villanueva, G., Bergman, H., Simarro, P., Kadima Ebeja, A., Priotto, G., Franco, J.R., 2020. New WHO guidelines for treatment of gambiense human African trypanosomiasis including fexinidazole: substantial changes for clinical practice. Lancet Infect. Dis. 20, e38–e46. https://doi.org/10.1016/S1473-3099(19)30612-7

Marcello, L., Barry, J.D., 2007. Analysis of the VSG gene silent archive in Trypanosoma brucei reveals that mosaic gene expression is prominent in antigenic variation and is favored by archive substructure. Genome Res. 17, 1344–1352. https://doi.org/10.1101/gr.6421207

Matthews, B.W., 1968. Solvent content of protein crystals. J. Mol. Biol. 33, 491–497. https://doi.org/10.1016/0022-2836(68)90205-2

Morris, J.C., Wang, Z., Drew, M.E., Englund, P.T., 2002. Glycolysis modulates trypanosome glycoprotein expression as revealed by an RNAi library. EMBO J. 21, 4429–4438. https://doi.org/10.1093/emboj/cdf474

Mowatt, M.R., Clayton, C.E., 1987. Developmental regulation of a novel repetitive protein of Trypanosoma brucei. Mol. Cell. Biol. 7, 2838–2844. https://doi.org/10.1128/mcb.7.8.2838

Mugnier, M.R., Cross, G.A.M., Papavasiliou, F.N., 2015. The in vivo dynamics of antigenic variation in Trypanosoma brucei. Science 347, 1470–1473. https://doi.org/10.1126/science.aaa4502

Mugnier, M.R., Stebbins, C.E., Papavasiliou, F.N., 2016. Masters of Disguise: Antigenic Variation and the VSG Coat in Trypanosoma brucei. PLOS Pathog. 12, e1005784. https://doi.org/10.1371/journal.ppat.1005784

Nilsson, D., Gunasekera, K., Mani, J., Osteras, M., Farinelli, L., Baerlocher, L., Roditi, I., Ochsenreiter, T., 2010. Spliced Leader Trapping Reveals Widespread Alternative Splicing Patterns in the Highly Dynamic Transcriptome of Trypanosoma brucei. PLOS Pathog. 6, e1001037. https://doi.org/10.1371/journal.ppat.1001037

Pays, E., Vanhollebeke, B., Vanhamme, L., Paturiaux-Hanocq, F., Nolan, D.P., Pérez-Morga, D., 2006. The trypanolytic factor of human serum. Nat. Rev. Microbiol. 4, 477–486. https://doi.org/10.1038/nrmicro1428

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., Ferrin, T.E., 2004. UCSF Chimera--a visualization system for exploratory research and analysis. J. Comput. Chem. 25, 1605–1612. https://doi.org/10.1002/jcc.20084

Pinger, J., Nešić, D., Ali, L., Aresta-Branco, F., Lilic, M., Chowdhury, S., Kim, H.-S., Verdi, J., Raper, J., Ferguson, M.A.J., Papavasiliou, F.N., Stebbins, C.E., 2018. African trypanosomes evade immune clearance by O -glycosylation of the VSG surface coat. Nat. Microbiol. 3, 932–938. https://doi.org/10.1038/s41564-018-0187-6

Potterton, L., Agirre, J., Ballard, C., Cowtan, K., Dodson, E., Evans, P.R., Jenkins, H.T., Keegan, R., Krissinel, E., Stevenson, K., Lebedev, A., McNicholas, S.J., Nicholls, R.A., Noble, M., Pannu, N.S., Roth, C., Sheldrick, G., Skubak, P., Turkenburg, J., Uski, V., von Delft, F., Waterman, D., Wilson, K., Winn, M., Wojdyr, M., 2018. CCP4i2: the new graphical user interface to the CCP4 program suite. Acta Crystallogr. Sect. Struct. Biol. 74, 68–84. https://doi.org/10.1107/S2059798317016035

Ramey-Butler, K., Ullu, E., Kolev, N.G., Tschudi, C., 2015. Synchronous expression of individual metacyclic variant surface glycoprotein genes in Trypanosoma brucei. Mol. Biochem. Parasitol. 200, 1–4. https://doi.org/10.1016/j.molbiopara.2015.04.001

Richardson, J.P., Beecroft, R.P., Tolson, D.L., Liu, M.K., Pearson, T.W., 1988. Procyclin: an unusual immunodominant glycoprotein surface antigen from the procyclic stage of African trypanosomes. Mol. Biochem. Parasitol. 31, 203–216. https://doi.org/10.1016/0166-6851(88)90150-8

Rico, E., Rojas, F., Mony, B.M., Szoor, B., MacGregor, P., Matthews, K.R., 2013. Bloodstream form pre-adaptation to the tsetse fly in Trypanosoma brucei. Front. Cell. Infect. Microbiol. 3. https://doi.org/10.3389/fcimb.2013.00078

Roditi, I., Clayton, C., 1999. An unambiguous nomenclature for the major surface glycoproteins of the procyclic form of Trypanosoma brucei. Mol. Biochem. Parasitol. 103, 99–100. https://doi.org/10.1016/S0166-6851(99)00124-3

Roditi, I., Furger, A., Ruepp, S., Schürch, N., Bütikofer, P., 1998. Unravelling the procyclin coat of Trypanosoma brucei. Mol. Biochem. Parasitol. 91, 117–130. https://doi.org/10.1016/s0166-6851(97)00195-3

Ross, D.T., Raibaud, A., Florent, I.C., Sather, S., Gross, M.K., Storm, D.R., Eisen, H., 1991. The trypanosome VSG expression site encodes adenylate cyclase and a leucine-rich putative regulatory gene. EMBO J. 10, 2047–2053.

Salmon, D., Hanocq-Quertier, J., Paturiaux-Hanocq, F., Pays, A., Tebabi, P., Nolan, D.P., Michel, A., 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. EMBO J. 16, 7272–7278. https://doi.org/10.1093/emboj/16.24.7272

Schell, D., Evers, R., Preis, D., Ziegelbauer, K., Kiefer, H., Lottspeich, F., Cornelissen, A.W., Overath, P., 1991. A transferrin-binding protein of Trypanosoma brucei is encoded by one of the genes in the variant surface glycoprotein gene expression site. EMBO J. 10, 1061–1066. https://doi.org/10.1002/j.1460-2075.1991.tb08045.x

Steverding, D., 2010. The development of drugs for treatment of sleeping sickness: a historical review. Parasit. Vectors 3, 15. https://doi.org/10.1186/1756-3305-3-15

Swallow, B.M., 1999. Impact of trypanosomiasis on African agriculture. OAU/STRC.

Tarentino, A.L., Plummer, T.H., 1994. [4] Enzymatic deglycosylation of asparagine-linked glycans: Purification, properties, and specificity of oligosaccharide-cleaving enzymes from Flavobacterium meningosepticum, in: Methods in

Enzymology, Guide to Techniques in Glycobiology. Academic Press, pp. 44–57. https://doi.org/10.1016/0076-6879(94)30006-2

Trypanosomiasis, human African (sleeping sickness) [WWW Document], n.d. URL https://www.who.int/news-room/fact-sheets/detail/trypanosomiasis-human-african-(sleeping-sickness) (accessed 4.13.21).

Tsiolaki, P.L., Louros, N.N., Iconomidou, V.A., 2018. Hexapeptide Tandem Repeats Dictate the Formation of Silkmoth Chorion, a Natural Protective Amyloid. J. Mol. Biol., Functional Amyloids in Health and Disease 430, 3774–3783. https://doi.org/10.1016/j.jmb.2018.06.042

Umaer, K., Aresta-Branco, F., Waxman, B., Chandra, M., van Straaten, M., Zeelen, J., Stebbins, C. E., Bangs, J. D., 2021. Dynamic Oligomerization and the Trafficking of Variant Surface Glycoproteins of Trypanosoma brucei (submitted)

Urwyler, S., Studer, E., Renggli, C.K., Roditi, I., 2007. A family of stage-specific alanine-rich proteins on the surface of epimastigote forms of Trypanosoma brucei. Mol. Microbiol. 63, 218–228. https://doi.org/10.1111/j.1365-2958.2006.05492.x

Van Den Abbeele, J., Claes, Y., van Bockstaele, D., Le Ray, D., Coosemans, M., 1999. Trypanosoma brucei spp. development in the tsetse fly: characterization of the post-mesocyclic stages in the foregut and proboscis. Parasitology 118 ( Pt 5), 469–478. https://doi.org/10.1017/s0031182099004217

Vanhamme, L., Pays, E., 1995. Control of Gene Expression in Trypanosomes. MICROBIOL REV 59, 18.

Vassella, E., Brun, R., Roditi, I., n.d. A major surface glycoprotein of Trypanosoma brucei is expressed transiently during development and can be regulated post-transcriptionally by glycerol or hypoxia 12.

Walker, J.M. (Ed.), 2005. The Proteomics Protocols Handbook. Humana Press. https://doi.org/10.1385/1592598900

Wang, S., Ma, J., Peng, J., Xu, J., 2013. Protein structure alignment beyond spatial proximity. Sci. Rep. 3, 1448. https://doi.org/10.1038/srep01448

Wang, S., Peng, J., Xu, J., 2011. Alignment of distantly related protein structures: algorithm, bound and implications to homology modeling. Bioinformatics 27, 2537–2545. https://doi.org/10.1093/bioinformatics/btr432

Weichenberger, C.X., Rupp, B., 2014. Ten years of probabilistic estimates of biocrystal solvent content: new insights via nonparametric kernel density estimate. Acta Crystallogr. D Biol. Crystallogr. 70, 1579–1588. https://doi.org/10.1107/S1399004714005550

Weirather, J.L., Wilson, M.E., Donelson, J.E., 2012. Mapping of VSG similarities in Trypanosoma brucei. Mol. Biochem. Parasitol. 181, 141–152. https://doi.org/10.1016/j.molbiopara.2011.10.011

Wiedemar, N., Graf, F.E., Zwyer, M., Ndomba, E., Renggli, C.K., Cal, M., Schmidt, R.S., Wenzler, T., Mäser, P., 2018. Beyond immune escape: a variant surface glycoprotein causes suramin resistance in Trypanosoma brucei. Mol. Microbiol. 107, 57–67. https://doi.org/10.1111/mmi.13854

Xu, J., Zhang, Y., 2010. How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics 26, 889–895. https://doi.org/10.1093/bioinformatics/btq066

Zeelen, J., van Straaten, M., Verdi, J., Hempelmann, A., Hashemi, H., Perez, K., Jeffrey, P.D., Hälg, S., Wiedemar, N., Mäser, P., Papavasiliou, F.N., Stebbins, C.E., 2021. Structure of trypanosome coat protein VSGsur and function in suramin resistance. Nat. Microbiol. 6, 392–400. https://doi.org/10.1038/s41564-020-00844-1

Zemla, A., 2003. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res. 31, 3370–3374.

Zhang, Y., 2005. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 33, 2302–2309. https://doi.org/10.1093/nar/gki524

Zimmermann, H., Subota, I., Batram, C., Kramer, S., Janzen, C.J., Jones, N.G., Engstler, M., 2017. A quorum sensing-independent path to stumpy development in Trypanosoma brucei. PLOS Pathog. 13, e1006324. https://doi.org/10.1371/journal.ppat.1006324

## Appendixes

### Appendix 1

***Data collection and refinement statistics for the mVSG531 structure.*** The table is generated by PHENIX crystallography software suite (Adams et al., 2010)

| Parameters | Values |
|---|---|
| **Wavelength** | 1.000 Å |
| **Resolution range** | 59.09 - 1.949 (2.019 - 1.949) |
| **Space group** | P 1 21 1 |
| **Unit cell** | 50.1069 170.885 163.652 90 90.7052 90 |
| **Total reflections** | 592350 (57625) |
| **Unique reflections** | 195472 (18665) |
| **Multiplicity** | 3.0 (3.1) |
| **Completeness (%)** | 97.68 (93.16) |
| **Mean I/sigma(I)** | 4.24 (1.13) |
| **Wilson B-factor** | 34.64 |
| **R-merge** | 0.1157 (0.7639) |
| **R-meas** | 0.1412 (0.9266) |
| **R-pim** | 0.07975 (0.5179) |
| **CC1/2** | 0.977 (0.629) |
| **CC*** | 0.994 (0.879) |
| **Reflections used in refinement** | 195256 (18624) |
| **Reflections used for R-free** | 9605 (847) |
| **R-work** | 0.2134 (0.3583) |
| **R-free** | 0.2438 (0.3808) |

| | |
|---|---|
| **CC(work)** | 0.916 (0.772) |
| **CC(free)** | 0.904 (0.745) |
| **Number of non-hydrogen atoms** | 22693 |
| **macromolecules** | 20778 |
| **ligands** | 618 |
| **solvent** | 1339 |
| **Protein residues** | 2866 |
| **RMS(bonds)** | 0.006 |
| **RMS(angles)** | 1.16 |
| **Ramachandran favored (%)** | 96.78 |
| **Ramachandran allowed (%)** | 3.22 |
| **Ramachandran outliers (%)** | 0.00 |
| **Rotamer outliers (%)** | 0.25 |
| **Clashscore** | 1.81 |
| **Average B-factor** | 46.61 |
| **macromolecules** | 47.27 |
| **ligands** | 30.97 |
| **solvent** | 43.09 |
| **Number of TLS groups** | 45 |

Statistics for the highest-resolution shell are shown in parentheses.

**Appendix 2**

***Data collection and refinement statistics for the mVSG1954structure.*** The table is generated by PHENIX crystallography software suite (Adams et al., 2010)

| Parameters | Values |
|---|---|
| **Wavelength** | 0.9198 Å |
| **Resolution range** | 60.29 -1.681 (1.741 -1.681) |
| **Space group** | P 3 2 1 |
| **Unit cell** | 69.6141 69.6141 115.36 90 90 120 |
| **Total reflections** | 723113 (71206) |
| **Unique reflections** | 37543 (3705) |
| **Multiplicity** | 19.3 (19.2) |
| **Completeness (%)** | 99.79 (97.89) |
| **Mean I/sigma(I)** | 7.45 (0.48) |
| **Wilson B-factor** | 19.13 |
| **R-merge** | 0.1453 (1.283) |
| **R-meas** | 0.1492 (1.318) |
| **R-pim** | 0.03383 (0.2986) |
| **CC1/2** | 0.999 (0.772) |
| **CC*** | 1 (0.934) |
| **Reflections used in refinement** | 37464 (3627) |
| **Reflections used for R-free** | 1877 (201) |
| **R-work** | 0.1757 (0.3505) |
| **R-free** | 0.2137 (0.3762) |
| **CC(work)** | 0.968 (0.829) |

| | |
|---|---|
| **CC(free)** | 0.971 (0.783) |
| **Number of non-hydrogen atoms** | 2798 |
| **macromolecules** | 2514 |
| **ligands** | 62 |
| **solvent** | 250 |
| **Protein residues** | 353 |
| **RMS(bonds)** | 0.011 |
| **RMS(angles)** | 1.34 |
| **Ramachandran favored (%)** | 98.84 |
| **Ramachandran allowed (%)** | 1.16 |
| **Ramachandran outliers (%)** | 0.00 |
| **Rotamer outliers (%)** | 0.41 |
| **Clashscore** | 3.19 |
| **Average B-factor** | 29.42 |
| **macromolecules** | 28.74 |
| **ligands** | 61.13 |
| **solvent** | 31.97 |
| **Number of TLS groups** | 5 |

Statistics for the highest-resolution shell are shown in parentheses.