University of Nottingham
UK | CHINA | MALAYSIA

Exploring the surface of *Trypanosoma brucei* through protein sorting and recombinant expression

Thomas Miller

MSci (Hons)

Thesis submitted to the University of Nottingham towards the degree of Doctor of Philosophy

School of Life Sciences

September 2019

Contents

I. Abstract

Surface membrane structure and composition define the biological niche of a unicellular organism, controlling how it interacts with and survives in its environment. The human and animal pathogen *Trypanosoma brucei* lives extracellularly in the blood of its mammalian host, where it must evade continual surveillance by the immune system whilst obtaining nutrients required for survival. It achieves this through antigenic variation of its major surface protein (the glycosylphosphatidylinositol (GPI)-anchored VSG) and surface compartmentalisation, retaining transporters and receptors essential for the uptake of nutrient in a specialised membrane invagination at the base of its flagellum – the flagellar pocket (the sole site of endocytosis and secretion in the parasite). This PhD exploits a high-confidence, validated cell surface proteome ('surfeome') for bloodstream-form *T. brucei* to test hypotheses about GPI-anchored protein sorting and flagellar pocket retention. It also attempts to contribute towards early-stage development of strategies for disease control through the recombinant production of surfeome components for testing as vaccine candidates.

It has been proposed that sorting of trypanosome surface proteins to their target membrane domain is influenced by protein abundance, glycosylation, or the number of GPI anchors attached ('GPI valence'). However, none of these hypotheses is sufficient to explain what we now know about the parasite cell surface. Instead, this project tests if the information required to direct GPI-anchored protein sorting is intrinsic to the GPI-insertion signal sequence itself. The GPI signal sequences from five *T. brucei* surface proteins (that localise to different domains on the parasite surface) were fused to exogenous fluorescent reporter proteins. These signal sequences allowed correct GPI attachment, but did not result in the differential localisations of the respective endogenous proteins, with all fusions diffused across the entire cell surface membrane. Significantly, results presented herein are incompatible with GPI valance being the primary mode of sorting of GPI-anchored proteins, raising further questions as to what controls the delivery of membrane components to the appropriate target membrane domain.

No vaccines exist for protection against African trypanosomiasis. For the development of an effective vaccine, native-like recombinant antigens must be produced and purified. Initial experiments in this PhD project used the commercially available, *Leishmania tarentolae*-based system LEXSY; but its underperformance led to the development of a novel system based on *Crithidia fasciculata* (CExSy). A single marker *C. fasciculata* line (SMC) that expresses the T7 RNA polymerase and the tetracycline repressor protein was generated, along with a suite of plasmids that allowed production of >10 milligrams of GFP per litre of cell culture. Subsequent expression of three invariant, surface-exposed *T. brucei* antigens enabled characterisation of glycosylation status and isolation of high purity protein. This system may prove useful for downstream biochemical, structural and pre-clinical applications.

II. Acknowledgments

Firstly, I would like to give a huge thank you to my supervisor Dr. Catarina Gadelha for her continued support and guidance throughout this process, without which this would not have been possible. Her diligence to teaching has made me a better scientist and writer. I will forever know how to use a compound adjective, even if the hyphen still manages to go mysteriously missing at times! I would also like to thank Dr. Bill Wickstead; his input and advice has been invaluable during this project.

Thank you to all the members of the Allers, Friel, Gadelha, Huett, and Wickstead labs for making the last four years enjoyable. In particular, special thanks goes to Simon for always providing help and encouragement when needed; to James for keeping me sane with the climbing sessions and Formula 1; to Sarah for her ‘crackin’ sense of humour; to Rob for keeping me topped up with coffee and chillies during thesis write up; and to Georgina for her wisdom and infectious positive outlook that helped keep me going through the final months of the PhD and the thesis.

A big thank you goes to my girlfriend Sarah, who has had to put up with me through my best and my worst moments. She has supported me through the last six years, including the final years of my undergraduate course and now my entire PhD project. I’m sure she will be grateful to having a bit more help around the house now the thesis is done! Finally, I would like to thank my family for their continued unconditional support.

This work was supported by a BBSRC DTP programme stipend. Work in the Gadelha lab is funded by the MRC.

III. Declaration

This thesis is the result of my own work, except where included data is explicitly mentioned, which has been undertaken during my period of registration for this degree at The University of Nottingham.

IV. Abbreviations

AAT	Animal African trypanosomiasis
AIS	Axonal initial segment
APOL1	Apolipoprotein L1
AQP	Aquaglyceroporin
Atg	Autophagy-related
BCA	Bicinchoninic Acid
BHI	Brain heart infusion
BSF	Bloodstream form
CATT	Card agglutination test for trypanosomiasis
CExSy	Crithidial expression system
CHO	Chinese hamster ovary
CnBr	Cyanogen bromide
Cryo-EM	Cryogenic electron microscopy
CTD	C-terminal domain
DNA	Deoxyribonucleic acid
rDNA	Ribosomal DNA
Dol-P-Man	Dolichol-phosphate-mannose
Endo-H	endoglycosidase-H
ERAD	ER associated protein degradation
ERES	ER exit sites
ESAG	Expression site associated gene
ESP	Enriched in surface-labelled proteome
FAZ	Flagellum attachment zone
FBS	Foetal bovine serum
FP	Flagellar pocket
GDI	Guanine nucleotide dissociation inhibitor
GalNAc	N-acetylagalactosamine
GIPL	Glycoinositolphospholipids
GlcNAc	N-acetylglucosamine
GPCR	G protein-coupled receptor
GPIAP	Glycosylphosphatidylinositol-anchored protein

GPI-MT	GPI-α 1-4-mannosyltransferase
GPI-PLC	Glycosylphosphatidylinositol phospholipase C
GPISS	Glycosylphosphatidylinositol signal sequence
GPI-T	GPI transamidation complex
GRESAG	Gene related to expression site associated gene
GSPS	Glutathionylspermidine synthase
HAT	Human African Trypanosomiasis
HC	Hook complex
HEK	Human embryonic kidney
Hh	Hedgehog
HpHbR	Haptoglobin-haemoglobin receptor
HRP	Horseradish peroxidase
HYG^R	Hygromycin phosphotransferase
IDT	Integrated DNA Technologies
IGS	Intergenic spacer
IMAC	Immobilized metal affinity chromatography
IMDM	Iscove's Modified Dulbecco's Medium
ISG	Invariant surface glycoprotein
LAC	Linear artificial chromosome
LAG	Lipoarabinogalactans
LEXSY	Leishmanial expression system
LPG	Lipophosphoglycans
MCS	Multiple cloning site
MDCK	Madin-Darby canine kidney
MMEJ	Microhomology-mediated end joining
MW	Molecular weight
mSca-I	mScarlet-I
NAT	Nourseothricin N-acetyltransferase
Ni-NTA	Nickel-nitrilotriacetic acid
NEO^R	Neomycin phosphotransferase
NTD	N-terminal domain
ORF	Open reading frame
OST	Oligosaccharyltransferase complex
PAD	Proteins associated with differentiation

PAGE	Polyacrylamide gel electrophoresis
PBS	Phosphate buffered saline
PCF	Procyclic form
PFR	Paraflagellar rod
PGKA	Phosphoglycerate kinase A
PGKB	Phosphoglycerate kinase B
PI	Phosphatidylinositol
PIC	Protease inhibitor cocktail
pNAL	Poly-N-acetyl lactosamine
PT	Pyruvate transporter
pNAL	Poly-N-acetyllactosamine
PNGase-F	Peptide-N-glycosidase F
PUR^R	Puromycin N-acetyltransferase
PSG	Phosphate saline glucose
RBP	RNA binding protein
RNA	Ribonucleic acid
RSCU	Relative synonymous codon usage
SDS	Sodium dodecyl sulphate
sfGFP	Superfolder GFP
SMC	Single marker *Crithidia*
SMO	Smoothened
SP	Signal peptide
SRA	Serum resistance-associated protein
SRP	Signal recognition particle
T7RNAP	T7 RNA polymerase
TBS-T	Tris-buffered saline with tween-20
tet	Tetracycline
TetO	Tetracycline operator
TETR	Tetracycline repressor
TEV	Tobacco Etch Virus
TfR	Transferrin receptor
TIS	Transcription initiation site
TL	Tomato lectin
TUBA	α-Tubulin

TUBB	β-Tubulin
UGGT	UDP-glucose:glycoprotein glucosyltransferase
UTR	Untranslated region
VSG	Variant surface glycoprotein

Chapter 1. Introduction

Lipid-bilayer membranes are essential to all cellular life. They are the molecular chassis that define cellular boundaries, maintain intracellular physiological conditions, and restrict cellular components to their required locations for the execution of crucial biological functions. In addition to phospholipids, membranes contain a diverse range of proteins that evolved for different roles: signalling, transport of molecules and ions across membranes, cellular and organellar structural integrity, cell adhesion and mobility, communication and metabolism. Specific lipid and protein composition dictate the characteristics and biochemical properties of biological membranes.

1.1. Membrane protein function and association

Membrane protein function is heavily influenced by its association to the lipid bilayer. Integral membrane proteins (Figure 1-1, examples 1-3) span the entire phospholipid bilayer and, thus, can act on either side of the membrane, featuring in roles such as transport of molecules across membranes and signalling between different environments. These proteins can be subdivided into single pass proteins, utilising a single hydrophobic transmembrane α-helix for membrane insertion (Figure 1-1, example 1), or multipass proteins, spanning the membrane twice or more with either β-barrel structures or multiple α-helices (Figure 1-1, examples 2 and 3). Single-pass proteins are called type I or type II depending on whether they are orientated with their C-terminus or N-terminus towards the cytoplasm respectively.

In contrast to integral membrane proteins, peripheral ones associate with only one leaflet of the phospholipid bilayer (Figure 1-1, examples 4–8). Hence, these proteins can only act on one side of the membrane, often acting in intracellular signalling, working in tandem with integral membrane proteins to pass on extracellular signals. Peripheral membrane association is usually not intrinsic to the protein’s amino acid structure and relies instead upon either interactions with other membrane-associated proteins, or on the attachment of

different lipid modifications (Figure 1-1, examples 5–8) (although some peripheral proteins utilise amphipathic α-helices for membrane interaction, with the hydrophobic face of the α-helix able to insert into the membrane, whilst the hydrophilic face interacts with the phospholipid head groups; Figure 1-1, example 4). Regarding attachment via a lipid moiety (Figure 1-1, examples 5 and 6), peripheral proteins can be modified with a range of different lipid groups, including fatty acids, isoprenoids, sterols, phospholipids and glycosyl phosphatidylinositol (GPI) anchors (Resh, 2016, 2013). These different lipid moieties are covered in more detail below.

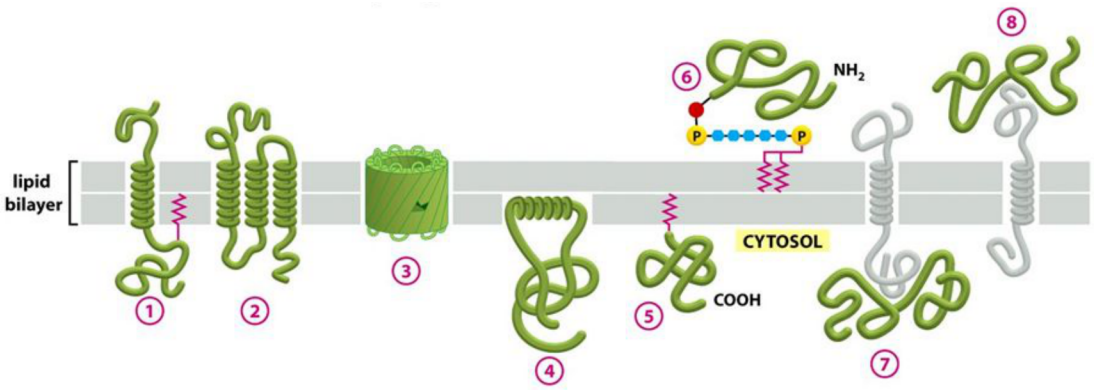


Figure 1-1 Structures by which proteins associate to membranes. Integral membrane proteins span the lipid bilayer with a single α-helix (1), multiple α-helices (2) or a β-barrel structure (3). Peripheral membrane proteins associate to the lipid bilayer through an amphipathic α-helix (4), through lipid modifications (5 and 6) or through interaction with another membrane associated protein (7–8). Lipid modifications include fatty acids, isoprenoids, sterols, phospholipids (5) and GPI anchors (6). Figure from Alberts *et al.*, 2008a.

1.1.1. Fatty acids

Fatty acid modifications on proteins vary in length from 8 to more than 20 carbons long (Resh, 2016). Most common are 16-carbon palmitate, and 14-carbon myristate moieties. Palmitate is reversibly attached to cysteine residues, usually on the cytoplasmic face of the ER, Golgi or plasma membrane. The modification dictates protein-membrane association, lipid raft localisation, trafficking and stability. Reversibility of palmitoylation allows regulation of protein localisation which can be important for function. H- and N-Ras are well-studied examples of this (Ahearn *et al.*, 2011b). Ras proteins are GTPases that

function in the signalling and regulation of multiple cellular processes. Reversible palmitoylation (along with irreversible farnesylation, see section 1.1.2) of H- and N-Ras allows cycling between the Golgi and the plasma membrane. Once at the plasma membrane, palmitate groups can be removed from Ras proteins (regulated by the isomerase FKBP12 and the deacylase ABHD17; Ian M. Ahearn *et al.*, 2011; Lin and Conibear, 2015), releasing them from the membrane and allowing recycling back to the Golgi. Evidence suggests that Ras signalling from the Golgi and plasma membranes differ (in terms of downstream pathway response), thus, reversible palmitoylation is believed to regulate signalling.

The myristate modification differs somewhat from palmitate. The lipid is attached in the cytosol to an N-terminal glycine, usually co-translationally (after removal of the preceding methionine residue) (Resh, 2016). As a shorter lipid moiety, myristate cannot stably associate proteins to membranes without the aid of other lipids (such as palmitate), hydrophobic amino acids, or positively charged amino acids for electrostatic interaction with negatively charged phospholipids. Additionally, unlike palmitate, the attachment of myristate is not reversible. Instead, to regulate membrane interaction, myristoylated proteins rely either on attachment and removal of additional palmitate moieties, or on the regulation of myristoyl switches. These switches can inhibit membrane interaction through three main mechanisms: conformational changes to sequester myristate moieties in hydrophobic clefts; multimerization of proteins to similarly block myristate-membrane interactions; or phosphorylation of basic amino acids to inhibit association-supporting interactions between the protein and charged phospholipids.

1.1.2. Isoprenoids

Isoprenoids (specifically 15-carbon farnesyl and 20-carbon geranylgeranyl) are attached to proteins at or near the C-terminus (Resh, 2013). This occurs post-translationally in the cytosol. Much like the case of palmitate, geranylgeranyl groups are long enough to stably associate proteins to membranes; whilst the shorter farnesyl groups require additional lipids or

charged residues for stable membrane association. Reversible palmitoylation of farnesylated proteins is used to regulate membrane attachment, as in the case of H- and N-Ras above, whereby the farnesyl group acts to enable the initial weak association of the proteins to the Golgi membrane for subsequent palmitoylation (Ahearn *et al.*, 2011b). In contrast to Ras GTPases, geranylgeranylated proteins such as Rabs (also signalling GTPases) can be removed from membranes through interactions with guanine nucleotide dissociation inhibitor proteins (GDIs) (Nishimura and Linder, 2013). GDIs have hydrophobic grooves that can bind to geranylgeranyl groups, leading to membrane dissociation and inactivation of Rab.

1.1.3. Sterols and phospholipids

Membrane association of proteins through covalent attachment of sterols or phospholipids is less well documented. Proteins reported to be directly attached to phospholipids are members of the autophagy-related protein 8 (Atg8) family (Bonnon *et al.*, 2010; Ichimura *et al.*, 2000; Kabeya *et al.*, 2004; Sou *et al.*, 2006), whilst the Hedgehog (Hh) signalling protein family is known to be covalently modified with cholesterol (Porter *et al.*, 1996). During Hh maturation, the protein goes through autocleavage in the ER and the N-terminal fragment is modified by covalent attachment of both cholesterol and palmitate moieties. Cholesterol is involved in the autocleavage, acting as a nucleophile to break apart a thioester intermediate, becoming attached to the C-terminal glycine of the N-terminal fragment (Ciulla *et al.*, 2018). Subsequent release of this protein fragment from the cell surface is thought to be controlled by sequestration of the lipid moieties, leading to the formation of extracellular multimers that act in signalling (Resh, 2016). More recently, smoothened (SMO, another protein in the Hh signalling pathway) has also been shown to be cholesterylated, and this modification is in fact regulated by Hh signalling (Xiao *et al.*, 2017). However, outside of these examples, this modification has not been observed.

1.1.4. GPI anchor

In eukaryotes, the discussed lipid modifications are largely thought to be attached to cytosolic proteins, associating them to the cytoplasmic face of membranes (with known exceptions including Hh and another secreted signalling molecule – Wnt; Porter *et al.*, 1996; Takada *et al.*, 2006). In contrast, GPI-anchored proteins (GPIAPs) only associate to the external face of the plasma membrane (and the lumen of organelles of the endocytic and secretory pathways), where they can feature in signalling, adhesion, nutrient uptake and enzymatic reactions (Ahmad *et al.*, 2012; Ferguson *et al.*, 2017; Varma and Hendrickson, 2010). The GPI anchor itself consist of lipids connected to a carbohydrate structure that is attached to the C-termini of proteins via an ethanolamine group (Ferguson *et al.*, 2017). Attachment of the anchor occurs in the lumen of the ER (Figure 1-2). Proteins competent for GPI attachment contain a C-terminal GPI signal sequence (GPISS) which acts as a transmembrane domain, associating proteins to the luminal side of the ER membrane. Upon recognition by the GPI transamidase complex, this GPISS is cleaved and replaced by a GPI anchor (for more detail on the structure, synthesis and protein attachment of GPI anchors, see sections 1.4.2.1. and 1.4.2.2). GPI-anchoring can lead to lipid raft association and also allows subsequent release of the attached protein in a soluble form upon activation of phospholipase enzymes.

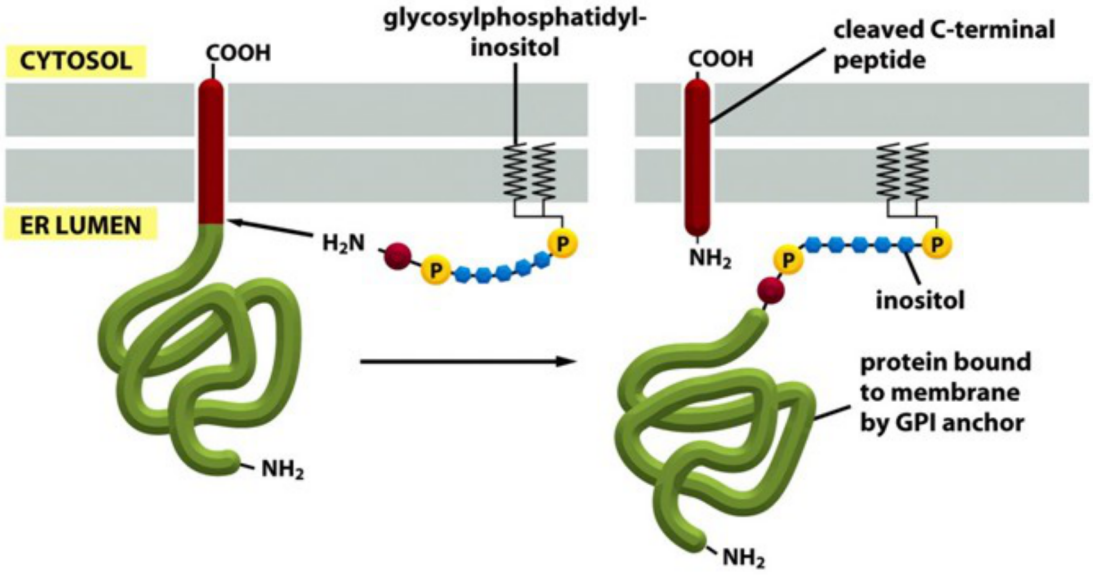


Figure 1-2 GPI anchor attachment in the ER. Proteins competent for GPI anchoring contain a C-terminal GPI signal sequence (GPISS). After translation and translocation into the ER, the GPISS (red) anchors the nascent protein to the luminal face of the ER membrane. Here the GPISS is cleaved and replaced with a GPI anchor consisting of lipids, phosphate groups (P; yellow), mannose residues (blue) and ethanolamine (red). Figure from Alberts *et al.*, 2008b.

1.2. Glycosylation of membrane proteins

Membrane proteins that are exposed to the extracellular space or to the lumen of organelles are often extensively glycosylated. The association of sugar moieties can modulate a protein's biochemical properties, influencing folding, solubility and function. These glycans can be attached to asparagine residues (N-glycosylation, for more detail see section 1.4.2.3) or to serine or threonine residues (O-glycosylation, more detail in section 1.4.4.3) (Varki, 2017). In rarer cases, a single mannose residue can be attached to the indole C2 carbon atom of tryptophan (C-mannosylation; Shcherbakova *et al.*, 2019). This is an unusual modification in that the sugar is attached via a carbon-carbon bond, rather than via a more reactive atom such as nitrogen or oxygen as in N- and O-glycosylation respectively. Glycans play a part in many different roles; but broadly, protein glycosylation functions to modulate structure (both of the protein itself and in the formation of physical barriers) and/or modulate molecular interactions, with some overlap between these two categories.

1.2.1. Protein glycosylation influences protein structure and function

Folding can be greatly enhanced by glycosylation. This is in part due to the association of ER folding chaperones that recognise certain types of glycosylation (specifically N-glycosylation), such as calnexin and calreticulin (Stanley *et al.*, 2017). C-mannosylation has also been implicated in protein folding through its influence on correct disulphide bond formation and protein stability (Shcherbakova *et al.*, 2019). Additionally, glycan structures convey enhanced solubility to proteins. Glycosylation is essential to the solubility and stability of many therapeutic molecules, such as erythropoietin (EPO), interferon-β (IFN-β) and various antibodies (Zhou and Qiu, 2019). Glycosylation has also been suggested to facilitate the high concentration of proteins in blood plasma (~50–70 mg/mL in humans; Varki, 2017).

Glycans have relatively rigid structures and can therefore impede other macromolecules from interacting with proteins through steric hindrance or negative charge (Varki, 2017). These properties allow glycans to protect proteins from proteases, form diffusion barriers, and protect cells from invasion by other microorganisms. For example, mucins (heavily O-glycosylated proteins commonly found secreted from or bound to the surfaces of epithelial cells) protect mucosal epithelium of the stomach and intestines from proteolytic and acidic damage (Loomes *et al.*, 1999). Mucins also act to inhibit cellular invasion by pathogens through shielding other surface receptors, functioning as releasable decoys that can self-cleave upon pathogen binding, and modulation of the inflammatory response (Dhar and McAuley, 2019). As for diffusion barriers, sialylation of nephrin and podocalyxin produced in podocytes appears to be vital for the integrity of blood filtration barriers in the glomerulus (Weinhold *et al.*, 2012). Other structural roles of glycans include modulation of tissue lubrication and elasticity, spatial organisation and sorting of proteins within and between membranes, and extracellular matrix organisation (Varki, 2017).

Glycosylation can have a large influence on molecular interactions, leading to regulation of signalling pathways, receptor-ligand interactions and immune modulation. For example, signalling by Notch receptors impacts differentiation of haematopoietic stem cells (Luca *et al.*, 2015). O-glycosylation can modulate the signalling of this receptor, both through modulation of Notch processing after

ligand attachment, and through altering the receptor's affinity for different ligands. The glycans have been shown to make specific contacts with Notch-interacting ligands, playing a direct role in molecular recognition (rather than having an allosteric effect). Similarly, deletion of Fut8 (an enzyme which fucosylates N-glycans) reduces binding of various growth factor receptors to their respective ligands (such as TGF-β and EGF receptors; Wang *et al.*, 2006, 2005).

1.3. Membrane proteins of the protozoan parasite *Trypanosoma brucei*

1.3.1. GPI-anchored proteins of bloodstream form *Trypanosoma brucei* and other trypanosomatids

The features of membrane proteins discussed above are common to all eukaryotes and can be successfully studied in many model organisms. For instance, GPIAPs have been extensively studied in Trypanosomatids, a group of parasitic protozoa in which GPIAPs are highly abundant. For example, the Chagas' disease causal agent *Trypanosoma cruzi,* which invades multiple mammalian cell types, utilises GPIAPs to aid evasion of compliment and to facilitate invasion of host cells (de Souza *et al.*, 2010). *T. cruzi* trans-sialidases are GPI-anchored enzymes, which transfer sialidase groups from host cells to parasite cells, conferring resistance to complement. One of the protein groups these sialidase groups are attached to are the GPI-anchored the mucins. These are the *T. cruzi* major surface glycoproteins and are thought to be involved in host cell interactions and invasion. Similarly, the surface of *Leishmania* spp. are rich in the GPIAP major surface protease GP63, which is a zinc metalloprotease shown to play a role in avoidance of compliment-mediated lysis, attachment to macrophage compliment receptors to aid in invasion, and survival within macrophages (Yao, 2010). Homologues of GP63 can also be found in all trypanosomatids studied to date, including monoxenous species that only infect insects, such as *Crithidia* spp. Here they are thought to be involved in nutrient acquisition and/or to mediate adhesion of the parasite to the insect host's gut wall (d'Avila-Levy *et al.*, 2006; D'Avila-Levy *et al.*, 2014; Filosa *et al.*, 2019).

Interestingly, in addition to GPIAPs, the above-mentioned species also contain abundant surface GPI molecules that are not attached to proteins (Schneider *et al.*, 1996; Valente *et al.*, 2019). These include glycoinositolphospholipids (GIPLs; found in *T. cruzi*, *Leishmania* spp. and *Crithidia* spp.), lipophosphoglycans (LPGs; found in *Leishmania* spp.) and lipoarabinogalactans (LAGs; found in *Crithidia* spp.). GIPLs are small GPI molecules involved in macrophage invasion and modulation. LPGs are the major surface glycocongugate of *Leishmania* spp. promastigotes, consisting of a GPI-anchor with a heptasaccharide glycan core joined to a long phosphoglycan polymer. In the insect vector, LPGs protect cells from hydrolytic enzymes and assist in adhesion to the gut epithelium. In the mammalian bloodstream, as with GP63, LPGs protect cells from complement mediated lysis and act in invasion of, and survival within, macrophages. Meanwhile, LAGs are thought to be the LPG equivalents in *Crithidia* spp., albeit with a different structure to their polymers (Alcolea *et al.*, 2014; Schneider *et al.*, 1996).

Another clinically relevant trypanosomatid that relies heavily on GPIAPs throughout its lifecycle is *Trypanosoma brucei*. In the bloodstream form (BSF) of *T. brucei*, a single GPIAP – the variant surface glycoprotein (VSG) – accounts for approximately 10% of this cell's proteome (and about 8% of its protein biosynthesis) (Cross, 1975). The structure of the GPI anchor was, in fact, first elucidated in this organism (Ferguson *et al.*, 1988, 1985).

T. brucei is a parasite of mammals in sub-Saharan Africa, where it is spread by tsetse fly bite, causing significant animal and human disease. Animal African trypanosomiasis (AAT or nagana) is caused by all subspecies of *T. brucei* (including *T. b. brucei*), whilst human African trypanosomiasis (HAT or sleeping sickness) is inflicted by *T. b. rhodesiense* and *T. b. gambiense*. Once inside its mammalian host, *T. brucei* lives solely extracellularly in the blood and, thus, is constantly under the surveillance of the host immune system. VSG (which exists on the surface of the parasite as a homodimer) is pivotal to the survival strategy of this organism, aiding in the evasion of both the innate and the adaptive immune responses. Key to this success is the presence of over 3000 VSG genes and pseudo genes in the *T. brucei* genome (Cross *et al.*, 2014). Despite primary sequence divergence, the different VSGs (~50–60 kDa) are predicted

to have similar tertiary structures consisting of N- and C-terminal domains connected by a flexible linker (Figure 1-3A) (Blum *et al.*, 1993; Carrington and Boothroyd, 1996). The hypervariable N-terminal domain (NTD) forms two long α-helices in a coiled-coil structure orientated perpendicular to the plasma membrane, whilst the GPI-anchored C-terminal domain (CTD) is more conserved and has been grouped into 6 classes based on amino acid sequence features (Berriman *et al.*, 2005; Carrington *et al.*, 1991). VSG is highly immunogenic and so, over time, elicits a strong humoral response from the mammalian host (Dempsey and Mansfield, 1983; Diffley, 1985). To evade this response, African trypanosomes have evolved an antigenic variation mechanism dependent on the monoallelic expression of *VSG* (from one of ~20 sub-telomeric expression sites) coupled with periodic switching of expression to one of the many structurally similar, yet immunologically distinct VSG variants. This prevents elimination of the parasite population, resulting in successive waves of parasitaemia and chronic infection.

Further to antigenic variation, VSG protein structure allows tight packing of VSG dimers on the plasma membrane of BSF *T. brucei*. Thus, the presence of ~$5x10^6$ VSG dimers on the parasite's surface creates a dense monolayer approximately 12–15 nm thick (Jackson *et al.*, 1985; Vickerman, 1969). This is thought to act as a physical barrier that protects the parasite from compliment-mediated lysis and shields invariant surface proteins from immune recognition whilst allowing nutrient uptake (Ferrante and Allison, 1983; Mehlert *et al.*, 2012). VSG flexible linkers likely facilitate this, permitting the formation of two different VSG conformations in response to obstacles in the membrane and changing protein density, whilst continually maintaining a physical barrier (Figure 1-3B) (Bartossek *et al.*, 2017). In addition, N-glycans associated to the NTD may act as space fillers, either between VSG dimers or at the membrane distal tip, to further occlude the approach of macromolecules (Mehlert *et al.*, 2002).

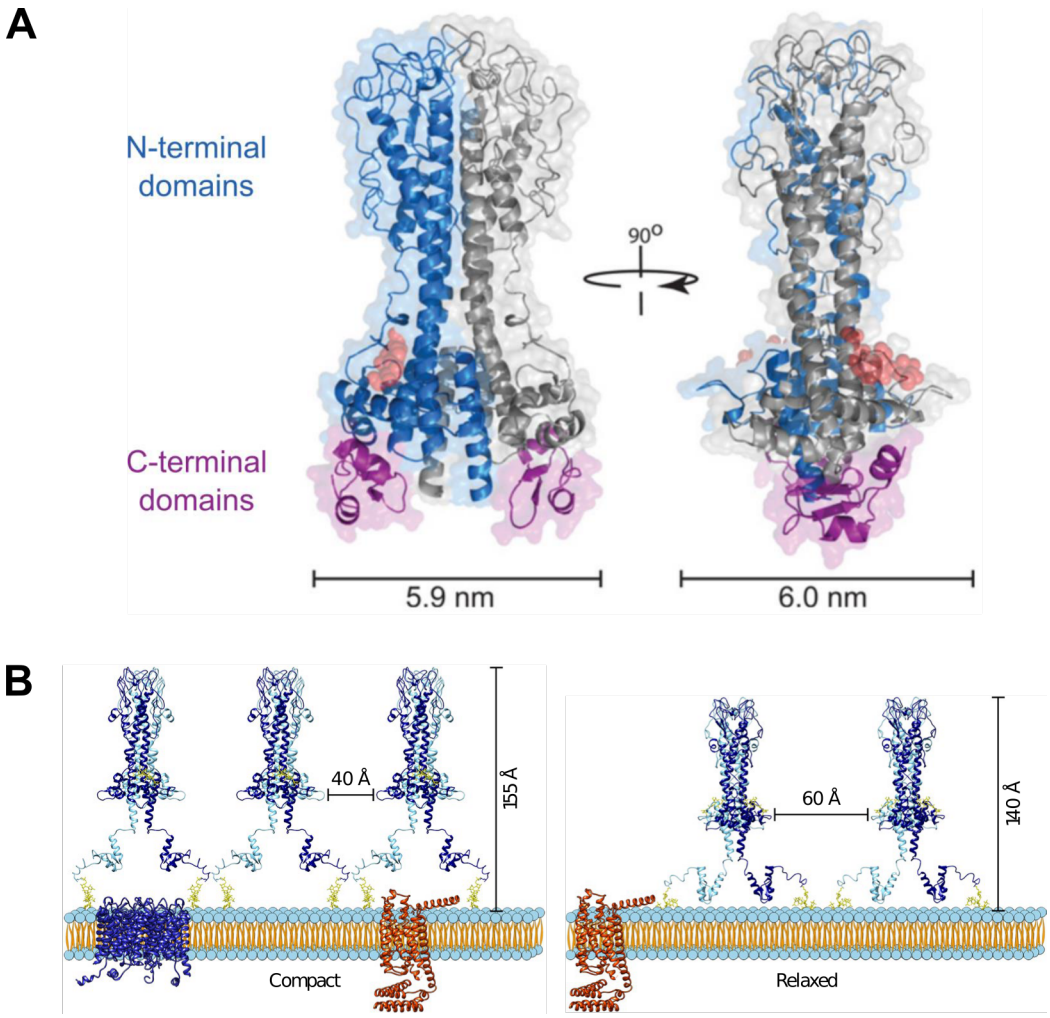


Figure 1-3 Structure and conformations of the VSG homodimer. A Space-filling model of the VSG-2 dimer (herein referred to as VSG221 in accordance with previous nomenclature) viewed from two different angles. Structures of the NTDs of each monomer are shown in blue or grey (Freymann *et al.*, 1990), whilst the two CTDs are shown in purple (Chattopadhyay *et al.*, 2005). The N-linked oligosaccharide in the N-terminal domain is shown in red. The relative positions of the N- and C-terminal domains in this model were estimated as their structures were resolved in separate studies. Modified from Schwede *et al.*, 2015. **B** Model of VSG packing at different densities. Tight packing of VSG is thought to lead to a compact structure (left) which elevates the VSG layer above transmembrane proteins (shown in dark blue and red). Lower densities of VSGs are thought to cause a more relaxed conformation (right), allowing the maintenance of a protective coat despite the greater space around each VSG dimer. Adapted from Bartossek *et al.*, 2017.

In addition to VSG, *T. brucei* is known to express VSG-related GPIAPs essential to parasite survival – e.g. the human serum resistance-associated protein (SRA) and the transferrin receptor (TfR). SRA is found specifically in *T. b. rhodesiense*. It resembles a truncated VSG, having shorter α-helices and lacking surface-exposed loops from the membrane distal-end of VSG (Figure

1-4A and B; Campillo and Carrington, 2003; Zoll *et al.*, 2018). It also contains an additional helix, which prevents dimerization (helix 3 in Figure 1-4A; Zoll *et al.*, 2018). Unlike VSG, SRA largely localises to the endosomal system, where it conveys resistance to the trypanolytic properties of the human apolipoprotein L1 (APOL1). APOL1 is found in the blood associated to haemoglobin and other components, allowing its uptake by the *T. brucei* haptoglobin-haemoglobin receptor (HpHbR, also a GPIAP; Vanhollebeke *et al.*, 2008). Once internalised, APOL1 forms pores in the lysosomal membrane, resulting in an influx of chloride ions, osmotic swelling and cell lysis (Pérez-Morga *et al.*, 2005). SRA can inhibit this pore-forming activity, encountering APOL1 in the endosomes and binding to it through the SRA NTD (Zoll *et al.*, 2018). This prevents cell lysis, allowing *T. b. rhodesiense* to survive and propagate in the human bloodstream. In contrast, *T. b. gambiense* survives APOL1 through reduced uptake by HpHbR (due to a single amino acid substitution and down-regulation of expression; DeJesus *et al.*, 2013; Kieft *et al.*, 2010), and expression of TgsGP, a *T. b. gambiense*-specific protein which causes membrane stiffening (Capewell *et al.*, 2013; Uzureau *et al.*, 2013). Meanwhile, the lack of either of these mechanisms in *T. b. brucei* prevents the subspecies from causing human disease.

TfR is a heterodimer of two expression site-associated genes – ESAG6 and ESAG7 – involved in growth of lab-adapted *T. brucei* in culture (Batram *et al.*, 2014; Tiengwe *et al.*, 2016). Only ESAG6 is GPI-anchored, and both subunits lack the VSG CTD, which predicts the receptor to be recessed into the VSG coat (Figure 1-4C and D; Mehlert *et al.*, 2012). Yet, amino acids 205-215 and 223-238 (which reside in the predicted surface-exposed loops of both ESAG6 and 7) have been shown to form the ligand-binding domain of the receptor (Salmon *et al.*, 1997), with extensive N-glycosylation (3 experimentally validated and 2 putative N-glycosylation sites on ESAG6; 1 experimentally validated and two putative N-glycosylation sites on ESAG7; Figure 1-4D) conferring space for transferrin binding without significantly disrupting the VSG monolayer.

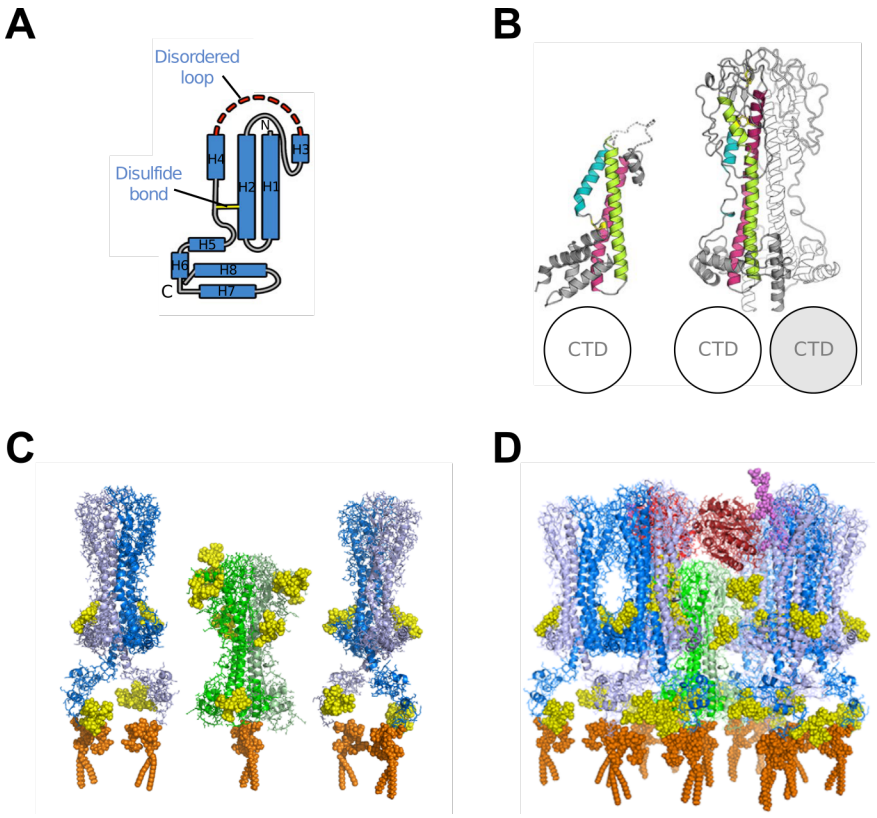


Figure 1-4 Structures of VSG-related GPIAPs SRA (A and B) and TfR (C and D). **A** Schematic model of the SRA NTD. Helices are represented as rectangles. **B** Structural comparison of the SRA NTD with that of VSG221. Both proteins contain a three-helical bundle fold. SRA lacks the VSG membrane distal loop structures but has extra helices. **C** Molecular model of the ESAG6/ESAG7 TfR heterodimer (based on homology between ESAG6/7 and the VSG221 N-terminal domain) flanked by two VSG221 homodimers. TfR peptide chains shown in green; VSG peptide chains in blue; GPI anchor in orange; N-glycans in yellow. **D** Modelling of TfR within the VSG coat bound to a transferrin molecule (red). **A** and **B** adapted from Zoll *et al.*, 2018; **C** and **D** adapted from Mehlert *et al.*, 2012.

1.3.2. Other *Trypanosoma brucei* surface proteins.

The plasma membrane of bloodstream-form *T. brucei* also contains other variant and invariant proteins. Perhaps the most abundant, non-VSG molecules on the surface are members of the invariant surface glycoprotein (ISG) family of type-I transmembrane proteins (Ziegelbauer and Overath, 1992, 1993). ISG65 and ISG75 have been the most extensively studied within this family, yet their functions remain largely unknown (although ISG75 has been linked to the uptake of the trypanocidal drug suramin; Alsford *et al.*, 2012). The functions of other non-GPI *T. brucei* surface proteins are better understood. These include aquaglyceroporins (AQP1–3), pyruvate transporters (TbPT1–5), glucose

transporters (THT1 and 2) and Proteins Associated with Differentiation (PAD1 and TbGPR89).

Aquaglyceroporins are important for the transport of water, small solutes and glycerol across the plasma membrane. Recently AQP2 has been studied with interest as it was shown to be involved in the uptake of trypanocidal drugs melarsoprol and pentamidine, and mutations to the ORF have been implicated in drug resistance (Graf *et al.*, 2013; Munday *et al.*, 2015). Meanwhile, both the pyruvate and glucose transporters are important for life in the mammalian host bloodstream, whereby *T. brucei* relies solely on glycolysis as a source of ATP (Mazet *et al.*, 2013). Thus, the parasite must be able to uptake sufficient glucose, whilst also efficiently exporting pyruvate (the major end-product of glycolysis), as build-up of pyruvate is toxic to cells.

While in the bloodstream, proliferative 'slender' form *T. brucei* has the capacity to differentiate into non-proliferative 'stumpy' forms in a density-dependent manner, avoiding rapid killing of the host and aiding parasite transmission. PAD1 is associated with this and has commonly been used as a surface-membrane marker of 'stumpy' forms (Dean *et al.*, 2009). More recently, the surface protein TbGPR89 has been implicated in quorum sensing of *T. brucei* (Rojas *et al.*, 2019). TbGPR89 is a multipass transmembrane protein related to G-protein-coupled receptors (GPCRs). Overexpression of TbGPR89 in pleomorphic BSF cells (capable of differentiating into stumpy form) led to growth arrest and morphological changes resembling stumpy formation. In contrast, overexpression of TbGPR89 in monomorphic cells (culture-adapted parasites that have lost sensitivity to stumpy-formation signals) led to no such effect, only resulting in a small growth defect. TbGPR89 was shown to transport oligopeptides which, at high concentrations, could trigger stumpy differentiation even at low parasite density (Rojas *et al.*, 2019).

Recently there has been a much greater interest in the trypanosomatid cell surface, through the use of bioinformatics and/or proteomics to identify the surface protein repertoire of African and American trypanosomes, leishmanias and also phytomonas. Jackson *et al* in 2013 focused on phylogenetic comparisons of the predicted surface protein contingents of *T. brucei*, *T. congolense* and *T. vivax,* in order to better understand the evolution of African

trypanosome surface architecture and identify proteins which might be important to each organism's specific biological niche. In contrast, Shimigawa and colleagues focussed on isolating surface proteins of both BSF and PCF (procyclic form; resides in the tsetse fly midgut) *T. brucei* through chemical labelling of the respective cell surfaces with biotin, prior to affinity purification with streptavidin and mass spectrometry (Shimogawa *et al.*, 2015). As expected, BSF and PCF present large differences in surface protein composition, likely adapting the parasite for survival in each lifecycle stage.

Meanwhile, Gadelha and colleagues took a different biochemical approach (Gadelha *et al.*, 2015). Focussing exclusively on BSF, they covalently labelled the *T. brucei* cell surface with activated fluorescein, deliberately avoiding a biotin-avidin system because trypanosomes have been reported to biotinylate their own endogenous proteins (a confounding factor in avidin purifications; Vigueira and Paul, 2011). Fluorescein-labelled cells were lysed, and fluoresceinated surface proteins captured by affinity chromatography. A semi-quantitative, comparative mass spectrometry approach was used to analyse fluorescein-labelled samples versus unlabelled and lysed cell controls, enabling the definition of a high-confidence 'surfeome' containing 175 putative components that were termed ESPs (for 'enriched in surfeome protein'). Validation by cellular localisation (through endogenous-locus tagging of 25 ESP and 12 ESAG ORFs with sfGFP) confirmed a ~80% positive hit rate. Together, these studies have greatly expanded our knowledge about the proteins residing at the host-parasite interface of *T. brucei*, providing new avenues for basic and therapeutic research.

1.4. The birth, life and death of surface membrane proteins

1.4.1. Signal peptide and translocation

All proteins on the surface membrane of BSF *T. brucei* start their lives as an mRNA that encodes an N-terminal signal peptide (SP) sequence. Upon translation, the SP allows a protein to take the first step required to become surface localised – entry into the secretory pathway (Figure 1-5). This occurs by translocation into the ER, followed by cleavage of the SP (although

transmembrane domains can also act as non-cleavable signal sequences). The SP peptide averages in length of ~22 residues and normally consists of three regions – the positively charged n-region, the hydrophobic h-region, and the polar c-region (von Heijne, 1985).

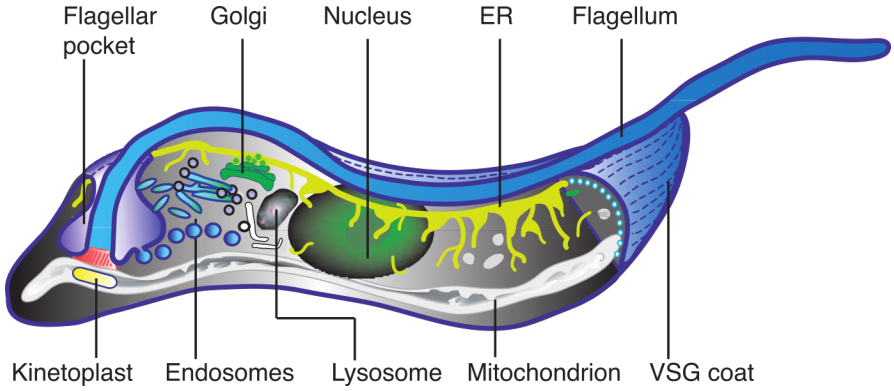


Figure 1-5 Secretory pathway and endocytic organelles of *T. brucei* BSF. To become surface localised, proteins must be translocated into the ER, for subsequent folding and modification, before being transported to the Golgi for further processing. From there proteins are transported to the flagellar pocket (FP; the sole site of endo and exocytosis in the cell). Once endocytosed, surface proteins can be recycled back to the FP or sent to the lysosome for destruction. The kinetoplast is the mitochondrial DNA. Adapted from Overath and Engstler, 2004.

Generally, translocation can occur either co- or post-translationally (summarised in Figure 1-6), depending on the organism and the protein involved. In mammals, the co-translational pathway is predominantly used. Only very small proteins (<100 amino acids), too small the be recognised for co-translational translocation, are translocated post-translationally (Johnson *et al.*, 2012; Lakkaraju *et al.*, 2012; Shao and Hegde, 2011). In yeast, the pathway chosen appears to be regulated by the hydrophobicity of the SP, with greater hydrophobicity directing proteins to the co-translational pathway (Davis *et al.*, 1996). A similar mechanism may be involved in trypanosomes, though it appears to be less strict (Goldshmidt *et al.*, 2008). In *T. brucei* PCFs, it was shown that the post-translational pathway is required for GPIAPs and that these proteins have SP sequences with lower hydrophobicity relative to non-GPIAPs. In contrast, the co-translational pathway was reported to be required for multipass transmembrane proteins (Lustig *et al.*, 2007); other proteins appear to be able to utilise either pathway. However, SP hydrophobicity may not be the

main controlling factor. SPs from different organisms (humans, *Saccharomyces cerevisiae*, *Escherichia coli* and *T. brucei*) have been shown to contain organism-specific sets of conserved-motifs in their h-regions (Duffy *et al*., 2010). Additionally, in trypanosomes, SPs that were mutated to be highly hydrophobic but not contain any of the trypanosome h-motifs, could not be translocated. These results could explain why some SPs are not interchangeable between different organisms (Al-Qahtani *et al.*, 1998; Duffy *et al.*, 2010).

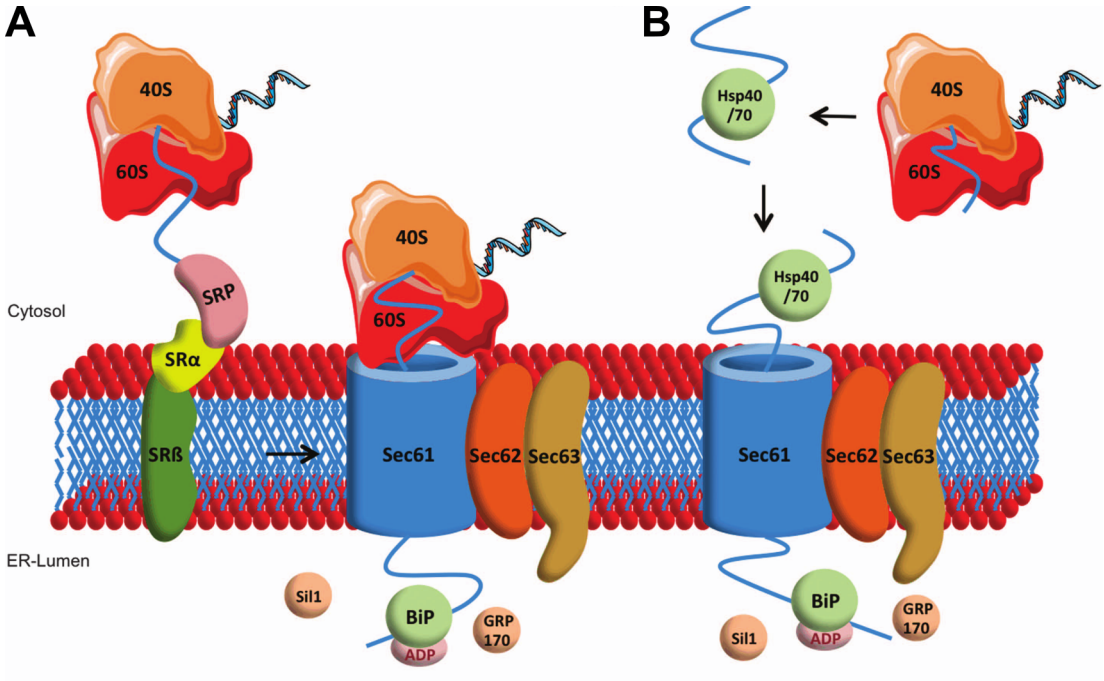


Figure 1-6 Protein translocation across the ER membrane. A Co-translational translocation. The signal recognition particle (SRP) recognises the signal peptide (SP) of the nascent protein as it is translated. Binding of SRP to its receptor (SRα/β) couples the translating complex to the Sec61 translocon for translocation across the ER membrane. Sec63 recruits BiP to the membrane where it opens the Sec61 channel and acts as a molecular ratchet, preventing reverse translocation of the nascent protein. Sil1 and GRP170 function as nucleotide exchange factors for BiP. **B** Post-translational translocation. Fully translated proteins are prevented from misfolding and aggregating in the cytosol by chaperones of the heat shock 40 and 70 protein families (HSP40/70). Sec62/63 (and Sec72/73 in yeast) are thought to be important for targeting these translated proteins to the translocon in the absence of SRP. Adapted from Linxweiler *et al.*, 2017.

The co-translational translocation pathway requires the signal recognition particle (SRP; Figure 1-7). SRP binds to the SP h-region of nascent proteins emerging from the ribosome, connecting them to the translocon in the ER membrane via interaction with the heterodimeric SRP receptor (Siegel and

Walter, 1988). The bacterial SRP (used to translocate proteins across the plasma membrane; Figure 1-7A) only contains one protein and a molecule of 4.5S RNA (Phillips and Silhavy, 1992). Eukaryotic SRPs are more complex: in mammals it consists of six different components and a molecule of 7SL RNA (Figure 1-7; Akopian *et al.*, 2013). SRP9 and SRP14 bind to the Alu domain of the 7SL RNA and act to stop translation of the nascent protein until the complex is assembled at the translocon. SRP54 is found in all SRPs and is the component which binds to SPs. SRP54, along with SRP19, SRP68 and SRP72 all bind to the S domain (signal recognition domain) of the RNA component. SRP19 is needed to enhance correct binding of SRP54. The functions SRP68 and SRP72 are still unclear, but they may be important in organising the 7SL RNA to allow SRP54 binding.

The *T. brucei* SRP is unusual in that it contains only four proteins and two molecules of RNA – the 7SL RNA and a tRNA-like molecule (Lustig *et al.*, 2005). It also lacks Alu-binding proteins, and it has been hypothesised that the tRNA-like molecule fulfils this role.

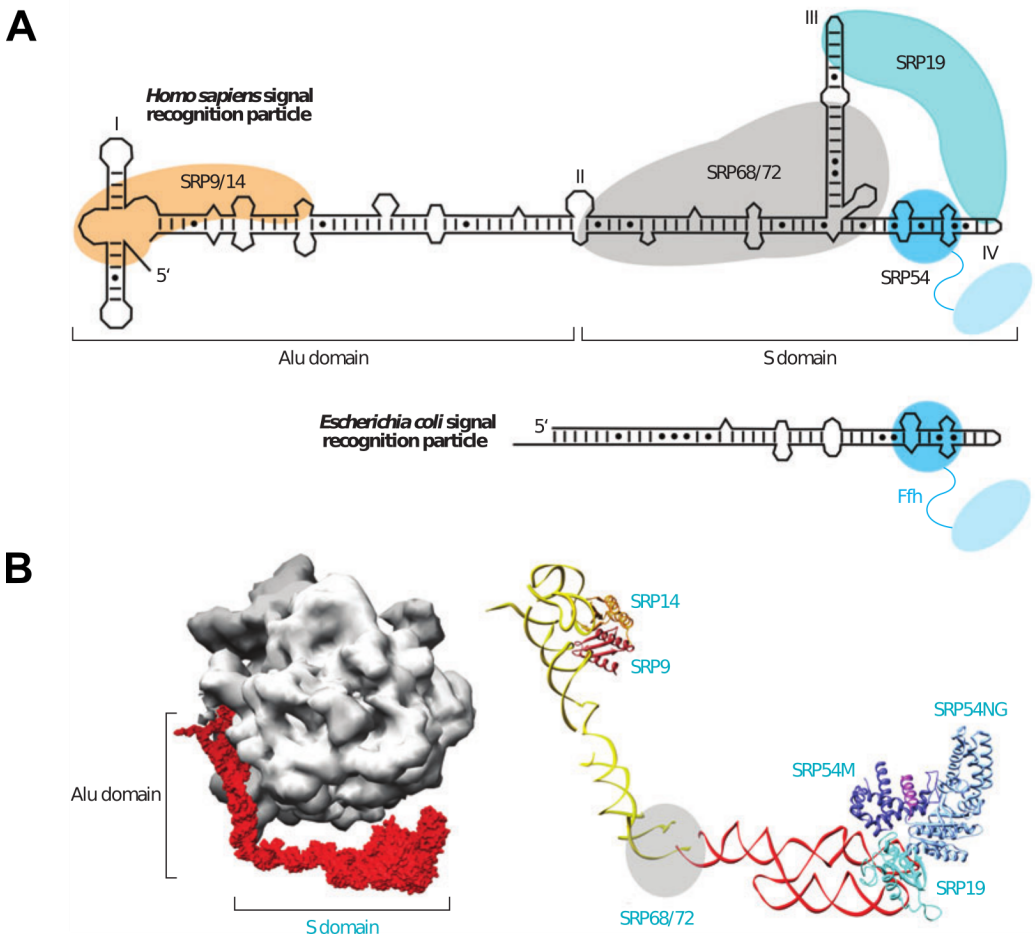


Figure 1-7 Composition of the signal recognition particle (SRP). A Comparison between the composition of the mammalian and bacterial SRPs. The signal-recognition (S) domain recognises N-terminal signal peptides. The Alu domain stalls ribosomal translation of nascent proteins until the SRP-ribosome complex has assembled at the translocon. **B** Structure of the mammalian SRP bound to the ribosome-nascent polypeptide chain complex, resolved by cryoelectron microscopy (cryo-EM) (left; Protein Data Bank in Europe EMD-1063), and molecular model of the mammalian SRP based on cryoEM and crystal structures of the individual proteins (right; Protein Data Bank 1RY1). Grey circle represents the yet unresolved structure of the SRP68/72 complex. The S and Alu domains of the SRP RNA are in red and yellow, respectively. Adapted from Akopian *et al.*, 2013.

The post-translational translocation pathway requires multiple chaperones to prevent misfolding and aggregation of precursor proteins in the cytosol (Dudek *et al.*, 2015). The chaperones include members of the heat shock protein (HSP) families (Ssa1p and farnesylated Ydj1P in yeast, Hsc70 and DNAJB12 in humans; A J Caplan *et al.*, 1992; Avrom J Caplan *et al.*, 1992; Deshaies *et al.*, 1988; Grove *et al.*, 2011). Calmodulin has also been shown to be used in humans (Shao and Hegde, 2011). The chaperones involved in *T. brucei* have not been characterised. However, trypanosomes do have an

extensive repertoire of HSPs (Bentley *et al.*, 2019), one of which has been shown to functionally complement Ydj1P in yeast (Edkins *et al.*, 2004). In addition to folding chaperones, other components are required to direct these fully translated proteins to the ER translocon (taking the place of the SRP receptor from the co-pathway). In yeast, the heterotrimeric Sec63/71/72 complex takes on this role (Dudek *et al.*, 2015). Humans do not have Sec71/72 homologues, but both Sec62 and Sec63 have been shown to be required for post-translational translocation (Haßdenteufel *et al.*, 2018). Meanwhile, the *T. brucei* genome encodes homologues of both Sec63 and 71, but not Sec62 or 72 (Goldshmidt *et al.*, 2008; Lustig *et al.*, 2007).

For both pathways in all organisms studied, the core of the translocon is made up of the heterotrimeric Sec61 complex (Sec61α, β and γ in mammals and trypanosomes; Sec61p, Sbh1p and Sss1p in yeast; Dudek *et al.*, 2015). The complex forms a gated, aqueous channel in the ER membrane, through which proteins are translocated. Also required are BiP, Sec63 (which has been shown to recruit BiP to the ER membrane in yeast; Lyman *et al.*, 1992) and nucleotide exchange factors Sil1 andGrp170 (required for the activity of BiP; Chung *et al.*, 2002; Weitzmann *et al.*, 2006). BiP acts both to open and close the Sec61 complex (Lyman and Schekman, 1997), and as a molecular ratchet, preventing translocating proteins from going backwards through the complex (Lyman *et al.*, 1995; Matlack *et al.*, 1999). In yeast, there is an alternative Sec61 complex which also acts in both translocation pathways (Ssh1p, Sbh2p and Sss1p) (Finke *et al.*, 1996; Zimmermann *et al.*, 2011), and there may also be one in mammals (based on the presence of *SEC61A2* gene; Dudek *et al.*, 2015). However, in trypanosomes, only one Sec61 complex is known. The roles played by these alternative complexes in yeast and mammals have not been elucidated, but they may act to process specific types of cargos with different SPs.

1.4.2. Protein maturation in the ER

The ER is a vital step in the life of a membrane protein. Major processing steps such as GPI anchoring and attachment of N-glycans occur here. In

addition, there are many chaperones to aid in protein folding, and molecular decisions are made as to whether to continue processing or to destroy terminally misfolded proteins. This section will cover the major processing steps that occur in the ER before transportation to the Golgi.

1.4.2.1. GPI-biosynthesis

As discussed, *T. brucei* has been extensively studied as a model for GPIAPs. As such, much is known about how the parasite synthesises the anchor and its highly conserved core structure (Ferguson *et al.*, 2017). In addition, the subtle differences in *T. brucei* and mammalian GPI synthesis have identified the biosynthetic pathway as a potential drug target.

The majority of GPI anchors have a highly conserved core structure: ethanolamine-phosphate-6-D-mannoseα(1-2)-D-mannoseα(1-6)-D-mannoseα(1-4)-D-glucoseamineα(1-6)-phosphatidylinositol (Hong and Kinoshita, 2009) (Figure 1-8). The ethanolamine residue is attached to the C terminus of the proteins via an amide bond. Variations on top of the core structure include different numbers of mannose residues, more ethanolamine residues attached to the other mannose residues, and variations in the phosphatidylinositol (PI) group. However, in *T. brucei* BSF only the core structure is found. In the PCF the core differs slightly in that the inositol of procyclins is acylated (Field *et al.*, 1991). The full GPI biosynthetic pathway for *T. brucei* BSF is shown in Figure 1-9. The core structure of GPIs can be modified further after protein attachment, by addition of sugar side chains to the mannose residues, but this occurs in the Golgi (Ferguson *et al.*, 1988; A Mehlert *et al.*, 1998; Mehlert and Ferguson, 2007; Zitzmann *et al.*, 2000).

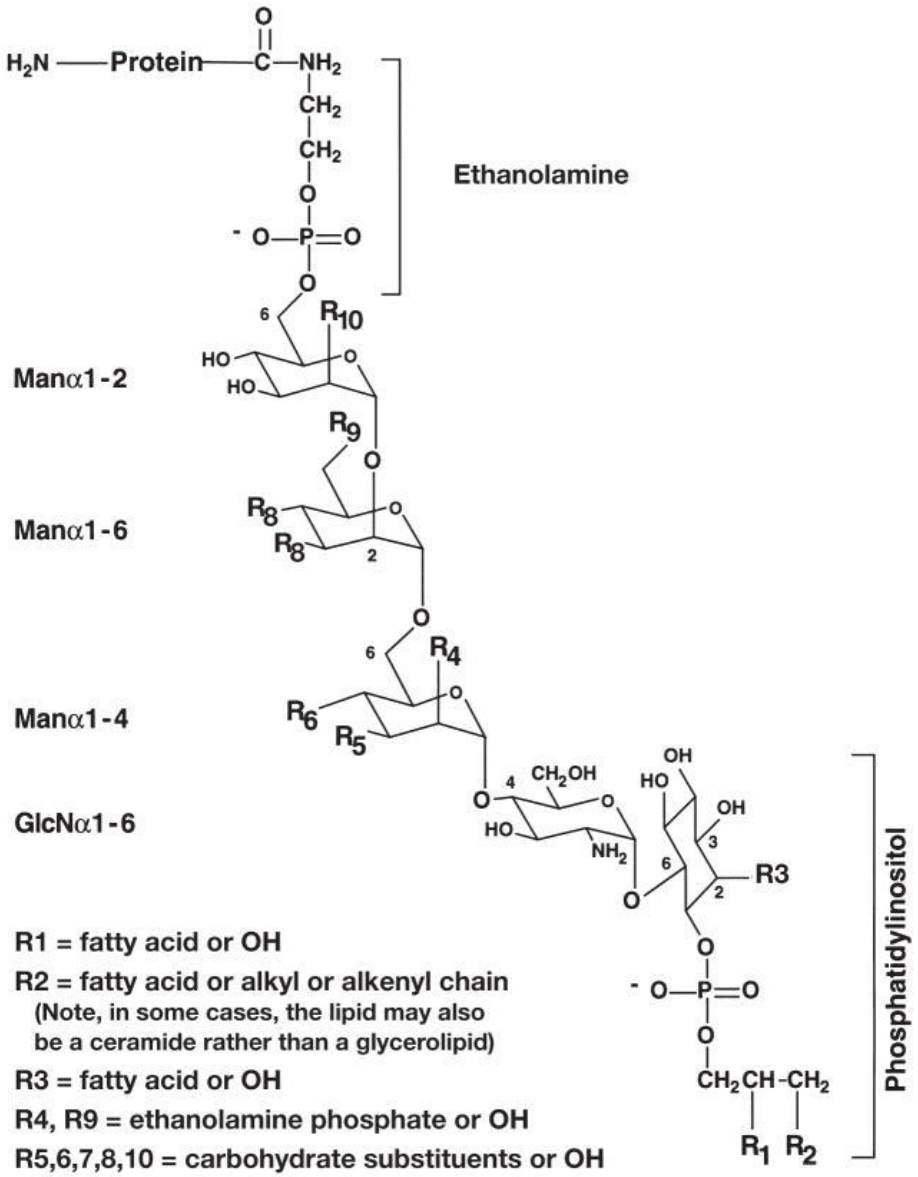


Figure 1-8 Basic core structure of the GPI anchor with possible modifications indicated. In *T. brucei* R3 is OH in BSF but is a fatty acid in PCF; and R4 and R9 are always OH. Figure taken from Ferguson *et al.*, 2009.

GPI biosynthesis in both mammals and *T. brucei* begins on the cytoplasmic side of the ER membrane, with attachment of N-acetylglucosamine (GlcNAc) to PI (Figure 1-9 reaction 1; Hong and Kinoshita, 2009). This occurs by transfer of GlcNAc from uridine diphosphate N-acetylglucosamine. In mammals this is carried out by a multi enzyme complex. In *T. brucei* the enzymes involved are unknown; however, sulphydryl alkylating agents have been shown to inhibit this step in a trypanosome-derived cell-free system (Milne *et al.*, 1992).

The next step involves the deacetylation of the glucosamine residue, for which the zinc metalloenzyme TbGPI12 is responsible (Figure 1-9 reaction 2; Chang *et al.*, 2002; Urbaniak *et al.*, 2005). Its substrate specificity has been shown to differ to the human homologue PIG-L (Sharma *et al.*, 1999). As a result, much work has been done towards developing TbGPI12 inhibitors (Abdelwahab *et al.*, 2012; Capes *et al.*, 2014; Smith *et al.*, 1999, 2001, 2004; Urbaniak *et al.*, 2014). In one study, cell-permeable GPI intermediate analogues were synthesised, which killed trypanosomes within hours (Smith *et al.*, 2004). These analogues were not toxic to HeLa cells even at high concentrations, but were synthetically complex to make and would likely be susceptible to enzymes present in human serum. However, they did highlight the potential of this step as a trypanosome-specific drug target.

For the third step, the GPI precursor is thought to be flipped to the luminal side of the ER membrane (Hong and Kinoshita, 2009). This is the point at which the mammalian and *T. brucei* pathways deviate from each other slightly. At this point in mammals, the inositol group is acylated and the acyl group remains on the anchor throughout the rest of the pathway until it is attached to a protein. Whereas in *T. brucei,* prior to acylation, the first mannose group is attached. After this step, the GPI intermediates can all be acylated and deacylated. The acyl group (predominantly palmitate; Mayor *et al.*, 1990) only seems to be necessary for the ethanolamine-phosphate attachment step (Güther and Ferguson, 1995; Smith *et al.*, 1999). The necessity of the mannose group for acylation, and subsequent removal in other steps, again points towards variations in enzyme substrate specificities between *T. brucei* and mammals. The acylase in *T. brucei* is unknown but can be inhibited by phenylmethysulfonyl fluoride (PMSF), unlike in HeLa cells (Guther *et al.*, 1994).

In humans, the GPI-α 1-4-mannosyltransferase I (GPI-MT-I) has been shown to be PIG-M and requires PIG-X for stability (Ashida *et al.*, 2005; Maeda *et al.*, 2001). It transfers mannose from dolichol-phosphate-mannose (Dol-P-Man) to the GPI intermediate via a 1-4 glycosidic bond with the glucoseamine residue. A PIG-M homologue has been identified in *T. brucei* as a putative GPI-MT-I (TbGPI14; acts in Figure 1-9 reaction 3; Maeda *et al.*, 2001). It has 30% amino acid identity and contains a highly conserved DXD motif, which is functionally important in PIG-M and other glycosyltransferases, where it is

thought to bind a manganese ion which, in turn, helps bind the sugar substrate (Maeda *et al.*, 2001; Wiggins and Munro, 1998). Again, synthetic compounds can inhibit this step in a *T. brucei* cell-free system but not in a HeLa-derived one (Smith *et al.*, 1999, 2000). GPI-α 1-6-mannosyltransferase I (GPI-MT-II) attaches the second mannose via a 1-6 glycosidic bond to the first mannose. In humans the enzyme is PIG-V (Kang *et al.*, 2005). In *T. brucei* a putative GPI-MT-II has been identified (TbGPI18) via its homology with PIG-V (acts in Figure 1-6 reaction 4; Smith and Bütikofer, 2010). GPI-α 1-2-mannosyltransferase I (GPI-MT-III) attaches the third mannose via a 1-2 glycosidic bond to the second mannose. The *T. brucei* GPI-MT-III (TbGPI10) was identified through homology with the mammalian PIG-B, and it was shown to be able to restore biosynthesis in a PIG-B-deficient mouse (acts in Figure 1-9 reaction 5; Nagamune *et al.*, 2000). TbGPI10 could not be knocked out in *T. brucei* without a supplied episomal source, indicating its essentiality.

As mentioned, the transfer of ethanolamine-phosphate to the GPI intermediate Man_3GlcN-IP is the only step in *T. brucei* which requires the acylation of the inositol group (Güther and Ferguson, 1995; Smith *et al.*, 1999). The ethanolamine-phosphate is transferred to the second position on the third mannose using phosphatidylethanolamine as the substrate (Figure 1-9 reaction 6). In mammals, ethanolamine-phosphate can also be added to the other two mannose residues (Hong *et al.*, 2000, 1999; Shishioh *et al.*, 2005; Stokes *et al.*, 2014). To do this, four enzymes work in different combinations in ethanolamine-phosphate transfer: PIG-N adds to the first mannose (Hong *et al.*, 1999), a PIG-F and PIG-G heterodimer adds to the second (Shishioh *et al.*, 2005), and a PIG-F and PIG-O heterodimer adds to the third (Hong *et al.*, 2000). PIG-O and PIG-G are believed to be the enzymes within their respective dimers that transfer the ethanolamine-phosphate. The function of PIG-F is unknown, though it appears to be important in the stable expression of PIG-O and PIG-G (Hong *et al.*, 2000; Shishioh *et al.*, 2005). Additionally, one study found a highly conserved motif not involved in dimer binding that is vital for function (Stokes *et al.*, 2014). In the same study, a *T. brucei* PIG-F homologue was found to selectively bind to a PIG-O homologue but not to the human form. In contrast, the human PIG-F could bind to TbPIG-O. Expression of TbPIG-O and/or TbPIG-F in mammalian PIG-F/PIG-O deficient cells did not appear to restore function.

This could be due to different substrate specificities, pointing towards this step as another potential drug target.

The final step in *T. brucei* GPI biosynthesis is lipid remodelling. Before this can happen, the anchor's inositol must be deacylated (Hong and Kinoshita, 2009). As mentioned, this can occur to any of the acylated intermediates. Two deacylases have been identified in *T. brucei* – TbGPIdeAc (Güther *et al.*, 2001) and TbGPIdeAc2 (Hong *et al.*, 2006). Inhibition of the deacylase reaction in a cell-free system using diisopropylfluorophosphate (DFP) lead to the accumulation of acylated non-lipid modified GPIs (glycolidpid C'-type GPIs, see Figure 1-9; Güther and Ferguson, 1995). These were transferred to VSGs but with lower efficiency of attachment than with the usual lipid-modified anchor. In another study, knockdown of TbGPIdeAc2 found a similar result, in that there was an accumulation of glycolipid C'-type glycolipids and that the expression of VSG on the parasite surface decreased (Hong *et al.*, 2006). Cell growth was also substantially decreased.

Once deacylation of glycolipid C' has occurred to form glycolipid A', lipid remodelling can commence (Hong and Kinoshita, 2009). Initially the lipid in position sn-2 on the anchor (a mix of unsaturated lipids with 18–22 carbons, R1 in Figure 1-8) is removed to create glycolipid θ (Figure 1-9 reaction 7.1). At this point, the GPI is thought to be flipped to the cytoplasmic side of the ER membrane (Ferguson *et al.*, 2017). Position sn-2 is then myristoylated by transfer of myristate from myristoyl-CoA to form glycolipid A'' (Figure 1-6 reaction 7.2; Hong and Kinoshita, 2009). The lipid in position sn-1 ($C_{18:0}$, R2 in Figure 1-8) is then removed, forming glycolipid θ' (Figure 1-9 reaction 7.3). Finally, the site is myristoylated, forming glycolipid A (Figure 1-9 reaction 7.4). This is then flipped back into the ER lumen where it can be attached to proteins (Figure 1-9 reaction 8; Ferguson *et al.*, 2015). The only enzyme known to date to be involved in the remodelling is TbGup1, which acts in the addition of myristate to the sn-2 position (Jaquenoud *et al.*, 2008). Once back on the lumen side of the ER, glycolipid A can be acylated to form glycolipid C. However, unless inositol deacylation is inhibited, only glycolipid A is attached to VSG (Güther and Ferguson, 1995). The core structure of the ESAG6 GPI anchor is also known to be glycolipid A (Mehlert and Ferguson, 2007). Glycolipid C may be the first substrate for the catabolic route of GPI anchors as the cell produces

them in excess (Güther and Ferguson, 1995). Alternatively, they may be a storage mechanism from which glycolipid A can be quickly produced. The exclusive use of myristate in BSF GPIs appears to be highly controlled, as a second myristoylation step occurs after GPIAPs have left the ER (Buxbaum *et al.*, 1996, 1994). Here, it was shown that on the VSG's GPI, myristate groups are replaced by new myristate groups (Buxbaum *et al.*, 1996). This could be a quality control mechanism to ensure the wrong lipids have not been incorporated. If that is the case, then clearly the use of myristate is highly important if BSF cells go through so much to ensure no other lipids are attached. Myristate is shorter than the lipids found in most GPIs in other organisms. The use of a shorter lipid may increase the mobility of GPIs in the membrane, as there are less hydrophobic contacts with other lipid chains in the bilayer. This would aid in hydrodynamic flow – a process thought to be utilised by *T. brucei* in the clearance of bound antibodies from the surface, avoiding complement-mediated lysis by the host's innate immune system (Engstler *et al.*, 2007). This works because of a specialised invagination of the plasma membrane at the base of the parasite's flagellum called the flagellar pocket (FP; Figure 1-5), which is the sole site of endocytosis and exocytosis in *T. brucei*. In BSF *T. brucei,* the FP is positioned towards the posterior end of the cell, and so forward trypanosome motility causes hydrodynamic drag towards the FP (Engstler *et al.*, 2007). When VSG is capped by an antibody, it is thought that the extra drag caused by the immunoglobulin 'molecular sail', leads to quick movement of the complex to the FP. Here it can be endocytosed, the VSG recycled, and the antibody degraded in lysosomes.

Knockout, knockdown, and inhibition of various steps of GPI biosynthesis have shown it to be essential in *T. brucei* (Chang *et al.*, 2002; Hong *et al.*, 2006; Nagamune *et al.*, 2000; Smith *et al.*, 2004). Subtle differences between the parasite's pathway and that of mammals identify it as a good target for anti-trypanosomal drugs (Abdelwahab *et al.*, 2012; Capes *et al.*, 2014; Guther *et al.*, 1994; Nagamune *et al.*, 2000; Sharma *et al.*, 1999; Smith *et al.*, 2004, 2001, 2000, 1999; Stokes *et al.*, 2014; Urbaniak *et al.*, 2014). Multiple enzymes remain to be identified, including those involved in GlcNAc transfer, inositol acylation and lipid remodelling. The enzyme(s) which flip the GPI intermediates to either side of the ER membrane are unknown. The full function of PIG-F is still to be

proven. Further dissection of the GPI biosynthesis pathway would potentially greatly assist in the development of effective drugs against trypanosomes in host blood.

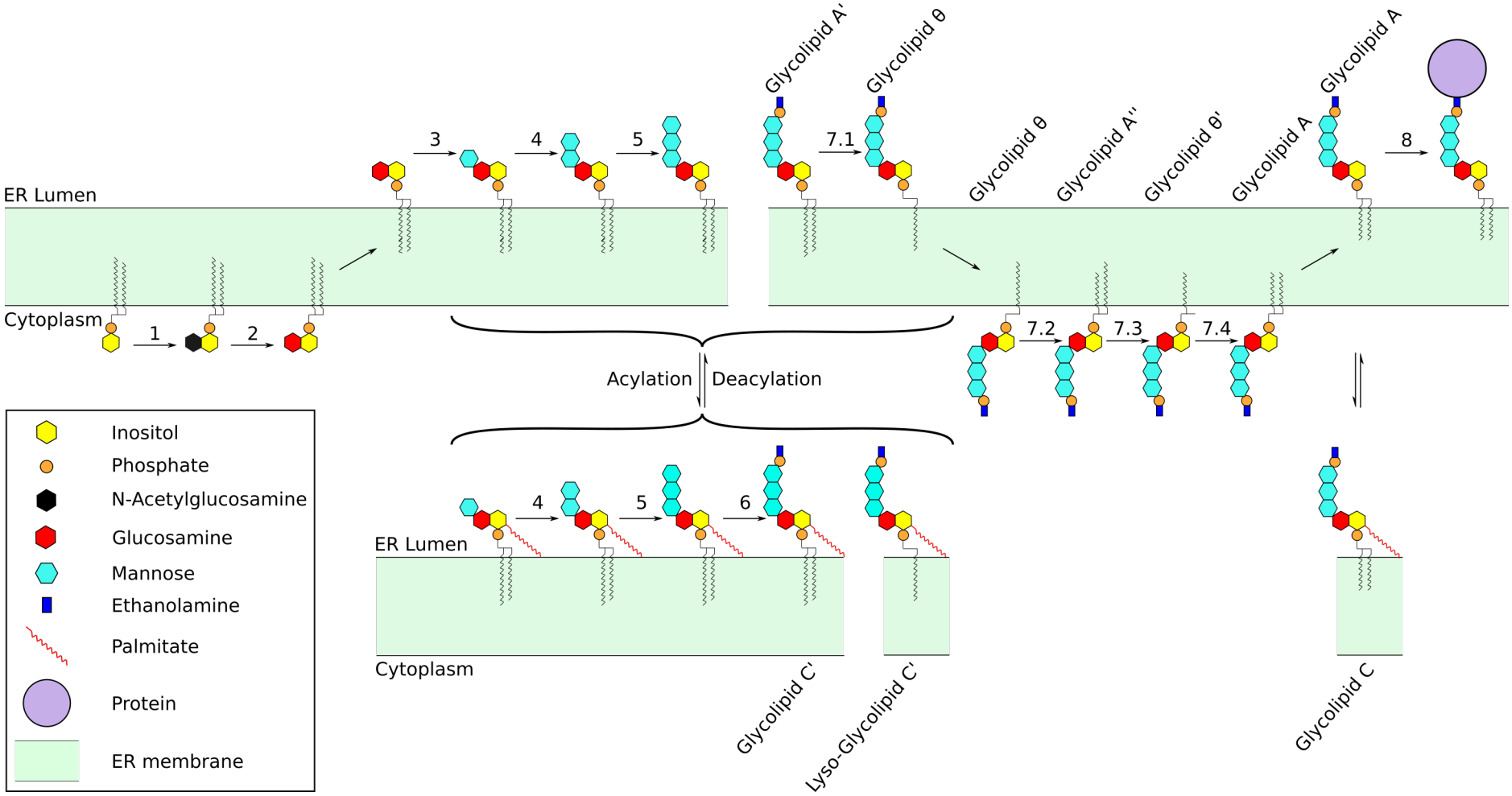


Figure 1-9 The *T. brucei* BSF GPI biosynthetic pathway is a multistep process carried out on the ER membrane. Intermediates that can be acylated and deacylatyed on their inositol groups are indicated. The side of the ER membrane upon which particular steps are thought to occur

are based on enzyme topologies (flipases unknown). Numbers correspond to the following reaction steps. 1. N-Acetylglucosamine addition to phosphatidylinositol (PI). 2. Deacetylation of N-Acetylglucosamine. 3-5. Mannose transfer. 6. Ethanolamine phosphate transfer (requires inositol acylation). Steps 7.1-7.4 are lipid-remodelling steps (require inositol deacylation): 7.1. Lipid removal from PI. 7.2. Myristoylation. 7.3. Lipid removal from PI. 7.4. Myristoylation. 8. Protein to GPI attachment by a transamidation reaction.

1.4.2.2. GPI-anchoring and the ω-site

As covered in section 1.1.4, for a protein to become GPI anchored, its amino acid sequence must contain a GPISS (Ferguson *et al.*, 2017) (Figure 1-10A). This sequence is cleaved off and replaced with the GPI anchor, which is attached directly to the new C-terminus. The GPISS begins with the ω-site. This is the amino acid to which the anchor becomes linked. The ω-site and nearby amino acids (ω -1, ω+1 and ω+2) tend to have small side chains (Ferguson *et al.*, 2009; Pierleoni *et al.*, 2008). Downstream of this site, there is a small hydrophilic linker region of 5-10 amino acids, followed by a C-terminal hydrophobic sequence of about 15-20 amino acids. Prior to cleavage, the hydrophobic region inserts into the ER membrane on the luminal side. Here the GPI transamidation complex (GPI-T) cleaves the signal sequence and replaces it with the GPI anchor (Ferguson *et al.*, 2009; Hong and Kinoshita, 2009; Pierleoni *et al.*, 2008) (Figure 1-10B). GPI-T in all organisms contains five subunits (Nagamune *et al.*, 2003). These are all homologous in humans and yeast. The human (and yeast) subunits are GPI8 (Gpi8p), GAA1 (Gaa1p), PIG-T (Gpi16p), PIG-S (Gpi17p) and PIG-U (Cdc91p) respectively. *T. brucei* have homologues for the first three components (TbGPI8, TbGAA1 and TbGPI16), whereas the other two GPI-T components appear to be trypanosomatid-specific – TTA1 and TTA2. Homologues of TTA1 and TTA2 are present in *Leishmania major* and a TTA2 homologue is found in *T. cruzi.*

GPI attachment by transamidation was first suggested following a cell-free system study showing that a small amount of a known GPIAP precursor could be cleaved but not linked to GPI (S. E. Maxwell *et al.*, 1995), and that nucleophiles such as hydrazine and hydroxylamine could increase this amount (S E Maxwell *et al.*, 1995). This indicates that, during GPI attachment, an activated carbonyl intermediate is formed, which is susceptible to nucleophilic attack. GPI8 is thought to be the component which cleaves the signal sequence and binds the activated intermediate for anchor attachment (Kang *et al.*, 2002; Meyer *et al.*, 2000). It contains conserved cysteine and histidine residues vital to function and believed to act as a catalytic dyad for protein cleavage. The protein would then be bound to the cysteine as the activated intermediate ready

for GPI attachment. GPI8 is linked to GPI16 by a disulphide bond. GPI16 is a structural component that stabilises the complex (Ohishi *et al.*, 2003).

GAA1 is stably associated with GPI8 and GPI16, and may be involved in recruiting the GPI anchor and/or catalysing its use as the nucleophile. A study showed that GAA1 could be co-immunoprecipitated with the GPI anchor through radioactive labelling of GPI using [^{3}H]mannose (Vainauskas and Menon, 2003). The same study also found that mutation of a conserved proline residue knocked out this ability. Since then, bioinformatic analysis points towards GAA1 having similarity with the M28 family of peptidases; and that it may contain a zinc ion that could activate the GPI ethanolamine-phosphate group for attack of the GPI8-protein intermediate (Eisenhaber *et al.*, 2014).

The function of the other two subunits in the complex is unknown. They are vital in GPI attachment. However, in yeast they are not stably associated with the rest of the complex, and so may not be involved in catalysis but rather recruitment (Fraering *et al.*, 2001; Zhu *et al.*, 2005). They may help to recruit proteins or GPIs for the transamidation process. PIG-U has similarity to fatty acid elongase and acyl transferase, and so has been suggested to recognise GPIs through their fatty acid chains (Eisenhaber *et al.*, 2018; Vainauskas and Menon, 2003). As mentioned, TTA1 and TTA2 are not homologous to the other eukaryotic GPI-T components. Unlike PIG-U/Cdc91p and PIG-S/Gpi17p, they are linked to each other by a disulphide bond (Nagamune *et al.*, 2003). They do have similar hydrophobicity profiles to PIG-U/Cdc91p and PIG-S/Gpi17p and may, through convergent evolution, carry out the same function. If they are indeed involved in recruitment, then this might explain the variation seen between mammalian and *T. brucei* specificity for GPI signal sequences. A study showed that *VSG* expressed in mammalian cells was inefficiently anchored (Caras and Moran, 1994). The cells also had problems anchoring a mammalian protein with the *T. brucei* VSG signal sequence attached. Based upon current knowledge of the GPI-T complex, a suggested reaction mechanism for GPIAP transamidation is shown in Figure 1-10C.

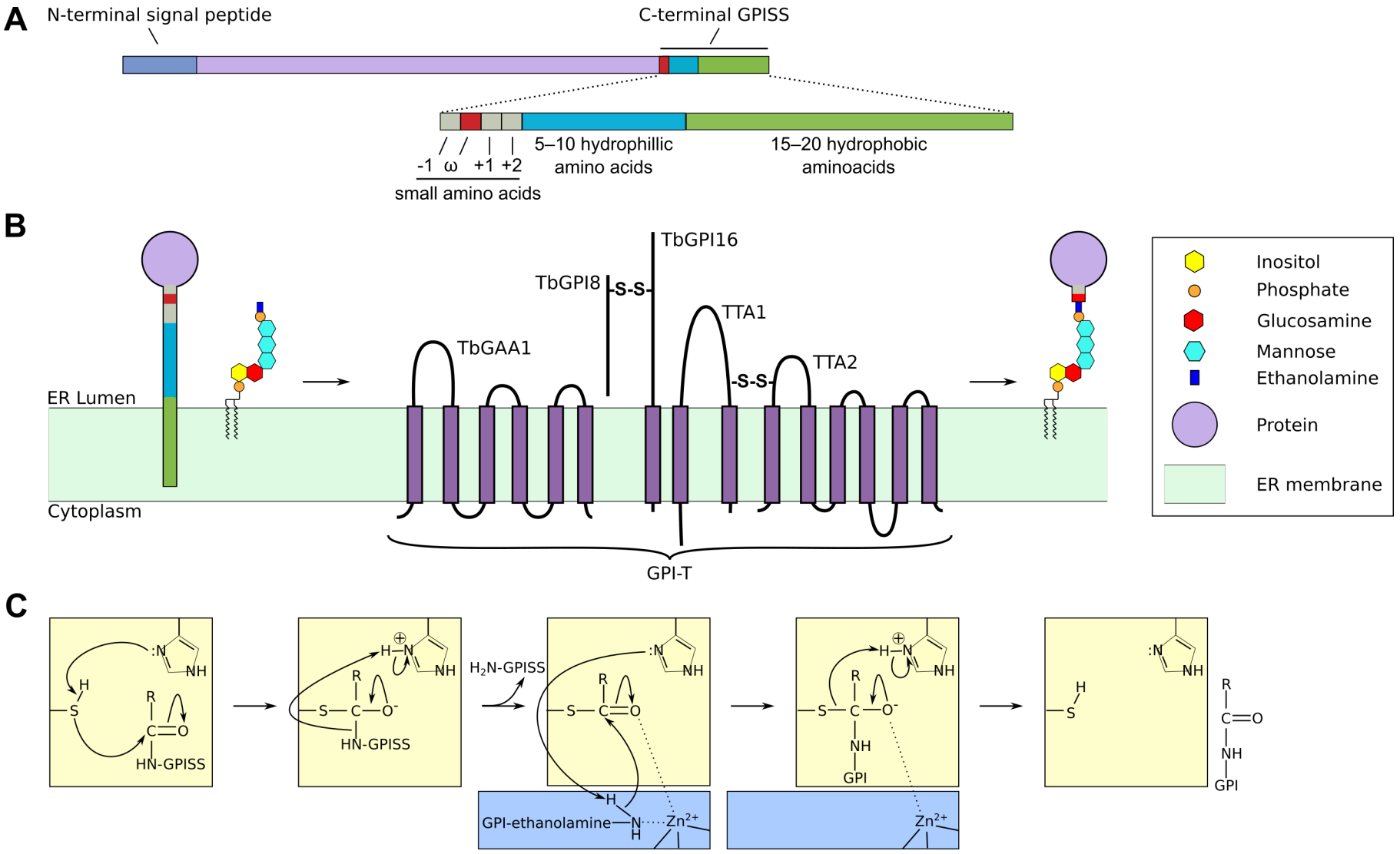

A
N-terminal signal peptide
C-terminal GPISS
-1 ω +1 +2
small amino acids
5–10 hydrophillic
amino acids
15–20 hydrophobic
aminoacids
B
TbGPI16
TbGPI8
-S-S-
TTA1
TbGAA1
TTA2
-S-S-
ER Lumen
Cytoplasm
GPI-T
Inositol
Phosphate
Glucosamine
Mannose
Ethanolamine
Protein
ER membrane
C
HN-GPISS
H2N-GPISS
GPI-ethanolamine
Zn2+
GPI

Figure 1-10 Protein-GPI transamidation. A Schematic of an unprocessed GPI-anchored protein (not to scale). The N-terminal signal peptide directs the nascent protein to the ER where it is cleaved off. The C-terminal signal peptide associates the protein to the ER membrane. The peptide is then cleaved off and a GPI anchor attached to the ω site (red). The C-terminal peptide consists of amino acids with small side chains (grey) around the ω site, a stretch of 5-10 hydrophilic amino acids (blue), and a hydrophobic stretch of 15-20 amino acids (green). The hydrophobic amino acids associate the protein to the luminal side of the ER membrane. **B** GPI transamidation by the *T. brucei* GPI transamidase complex (GPI-T). TbGPI8 (possibly along with TbGAA1) is thought catalyse GPI-transamidation. TbGPI16 stabilises the complex. The functions of TTA1 and TTA2 are unknown but likely involve recruiting the GPI anchor. **C** Proposed reaction mechanism of GPI transamidation based on current knowledge. Yellow box is the TbGPI8 active site. Blue is the TbGAA1 active site. Regarding the protein to be anchored, 'R' represents the amino acids upstream of the amide bond involved in the transamidation reaction. TbGPI8 cleaves the C-terminal signal sequence, forming an activated carbonyl intermediate with the protein substrate. TbGAA1 activates the amine group in the ethanolamine residue of the GPI anchor, allowing it to carry out nucleophilic attack on the activated carbonyl, forming the final GPIAP.

1.4.2.3. N-glycosylation

N-glycans are attached to asparagine residues of proteins via molecular recognition of N-X-S/T sequons (in which X can be any amino acid except proline; Stanley *et al.*, 2017). These glycans vary largely but can generally be defined as one of four different types: oligomannose, paucimannose, complex and hybrid (Figure 1-11). N-glycosylated fungal proteins tend to have highly elaborated oligomannose structures with up to 200 mannose residues (Conde *et al.*, 2004). Mammalian proteins largely have complex-type N-glycans. These can be highly branched with a myriad of different structures and sugar moieties used (Croset *et al.*, 2012; Stanley *et al.*, 2017). In contrast, *T. brucei* proteins have been shown to receive oligomannose, paucimannose and complex-type N-glycan; their complex N-glycans are only biantennary and are not synthesised with fucose or sialic acid residues (Mehlert *et al.*, 1998; Mehlert *et al.*, 2012) (although PCF cells can scavenge sialic acid residues from the host and attach them to parasite surface N-glycans using trans-sialidases; Ammar *et al.*, 2013). Instead, *T. brucei* complex N-glycans consist of chains of poly-N-acetyl lactosamine (pNAL; alternating residues of GlcNAc and galactose) of varying lengths. For example, VSG has been shown to contain 2–8 repeat units of N-acetyl lactosamine (Mehlert *et al.*, 1998). Other *T. brucei* proteins have been shown to have much larger chains of pNAL, reaching >50 units (Atrih *et al.*, 2005), and are predicted to be FP/endosomal-resident based on recognition by the sugar binding protein tomato lectin (Nolan *et al.*, 1999).

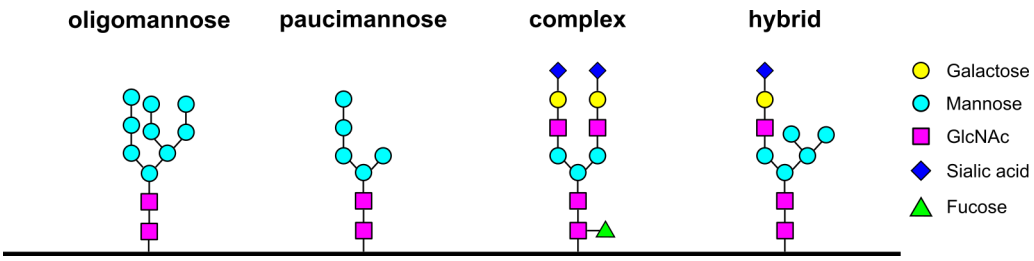


Figure 1-11 N-glycan types attached to proteins in eukaryotes. Yeast and fungi attach oligomannose N-glycans. Mammals typically use complex and hybrid-type N-glycans. *T. brucei* utilises oligomannose, paucimannose and complex-type, although they do not use fucose residues and can only attach sialic acid residues through scavenging from the host.

All types of N-glycans begin by synthesis and attachment to proteins in the ER (summarised in Figure 1-12A; Stanley *et al.*, 2017) prior to further modification and elaboration in the Golgi. This process is conserved between eukaryotes, but trypanosomes differ at particular steps. N-glycan precursors are constructed on dolichol phosphate (Dol-P) lipid carriers. Dolichol is a polyisoprene. In most eukaryotes it consists of 14–19 5-carbon isoprene units. However, trypanosomatids synthesise unusually short dolichol molecules of 11–13 units (Low *et al.*, 1991; Parodit and Quesada-Allue, 1982; Quesada-Allue and Parodi, 1983). Dol-P anchors the N-glycan precursors to the ER membrane, where different ALG proteins (Asn-linked glycosylation) can sequentially transfer residues from nucleotide sugars to the precursors. The first steps of this process occur on the cytoplasmic leaflet of the ER membrane. Here, GlcNAc-1-phosphotransferase (ALG7) transfers GlcNAc-1-P from UDP-GlcNAc to form Dol-P-P-GlcNAc. Subsequently, another GlcNAc residue is transferred (by ALG13/14 from UDP-GlcNAc) followed by five mannose residues (by ALG1, 2, and 11 from GDP-mannose) resulting in $Man_5GlcNAc_2$-P-P-Dol.

At this point, the precursor is flipped to the luminal leaflet of the ER membrane by an unknown mechanism. Flipase activity was once attributed to RFT1 in yeast, as inhibiting its expression lead to glycosylation deficiencies and the build-up of the $Man_5GlcNAc_2$-P-P-Dol precursor (Helenius *et al.*, 2002). Subsequently, however, cells lacking RFT1 were shown to retain flipase activity (Sanyal *et al.*, 2008). In addition, knockout of the *T. brucei* homologue (TbRFT1, reported to functionally complement the yeast knockout) also led to the build-up of $Man_5GlcNAc_2$-P-P-Dol. Yet, glycans attached to proteins were shown to originate from a fully processed precursor (Jelk *et al.*, 2013). RFT1 knockout has also been shown to effect GPI anchor side-chain modifications in *T. brucei* PCF (Gottier *et al.*, 2017). Thus, the protein is now thought to act in some way to promote further processing steps of $Man_5GlcNAc_2$-P-P-Dol and GPI anchors, rather than in flipping or in directly glycosylating these molecules.

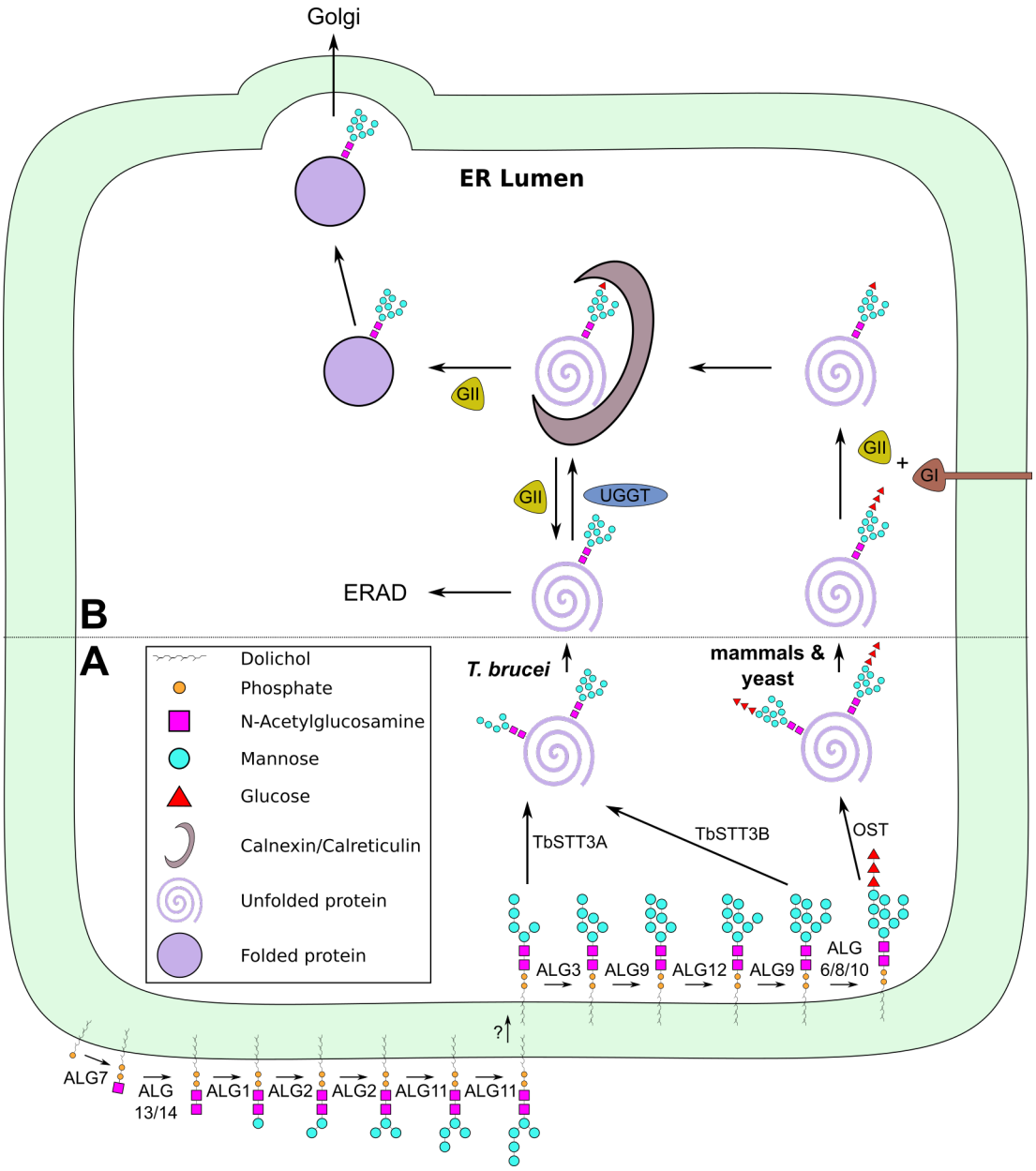


Figure 1-12 Processing and folding of N-glycosylated proteins in the ER. The processes shown in the figure also occur to membrane-associated proteins, but a soluble protein example was used for clarity. **A** Illustration of the biosynthesis of N-glycan precursors and their attachment to nascent proteins. Enzymes used at each step are indicated. Mammalian and yeast transfer $Glc_3Man_9GlcNAc_2$ to proteins using oligosaccharyltransferase (OST) complexes. *T. brucei* lack ALG6, 8 and 10. Thus, they transfer $Man_5GlcNAc_2$ or $Man_9GlcNAc_2$ to nascent proteins using TbSTT3A and TbSTT3B respectively (only known components of the *T. brucei* OST). **B** Illustration of glycoprotein folding in the ER prior to forward trafficking to the Golgi. Mammalian and yeast proteins first have two glucose residues removed by glucosidase I and II (GI and GII), whilst *T. brucei* proteins initially have one glucose attached by UDP-glucose:glycoprotein glucosyltransferase (UGGT). Proteins containing N-glycans with a single glucose residue are recognised by lectin folding chaperones (calnexin or calreticulin in mammals and yeast; calreticulin only in *T. brucei*). Proteins are then deglucosylated by GII. Folded proteins are transported on to the Golgi. Proteins that remain unfolded are either reglucosylated or targeted for ER-associated degradation (ERAD).

Once on the luminal face of the ER membrane, Man5GlcNAc2-P-P-Dol is further modified by various ALG enzymes (see Figure 1-12A for details; Stanley *et al.*, 2017). In mammals and yeast, this results in the generation of $Glc_3Man_9GlcNAc_2$-P-P-Dol, from which the glycan structure is subsequently transferred to proteins. In *T. brucei* the process differs slightly. Firstly, glucose residues are not attached to the N-glycan precursors (Izquierdo *et al.*, 2009a, 2012; Jinnelov *et al.*, 2017). Thus, the end-product of this pathway is instead $Man_9GlcNAc_2$-P-P-Dol. Secondly, *T. brucei* can transfer $Man_5GlcNAc_2$ or $Man_9GlcNAc_2$ to protein glycosylation sites. Transfer of $Man_9GlcNAc_2$ leads to sites with oligomannose-type N-glycosylation; whilst transfer of $Man_5GlcNAc_2$ leads to sites with paucimannose or complex-type N-glycosylation depending upon downstream processing. This difference in N-glycan transfer is likely due to their distinct oligosaccharyltransferase (OST) complexes. OSTs act to transfer N-glycans to proteins, and in mammals and yeast consist of 8 or 9 subunits (Cherepanova *et al.*, 2016). Mammalian OST isoforms have two different catalytic subunits (STT3A and STT3B), but these function to glycosylate proteins either during or after translocation respectively. Indeed, the mammalian OST complex is known to associate to the translocon. In contrast, the *T. brucei* genome lacks recognisable homologues for most of the OST subunits, only encoding three STT3 homologues (TbSTT3A, B and C). Based on immunoprecipitation studies, the BSF OST complex is hetero-oligomeric but only contains TbSTT3A and B (Jinnelov *et al.*, 2017). *TbSTT3C* mRNA is reportedly not detected in BSF or PCF cells (Izquierdo *et al.*, 2009a). TbSTT3A and TbSTT3B have been shown to have different substrate preferences, with TbSTT3A attaching $Man_5GlcNAc_2$ to sequons in acidic amino acid environments, and TbSTT3B attaching $Man_9GlcNAc_2$ to sequons in basic to neutral amino acid environments (Izquierdo *et al.*, 2009a, 2012; Jinnelov *et al.*, 2017). Despite these preferences, upon deletion of *TbALG12* (TbALG12 usually converts $Man_7GlcNAc_2$-P-P-Dol to $Man_8GlcNAc_2$-P-P-Dol), TbSTT3B was shown to attach either $Man_5GlcNAc_2$ or $Man_7GlcNAc_2$ to its preferred glycosylation site in VSG221, albeit with a lower efficiency (Izquierdo *et al.*, 2012). In addition, TbSTT3B is thought to be more promiscuous in its sequon-

site recognition, as knockdown of *TbSTT3A* leads to the appearance of oligomannose-type glycosylation in more acidic sequons (Izquierdo *et al.*, 2009a). Conversely, upon knockdown of *TbSTT3B*, non-acidic sequons were poorly glycosylated. Thus, it was suggested that TbSTT3A acts to glycosylate proteins as they are translocated, whilst TbSTT3B subsequently glycosylates any remaining accessible sites after translocation (and possibly after folding). The specific preference of TbSTT3A for acidic sequons is attributed to the presence of an arginine residue in its predicted active site based on molecular models (as opposed to a histidine in TbSTT3B; Jinnelov *et al.*, 2017).

Once attached to proteins, N-glycans play a large role in glycoprotein folding. In mammals and yeast, this is regulated by the presence, and removal, of glucose residues of the $Glc_3Man_9GlcNAc_2$ structure (Caramelo and Parodi, 2015). Initially two of the glucose residues are removed sequentially by the membrane-bound glucosidase I (GI), followed by the soluble heterodimeric glucosidase II (GII). Subsequently, one of two folding chaperones (the soluble calreticulin or its membrane-associated paralogue calnexin) recognises the $Glc_1Man_9GlcNAc_2$ protein-linked structure. Association with calnexin/calreticulin improves folding efficiency, reduces aggregation and facilitates disulphide bond isomerisation of N-glycosylated proteins. Eventually, the final glucose is removed by GII. If the protein is correctly folded, it is transported to the Golgi. If not, it is recognised by UDP-glucose:glycoprotein glucosyltransferase (UGGT), which functions to reattach the final glucose residue; thus, reassigning the misfolded protein for further modulation by calnexin/calreticulin. This cycle continues until the protein is properly folded or until it is sent for degradation by the ER-associated degradation (ERAD) machinery. As discussed, N-glycans attached to proteins in the ER do not contain glucose residues in *T. brucei*. Instead, N-glycans are glycosylated after protein attachment. Just as in mammals, this glycosylation occurs via a UGGT homologue (Izquierdo *et al.*, 2009c; Jones *et al.*, 2005). Concurrent with the attachment of $Man_{5\text{-}7}GlcNAc_2$ and $Man_9GlcNAc_2$ glycans to proteins, *T. brucei* UGGT has been shown to transfer glucose to each structure (Izquierdo *et al.*, 2009c). Regarding the other proteins involved in this step, the*T. brucei* genome does not encode calnexin but does calreticulin and GII homologues (Jones *et al.*, 2005). Thus, the

glucosylation-deglucosylation folding cycle is also known to occur in this parasite.

1.4.3. Trafficking ER to Golgi

Once proteins destined for the cell surface are processed and folded in the ER, they must carry on their journey via the Golgi. In eukaryotes, ER exit is normally carried out by COPII vesicles, which traffic cargo to the Golgi apparatus (Sevova and Bangs, 2009). The basic steps for COPII vesicle formation are as follows (Figure 1-13 ; as reviewed by Jensen *et al.*, 2011). On the cytoplasmic face of the ER, the transmembrane protein Sec12 stimulates exchange of GDP for GTP on the cytosolic protein Sar1. Sar1-GTP now has an amphipathic α-helix exposed that helps it bind to the ER membrane, initiating membrane deformation. The Sec23/Sec24 heterodimer will then bind to Sar1-GTP via the Sec23 subunit. Sec24 can binds to transmembrane cargo proteins. The Sec13/Sec31 heterodimer now binds to the Sec subunits to form the outer coat of the vesicle and buds off, which occurs at specific locations in the ER known as ER exit sites (ERES). As GPIAPs do not span the whole membrane and are only embedded by their lipid moieties, additional proteins are required to facilitate interaction with the cytosolic COPII components and subsequent sorting into these vesicles. In yeast and mammals, this role is carried out by the p24 family of proteins (Bonnon *et al.*, 2010; Castillon *et al.*, 2011). They form hetero-oligomeric complexes which span the ER membrane, binding to both the Sec24 COPII subunit and to GPIAPs (Bonnon *et al.*, 2010).

It has been shown that export of ER proteins in *T. brucei* also utilises COPII vesicles (Sevova and Bangs, 2009). Expression of a TbSar1 dominant negative mutant (that is stuck in the GDP-bound form) caused lack of transport-dependant processing of reporter proteins, indicating retention within the ER. The study also found that *T. brucei* has two isoforms of both TbSec23 and TbSec24, forming the heterodimers of TbSec23.1/TbSec24.2 and TbSec23.2/TbSec24.1. Knockdown of any of the four components individually affected parasite growth, but only minimally affected the transport of soluble or transmembrane reporter proteins, indicating redundancy for transport of certain

cargo. Ablation of both TbSec23.2/TbSec24.1 subunits, however, affected transport of VSG and a soluble protein fused to a GPI anchor. This specific inhibition of GPIAP transport was suggested to be caused by the differences in the Sec24 isoform sequences in the region that binds to cargo proteins. It would be this region that interacts with p24 proteins. The *T. brucei* genome encodes eight p24 ORFs (TbERP1–8), four of which (TbERP1, 2, 3 and 8) are expressed in BSF (Kruzel *et al.*, 2017). These four proteins were shown to localise to ERES and to be required for VSG transport. They were also shown to be dependent on each other for stability, indicating complex formation as in mammals (Bonnon *et al.*, 2010). After budding from ER exit sites, COPII vesicles traffic proteins to the Golgi where they can be further modified.

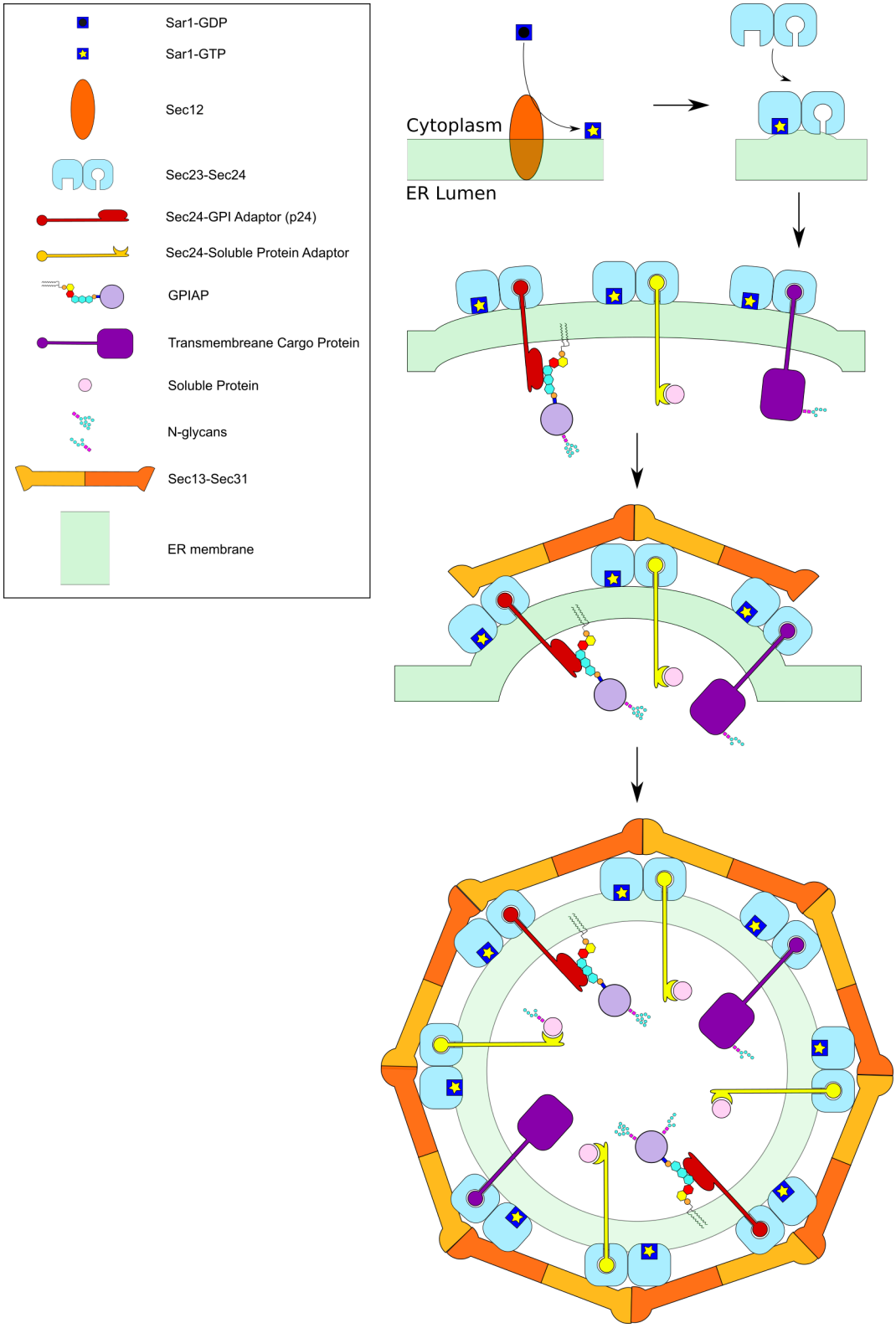


Figure 1-13. COPII mediated vesicle formation. Sec12 induces exchange of GDP for GTP on Sar1. Sar1-GTP binds to the ER membrane, initiating deformation. The Sec23-Sec24 complex binds to Sar1-GTP. Cargo and adaptor proteins bind to the Sec23-Sec24 complex via Sec24. The Sec13-Sec31 heterodimer then binds to the Sec23-Sec24-cargo complexes and concentrates them into a vesicle which then buds off the ER membrane.

1.4.4. Golgi apparatus – the finishing touches

Once in the Golgi, further modifications and elaborations to both N-glycans and GPI anchors can occur. O-glycosylation most commonly features here but can also occur in the cytoplasm or the ER (Steen *et al.*, 1998). Little is known of O-glycosylation in *T. brucei*; what is known will be discussed in this section.

1.4.4.1. GPIAPs

In mammals GPI lipid remodelling occurs in the Golgi (Kinoshita and Fujita, 2016). This process requires the enzymes PGAP3 (for removal of the lipid in the sn-2 position) and PGAP2 (for reacylation of the sn-2 position, usually with the 18-carbon stearic acid). In addition to this, the first mannose residue can be decorated with additional sugar residues. This elaboration always begins with a N-acetylagalactosamine (GalNAc) residue which can subsequently be extended by the addition of a galactose and/or a sialic acid residue (Hirata *et al.*, 2018). Until recently, none of the enzymes involved were known. However, Hirata and colleagues discovered the first enzyme in this process, which attaches the GalNAc residue. They did this by studying a CHO cell line deficient in the transport of UDP-Galactose. This line cannot modify GPI anchor side chains past the initial addition of GalNAc. Random mutagenesis of this line's genome using gene trapping, coupled with the use of an antibody that specifically recognises GPI anchors with the GalNAc residue, allowed identification of the enzyme required for this step, termed PGAP4. Other enzymes involved in the addition of further residues are yet to be revealed.

In contrast to mammals, *T. brucei* GPI anchors are not known to be lipid-remodelled in the Golgi (Hong and Kinoshita, 2009). Sugar sidechains can indeed be attached but differently to mammalian cells. Only galactose has been shown to be used for GPI-anchor side chains in *T. brucei* BSF (Ferguson *et al.*, 1988; A Mehlert *et al.*, 1998; Mehlert and Ferguson, 2007; Zitzmann *et al.*, 2000). Galactose can be added to any of the mannose residues, and anchors have been characterised that contain from none to six galactose residues. It has been suggested that the number of residues attached is regulated by the

accessibility of enzymes to the GPI anchor, based on the attached protein structure (A Mehlert *et al.*, 1998; Zitzmann *et al.*, 2000). This might further accommodate different classes of VSG CTDs, functioning to fill space and aid in VSG packing onto the cell surface. The enzymes involved in these anchor modifications have not yet been characterised.

1.4.4.2. N-glycans

The formation of complex N-glycans in mammalian cells begins with mannose trimming before the addition of a GlcNAc residue to $Man_5GlcNAc_2$ by GlcNAc transferase I (GnTI) (Figure 1-14; Damerow *et al.*, 2014; Stanley *et al.*, 2017). Two mannose residues are subsequently removed by α-mannosidase II enzymes (MAN2A1 and MAN2A2) prior to addition of a second GlcNAc by GlcNAc transferase II (GnTII). In contrast, it has been suggested that *T. brucei* trims the mannose residues of $Man_5GlcNAc_2$ down to a $Man_3GlcNAc_2$ (through unknown enzymes) prior to addition of either GlcNAc residue (Damerow *et al.*, 2014). Thus, whilst all downstream complex-glycan processing in mammals relies on GnTI activity (hence, lack of GnTI causes embryonic lethality in mice), TbGnTI and II are thought to act independently on the same substrate. Initially, TbGnTI and II could not be identified by sequence similarity with eukaryotic homologues; although, through mining *T. brucei* encoded proteins with the human β1-3-N-acetylglucosaminyltransferase sequence (β3GnT5), Izquierdo and colleagues discovered a family of 21 putative UDP sugar-dependent glycosyltransferases of unknown function in *T. brucei* (Izquierdo *et al.*, 2009b). Three of these have been characterised, revealing the identities of TbGnTI and II (Damerow *et al.*, 2014, 2016), and an additional bifunctional glycosyltransferase (TbGT8) involved in N-glycan elaboration in BSF and GPI anchor modification in PCF (Izquierdo *et al.*, 2009b; Nakanishi *et al.*, 2014).

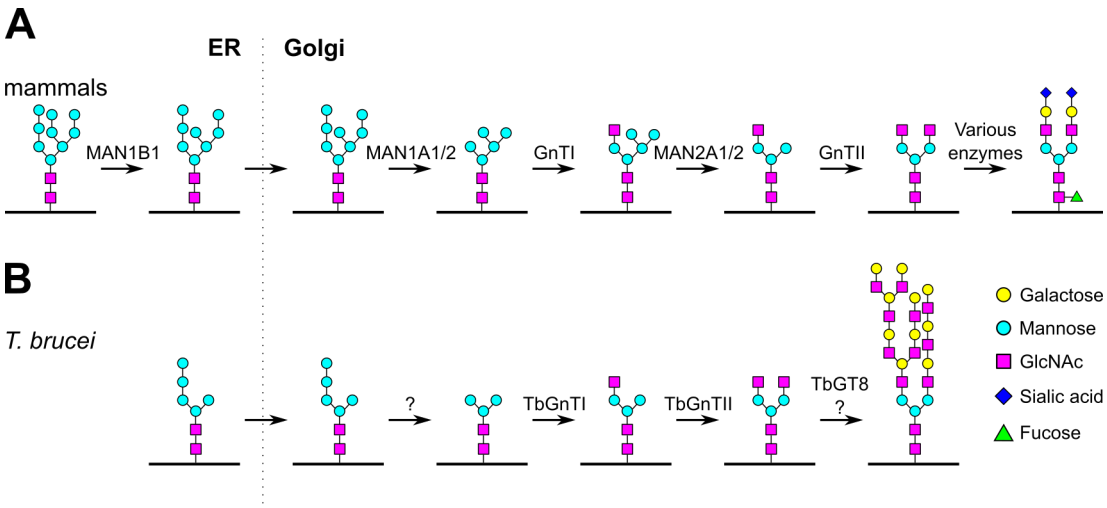


Figure 1-14 Biosynthesis of complex-type N-glycans in mammals and *T. brucei*. Illustration of N-glycan processing in the ER and Golgi, and the enzymes involved in **A** mammals, and **B** *T. brucei*.

1.4.4.3. O-glycosylation

O-glycosylation differs from N-glycosylation in so far as sugar residues are sequentially attached to proteins to create larger structures, rather than the en-bloc transfer of synthesised precursors (Steen *et al.*, 1998). In addition, O-glycosylation attachment sites are more varied than the common N-glycosylation sequons of N-X-S/T and, thus, can be harder to predict *in silico* (although prediction algorithms do exist; Steentoft *et al.*, 2013). O-glycosylation is initiated by the attachment of a variety of different residues: GlcNAc, GalNAc, mannose, galactose, glucose and fucose (Mendonça-Previato *et al.*, 2013; Steen *et al.*, 1998; Takeuchi *et al.*, 2012). These can be further elaborated with sugar residues including the aforementioned sugars, as well as sialic acid and xylose, resulting in branched or linear structures. O-glycosylation does occur in the cytoplasm to non-surface/non-secreted proteins, but this will not be covered here. O-glycosylation of surface proteins is initiated in the ER or the Golgi, depending on the first residue attached, but further elaboration generally happens in the Golgi. GlcNAc is the most common starting residue, and its attachment occurs in the Golgi. The other O-glycan types are initiated in the ER.

Until recently, the only example of O-glycosylation known in trypanosomes was that of mucin-type glycoproteins of *T. cruzi* (the aetiological agent of Chagas' disease; Mendonça-Previato *et al.*, 2013). *T. cruzi* mucin O-

glycosylation starts with the addition of GlcNAc to threonine by UDP-GlcNAc:polypeptide α-N-acetylglucosaminyltransferase (pp-α-GlcNAcT) in the Golgi. Following processing, the resulting O-glycans have a large amount of structural diversity, including galactose (in both the pyranose and furanose configurations) glucofuranose and sialic acid residues (with sialic acid as terminal residues scavenged from the host). Recently, it was reported that some VSGs in *T. brucei* also have O-glycans attached (VSG3, 11 and 615; Pinger *et al.*, 2018). The VSGs studied had 0–3 hexose sugars attached to serine in a surface-exposed loop of the protein 3D structure. The first residue in VSG3 was shown to be glucose. In mammals, O-glucosylation occurs in the ER, in some cases followed by the attachment of one or two xylose residues (Takeuchi *et al.*, 2012). Based on this, O-glucosylation in *T. brucei* may also occur in the ER, although it does differ slightly from mammals to the extent that *T. brucei* attaches glucose in a Glcα1-O-Ser linkage, whilst mammals use a Glcβ1-O-Ser linkage (Pinger *et al.*, 2018; Takeuchi *et al.*, 2012). As the O-glycan of VSG occurred on a surface-exposed loop, Pinger and colleagues decided to investigate its role in parasitaemia, removing the O-glycan through a S317A point mutation. Intriguingly, mice infected with the O-glycan-deficient VSG mutant showed greater survival than those infected with the wildtype version, suggesting the O-glycan to be important for inhibition of host antibody clearance of the parasite. Thus, O-glycosylation may become an important area of study for control sleeping sickness and animal trypanosomiasis.

1.4.5. Exocytosis and the endosomal trafficking system

1.4.5.1. Exocytosis and the exocyst

Once membrane proteins have traversed the Golgi stack, they are transported to their final destination – the flagellar pocket (FP), the sole site of exocytosis in African trypanosomes. Membrane trafficking within eukaryotes is generally controlled by various protein families, including Rabs, SNAREs, vesicle coats and tethering complexes. Rabs are a family of proteins within the Ras superfamily of small GTPases, implicated in vesicle budding, fusion and flagellar/ciliary function (Kelly *et al.*, 2012). Whereas SNAREs are primarily

involved in vesicle fusion. Different Rabs and SNAREs associate with specific organelles and membrane compartments, facilitating vesicular transport between them. A membrane-tethering complex that is important for exocytic vesicle tethering to the plasma membrane is the exocyst (Boehm *et al.*, 2017; Wu and Guo, 2015). The complex contains eight different subunits (Exoc1–8 in mammals; Sec3, 5, 6, 8, 10, 15, Exo70 and Exo84 in yeast). Sec3 (Exoc1) and Exo70 (Exoc7) interact with phosphatidylinositol 4,5-bisphosphate (PIP_2) on the inner leaflet of the plasma membrane; whilst Sec15 (Exoc6) interacts with Rab GTPases associated with the vesicle to be tethered (Figure 1-15; Wu and Guo, 2015). Once the vesicle has been tethered, SNARE proteins mediate its fusion with the plasma membrane.

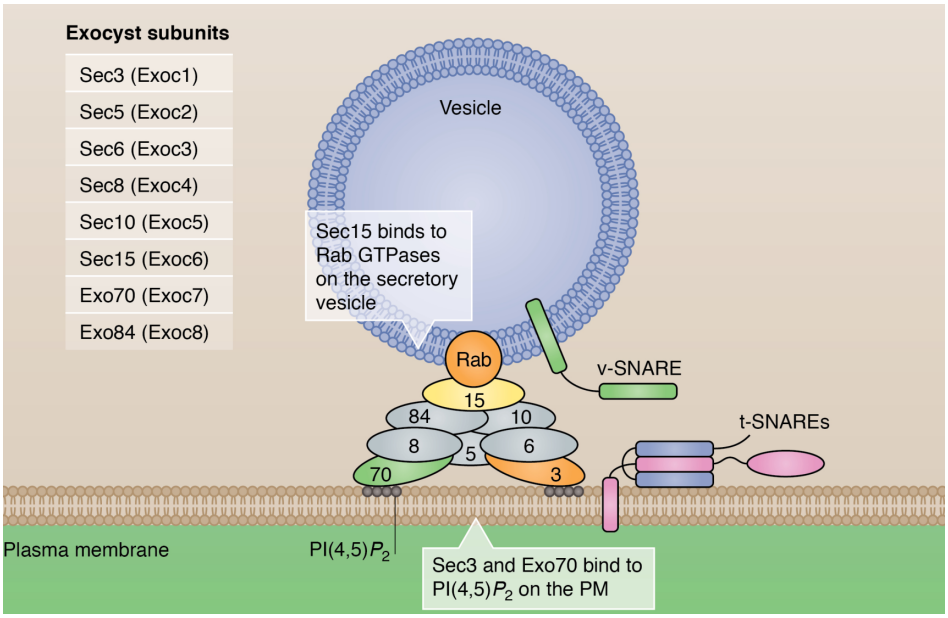


Figure 1-15 Structure and function of the exocyst complex. The exocyst mediates vesicle-plasma membrane tethering. Table on the left shows the names of the yeast (and mammalian) exocyst subunits. The *T. brucei* complex contains an additional ninth subunit (Exo99). Sec15 associates with Rab GTPases on the vesicle, while Exo70 and Sec3 interact with PIP_2 on the plasma membrane. Figure adapted from (Wu and Guo, 2015)

Recently, the exocyst of *T. brucei* has been identified and characterised through sequence homology, affinity purification and mass-spectrometry (Boehm *et al.*, 2017). The complex contains homologues for all 8 previously identified exosome components; albeit with TbSec5 and TbExo84 being

significantly larger than their fungal and mammalian counterparts. In addition to these 8 proteins, the *T. brucei* exocyst contains a novel, ninth subunit named Exo99. Co-localisation of TbSec15 and Exo99 (in both PCF and BSF) by epitope-tagging and immunofluorescence confirmed its role as an additional excocyst component. Both proteins localised to one or two puncta neighbouring the FP. Ablation of either TbSec15 or Exo99 was shown to be lethal. Intriguingly, both knockdowns also resulted in a proportion of cells exhibiting a 'big-eye' phenotype, whereby the FP becomes enlarged. This phenotype is associated with endocytosis defects. Inhibition of endocytosis by TbSec15 or Exo99 depletion was confirmed using fluorescent endocytic markers, highlighting a previously unexplored role for the exocyst. Indeed, the authors went on to show that ablation of Exoc6 (the mammalian Sec15 homologue) also reduced endocytosis in HeLa cells.

1.4.5.2. The final destination and back again

The 'end-point' of the secretory pathway is the delivery of membrane-associated or luminal proteins to the FP. From there, molecules may be secreted to the extracellular milieu, reside within the FP membrane, migrate out of the FP across the flagellar and/or cell body membranes, or be rapidly internalised through the endocytic pathway, which in trypanosomes is entirely dependent on clathrin. Even then, endocytic carriers may recycle molecules back to the cell surface or direct them all the way to lysosomal degradation. Endocytosis in African trypanosomes is fast enough to turn-over the entire VSG surface coat in just 12 minutes (Engstler *et al.*, 2004). This rapid membrane flux facilitates antibody clearance from the surface, aiding in immune evasion. Internalisation of VSG is thought to be a passive process (Grunfelder *et al.*, 2003), whilst endocytosis of transmembrane proteins may be regulated by ubiquitin, as ubiquitinylation of ISG65 and ISG75 leads to their endocytosis and lysosomal degradation (Chung *et al.*, 2008; Leung *et al.*, 2011). Recycling of GPIAPs is likely important for maintenance of the VSG coat density, as the concentration of VSG increases roughly 20-fold between the Golgi and the surface (Grunfelder *et al.*, 2002; Overath and Engstler, 2004). Indeed, through selective biotinylation of surface VSG, it was revealed that ~90% of the VSG

pool resides on the surface of the cells; immunofluorescence and immunoelectron microscopy showed ~75% of the internal VSG to be present in endosomes (Grunfelder *et al.*, 2002).

The *T. brucei* endosomal system is regulated by Rabs (Manna *et al.*, 2014). Figure 1-16 summarises the different Rab proteins associated with specific endocytic compartments. GPIAPs and transmembrane proteins are differentially sorted within this system. TbRAB5A has been associated with internalisation of extracellular molecules to the early endosomes, as its ablation results in big-eye cells (Hall *et al.*, 2004). TbRAB5B, on the other hand, has i) been associated with endocytosis (Hall *et al.*, 2004); ii) shown to localise in a different endosomal compartment to TbRAB5A (Pal *et al.*, 2002), and iii) to co-localise specifically with a non-GPI anchored protein (ISG_{100}) but not VSG or transferrin (Pal *et al.*, 2002). In addition, depletion of TbRab11 (which is associated with the sorting endosomes; Figure 1-16) does not affect VSG or TfR trafficking, but it does reduce recycling of ISG65, increasing its lysosomal degradation (Umaer *et al.*, 2018).

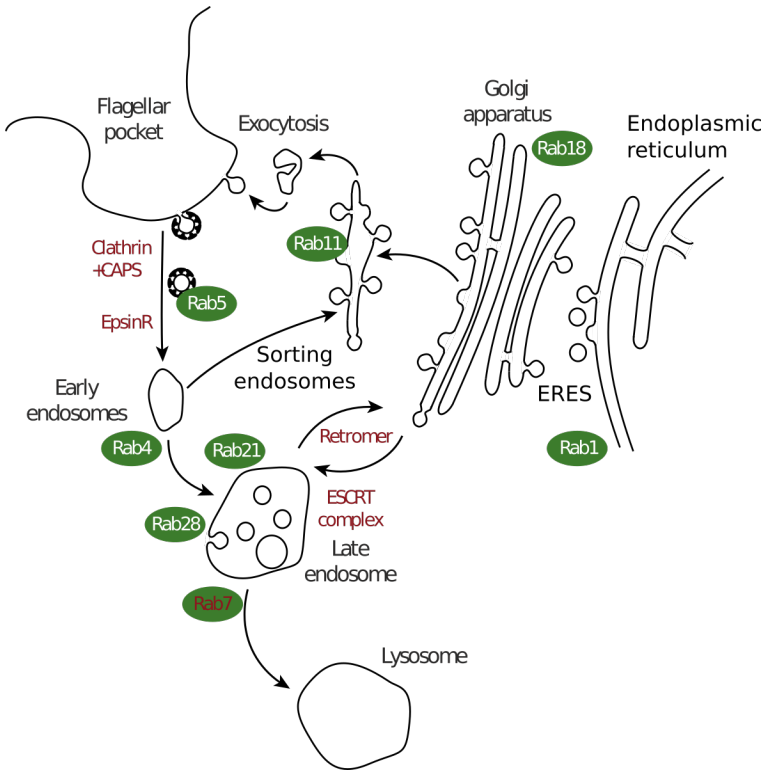


Figure 1-16 The endosomal trafficking system of *T. brucei*. Intracellular compartments involved in the trafficking of surface proteins. Rab GTPases regulate

vesicle trafficking and serve as markers for subcellular compartments. Other proteins and complexes include: clathrin-associated proteins (CAPS); EpsinR, a clathrin-interacting protein involved in vesicle trafficking; the retromer complex, known to be important for recycling proteins to the Golgi; and the edosomal-sorting complex required for transport (ESCRT). Adapted from Manna *et al.*, 2014.

1.4.6. The surface of *Trypanosoma brucei* BSF

The plasma membrane of bloodstream-form *T. brucei* is highly specialised, and surface molecules localise to one or combinations of domains – the cell body, the flagellum, and the flagellar pocket membranes – within this single contiguous lipid bilayer (Gadelha *et al.*, 2015, 2009; regions highlighted by different colours in Figure 1-17A). Surface membrane specialisation is again thought to aid in the extracellular lifestyle of the parasite, whereby vital receptors are contained within the FP, where they are thought to be shielded from host immune recognition (e.g. TfR and HpHbR; Salmon *et al.*, 1994; Vanhollebeke *et al.*, 2008). Indeed, the solved crystal structure of HpHbR indicates that it would extend above the VSG coat for interaction with its ligand, potentially exposing it to the immune system (Higgins *et al.*, 2013; Stødkilde *et al.*, 2014). Thus, FP retention of this receptor may function both to restrict it to the only endocytic-active on the cell surface, and to shield it from immune attack. However, ISG65, a molecule known to localise to the entire cell surface (the FP, flagellum and cell body membranes), has had its primary sequence aligned to the TcoHpHbR structure (protein threading), revealing its longest axis to be of similar length to that of VSG (Schwede *et al.*, 2015). This could imply ISGs to reach most or all of the way through the VSG coat. Invariant surface molecules with elongated structures may prove effective in the development of vaccines against this deadly parasite (this will be expanded upon in Chapter 4).

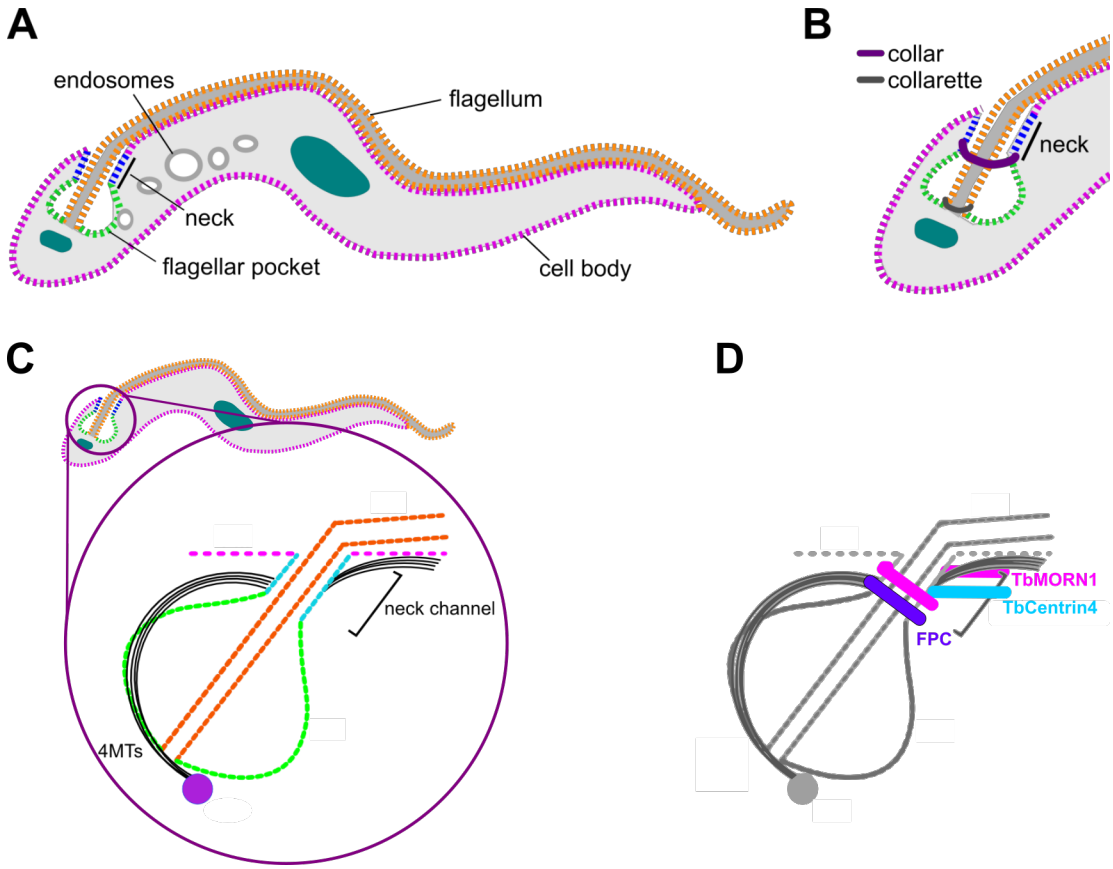


Figure 1-17 Surface membrane architecture and underlying structural features of BSF *T. brucei*. A Illustration indicating the specialised surface domains of the *T. brucei* plasma membrane, which differ in both lipid and protein composition: cell body (magenta); flagellum (orange); flagellar pocket (green); neck (blue). Mitochondrial and nuclear DNA shown in teal. **B** The FP region showing two cytoskeletal structures located at the boundaries between the FP and neck membranes (collar) and the FP and flagellum membranes (collarette). **C** and **D** Insets highlighting the structural features underlying the FP. A set of 4 specialised microtubules (MTs) originate between the basal and probasal bodies (magenta circle) and wrap around the FP in a left-handed helix. The major cytoskeletal features (**D**) include the flagellar pocket collar, the hook complex (depicted by its major protein, TbMORN1) and the centrin arm. Figure kindly provided by Sarah Whipple.

All proteins that reach the surface of *T. brucei* reach first the FP membrane, from where they can move by two-dimensional lateral diffusion across the entire plasma membrane. The differential localisation of surface molecules on the membrane, defining distinct specialised domains, implies the existence of physical barriers that maintain the molecular identity of these domains and may regulate the inclusion/exclusion of specific proteins (Gadelha *et al.*, 2015). Transmembrane protein arrangements resembling structural barriers have been observed by freeze-fracture electron microscopy at the

junction between flagellum:flagellar pocket and pocket:cell body membranes (Gadelha *et al.*, 2009). Examples of barriers in other systems include the polarisation of mammalian epithelial cells into apical and basolateral domains separated by tight junctions. In addition, the apical domain can contain cilia with distinct protein compositions relative to the rest of the apical surface membrane, also maintained by physical structures (Hu *et al.*, 2010). Tight junctions consist of a multiprotein complex composed largely of tetra-spanning transmembrane proteins, claudins and occludins (Furuse, 2010). Claudins are known to form protein strands that cross the lateral junction between cells, preventing the movement of lipids and proteins between the outer leaflets of the apical and basolateral membranes (Furuse, 2010). The protein composition of cilia is maintained by septins, which are GTP-binding proteins often associated with regions of membrane curvature (Hu *et al.*, 2010). Septins are also involved in the budding of yeast (Longtine and Bi, 2003), and in the maintenance of protein distribution between the posterior and anterior regions of sperm cell tail membranes (Ihara *et al.*, 2005). In addition to membrane barriers, the underlying cytoskeleton can also play a role in the distribution of membrane proteins. For example, in neurons dense regions of membrane proteins are formed between the axonal and somatodendritic domains through linkage to F-actin within the cytoskeleton, preventing molecular diffusion between the two domains (Nakada *et al.*, 2003).

In trypanosomatids, overall cell shape is defined by an array of subpellicular microtubules beneath the plasma membrane. Microtubules are mostly absent from the FP region. Instead the FP collar (purple in Figure 1-17D), the hook complex (depicted by TbMORN1 in Figure 1-17D), the centrin arm (blue in Figure 1-17D) and a set of four rootlet microtubules (4MTs; Figure 1-17C and D) are thought to maintain this specialised domain. The hook complex (HC; named for its hook-like shape around the FP neck) and the centrin arm form two domains of what is known as the bilobe structure, through which the 4MTs pass (Esson *et al.*, 2012). Three HC proteins have been characterised to date – TBCCD1, LRRP1 and TbMORN1 (André *et al.*, 2013; Morriswood and Schmidt, 2015; Zhou *et al.*, 2010). TBCCD1 and LRRP1 have been studied in PCF cells, where TBCCD1 localises to the centriole, the bilobe and the anterior

end of the cell (André *et al.*, 2013), whilst LRRP1 regulates FAZ assembly, flagellum inheritance, and cell division (Zhou *et al.*, 2010). TbMORN1 was characterised in BSF, where its ablation resulted in an enlarged FP and a defect in FP entry and endocytosis (Morriswood and Schmidt, 2015).

The collar forms a horseshoe shaped electron dense structure around the FP neck region (Bonhivers *et al.*, 2008; Gadelha *et al.*, 2009). Two components have been defined, the first of which was BILBO1 (Bonhivers *et al.*, 2008). BILBO1 is able to form homodimers through interactions between coiled-coil domains (Florimond *et al.*, 2015). It has been suggested that assembly of BILBO1 into a ring-like structure provides a platform upon which the rest of the collar structure can be built. Ablation of BILBO1 in PCF cells led to cells with abnormal morphology and inhibited the biogenesis of a new collar, FP and FAZ (Bonhivers *et al.*, 2008). Data from our lab revealed ablation of BILBO1 in BSF to cause the loss of FP localisation of a set of four distinct transmembrane surface proteins (ESP10, 11, 21 and TbPT0; Whipple, 2017). Thus, the collar may function as a barrier to protein diffusion across the plasma membrane, although it remains unclear if other structures are still present around the FP upon BILBO1 ablation. The second collar component was named FPC4 (Albisetti *et al.*, 2017). This protein co-localises with the collar and the HC, and shown to both interact with BILBO1 and bind to microtubules. Hence, it was suggested to link together the FPC, HC and the 4MT.

The discussed structures clearly play important roles in the formation and maintenance of the FP. Studies using endocytic markers also implicate these cytoskeletal features in the possible regulation of FP protein retention and access to the FP lumen. However, determinants of membrane barriers remain elusive. As does a definite mechanism of molecular sorting across the plasma membrane. Possible factors influencing sorting include glycosylation, the number of GPI anchors, and protein amount. These will be expanded upon more in Chapter 3.

1.5. Conclusions

The surface biology of this clinically-relevant parasite contributes substantially to its extracellular lifestyle, utilising multiple mechanisms for evasion of the host immune system. This is modulated, among other factors, by how trypanosomes process its surface constituents, in similar yet distinct ways to other well-studied eukaryotes. Our knowledge about the molecular composition of the parasite cell surface has also been significantly augmented in recent years, making timely the exploitation of novel eukaryotic surface biology, as well as interventions to tackle trypanosomiasis and related diseases. In this regard, invariant, validated surface-exposed proteins could potentially be exploited in the development of vaccines against this deadly parasite.

1.6. Aims

The present study has two main aims, both of which take advantage of the expanded knowledge that now exists on surface proteins of bloodstream-form *T. brucei*. The first aim is a question of fundamental biology, that asks if the GPISSs from proteins of distinct cellular localisations influence their delivery to the appropriate target membrane. The second aim looks at contributing towards a protective vaccine against trypanosomes with the development of tools for recombinant surface antigen expression.

Chapter 2. Methods

Unless stated, the majority of chemicals and reagents used in this study were obtained from Sigma, including brain heart infusion (BHI) powder. Foetal bovine serum (FBS) came from Sigma or Thermo Fisher Scientific (Gibco), whilst Iscove's Modified Dulbecco's Medium (IMDM) came from the latter. Selection drugs were also purchased from Thermo Fisher Scientific (Invitrogen). Enzymes for molecular biology came from NEB.

2.1. Cell lines, cell culture and transfection

2.1.1. *T. brucei*

Bloodstream-form *T. brucei* cell lines were derived from *T. b. brucei* Lister 427 single marker line known as S16 (Wirtz *et al.*, 1999). Cells were maintained in IMDM supplemented with 1 mM cysteine, 50 µM bathocuproine, 192 µM β-mercaptoethanol, 90 µM cytosine, 1 mM hypoxanthine, 90 µM uracil and 15% FBS, at 37°C and 5% CO_2. Cells were maintained within mid-log phase (~10^3–10^6 cells/mL). Transfected cells were maintained under drug selection pressure. Selection markers, drugs and concentrations used are shown below:

Table 2-1. Selection markers and drugs used for *T. brucei* cell lines in this study

Selection marker	Selection drug	Concentration (µg/mL)
HYG^R (hygromycin phosphotransferase)	Hygromycin B	5
PUR^R (puromycin N-acetyltransferase)	Puromycin	0.1

2.1.2. *L. tarentolae* and *C. fasciculata*

Promastigote-form *L. tarentolae* cell lines were derived from the T7TR line purchased from Jena Bioscience. Choanomastigote form *C. fasciculata* cell lines were derived from a wild-type clone used by Gadelha *et al.*, 2005. Both *L. tarentolae* and *C. fasciculata* were maintained in BHI medium supplemented with 10 µg/mL hemin and 5% FBS, at 27°C. Cells were maintained within mid-

log phase (~10^5–10^8 cells/mL). Inducible cell lines were induced with 10 µg/mL tetracycline. Transfected cells were maintained under drug selection pressure. Selection markers, drugs and concentrations used are shown below:

Table 2-2. Selection markers and drugs used for *L. tarentolae* cell lines in this study

Selection Markers	Selection drug	Concentration (µg/mL)
HYG^R (hygromycin phosphotransferase)	Hygromycin B	50
NAT^R (nourseothricin N-acetyltransferase)	Nourseothricin	50
NEO^R (neomycin phosphtransferase)	G418	100

Table 2-3. Selection markers and drugs used for *C. fasciculata* cell lines in this study

Selection Markers	Selection drug	Concentration (µg/mL)
HYG^R (hygromycin phosphotransferase)	Hygromycin B	100
NEO^R (neomycin phosphtransferase)	G418	50

2.1.3. Transfection

2–5x10^7 cells in mid-log phase were harvested and resuspended with 10–20 µg/mL linearised plasmid DNA in 120 µL Tb-BSF buffer (90 mM $NaHPO_4$, 5 mM KCl, 0.15 mM $CaCl_2$, 50 mM HEPES pH 7.3; Burkard *et al.*, 2011). This cell/DNA suspension was electroporated with an Amaxa Nucleofector 2b (Lonza) programme Z-001. Cells were then transferred to fresh medium to recover for 8–16 hours (*T. brucei* and *L. tarentolae*) or 4–8 hours (*C. fasciculata*) before selection drug was added to the culture. Clonal transfectants were obtained by limiting dilution in 96-well plates. Cells were also transfected with circular plasmids by the same method but kept as populations of transfectants.

In addition to the above protocol, in some experiments *L. tarentolae* and *C. fasciculata* were transfected as follows: 2–5x10^7 cells in mid-log phase were harvested and resuspended with 10–20 µg/mL linearised or circular plasmid DNA in 400 µL culture medium. This cell/DNA suspension was electroporated

with an Eppendorf Eporator set at 1700V, using two pulses 10 seconds apart. Cells were transferred to fresh medium to recover for 8–16 hours (*L. tarentole*) or 4–8 hours (*C. fasciculata*) before selection drug was added. Clonal transfectants were obtained by limiting dilution in 96-well plates.

2.1.4. Analysis of growth rate

Growth rate was monitored using an Improved Neubauer haemocytometer or a FastRead Disposable Counting Slide. Cells were counted either live or fixed with 2.5% formaldehyde. For growth curves, approximately 100 cells were counted at each 8-hour time point across three days.

2.2. Bioinformatics

2.2.1. Protein sequence analysis

Signal peptide amino acid sequences were predicted using SignalP v3.0 (Dyrløv Bendtsen *et al.*, 2004) and GPI-anchor attachment signal sequence was predicted with PredGPI (Pierleoni *et al.*, 2008). Transmembrane domain predictions were made using TMHMM v2.0 (Krogh *et al.*, 2001; Sonnhammer *et al.*, 1998).

2.2.2. Codon usage bias

C. fasciculata (strain Cf-Cl) ribosomal protein DNA sequences were downloaded from TriTrypDB (www.tritrypdb.org). Relative synonymous codon usage (RSCU) was calculated using the following formula:

$$RSCU = \left(\frac{n_i X_{ij}}{\sum_{j=1}^{n_i} X_{ij}} \right) - 1$$

Where n_i = number of synonymous codons for amino acid i, Xij = number of occurrences of codon j for amino acid i. A value of 0 indicates no codon bias

for that particular amino acid. RSCU values were used to codon optimise ORF sequences for expression in *C. fasciculata*.

2.3. Molecular Biology

2.3.1. Molecular cloning

Unless otherwise stated, *T. brucei* DNA fragments were amplified using *T. b. brucei* Lister 427 genomic DNA as template (previously purified in the lab from DNA agarose plugs). For *C. fasciculata* DNA fragments, *C. fasciculata* genomic DNA from a clone derived from Gadelha *et al.*, 2005, was used as template. This *C. fasciculata* genomic DNA was obtained from $1x10^8$ cells using a DNeasy blood & tissue kit (Qiagen). A typical PCR reaction included 25 ng of template gDNA, 0.2 mM dNTPs, 3 molar excess of each primer and 1.25 units of Phusion High-Fidelity DNA polymerase (NEB). Restriction digestion of DNA was carried out with 10- or 20-fold excess of restriction enzymes for plasmid and insert DNA respectively. DNA ligations were performed using T4 DNA ligase at 4°C, 14°C or room temperature for 16 hours, 2 hours or 1 hour respectively. For two-way ligations 3–5:1 molar ratio of insert:vector was used. Plasmids were transformed into *Escherichia coli* strain XL1-Blue by heat shock, which were subsequently grown at 37°C on Luria Broth agar plates supplemented with 100 µg/mL ampicillin. Plasmids were isolated from liquid cultures by alkaline lysis and silica-adsorption-based DNA purification. Purified plasmids were Sanger-sequenced across ligation joins to ensure correct fragment insertion. Prior to eukaryotic transfections with linear DNA, plasmids were digested with 1.5-fold excess of NotI or SwaI (depending on plasmid linearization site), purified by anion-exchange chromatography and sterilised by ethanol precipitation. Primers and synthetic DNA fragments (gBlocks) were manufactured by Integrated DNA Technologies (IDT).

2.3.1.1. Fusion PCR

In the production of plasmids, certain fragments were joined together by fusion PCR. 2 or 3 fragments to be joined were initially amplified with complementary overlapping ends of 15–25 bp. Fragments (~100 ng total) were

then mixed into a second PCR reaction which included a forward primer to anneal to the most 5' fragment, and a reverse primer to anneal to the most 3' fragment. For fusion of two fragments, equimolar ratios were used. For fusion of three fragments, the molar ratios used were 1:4:1 (with 4x excess of the middle fragment).

2.3.1.2. Gibson Assembly

Gibson Assembly was used to combine 3 to 5 inserts and a vector backbone for the generation of various plasmids. Fragments to be combined were amplified with complementary overlapping ends of 20–30 bp. Gibson Assembly reactions (using the NEB Gibson Assembly Master Mix) were carried out with 3:1 insert:vector molar ratio in a final volume of 4 µL using a total of ~0.05 pmol of DNA. 2 µL from each reaction were used for transformation of XL1-Blue *E. coli* by heat shock.

2.3.2. GPI-anchored GFP plasmids

The pSiG vector series (Gadelha *et al.*, 2015) is a suite of vectors for endogenous locus tagging in bloodstream-form *T. brucei*. pSiG stands for SP at the N-terminus and GPISS at the C-terminus, both derived from the VSG221 ORF (gene ID Tb427.BES40.22). The base vector pSiG-HhsfG also contains *HYG*^{R}, 3xHA tags, and *sfGFP*. By adding the *TUBB* ORF with a linearization site in the middle – pSiG-HhsfG-Tub – it is possible to direct ectopic expression of membrane bound sfGFP (with a GPI anchor).

pSiG-HhsfG-Tub derivatives, each containing a different GPISS, were previously generated in the lab by Sabine Schiessler. GPISSs were derived from ESAG2, ESP5, ESAG6 and GRESAG9. Further modifications were made in this study to the ω-4 to ω-1 positions of pSiG-HhsfG_ESAG2_omega_m4-Ct and pSiG-HhsfG_ESAG6_omega_m4-Ct, using primers with different tags (see Table 2-4 for fragments, primers and resultant constructs). Fragments were amplified from their respective plasmids and ligated back in by two-way ligation using XbaI and BamHI restriction sites and T4 DNA ligase.

2.3.3. LEXSY plasmid modifcations

pLEXSY_IE-sfGO-N and pLEXSY_IE-sfG-N were generated by cloning gBlocks encoding TEV-*sfGFP*^{op}-His_6 or TEV-*sfGFP*-His_6 respectively between NcoI and NotI restriction sites of pLEXSY_IE-egfp-red-neo4 (Jena Bioscience).

pLEXSY_neo2_LiTat1.3 and pLEXSY_neo2_ISG65 were generated by isolating *NEO*^{R} from pLEXSY_IE-sfGO-N using restriction enzymes SpeI and BamHI, and cloning it between SpeI and BamHI restriction sites of pLEXSY_hyg2_LiTat1.3 and pLEXSY_hyg2_ISG65 (kindly provided by Barrie Rooney, now at TroZon X17)

2.3.4. CExSy – Crithidial expression system

2.3.4.1. Single marker *Crithidia* plasmid (pSMC) construction

pSMC was generated de novo. *TUBB* UTR sequences and *PFR1* and *PFR2* intergenic regions were amplified from *C. fasciculata* genomic DNA. *TetR* and *T7RNAP* ORFs were obtained by restriction digestion of pSmOx (Poon *et al.*, 2012). ColE1 origin and ampicillin resistance marker obtained by digestion of pSPR0 (Daniels *et al.*, 2012). The following fragments were combined by fusion PCR: *TUBB* 5' UTR to *T7RNAP*; *PFR1* intergenic to *NEO*^{R}; *PFR2* intergenic to *TETR* to *TUBB* 3' UTR. The resultant fragments were then assembled into pSMC using Gibson Assembly.

2.3.4.2. *Crithidia* expression plasmid (pCEx) construction

pCEx was constructed de novo. *TUBA* and *TUBB* UTR sequences were amplified from *C. fasciculata* genomic DNA. A DNA fragment encoding a '10% strength' T7 promotor (Wirtz *et al.*, 1998), the *PGKB* 5' UTR, the *HYG*^{R} resistance marker and the *GSPS* 3' UTR was amplified from pNUS-GFPcH (Tetaud *et al.*, 2002). TEV-*sfGFP*^{op}-His_6 was amplified from pLEXSY_IE-sfGO-N. T7 terminators were amplified from pLEXSY_IE-sfGO-N. A gBlock DNA fragment encoding a targeting sequence for the *C. fasciculata* rDNA intergenic

spacer was manufactured by IDT. ColE1 origin and ampicillin resistance marker were amplified from pSPR5 (Daniels *et al.*, 2012). The following fragments were combined by a single fusion PCR reaction: *TUBB* 5' UTR to TEV-*sfGFP*^{op}-His_6 to *TUBA* 3' UTR. The resultant fragment was then assembled with the remaining fragments into a vector using Gibson Assembly. Finally, a full strength T7 promotor sequence was inserted into the vector between EcoRI and HindIII restriction sites, using two complimentary primers, finally creating pCEx.

2.3.4.3. pCEx modification – UTRs and tet operators

The following UTRs were amplified from *C. fasciculata* genomic DNA: *PFR2* 5'; ribosomal protein *L6* 5'; *PFR1* 3'; ribosomal protein *L6* 3'. *PGKB* 5' and *GSPS* 3' UTRs were amplified from pNUS-GFPcH. 5' UTRs were cloned into pCEx using HindIII and BglII sites. 3' UTRs were cloned into pCEx using NdeI and NsiI restriction sites.

To create pCEx with the 1T and 2T arrangements, complimentary primers were used that encoded the corresponding T7-tet operator arrangements. These were cloned between EcoRI and HindIII restriction sites of pCEx. To create pCEx with the 3T arrangement, complimentary primers encoding one tet operator was cloned into the EcoRI restriction site of the pCEx_2T derivative.

2.3.4.4. pCEx modification – pCExP

rDNA IGS targeting fragments were amplified from *C. fasciculata* genomic DNA. A single DNA fragment encoding *PGKB* 5' UTR, *HYG*^{R} and *GSPS* 3' UTR was amplified from pCExC. As a first step, the downstream rDNA IGS targeting fragment was cloned into EcoRI and HindII sites of pCExC. Next, the *PGKB* 5'-*HYG*^{R}-*GSPS* 3' fragment was digested with NsiI and XbaI; the upstream rDNA IGS targeting fragment digested with XbaI and SacI. These two fragments were cloned between the modified pCExC (containing the downstream rDNA IGS targeting fragment) by three-way DNA ligation between NsiI and SacI restriction sites.

2.3.5. Diagnostic PCR

Diagnostic PCR was used to investigate integration of pSMC and pCEx constructs into the *C. fasciculata* genome of relevant transfectant lines. For genomic DNA extraction, $2x10^5$ cells were lysed by hot alkaline lysis in 120 µL of 50 mM NaOH, 2 mM EDTA, heated to 95°C for 10 minutes, before neutralisation with 24 µL of 450 mM Tris-HCl pH 6.8. 2 µL of this preparation was used as template DNA in 25 µL-diagnostic PCR reactions.

2.4. Protein biochemistry

2.4.1. Whole cell lysates

For preparation of whole cell lysates, mid-log phase cells were harvested at 1800 g for 2 minutes and washed twice with phosphate buffered saline (PBS; 140 mM NaCl, 10 mM Na_2HPO_4, 3 mM KCl, pH 7.5). One volume of two-times concentrated ([2x]), hot (95°C) Laemmli buffer modified to pH 7.2 (100 mM Tris-HCl, 4% SDS, 20% glycerol, 0.8 M β-mercaptoethanol) was added to cells suspended in PBS, giving a final concentration of $2.5x10^5$ cells/µL of Laemmli buffer ($5x10^5$ cells/µL in Chapter 4). Samples were heated at 95 °C for 5 minutes. Cells expressing GFP-tagged multipass transmembrane proteins were heated at 50 °C for 10 minutes to prevent hydrophobic protein aggregation.

2.4.2. GFP ruler

Laemmli samples of recombinant GFP at different protein concentrations were made for protein quantification by immunoblot. This was done using His_6-eGFP previously produced in our lab (Simon D'Archivio and Catarina Gadelha), expressed in *E. coli* strain M15 using the pQE30 expression system (Qiagen), and purified by Nickel affinity chromatography.

2.4.3. Hypotonic lysis

$3x10^7$ mid-log phase *T. brucei* cells were harvested and washed twice in-ice cold (0 °C) PSG (PBS supplemented with 20 mM glucose). Cells were then resuspended in 90 µL ice-cold H_2O with the following mixture of protease inhibitors: 1 mM EDTA, 5 µM E-64d, 7.5 µM pepstatin A, 50 µM leupeptin, 0.5 mM phenylmethysulfonyl fluoride (PMSF), and 2 mM 1,10-phenanthroline. The cell suspension was left on ice for 15 minutes and then split into three samples (**1**, **3** and **4** as shown in Figure 2-1). To sample **1,** hot (95 °C) Laemmli buffer was added to prepare whole cell lysate. Sample **3** was fractionated into supernatant (**2**) and pellet (**3**) fractions by centrifugation at 3400 g at 4 °C for 5 minutes; hot Laemmli buffer was then added to each fraction. Sample **4** was transferred to 37 °C for 15 minutes before being fractionated into supernatant (**5**) and pellet (**4**) fractions at 3400 g at room temperature for 5 minutes; and hot Laemmli buffer then added to each fraction. All samples were heated to 95°C for 5 minutes (except for negative control cells expressing GFP-tagged multipass protein ESAG10 which was treated at 50 °C for 10 minutes to prevent protein aggregation).

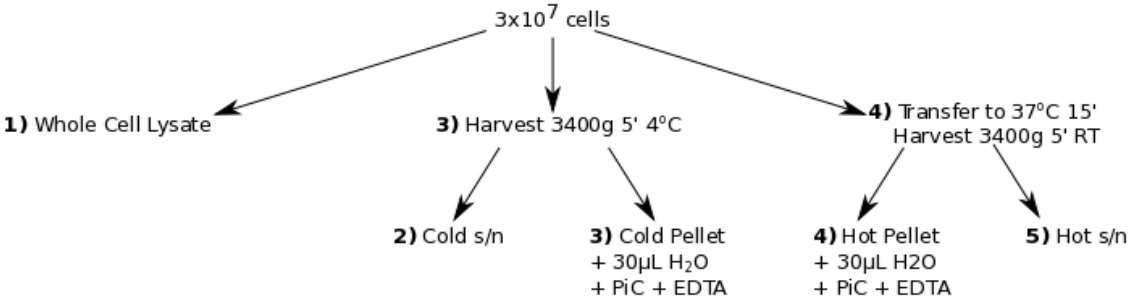


Figure 2-1 Workflow of hypotonic lysis assay. Schematic showing the key steps of hypotonic lysis of *T. brucei* and the separation of different fractions that become samples **1**–**5**.

2.4.4. Detergent extraction of membrane proteins

L. tarentolae cells were cultured for 24 h in the presence of 10 µg/mL tetracycline, washed twice with PBS before resuspension in PBS at $1x10^6$ cells/µL. 1 volume of [2x] detergent lysis buffer was added (2 or 4% detergent, 400 µg/mL DNAseI, protease inhibitors (10 µM E64-d, 4 mM 1,10-

phenanthroline, 1 mM PMSF, 100 µM leupeptin and 15 µM pepstatin A) in PBS). Detergents tested were Triton X-100, Octyl glucoside and Nonidet P40. Cells were lysed at 0°C for 5 or 30 minutes before separation by centrifugation into soluble and insoluble fractions. The insoluble fractions were resuspended in PBS before the addition of 1 volume of hot (95°C) [2x] Laemmli buffer, giving a final concentration of $5x10^5$ cell equivalents/µL. The soluble fractions were precipitated with super cold acetone (-80°C; as described in 2.4.6) before resuspension in hot (95 °C) Laemmli buffer at $5x10^5$ cell equivalents/µL. Laemmli samples were heated at 95°C for 5 minutes before storage at -80 °C.

2.4.5. Removal of glycans

Endo-H and PNGase-F glycosidases were used to assess the glycosylation status of sfGFP-tagged proteins in *T. brucei,* and of partially-purified $sfGFP^{op}$-tagged recombinant proteins produced in *C. fasciculata* or *L. tarentolae*. The assay for *T. brucei* whole cells and partially-purified proteins differed. Regarding *T. brucei,* $2x10^7$ (cells expressing sfGFP-tagged ESP13 and ESP14) or $4x10^7$ cells (cells expressing sfGFP-tagged ESP10) were harvested and washed twice with PSG prior to resuspension in PSG at $2x10^6$ or $4x10^6$ cells/µL respectively (final volume, 10 µL). Conversely, ~40 ng of each partially-purified protein in PBS were topped up to 10 µL total volume. From here on, all samples were treated as follows. One volume (10 µL) of [2x] NEB Denaturing buffer (1% SDS, 80 mM dithioreitol) was added and samples heated to 95 °C for 10 minutes. Each sample was divided into 3 sub-samples of ~6 µL, which was treated with no enzyme, with Endo-H or with PNGase-F. For Endo-H treatment, 2000 U of Endo-H (NEB) and 1 µL of [10x] NEB Glycobuffer 3 (50 mM sodium acetate, pH 6.0) were added. For PNGase-F treatment, 0.1% Nonidet P40, 1000 U of PNGas-F (NEB) and 1 µL of [10x] NEB Glycobuffer 2 (50 mM sodium phosphate, pH 7.5) were added. H_2O was added to untreated samples instead of glycosidase enzymes. Final reaction volume was 10 µL for each treatment. Sub-samples were subsequently incubated at 37°C for 1 hour before addition of one volume of Laemmli buffer and heating to 95°C for 5 minutes.

2.4.6. Total protein precipitation by cold acetone ('medium samples')

Cell cultures were spun at 1800 g for 2 minutes, and the supernatant was seperated from the cell pellet. Cells were used to make whole cell lysates as above. To the supernatant, 10 volumes of super cold acetone (-80 ^{o}C) was added, and the mixture incubated at -80 ^{o}C for at least 2 hours. Precipitated protein was harvested at 18,000 g 4 ^{o}C for 10 minutes, washed twice with -80^{o}C super cold acetone and concentrated at 18,000 g 4 ^{o}C for 2 minutes. Pellets were air-dried for 1 minute, resuspended in hot (95 ^{o}C) Laemmli buffer and heated at 95^{o}C for 5 minutes before storage at -80 ^{o}C.

2.4.7. Selective protein precipitation by ammonium sulphate

Cells were grown under orbital shaking (140 rpm) to top density ($2x10^{8}$ cells/mL) and left in stationary phase for 48 hours. A sample was taken to prepare whole cell lysates and acetone-precipitated medium samples (as described in 2.4.6). The remaining culture was harvested at 800 g for 10 minutes, and the supernatant collected. From this point onward, supernatant was kept at 4^{o}C. Solid ammonium sulphate was added gradually to the culture supernatant, with magnetic stirring, to the desired percentage saturation. Once all the ammonium sulphate dissolved, the solution was left stirring for 30 minutes. Precipitated protein was then harvested at 18,000g for 30 minutes. The supernatant was removed and stored. The pellet was resuspended in the minimum volume of PBS and sampled at the same relative concentration as the initial precipitated medium sample. Ammonium sulphate precipitation was then repeated on the supernatant as many times as required to obtain the desired fractionation.

2.4.8. Immobilised metal affinity chromatography (IMAC)

IMAC experiments were carried out using Ni-NTA Agarose (Qiagen). After ammonium sulphate precipitation, the relevant protein fractions were resuspended in binding buffer (500 mM NaCl, 0.2% Tween-20, 50 mM sodium phosphate pH 7.5) and allowed to bind to 2–100x excess of Ni-NTA resin (based

on the binding capacity of 25 mg of protein per mL of resin stated by the manufacturer) for 2 hours at 4 ^{o}C with gentle agitation. After 2 hours, unbound material was collected as flow-through, and the resin was washed 3 times with binding buffer containing 20 mM imidazole, and then once with binding buffer containing 40 mM imidazole, prior to elution with 250 mM imidazole.

2.4.9. Immunoprecipitation

Anti-GFP nanobodies were previously expressed and purified in the lab from BL21 (DE3) *E. coli* by Sarah Whipple. Nanobodies used in this study were LaG16 (Fridy *et al.*, 2014) and the GFP binding protein (GFPBP) commercially available from Chromotek as GFP-trap (Kirchhofer *et al.*, 2010). For use in protein pulldown experiments, nanobodies were coupled to CNBr-Activated Sepharose 4B (GE Healthcare) as follows. The sepharose powder was swelled in 60 bed volumes of 1 mM HCl for 30 minutes. Sepharose was then washed with 25 bed volumes of distilled H_2O followed by 25 bed volumes of coupling buffer (500 mM NaCl, 100 mM $NaHCO_3$, pH 8.3). Nanobodies (10 µg/µL of sepharose) in coupling buffer were allowed to bind to sepharose for 2 hours at room temperature with agitation on a tube roller. Unbound material was collected by centrifugation at 100 g for 10 seconds, and the sepharose washed with 25 bed volumes of coupling buffer. Uncoupled sites within the sepharose were blocked by incubation using 25 bed volumes of 1 M Tris-HCl, pH 8, at room temperature for 2 hours with agitation on a tube roller. Blocked sepharose was washed with 25 bed volumes of Tris-buffered saline (TBS; 20 mM Tris-HCl, pH 7.5, 150 mM NaCl) with 0.05% Tween-20 (TBS-T). It was then subjected to a ‘mock elution’ with 2 rounds of alternating pH, using 25 bed volumes of 100 mM glycine pH 2.7 followed by 25 bed volumes of 50 mM Tris-HCl, pH 8, and finally equilibration with 25 bed volumes of 50 mM Tris-HCl, pH 7.4.

Eight times molar excess of affinity-purified nanobodies covalently attached to CNBr-activated sepharose were used to pull down eGFP (in PBS) or sfGFP-tagged proteins (after ammonium sulphate precipitation and resuspension in binding buffer, see section 2.4.8 for diluent composition). Binding was allowed to take place for 16 hours (overnight) at 4 ^{o}C with rotary mixing. Nanobody-

sepharose was then washed with 25 bed volumes of TBS-T before elution (either once or twice) with 5 bed volumes of 100 mM glycine pH 2.7. Elutions were neutralised with $^1/_{10}$ volume of 1 M Tris-HCl, pH 8.

2.4.10. Micro-dialysis

Resuspended proteins precipitated from cell culture supernatants using ammonium sulphate were dialysed against binding buffer (see 2.4.8 for composition) using a Tube-O-DIALYZER mini dialysis system (Sigma), through a membrane with a MW cut-off of 8 kDa. Samples were dialysed at 4^{o}C against 100x volume of binding buffer with magnetic stirring. Dialysis buffer was changed every 2 hours for a total of 3 times. After the third change, the samples were left at 4^{o}C overnight as a final dialysis step.

2.4.11. Protein quantification

For quantification of total amount of protein in a sample, the BCA method was used, in which proteins can reduce Cu^{2+} to Cu^{1+} in an alkaline solution (the biuret reaction) and result in the formation of a purple colour by bicinchoninic acid. 10 µL of protein samples were mixed with 200 µL bicinchoninic acid (BCA) and 4 µL copper (II) sulphate, and incubated at 37^{o}C or 60^{o}C for 30 minutes before absorbance at 562 nm was measured and compared to BSA protein standards (ranging from 30 to 2000 µg/mL).

2.4.12. SDS-PAGE and immunoblotting

Protein samples were resolved by SDS-PAGE either on Criterion TGX pre-cast gels or on Novex Tris-Glycine gels. $2.5x10^6$ ($5x10^6$ in Chapter 4) cell equivalents or 25 µL medium equivalents were loaded per lane (unless otherwise stated). Gels were run in buffer containing 25 mM Tris, 250 mM glycine and 0.1% SDS. Resolved proteins were electrotransferred onto nitrocellulose membranes (0.45 µL pore size) by immersion in buffer containing 25 mM Tris, 192 mM glycine, 0.02% SDS and 10% methanol, at 40 mA for 14

hours. Nitrocellulose membranes were stained with 0.1% Ponceau-S in 1% acetic acid for visualisation. Membranes were then blocked with 5% semi-skimmed milk in TBS-T.

To detect GFP-tagged proteins, membranes were incubated with monoclonal anti-GFP antibodies (clones 7.1 and 13.1; Roche) at 0.8 µg/mL, for 1 hour at room temperature with agitation (end-to-end rocking). Membranes were then washed 3 times of 5 minutes in TBS-T, before addition of a secondary antibody (polyclonal goat anti-mouse IgG, horseradish peroxidase (HRP) conjugated; Sigma) at 160 ng/mL in TBS-T, for 1 hour at room temperature with agitation. Unbound secondary antibody was washed off 4 times of 5 minutes in TBS-T, before membranes were incubated for 1 minute with a luminol-based chemiluminescent substrate of HRP and H_2O_2 (Western Lightning Plus-ECL, PerkinElmer). Chemiluminescence was detected either by x-ray film or by using a Fusion FX system (Vilber).

For detection of His_6-tagged proteins, milk-blocked membranes were incubated with an HRP conjugated anti-His_6 monoclonal antibody (120 ng/mL, Protein Tech) overnight at 4 ^oC with agitation. As above, after washing 4 times of 5 minutes each in TBS-T, membranes were exposed to Western Lightning Plus-ECL. Chemiluminescence was detected using a Fusion FX system (Vilber).

To strip bound antibodies from membranes for re-probing with a different antibody, membranes were washed 4 times of 5 minutes in TBS followed by incubation in stripping buffer (62.5 mM Tris-HCl, pH 6.8, 2% SDS, 100 mM β-mercaptoethanol) at 60^oC for 2 times of 15 minutes. Membranes were washed again 4 times of 5 minutes before proceeding to immunoblotting as usual, starting from the blocking step.

2.5. Native live fluorescence microscopy

10^6 mid-log phase cells were harvested, washed in PBS, and resuspended in ~5–10 µL of PBS. 2 µL of concentrated cell suspension was transferred to a clean glass slide, and a glass coverslip placed on top with slight pressure. Images were captured using a BX51 microscope with a 100x oil immersion

UPlanApo objective (NA 1.35, Olympus) and a CoolSnap-HQ CCD camera (6.45 µm/pixel, Photometrics). Images were processed using ImageJ (Schneider *et al.*, 2012).

2.6. Flow cytometry

To quantify fluorescence intensity of CExC cell lines producing sfGFP^{op}, ~3x10^6 cells were washed twice with PBS before being resuspended at 5x10^6 cells/µL in PBS with 1 µg/mL propidium iodide as a dead cell marker. Samples were analysed on a MoFo Astrios (Beckman Coulter) flow cytometer using a 561nm laser with a 614/20 filter for detection of propidium iodide fluorescence, and a 488 nm laser with a 513/26 filter for detection of GFP fluorescence. At least 20,000 events were captured per sample. Data analysis was performed using FlowJo v10.6.0 (TreeStar Software, Ashland, Oregon).

2.7. Tables of primers, plasmids and cell lines

Table 2-4 Primers used in this study. Uppercase indicates sequence complimentary to template DNA. Lowercase indicates primer tags, with restriction enzyme sites underlined.

pSiG-HhsfG-Tub derivatives

Fragment	Forward primer (XbaI)	Reverse primer (BamHI)	Construct
ESAG2 ω-4 to Ct	catctagaGCTAGGGTAGATAATGATGG	caggatccCTAAAGCGTACAAAAAAGGG	pSiG-HhsfG-ESAG2_ω-4_Ct-Tub
ESP5 GPISS ω-4 to Ct	catctagaGTTTCTCCTGCAAGAC	caggatccTCACAAAATTAATGAAAACAATACC	pSiG-HhsfG-ESP5_ ω-4_Ct-Tub
ESAG6 GPISS ω-4 to Ct	catctagaCGTGGACCTTTCACGGTA	caggatccTCACAGCACTCCCAACAAT	pSiG-HhsfG-ESAG6_ ω-4_Ct-Tub
GRESAG9 GPISS ω-4 to Ct	catctagaGCAAACAGTACACAAGTGG	caggatccCTAGACCAACAAGCGTTGA	pSiG-HhsfG-GRESAG9_ ω-4_Ct-Tub
ESAG2 ω-3 to Ct	cctctagaAGGGTAGATAATGATGGGC	caggatccCTAAAGCGTACAAAAAAGGG	pSiG-HhsfG-ESAG2_ ω-3_Ct-Tub
ESAG2 ω-2 to Ct	cctctagaGTAGATAATGATGGGCTTACTC	caggatccCTAAAGCGTACAAAAAAGGG	pSiG-HhsfG-ESAG2_ ω-2_Ct-Tub
ESAG2 ω-1 to Ct	cctctagaGATAATGATGGGCTTACTCATT	caggatccCTAAAGCGTACAAAAAAGGG	pSiG-HhsfG-ESAG2_ ω-1_Ct-Tub
ESAG2 ω to Ct	cctctagaAATGATGGGCTTACTCATTTG	caggatccCTAAAGCGTACAAAAAAGGG	pSiG-HhsfG-ESAG2_ ω_Ct-Tub
VSG221 ω-4 to ω-1, *ESAG2* ω to Ct	cctctagagggaaaactggaAATGATGGGCTTACTCATTTG	caggatccCTAAAGCGTACAAAAAAGGG	pSiG-HhsfG-VSG221_ ω-4_ ω-1_ESAG2_ ω_Ct-Tub
ESP5 ω-4 to ω-1, *ESAG2* ω to Ct	cctctagagtttctcctgcaagaAATGATGGGCTTACTCATTTG	caggatccCTAAAGCGTACAAAAAAGGG	pSiG-HhsfG-ESP5_ ω-4_ ω-1_ESAG2_ ω_Ct-Tub
ESAG6 ω-4 to ω-1, *ESAG2* ω to Ct	cctctagacgtggacctttcAATGATGGGCTTACTCATTTG	caggatccCTAAAGCGTACAAAAAAGGG	pSiG-HhsfG-ESAG6_ ω-4_ ω-1_ESAG2_ ω_Ct-Tub
GRESAG9 ω-4 to ω-1, *ESAG2* ω to Ct	cctctagagcaaacagtacaAATGATGGGCTTACTCATTT	caggatccCTAAAGCGTACAAAAAAGGG	pSiG-HhsfG-GRESAG9_ ω-4_ ω-1_ESAG2_ ω_Ct-Tub
AAAG ω-4 to ω-1, *ESAG2* ω to Ct	cctctagagctgctgctggtAATGATGGGCTTACTCATTTG	caggatccCTAAAGCGTACAAAAAAGGG	pSiG-HhsfG-AAAG_ ω-4_ ω-1_ESAG2_ ω_Ct-Tub
A ω-4, *ESAG6* ω-3 to Ct	cctctagagctGGACCTTTCACGGTAGC	caggatccTCACAGCACTCCCAACAAT	pSiG-HhsfG-A_ ω-4_ESAG6_ ω-3_Ct-Tub

AA ω-4 to ω-3, *ESAG6* ω-2 to Ct	cctctagagctgcaCCTTTCACGGTAGCGG	caggatccTCACAGCACTCCCAACAAT	pSiG-HhsfG-AA_ ω-4_ ω-3_ESAG6_ω-2_Ct-Tub
AAA ω-4 to ω-2, *ESAG6* ω-1 to Ct	cctctagagctgcagctTTCACGGTAGCGGGG	caggatccTCACAGCACTCCCAACAAT	pSiG-HhsfG-AAA_ ω-4_ ω-2_ESAG6_ ω-1_Ct-Tub
AAAV ω-4 to ω-1, *ESAG6* ω to Ct	cctctagagctgcagctgtcACGGTAGCGGGGTCC	caggatccTCACAGCACTCCCAACAAT	pSiG-HhsfG-AAAV_ ω-4_ ω-1_ESAG6_ ω_Ct-Tub
ESAG6 ω to C-terminus	cctctagaACGGTAGCGGGGTCC	caggatccTCACAGCACTCCCAACAAT	pSiG-HhsfG_ESAG6_ ω_Ct-Tub
ESAG2 ω-4 to ω-1, *ESAG6* ω to Ct	cctctagagctagggtagatACGGTAGCGGGGTCC	caggatccTCACAGCACTCCCAACAAT	pSiG-HhsfG-ESAG2_ ω-4_ ω-1_ESAG6_ ω_Ct-Tub

LEXSY recombinant proteins

Fragment	Forward primer	Reverse primer	Construct
ISG65FL	gtaccatggTGAAGTATTTGCTGGTATTTGCA	gtagctagcCATTACTACTTTTACGCTAGAAACCC	pLEXSY_IE-sfGO-N-ISG65FL
ISG65ΔC	gtaccatggTGAAGTATTTGCTGGTATTTGCA	gtagctagcTATGAAGAATGCAACGGCC	pLEXSY_IE-sfGO-N-ISG65ΔC
ISG65Cy	gtaccatggATAACCGTGTTCCAGGAGA	gtagctagcTCTCTGATGTCTGCTTTTTACTTC	pLEXSY_IE-sfGO-N-ISG65Cy
ISG65SSec	taatctaGACTTGTTAGTAATTGGCAGTGAG	ataggtacCTCTCTGATGTCTGCTTTTTACTTC	pLEXSY_neo2-ISG65SSec
ESP10FL	gtaccatggTTGTCGGGGTGCTTG	gtagctagcTTCCTTGTTAGGGAACCCTT	pLEXSY_IE-sfGO-N-ESP10FL
ESP10ΔC	gtaccatggTTGTCGGGGTGCTTG	gtagctagcTAGAAGGGATATGATGACAGCG	pLEXSY_IE-sfGO-N-ESP10ΔC
ESP10Cy	gtaccatggGTAAGACTGTCGTGACCAAAAG	gtagctagcGTCAGTGCCGGACTG	pLEXSY_IE-sfGO-N-ESP10 Cy
ESP10Sec	gtaccatggGTAAGACTGTCGTGACCAAAAG	gtaggtaccGTCAGTGCCGGACTG	pLEXSY_neo2-ESP10Sec
ESP13FL	gtaccatggCAGTGAAGGTTTCATTTTCTCTTACA	gtagctagcCTTCGGATTGTTGTCAGCA	pLEXSY_IE-sfGO-N-ESP13FL
ESP13ΔC	gtaccatggCAGTGAAGGTTTCATTTTCTCTTACA	gtagctagcTAAGAATACGCAAATTGCAGAAACG	pLEXSY_IE-sfGO-N-ESP13ΔC
ESP13Cy	gtaccatggGCTTTGCAACACCGAGGGACAATAATTC	gtagctagcCGCGCAGCCCTCCC	pLEXSY_IE-sfGO-N-ESP13 Cy

ESP13Sec	gtaccatggGCTTTGCAACACCGAGGGACAATAATTC	gtaggtaccCGCGCAGCCCTCCC	pLEXSY_neo2-ESP13Sec
ESP14Sec	attctcgagGCTCCTCAAGAAAGCAATGG	aatggtaccATTGTCGCTGTTCATCATCCT	pLEXSY_neo2-ESP14Sec
ESP31FL	catgtcgacATGTCCTCCGTTACCACT	tgtgctagcCTGTTCTTCTTTCTTCCCATAC	pLEXSY_IE-sfGO-N-ESP31FL

pSMC Construction

Fragment	Forward primer	Reverse primer	Construct
TUBB 5' UTR	cttgacggggaaagccatcgaagcttCTCCGCGGAGTAGATCT	gcttgtggccgcgcatactagtATGGTTGTAAAGCGGTAC	pSMC
TUBB 3' UTR	agtgcgagtcgggttcgtaacatatgAGGGCTATTCTGGCAGGGC	ctaaagggaacaaaagctggagaagcttCGGAGCACGCAAACACGC	pSMC
PFR1 Intergenic	ctgacttcgcgttcgcgtaaggatccGAGCCCTCGCTTCTTTGC	gtgggcttgtactcggtcatgctagCGTGCCGTGCAGTGGG	pSMC
PFR2 Intergenic	cgcaagcccggtgcctaaccatggGGCCAGCAGCAGACGATGT	ccgcgcatcctagcttgcatcctaggCGCTGTGGGGGAGAGAGT	pSMC
NEO^R	cacgctagcATGGGATCGGCCATTGA	gccccatggTCAGAAGAACTCGTCAAGAAGG	pSMC

pCEx Construction

Fragment	Forward primer	Reverse primer	Construct
TUBB 5'UTR	aagcttAGCAAAACGAAGACATGCG	catggtagatctATGGTTGTAAAGCGGTACG	pCEx
TEV-*sfGFP*-His$_6$ ORF	gctttacaaccatagatctaccatgCTCGAGGGCGCTAGC	agcaagcaccaaccaccatatgaCGGCCGCTTAGTGGTG	pCEx
TUBA 3'UTR	GTGGTTGGTGCTTGCTTTATCC	CAATGGAATGGTGGAGAGGAGG	pCEx
T7Ter	ctccaccattccattgatgcatGGCCCGTTTGTTATCTATGC	tctctctgctgtgtctagaGGCCCAATCCGGATATAGT	pCEx
T7Ter	gtgccacagttctgcgAGCTCCGGTTCGTCC	gggaaagtaaacaggtaccAATCCGGATATAGTTCTCCTTTCAG	pCEx
10%T7-*GSPS* 3'UTR-*HYG*^{R}-*PGKB* 5'UTR	taatacgtctcactatagggcctaggCTAGAGTTTACCGACAAGACCAG	ggagctCGCAGAACTGTGGCAC	pCEx
pSPR5 backbone	cctaggccctatagtgagacgtattaGAGCTCCAGCTTTTGTTCC	catgtcttcgttttgctaagcttCAGGAATTCGATGGCTTTCC	pCEx

T7 promoter	AATTCTAATACGACTCACTATAGGGATATCA	AGCTTGATATCCCTATAGTGAGTCGTATTAG	pCEx

pCEx derivatives

Fragment	Forward primer	Reverse primer	Construct
PGKB 5' UTR	tctAAGCTTACCGACAAGACCAGA	ggtagatctGCTTGACAAGTGGAAGATAGTTGG	pCEx_PGKB5
PFR2 5' UTR	tctaagcttAAGAACAACATTACTGGAAGCAGCC	ggtagatctCGCTGTGGGGGAGAGA	pCEx_PFR25
L6 5' UTR	tctaagcttCTTGCTTGTTGTGCCTTCTTTTCT	ggtagatctGGTGGCTGGATACGGATAC	pCEx_L65
GSPS 3' UTR	atacatatgAGCAGGCGGAGAAAGAG	ataatgcatGCAGAACTGTGGCACG	pCEx_GSPS3
PGKA 3' UTR	atacatatgATTCTGTATTACGCCGTTTTAAGAG	ataatgcatAGAGAAGTGGTGAGTGGA	pCEx_PGKA3
PFR1 3' UTR	atacatatgGAGCCCTCGCTTCTTTG	ataatgcatAGGGGATAAGAAAACTTCGTCG	pCEx_PFR13
L6 3' UTR	atacatatgGCGATGGTTCGACAGG	ataatgcatTGTGCGAGAGGAAGAATGG	pCEx_L63 (pCExC)
TetO-T7-TetO	AATTCCCTATCAGTGATAGAGATAATACGAC TCACTATAGGGTCCCTATCAGTGATAGAGA	AGCTTCTCTATCACTGATAGGGACCCTATAG TGAGTCGTATTATCTCTATCACTGATAGGG	pCEx_1T
T7-TetO-TetO	ATCTCCCTATCAGTGATAGAGATATCCCTAT CAGTGATAGAGA	AGCTTCTCTATCACTGATAGGGATATCTCTAT CACTGATAGGGAGAT	pCEx_2T
TetO	AATTCTCCCTATCAGTGATAGAG	AATTCTCTATCACTGATAGGGAG	pCEx_3T (pCExT)
GSPS 3'UTR-*HYG*^{R}-*PGKB* 5'UTR	ttgatgcatACCGACAAGACCAGAATAGC	tactctagaGCAGAACTGTGGCACG	pCExP
rDNA upstream targeting fragment	gaattcatttaaatCCTGCGCGTGCTGTGTA	tataagcttTGCATGTGGGTGTGGGTG	pCExP
rDNA downstream targeting fragment	tgctctagaAATTCTATGGACGATTGTGGGC	atagagctcatttaaatCAAGAAGCAGGGGTACAAAAA	pCExP
CfSAP SP 5'	GATCAACCATGGCCAGCAAGATCAGCCGCC TGCTGCT	CGGCCACCAGCAGGGCGGCCAGCAGCAGG CGGCTGATCTTGCTGGCCATGGTT	pCEx_CfSAP-SP
CfSAP SP 3'	GGCCGCCCTGCTGGTGGCCGCCGCCATCA CGGCCGACGCCC	TCGAGGGCGTCGGCCGTGATGGCGG	pCEx_CfSAP-SP
CfGP63 SP 5'	GATCTACCATGCACGCCTTCCAGTGGAAGC GCCACCGCGCCGCCG	GCGGCGGCGGCGCGGTGGCGCTTCCACTG GAAGGCGTGCATGGTA	pCEx_CfGP63-SP

CfGP63 SP 3'	CCGCCCTGTTCTTCCTGTGCCTGCTGCTGG CCACGGCCCTGGCCC	TCGAGGGCCAGGGCCGTGGCCAGCAGCAG GCACAGGAAGAACAGG	pCEx_CfGP63-SP

Diagnostic PCR

Fragment	Forward primer	Reverse primer	Construct
pSMC integrated	GGTGCCATCGATGTTC	AGCTCGATGTCAGAGAAC	
TUBB ORF 3'-3'UTR (SMC control PCR)	GGTGCCATCGATGTTC	CCCAAAGATGAAGTTGTCC	
pCEx circular	GGAGAACTATATCCGGATTGG	GGAGAACTATATCCGGATTGG	
pCEx integrated	GGAGAACTATATCCGGATTGG	ATATGCCCACAATCGTCC	

CExSy recombinant proteins

Fragment	Forward primer	Reverse primer	Construct
ESP10Sec (endogenous SP)	atagatctaccATGCTTGTCGGGGTGCTTG	gtagctagcGTCAGTGCCGGACTG	pCExC-ESP10Sec
ESP10Sec (no SP)	gtactcgagGTCGGAAGCATTGCAAAGTAC	gtagctagcGTCAGTGCCGGACTG	pCExS-ESP10Sec
ESP10FL^{op} modification	ATCAAGCTTAGCAAAACGAAGACATGCGGC	TACCCGGGCGAGCCACAGCACTAGGGACA ATCATACTAGCATTGTCGCCCATGCCGTAG C	pCExS-ESP10FLO
ESP10Sec^{op}	ATCAAGCTTAGCAAAACGAAGACATGCGGC	attgctagcGTCCGTGCCGCTCTG	pCExS-ESP10SecO
ESP13Sec	gtagtcgacAAAAATTGCTTTGCAACACCGAGGGAC	gtagctagcCGCGCAGCCCTCCC	pCExS-ESP13Sec
ESP14Sec	attctcgagGCTCCTCAAGAAAGCAATGG	aatgctagcATTGTCGCTGTTCATCATCCT	pCExS-ESP14Sec
ISG65Sec	gtagtcgacTTGTTAGTAATTGGCAGTGAGG	gtagctagcTCTCTGATGTCTGCTTTTTACTTC	pCExS-ISG65Sec
HpHbRSec (endogenous SP)	caagatctATGGAGAAACCGTCTTGCAGG	tgtgctagcAACCACGTCAACGGGC	pCExC-HpHbRSec
HpHbRSec (no SP)	attctcgagGCTGAGGGTTTAAAAACCAAAGACG	tgtgctagcAACCACGTCAACGGGC	pCExS-HpHbRSec

Table 2-5 Cell lines generated in this study

T. brucei

Plasmid	Parental cell line	Resultant cell line	Selection marker
pSiG-HhsfG_ESAG2_ω-4_Ct-Tub	S16	GG-ESAG2	HYG^R
pSiG-HhsfG_ESP5_ ω-4_Ct-Tub	S16	GG-ESAG5	HYG^R
pSiG-HhsfG_ESAG6_ ω-4_Ct-Tub	S16	GG-ESAG6	HYG^R
pSiG-HhsfG_GRESAG9_ ω-4_Ct-Tub	S16	GG-GRESAG9	HYG^R
pSiG-HhsfG_ESAG2_ ω-3_Ct-Tub	S16	GG-ESAG2_ ω-3_Ct	HYG^R
pSiG-HhsfG_ESAG2_ ω-2_Ct-Tub	S16	GG-ESAG2_ ω-2_Ct	HYG^R
pSiG-HhsfG_ESAG2_ ω-1_Ct-Tub	S16	GG-ESAG2_ ω-1_Ct	HYG^R
pSiG-HhsfG_ESAG2_ ω_Ct-Tub	S16	GG-ESAG2_ ω_Ct	HYG^R
pSiG-HhsfG_VSG221_ ω-4_ ω-1_ESAG2_ ω_Ct-Tub	S16	GG-ESAG2_ VSG221ω-4_ ω-1	HYG^R
pSiG-HhsfG_ESP5_ ω-4_ ω-1_ESAG2_ ω_Ct-Tub	S16	GG-ESAG2_ ESP5ω-4_ ω-1	HYG^R
pSiG-HhsfG_ESAG6_ ω-4_ ω-1_ESAG2_ ω_Ct-Tub	S16	GG-ESAG2_ ESAG6ω-4_ ω-1	HYG^R
pSiG-HhsfG_GRESAG9_ ω-4_ ω-1_ESAG2_ ω_Ct-Tub	S16	GG-ESAG2_ GRESAG9ω-4_ ω-1	HYG^R
pSiG-HhsfG_AAAG_ ω-4_ ω-1_ESAG2_ ω_Ct-Tub	S16	GG-ESAG2_ AAAGω-4_ ω-1	HYG^R
pSiG-HhsfG_A_ ω-4_ESAG6_ ω-3_Ct-Tub	S16	GG-ESAG6_ A_ω-3_Ct	HYG^R
pSiG-HhsfG_AA_ ω-4_ ω-3_ESAG6_ ω-2_Ct-Tub	S16	GG-ESAG6_AA_ ω-2_Ct	HYG^R
pSiG-HhsfG_AAA_ ω-4_ ω-2_ESAG6_ ω-1_Ct-Tub	S16	GG-ESAG6_ AAA_ω-1_Ct	HYG^R
pSiG-HhsfG_AAAV_ ω-4_ ω-1_ESAG6_ ω_Ct-Tub	S16	GG-ESAG6_ AAAV_ω_Ct	HYG^R
pSiG-HhsfG_ESAG6_ ω_Ct-Tub	S16	GG-ESAG6_ ω_Ct	HYG^R
pSiG-HhsfG_ESAG2_ ω-4_ ω-1_ESAG6_ ω_Ct-Tub	S16	GG-ESAG6_ ESAG2ω-4_ ω-1	HYG^R

L. tarentolae

Plasmid	Parental cell line	Resultant cell line	Selection marker
pLEXSY_IE-egfp-red-neo4	T7TR	eGFP-DsRed-Lt	NEO^R
pLEXSY_IE-sfG-N	T7TR	sfGFP-Lt	NEO^R
pLEXSY_IE-sfGO-N	T7TR	$sfGFP^{op}$-Lt	NEO^R
pLEXSY_IE-sfGO-ISG65FL-N	T7TR	rISG65FL-Lt	NEO^R
pLEXSY_IE-sfGO-ISG65ΔC-N	T7TR	rISG65ΔC-Lt	NEO^R
pLEXSY_IE-sfGO-ISG65Cy-N	T7TR	rISG65Cy-Lt	NEO^R
pLEXSY_IE-sfGO-ESP10FL-N	T7TR	rESP10FL-Lt	NEO^R

pLEXSY_IE-sfGO-ESP10ΔCL-N	T7TR	rESP10ΔC-LT	NEO^R
pLEXSY_IE-sfGO-ESP10CyL-N	T7TR	rESP10ΔCy-LT	NEO^R
pLEXSY_IE-sfGO-ESP13FL-N	T7TR	rESP13FL-Lt	NEO^R
pLEXSY_IE-sfGO-ESP13ΔCL-N	T7TR	rESP13ΔC-LT	NEO^R
pLEXSY_IE-sfGO-ESP13CyL-N	T7TR	rESP13ΔCy-LT	NEO^R
pLEXSY_IE-sfGO-ESP31FL-N	T7TR	rESP31FL-Lt	NEO^R
pLEXSY_neo2_LiTat1.3Sec	T7TR	rLiTat1.3Sec-Lt	NEO^R
pLEXSY_neo2_TbgISG65Sec	T7TR	rTbgISG65Sec-Lt	NEO^R
pLEXSY_neo2_TbbISG65Sec	T7TR	rTbbISG65Sec-Lt	NEO^R
pLEXSY_neo2_ESP10Sec	T7TR	rESP10Sec-Lt	NEO^R
pLEXSY_neo2_ESP13Sec	T7TR	rESP13Sec-Lt	NEO^R
pLEXSY_neo2_ESP14Sec	T7TR	rESP14Sec-Lt	NEO^R

C. fasciculata

Plasmid	Parental cell line	Resultant cell line	Selection marker
pSMC	*C. fasciculata*	SMC	NEO^R
pCEx	SMC	CEx	HYG^R
pCEx_PGKB5	SMC	CEx_PGKB5	HYG^R
pCEx_PFR25	SMC	CEx_PFR25	HYG^R
pCEx_L65	SMC	CEx_L65	HYG^R
pCEx_GSPS3	SMC	CEx_GSPS3	HYG^R
pCEx_PGKA3	SMC	CEx_PGKA3	HYG^R
pCEx_PFR13	SMC	CEx_PFR13	HYG^R
pCExC	SMC	CExC	HYG^R
pCEx_1T	SMC	CEx_1T	HYG^R
pCEx_2T	SMC	CEx_2T	HYG^R
pCExT	SMC	CExT	HYG^R
pCExP	SMC	CExP	HYG^R
pCExS	SMC	CExS	HYG^R
pCExC-ISG65FL	SMC	rISG65FL-Cf	HYG^R
pCExC-ISG65ΔC	SMC	rISG65ΔC-Cf	HYG^R

pCExC-ESP10FL	SMC	rESP10FL-Cf	HYG^R
pCExC-ESP10ΔC	SMC	rESP10ΔC-Cf	HYG^R
pCExC-ESP13FL	SMC	rESP13FL-Cf	HYG^R
pCExC-ESP13ΔC	SMC	rESP13ΔC-Cf	HYG^R
pCExC_HpHbRSec	SMC	rHpHbRSec-Cf (TbSP)	HYG^R
pCExS_HpHbRSec	SMC	rHpHbRSec-Cf (LmSAPSP)	HYG^R
pCExC_ESP10Sec	SMC	rESP10Sec-Cf (TbSP)	HYG^R
pCExS_ESP10Sec	SMC	rESP10Sec-Cf (LmSAPSP)	HYG^R
pCEx_CfSAP-SP-ESP10Sec	SMC	rESP10Sec-Cf (CfSAPSP)	HYG^R
pCEx_CfGP63-SP-ESP10Sec	SMC	rESP10Sec-Cf (CfGP63SP)	HYG^R
pCExS_ESP10Sec^{op}	SMC	rESP10Sec^{op}-Cf (LmSAPSP)	HYG^R
pCExS_ESP13Sec	SMC	rESP13Sec-Cf	HYG^R
pCExS_ESP13Sec^{op}	SMC	rESP13Sec^{op}-Cf	HYG^R
pCExS_ESP14Sec	SMC	rESP14Sec-Cf	HYG^R
pCExS_ESP14Sec^{op}	SMC	rESP14Sec^{op}-Cf	HYG^R
pCExS_TbbISG65Sec	SMC	rTbbISG65-sfGFP^{op}-Cf	HYG^R
pCExS2_TbbISG65Sec	SMC	rTbbISG65-Cf	HYG^R
pCExS2_TbgISG65Sec	SMC	rTbgISG65-Cf	HYG^R
pCExS_ESP10Sec	SMC	rESP10Sec-Cf (no GFP)	HYG^R
pCExS2_ESP13Sec	SMC	rESP13Sec-Cf (no GFP)	HYG^R
pCExS2_ESP14Sec	SMC	rESP14Sec-Cf (no GFP)	HYG^R

Chapter 3. GPI-anchored protein sorting in *Trypanosoma brucei*

3.1. Introduction

The studied GPIAPs of *T. brucei* play varied, important roles. To carry out these roles, distinct surface localisations appear to be important. BSF immune system evasion protein VSG, and PCF protease-resistance protein procyclin are distributed across the whole of their respective cell surfaces to protect the two forms in their disparate environments (Gadelha *et al.*, 2011). Two important BSF receptors – HpHbR and TfR – are constrained within the flagellar pocket, shielding them from the immune system and allowing rapid endocytosis (Schwartz *et al.*, 2005; Vanhollebeke *et al.*, 2008). SRA (present specifically in *T. brucei. rhodesiense*) localises to endosomes and the lysosome, where it protects BSF cells from ApoL1 mediated cell lysis (Bart *et al.*, 2015). It is largely unknown how these different proteins are directed to their specific membrane sub compartments.

N-glycosylation of membrane associated proteins has previously been suggested to play a role in cellular sorting in *T. brucei*. An early study had shown proteins associated with linear poly-N-acetyllactosamine (pNAL) to exclusively associate with the FP and endocytic pathway (Nolan *et al.*, 1999). This was highlighted by their specific association with the sugar-binding protein tomato lectin (TL). As a sole sorting mechanism, N-glycosylation has been more recently challenged, as mapping of the N-glycan structures of TfR revealed it to contain no pNAL but only paucimannose and oligomannose (although there is evidence TfR binds to a protein that does contain pNAL) (Mehlert *et al.*, 2012). Additionally, work in our lab has shown a mix of N-glycan types on proteins of distinct surface membrane localisation (Whipple, 2017).

Two other hypotheses were put forward to explain surface membrane protein sorting: one based on protein levels as a saturable mechanism in the flagellar pocket (Engstler *et al.*, 2005; Mussmann *et al.*, 2003, 2004), and the other on GPI valence (Schwartz *et al.*, 2005; Tiengwe *et al.*, 2017; Triggs and

Bangs, 2003). The FP saturation hypothesis is based on two examples – TbMBAP1 and TfR. TbMBAP1 localises to the endosomal system, but tetracycline-regulated overexpression led to its appearance in the flagellar pocket (upon 3-4 fold overexpression), and on the cell body and flagellar membranes (upon 10-20 fold overexpression) (Engstler *et al.*, 2005). Meanwhile, TfR has been shown to be upregulated in response to iron starvation, leading to its detection across the cell body membrane, again linking an increase in protein levels with FP escape (Mussmann *et al.*, 2003, 2004). However, past these two proteins, no further evidence has been reported to support the pocket saturation hypothesis.

With respect to GPIAPs, the valence hypothesis has gained traction within recent years. It suggests that the number of GPI anchors on a protein or protein complex dictates retention or escape from the flagellar pocket. This was based upon the fact that VSG, which localises across the entire plasma membrane, is a homodimer and, hence, has two GPI anchors, whilst FP localised TfR is a heterodimer of ESAG6/7 but only ESAG6 is GPI-anchored (Triggs and Bangs, 2003). Yet, as mentioned above, TfR can be detected on the cell body membrane as a result of iron starvation (Mussmann *et al.*, 2003, 2004). However, cell body localised TfR has subsequently been shown to be a non-functional ESAG6 homodimer (having two GPI anchors), further supporting the valence hypothesis (Schwartz *et al.*, 2005; Tiengwe *et al.*, 2017). More recently, Tiengwe *et al* reported that attaching a GPI anchor onto ESAG7 through endogenous-locus tagging of ESAG7 with the ESAG6 GPISS, led to the detection of functional ESAG6/7 heterodimers across the cell surface. Whilst the study provides fairly compelling evidence for the valence hypothesis, it is still only based on VSG and TfR and may not hold for other GPIAPs. For example, the HpHbR is a GPIAP monomer (thus, one anchor), which binds to its ligand cooperatively with a second HpHbR molecule (leading to a complex with two anchors) (Lane-Serff *et al.*, 2014; Stødkilde *et al.*, 2014). Whilst *T. brucei* BSF endocytosis is rapid (Engstler *et al.*, 2004), it would seem counter intuitive to have this cooperative ligand binding if it would then risk escape of this complex out of the FP.

In other organisms, sorting of GPIAPs can be dictated by the attachment of different GPISSs (Miyagawa-Yamaguchi *et al.*, 2015, 2014; Paladino *et al.*, 2008). In Madin-Darby canine kidney (MDCK) cells, GPISSs derived from differentially localised proteins (the apically sorted folate receptor (FR) and the basolaterally sorted prion protein (PrP)) drove GFP to the endogenous locations of these respective proteins (Paladino *et al.*, 2008). This was shown to be linked to the oligomerisation state of the GFP. Similarly in HeLa cells, attachment of two different GPISSs (from decay accelerating factor (DAF) and Thy-I) to two florescent proteins (red and green respectively) led to partitioning into distinct membrane microdomains (Miyagawa-Yamaguchi *et al.*, 2014). In addition, fusion of these two GPISSs to horse radish peroxidase (HRP) led to fluorescein labelling of distinct sets of membrane proteins. These sets were subsequently shown to include the respective proteins from which the GPISSs were derived (Miyagawa-Yamaguchi *et al.*, 2015). Interestingly, the HRP fusions with different GPISSs were also shown to have different N-glycan types attached, again matching the respective GPISS proteins. This was shown likely to be the result, rather than the cause of the differential clustering of these proteins, as inhibition of Golgi N-glycan modifying enzymes did not disrupt the respective clustering patterns (Miyagawa-Yamaguchi *et al.*, 2014).

Coupled with the fact that mammalian cells can modify their GPI anchors to have differences in their lipids and sugar side-chains (Ferguson *et al.*, 2017; Kinoshita and Fujita, 2016), the above results suggest that different GPISSs can dictate the attachment/modification of distinct GPI anchors with individual clustering properties. Indeed, PrP and Thy-I have been shown to contain anchors that differ in their sugar moiety, with the PrP anchor containing galactose and sialic acid residues that are absent in the anchor of Thy-I (Homans *et al.*, 1988; Stahl *et al.*, 1992). This sialic acid residue is vital in the synaptic sorting of PrP in neuronal cells (Bate *et al.*, 2016). Further to this, it has been reported that PrP with the Thy-I GPISS expressed in mouse neuronal cells did not contain a sialic acid residue, whilst PrP expressed with its native GPISS did contain sialic acid (Puig *et al.*, 2019). This latter study reports the first direct evidence that different GPISSs on the same protein can result in different GPI anchor modifications.

It is somewhat surprising that GPISSs appears to have such a specific control over GPI anchor structure, as modification of the GPI anchor occurs in the Golgi apparatus, after the GPISS has been removed and the unmodified anchor attached to the protein in the ER (Ferguson *et al.*, 2017; Kinoshita and Fujita, 2016). The unmodified anchor attached to the protein in the ER can vary to some extent, containing either 3 or 4 mannose residues and 2 or 3 ethanolamine-phosphate groups. However, the PrP and Thy-I anchors both contain 2 ethanolamine residues (in the same positions) and can be found with 3 or 4 mannose residues, implying that neither of these modifications have a bearing on anchor modification in the Golgi (Homans *et al.*, 1988; Stahl *et al.*, 1992). As such, the mechanism by which GPISSs affect GPI anchor modification is still unknown.

Similarly to mammalian cells, the GPI anchor of *T. brucei* BSF is not known to be further modified prior to ER exit (Güther and Ferguson, 1995; Mehlert and Ferguson, 2007). In fact, there is even less variation in the unmodified anchor structure initially attached to proteins than in mammalian cells (see Figure 1-8 and 1-9 in Chapter1). Nevertheless, the anchor can be modified post-attachment by addition of galactose to the mannose residues in the Golgi (Ferguson *et al.*, 1988; A Mehlert *et al.*, 1998; Mehlert and Ferguson, 2007; Zitzmann *et al.*, 2000). This is thought to be regulated by the accessibility of enzymes to the GPI anchor based upon the attached protein structure rather than the GPISS (A Mehlert *et al.*, 1998; Zitzmann *et al.*, 2000). This is in part because there is large variation in the level of galactosylation of anchors attached to different classes of VSGs (defined according to homologies in their C-terminal domains), varying from none at all, up to at least six galactose residues (A Mehlert *et al.*, 1998). TfR has also been shown to have six hexose sugars attached to the core GPI structure (Mehlert and Ferguson, 2007), implying that the anchor may not play a role in the differential sorting of these proteins. However, these sugars have not specifically been shown to be galactose. Additionally, as mammalian PrP can be differentially sorted simply through the difference of a single residue in its anchor (Bate *et al.*, 2016; Puig *et al.*, 2019), it is not implausible that subtle differences in the structure of attached side chains of TfR could play a role in its sorting.

The multifaceted evidence above shows that sorting of GPIAPs is complex, with multiple features potentially playing a role. Whilst N-glycosylation has largely been disregarded as a sorting mechanism for these proteins in *T. brucei*, it may still play an indirect role (Mehlert *et al.*, 2012). Compelling evidence exists towards the valence hypothesis (Tiengwe *et al.*, 2017), but this has not yet been definitively proven, and if correct, may not dictate sorting for all GPIAPs. There is mounting evidence in mammalian cells that the GPISS plays a large role in GPIAP sorting through dictating anchor modification. Yet, this mechanism has not been tested in *T. brucei*. This chapter looks to elucidate whether GPISS affects GPIAP sorting in *T. brucei* BSF through the use of an exogenous reporter protein – 'superfolder' GFP (sfGFP). Utilisation of this protein and another monomeric fluorescent protein – mScarlet-I (mSca-I) – also allowed investigation of the valence hypothesis.

3.2. Results

3.2.1. *T. brucei* GPIAP candidates

Prior studies investigating GPIAP sorting in *T. brucei* have been hindered by the lack of GPIAPs of known localisation, relying on the same two examples – VSG and TfR (Mehlert *et al.*, 2012; Mussmann *et al.*, 2004, 2003; Schwartz *et al.*, 2005; Tiengwe *et al.*, 2017). The recent surfeome study has rectified this issue, providing a larger set of GPIAPs with validated localisation across different parts of the secretory pathway/cell surface (Gadelha *et al.*, 2015). A summary of the GPIAPs of known localisation are shown in Figure 3-1.

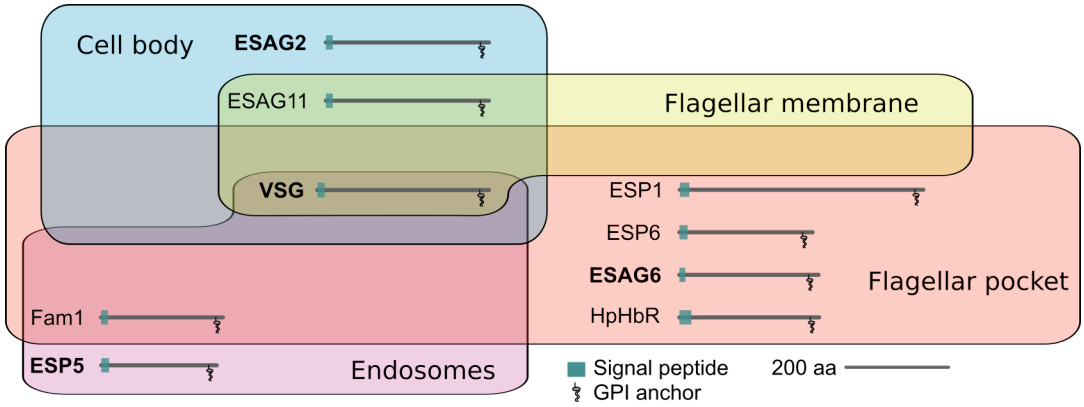


Figure 3-1 GPI-anchored proteins localise to different membrane domains on the surface of BSF *T. brucei*. Venn diagram indicating the localisation of GPI-anchored proteins in *T. brucei* BSF – the cell body, flagellar membrane, flagellar pocket, endosomal membranes, or a combination of these domains. Proteins in bold are representatives of different GPIAP localisations selected for this study. Figure modified from Gadelha *et al*, 2015.

To study GPIAP sorting mechanisms, test candidates were selected as representatives of different GPIAP localisations – VSG, ESAG2, ESP5 and ESAG6 (bold proteins in Figure 3-1). GRESAG9 was also selected. GRESAG9 is predicted to be GPI-anchored but localises to the ER when tagged at its endogenous locus either N-terminally or C-terminally (Chamberlain and Gadelha, personal communication). To define the GPISSs of this set of proteins, the algorithm PredGPI was used to predict their respective ω-sites (Pierleoni *et al.*, 2008). The set of proteins, their gene IDs, localisations and predicted GPISSs are summarised in Table 3-1. Alignment of GPISSs from the selected GPIAPs revealed little conservation other than in the C-terminal hydrophobic region.

To investigate whether GPISSs alone could dictate surface sorting of GPIAPs, these GPISSs were fused to an exogenous reporter protein – sfGFP. To ensure correct anchoring, the 4 amino acids upstream of the predicted ω-sites were included, reducing the risk of no anchoring due to an incorrectly predicted ω-site or interference by the sfGFP, and preventing a change in ω-site prediction as a result of the C-terminal sfGFP amino acids.

Table 3-1 GPI-anchored proteins of different cellular locations used in this study. Cellular locations are indicated (+). GPI signal peptides shown from ω-4 to C terminal end, with the ω-site indicated in red (predicted by PredGPI) and 4 upstream amino acids in bold.

PROTEIN		CELLULAR LOCATION					GPI SIGNAL SEQUENCE PREDICTION	
NAME	**Gene ID**	**Endosomes**	**Flagellar Pocket**	**Flagellar Membrane**	**Cell body**	**ER**	**ω-4 to Ct**	**sfGFP-GPI fusion nickname**
VSG221	Tb427.BES40.22	+	+	+	+		**G K T G** N T N T T G S S N S F V I S K T P L W L A V L L F	GG-VSG221
ESAG2	Tb427.BES40.18				+		**A R V D** N D G L T H L N I A T G V V M L L V L S L F C T L	GG-ESAG2
ESP5	Tb927.5.291b	+					**S P A R** P A G E Q K S Y S Q V I S I P L R V L F S L I L	GG-ESP5
ESAG6	Tb427.BES40.3	+	+				**R G P F** T V A G S N T V A V H L S L F T A A L C C S A L L L G V L	GG-ESAG6
GRESAG9	Tb927.5.120					+	**A N S T** Q V G R T T A H K R H V M L I T A T L S F I C V Q R L L V	GG-GRESAG9

sfGFP fusions were constructed through modification of the endogenous locus tagging vector pSiG (Figure 3-2A) (Gadelha *et al.*, 2015). This vector contains the *sfGFP* ORF with the VSG221 N-terminal signal peptide and C-terminal GPISS. It also contains restriction sites for exchange of the C-terminal GPISS, allowing insertion of other GPISSs from the selected GPIAP candidates. Candidate GPISSs depicted in Table 3-1 were amplified from *T. brucei* gDNA for insertion into pSiG, creating the sfGFP derivatives shown in Figure 3-2B.

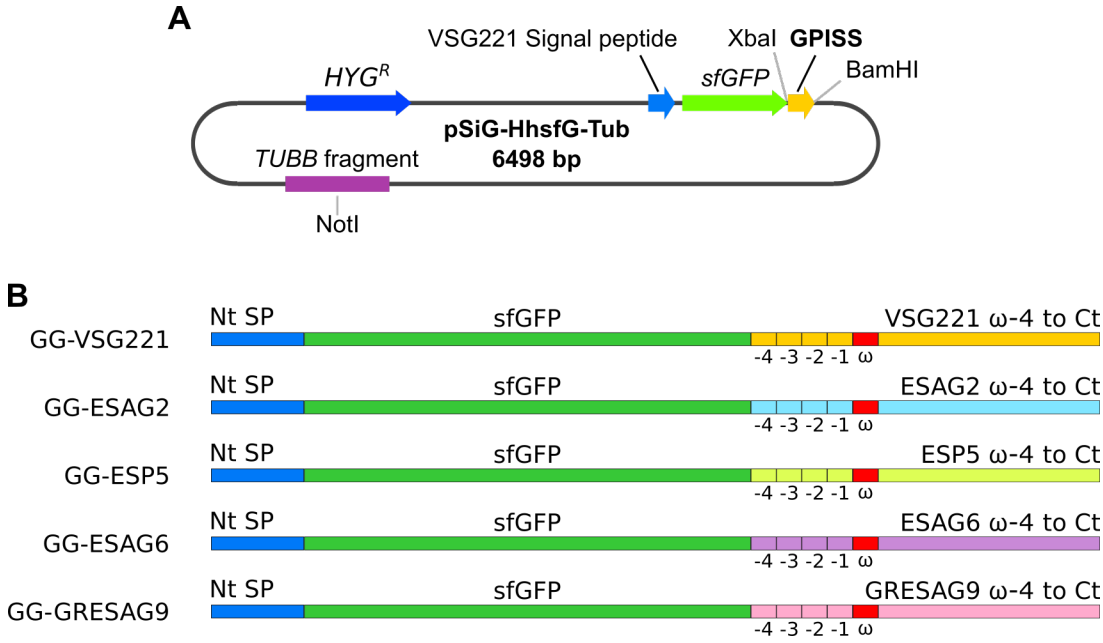


Figure 3-2. The pSiG endogenous locus tagging vector was used as a tool for investigating GPI-anchored protein sorting. A The pSiG vector contains an open reading frame consisting of the VSG221 N-terminal signal peptide, the *sfGFP* ORF and the VSG221 C-terminal GPI signal sequence. The signal peptide directs sfGFP to the secretory pathway and the GPISS allows the newly translocated protein to become GPI-anchored in the lumen of the ER. The vector contains a *TUBB* fragment for integration into the *T. brucei* tubulin locus, enabling constitutive readthrough expression; and Hygromycin B resistance selectable marker. Restriction sites BamHI and XbaI can be used for swapping the GPI signal peptide. The NotI site is used for plasmid linearization prior to transfection. **B** Schematics of sfGFP-GPI fusions. Each fusion consists of the VSG221 N-terminal signal peptide (Nt SP), sfGFP and a C-terminal GPI signal sequence derived from VSG221, ESAG2, ESP5, ESAG6 or GRESAG9. C-terminal GPI signal sequences amplified from *T. brucei* gDNA and inserted into pSIG. Cartoon not to scale.

3.2.2. The sfGFP-GPI fusions produce proteins of different sizes

pSiG contains a *TUBB* fragment for insertion of the vector DNA into the *T. brucei* genome, driving *sfGFP* transcription by constitutive PolII readthrough. Thus, upon construction of the pSiG derivatives, the vectors were linearised with NotI and transfected into *T. brucei* BSF S16 cells (Wirtz *et al.*, 1999). Transfectants were selected using 5 µg/mL Hygromycin B, and 3 independent clones from each transfection were screened for sfGFP expression by anti-GFP immunoblotting. sfGFP levels were consistent between different clones derived from each individual transfection (data not shown). Thus, a representative clone from each transfection was taken forward for further analysis.

Figure 3-3 shows an anti-GFP immunoblot of the representative clones. The predicted MW for all fusions is ~34 kDa (this includes the GPI anchor). Interestingly, the different fusions migrate at different rates to each other, suggesting distinct post-translational modifications. GG-VSG221 and GG-GRESAG9 appear ~3 kDa heavier than predicted. GG-ESP5 and GG-ESAG6 appear at approximately the predicted MW but also present an additional, less intense band at ~45 kDa. GG-ESAG2 shows only one band matching the higher MW (45 kDa). These 45 kDa bands will be discussed further in 3.2.5. Bands also appear below the predicted MW, likely as a result of proteolysis.

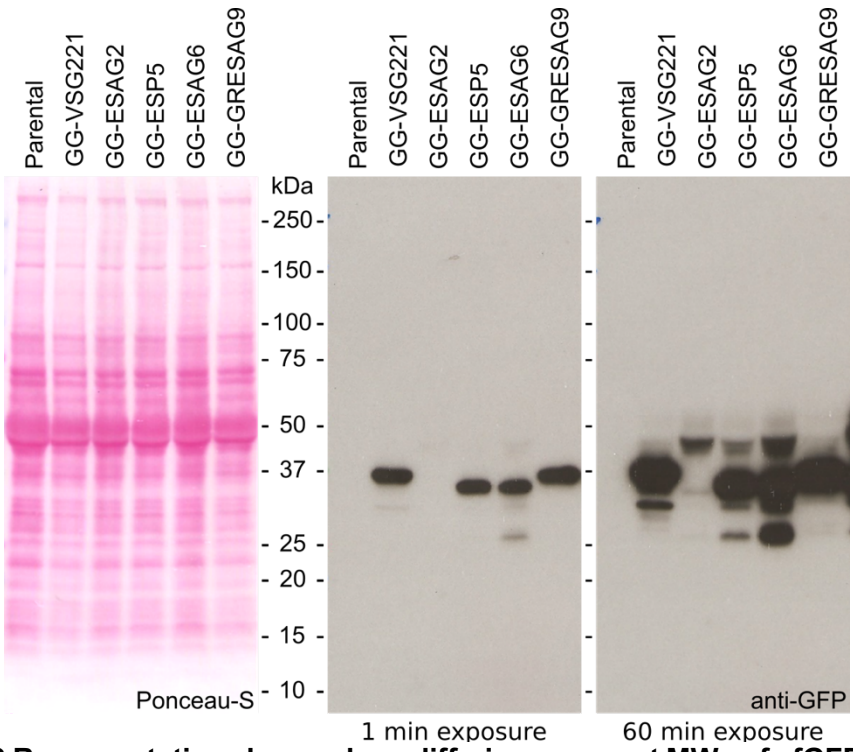


Figure 3-3 Representative clones show differing apparent MWs of sfGFP fusions. Immunoblot of whole cell lysates from sfGFP fusion cell lines. $2x10^6$ cell equivalents were loaded on each lane. Fusion proteins were detected with anti-GFP antibodies. Ponceau-S stained nitrocellulose membrane shown as a loading control. Predicted MW for sfGFP fusions is approximately 34 kDa. GG-VSG221, GG-ESAG2 and GG-GRESAG9 migrate slower than predicted. GG-ESP5 and GG-ESAG6 also show a higher MW band of ~45kDa on longer exposure of X-ray film to chemiluminescence. GG-ESAG2 only appears on longer exposure, showing the higher MW band and a very weak band at the predicted MW.

Along with further modifications to the GPI anchor, GPIAPs can be N-glycosylated as they progress through the secretory pathway. This type of modification is large enough to cause a visible shift in MW by Western blot. To analyse whether the observed differences in size of the sfGFP-GPI fusions could have been caused by N-glycosylation, NetNGlyc (www.cbs.dtu.dk/services/NetNGlyc) was used to predict whether the GPISSs (from ω-4 to C-terminus) contained putative N-glycosylation sites. Both GG-VSG221 and GG-GRESAG9 were predicted to contain N-glycans with high probability (>73%; Figure 3-4A), potentially explaining the ~3 kDa shifts observed in Figure 3-3.

N-glycans can be removed from glycoproteins using the glycosidase PNGase-F. Thus, to confirm the glycosylation states of the sfGFP-GPI fusions, this enzyme was utilised. Whole cell lysates were prepared and split into two

samples, one with added PNGase-F and one without. These samples were then resolved by SDS-PAGE and electro-transferred to a nitrocellulose membrane for analysis by anti-GFP immunoblot (Figure 3-4B). Both GG-VSG221 and GG-GRESAG9 shifted to the predicted MW of ~34 kDa upon treatment with PNGase-F, confirming the presence of N-glycosylation. None of the other sfGFP fusions showed a change in MW upon PNGase-F treatment. Additional lower MW bands appeared in all lanes, likely as a result of proteolysis.

Interestingly, the predicted N-glycosylation site for GG-VSG221 is downstream of the predicted ω-site in the amino acid sequence (Figure 3-4A). This would imply that the ω-site prediction is incorrect, otherwise the N-glycosylation site would have been removed during GPI-attachment and the resultant GG-VSG221 protein would not have been N-glycosylated. An alternative ω-site could be the second Ser residue within the sequence GSSNS as all known VSG ω-sites are Ser, Asp or Asn and commonly have either 17 or 23 amino acids removed upon GPI-anchoring (Böhme and Cross, 2002; Wang *et al.*, 2003).

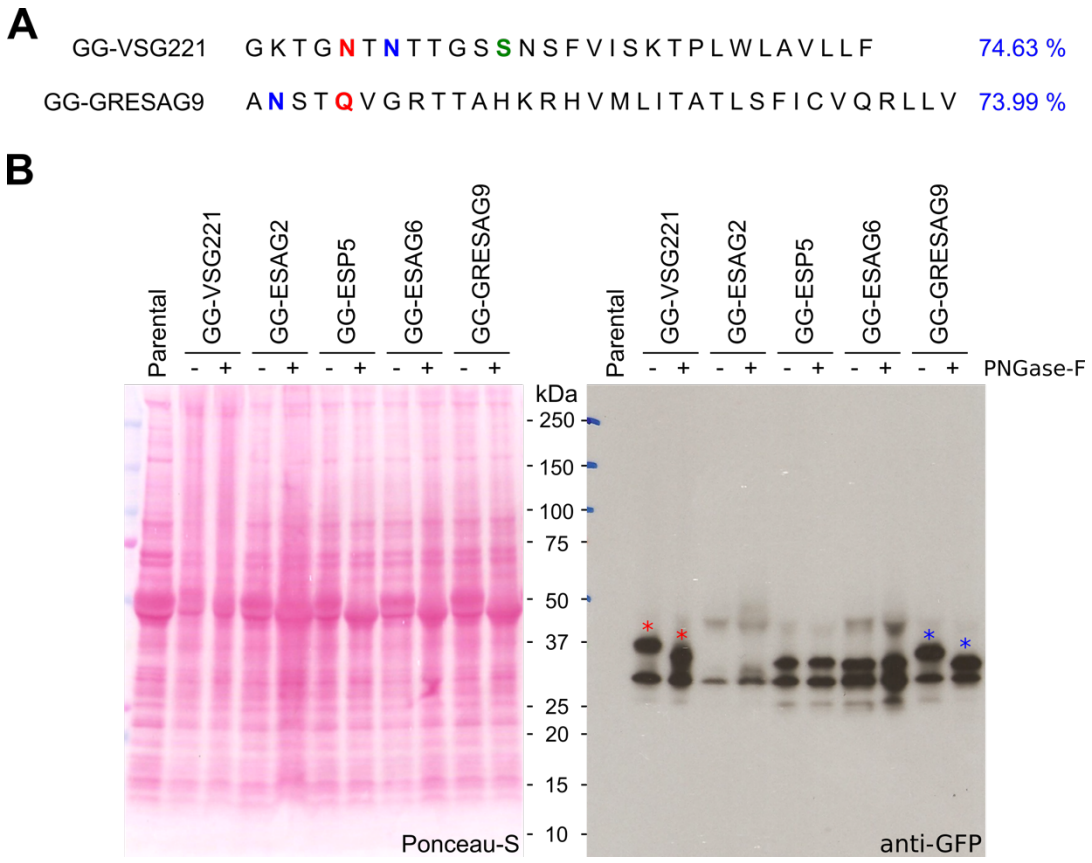


Figure 3-4 PNGase-F treatment reveals GG-VSG221 and GG-GRESAG9 to be N-glycosylated. A Amino acid sequences from ω-4 to C-terminus of GG-VSG221 and GG-GRESAG9. Predicted ω-sites shown in red. Predicted N-glycosylation sites (and prediction probabilities) shown in blue. Likely alternative GG-VSG221 ω-site, based on presence of N-glycosylation and on previous studies, shown in green (Böhme and Cross, 2002; Wang *et al.*, 2003). **B** Ponceau-S stained membrane and anti-GFP immunoblot of whole cell lysates, treated (+) or untreated (-) with the glycosidase PNGase-F. GG-VSG221 (*) and GG-GRESAG9 (*) show a shift in molecular weight, indicating presence of N-glycosylation.

3.2.3. sfGFP-GPI fusions are GPI-anchored

T. brucei has a GPI-specific phospholipase C (GPI-PLC) that localises to the flagellum membrane and has been shown to act as a virulence factor (Sunter *et al.*, 2013; Webb *et al.*, 1997). Upon cell death, GPI-PLC releases GPIAPs from the plasma membrane of the parasite by cleaving the lipids off of GPI anchors. The mechanism is thought to aid in VSG switching, distracting the immune system with epitopes of the released VSG whilst other cells begin to express novel VSGs. This release can be triggered by hypotonic lysis of cells at 37°C (but not at 0°C as GPI-PLC is inactive at this temperature). This feature is commonly used as an assay to assess GPI-anchoring of proteins in *T. brucei*.

Cells are lysed in water, supplemented with protease inhibitors, at either 37°C or 0°C before separation into supernatant and pellet fractions (Figure 3-5A). At 0°C, GPI-PLC is not active, and GPIAPs remain in the pellet fraction. At 37°C, GPI-PLC is active and the resultant cleaved GPIAPs are present in the supernatant fraction. This hypotonic lysis assay was used to test whether the sfGFP-GPI fusions were indeed GPI-anchored.

Figure 3-5B shows the results of the hypotonic lysis assay. The GFP-tagged multi-pass transmembrane protein ESAG10 was included in the experiment as a non-GPI negative control. The VSG band (~50 kDa) visible on the Ponceau-S stained membrane functions as an internal GPIAP positive control. All of the sfGFP-GPI fusions moved from the insoluble pellet fraction at 0°C to the supernatant fraction at 37°C, showing them to be GPI-anchored.

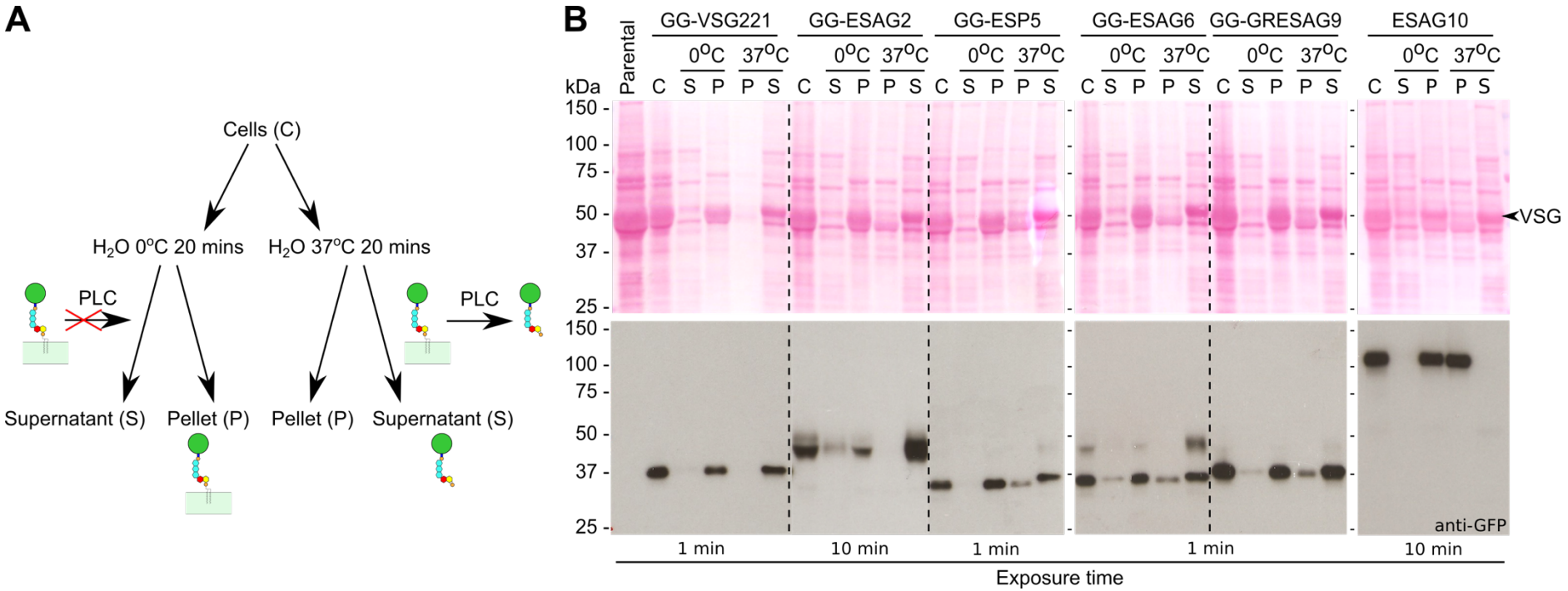


Figure 3-5. Cleavage by GPI-PLC demonstrates GPI anchoring of the sfGFP-GPI fusions. A Flow diagram of the hypotonic lysis assay. *T. brucei* contains a GPI phospholipase C (GPI-PLC) which cleaves GPI anchors upon hypotonic cell lysis at 37°C. Cells are lysed at either 37°C or 0°C and the whole cell lysates fractionated into pellet and supernatant. At 0°C GPI-PLC is inactive, the anchors are not cleaved, and GPI-APs are retained in the pellet. At 37°C, GPI-PLC cleaves the lipids of GPI anchors, and GPIAPs move to the supernatant. **B** Ponceau-S stained membrane and immunoblot of hypotonic lysis samples of sfGFP-GPI fusions. The visible movement of the VSG band (50 kDa) from the 0°C pellet to the 37°C supernatant acts as an internal control. Movement from pellet to supernatant is seen for all sfGFP-GPI fusions, indicating that they are all GPI-anchored. ESAG10 is a multipass protein, serving as a non-GPI negative control.

3.2.4. Fluorescent protein-GPI fusions localise to the endosome, flagellar pocket and cell body membranes in *T. brucei* BSF

To determine the cellular localisations of the sfGFP-GPI fusions, the cell lines were observed by live native fluorescence microscopy. Representative micrographs are shown in Figure 3-6. All 5 sfGFP-GPI fusions localised to the endosomes, flagellar pocket and across the whole cell surface membrane, implying that sorting of GPIAPs in *T. brucei* is not dictated by GPISS alone. However, the localisations were not identical, with some of the constructs showing much stronger endosomal signal relative to surface signal. This could have been due to increased protein levels, as the weakest endosomal signal appears in GG-ESAG2, which was the line with the lowest GG-fusion levels according to immunoblot (Figure 3-3). In addition, the different cell lines (most notably GG-ESP5 and GG_ESAG6) contained punctate dots within the cell body. These dots may have appeared due to the efficiency of processing of the different constructs, with some protein aggregating within the ER; however, without a marker for the different organelles, nothing more can be said about their precise localisations. One option would be to stain the cells by immunofluorescence, using anti-GFP antibodies, and antibodies against organelle-specific markers (such as BiP for the ER and p67 for the lysosome). This technique would also provide further evidence towards the surface localisations of the different constructs.

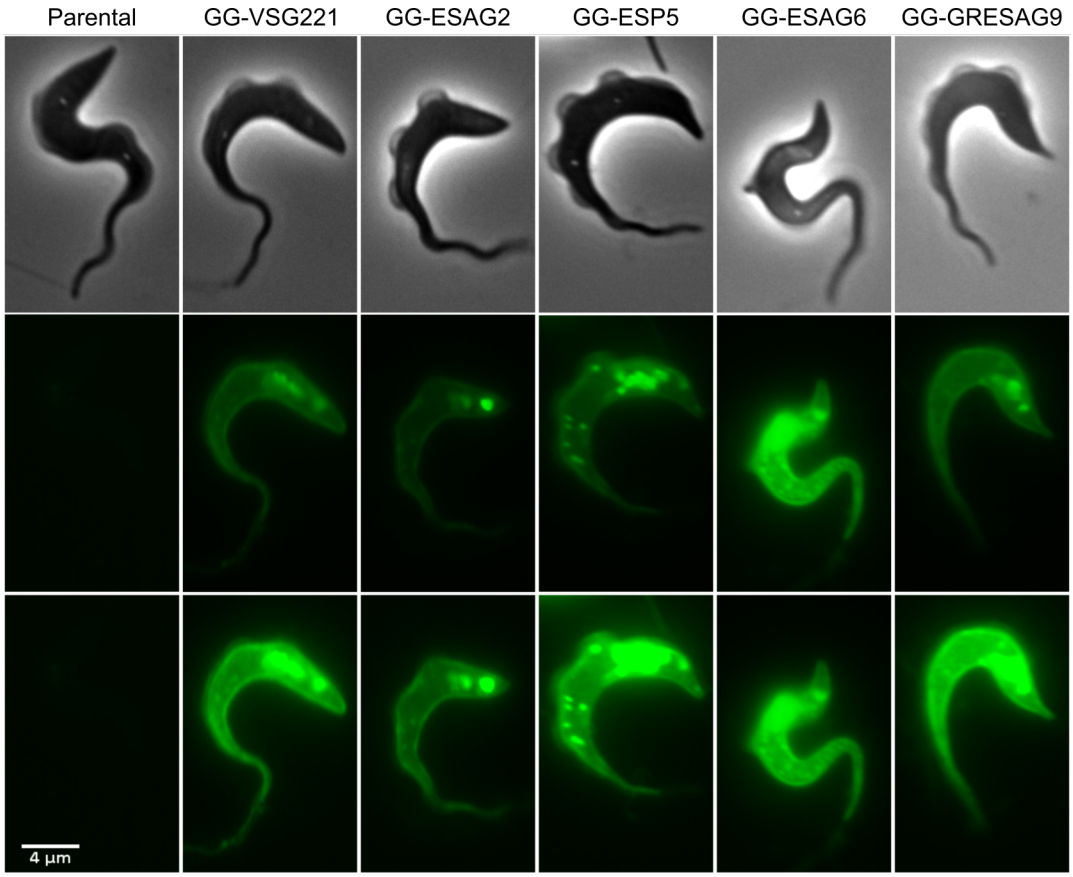


Figure 3-6 sfGFP-GPI fusions all locate to the endosomes, flagellar pocket and cell body. Live microscopy of cell lines containing sfGFP-GPI fusions. All images were captured with the same exposure. The florescence micrographs in row two are processed equally. The florescence micrographs in row three are the same images again but processed to make observation of protein localisation clearer.

The microscopy data show that GPIAP sorting may not be as straightforward as the valence hypothesis suggests, as sfGFP is thought to be monomeric (Costantini *et al.*, 2012; Pédelacq *et al.*, 2006). This would leave it with one GPI anchor and a predicted localisation exclusively within the flagellar pocket. There is some evidence, however, that sfGFP has a weak ability to dimerise when expressed membrane bound (Cranfill *et al.*, 2016). mScarlet-I, on the other hand, is a bright, mid-red fluorescent protein specifically engineered to be monomeric at all times (Bindels *et al.*, 2017). Hence, mScarlet-I was used to further test the valence hypothesis.

mScarlet-I was cloned into pSiG (VSG221 GPISS), transfected into S16 cells and Transfectants were selected using 5 µg/mL Hygromycin B. As mScarlet-I was engineered using a sequence derived from mCherry,

independent clones could be screened by anti-mCherry immunoblotting. A representative mScarlet-I clone was taken forward for further analysis. Much like its sfGFP equivalent, the mScarlet-I-GPI fusion migrated slightly slower than predicted by SDS-PAGE, likely due to N-glycosylation of the VSG221 GPISS (Figure 3-7A). Upon observation by live, native fluorescence microscopy, mScarlet-I-GPI showed the same localisation as the sfGFP-GPI fusions (Figure 3-7B). This result confirms that GPIAP sorting is at least not solely dictated by the valence hypothesis.

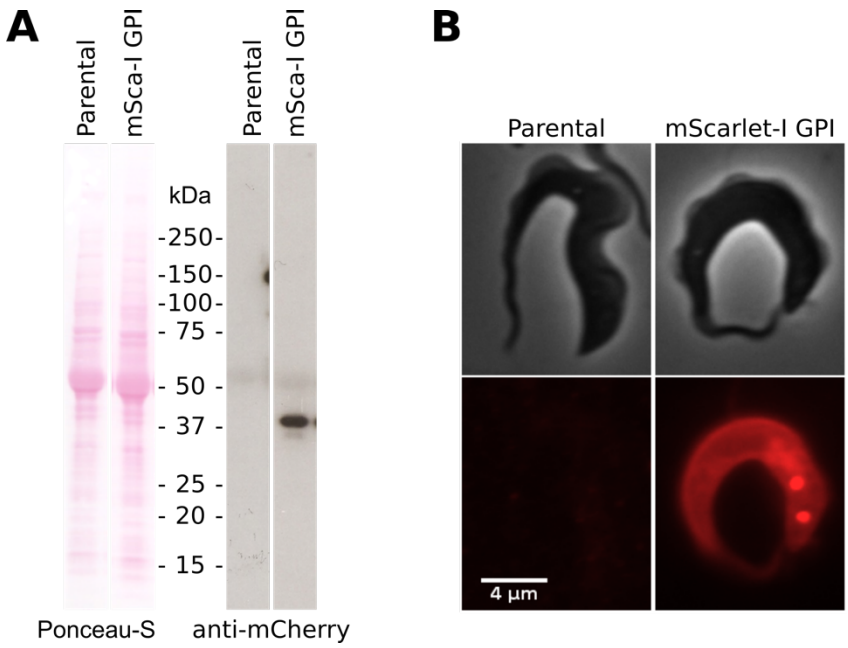


Figure 3-7 A second monomeric reporter protein (mScarlet-I) also localises to the whole cell surface when fused to a GPI signal sequence. A Ponceau-S stained membrane and anti-mCherry immunoblot of whole cell lysates of a representative clone expressing an mScarlet-GPI fusion. Predicted MW is ~34 kDa. mScarlet-I band appears slightly higher than predicted, likely due to the N-glycosylation of the VSG221 GPI signal peptide. **B** Live microscopy of the mScarlet-I-GPI cell line. The GPI signal peptide comes from VSG221. The fusion localises to the endosomes, flagellar pocket and cell surface membranes.

3.2.5. Environment around the ω-site effects GPI anchor processing

Despite the similar localisations of the different sfGFP-GPI fusions, the appearance of the 45 kDa bands in Figure 3-3 suggest differential processing of the GPI anchor, such as through the addition of sugar side chains to the mannose residues (Ferguson *et al.*, 1988; A Mehlert *et al.*, 1998; Mehlert and

Ferguson, 2007; Zitzmann *et al.*, 2000). As discussed in section 3.1, different GPI-anchor modifications in *T. brucei* are thought to be dictated by the amino acid environment directly upstream of the ω-site affecting accessibility by modifying enzymes. (A Mehlert *et al.*, 1998; Zitzmann *et al.*, 2000). Each of the sfGFP-GPI fusions were designed to include 4 amino acids upstream of their respective ω-sites and, thus, should retain these 5 amino acids (including the ω-site) at the C-terminus of the sfGFP sequence after GPI-anchor attachment. Therefore, to investigate whether these amino acids contribute to the appearance of the 45 kDa bands, modifications were made to ω-4 to ω-1 of GG-ESAG2 and GG-ESAG6 (summarised in Figure 3-8 and Table 3-2). GG-ESAG2 was selected as the only sfGFP-GPI fusion to show exclusively a ~45 kDa band upon immunoblot. GG-ESAG6 was chosen as a candidate that showed a band of predicted size; GG-VSG221 was avoided as the ω-site predicted by PredGPI was shown here to be incorrect, whilst GG-ESP5 and GG-GRESAG9 have predicted ω-sites consisting of amino acids not previously reported in this position (Pro and Gln).

Table 3-2 Amino acids in positions ω-4 to ω for the modified GG-ESAG2 and GG-ESAG6 fusions. Modifications correspond to those in Figure 3-8. ω-site shown in red.

GG-ESAG2 ω-4 to ω Modification	ARVDN	**GG-ESAG6 ω-4 to ω Modification**	RGPFT
ω-3_Ct	RRVDN	A	AGPFT
ω-2_Ct	SRVDN	AA	AAPFT
ω-1_Ct	TSRDN	AAA	AAAFT
ω_Ct	GTSRN	AAAV	AAAGT
VSG221	GKTGN	ω	GTSRT
ESP5	SPARN	ESAG2	ARVDT
ESAG6	RGPFN		
GRESAG9	ANSTN		
AAAG	AAAGN		

To GG-ESAG2, the 4 amino acids upstream of the ω-site were sequentially removed, leaving the GPISS fused to sfGFP with different length linkers. The 4 amino acids were also swapped for those of the other GPISSs (VSG, ESP5, ESAG6 and GRESAG9), and for 4 small neutral amino acids (AAAG). AAAG was included as a small, flexible linker that should theoretically have minimal interference with the anchor. All of the above modifications were analysed by PredGPI to ensure that the predicted ω-site did not change. The reciprocal experiment was to be carried out on GG-ESAG6. However, some of the subsequent deletions and swaps were predicted by PredGPI to have a different ω-site. Thus, instead of sequential removal of ω-4 to ω-1, these 4 amino acids were either entirely removed or sequentially swapped for small neutral amino acids – AAAV. As a complimentary experiment to the GG-ESAG2 swaps, GG-ESAG6 ω-4 to ω-1 was also exchanged for ω-4 to ω-1 from GG-ESAG2. Figure 3-8 shows schematics of all of the above modifications, whilst Table3-2 shows the amino acids in positions ω-4 to ω-1 as a result of the modifications.

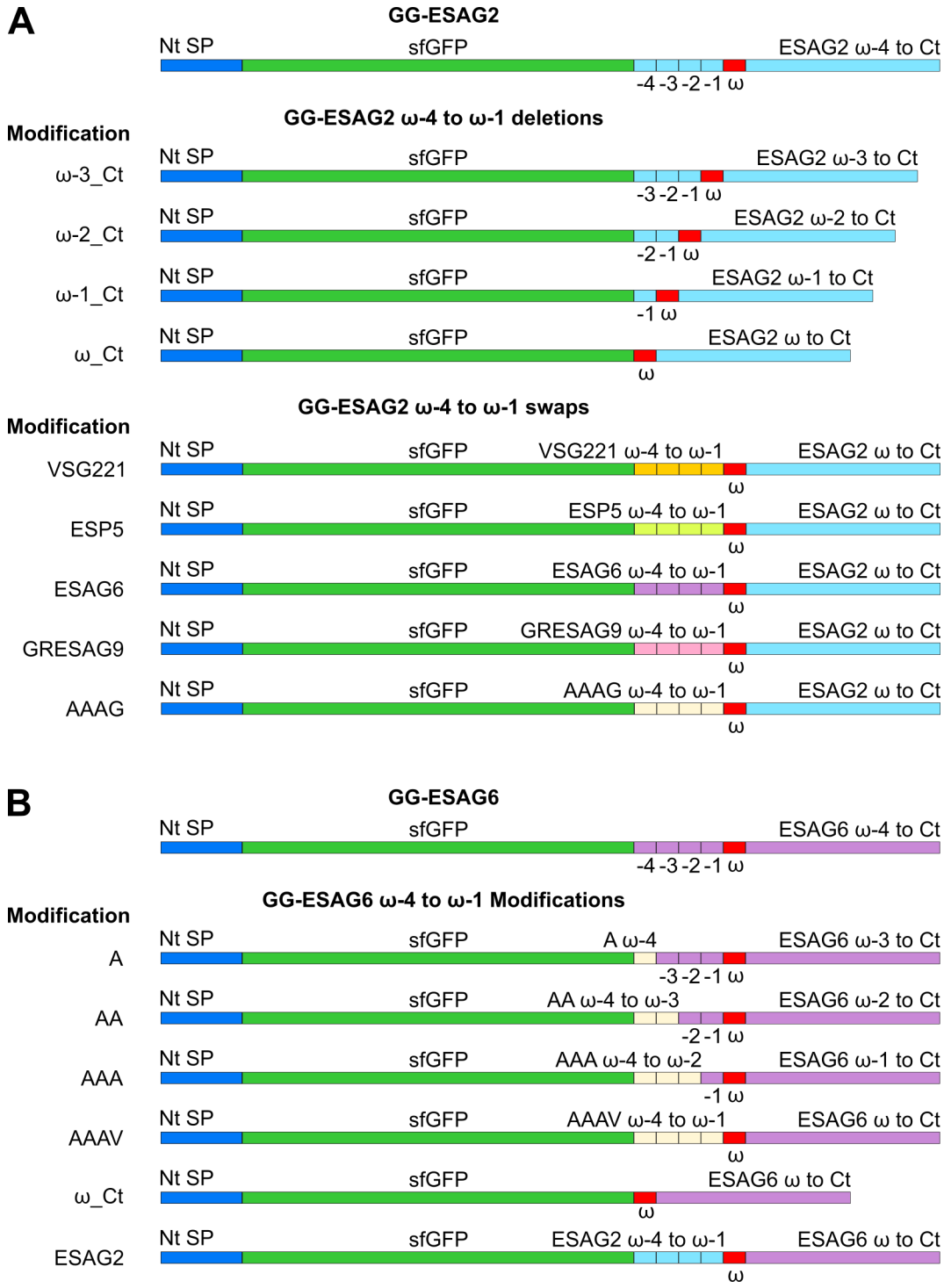


Figure 3-8 Schematics of sfGFP-GPI fusion modification. To investigate the importance of amino acids upstream of the ω site, modifications were made to the sfGFP-GPI fusions GG-ESAG2 (**A**) and GG-ESAG6 (**B**). GG-ESAG2 was modified by sequentially deleting ω-4 to ω-1, or by swapping ω-4 to ω-1 for amino acids from the other sfGFP-GPI fusions, or for four small neutral amino acids. I*n silico* analysis of some of the corresponding modifications to the GG-ESAG6 fusion show a change the predicted ω-site. For this reason, instead of deletions, ω-4 to ω-1 were sequentially swapped for four neutral amino acids, completely deleted, or swapped for the ESAG2 ω-4 to ω-1.

To create the sfGFP-GPI modifications, the GPISSs were PCR amplified from their respective pSiG-HhsfG derivatives using primers that included tags corresponding to the different ω-4 to ω-1 modifications. These amplicons were then ligated back into pSiG-HhsfG and sequences confirmed by Sanger sequencing. Once these constructs were all created, they were linearised with NotI and transfected into S16s. Transfectants were selected using 5 μg/mL Hygromycin B, and 3 independent clones from each transfection were screened for sfGFP expression by anti-GFP immunoblotting. sfGFP levels were consistent between different clones derived from each individual transfection (data not shown). Thus, a representative clone from each transfection was taken forward for further analysis.

To ensure that the modified sfGFP-GPI fusions were still GPI-anchored, and that the changes had not affected localisations, live fluorescence microscopy was carried out on representative clones (Figure 3-9). All lines produced surface localised GFP with no change in distribution, implying that the modified sfGFP fusions are still GPI-anchored; although, the GG-ESAG2 ω-4_ω swaps did seem to have some impact on protein levels. As before, the modified GG-ESAG6 lines (Figure 3-9C) contained a greater amount of punctate dots in their cell bodies compared to the modified GG-ESAG2 lines (Figure 3-9A and B).

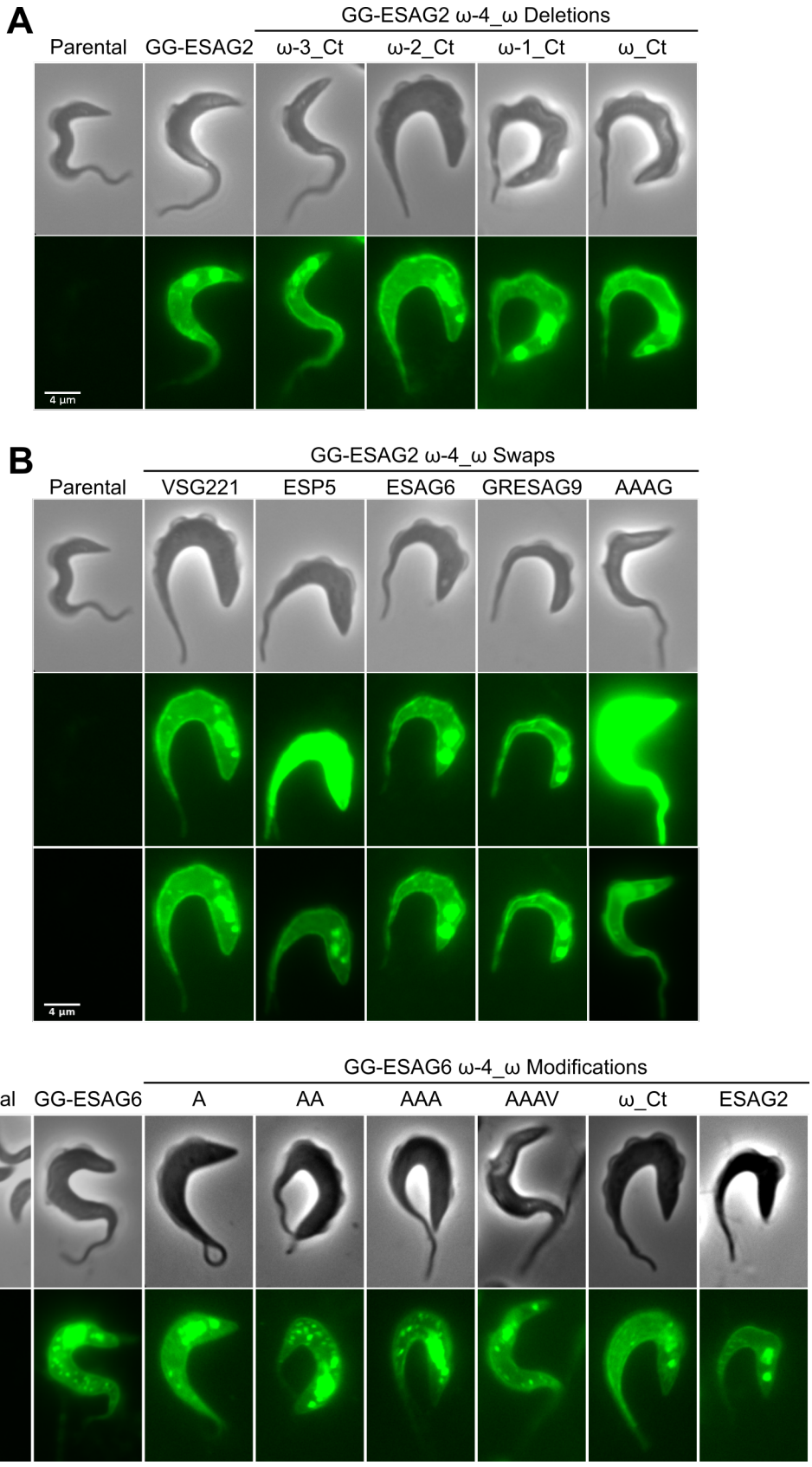


Figure 3-9 Modifications to sfGFP-GPI fusions do not alter their cellular locations. Live microscopy of cell lines expressing modified sfGFP-GPI constructs. All lines show the GFP fusions localising to the endosomes, flagellar pocket and cell body. **A** GG-ESAG2 deletions. **B** GG-ESAG2 swaps. **C** GG-ESAG6 modifications. All images within

A B or **C** were captured with the same exposure. Micrographs in **A** and **C** were processed equally. In **B,** the florescence micrographs in row two are processed equally, whilst the florescence micrographs in row three are the same images again but processed to make observation of protein localisation clearer.

An anti-GFP immunoblot of representative clones is shown in Figure 3-10. Regarding GG-ESAG2, deletions of ω-4 to ω-1 had minimal effect on the apparent MW of the sfGFP-GPI fusions. Interestingly, however, swapping these amino acids for those from the other GPISSs (other than ESAG6) led to the appearance of bands at the expected MW of ~34 kDa (and ~37 kDa for GRESAG9 due to N-glycosylation). In contrast to GG-ESAG2, the GG-ESAG6 modifications had little to no effect on the processing of the sfGFP-GPI fusions, but did lead to variation in the detectable sfGFP levels by Western blot, potentially revealing differing efficiency in GPI anchor attachment, as unanchored sfGFP would likely be secreted from the cells and not appear on the blot. These disparate results between GG-ESAG6 and GG-ESAG2 could perhaps be explained by a wrong prediction of the GG-ESAG6 ω-site by PredGPI. This would leave more amino acids on the fusion protein after anchoring than predicted. Nevertheless, as a whole, the results show that the amino acids upstream of the ω-site can impact the production of the sfGFP-GPI fusions, whether that be through apparent MW (possibly through influencing modification of the GPI anchor) or through observed protein abundance (possibly through an effect on efficiency of anchor attachment).

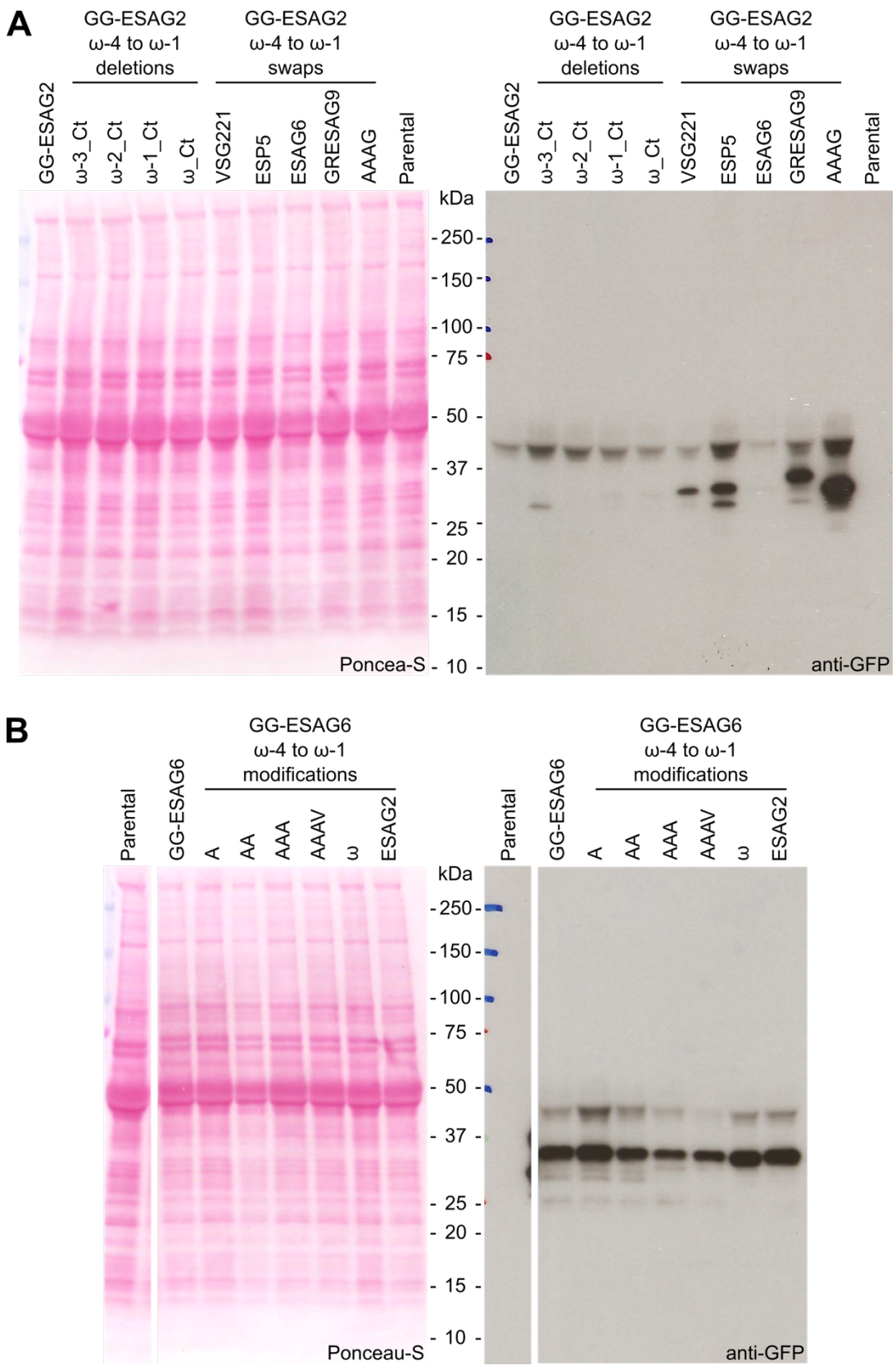


Figure 3-10 Effect of amino acid modifications on the apparent mobilities of sfGFP-GPI fusions in SDS-PAGE. Ponceau-S stained membranes and anti-GFP immunoblots of whole cell lysates of cells expressing modified sfGFP-GPI fusions. **A** GG-ESAG2 modifications. The ω-4 to ω-1 deletions do not affect relative migration of the GG-ESAG2 fusion. However, swapping ω-4 to ω-1 causes bands to appear at

~32kDa. The GRESAG9 modification causes a slightly higher band due to the N-glycosylation site within ω-4 to ω-1. **B** GG-ESAG6 modifications. MWs equal to the unmodified fusion (32 kDa). Modifying ω-4 to ω-1 does not affect relative migration of the GG-ESAG6 fusion but does affect protein level.

3.3. Discussion

This chapter describes attempts to contribute towards our understanding of how GPIAPs are sorted to distinct domains on the plasma membrane of *T. brucei* BSF. Experiments were designed to assess both the effect of the GPISS on sorting of a reporter protein and to further explore the GPI valence hypothesis. Thus, an in-silico analysis was used to determine predicted ω-sites of *T.* brucei GPIAPs with known distinct localisations, and their respective GPISSs attached to a reporter protein – sfGFP. All resultant fusions localised across the entire plasma membrane, revealing that the GPISS is not sufficient to drive differential surface sorting of GPI-anchored sfGFP in *T. brucei* BSF. Further to this, the results indicated that the valence hypothesis (Schwartz *et al.*, 2005) may also be insufficient to explain subcellular sorting of GPIAPs alone. This was confirmed with a second monomeric fluorescent protein – mScarlet-I.

Despite these observations, it is possible that either hypothesis could be involved in GPIAP sorting, but that the results in this Chapter are obscured by the effect of flagellar pocket saturation. To address this, the sfGFP-GPI fusions could be integrated into the endogenous loci of known pocket proteins, utilising the same DNA processing signals and ensuring similar expression levels. Nevertheless, in contrast to the results here, evidence backing up the valence hypothesis (Tiengwe *et al.*, 2017) and the possible role that glycosylation may still play (Mehlert *et al.*, 2012) show that *T. brucei* surface protein sorting is likely a complex, multi-factorial process and may not rely exclusively on any one of these elements.

Whilst surface localisations of the different sfGFP-GPI fusions were similar, different apparent MWs were revealed by immunoblot. This was in part due to N-glycosylation of GPISS amino acids upstream of the ω-site, retained on sfGFP fusions after GPI-anchoring (Figure 3-4). However, the appearance of a

~45 kDa band was shown not to be caused by N-glycosylation, but was influenced by the presence of different amino acids directly upstream of the ω-site (Figure 3-10). This could be due to differential modification of the GPI anchor, as *T. brucei* is able to attach sugar sidechains to protein-linked GPI anchors (Ferguson *et al.*, 1988; A Mehlert *et al.*, 1998; Mehlert and Ferguson, 2007; Zitzmann *et al.*, 2000). Thus, amino acids upstream of the ω-site may result in distinct secondary structures around the GPI anchor, limiting access by anchor modifying enzymes, as has previously been suggested for different classes of VSG (A Mehlert *et al.*, 1998; Zitzmann *et al.*, 2000). Yet, the results of this Chapter were not unambiguous. GG-ESAG2 with or without a 4 amino acid linker between the sfGFP ORF and the predicted ω-site both solely resulted in the larger ~45 kDa product, whilst insertion of particular alternative amino acids (other than those of GG-ESAG6) resulted in the appearance of smaller ~34-37 kDa bands. In contrast, removal of the linker from GG-ESAG6, or insertion of the GG-ESAG2 linker into GG-ESAG6, did not prevent the production of the ~34 kDa band. However, the contrasting results could be caused by an incorrect ω-site prediction by PredGPI, thus resulting in amino acids proximal to the ω-site that were different than expected.

Possible miss-calling of ω-sites by PredGPI was highlighted by its incorrect annotation of the GG-VSG221 ω-site (based on the position of proven N-glycosylation) and its likely erroneous annotation of the GG-ESP5 and GG-GRESAG9 ω-sites. This is likely due to the training set of proteins used in the creation of the PredGPI algorithm (Pierleoni *et al.*, 2008). At the time of set-up of PredGPI, of the 340 proteins within the SwissProt database (from which the training protein sets were taken) that were experimentally proven to be GPI-anchored, only 26 had had their ω-sites experimentally confirmed. In addition, only 2 of the 26, and fewer than 20 of the 340, were proteins from trypanosomatids. There is some evidence that trypanosomatid GPISSs can differ in nature to common GPISSs from other eukaryotes (Eisenhaber *et al.*, 1998; Moran and Caras, 1994).

Other algorithms exist for the prediction of GPI-anchoring of proteins and their ω-sites. These include BigPI, DGPI and GPI-SOM. When PredGPI was created, the abilities of the four algorithms to correctly assign the 26

experimentally proven ω-sites were assessed (Pierleoni *et al.*, 2008). PredGPI correctly assigned 24 ω-sites, BigPI 23 ω-sites, DGPI 16 ω-sites, and GPI-SOM 15 ω-sites. As the only algorithm of comparable performance, BigPI was also used to assess the ω-sites of the sfGFP-GPI fusions of this chapter. However, only GG-VSG221 and GG-ESAG2 were predicted to be anchored. For this reason, BigPI was not used to further inform the results here.

To overcome the issue of ω-site annotation and further disentangle the results of this chapter, accurate experimental identification of the ω-sites would be required. One way to do this would be to isolate the sfGFP-GPI fusions and use hydrofluoric acid to cleave phosphodiester bonds of the GPI anchors, releasing sfGFP retaining only the ethanolamine group from the GPI anchor. Released sfGFP could then be analysed by mass spectrometry to confirm ω-site identity. Alternatively, a high throughput version of this method could be used to identify the 'GPIome' of *T. brucei*, whilst simultaneously confirming the ω-sites of these GPIAPs. Masuishi *et al* developed a method for this very purpose (Masuishi *et al.*, 2016). They isolated GPIAPs from human cancer cell lines using detergents, before cleaving the GPI lipids with a bacterial PLC. This cleavage leaves the majority of the GPI anchor on proteins, with a terminal phosphate exposed. The authors then digested these proteins with trypsin/chymotrypsin and enriched for peptides attached to the remainder of the GPI anchor by use of titanium dioxide (which can pulldown phosphate-linked peptides). Finally, this was followed by hydrofluoric acid treatment and mass spectrometry analysis. A total of 49 GPIAPs were identified, with ω-sites including the following amino acids in order of abundance: Ser>Gly>Asn>Ala>Asp>Cys>Met>Leu>Thr. This methodology could provide a useful tool for identification of trypanosomatid GPIAPs and their ω-sites. However, an analysis of this kind is beyond the scope of this project and was not attempted here.

This chapter has provided further evidence towards unveiling the sorting mechanisms employed by *T. brucei*. It has revealed that neither GPISS nor GPI-valence are sufficient to dictate localisations of surface proteins. However, more work is still to be done to unravel this complex multifaceted process.

Chapter 4. Recombinant *Trypanosoma brucei* protein expression in a system based on *Leishmania tarentolae* (LEXSY)

4.1. Introduction

In addition to investigating possible cell surface sorting mechanisms, information gleaned from the BSF *T. brucei* surfeome will be vital in elucidating host parasite interactions, discovering new essential proteins for parasite survival and in opening up new avenues for tackling HAT. Our lab explores these related themes to ultimately better understand *T. brucei* biology and how it survives in the mammalian host. Ongoing projects in the lab are addressing the question of protein function and essentiality through genome-scale RNAi screens and individual knockdowns. My project takes a different approach and revolves around the development of tools required to evaluate the potential of surfeome proteins as vaccination candidates.

There are currently no available vaccinations against African trypanosomiasis (Black and Mansfield, 2016). VSG itself is highly immunogenic. However, the rate of VSG switching, coupled with the highly variable nature of exposed VSG epitopes and the buried shielded nature of conserved VSG peptide sequences, prohibits the development of an effective VSG vaccine (Schwede *et al.*, 2015). Studies that have looked towards invariant surface proteins have had mixed success. Vaccination of mice with the extracellular domains of ISG65 and ISG75 recombinantly produced in *Escherichia coli* did not show a protective effect (Ziegelbauer and Overath, 1993). Immunisation with a DNA plasmid encoding an ISG65 ORF did induce a protective effect in mice, albeit only partial and for a low dose parasite infection (Lança *et al.*, 2011). The lack of a strong protective effect by ISG65 is thought largely to be down to limited access by the immune system due to rapid recycling of ISG65 and shielding by VSG (Schwede *et al.*, 2015).

Whilst individual ISGs are present at relatively low levels across the parasite surface (~1 for every 200 VSG, Schwede *et al.*, 2015), an alternative

source of highly concentrated invariant proteins can be found in the flagellar pocket (Field and Carrington, 2009). Access to the pocket is via a channel of 25 nm in diameter (Gadelha *et al.*, 2009). This is wide enough to allow entry to members of the complement pathway and also IgG antibodies. As a result, the invariant proteins that reside here could act as potential vaccines against HAT. One study attempted to immunise mice with a flagellar pocket fraction prepared from *T. brucei* BSF (Radwanska *et al.*, 2000). As found for the ISGs, a partial protective effect was observed and only for low dose infection. However, this does highlight the potential for flagellar pocket antigens as possible vaccination targets.

If an effective vaccine against HAT is to be developed, new surface antigens must be found that can bypass the current pitfalls. An effective vaccination candidate must therefore be invariant (to avoid variation through switching); it must be abundant and of high molecular weight (to avoid masking by the VSG coat); it must be immunogenic; and it must be parasite specific to avoid any antigens that might provoke a host auto-immune response. The surfeome has provided us with the largest pool to date of parasite-specific, invariant, validated surface proteins that can be assessed for their ability to overcome the current challenges (Gadelha *et al.*, 2015). In order to carry out vaccination studies on selected proteins, sufficient pure protein is required. To enable a prime-boost vaccination study, 50 µg is desired per injection to enable a maximal chance of protection (Harlow and Lane, 2014). Based upon power calculations (ClinCalc) with alpha set at 0.05 and power set at 0.9 (assuming a 60% incidence of survival in the vaccinated group *vs.* 0% in the control group), 10 mice are required in each group. Therefore, for each recombinant protein ~1 mg is required.

When attempting to recombinantly express proteins, there are many options commercially available, each with their own advantages and disadvantages. The amount of functional protein produced can vary greatly between hosts and on a protein-to-protein basis (Gomes *et al.*, 2016). Whilst high yield of recombinant protein is desirable, if the protein is incorrectly folded or processed it may not contain vital epitopes required to elicit an immune response against the native protein. Ideally, a recombinant protein expression system would have

the following features: cheap and easy to grow; genetic tractability; have the option of inducibility in case of toxicity (especially when working with proteins of unknown function); have the ability to fold, process and modify proteins as close to the endogenous versions as possible.

4.1.1. Commonly used recombinant protein expression systems

The most commonly used systems to date are bacterial systems based on species such as *E. coli* and *Bacillus subtilis* (Gomes *et al.*, 2016; Rosano and Ceccarelli, 2014). These systems have big advantages in terms of their ease of use, growth rate, protein production levels and low costs. However, prokaryotes have been shown to present some limitations when trying to express eukaryotic proteins. Notably, improper folding caused by high expression levels coupled with a lack of eukaryotic folding chaperones, can lead to non-functional or insoluble proteins and the formation of inclusion bodies (Carrió *et al.*, 2000). Additionally, there are various post-translational modifications that differ greatly between prokaryotes and eukaryotes. These include N- and O-linked glycosylation, disulphide bond formation, phosphorylation and fatty acid acylation. For example, N-glycosylation appears rarely in bacteria, tends to attach chains of N-acetylgalactosamine, and requires a longer recognition sequon than eukaryotes (Nothaft and Szymanski, 2013). For this reason, prokaryotic systems are often not the first choice when trying to study eukaryotic surface proteins as they tend to be highly glycosylated.

Other expression systems that try to get around some of these issues involve a variety of different species, including yeast, insect cells and mammalian cells. Yeast, like bacteria, have the advantages of fast growth and high protein yields, as well as being easy to genetically manipulate (Gomes *et al.*, 2016; Mattanovich *et al.*, 2012; Vieira Gomes *et al.*, 2018). They carry out eukaryotic post-translational modifications and are commonly used to produce therapeutics such as recombinant human insulin (Baeshen *et al.*, 2014). Despite this, they still differ in their N- and O-glycosylation patterns when compared to other eukaryotes such as mammals and most notably trypanosomes. Yeast, particularly *Saccharomyces cerevisiae*, have a tendency to hyper-mannosylate their N-glycans, adding up to 200 mannose residues (Conde *et al.*, 2004).

Alternative yeast species, such as *Pichia pastoris,* do this to a lesser extent, usually adding 8–14 mannose residues (Bretthauer and Castellino, 1999; Vieira Gomes *et al.*, 2018), and a *P. pastoris* strain has even been engineered to carry out mammalian-like N-glycosylation (Hamilton and Gerngross, 2007).

Insect systems are also used for the production of recombinant eukaryotic proteins (Chambers *et al.*, 2018; Gomes *et al.*, 2016). These systems utilise baculoviruses to infect insect cells and force over-production of the recombinant protein. The viruses are arthropod-specific and so considered safe to work with (as they cannot infect mammalian cells). These systems can produce high levels of recombinant proteins and have been used to produce various viral vaccines, such as the human papillomavirus (Monie *et al.*, 2008). However, they also have their drawbacks. Creation of recombinant virus particles requires multiple rounds of molecular cloning, followed by infection of insect cells with a baculovirus shuttle vector (bacmid), and wild-type virus to create the recombinant virus stock (Chambers *et al.*, 2018). This virus stock is then used to infect a large number of insect cells for recombinant protein production. This set up is lengthy, and the transfected insect cells producing the recombinant protein eventually die. This means that every round of protein production requires subsequent infection with the recombinant virus stock. Regarding glycosylation, insect cells do N-glycosylate their proteins but in a very simplified manner, utilising fucosylated paucimannose-type structures (Harrison and Jarvis, 2006; Khan *et al.*, 2017).

If complex-type N-glycosylation is necessary, mammalian-based protein expression systems can be used. (Croset *et al.*, 2012). In the last two decades mammalian cells have overtaken other protein expression platforms in terms of recombinant drugs approved for use in humans, and this is in large part due to their glycosylation abilities (Sanchez-Garcia *et al.*, 2016). The most common mammalian cell lines used for this purpose are Chinese Hamster Ovary cells (CHO) and Human Embryonic Kidney cells (HEK293). However, even between these two lines, specific complex-type glycosylation patterns can differ greatly for the same protein (Croset *et al.*, 2012). This is perhaps unsurprising, as these lines originate not only from different hosts, but also from different tissues. Even so, within these cell lines, endogenous and recombinant proteins can also be modified to different degrees with different glycosylation patterns, resulting in

the production of a variety of isoforms. This is likely due to the wide range of complex glycan patterns available to mammalian cells.

4.1.2. Leishmanial expression system (LEXSY)

A perhaps more unusual system for recombinant protein production is the **L**eishmanial **ex**pression **sy**stem (LEXSY, Jena Bioscience) based on the non-human infective kinetoplastid *Leishmania tarentolae* (Breitling *et al.*, 2002; JenaBioscience; Raymond *et al.*, 2012). As a kinetoplastid, *L. tarentolae* is closely related to *T. brucei*. It therefore likely contains the molecular machinery required to process recombinant proteins as close to the endogenous versions as possible, including folding, dimerization and post translational modifications. *L. tarentolae* grows with a relatively short doubling time and to high cell densities in culture. Through the production of human erythropoietin, LEXSY was shown to be able to N-glycosylate proteins with both paucimannose and complex-type glycans of remarkable homogeneity (Breitling *et al.*, 2002). The paucimannose glycans consisted of the core Man3GlcNAc2 structure, whilst the complex-type N-glycan structure was a non-sialylated, biantennary, bi-β-1,4-galactosylated, core-α-1,6-fucosylated structure (Figure 4-1). This N-glycosylation ability and the host's close relatedness to African trypanosomes could make LEXSY ideal for the production of *T. brucei* cell surface proteins. This chapter will focus on setting up the LEXSY system in the lab and attempts to use it to express recombinant *T. brucei* cell surface antigens.

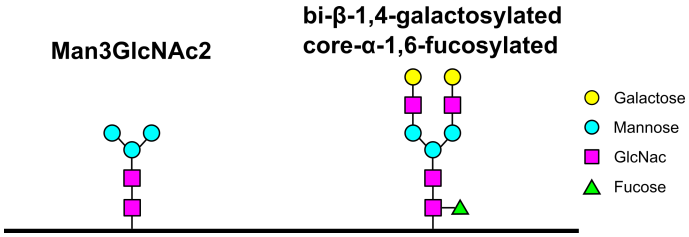


Figure 4-1 N-glycosylation structures found on recombinant proteins produced in *L. tarentolae*. On the left is the Man3GlcNAc2 glycan structure. On the right is the bi-β-1,4-galactosylated, core-α-1,6-fucosylated glycan structure.

There are multiple different configurations of the LEXSY system commercially available, including constitutive or inducible expression with

genome integrated constructs and inducible expression with episomal constructs (JenaBioscience). The inducible systems rely on the highly processive bacteriophage T7 RNA polymerase to drive recombinant gene expression. The T7TR *L. tarenolae* cell line has the T7 RNA polymerase (*T7RNAP*) and Tetracycline repressor (*TETR*) genes stably integrated into the 18S rDNA loci of its genome, utilising constitutive PolI transcription to drive expression of these genes. TETR binding to the tet-operator sequence in the expression vectors prevents transcription, unless in the presence of tetracycline which binds to TETR, releasing it from the tet-operator and allowing inducible expression of transgenes.

The episomal configuration of LEXSY (Figure 4-2) is known as a linear artificial chromosome (LAC) (JenaBioscience; Kushnir *et al.*, 2011). The LAC construct contains two telomere-like regions separated by a 900bp stuffer. When cut with SwaI, the stuffer is removed, and the linearised vector is transfected into the host cells to be propagated as a stable linear episome. This was shown to produce high expression clones (5-10 fold higher than integrated inducible expression) with low clone-to-clone heterogeneity (Kushnir *et al.*, 2011). For these reasons the LAC configuration of LEXSY (pLEXSY_IE IE-egfp-red-neo4) was used for the following experiments.

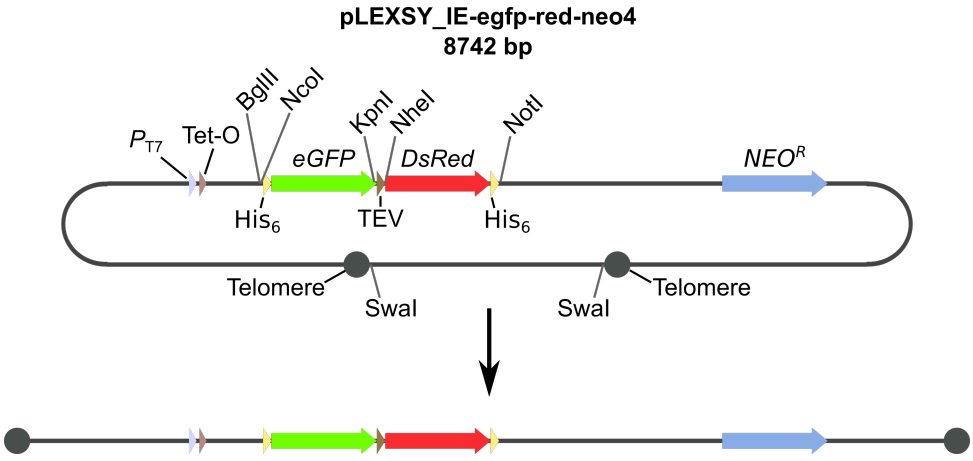


Figure 4-2 Schematic of the recombinant protein expression vector pLEXSY_IE-egfp-red-neo4 from Jena Bioscience. Genes can be cloned into the vector to create a fusion with either an N-terminal His_6-eGFP-TEVsite or a C-terminal TEVsite-DsRed-His_6. Prior to transfection the vector is linearised with SwaI, exposing its telomere-like regions. The vector is subsequently propagated as a LAC within *L. tarentolae* cells. Recombinant protein expression is driven by the T7 promotor in the presence of tetracycline.

4.2. Results

4.2.1. Characterisation of the T7TR cell line

Trypanosomatids do not synthesise haem but scavenge it from their environment (Kořený *et al.*, 2010). Some trypanosomatids, such as *Leishmania* and *Crithidia* sp., contain the last three enzymes of the biosynthetic pathway. As a result, they can survive on the addition of haemin or protoporphyrin IX to their growth media (Kořený *et al.*, 2010). According to the LEXSY manual (JenaBioscience), *L. tarentolae* can be propagated in brain heart infusion (BHI) medium supplemented with 5 µg/mL haemin. Initial attempts to grow T7TR in BHI + haemin led to slow growth (data not shown). The medium was supplemented with 5% foetal bovine serum (FBS) to improve culturing conditions. T7TR cells growing in FBS-supplemented medium doubled every 6.5 h and reached a top density of $2x10^8$ cells/mL (Figure 4-3).

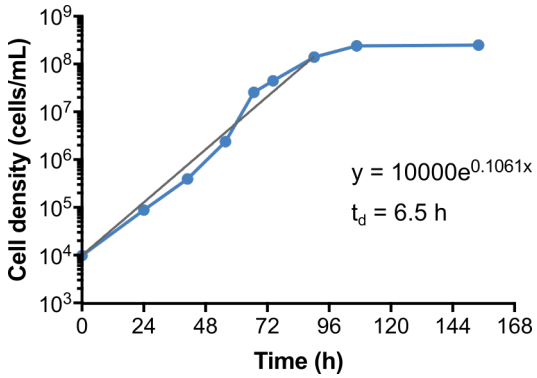


Figure 4-3 T7TR cells grow to $2x10^8$ cells/mL with a log phase doubling time of 6.5 h in BHI supplimented with 5 µg/mL Hemin and 5% FBS. Growth assesment of T7TR cell line across 7 days. Log-phase growth fitted to calculate doubling time (grey line). Equation of the fit shown, where y is cell density and x is time.

4.2.2. Modifying the original pLEXSY vector

pLEXSY_IE-egfp-red-neo4 (Figure 4-2) allows expression of a chosen transgenic protein fused to either eGFP or DsRed with a His_6 tag (to aid in screening and purification), and a TEV protease site (for removal of the fluorescent proteins). The proteins to be expressed here are surface proteins with signal peptides. As a result, these proteins will enter the secretory pathway, where they will encounter a change in redox environment as they are

translocated into the ER. ‘Superfolder’ GFP (sfGFP) is a variant of GFP that has been engineered to have improved folding kinetics, allowing it to be more resistant to the reducing environments of the ER and the extracellular space compared to other variants (Pédelacq *et al.*, 2006). For this reason, *sfGFP* was chosen to replace *eGFP* and *DsRed* in pLEXSY_IE-egfp-red-neo4. A TEV site and His_6 were to be included to create ORF TEV-*sfGFP*-His_6 for C-terminal tagging of recombinant proteins.

When designing DNA sequences for protein production, an important factor to consider is codon bias. Codon bias is an organism’s inherent bias for using particular codons for specific amino acids more frequently than others (Quax *et al.*, 2015; Sharp and Li, 1986). This bias can affect the efficiency of recognition of particular codons by tRNAs. Usage of rare codons can lead to a depletion of low abundance tRNAs and subsequently cause ribosome stalling and detachment. Previous studies have looked at various aspects of codon influence on gene and protein expression, including frequency of codon use, copy number of tRNA genes, translation efficiency and mRNA stability (Quax *et al.*, 2015). Recent studies have been done specifically on the effect of codon bias on expression levels in *T. brucei* (de Freitas Nascimento *et al.*, 2018; Jeacock *et al.*, 2018). In both studies, using different methods they showed that there is a degree of predictive power in the sequence of an ORF towards its mRNA expression levels in *T. brucei*. These predictions could be used to increase expression of GFP or luciferase. Jeacock *et al* showed that by increasing the number of GC3 codons in a sequence (codons with a G or C in the third position), both mRNA and protein levels could be increased in *T. brucei* BSF. Freitas Nascimento *et al* developed a gene expression codon adaptation index (geCAI), which looks at the relationship between used codons and transcript abundance. Each codon is given a score between 0 and 1 based on its correlation with highly abundant transcripts; a codon with a score of 1 is predicted to have the highest positive impact on mRNA abundance, whilst a score of 0 predicts the opposite. A geCAI score is then given to an ORF based on geometric mean of codon scores across the its length. This was used to recode the GFP ORF to have a range of different geCAI scores, which were highly predictive of GFP protein and mRNA levels in *T. brucei* PCF.

To test if codon bias would have an effect on recombinant protein level in *L. tarentolae*, two TEV-*sfGFP*-His_6 sequence fragments were designed (Figure 4-4A). One fragment was based on the original *sfGFP* sequence, which itself had been created with a bacterial codon bias (Pédelacq *et al.*, 2006). Alongside this, a fragment was designed to be codon optimised for *L. tarentolae*. To create this *L. tarentolae* codon-optimised sequence, the codon bias of a highly expressed subset of genes was used – the ribosomal protein genes. To assess the frequency of use of synonymous codons within this set, the Relative Synonymous Codon Usage (RSCU) was calculated as described in Chapter 2. RSCU assigns a score to each codon relative to the expected frequency if the distribution of codon usage was random for that particular amino acid. Non-biased use of a codon results in a score of 0. A positive bias towards use of a codon results in a score >0, whilst a negative bias a score <0. The codon-optimised TEV-*sfGFP*-His_6 sequence uses the codons with the highest scores, differing from the optimal codons only at restriction-enzyme sites and His_6 as indicated. Codon-optimised *sfGFP* is referred to as $sfGFP^{op}$. Figure 4-4B shows the comparison of the RSCU values across the TEV-*sfGFP*-His_6 ORF for the original, codon-optimised and optimal sequences.

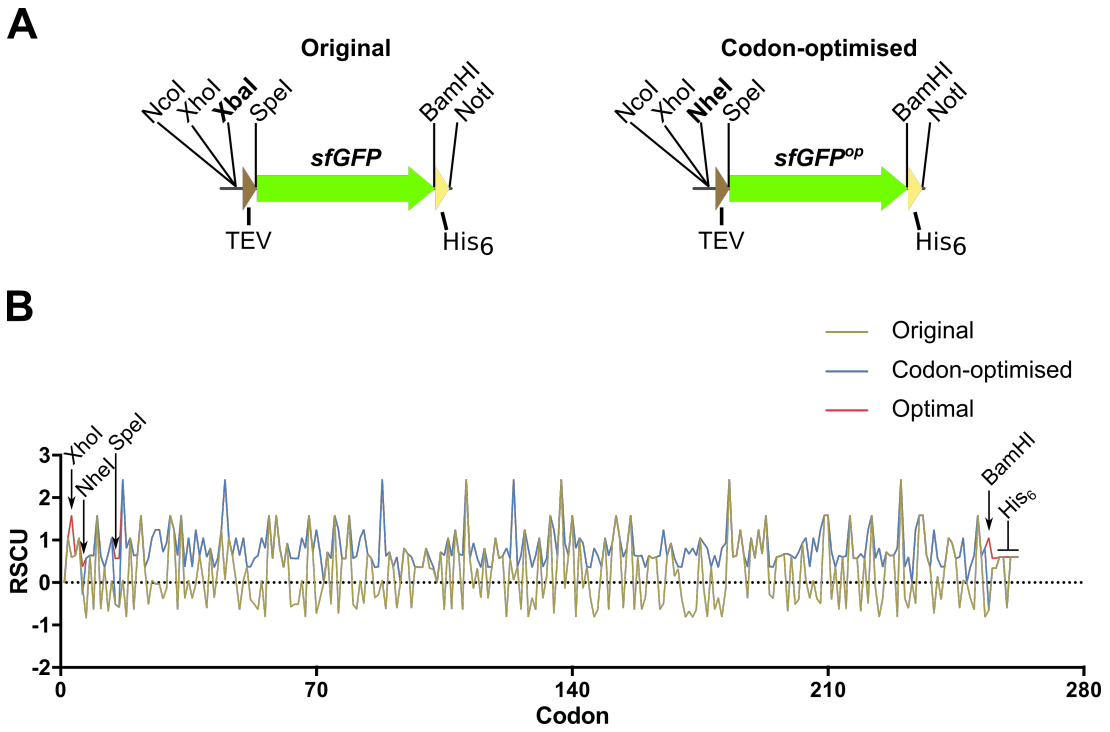


Figure 4-4 Comparison of RSCU values across the TEV-*sfGFP*-His$_6$ fragment for the original, the codon-optimised and the optimal sequences A Schematics of the original and codon-optimised TEV-*sfGFP*-His$_6$ sequences. Differing restiction sites in bold. **B** The codons used across the original sequence are shown in yellow. The optimal choice of codons according the *L. tarentolae* ribosomal codon bias are shown in red. The codons of the codon-optimised sequence are shown in blue. The optimised codons differ from the optimal only at the restriction sites and the His$_6$ as indicated.

Both the original and the codon-optimised DNA fragments were ordered as gBlocks from Integrated DNA Technologies (IDT). These were then cloned into pLEXSY_IE IE-egfp-red-neo4 cut with NcoI and NotI to create pLEXSY_IE-sfG-N and pLEXSY_IE-sfGO-N (Figure 4-5A). These three vectors were linearised and transfected into the T7TR cell line, using G418 for selection of transgenic cell lines. To assess relative protein levels derived from each vector sequence, cell lines were induced for 24 h with 10 µg/mL tetracycline (tet) and whole cell lysates were prepared. These lysates were resolved by SDS-PAGE and electro-transferred to nitrocellulose membrane. Following Ponceau-S staining to visualise whole cell proteomes, the membrane was probed by immunoblot using anti-GFP antibodies (Figure 4-5B).

The immunoblot showed codon optimisation to have a substantial effect on protein levels, with cells containing pLEXSY_IE-sfGO-N expressing a much higher amount of sfGFP^{op} relative to sfGFP from cells containing pLEXSY_IE-

sfG-N. They also expressed a much higher amount of sfGFP^{op} than the amount of eGFP-DsRed expressed by cells containing pLEXSY_IE-egfp-red-neo4. These cells only produced a weak band of expected size at 56 kDa along with other bands, likely to be degradation products. The correct molecular weight was barely picked up by a short exposure, whilst the strongest band was a degradation product picked up at 25 kDa, close to the expected size for eGFP of 27 kDa. Even this band only managed to match the levels of the sfGFP from the non-optimised pLEXSY_IE-sfG-N cell line.

Due to the much higher recombinant protein levels, pLEXSY_IE-sfGO-N was taken forward. However, the leakiness of this construct should be noted. The level of sfGFP^{op} seen from non-induced cells containing pLEXSY_IE-sfGO-N was higher than the level observed from induced cells containing pLEXSY_IE-sfG-N. This leakiness is likely down to the design of the vector and the way that gene expression functions in trypanosomatids. They express their genes as long polycistrons which are then processed into the individual mature mRNA (Wirtz *et al.*, 1998). *NEO^{R}* in the pLEXSY vectors does not have its own promoter, and instead is expressed in a polycistron along with the recombinant protein ORF. G418 pressure, therefore, selects for leakiness of the construct, as without this, transgenic cells would die when exposed to the selection drug. The extent of this leakiness has likely been accentuated by the codon optimisation of *sfGFP^{op}*. This could be an issue if the construct is used to express toxic protein products.

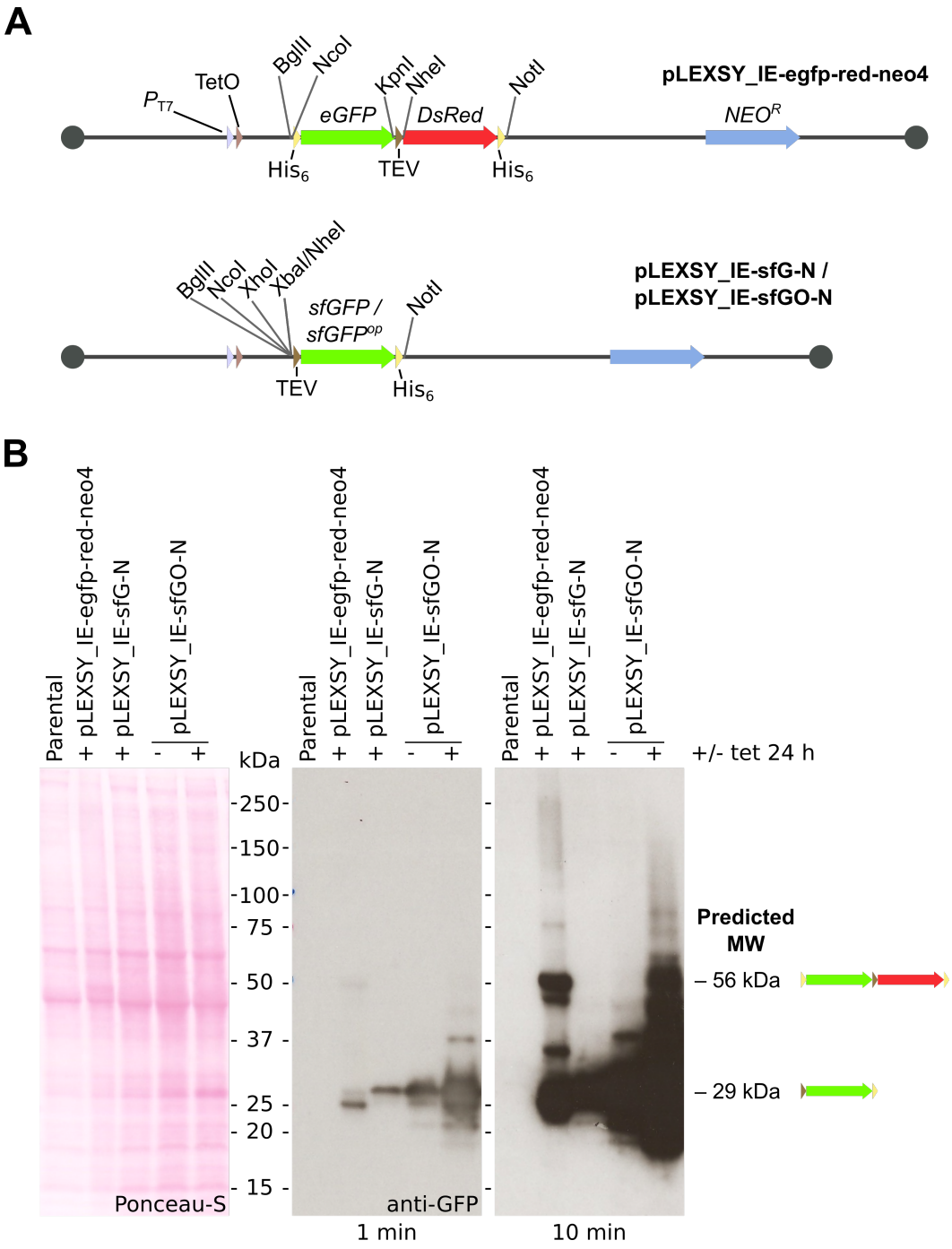


Figure 4-5 Codon bias affects recombinant protein levels of a reporter GFP in *L. tarentolae*. A Schematics of linearised pLEXSY_IE-egfp-red-neo4, pLEXSY_IE-sfG-N and pLEXSY_IE-sfGO-N. The latter two constructs were made through swapping the His_6-*eGFP*-TEV-*DsRed*-His_6 for TEV-*sfGFP*-His_6 with either its original sequence or a sequence codon optimised for *L.tarentolae* respectively. **B** Ponceau-S stained membrane and anti-GFP immunoblot of whole cell lysates of T7TR lines expressing His_6-eGFP-TEV-DsRed-His_6, TEV-sfGFP-His_6 or TEV-sfGFP^{op}-His_6. Cells either grown in the presence (+) or absence (-) of tetracycline for 24 h. GFP signal seen from cells expressing *sfGFP^{op}* is substantially stronger relative to cells expressing *sfGFP* or *eGFP-DsRed*, showing that codon optimisation can positively affect *L. tarentolae* recombinant protein levels.

4.2.3. *T. brucei* recombinant protein expression in pLEXSY_IE-sfGO-N

Other than Figure 4-6B, the work in section 4.2.3 of this chapter was carried out by three MSci students under my supervision – Charlotte Day, Frazer Whittaker and Ryan Beazley.

The surfeome includes a set of proteins whose surface localisations were validated by endogenous locus tagging and fluorescence microscopy – ESPs (enriched in surface-labelled proteome) and ESAGs (Gadelha *et al.*, 2015). This set was therefore examined to find potential vaccination candidates that addressed the selection criteria of being large, abundant, parasite-specific, invariant surface proteins. In addition to these criteria, a combination of cell body and flagellar pocket-localised proteins were selected to test whether pocket proteins are indeed viable targets for vaccination. ESAGs are multigene families that are co-expressed along with VSG from the active expression site and, as such, are subject to variable expression. For this reason, the ESAGs were not considered here. Within the set of proteins under consideration, candidates were excluded which had a large portion of their structure predicted to be on the cytoplasmic side of the plasma membrane (thus not surface exposed). Those with only weak surface signal observed by fluorescence microscopy were also disregarded. GPIAPs were avoided at this point to avoid the added complication of including a GPISS (which would otherwise be disrupted by C-terminal tagging).

Ultimately, three candidates were chosen: ESP10, ESP13, and ESP31. All three are parasite-specific, with the bulk of their sequences predicted to be extracellular (Gadelha *et al.*, 2015). ESP10 is a large type 1 transmembrane protein which localises to the flagellar pocket. When tagged at its endogenous locus with GFP, the majority of the signal is seen in the flagellar pocket with very little endosomal signal, indicating there may be a low level of recycling of this particular protein. This would leave individual molecules exposed on the surface (and, therefore, the immune system) for longer, highlighting ESP10 as an ideal flagellar pocket candidate for vaccination tests. ESP13 is a type 1 transmembrane protein that localises to the whole cell surface. ESP31 is a multi-pass transmembrane protein that localises to the FP and the flagellum (Whipple and Gadelha, personal communication). Its codon sequence has five

transmembrane domains, but as stated, most of its length is predicted to be extracellular.

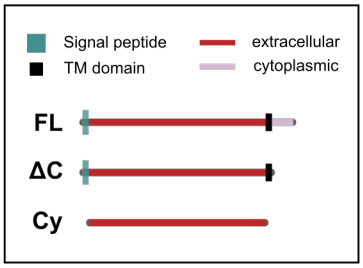
These three proteins represent a subset of potential surfeome components that we wish to exploit in vaccination studies. The surfeome may contain other proteins that present as equally promising candidates for these studies. Thus, the lab intends to tag further members of the surfeome to confirm specific localisations and identify other vaccine candidates. However, at present, these three proteins fulfil the above selection criteria and are therefore good candidates for pilot vaccination experiments. In addition, the inclusion of type 1 and multi-pass transmembrane proteins enables testing of the robustness of the LEXSY system for expression of different membrane topologies (although this is not a primary aim). Finally, an ISG65 was also included in this initial set as a positive control for recombinant protein expression, as it has previously been recombinantly expressed in *L. tarentolae* (Rooney *et al.*, 2015).

For vaccination, an important part of a membrane protein is the extracellular domain. As such, the candidates listed above were to be expressed in LEXSY as three different versions, to test which would be the best option for recombinant expression (Figure 4-6): full length protein (FL); a membrane-bound version lacking its cytoplasmic C-terminus (ΔC); a cytoplasmic version lacking its signal peptide and its transmembrane domain to C-terminus (Cy). The cytoplasmic version in not expected to go through the ER and be glycosylated or folded by ER chaperones. The multi-pass ESP31 was to be expressed as full length only.

All of the above variations were amplified from *T. brucei* genomic DNA (*Trypanosoma brucei* Lister 427), cloned into pLEXSY_IE-sfGO-N and transfected into T7TR. After selection in G418, the recombinant cell lines were induced with 10 µg/mL tetracycline for 24 h. As before, whole cell lysates were made and resolved by SDS-PAGE, and probed for by immunoblot using anti-GFP antibodies (Figure 4-6A). Other than for rISG65Cy, the level of protein expression was not dependent upon the form produced (although rESP13FL appears to have had more sfGFP cleaved off of it compare to the ΔC and Cy forms). rISG65Cy showed an increase in amount but also formed a smeary band, indicative of possible degradation by the proteosome. Each of the other

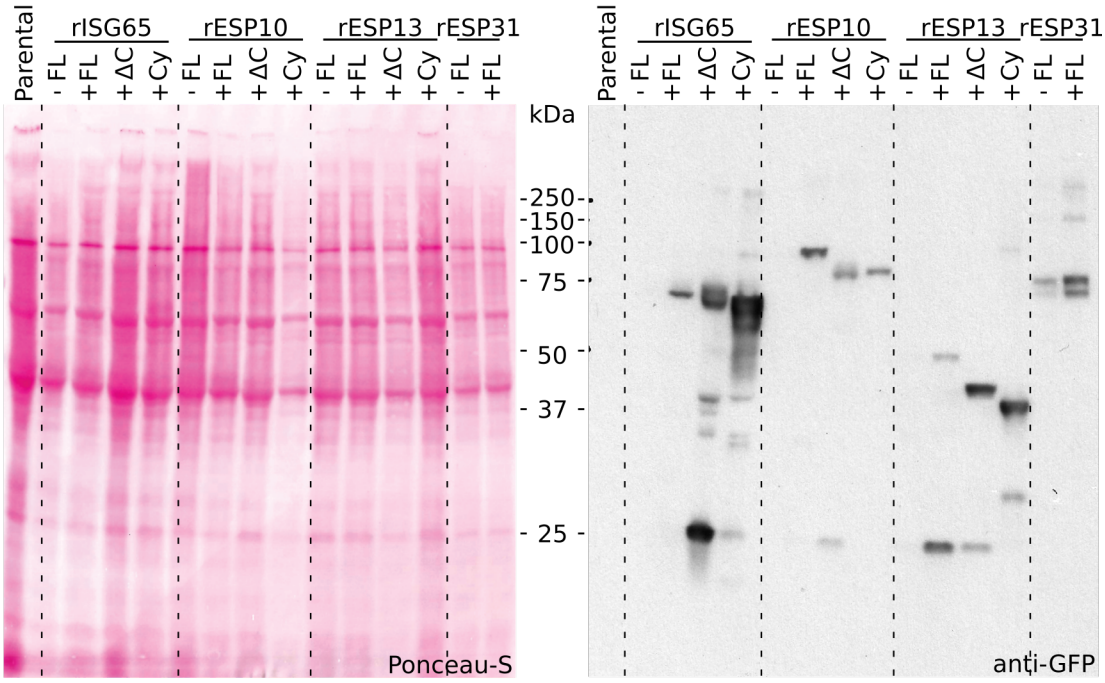
recombinant proteins all expressed to relatively similar levels. These levels were quantified by resolving rESP10FL next to known concentrations of eGFP (Figure 4-6B). Based upon this blot, rESP10FL amounted to ~500 µg/L before purification.

A

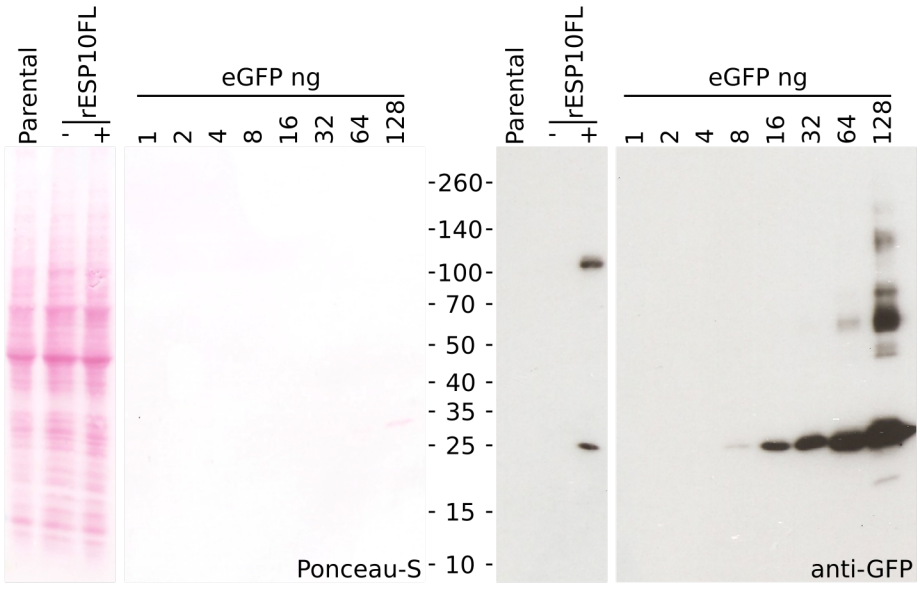


	MW (kDa)		
	FL::sfGFP	ΔC::sfGFP	Cy::sfGFP
ISG65	74	71	69
ESP10	109	101	98
ESP13	55	49	46
ESP31	97	n/a	n/a

B



C



$5x10^6$ cells/lane = 13 ng rESP10FL
∴1L cells @ $2x10^8$ cells/mL = 500 µg rESP10FL

Figure 4-6 T7TR cell lines recombinantly express membrane-bound *T. brucei* surface proteins to relatively low levels. Ponceau-S stained membrane and anti-

GFP immunoblot of whole cell lysates of cell lines expressing rIGG65, rESP10, rESP13 or rESP31 in FL, ΔC or Cy forms. **A** Left panel shows structural schematics of FL, ΔC and Cy. Table contains the predicted molecular weights of each of the fusion proteins. **B** The immunoblot shows bands at the expected size for the recombinant proteins in their different forms. The form of protein produced did not have a large effect on protein levels. **C** Ponceau-S stained membrane and anti-GFP immunoblot of whole cell lysates of the rESP10FL cell line. Known quantities of pure eGFP were loaded for quantification of rESP10FL, indicating its expression to be approximately 500 µg/L. Cell lines in **B** and **C** grown with (+) or without (-) tet for 24 h.

4.2.4. rESP10FL purification attempts

As previously estimated, we would require approximately 1 mg of recombinant protein for each vaccination study. Taking into account possible protein loss during purification steps, the amount generated prior to purification should ideally be closer to 2-3 mg. This would require growth of 5-10 litres of *L. tarentolae* culture, followed by detergent lysis to extract recombinant proteins from membranes prior to affinity purification. Towards this, detergents were selected to test the extraction of rESP10FL as a representative of the candidates expressed thus far.

To preserve the structure and conformation of the extracted recombinant proteins, mild non-ionic detergents were selected as they do not tend to disrupt protein-protein interactions and therefore are less likely to denature proteins. NP-40 (IGEPAL CA-630) was selected as it was the detergent originally used to extract *T. brucei* proteins in the surfeome study (Gadelha *et al.*, 2015). Triton X-100 was selected as it has routinely been used with kinetoplastid proteins. Octyl glucoside (OG) was also selected; OG is another mild, particularly small detergent. It has a high critical micelle concentration with a low aggregation number, allowing it to form smaller micelles that are easy to remove via dialysis. OG is commonly used in crystallisation studies to extract fully folded membrane proteins (Birch *et al.*, 2018; Stetsenko *et al.*, 2017). A small-scale experiment was set up to test recombinant protein solubilisation from *L. tarentolae* using the selected detergents. Figure 4-7A depicts a flowchart of the test experiment. It consists of growing cells for 24 h under tet induction, followed by treatment with a detergent solution containing protease inhibitors and DNAseI, and fractionation into soluble and insoluble fractions. Fractionation of the recombinant protein is followed by anti-GFP immunoblot.

The initial experiment tested all three detergents at 1% w/v, lysing cells for 5 minutes at 0°C (Figure 4-7B). Under these conditions, OG was the worst at extracting rESP10FL. The poorer performance of OG may have been down to its concentration, as its critical micelle concentration is ~0.7% w/v, and therefore 1% is not in great excess of this (elsewhere when used with trypanosomatids OG has been used at 2%). However, when tested again at 2%, its performance did not improve (data not shown).

NP-40 extracted the greatest amount of rESP10FL into the soluble fraction, with ~25% remaining insoluble. To improve upon this, a second experiment was carried out, using NP-40 at 2% w/v and lysing cells for either 5 or 30 minutes at 0°C (Figure 4-7C). The concentration did not appear to increase the amount of soluble rESP10FL relative to whole cells. Leaving to extract for a longer time did increase the amount of soluble protein. However, it should be noted that the amount of protein aggregation seen in the well of the gel also increased.

The experiments showed NP-40 to be a viable option for extracting recombinant *T. brucei* surface proteins from *L. tarentolae* membranes. Based on the observed results, a 1% solution would be used to lyse for 30 minutes (although this could likely be further optimised). Despite recovering 75% of the recombinant protein into the soluble fraction, the amount of cell culture required for each of the recombinant proteins remains at several litres. The next chapter explores an alternative approach to improve protein recombinant protein yield.

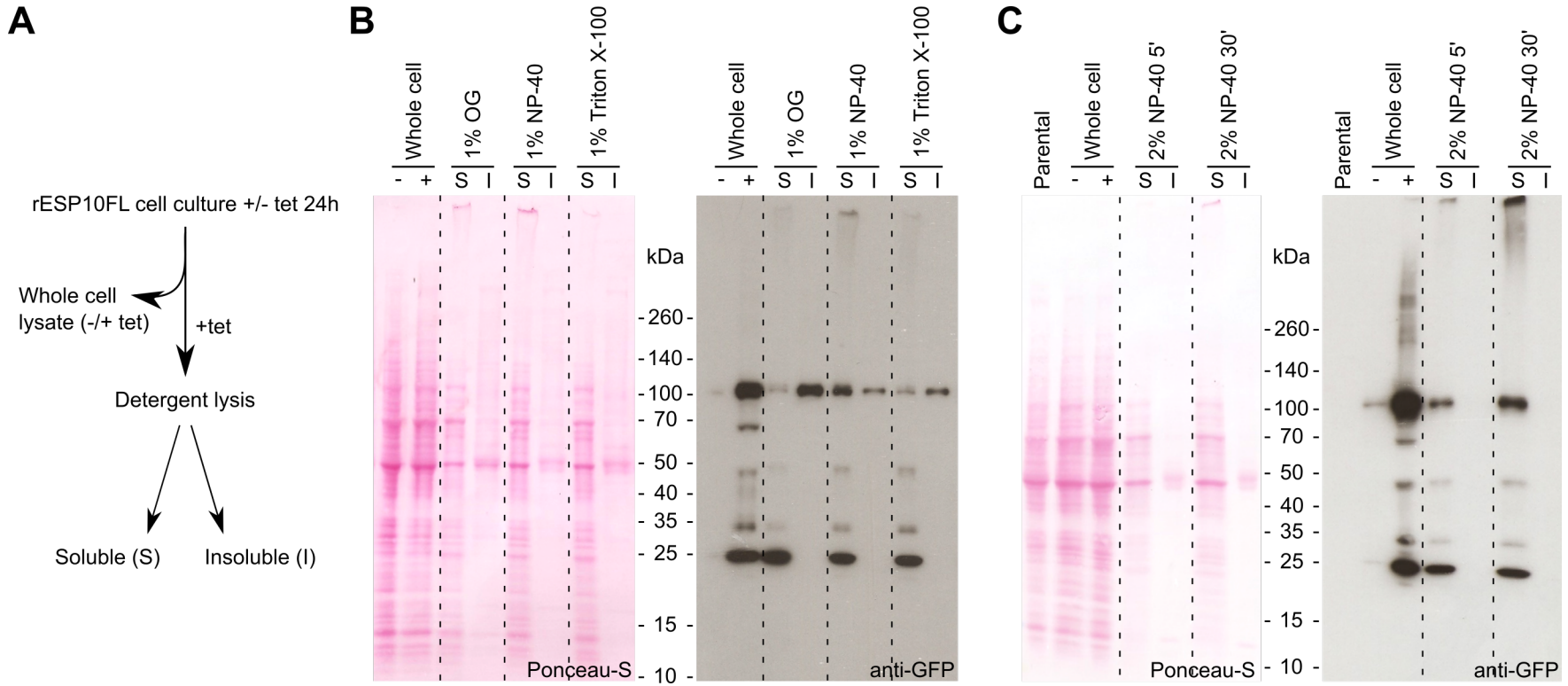


Figure 4-7 NP-40 extracts ~75% of rESP10FL from *L. tarentolae* membranes. A Flow of detergent lysis experiment. **B** Ponceau-S stained membrane and anti-GFP immunoblot of detergent lysis experiment, comparing protein extraction with 1% of OG, NP-40 or Triton X-100. Lysate is split into soluble (S) and insoluble (I) fractions. The rESP10FL band appears just above 100 kDa (predicted MW = 109 kDa). NP-40 extracts ~75% rESP10FL into the soluble fraction. OG and Triton X-100 leave the majority of rESP10FL in the insoluble fraction. **C** Ponceau-S stained membrane and anti-GFP immunoblot of the detergent lysis experiment testing NP-40 at 2% concentration and extracting for different lengths of time. 2% NP-40 does not appear to improve extraction but lysing for longer leads to a stronger GFP signal in the soluble fraction. Cell lines in **A** and **B** grown with (+) or without (-) tet for 24 h.

4.3. Discussion

This chapter describes the establishment of the LEXSY recombinant protein expression system in our lab. LEXSY was easy to handle and manipulate. Initially it was used to test the effect of codon optimisation in expression of the reporter protein sfGFP. It was shown that a codon-optimised sequence led to a substantial increase in protein levels when compared to a bacterial-optimised sequence. This agreed with prior studies in other trypanosomatids, whereby codon bias can affect protein amount (de Freitas Nascimento *et al.*, 2018; Jeacock *et al.*, 2018). The choice of designing the *L. tarentolae* codon-optimised sequence based on the ribosomal codon bias may have played a part in the large relative effect observed here. Indeed, Freitas Nascimento *et al* found ribosomal proteins to have a particularly high geCAI, despite not being used in the original calculation of codon weights (due to the fact that they are multicopy genes). In spite of the good expression levels shown, codon optimisation did also reveal the leakiness of the system. Whilst this did not cause issues here, it should be noted as it could cause problems if toxic recombinant proteins are expressed.

When recombinantly expressing *T. brucei* surface proteins, the yield was lower than for sfGFP^{op} on its own. These recombinant proteins were amplified from *T. brucei* gDNA and therefore not codon optimised for *Leishmania*. This might explain their lower levels. However, codon optimisation is not a guarantee of success. Some organisms have been shown to use poor codons for both optimisation of protein secretion and optimal protein folding (Quax *et al.*, 2015). In yeast it was shown that non-optimal codon cluster of 35-40 codons are frequently used downstream of the SRP recognition site in transmembrane proteins (Pechmann *et al.*, 2014). It was suggested that this was to allow more time for the SRP to bind to the protein as it emerges from the ribosome exit tunnel. This would allow efficient translocation into the ER and reduce the chance of misfolding in the cytoplasm. Similarly, poor codons have been linked with particular protein structures (Pechmann and Frydman, 2013). For example, alpha helices were shown to have enriched poor codon choice at specific positions in the helix, likely to enable efficient correct folding as proteins emerge from the ribosome.

A previous study showed codon optimisation not to be required to recombinantly express *T. brucei* surface proteins to high levels in *L. tarentole* (Rooney *et al.*, 2015). In the study, two different VSGs and an ISG65 from *T. brucei. gambiense* were recombinantly expressed in *L. tarentolae* cells. Using *T. b. gambiense* sequences, 10 mg of each recombinant protein was purified per litre of cell culture (20-fold higher yield than observed here). There are a few possible reasons for the differences in protein levels seen by Rooney *et al* and the levels seen in this chapter. The rISG65 that Rooney *et al* expressed came from *T. b. gambiense,* whilst the rISG65 described in this chapter derived from *T. b. brucei*. Although there are some differences between the amino acid sequences of these two proteins, overall they are very similar, and this is unlikely to be the reason for the difference seen. Instead, it could be due to one of the three main differences in the expression strategies used (see below).

Firstly, Rooney *et al* used a different LEXSY vector; one that integrates into the ssu rDNA locus of the genome, relying on constitutive PolI readthrough to express the recombinant proteins instead of the T7RNAP. However, this is unlikely to have caused the difference in levels, as the T7RNAP is highly processive and successfully expressed a large relative amount of sfGFP^{op} in these cells.

Secondly, Rooney *et al* expressed recombinant proteins with no GFP fusion. Whilst lack of a sfGFP^{op} fusion could be the reason for the higher levels, the fast folding kinetics of sfGFP^{op} coupled with its previous use in tagging endogenous *T. brucei* surface proteins implies that this may not be the case (Gadelha *et al.*, 2015; Pédelacq *et al.*, 2006).

Thirdly, these proteins were expressed with no membrane domains but with their signal peptides and so were secreted from the *L. tarentolae* cells into the culture medium. Membrane-bound proteins can be hard to overexpress as the hydrophobic nature of their transmembrane domains can lead to protein aggregation (Gutmann *et al.*, 2007). In addition, there is only a finite amount of surface area in a membrane, and this could limit the amount of overexpression possible. Indeed, in this work, expression of cytoplasmic rISG65 did show higher levels relative to the membrane-associated forms. However, this was not the

case for rESP10 and rESP13. These disparate results suggest that the increased protein levels reported by Rooney *et al* may indeed by down to loss of membrane association, but that it might also be protein specific.

Another possible reason that *L. tarentolae* did not express these proteins to higher levels could be due to fundamental differences in its secretory pathway. *L. tarentolae* lacks the ER chaperone calreticulin (Raymond *et al.*, 2012). As previously discussed, this chaperone plays a role in the folding of N-glycosylated proteins, and it is found in all other trypanosomatids, including *Leishmania*. This implies that *L. tarentolae* may differ to other trypanosomatids in some of the vital machinery of the secretory pathway. This did not cause a problem for Rooney *et al* but, as discussed, this could be down to the specific proteins they chose to express. If this is the case, then more success could be achieved by establishing a new expression system in an alternative non-human infective trypanosomatid.

Despite the promising results by Rooney *et al*, the evidence that the results may be protein specific, the possible differences between *L. tarentolae* and other trypanosomatids, and the problems with the design of pLEXSY_IE-egfp-red-neo4 regarding its leakiness, suggest that LEXSY may not be a good option for expression of recombinant *T. brucei* proteins. Whilst the results of this chapter show that recombinant *T. brucei* membrane-bound proteins can be expressed in *L. tarentolae*, to obtain the quantities required for vaccination, 5-10 litres of culture would be required. In light of this, an alternative option was explored, namely an attempt to set up a recombinant protein expression system in an alternative trypanosomatid – *C. fasciculata*. This will be covered in the next chapter.

Chapter 5. Establishment of a *Crithidia fasciculata* expression system – CExSy

5.1. Introduction

Chapter 4 highlighted various difficulties in the use of LEXSY. Other trypanosomatid research labs were also unsuccessful in using LEXSY for recombinant protein production (Theimann *et al*; Barret *et al*, personal communications). Theimann's group (São Carlos Institute of Physics, University of São Paulo) attempted to express nuclear proteins (Rad1, Rad9 and Hus1) with LEXSY, trying expression from all of the different LEXSY vector kits commercially available, with very little to no recombinant protein obtained. In trying to solve this problem, they turned to Jena Bioscience for help. However, the LEXSY team at Jena Bioscience also failed to produce these nuclear proteins in their lab in Germany. The Barret lab (School of Life Sciences, University of Glasgow) also attempted to express a carboxypeptidase in LEXSY, but with little success. These drawbacks, coupled with the results of Chapter 4, suggest that LEXSY may not be the best option for recombinant expression of trypanosomatid proteins (although it may still be a suitable system for not-protist proteins).

Given the above, a different approach was pursued – that of developing an entirely new expression system for trypanosome membrane proteins. Some of the features of LEXSY were still desired in this new approach, namely a closely-related, monogenetic species which is easy to grow, genetically tractable and has a fully sequenced genome. However, when selecting a new species, particular attention was paid to two important aspects. Firstly, validation of potential model organisms through extensive use in different labs and studies, providing reliable knowledge about culturing and transfectability. Secondly, capacity of the model organisms to process and glycosylate proteins like *T. brucei*.

Figure 5-1 shows a phylogenetic tree of the trypanosomatids, with the main human infective species in red. *Crithidia fasciculata* is an insect only parasite

that can be grown in cheap medium to high densities (>10^8 cells/mL) and has a doubling time of less than 4 hours (Gadelha, 2005 PhD thesis). Its genome has been fully sequenced by the Beverley group (Washington University) and is available at TriTrypDB (Aslett *et al.*, 2010). *C. fasciculata* has been well studied as a model for kinetoplastid biology (Comini *et al.*, 2005; Filosa *et al.*, 2019; Gadelha *et al.*, 2005; Kipandula *et al.*, 2017), and it has been used as a system for heterologous expression (Tetaud *et al.*, 2002). Importantly, *C. fasciculata* encodes the folding chaperone calreticulin (unlike *L. tarentolae*), and it has been shown to be able to carry out oligomannose-type glycosylation (Parodi, 1993). A rather convenient aspect of *C. fasciculata* biology is that it can survive in cold temperatures for long periods of time, establishing long term storage at 4^{o}C. Whilst not a necessity, this feature makes handling and storage of the organism much simpler. Based upon the above, *C. fasciculata* was selected as a trypanosomatid species that could potentially be turned into a recombinant expression system.

This chapter describes the creation of a *C. fasciculata* cell line suitable for transgene expression and an accompanying suite of molecular tools for recombinant protein expression.

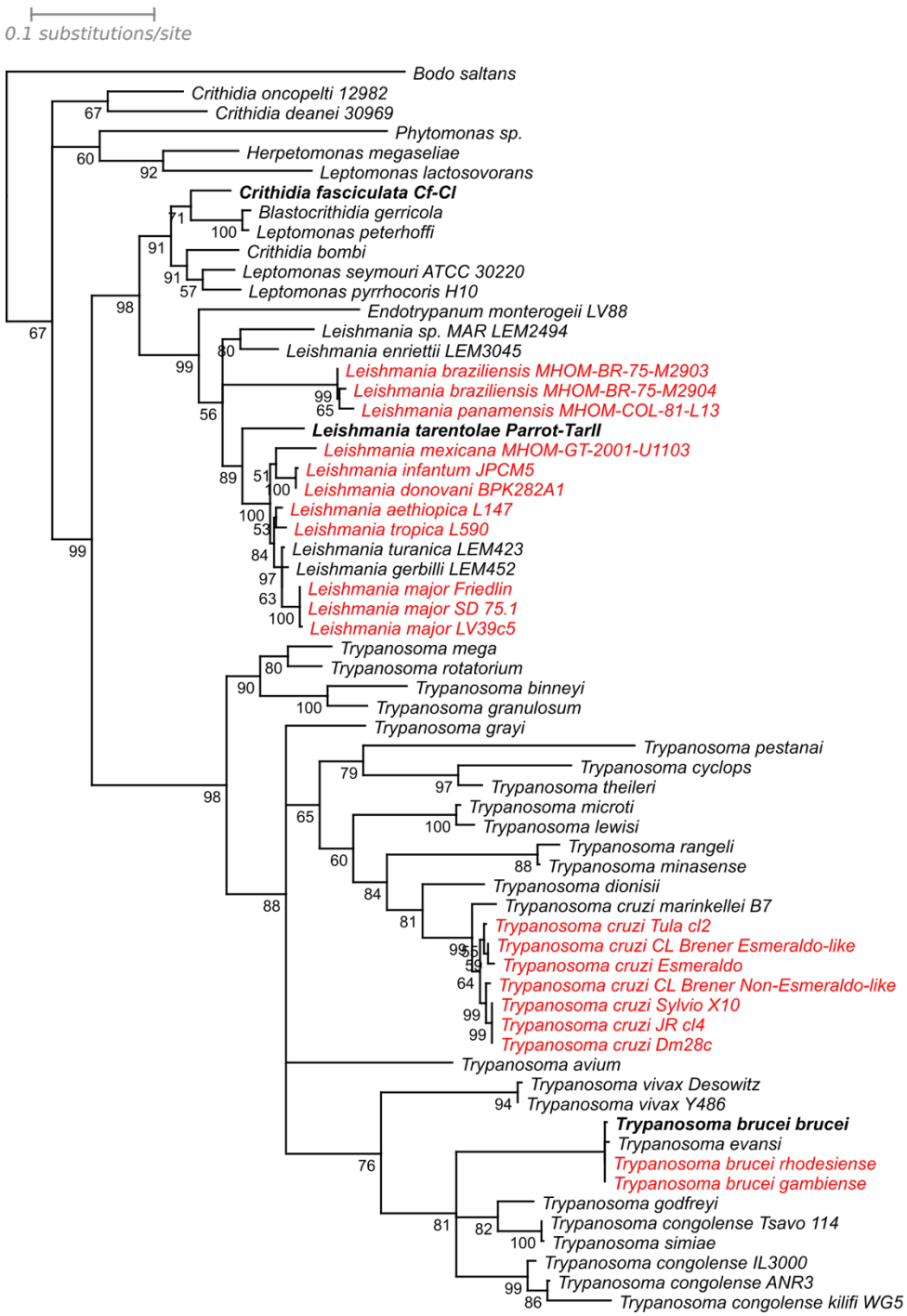
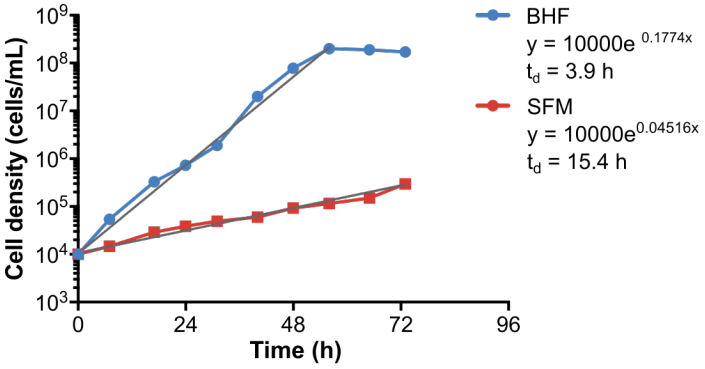


Figure 5-1 Phylogeny of the Trypanosomatida. Shown is a maximum likelihood tree based on the alignment of GAPDH genes. Node numbers are support values from 100 pseudoreplicates of the analysis. Highlighted in red are human infective species. In bold are the three organisms used in this thesis – *C. fasciculata, L. tarentolae* and *T. brucei*. Figure kindly provided by Bill Wickstead.

5.2. Results

5.2.1. Characterisation of *Crithidia fasciculata*

C. fasciculata is routinely grown in BHI medium supplemented with either FBS (5%) or haemin (2.5–20 µg/mL) (Biebinger and Clayton, 1996; Gadelha *et al.*, 2005; Schnare *et al.*, 2000; Tetaud *et al.*, 2002). It has also been reported to grow in a serum-free defined medium (Kipandula *et al.*, 2017). Maximum attainable growth and use of cheap medium are important for large-scale culture. Thus, both BHI medium (supplemented with FBS and Haemin) and Kipandula's serum-free medium (media referred to here as BHF and SFM respectively, see Figure 5-2 for compositions) were tested for culturing *C. fasciculata*. The *C. fasciculata* clone used in this study is derived from Gadelha *et al.*, 2005. To asses growth, cultures were inoculated at 10^4 cells/mL and monitored across 3 days (Figure 5-2). Cells in BHF reached stationary phase with a top density of $2x10^8$ cells/mL. Doubling time during exponential phase was 3.9 h. Cells in SFM did not reach stationary phase within the time period of the experiment, only reaching $3x10^5$ cells/mL with a much slower doubling time of 15.4 h. From here onwards, BHF was used to culture *C. fasciculata*.



BHF	
BHI	37 g/L
Haemin	2.5 mg/L
FBS	5% (v/v)

Serum free medium (SFM)	
Yeast extract	5 g/L
Tryptone	4 g/L
Sucrose	15 g/L
Triethanolamine	4.4 /L
Tween 80	0.5% (v/v)
Haemin	10 mg/L

Figure 5-2 *C. fasciculata* grows to $2x10^8$ cells/mL with log-phase doubling time of 3.9 h in BHF medium. Growth assessment of *C. fasciculata* over three days in either

BHF (blue line) or SFM (red line). Media composition shown beneath the graph. Log-phase growth fitted to calculate doubling time (grey line). Equation of the fit shown, where y is cell density and x is time. In BHF, cells grew with a doubling time of 3.9 h to a top density of $2x10^8$ cells/mL. In SFM, cells grew with a doubling time of 15.4 h. Stationary phase was not reached by cells in SFM during this experiment.

To ascertain which antibiotics could be used to select for transgenic *C. fasciculata* lines, six drugs routinely used in trypanosomatid research were tested at a range of concentrations – Hygromycin B, G418, phleomycin, puromycin, blasticidin and nourseothricin. The initial test was aimed at finding the antibiotics with the strongest selection effects and the range within which those could be used in culture. A $\sqrt{10}$ dilution series of the different drugs was set up, from 2.5 to 250 µg/mL, in 96-well plates (Figure 5-3A). Cells were seeded at 10^5 and 10^6 cells/mL. This relatively high starting density would allow observation of which drugs select particularly quickly, as cultures of *C. fasciculata* become turbid when they exceed 10^7 cells/mL. The experiment was visually monitored for 12 days. Perhaps unsurprisingly, the three drugs with previous reported use in *C. fasciculata* were the strongest selectors – Hygromycin B, G418 and phleomycin (Tetaud *et al.*, 2002). For Hygromycin B at 80 and 250 µg/mL, no culture medium turbidity was seen across the course of the experiment; complete cell death was observed for both concentrations after 6 days. A similar result occurred for G418, although some growth was seen at 80 µg/mL, and complete cell death at this concentration not observed until day 8. Phleomycin performed less well than the other two drugs, only selecting strongly at 250 µg/mL, yet with some turbidity in the 10^6 well at that concentration. Whilst cell growth had clearly been slowed in this well, visual inspection down the light microscope revealed unhealthy rounded cells that were still alive by day 12.

As the two drugs with the strongest selection pressure, Hygromycin B and G418 concentrations were fine-tuned in a subsequent experiment (Figure 5-3B). Five concentrations, from 50 to 250 µg/mL, were used for each drug. As in the previous experiment, complete cell death was observed in all wells after 6-8 days. Based upon the results, it was concluded that individual clones of transgenic lines could be selected with 100 µg/mL Hygromycin B or G418. As Hygromycin B produces the strongest selection pressure, G418 was to be used

for the generation of a parental tet-inducible *C. fasciculata* cell line (section 5.2.2), allowing the stronger selection of Hygromycin B to be reserved for selecting all subsequent recombinant protein expression cell lines.

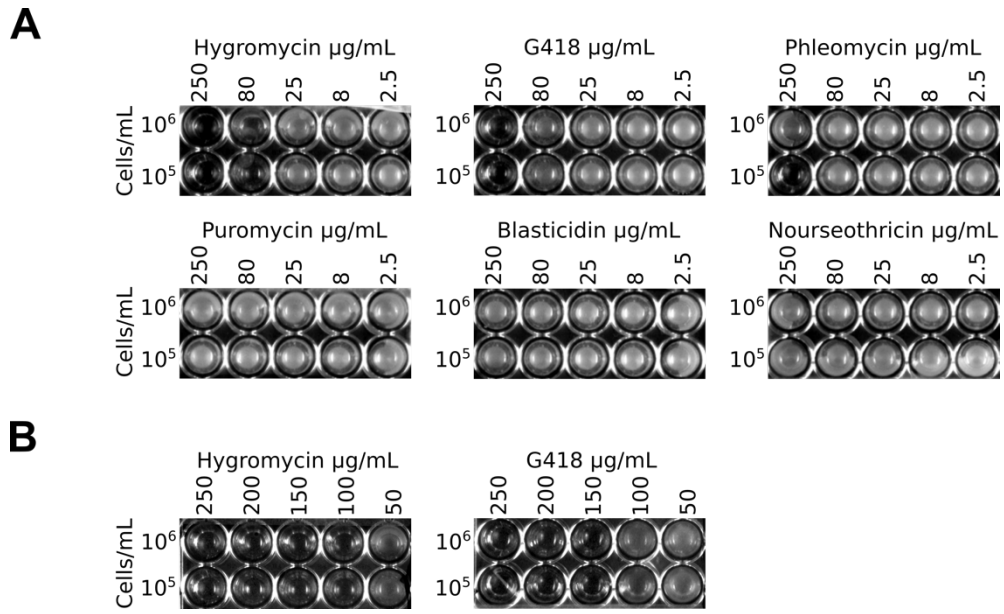


Figure 5-3 *C. fasciculata* is sensitive to Hygromycin B and G418 concentrations above 100 µg/mL. Experiment to find the concentration at which commonly used antibiotics might be used for selecting transgenic *C. fasciculata* cell lines. Cell growth can be seen as turbid wells, whilst cell death appears as clear wells. **A**. A √10 drug dilution series was set up for six commonly used selection drugs. *C. fasciculata* was seeded at relatively high densities to observe which drugs select quickly. The strongest selectors were Hygromycin B, G418 and phleomycin. **B** To fine tune drug concentration, Hygromycin B and G418 were tested at a series of concentrations across the 50–250 µg/mL range. 100 µg/mL was chosen as the optimal selection concentration for both drugs.

5.2.2. Creating a *Crithidia* cell line suitable for inducible transgene expression

A suite of community-resource expression vectors (pNUS and derivatives) already exists for *C. fasciculata* (Tetaud *et al.*, 2002). They allow heterologous expression of proteins in *Crithidia* and *Leishmania* species, with a few different selection markers and fluorescent protein tags. However, as these vectors were designed for protein tagging and functional studies rather than large-scale recombinant protein production, they do not contain promoter sequences or elements for regulation of gene expression. As such, in this project, to enable

inducible expression of recombinant proteins to as high a level as possible, the *T7RNAP* and *TETR* ORFs were introduced into the *C. fasciculata* genome.

When integrating transgenes into trypanosomatid genomes for constitutive readthrough transcription, two details to consider are the integration locus and UTRs. Integration locus dictates which promoter regulates the expression of the transgene(s). Trypanosomatid PolII promoters are thought to be active at similar levels, driving the expression of long polycistronic transcripts prior to processing into individual mRNAs (Clayton, 2019). The individual protein levels are controlled through processing and stability of these individual mRNAs, which is in turn dictated by elements within their UTRs and RNA binding proteins (RBPs). In contrast to this, PolI expression has been shown to be epigenetically regulated. In *T. brucei*, PolI drives not only the expression of the rRNA genes, but also the monoallelic expression of *VSG* in BSF, and *procyclin* in PCF (Günzl *et al.*, 2003). The transcriptional control of these two major surface proteins is essential to the lifecycle of the parasite, and their expression is tightly regulated. Specific transcription from the multiple rDNA loci has also been shown to be epigenetically regulated by the histone acetylase Elp3b (Alsford and Horn, 2011). Through transcription run-on analysis, Elp3b was shown to inhibit elongation of nascent rRNA transcripts as opposed to transcription initiation. Elp3b is also found in *C. fasciulata* and *L. tarentolae*. The rDNA repeat unit of *C. fasciculata* is shown in Figure 5-4.

The LEXSY T7TR cell line discussed in Chapter 4 has the *T7RNAP* and *TETR* transgenes integrated into 18S rRNA genes of rDNA repeat loci (Kushnir *et al.*, 2005). They are integrated into separate copies of the 18S gene, requiring two selection markers. Due to the different integration loci of *T7RNAP* and *TETR*, it is possible that they could be differentially regulated through epigenetic changes (Alsford and Horn, 2011), potentially leading to the leakiness seen in Chapter 4. Indeed, unpublished observations in our lab have seen large variability in expression levels between clones when transfecting constructs that target the rDNA loci.

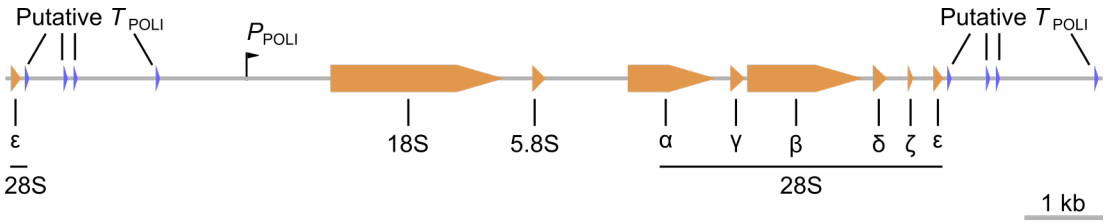


Figure 5-4 The rDNA repeat unit of *C. fasciculata*. The rDNA repeat locus consists of the 18S, 5.8S and 28S rRNA genes. A characteristic feature of trypanosomatid 28S rRNA is its fragmentation into 6 separate genes (Schnare *et al.*, 2000). The transcription initiation site has been mapped (P_{POLI}). The putative terminators consist of a 55-57 bp repeat containing a hairpin structure (T_{POLI}). Scale bar shown at bottom right of figure.

In other studies, the alpha- and beta-tubulin (*TUBA* and *TUBB*) loci have commonly been used as sites for integration of genes to be constitutively expressed by PolII readthrough (Poon et al., 2012; Wirtz et al., 1999). Within trypanosomatids, *TUBA* and *TUBB* genes are arranged in multicopy tandem arrays. These arrays can either be alternating between *TUBA* and *TUBB*, or they can be monotypic arrays of only *TUBA* or *TUBB* depending on the species (Jackson et al., 2006). *Trypanosoma* have alternating arrays, whilst *Leishmania* and related species have monotypic arrays.

One such example of using these loci is the Single Marker Oxford (SMOx) cell line (Poon et al., 2012). They introduced the *T7RNAP* and *TETR,* under a single selection marker, into the alternating array of *TUBB-TUBA* genes of *T. brucei,* creating a tet-inducible cell line. High levels of T7RNAP and TETR were obtained in SMOx relative to other *T. brucei* lines expressing these two proteins (Poon *et al.*, 2012; Wirtz *et al.*, 1999). This was likely due to the UTRs used and the codon optimisation of the *T7RNAP* and *TETR* genes. The UTRs included the *TUBB* UTRs and the paraflagellar rod protein 2 (*PFR2*) UTRs. Using tet-induced GFP expression and flow cytometry, SMOx cells were shown to have tight regulation, with non-induced cells showing GFP signal similar to background, and tet-induced cells showing a 100 to 200-fold increase in GFP signal.

Due to the effectiveness of the SMOx cell line, a Single Marker Crithidia plasmid was designed based on the SMOx stategy, utilising equivalent crithidial UTRs and targeting sequences (pSMC, Figure 5-5A). The *T. brucei* codon-optimised *T7RNAP* and *TETR* from pSMOx were compared to the *C. fasciculata*

ribosomal protein codon bias to assess their suitability for pSMC. They were found to be reasonably well coded for *C. fasciculata* and so would be taken directly from pSMOx. Being more closely related to *Leishmania* species than *Trypanosoma*, *C. fasciculata* contains monotypic tubulin arrays. It has an array of four *TUBA* genes on chromosome 8 and one gene on chromosome 14. Regarding *TUBB*, it has two tandem copies on chromosome 13 and single copies on chromosomes 17 and 28. pSMC was targeted to the *TUBB* intergenic region on chromosome 13 (Figure 5-5B). This allowed use of *TUBB* UTRs for both regulation of mRNA processing and integration locus targeting, without the need to knock-out a *TUBB* gene.

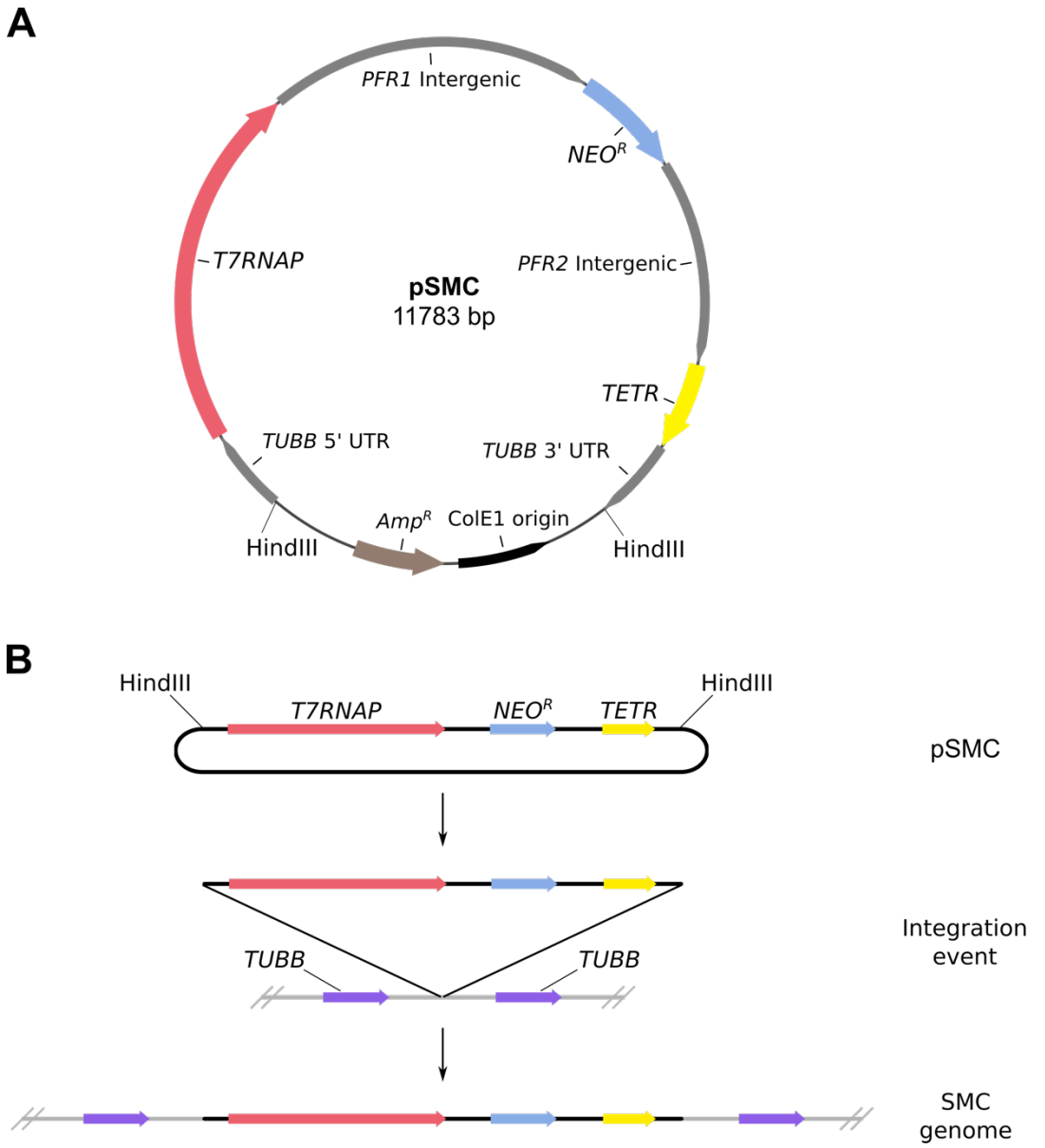


Figure 5-5 The Single Marker *Crithidia* plasmid pSMC for introduction of the *T7RNAP* and *TETR* genes into the *C. fasciculata genome*. A Vector map of the pSMC plasmid. It contains the neomycin resistance gene for selection by G418, and intergenic regions and UTR sequences from *C. fasciculata*. **B** Strategy for integration of pSMC into the *C. fasciculata* genome. HindIII restriction sites are used to linearise the vector before integration into an intergenic region of the β-tubulin (*TUBB*) locus of the genome. *T7RNAP* and *TETR* are transcribed by RNA PolII read-through.

To construct pSMC, a variety of molecular biology techniques were used, including PCR, fusion PCR and Gibson assembly. Fusion PCR and Gibson assembly are summarised in Figure 5-6. Both techniques require overlap of complimentary sequences between different fragments. The overlaps can be introduced by PCR primers. In fusion PCR, these overlaps can be used to allow

two different fragments to prime each other (Figure 5-6A). These fragments can then extend in another round of PCR to form one complete fragment. Combined with primers at the far ends of each fragment, this amplifies the entire combined fragment. Gibson assembly, on the other hand, relies on three different enzymes to do the work in a single isothermal reaction (Gibson *et al.*, 2009); a 5'-3' exonuclease, a polymerase and a ligase (Figure 5-6B). The exonuclease initially digests the DNA to create 3' overhangs. Once digested, the 3' overhangs across the complimentary region of the two fragments can then anneal. The 3' ends are extended by the polymerase and the fragments sealed by the ligase. The advantages of Gibson assembly are that it can combine multiple fragments at once without the need for primers, and that it can combine these fragments into an entire plasmid in one reaction. However, efficiency of correct assembly is inversely related to fragment length and number. This is why fusion PCR was initially used to combine fragments, reducing the number and making Gibson assembly simpler.

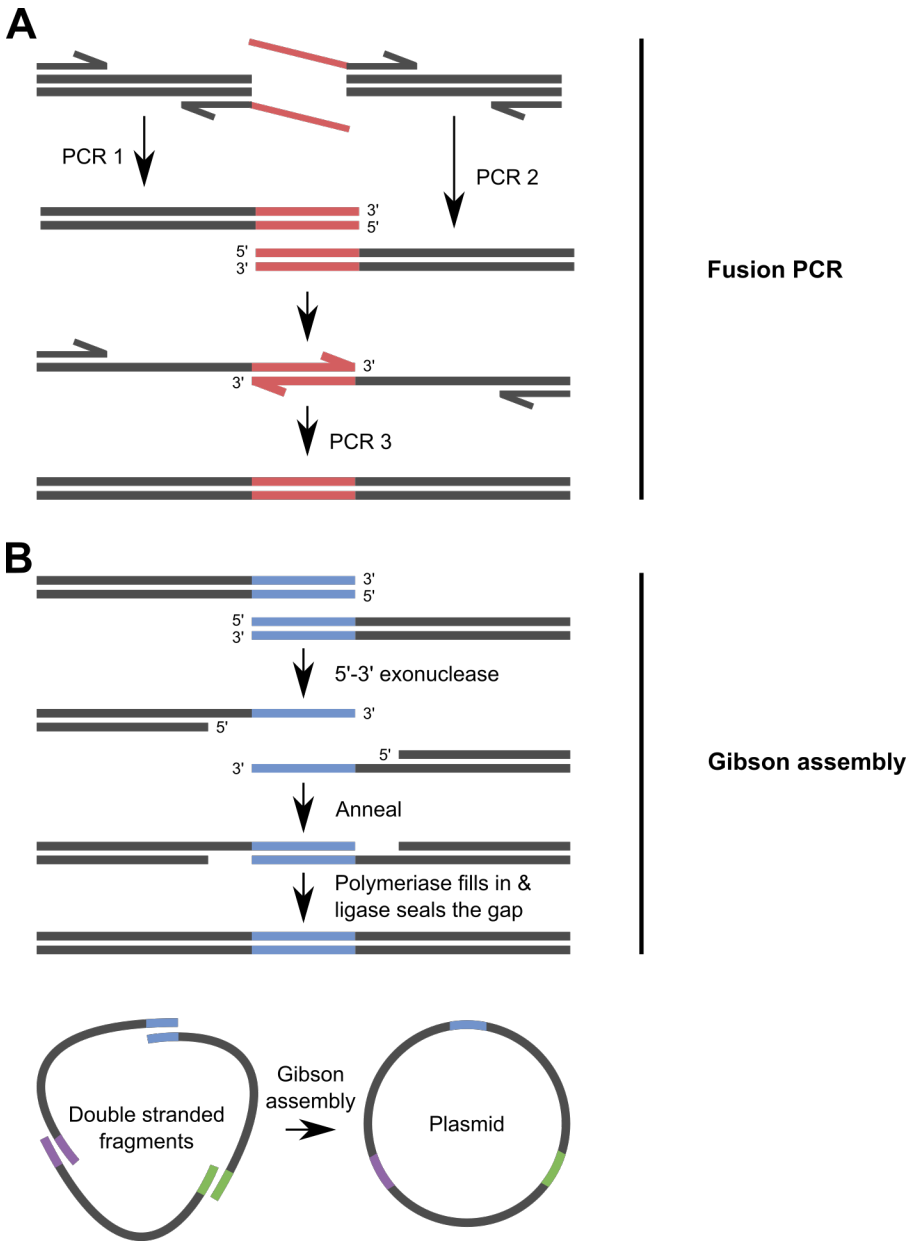


Figure 5-6 Schematics of molecular biology techniques used to construct pSMC **A** Fusion PCR: Two or more DNA fragments are amplified using primers with a complimentary tag (red). The complimentary region introduced by the tag allows the DNA fragments to anneal to each other and extend to form one complete fragment in a second round of PCR. **B** Gibson assembly: Combines fragments with complimentary regions (blue) using a 5'-3'exonuclease, a polymerase and a ligase. The exonuclease digests the DNA to create 3' overhangs. The complimentary regions can then anneal and be extended by the polymerase. Fragments are then sealed by the ligase. Gibson assembly can be used to combine multiple fragments to create a plasmid.

After its construction, pSMC was linearised with HindIII and transfected into *C. fasciculata*. The high voltage transfection protocol described by Jena Bioscience for LEXSY was followed, transfecting cells in growth medium (BHF)

and using an Eppendorf Eporator. Approximately $2.4x10^7$ cells and 10 µg of linearised pSMC DNA were used for transfection. G418 drug selection (100 µg/mL) was applied after 4 h recovery. To assess transfection efficiency, one half and one tenth of the recovery culture were separated into two 96-well plates, and the number of positive wells used to estimate efficiency by Poisson distribution. Based on the plates, the transfection resulted in a total of 6-10 independent clones.

To test for integration of pSMC, multiplex diagnostic PCR was done on three of the SMC clones obtained (strategy depicted in Figure 5-7A). Primers that anneal in *T7RNAP* (encoded in pSMC) and the *TUBB* 3' UTR were used to amplify a product of ~0.8 kb if correct integration had taken place. Another reverse primer was included which anneals in the *TUBB* gene and, in combination with the 3' UTR forward primer, was used to amplify a control product of ~1 kb whether integration had taken place or not. Genomic DNA purified from pSMC transfectants and parental cells was used as template in this multiplex diagnostic PCR reaction. Figure 5-7B shows two of the three clones with a clear integration product. Clone 1, named SMC, was taken forward for further testing. Unless stated, SMC was subsequently grown in 50 µg/mL G418.

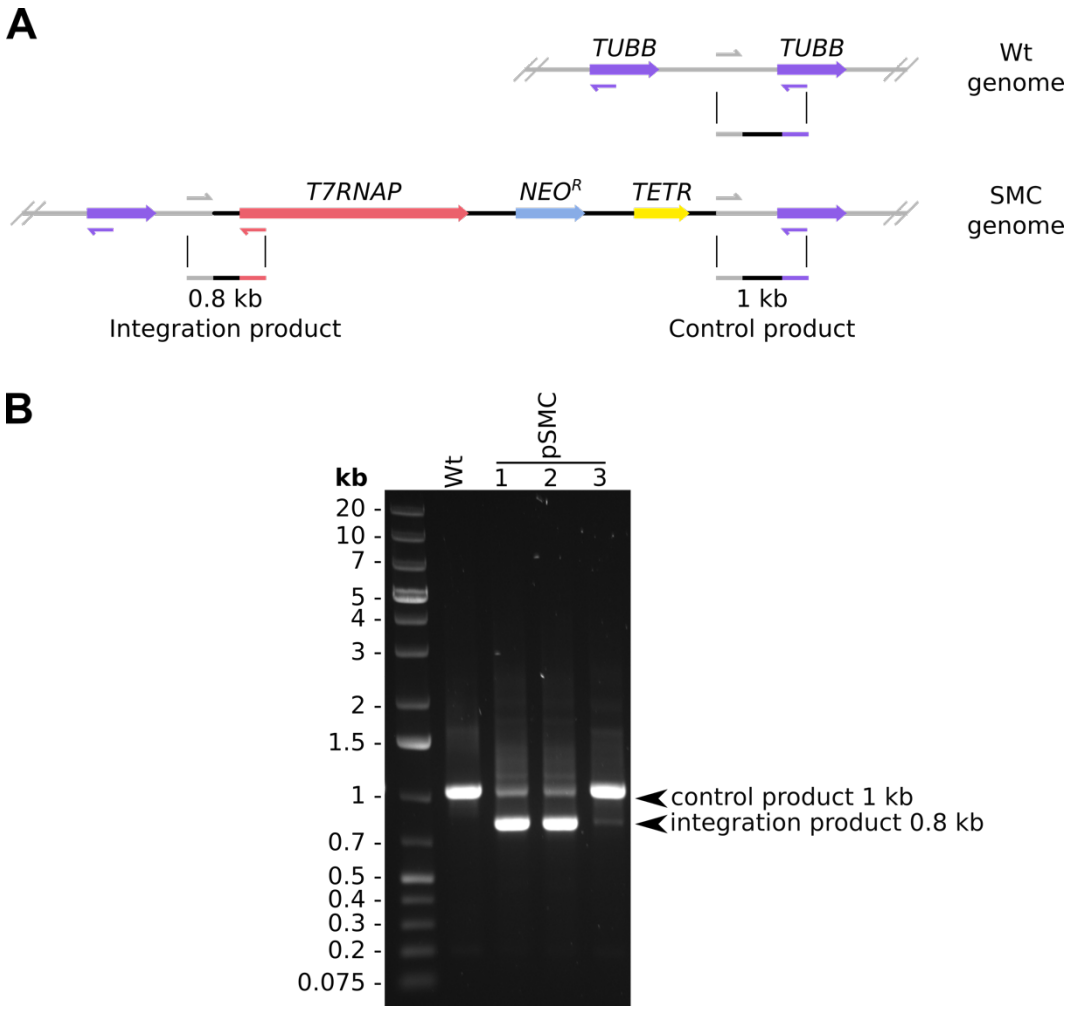


Figure 5-7 Diagnostic PCR shows pSMC to be successfully integrated into the *C. fasciculata* genome. A Schematic illustrating the multiplex diagnostic PCR strategy to identify integration of pSMC into the correct locus. Primers produce a 1 kb control product whether pSMC has correctly integrated or not, and a 0.8 kb product only if pSMC has correctly integrated. **B** Agarose gel electrophoresis was used to resolve diagnostic PCR amplicons from DNA of clones resulting from transfection with pSMC. Clones 1 and 2 show a strong integration product.

To test the inducibility of the newly generated SMC line, pLEXSY_IE-sfGO-N (used in Chapter 4) was transfected into these cells. As pLEXSY_IE-sfGO-N has the same selection marker as pSMC, following transfection the cells were kept as a population under 100 µg/mL G418 selection pressure. The population was split into two 5 mL cultures and grown overnight, either with or without 10 µg/mL tet. These two cultures were then observed by live fluorescence microscopy (Figure 5-8). Parental (untransfected) SMC cells were used as a negative control for background fluorescence. Non-induced cells showed little

GFP signal relative to background, whilst the induced cells showed much stronger GFP signal. This result demonstrates the SMC cell line to be inducible.

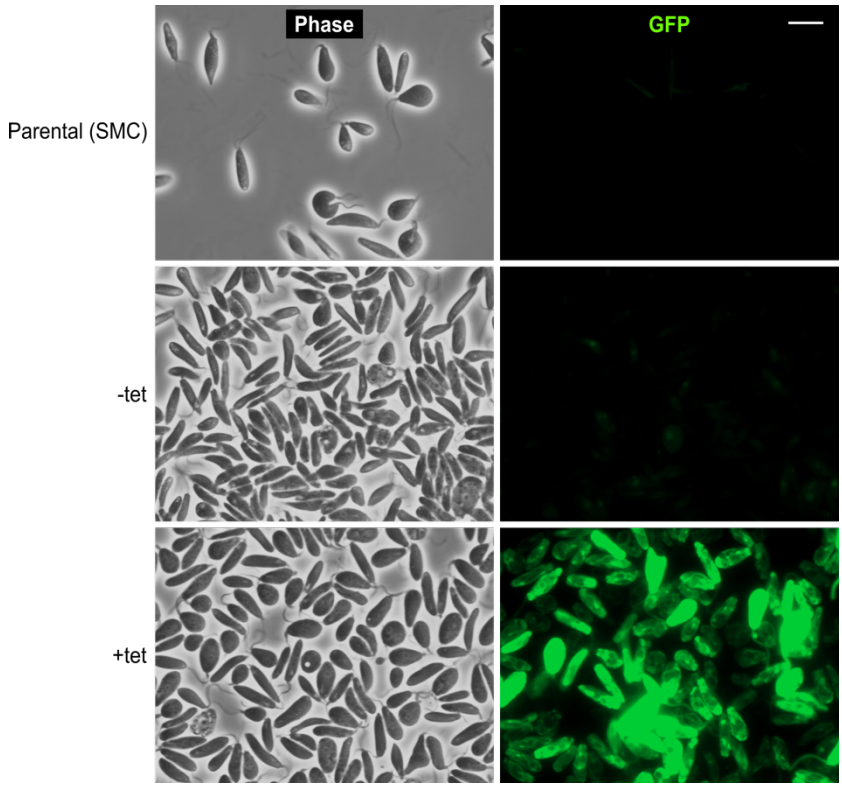


Figure 5-8 SMC cells transfected with pLEXSY_sfGO-N express sfGFP^{op} when grown in the presence of tetracycline. Live native fluorescence microscopy of SMCs transfected with the sfGFP^{op} expression plasmid pLEXSY_IE-sfGO-N. Cells cultured with or without 10 µg/mL tetracycline for 24 h. Cells in the presence of tetracycline show strong green fluorescence signal above background. Micrographs taken of transfected populations, likely explaining the cell-to-cell signal variability. Images were acquired and processed equally. Scale bar = 10 µm.

5.2.3. p**CEx** – a recombinant protein **C**rithidial **Ex**pression vector

For optimal expression of recombinant proteins in SMC cells, a **C**rithidal **Ex**pression vector named p**CEx** was designed (Figure 5-9A). This vector has been designed to be modular, so that any individual component could be easily removed with restriction enzymes, allowing further modification and optimisation of the vector. The initial vector design contained no tet operator, so that optimisation of UTRs could be carried out under a constitutive T7 promoter before subsequent tests of expression regulation (more details in sections 5.2.4 and 5.2.5). The ORF for recombinant protein expression contained TEV-

sfGFPO-His_6 with the same sequence as in pLEXSY_IE-sfGO-N (Chapter 4) as its sequence is also predicted to be optimal when compared to the *C. fasciculata* ribosomal protein codon bias. Importantly, it was ensured that the pre-ATG triplet for these ORFs was retained as ACC, as it has been previously reported in *L. tarentolae* that ACC positively influences eGFP expression, and is one of the most common pre-ATG triplets found in the *L. major* genome (Lukeš *et al.*, 2006). The UTRs chosen for the $sfGFP^{op}$ ORF were the *TUBB* 5' and the *TUBA* 3', two commonly used UTRs of highly expressed proteins.

The selection marker in pCEx is HYG^R, was placed in the antisense orientation relative to the $sfGFP^{op}$ ORF to reduce the chance of leaky transcription of $sfGFP^{op}$ selected for by expression of HYG^R (once tet operators were added to the plasmid). The UTRs used for HYG^R were the same that flank the resistance markers in the pNUS vectors as they have been shown to work; namely the phosphoglycerate kinase B (*PGKB*) 5' and glutathione synthase (*GSPS*) 3' UTRs (Tetaud *et al.*, 2002). A 10% strength T7 promoter was used to drive HYG^R expression (Wirtz *et al.*, 1998).

Regarding integration locus, the potentially silent rDNA intergenic spacer (IGS) was chosen (for full repeat locus see Figure 5-4, for integration strategy see Figure 5-9B). The transcription initiation site (TIS) for the rRNA genes has previously been mapped for this locus (Schnare *et al.*, 2000). In that study, it was also suggested that four 55-57 base pair repeats present upstream of the TIS play a role as the PolI terminators for the upstream rDNA repeat. These sequences are conserved in *Leishmania*, and in *L. infantum* it was shown through transcriptional run-on assays that little to no transcription happens downstream of these repeats (Requena *et al.*, 1997). In addition, it was pointed out that their internal structure is similar to that of bacterial rho-independent terminators, i.e. a hairpin structure followed by a uracil-rich sequence. For these reasons, the region downstream of these repeats but upstream of the TIS was chosen for pCEx integration as a locus likely to have minimal transcription, and thus lower chance of leaky transcription by PolI readthrough. To minimise the risk of leaky expression further, the integration strategy would have $sfGFP^{op}$ integrated in the antisense orientation to the rRNA genes (Figure 5-9B).

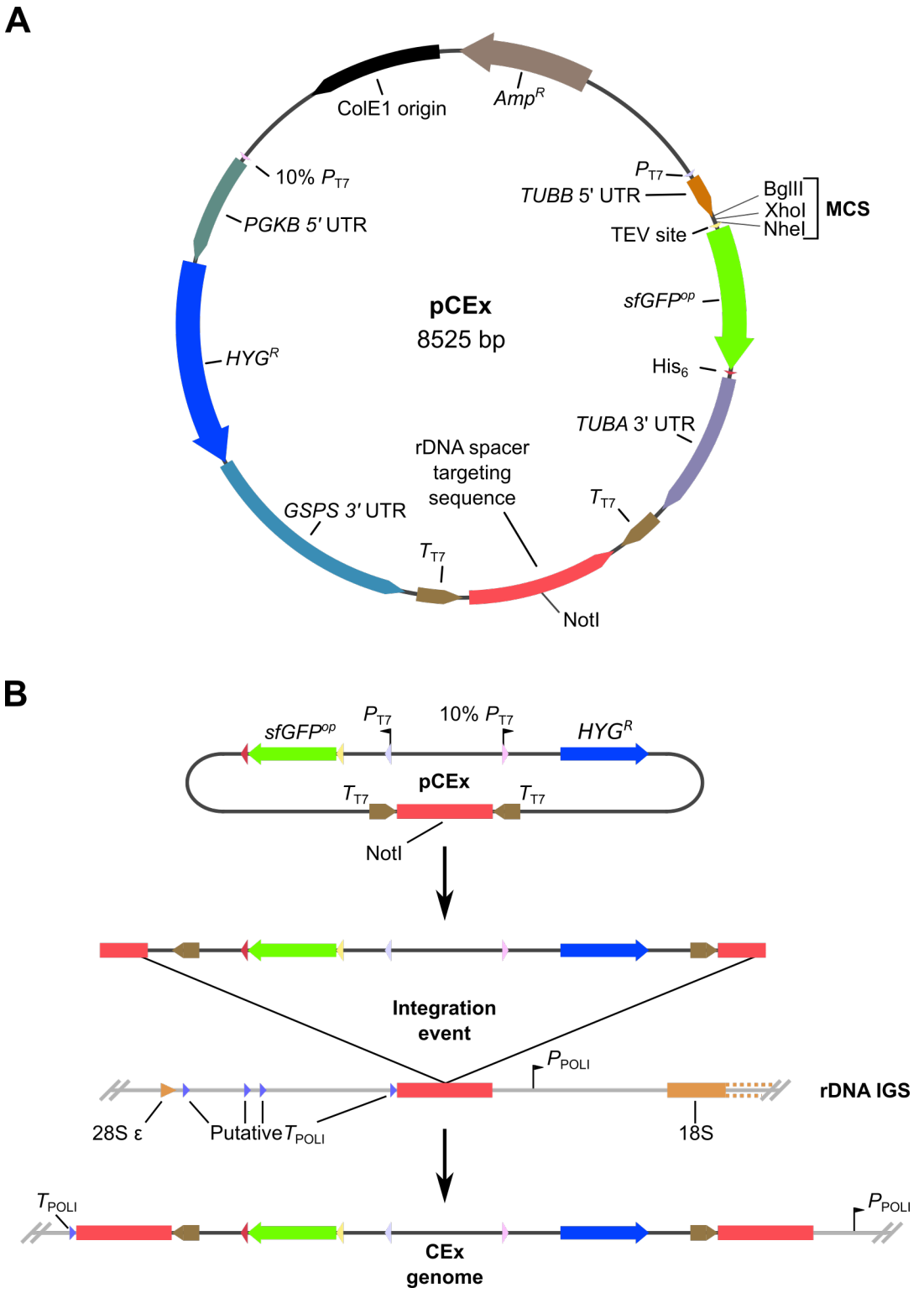


Figure 5-9 The *Crithidia* Expression plasmid pCEx for recombinant protein expression in SMC cells. A Vector map of pCEx. UTRs amplified from *C. fasciculata* genomic DNA. Transgene expression is driven by the T7 promoter. Hygromycin B drug resistance marker is driven by a 10% strength T7 promoter. Multiple cloning site (MCS) and NotI linearisation site for integration are indicated. Gene of interest is cloned in as a fusion with a C-terminal TEV-*sfGFPO*-His_6. **B** Schematics of the integration of pCEx into the rDNA intergenic spacer (IGS) of *C. fasciculata*. Integration targets the theoretically transcriptionally silent region of the IGS.

As for pSMC, pCEx was constructed through a combination of PCR, fusion PCR and Gibson assembly. Whilst pCEx was designed to integrate into the *C. fasciculata* genome if linearised prior to transfection, the vector can be propagated as an episome if transfected as uncut DNA. To test which form leads to higher sfGFP^{op} levels, pCEx was transfected into SMC cells, either circular or linearised with NotI.

Due to the low transfection efficiency seen when transfecting pSMC into *C. fasciculata* using the Jena Bioscience protocol, an alternative protocol was used here, one routinely used in African trypanosomes (Burkard *et al.*, 2011). Approximately 2x10^7 cells were transfected with 10 µg of pCEx DNA in Tb-BSF buffer using an Amaxa Nucleofector 2b device, programme Z-001. Hygromycin B drug selection (100 µg/mL) was applied after 4 h recovery. Once again, transfection efficiency was assessed by plating proportions of transfected populations (one half and one twentieth) into 96-well plates, and the probability of clonal distribution estimated by Poisson. Nucleofection coupled with Tb-BSF buffer resulted in higher efficiency: transfection of circular DNA resulted in 80-90 independent clones, whilst transfection of linearised vector resulted in 96-180 independent clones. Due to the higher transfection efficiency achieved here, all future transfections were done using the method above.

Four clones from each transfection were taken forward to assess relative sfGFP^{op} levels by immunoblotting with anti-GFP antibodies (Figure 5-10). Cells transfected with circular DNA showed similar sfGFP^{op} levels between clones. In comparison, cells transfected with linearised DNA showed lower expression and variability in sfGFP^{op} levels between clones. This could be caused by integration of pCEx into different rDNA loci. However, the variability observed appears to correlate with the sample loading revealed by the Ponceau-S stained membrane.

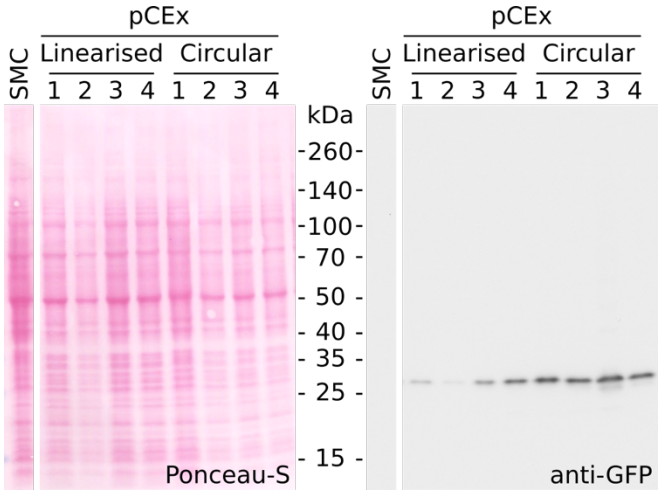


Figure 5-10 SMC cells transfected with pCEx express sfGFP^{op}. Ponceau-S stained membrane and anti-GFP immunoblot of whole cell lysates of SMC clones transfected with pCEx, either circular, or linearised for integration the rDNA intergenic spacer (IGS). Numbered lanes indicate independent clones. Expression levels from cells transfected with circular DNA appear consistent between clones. Levels from clones transfected with linearised DNA appear to vary between clones. This could in part be due to differential sample amount (according to the Ponceau-S loading control). It could also be due to construct integrating into different rDNA IGS loci in different clones.

The result above demonstrates that the newly generated pCEx plasmid successfully drives recombinant expression of sfGFP^{op} in SMC cells. Unless stated, further modifications to the plasmid were assessed using populations transfected with circular DNA (detailed below).

5.2.4. pCExC – optimisation of UTR choice for high recombinant protein expression

In the processing of trypanosomatid polycistrons, the 5' and 3' UTRs play different important roles but can both effect mRNA levels. The 5' UTR is important for both removal of mRNAs from the polycistronic transcript and for mRNA capping. It contains various sequence elements that are recognised by nuclear machinery, leading to the removal of the UTR and the trans-splicing of a 39-nucleotide 5'-capped leader sequence (Siegel *et al.*, 2005). It has been shown in *T. brucei* PCF cells, through a readout of luciferase activity, that variation in the length and relative positioning of these different elements can affect trans-splicing efficiency. In addition, attaching the relevant region of the 5' UTRs from different genes to this luciferase reporter led to differing splicing

efficiencies. This included UTRs from the *PGKA*, *PGKB* and *PGKC* genes which, despite being consecutive on the same polycistron, have different levels of mature mRNA. Consistent with PGKB having the highest mRNA levels in PCF cells, its 5' UTR also drove the highest luciferase levels. Therefore, mRNA levels can be controlled, to an extent, through the efficiency of 5' UTR trans-splicing.

The 3' UTR, meanwhile, is important for polyadenylation. It also contains structural elements that, in combination with RNA binding proteins, regulate the stability of the mRNA. Furger *et al* (1997) showed that the *procyclin* 3' UTR contains both negative and positive regulatory structural motifs that, when mutated, influenced steady-state mRNA and protein levels of the glutamic acid/alanine-rich protein GARP, and enzyme activity (and thus levels) of chloramphenicol acetyltransferase (CAT). The regulation did not occur if the mRNA was transcribed and translated in vitro, implying that other *T. brucei*-derived factors were required for this regulation (Furger *et al.*, 1997). Subsequently, many RBPs have been identified and studied (Clayton, 2019). For example, RBP10 has been shown to bind to UA(U)6 motif in the 3' UTR of many mRNAs specific to PCF *T. brucei* (including *procyclin*), targeting them for destruction (Mugo and Clayton, 2017). Conversely, the RBP TbDRBD3 stabilises PCF specific mRNAs (Estévez, 2008).

The regulation of mRNA levels by RBPs may not be restricted to the 3' UTR. In *C. fasiculata* it has been shown that cycling element binding proteins (CEBPs) bind to the consensus sequence (C/A)AUAGAA(G/A), which can be in either the 5' or 3' UTR, leading to cyclical build-up of proteins involved in DNA metabolism (Mittra and Ray, 2004).

Based upon the evidence that both 5' and 3' UTRs play a role in trypanosomatid mRNA levels, different 5' and 3' UTRs were tested to maximise expression levels attainable from pCEx. The list of chosen UTRs is shown in Figure 5-11A (blue boxes). The *PGKA/B* and *GSPS* UTRs were selected because they have been shown to drive transgene expression in *C. fasciculata* (Tetaud *et al.*, 2002). The *TUBB*, *TUBA*, and *PFR1/2* UTRs were chosen as they are from highly expressed proteins, have previously been used effectively in other trypanosomatids (Poon *et al.*, 2012), and were used successfully in

pSMC (other than *TUBA*). As ribosomal proteins are highly abundant, and their codon bias was shown to have a strong effect on recombinant sfGFP^{op} levels in Chapter 4, it was decided to also test a ribosomal protein UTR.

Four 5' UTRs were individually cloned into pCEx containing the *TUBA* 3' UTR, whilst five 3' UTRs were individually cloned into pCEx containing the *TUBB* 5' UTR. The 9 different constructs were transfected as circular DNA into SMC cells and selected as populations. Relative levels of sfGFP^{op} were assessed by immunoblotting with anti-GFP antibodies (Figure 5-12A). The 5' UTR swaps had a minimal if any effect on sfGFP^{op} levels. In contrast, variable levels were observed for the 3' UTRs. The highest sfGFP^{op} levels were seen when using the *L6* 3' UTR. It caused a substantial increase in sfGFP^{op} relative to the other 3' UTRs. The levels were high enough to be seen on the Ponceau-S stained membrane and were equivalent to >10 mg/L sfGFP^{op} based on comparison with known amounts of eGFP. Therefore, the pCEx variant containing the *L6* 3' UTR (and *TUBB* 5' UTR) was taken forward and renamed pCExC, the C standing for constitutive. The production of sfGFP^{op} by CExC cells was also confirmed by fluorescence microscopy (Figure 5-12B).

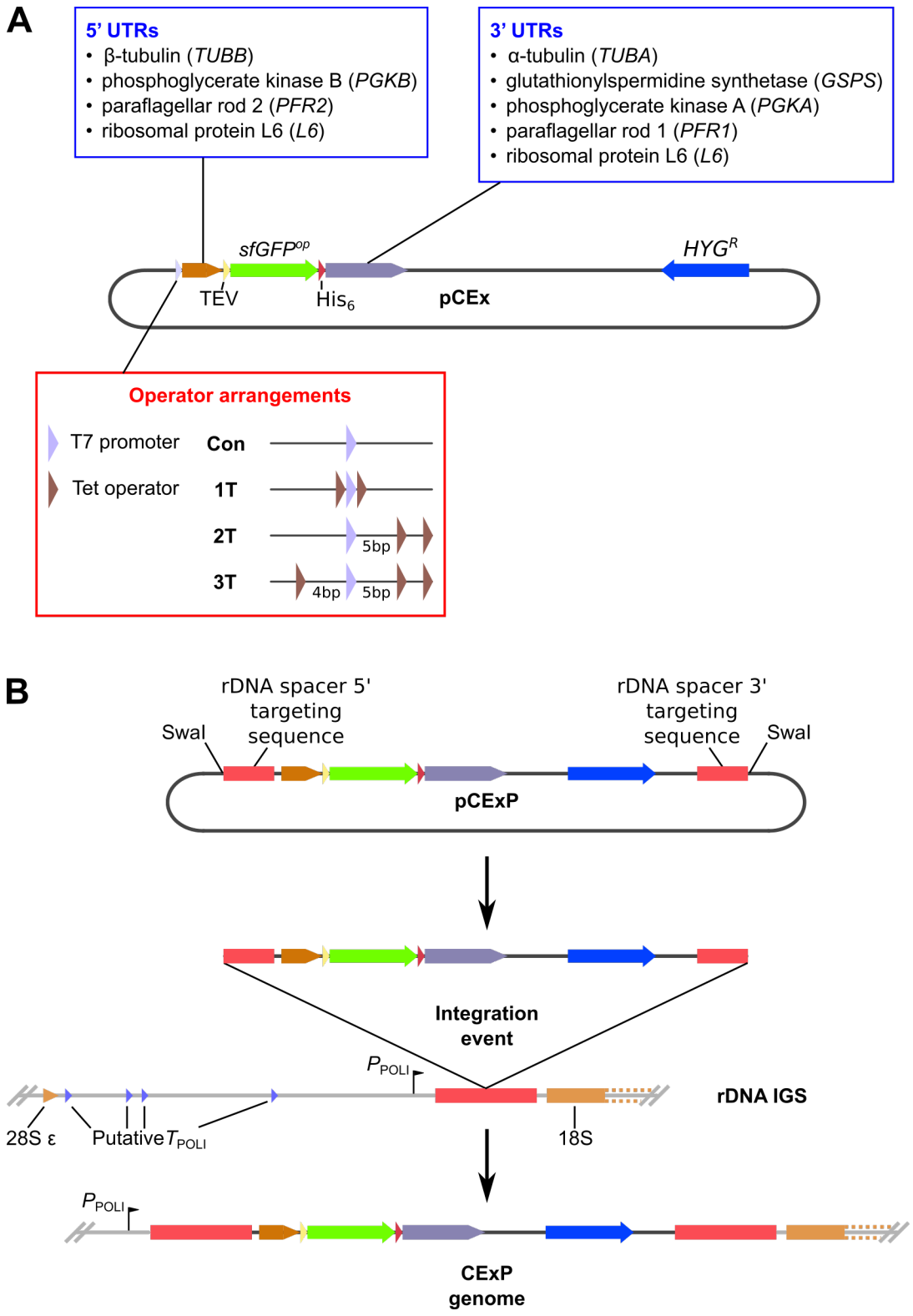


Figure 5-11 Schematics of optimisation of pCEx through modifying the DNA processing sequences and integration locus. A *sfGFP*^{op} expression from pCEx was optimised by independently swapping in different 5' and 3' UTRs of highly abundant proteins. Subsequently, regulation of expression was optimised through testing different numbers and arrangements of tet operators. **B** To test an alternative method of *sfGFP*^{op} expression, pCEx was modified to create pCExP (after UTR optimisation in **A**). pCExP integrates downstream of the PolI promoter in the intergenic spacer (IGS) of the rDNA locus, relying on PolI readthrough transcription.

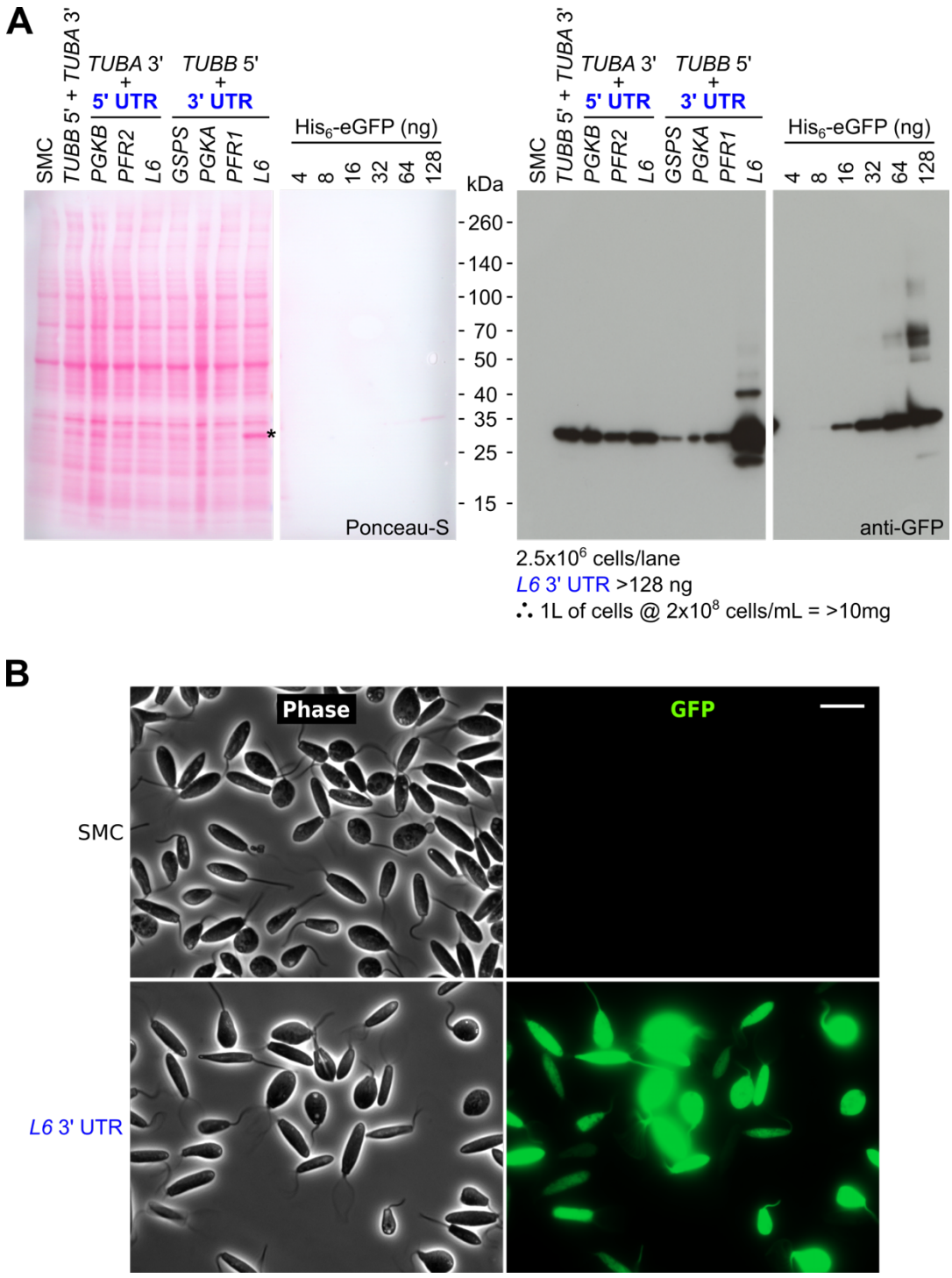


Figure 5-12 The *L6* 3' UTR substantially increases sfGFP^{op} expression levels in SMC cells. A Ponceau-S stained membrane and anti-GFP immunoblot of SMC cells containing pCEx with different 5' and 3' UTRs as indicated. eGFP of known concentrations was loaded for protein quantification. All 5' UTRs and most 3' UTRs make minimal difference to sfGFP^{op} levels, whilst the *L6* 3' UTR causes a large increase, even visible by Ponceau-S (indicated by *) **B** Live native fluorescence microscopy of parental SMCs and SMCs transfected with pCExC (constitutive promoter and *TUBB* 5' + *L6* 3' UTRs), showing sfGFP^{op} signal. Scale bar = 10 μm.

5.2.5. pCExT – optimisation of recombinant protein expression regulation

Following tests and validation of the pCExC UTRs, the next step was to optimise regulation of recombinant protein expression. As shown in Chapter 4, pLEXSY_IE-sfGO-N showed expression in the absence of tet (Figure 4-5). Factors contributing to this were discussed there, but one aspect that has not yet been explored here is the arrangement of the promoter and operator. pLEXSY_IE-sfGO-N has a T7 promoter followed by a single tet operator 5 base pairs downstream. Other studies have looked at adding additional tet operators downstream to tighten regulation with minimal success (Cross, 2017; Wirtz *et al.*, 1998). This is likely because TETR is more effective at inhibition of binding of the RNA polymerase rather than inhibition of transcript elongation (Dingermann *et al.*, 1992; Gatz and Quailt, 1988; Heins *et al.*, 1992). This implies that inhibition of transcription can occur when the tet operator is positioned either side of the TIS. In *T. brucei* this was shown when tet operators were placed either a few base pairs upstream or downstream of the *procyclin* promoter TIS (Wirtz and Clayton, 1995).

Based upon the above evidence, three distinct T7 promoter tet operator arrangements (1T, 2T and 3T) were designed in an attempt to create a pCEx vector with tightly regulatable expression (Figure 5-11A red box). 1T has a tet operator immediately upstream and downstream of the T7 promoter. In case this arrangement caused issues with induction through interaction between the operator sequences and the promoter, 2T and 3T were designed with their tet operators a few base pairs away from the promoter. 2T has two tet operators five base pairs downstream of the promoter. 3T is the same as 2T but with an additional tet operator four base pairs upstream of the promoter.

pCExC was modified, using oligonucleotide linkers containing the three different promoter-operator arrangements. These were transfected into SMCs. The transfected population were cultured for 24 h with or without 10 µg/mL of tet before whole cell lysates were made for assessment of $sfGFP^{op}$ levels. For the 1T arrangement, the levels of $sfGFP^{op}$ were barely detectable, even in the presence of tetracycline. Both 2T and 3T were inducible, with $sfGFP^{op}$ levels equivalent to 5-6 mg/L. 3T showed slightly tighter regulation relative to 2T but

not by a large amount. 3T gave an approximate 6-fold increase upon induction, whilst 2T gave a ~4-fold increase. The amount of signal still seen for non-induced cells is likely down to the optimised codon usage and *L6* UTR, causing high mRNA stability. As the construct with the tightest regulation, pCEx with the 3T arrangement was renamed pCExT, the T standing for tet-inducible.

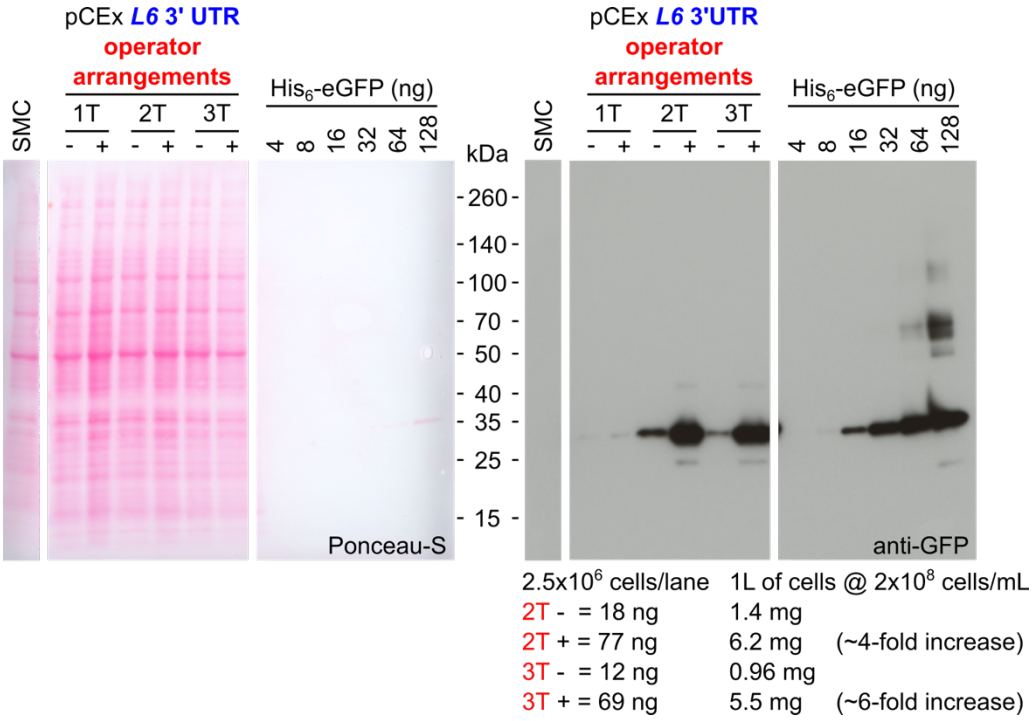


Figure 5-13 The 3T operator arrangement allows for the tightest regulation of sfGFP^{op} expression from pCEx. Ponceau-S stained membrane and anti-GFP immunoblot of SMC cells containing pCEx *L6* 3' UTR with different operator arrangements (1T, 2T and 3T). Cells were cultured with (+) or without (-) 10 µg/mL tet for 24 h. Known amounts of eGFP were loaded alongside for protein quantification. The 1T arrangement caused barely detectable levels of sfGFP^{op}, even in the presence of tet. The 3T arrangement presented the tightest regulation of sfGFP^{op} expression.

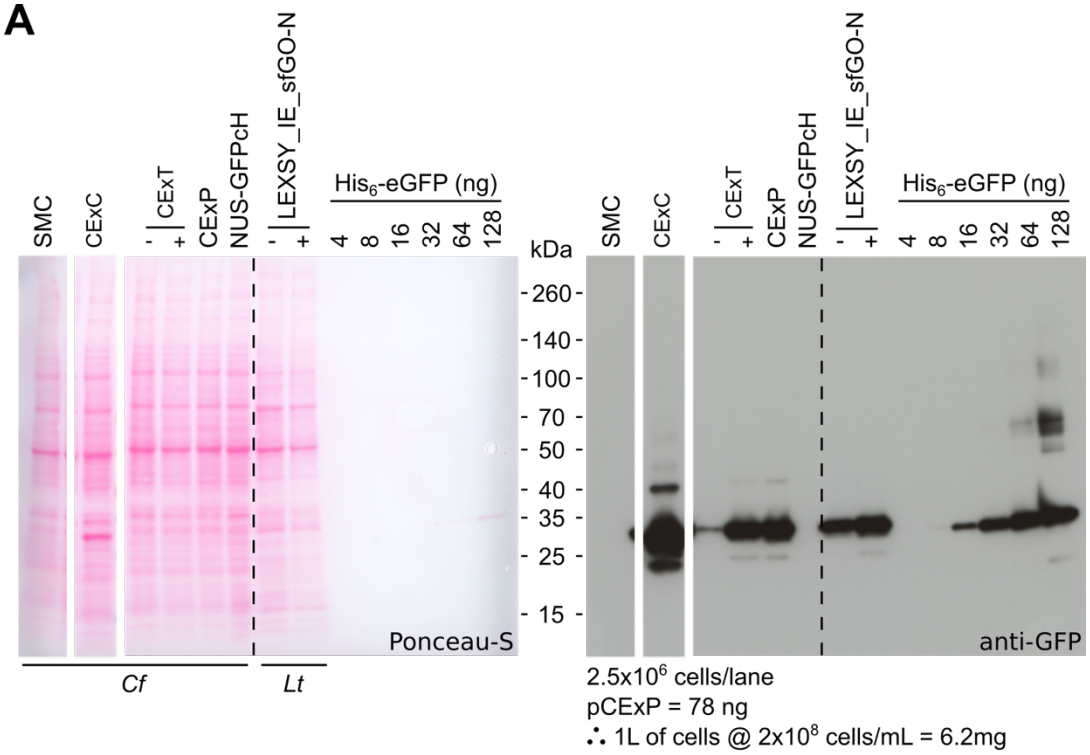
5.2.6. pCExP – integrated, PolI driven expression

pCEx was further modified to test reporter protein level upon integration downstream of the PolI promoter in the rDNA IGS (Figure 5-11B), relying on constitutive PolI readthrough for expression. To enable this, the original targeting sequence and the T7 terminators were removed from pCExC, the orientation of *HYG^R* was reversed, and new targeting sequences were inserted upstream of the T7 promoter and downstream of the *HYG^R* 3' UTR.

For transfection into cells, pCExP was linearised with SwaI. After transfection and clonal selection, three independent clones were taken forward. When assessing sfGFP^{op} amounts, all three clones had variable levels, likely due to integration into different rDNA repeat units (data not shown). The clone with the highest sfGFP^{op} levels was further analysed to quantify sfGFP^{op} abundance (Figure 5-14). These levels reached approximately 6 mg/L.

Figure 5-14 also shows CExP cells next to CExC and CExT cells for comparison of sfGFP^{op} levels. CExP shows similar expression levels to tet-induced CExT. These lines were further compared to SMC cells containing pNUS-GFPcH (expressing eGFP) and T7TR *L. tarentolae* cells containing pLEXSY_IE-sfGO-N (grown with (+) or without (-) 10 µg/mL tet 24h). eGFP expression from NUS-GFPcH cells were undetectable. sfGFP^{op} signal from LEXSY_IE-sfGO-N tet-induced cells was lower than that seen for the CEx lines. Regulation of expression of LEXSY_IE-sfGO-N cells was not as tight as pCExT, only showing a <2-fold increase of signal upon induction.

A



B

	GFP mg/L		
	Uninduced	**Induced**	**Induction fold-change**
pCExC	>10	n/a	n/a
pCExT	0.96	5.5	6
pCExP	6.2	n/a	n/a
pNUS	-	n/a	n/a
pLEXSY	2.2	4.2	2

Figure 5-14 A summary of the suite of CExSy vectors and their recombinant protein levels. A Ponceau-S stained membrane and anti-GFP immunoblot of SMC cells containing pCExC/T/P. *L. tarentolae* (*Lt*) T7TR cells containing pLEXSY_sfGO-N and *C. fasciculata* (*Cf*) SMC cells containing pNUS-GFPcH, expressing sfGFP^{op} and eGFP respectively, are included for comparison. Inducible cell lines were cultured with (+) or without (-) 10 µg/mL tet for 24 h. Known amounts of eGFP were run alongside for protein quantification. **B** Table summarising the recombinant protein levels from cell lines containing the different constructs in **A**.

5.2.7. pCExS – secretory expression

As discussed at the end of Chapter 4, two *T. brucei* surface proteins were expressed recombinantly to high levels when lacking their membrane association domains (Rooney *et al.*, 2015). As opposed to cytoplasmic

expression, secretion of these soluble recombinant proteins would be preferable, as this would drive them through the secretory pathway, allowing protein folding and processing as close to the native proteins as possible. To test secretion of recombinant proteins from SMC cells, the pCExS plasmid was created (Figure 5-15A). This was achieved through insertion of a signal peptide at the N-terminus of the *sfGFP^{op}* ORF in pCExC. The signal peptide used was from the *L. mexicana* secreted acid phosphatase (LmSAP). This is the same signal peptide as in the LEXSY vector used by Rooney *et al* (pLEXSY_hyg2). The DNA sequence was cloned in using codon-optimised oligonucleotide linkers.

To examine the level of sfGFP^{op} being secreted from CExS cells, the line was cultured for 24 h, reaching near to top density ($7x10^7$ cells/mL). The culture medium was then subjected to cold acetone treatment to precipitate proteins, which were then solubilised in Laemmli buffer for SDS-PAGE and immunoblot analysis (Figure 5-15). The cells themselves were also prepared for SDS-PAGE to enable direct comparison of intracellular *vs.* secreted protein. The amount of precipitated media resolved on the gel was equivalent to the number of cells loaded (based on the culture density at the time of harvest). Figure 5-15B shows the precipitated medium sample (M) to contain an equivalent of >3 mg/L, whilst the whole cell sample (C) contains an equivalent of 1.8 mg/L. It was assumed that the secreted protein yield could have been higher had the cell culture been allowed to reach top density ($2x10^8$ cells/mL) given the remaining amount of intracellular protein in sample C.

In conclusion, pCExS provides an option for secretion of recombinant proteins, reaching the range of mg/L for sfGFP^{op}; yet, pCExC provides the highest level of sfGFP^{op} among all pCEx vectors described here.

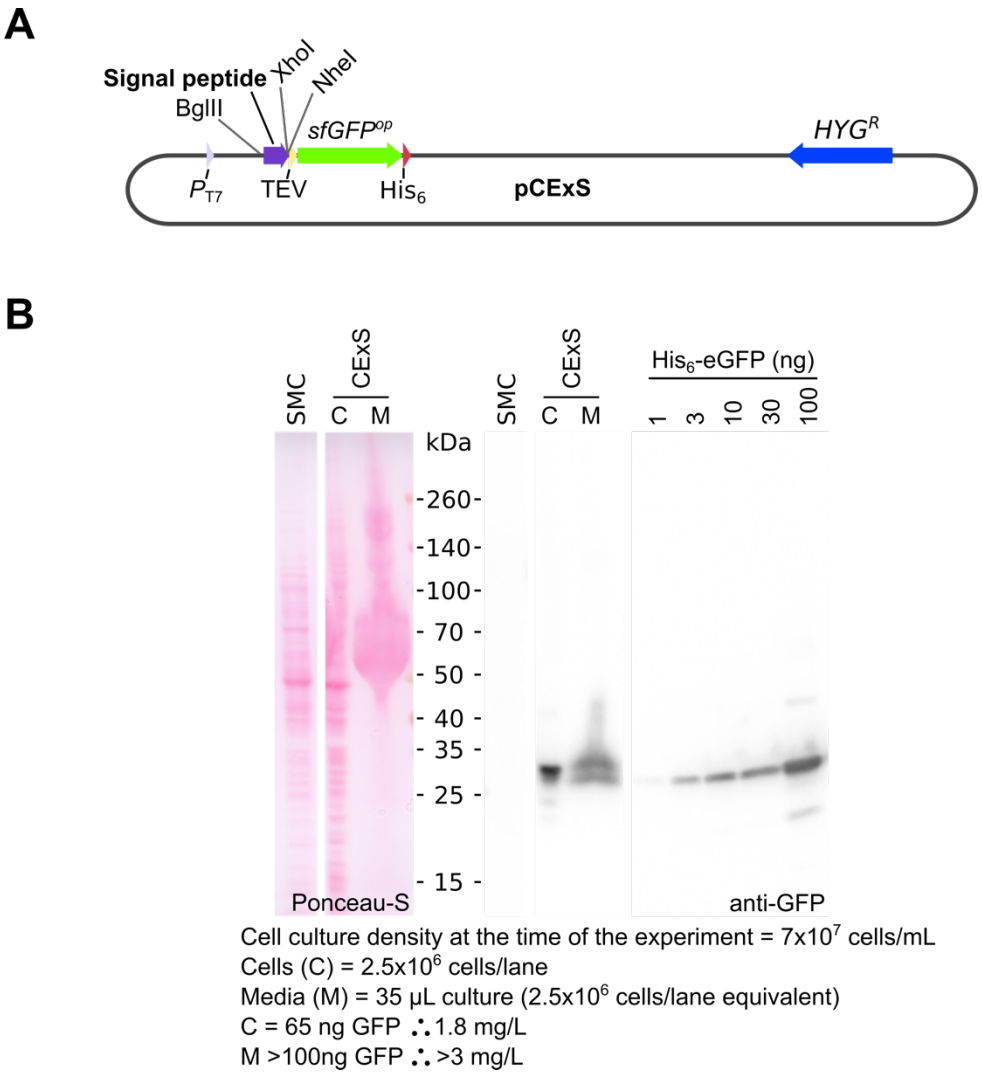


Figure 5-15 SMC cells containing pCExS secrete sfGFP^{op} into the medium. **A** Schematic of pCExS. The plasmid is derived from pCExC but modified with the addition of the *L. mexicana* secreted acid phosphatase (LmSAP) signal peptide to drive secretion of sfGFP^{op}. **B** Ponceau-S stained membrane and anti-GFP immunoblot of SMC cells containing pCExS. Expression of sfGFP^{op} was monitored through whole cell lysates (C) and acetone-precipitated media (M). Known amounts of eGFP were run alongside for protein quantification.

5.2.8. Integration and stability of pCExC

The experiments above describe the generation of a suite of vectors that enable recombinant protein expression in *C. fasciculata*. They were all characterised using the same reporter protein – sfGFP^{op} – and compared to the commercially available LEXSY system and the community resource pNUS. These experiments were, in the large part, carried out using cells transfected with circular plasmids (to be maintained as episomes). However, pCExC/T were

also designed with the option of integration into the *C. fasciculata* genome. Stable integration can be advantageous as it would allow large-scale culture without the need for (expensive) drug selection. As such, the optimised pCExC was used to test the integration stability of these vectors, and to directly compare sfGFP^{op} levels obtainable from the integrated construct against the episomal.

For integration of pCExC, the vector was linearised and transfected into SMCs. After 4 h recovery, individual clones were selected for with 100 µg/mL Hygromycin B. Six clones were screened by immunoblot (Figure 5-16A). Some degree of variation was seen between the clones, but the highest expressors reached similar levels as the SMC population transfected with episomal pCExC. Based on the immunoblot in Figure 5-16B, both integrated and episomal pCExC led to the production of >20 mg/L sfGFP^{op}.

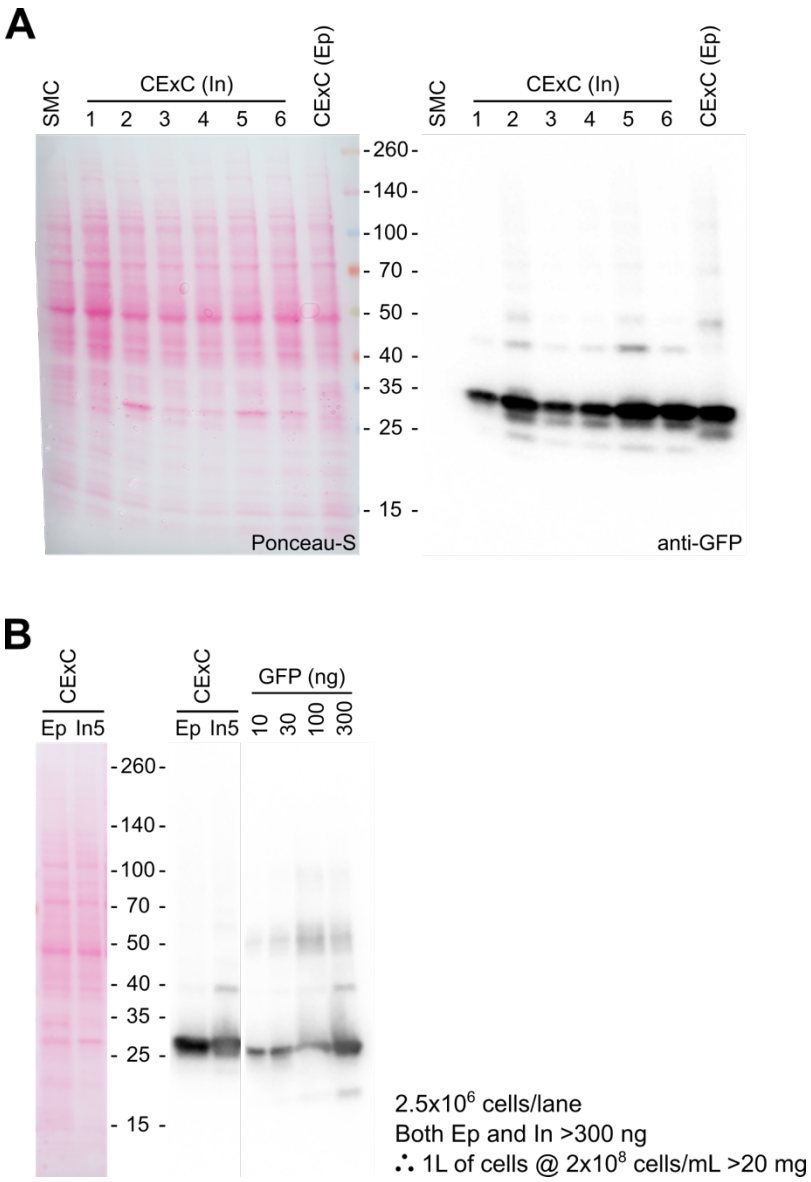


Figure 5-16 Genome-integrated pCExC produces clone-to-clone variation, with the highest expressors matching recombinant protein expression from episomal pCExC. A Ponceau-S stained membrane and anti-GFP immunoblot of SMC cells containing pCExC integrated into the genome (In) or maintained as an episomal (Ep). Six different genome-integrated clones and 1 episomal population were screened for sfGFP^{op} level. A degree of clonal variation was observed between the CExC (In) lines. Clone 5 showed the highest amount of sfGFP^{op} signal, comparable to that detected in the pCEx (Ep) population. **B** As the highest expressor, clone 5 was re-analysed (along with the episomal population) next to known amounts of eGFP for protein quantification. Both lines produce >20 mg/L.

To assess the stability of pCExC (genome-integrated and episomal), cells were grown with or without selection pressure across 1, 3, 5 and 7 days. Loss of pCExC was monitored through measurement of sfGFP^{op} florescence using flow cytometry (to count the number GFP-positive *vs.* GFP-negative

cells). The episomal cell line was expected to lose sfGFP^{op} fluorescence over time, whilst the integrated cell line was expected to lose less fluorescence (if any), showing stable integration. Figure 5-17 shows the flow cytometry results. Length of time grown without drug is displayed as number of generations, with 0 generations referring to cells constantly grown under selection pressure (Hygromycin B) as a positive control. The SMC cell line was used as a negative control to define background fluorescence. Surprisingly, both CExC (Ep) and CExC (In) lines lost sfGFP^{op} fluorescence at the same rate, with ~2% loss/generation, reaching only ~10% green cells by 42 generations. Although the two lines did differ in the relative number of GFP negative cells present when under constant selection pressure. The result suggests that the 'integrated' cell line either contains a non-integrated form of pCExC, or that pCExC was quickly removed by recombination upon the removal of selection drug. The difference observed between the relative number of GFP-negative cells for the cultures grown under selection pressure, suggests that the latter might be the case. It is possible that the cells up took circular DNA in the transfection, but this is unlikely, as the cut DNA was resolved by agarose gel electrophoresis prior to transfection and appeared to be completely linearised (Figure 5-17C).

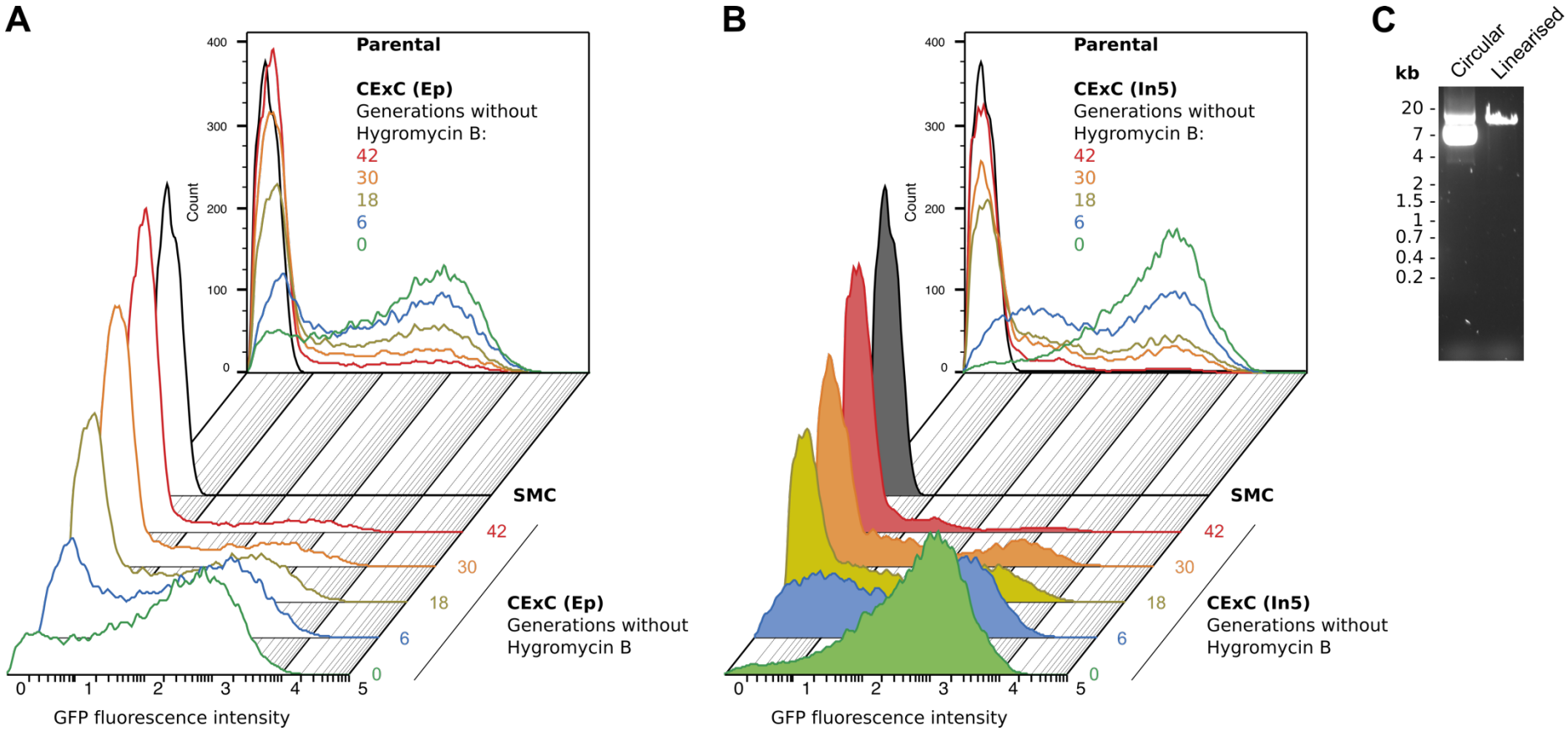


Figure 5-17 Rate of loss of green fluoresence is similar between SMC cells containing pCExC either genome-integrated or episomal. Flow cytometric analysis of SMC cells transfected with circular or linearised pCExC. Cell line containing circular pCExC is a population, whilst cell line containing linearised pCExC is clonal (clone 5 in Figure 5-16). Cells cultured for 0, 6, 18, 30 and 42 generations without selection pressure. 0 generations indicates cells grown with selection pressure (Hygromycin B). SMC cells were used as a negative control. **A** Data for the cell line transfected with circular DNA. **B** Data for the cell line transfected with linear DNA. Both cell lines had a green fluoresence loss/generation of ~2%, impling that the linearised pCExC vector was not stably integrated and quickly lost upon removal of selection pressure. **C** Uncut and NotI

cut pCExC plasmid DNA resolved by gel electrophoresis, prior to transfection into SMC cells. The cut DNA resolved here is a proportion of the exact same DNA sample used for transfection which resulted in CExC (In5) in **B.**

Whilst flow cytometry was a useful technique to use here as it can count thousands of cells at a time for analysis of GFP signal, the data cannot definitively distinguish between the two scenarios of 'integrated' pCExC either being present as an episome or being genome-integrated but quickly removed from the genome upon removal of selection pressure. Instead, diagnostic PCR was carried out (strategy depicted in Figure 5-18A). A single primer (P1) that anneals both up and downstream of the integration sequence in pCExC (acting as a forward and as a reverse primer) was used to amplify a product of ~0.8 kb if the plasmid was circular inside cells. An additional reverse primer was included (P2), that anneals downstream of the pCExC integration locus in the *C. fasciculata* genome. In combination with P1, P2 amplifies a product of ~1 kb if plasmid integration had taken place. Genomic DNA was purified from all six clones of integrated pCExC, from the population of episomal pCExC cells, and from parental SMC cells, to be used as templates in this diagnostic PCR reaction. Uncut pCExC plasmid was also included as a control.

Figure 5-18B shows that PCR of genomic DNA from all 6 clones of the linearised pCExC transfection contain a product at ~0.8 kb, matching that from both PCR of the plasmid and PCR of DNA from cells containing episomal pCExC. PCR of the plasmid also produced multiple amplicons at ~1 kb and above. These additional products were taken to be a result of non-specific primer binding, likely not appearing in the other PCR reactions due to lower amplification efficiency. This result provides strong evidence that clones resulting from transfection with linearised pCExC contain circular pCExC as an episome. It cannot be ruled out that these clones may also contain genome-integrated pCExC, as no positive control could be included for integration of the plasmid. However, this result, coupled with the flow cytometry data, implies that expression of sfGFP^{op} in *C. fasciculata* cells transfected with linearised pCExC, is largely driven by episomal DNA, resulting in the rapid loss observed in Figure 5-17 upon culture of cells without selection pressure.

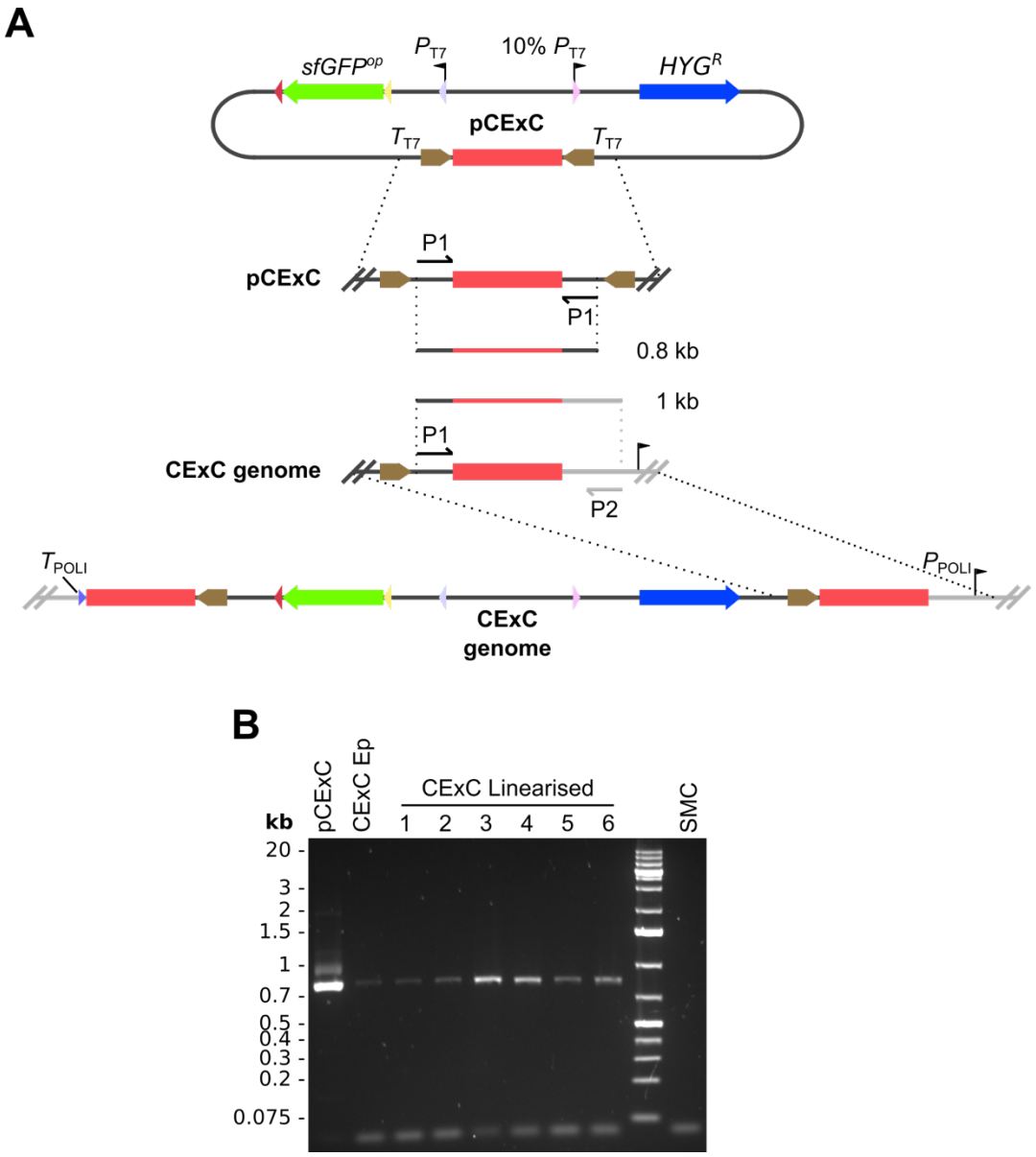


Figure 5-18 SMCs transfected with linearised pCExC contain circular pCExC. **A** Schematic illustrating the diagnostic PCR strategy to identify either episomal pCExC or pCExC integrated into the rDNA IGS of *C. fasciculata*. Primer P1 produces a ~0.8 kb product if pCExC is circular, whilst primers P1 and P2 together produce a ~1 kb product if pCExC has integrated. **B** Diagnostic PCR on gDNA from SMC cells transfected with circular (CExCEp) or linearised pCExC. Uncut plasmid DNA was also used as a template as a positive control, and gDNA from parental SMC cells as a negative control. Agarose gel electrophoresis to resolve PCR products reveals a 0.8 kb band, indicative of circular pCExC.

5.3. Discussion

This chapter describes the development of the ***C**rithidia* **Ex**pression **Sy**stem (CExSy). In terms of handling, *C. fasciculata* was shown to be just as simple as *L. tarentolae* with the added bonuses of a faster doubling time and the ability to

store cell cultures at 4 ^{o}C. An attempt was made to see if *L. tarentolae* could survive 4 ^{o}C storage for 2 weeks, but this culture did not recover when returned to 26 ^{o}C.

With wild type *C. fasciculata* culturing in the lab, the SMC cell line was generated. The initial goal of creating an inducible cell line was successful. Although it might be that future improvements could be made to this cell line. For example, PolI-driven expression in trypanosomatids is known to be higher than PolII-driven expression. In *T. brucei*, PolI was shown to drive ~10-fold higher luciferase levels than PolII (Wirtz *et al.*, 1999; Wirtz and Clayton, 1995). Therefore, integration at the rDNA locus using the single marker strategy (as opposed to the LEXSY T7TR dual marker strategy) may increase the T7RNAP and TETR levels obtained. However, increase in T7RNAP levels might not increase recombinant protein yield, as T7RNAP may not be limiting. In creation of the T7TR line, an alternative cell line was also created which expressed T7RNAP to roughly 10 times lower levels due to sub-optimal UTRs (Kushnir *et al.*, 2005). The levels of an eGFP reporter were only marginally reduced, suggesting the 10-fold lower T7RNAP expression was already near the levels required for maximal expression. Additionally, as demonstrated by the modulation of the T7RNAP levels with UTRs (Kushnir *et al.*, 2005; Poon *et al.*, 2012), and the results in this Chapter, increasing levels of T7RNAP could simply be done through selecting different UTRs. As the SMC cell line created was suitable to the aims of this project, further optimisation was not carried out. However, with the pCEx vectors now in hand, the SMC cell line could be further refined if necessary.

The pCEx vector design was intended to allow optimal expression of recombinant proteins from SMC cells. Initial optimisation of this vector was done through testing different UTRs. As already discussed, the 5' UTR can play a role in mRNA levels through efficiency of trans-splicing and through RBPs (Clayton, 2019; Mittra and Ray, 2004; Siegel *et al.*, 2005) (although RBP recognition motifs in the 5' UTR for differential regulation of mRNA levels have only been well documented in *C. fasciculata* by Mittra and Ray, 2004). However, in general in trypanosomatids, it is believed that the 3' UTR plays more of a role in the steady-state mRNA levels than the 5' UTR (Delhi *et al.*, 2011). This is in part

because the 5' UTR is restricted by structural requirements for translation initiation. Based on this, it is perhaps unsurprising that swapping the 5' UTRs of pCEx did not show a strong effect on sfGFP^{op} levels, whilst usage of the *L6* 3' UTR did cause a large increase in sfGFP^{op} amount. Ribosomal protein UTRs are good candidates for increasing mRNA stability as the ribosomal protein mRNAs have been shown in *T. brucei* to be highly stable with long half-lives (Manful *et al.*, 2011). Additionally, the RBP TbDRBD3, which is known to stabilise mRNA, has been shown to bind to ribosomal mRNA. It could be beneficial to screen multiple ribosomal protein 3' UTRs. This was not done here as the *L6* 3' UTR was sufficient for the desired yield (when assessed with GFP reporter proteins), with pCExC allowing production of recombinant sfGFP^{op} in the range of tens of milligrams per litre of *C. fasciculata* culture.

As discussed in Chapter 4, certain aspects of the pLEXSY_IE-sfGO-N plasmid were to be avoided, namely the arrangement of the drug resistance marker downstream of the recombinant protein ORF in the same polycistron, as this potentially contributed to the leaky expression. In pCEx, expression of the drug resistance marker was uncoupled from expression of the recombinant protein ORF by giving each of them their own promoters, and by placing them on the antisense strand relative to one another. This, along with testing different arrangement of the T7 promoter and tet operators, was used to minimise transcription leakiness from pCEx. This strategy produced the pCExT construct which had marginally tighter regulation than pLEXSY_IE-sfGO-N. However, pCExT was still particularly leaky. This was likely due to the optimised coding sequence coupled with the *L6* 3' UTR, providing high efficiency of translation and mRNA stability, and leading to higher expression in the off-state than would usually be seen. This leakiness might be unavoidable. However, further tests could be done in an attempt to tighten expression regulation, such as increasing the expression level of TETR. In one study in *T. brucei*, both the *T7RNAP* and the *TETR* gens were introduced into the *TUBB* locus, but with the *TETR* under the control of the 10% T7 promoter (Wirtz *et al.*, 1999). This promoter was shown to drive ~2x higher expression than constitutive PolII readthrough. As a result, TETR was produced to higher levels than T7RNAP, allowing tighter regulation than when they were both driven by the same promoter. The SMOx

cell line, on the other hand, has subsequently been shown to have tighter regulation in the off-state (Poon *et al.*, 2012). This may have been down to the overall higher levels of both proteins, and so even tighter regulation may be seen if TETR was driven to even higher levels relative to T7RNAP. In the SMC cell line, the absolute levels of T7RNAP and TETR have not been assessed. Neither was the regulation of SMC/pCExT expression optimised further here. However, this could be an option worth exploring if problems with toxic proteins are encountered.

An alternative approach would be to use a different inducible system altogether. Vanillic acid, along with the vanillic acid repressor protein VanR, has been used for inducible expression in *T. brucei* (Sunter, 2016). Using a modified SmOx vector containing both *VanR* as well as *TETR*, Sunter created a cell line that could be induced by vanillic or tetracyline depending on the operator sequence placed downstream of the T7 promoter. Whilst the VanON inducible system did not produce as strong of an induction effect as the tetON system, it did show tight regulation and, thus, is another option if tight regulation of expression is required.

The integration locus chosen for the pCExC/T/S vectors was the rDNA ISG, upstream of the TIS, in an attempt to minimise leaky transcription from PolI readthrough. However, based upon diagnostic PCR shown in Figure 5-18, cells transfected wth linearised pCExC contained circular pCExC DNA. This could be through direct re-circularisation of the vector. Trypanosomatids are not thought to carry out nonhomologous end joining due to lack of some of the proteins involved; yet, they have been shown to join linear DNA fragments through microhomology-mediated end joining (MMEJ), utilising homologous regions as short as 2–20 bp in length (Burton *et al.*, 2007; Laffitte *et al.*, 2016). Despite this, homolgous recombination occurs much more frequently in these organisms; thus occurance of MMEJ would likely require a failure of pCExC genomic integration by homologous recombination. If linearised pCExC is failing to integrate into the *C. fasciculata* genome, this is unlikely to be due to the length of the targeting sequence (384 base pairs each end once the vector is linearised), as it has been shown in *L. mexicana* that linear DNA with homologous sequences of lengths as short as 100 base pairs can successfully

integrate (Dean *et al.*, 2015) (and even lengths as low as 24 bp if Cas9 is used to target DNA cleavage; Beneke *et al.*, 2017). Instead, recombination may be of low efficiency at the locus targeted by pCExC/T/S. Indeed, Papadopoulou and Dumas (1997) reported that choice of integration locus can have a large effect on transfection efficiency.

An alternative explanation to re-circularisation of pCExC is successful genome integration of the vector, followed by quick subsequent removal by a second round of homologous recombination. This is supported by evidence that *Leishmania* have been shown to recombine regions of their genome and propagate them as multicopy circular elements as a mechanism of gene amplification, in response to drug selection pressure (Ubeda *et al.*, 2014). Appearance of these circular elements was reported to occur relatively rapidly (~10% of the population after 54 generations), as was subsequent disappearance upon removal of drug pressure. In addition to this, *Leishmania* have also been reported to carry out gene amplification through the production of extrachromasomal linear elements that can subsequently lead to formation of circular elements (Grondin *et al.*, 1998; Ubeda *et al.*, 2014). Production of linear amplicons relys on neighbouring inverted repeats. The selected integration locus of pCExC/T/S is directly downstream of the four putative PolI terminators which consist of 55-57 bp with internal inverted repeats. Thus, this mechanism of extrachromosomal gene amplification could also play a role here. In either senario, production of differential copy numbers of linear or circular elements could potentially explain the clone-to-clone variability observed for the pCExC 'integrated' cell lines (Figure 5-16).

Whilst the definite reason for the presence of circular pCExC DNA in these cells is perhaps unclear, it does reveal the need for more work if a stably integrated cell line is required. pCExP may function as an alternative option for this purpose; successful integration of pCExP has not been proven, but the vector relies on integration to allow expression of both the reporter sfGFP^{op} and the selectable marker by PolI readthrough. Therefore, the successful selection of CExP clones, and the expression of sfGFP^{op} by these clones, implies that integration of pCExP may have been successful. Despite this, in light of the

pCExC results, stable integration of pCExP cannot be assumed and should be confirmed by diagnostic PCR and flow cytometry prior to use for such a purpose.

CExSY and the suite of plasmids designed here provide a good base for recombinant protein expression from *C. fasciculata*. Whilst there is scope for further refinement of the system, in its current form it allows expression of mg/L of recombinant protein in multiple flexible configurations. In the next chapter, the use of this system in an attempt to express recombinant *T. brucei* surface proteins is described.

Chapter 6. Recombinant *Trypanosoma brucei* protein expression in CExSy

6.1. Introduction

With CExSy setup, the various configurations could now be used to recombinantly express *T. brucei* surface proteins. This chapter describes attempts to express these proteins, exploring the different options provided by CExSy and comparing them to expression using LEXSY.

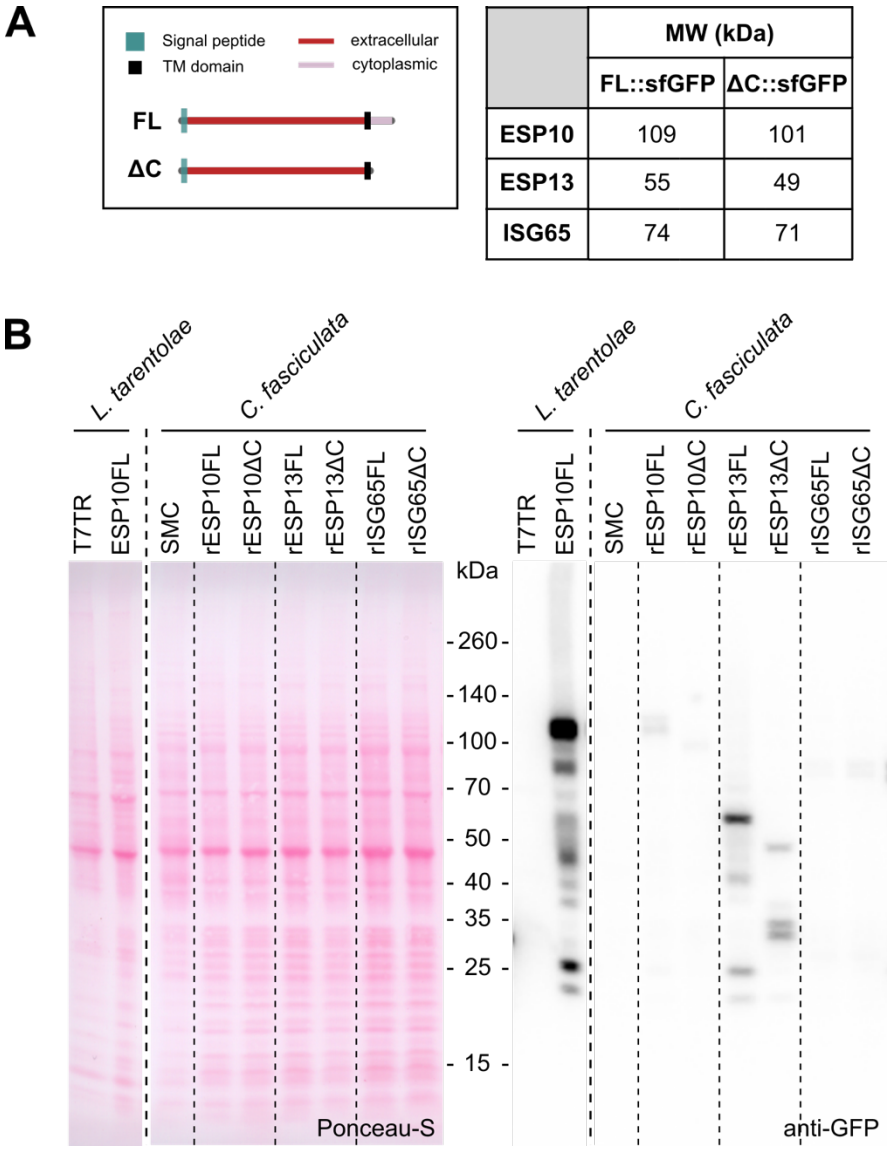
6.2. Results

6.2.1. Recombinant membrane-associated protein expression

Initial attempts to use CExSy for recombinant expression of *T. brucei* surface proteins largely followed the same strategy as used in LEXSY (Chapter 4), allowing direct comparison between the two systems. The candidate proteins ESP10, ESP13 and ISG65 were to be tested as both full length (FL) and with no cytoplasmic domain (ΔC). Cytoplasmic (Cy) forms were avoided, as it was decided that if non-membranous forms were required, the proteins would be secreted from the cells in order to drive them through the secretory pathway for folding and processing (see section 6.2.2).

ESP10, *13* and *ISG65*, either *FL* or *ΔC*, were cloned into pCExC as the plasmid that drove the highest level of sfGFP^{op} in Chapter 5. As there was still a possibility that overexpression of these proteins could be detrimental to the growth of *C. fasciculata,* one of the candidate proteins – ESP10FL – was also cloned into pCExT to test inducible expression, but this construct led to similar expression levels in SMCs as pCExC (and neither caused growth defects, data not shown). The different constructs were transfected as circular DNA into SMC cells. After transfection and drug selection, recombinant protein levels of these cell populations were assessed by immunoblotting with anti-GFP antibodies (Figure 6-1). In addition to the *C. fasciculata* lines, *L. tarentolae* T7TR cells expressing rESP10FL (from Chapter 2) were included in the analysis for

comparison. The signal detected on the blot for the different SMC cell lines was much lower than that of the T7TR rESP10FL cell line, with rESP13FL having the highest SMC protein level. Due to the low levels observed, alternative CExSy configurations were explored.



	MW (kDa)	
	FL::sfGFP	ΔC::sfGFP
ESP10	109	101
ESP13	55	49
ISG65	74	71

Figure 6-1 *C. fasciculata* SMC cells, containing pCExC derivatives, express recombinant *T. brucei* surface proteins to lower levels than *L. tarentolae* T7TR cells containing pLEXSY_IE-sfGFPO-N derivatives. Ponceau-S stained membrane and anti-GFP immunoblot of SMC and T7TR cells expressing *T. brucei* surface proteins as sfGFP^{op} fusions. **A** Left panel shows topology schematics of full length (FL) proteins and proteins lacking a cytoplasmic domain (ΔC). Table contains predicted molecular weights of each fusion protein. **B** Immunoblot shows a major band at the expected MW for each recombinant protein. Additional lower bands also appear, likely a result of proteolysis. Both rESP10 and rESP13 showed greater protein levels when expressed in their respective FL forms relative to their ΔC forms. rESP10FL levels were much lower in *C. fasciculata* compared to *L. tarentolae*.

6.2.2. Secretion from *C. fasciculata* and signal peptide optimisation

As discussed in Chapter 4, membrane proteins can be challenging to produce recombinantly to high levels. However, transmembrane proteins have been expressed to high amounts without their membrane association domains (being instead secreted from the cells), as reported for rVSG and rISG65 in LEXSY (Rooney *et al.*, 2015).

In addition to membrane proteins, secretion can be an effective strategy for the production of GPIAPs (as also demonstrated by the expression of rVSG by Rooney *et al*). Recombinant expression of membrane-associated GPIAPs would require a C-terminal GPISS to direct attachment of a GPI anchor. However, this anchor is attached to the mature folded protein and is not required for protein processing. Thus, secretion is a viable option for GPIAPs, simplifying their production and purification. Whilst expression of GPIAPs had not been attempted in this project up to this point, they are of interest, as they have the potential to be good vaccination candidates. When sorted to the plasma membrane, the GPI anchor causes the entirety of the protein to be extracellular and extend into the VSG coat. In addition, as GPIAPs often act as receptors and enzymes (Grandgenett *et al.*, 2007; Montagna *et al.*, 2002; Salmon *et al.*, 1994; Vanhollebeke *et al.*, 2008), they must be accessible to their ligand/substrate through the VSG coat and, therefore, may be accessible to antibodies. Although this does not guarantee antibody access, as small ligands may be able to pass the VSG 'molecular sieve' (with gaps between VSG dimers modelled to be 4–6 nm; Bartossek *et al.*, 2017) and would not require an exposed surface protein to interact with.

In Chapter 5, pCExS was created for secretion of recombinant proteins from SMCs into the culture medium. This plasmid contains the LmSAP signal peptide sequence at the 5'-terminus of the *sfGFP^{op}* ORF. When transfected into SMCs, pCExS led to the secretion of sfGFP^{op} in the range of milligrams per litre of cell culture, presenting a viable option for secretory production of surface membrane proteins and GPIAPs lacking their membrane association domains. Whilst use of the LmSAP signal peptide successfully yielded high level of sfGFP^{op} secretion, some proteins which naturally proceed through the secretory pathway may be more efficiently processed if they retain their endogenous signal

peptides. One study using LEXSY to secrete haemagglutinin of different influenza strains found that some of the resultant recombinant proteins would only be secreted when using their respective endogenous signal peptides rather than the LmSAP signal peptide (Pion *et al.*, 2014). As such, it was decided to test secretory production of recombinant *T. brucei* surface proteins (lacking their membrane association domains) using either their endogenous signal peptides or using the LmSAP signal peptide.

Two *T. brucei* proteins were chosen to test this. The first was ESP10 as one of the previously expressed candidate transmembrane proteins. The second was HpHbR as a candidate GPIAP. HpHbR is an invariant FP protein. Its structure has been solved, showing it to extend through the VSG coat and likely be accessible to antibodies (Higgins *et al.*, 2013; Stødkilde *et al.*, 2014). This accessibility has been further shown through HpHbR receptor-dependant killing of *T. brucei* cells with an anti-HpHbR antibody-drug conjugate (MacGregor *et al.*, 2019). In addition, HpHbR can be used as a preliminary test to see if *C. fasciculata* correctly glycosylates recombinant *T. brucei* proteins, as endogenous HpHbR migrates slower than expected by SDS-PAGE, showing an apparent molecular weight ~20 kDa larger than predicted by primary sequence due to N-glycosylation (Vanhollebeke *et al.*, 2008).

ESP10Sec (no transmembrane or cytoplasmic domain) and *HpHbRSec* (no GPISS) were cloned either into pCExS with the LmSAP signal peptide, or into pCExC with their endogenous signal peptides. These plasmids were transfected as circular DNA into SMC cells and the populations selected. To asses recombinant protein secretion, the four cell lines were cultured to $2x10^8$ cells/mL prior to preparation of whole cell lysates (C) and cold acetone precipitated medium samples (M). Recombinant protein levels were then analysed by anti-GFP immunoblot (Figure 6-2).

Surprisingly, no secretion was observed for proteins with their endogenous signal peptides (Tb lane M). In contrast, both of the recombinant proteins showed secretion when using the leishmanial signal peptide (Lm lane M). This suggests that *T. brucei* signal peptides may not be recognised by *C. fasciculata*, likely contributing to the low levels of recombinant membrane proteins seen in Figure 6-1.

Both rESP10Sec (Lm) and rHpHbRSec (Lm) show two bands in their respective C samples (albeit rESP10Sec is very faint), but only the higher MW band is detected in their respective M samples, suggesting that the two proteins are being modified prior to secretion from the cells. This is in agreement with previous work that shows both proteins to be N-glycosylated in *T. brucei* (Vanhollebeke *et al.*, 2008; Whipple and Gadelha, personal communication), revealing *C. fasciculata* to be a promising system for production of correctly processed *T. brucei* proteins. Encouragingly, the shift in size of rHpHbRSec does indeed match the ~20 kDa increase in size expected based on N-glycosylation of the native protein (Vanhollebeke *et al.*, 2008).

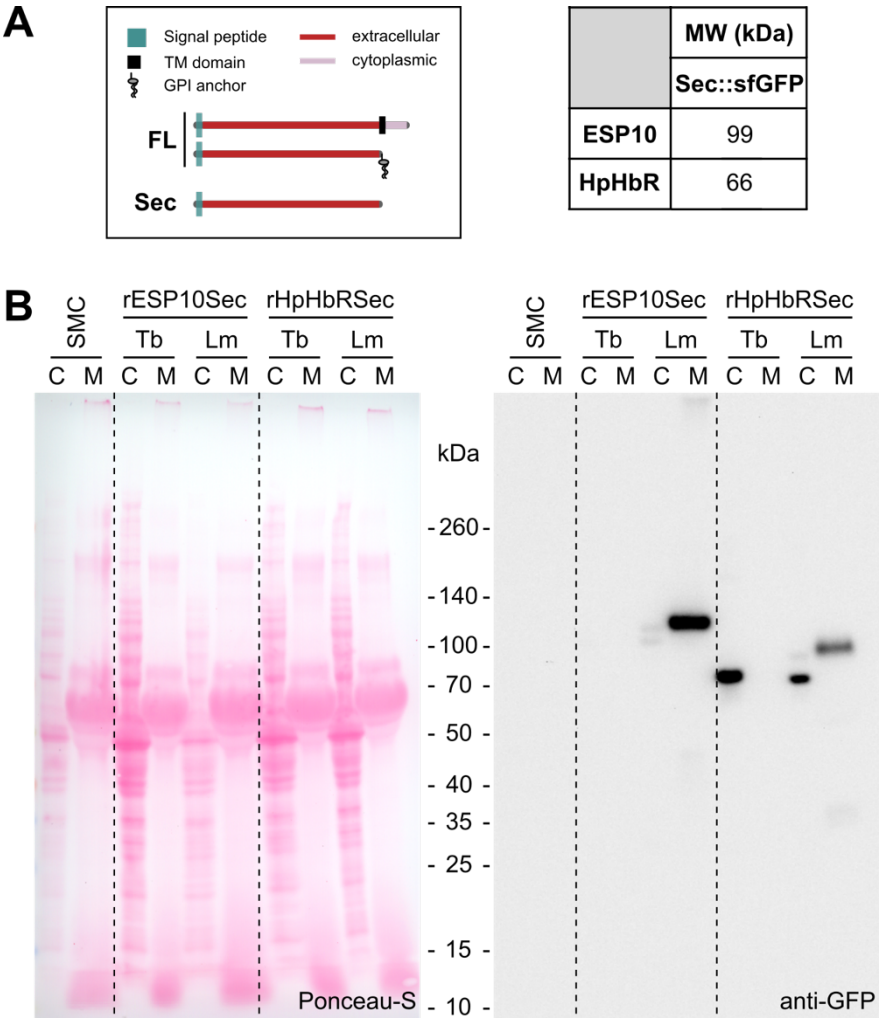


Figure 6-2 *C. fasciculata* does not recognise *T. brucei* signal peptides. Ponceau-S stained membrane and anti-GFP immunoblot of SMC cells secreting recombinant *T. brucei* surface proteins as sfGFP^{op} fusions with either their endogenous (Tb) or a *L. mexicana* (Lm) signal peptide. Samples consist of whole cell lysates (C) and acetone precipitated media (M) **A** Left panel shows topology schematics of full length transmembrane and GPI-anchored proteins (FL) and secreted (Sec) proteins. Table contains predicted molecular weights of eacg fusion protein. **B** Immunoblot reveals that recombinant proteins with Tb signal peptides present no signal in M samples, showing no secretion SMC cells. Recombinant proteins expressed with the Lm signal peptide do secrete, showing signal in M. Shift in size of secreted proteins likely due to post-translational modification such as N-glycosylation.

The lack of recognition of *T. brucei* signal peptides by *C. fasciculata* was surprising but not unprecedented. It has been reported that proteins with *T. brucei* or *L. infantum* signal peptides are not imported by canine-derived microsomes (fragments of the ER obtained through centrifugation of homogenised cells) (Al-Qahtani *et al.*, 1998). This was shown to be caused by residues in the h-region of the signal peptides, as swapping this region for a

non-trypanosomatid h-region led to protein import. The importance of the h-region has been further shown using *T. brucei* derived microsomes (Duffy *et al.*, 2010). Duffy *et al* found that changing amino acids in the h-region of a VSG signal peptide affected the efficiency of uptake into microsomes to varying degrees. On top of this, simply scrambling the order of amino acids in h-regions of functional signal peptides proved sufficient to completely inhibit import of VSG. It was therefore suggested that signal peptide recognition and import rely on specific motifs within the h-region.

Based upon the above evidence, despite the recognition of the LmSAP signal peptide by *C. fasciculata*, secretion efficiency may be improved through use of *C. fasciculata*-derived signal peptides (containing *C. fasciculata*-specific h-motifs). To test this, Two *C. fasciculata* signal peptides were chosen, to be tested on rESP10Sec as a model protein. The first signal peptide derives from the *C. fasciculata* homologue of LmSAP as a direct comparison (CfSAP – CFAC1_280074200). The second chosen signal peptide came from a copy of GP63. GP63 is the major surface protein of *Leishmanial* species, with ~$5x10^5$ copies per cell (~1% of total cellular protein) (Bouvier *et al.*, 1985). As such, it likely requires efficient secretion to allow the export of such a large amount of protein. GP63 is also present in *C. fasciculata* and likely acts in the adherence to the insect gut wall (d'Avila-Levy *et al.*, 2006; Yao, 2010).

GP63 is a multicopy gene. To decide which signal peptide to use, SignalP version 3.0 was used (Dyrløv Bendtsen *et al.*, 2004). SignalP predicts the presence of signal peptides and their likely cleavage sites in polypeptides. Different signal peptides were inserted *in silico* into pCExS_rESP10Sec. SignalP was then used to predict likely cleavage sites and their cleavage probability scores. Ideally only one cleavage site would be predicted, and with high probability score, as there is evidence that this can positively influence efficiency of secretion (Klatt and Konthur, 2012; Mori *et al.*, 2015). A GP63 signal peptide was chosen (CFAC1_140029100) that had only one predicted cleavage site with a probability score of 0.983 (when inserted into pCExS_rESP10Sec). In comparison, pCExS_rESP10Sec with either the LmSAP or CfSAP signal peptide has a predicted cleavage site with a probability score of 0.62 and 0.815 respectively (Figure 6-3).

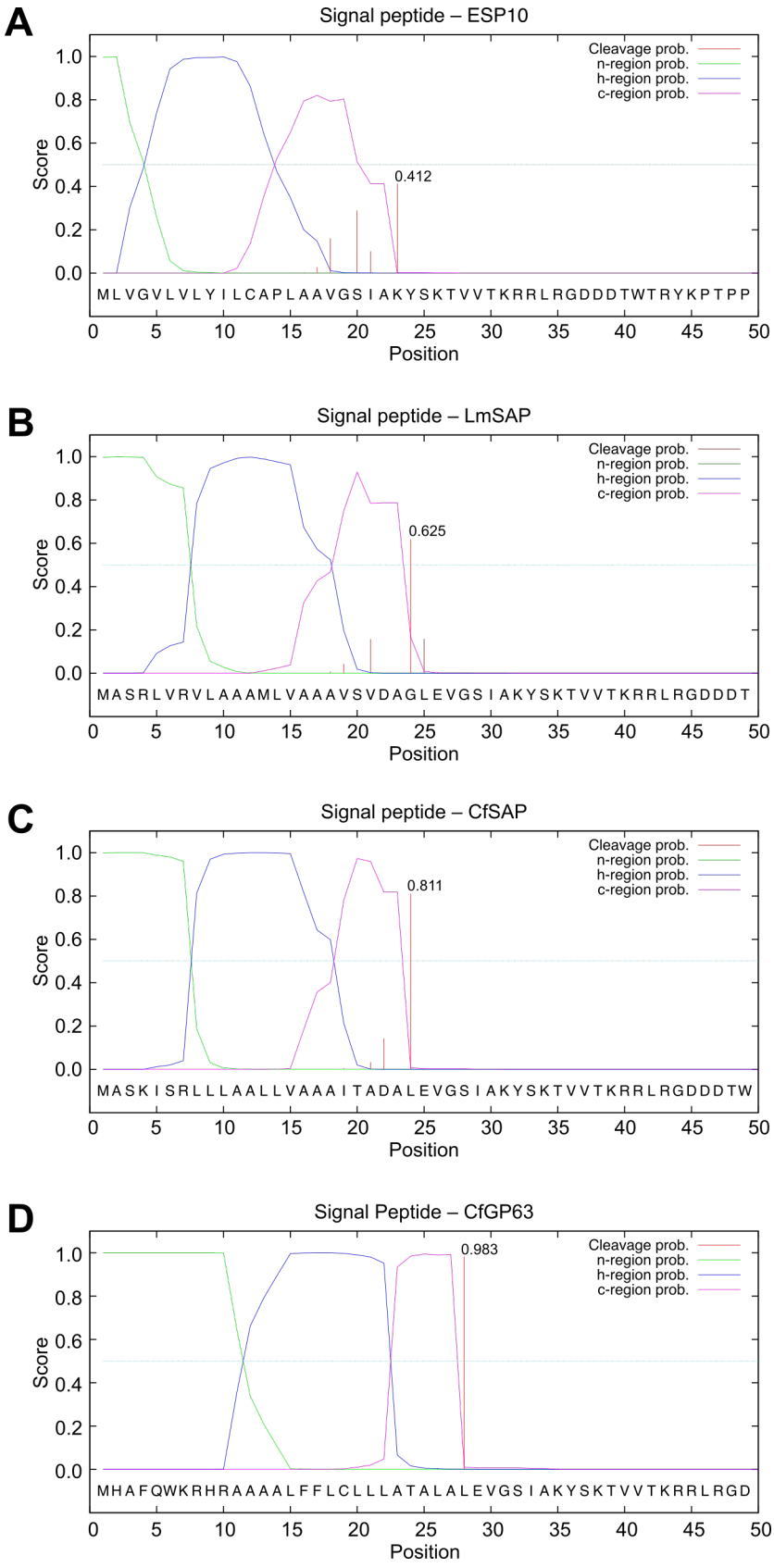


Figure 6-3 SignalP 3.0 signal peptide and cleavage site probability score predictions for ESP10 with different signal peptides. Signal peptides derive from **A** ESP10, **B** LmSAP, **C** CfSAP and **D** CfGP3. Cleavage probability scores indicated for cleavage sites of highest predicted probability.

The respective signal peptides from CfSAP and CfGP63 were cloned into pCExS_rESP10Sec before transfection into SMCs. After drug selection of the population, these cell lines were grown to $2x10^8$ cells/mL and whole cell lysates (C) and acetone precipitated media (M) prepared as before. Recombinant protein levels were then assessed by anti-GFP immunoblot (Figure 6-4). The CfSAP signal peptide made little if any difference to the secretion levels of rESP10Sec relative to the LmSAP signal peptide, whilst the CfGP63 signal peptide decreased the amount of secreted rESP10Sec. The results suggest that *Crithidia* and *Leishmania* have signal peptides with similar properties that can be equally processed by *C. fasciculata*. They also demonstrate that different signal peptides from the same organism are not all equally processed (whether that be because of the signal peptides themselves or the combination of signal peptide and attached protein). As such, the original pCExS with the LmSAP signal peptide was taken forward.

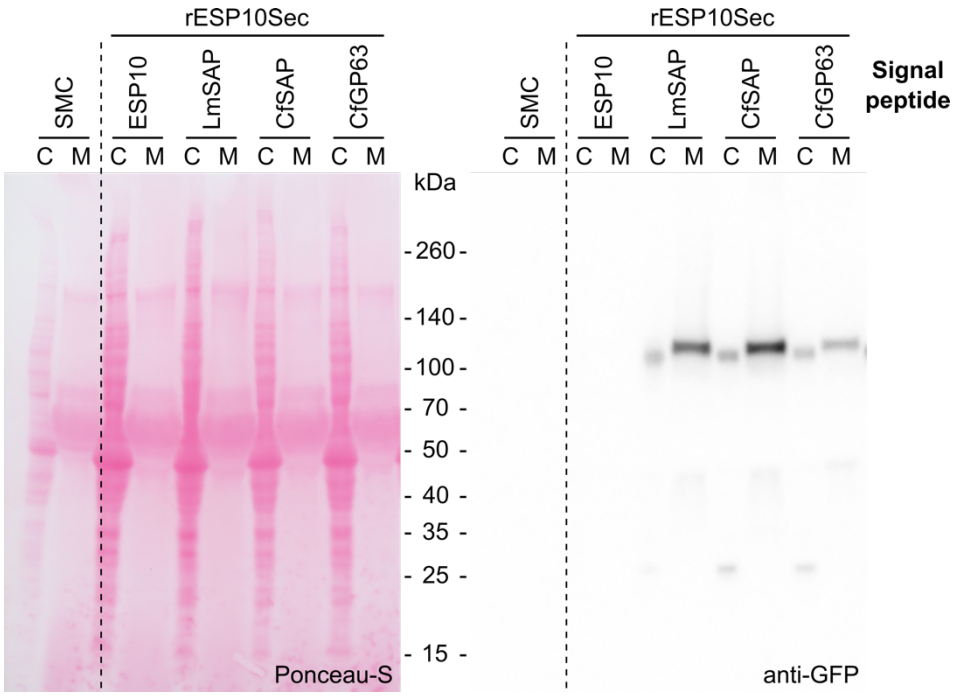


Figure 6-4 Efficiency of secretion from *C. fasciculata* can be effected by the choice of signal peptide. Ponceau-S stained membrane and anti-GFP immunoblot of SMC cells expressing recombinant *T. brucei* ESP10Sec as $sfGFP^{op}$ fusions, with either its endogenous signal peptide, the *L. mexicana* secreted acid phosphatase signal peptide (LmSAP), the *C. fasciculata* secreted acid phosphatase signal peptide (CfSAP), or the *C. fasciculata* GP63 signal peptide (CfGP63). Samples consist of whole cell lysates (C) and acetone precipitated media (M). The ESP10 signal peptide is not

recognised by *C. fasciculata*. The LmSAP and CfSAP signal peptides lead to similar levels of secretion, whilst the CfGP63 signal peptide yielded lower level.

6.2.3. Growth conditions for optimal yield of secreted proteins from *C. fasciculata*

As cells reach top density, recombinant protein will still be moving through the secretory pathway (as emphasised by the signal seen in the cellular fractions of Figure 5-15, Figure 6-2 and Figure 6-4). Thus, yield of secreted protein may be increased if cultures are left at top density for a time to allow the secretory pathway to empty of recombinant protein into the culture medium.

To test culture conditions required for obtaining optimal yields of secreted proteins from *C. fasciculata*, cells expressing rESP10Sec were grown, in agitated cultures, to $2x10^8$ cells/mL and left for 24, 48 and 72h before acetone precipitated media samples were made. In parallel to this, another culture was grown to top density, and each day a sample was taken before the cells were harvested and resuspended in fresh media at $2x10^8$ cells/mL. Protein levels were then quantified by anti-GFP immunoblot (Figure 6-5A). The blot revealed cells continued to secrete rESP10Sec across 72 h, but that the amount of rESP10Sec secreted after 48 h was negligible. From this point onwards, cells secreting recombinant proteins were therefore left for 48 h after reaching top density for further experiments.

To quantify the level of rESP10Sec secreted from SMCs after 48 h at top density, samples were resolved, along with GFP of known concentrations, by SDS-PAGE (Figure 6-5B). The amount of rESP10Sec that was secreted was equivalent to 200 µg/L.

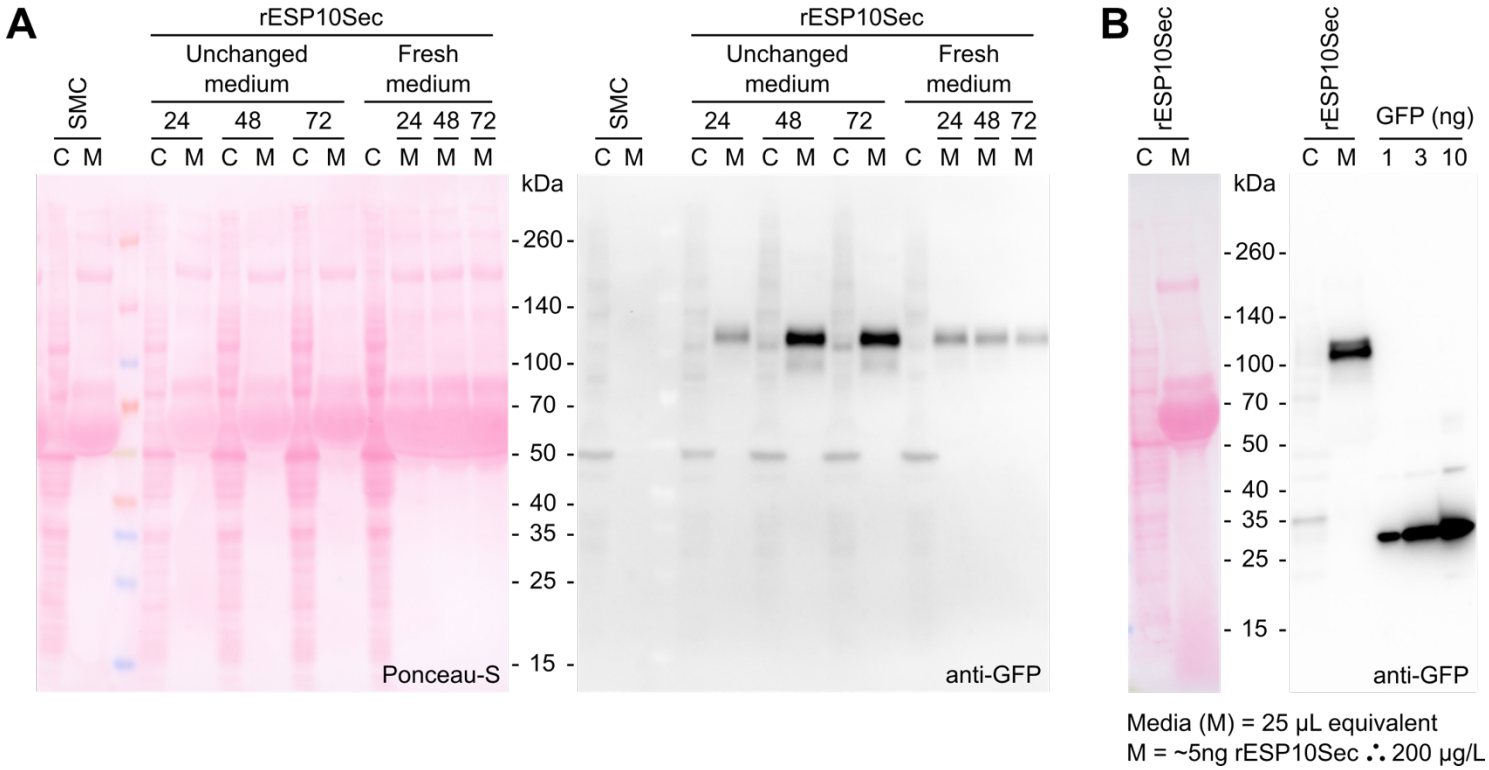


Figure 6-5 *C. fasciculata* secretes rESP10Sec for 72 h during stationary phase of growth. A Ponceau-S stained membrane and anti-GFP immunoblot of SMC cells expressing recombinant *T. brucei* ESP10Sec as a sfGFP^{op} fusion. Cells were either left in stationary phase (2x10^8 cells/mL) in the same medium for 24, 48 and 72 h, or cells were resuspended (at 2x10^8 cells/mL) in fresh medium every 24 h. Samples consist of whole cell lysates (C) and acetone precipitated media (M). **B** rESP10Sec culture samples resolved next to known amounts of eGFP for protein quantification. Quantity of rESP10Sec in M is equivalent to approximately 200 µg/L.

6.2.4. Codon bias and optimisation – using crithidial ribosomal codon bias

Despite the ability of *C. fasciculata* to secrete processed recombinant *T. brucei* surface proteins, based on rESP10Sec, the quantities obtainable would still require ~10 litres of cell culture to reach the desired amount of protein (2–3 mg prior to purification). However, as *C. fasciculata* appears to have the ability to correctly modify these proteins (Figure 6-2), it may still be advantageous to utilise this system. Thus, in another attempt to increase the recombinant protein levels produced in SMC cells, codon optimisation was revisited.

As shown in Chapter 4, codon optimisation has a large effect on the levels of sfGFP^{op} in *L. tarentolae.* Following the same strategy, the *C. fasciculata* ribosomal protein codon bias was used to calculate RSCU values. The *sfGFP^{op}* sequence in the pCEx vectors is the same as used in pLEXSY_IE-sfGO-N and, as mentioned in Chapter 5, is already optimised according to the crithidial ribosomal protein codon bias (Figure 6-6A). Instead, *ESP10* was used to test the effect of codon optimisation on protein levels in *C. fasciculata.* Figure 6-2 revealed that *T. brucei* signal peptides were not recognised in *C. fasciculata*, but the use of an alternative signal peptide in combination with rESP10FL had not yet been addressed. As such, an *ESP10* codon optimised sequence was designed for *ESP10FL^{op}*, using the LmSAP signal peptide, with the intention of testing the yield of recombinant protein both membrane-bound and secreted (*ESP10Sec^{op}*) (Figure 6-6B). In designing the sequence, some sub-optimal codons had to be used to allow successful DNA synthesis.

Initial attempts to clone *ESP10FL^{op}* and *ESP10Sec^{op}* into pCExC failed to produce constructs with the correct sequences (more details later). Due to problems encountered during cloning, two other ORFs were also codon optimised; *ESP13Sec* and *ESP14Sec* (again using the LmSAP signal peptide) (Figure 6-6C and D). Whilst ESP13 localises to the whole cell surface of *T. brucei*, ESP14 is an abundant parasite-specific transmembrane protein that, like ESP10, localises to the flagellar pocket (Gadelha *et al.*, 2015). As smaller proteins than ESP10, these two proteins could be quickly and cheaply synthesised, and were less likely to cause issues in cloning. Like *ESP10FL^{op},*

some sub-optimal codons had to be used to allow successful DNA synthesis of *ESP14Sec*^{op} (Figure 6-6D).

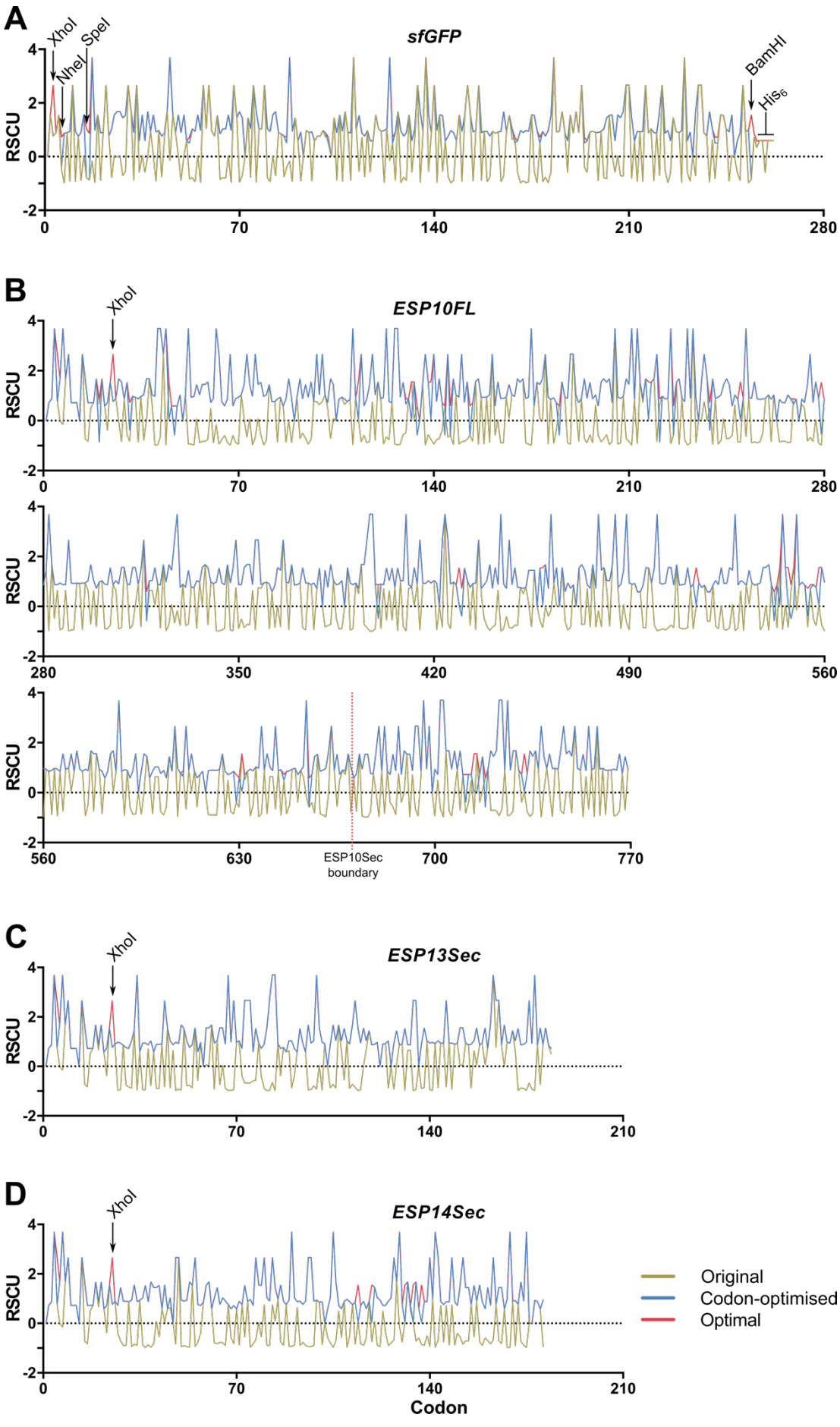

A
sfGFP
XhoI
NheI
SpeI
BamHI
His6
RSCU
B
ESP10FL
XhoI
ESP10Sec boundary
C
ESP13Sec
XhoI
D
ESP14Sec
XhoI
Codon
Original
Codon-optimised
Optimal

Figure 6-6 Comparison of RSCU values across different ORFs according to the *C. fasciculata* ribosomal codon bias. Values shown for **A** *sfGFP^{op}* **B** *ESP10FL* **C** *ESP13Sec* **D** *ESP14Sec*. **B-D** all contain the LmSAP signal peptide. Codons used across the original sequences are shown in yellow. The optimal choice of codons according the *C. fasciculata* ribosomal codon bias are shown in red. The codons used for DNA synthesis of the codon-optimised sequences are shown in blue. Sub-optimal codons required for restriction enzyme sites for cloning and the His_6 tag are indicated.

As mentioned above, initial attempts to clone *ESP10FL^{op}* and *ESP10Sec^{op}* sequences failed. The strategy for cloning *ESP10FL^{op}* was to directly digest the synthesised gBlock DNA and ligate the fragment into pCExC. The strategy for cloning *ESP10Sec^{op}* was to amplify an *ESP10Sec^{op}* fragment from the *ESP10FL^{op}* gBlock DNA by PCR, prior to digestion and ligation into pCExC. Multiple attempts to amplify *ESP10Sec^{op}* yielded no PCR product. Meanwhile, attempts to ligate *ESP10FL^{op}* into pCExC led to very low bacterial transformation efficiency of XL1-Blue *E. coli*. Of the bacterial colonies that did grow, 6 were screened by diagnostic restriction enzymes, revealing 3 clones with an insert of the correct size (the other 3 clones containing no insert). However, all three of these clones were shown, upon sequencing, to have deletions of different lengths within the same 45 bp region (Figure 6-7A XL1B clones A-C).

To check if the 45 bp region was present in the initial *ESP10FL^{op}* sequence prior to cloning, the synthesised fragment was Sanger sequenced, revealing the region to indeed be present (Figure 6-7A gBlock). It is possible that the region was forming secondary structure, leading to its removal upon cloning in bacteria. This secondary structure could also explain the difficulty in amplifying the *ESP10Sec^{op}* fragment. An attempt was made to clone *ESP10FL^{op}* in SURE2 *E. coli* instead. SURE2 cells lack proteins involved in the rearrangement and deletion of DNA secondary structures. However, these cells did not lead to successful cloning of *ESP10FL^{op}*. The transformation efficiency was still very low. After screening 20 clones by PCR across the insert region of pCExC, only two clones were show to contain an insert of the correct size, and these were revealed to be lacking in the same 45 bp DNA region upon sequencing (Figure 6-7A, SURE2 clones A and B).

As a final attempt to clone *ESP10FL^{op}*, the 45 bp region was modified by PCR of the two halves of *ESP10FL^{op}*, using primers with modified tags that covered the 45 bp region (Figure 6-7, “Modified region”). This modification did allow the successful cloning of *ESP10FL^{op}*. In addition, it allowed both the amplification of *ESP10Sec^{op}* from *ESP10FL^{op}*, and its subsequent cloning. Modification of the region necessitated the use of some sub-optimal codons. The change in RSCU values for the codons used across this region is shown in Figure 6-7B.

No issues were encountered when cloning *ESP13Sec^{op}* and *ESP14Sec^{op}*. In addition to the codon-optimised sequences, the original sequences for *ESP13Sec* and *ESP14Sec* from *T. brucei* gDNA were cloned into pCExS as a comparison, to assess the effect of codon usage on recombinant protein level.

With the desired sequences cloned into the pCEx vectors, these were all transfected into SMC cells as circular DNA. After drug selection of the transfected populations, these lines were grown to top density and whole cell lysates (C) and acetone precipitated media (M) prepared. Recombinant protein levels were then assessed by anti-GFP immunoblot (Figure 6-8). Codon-optimisation did not lead to significantly higher levels of the recombinant proteins. rESP10FL^{op} did express to higher levels than rESP10FL. However, unlike rESP10FL^{op}, rESP10FL contained its endogenous signal peptide, and so the difference in protein level seen is likely due to signal peptide recognition. The rESP10FL^{op} protein level was not as high as the rESP10Sec levels. The results show that, for this set of proteins at least, codon optimisation does not significantly affect recombinant protein levels in *C. fasciculata* SMC cells. Despite this, rESP14Sec showed much stronger expression levels relative to the other recombinant proteins. This was assessed by immunoblot to be ~800 µg/L (data not shown).

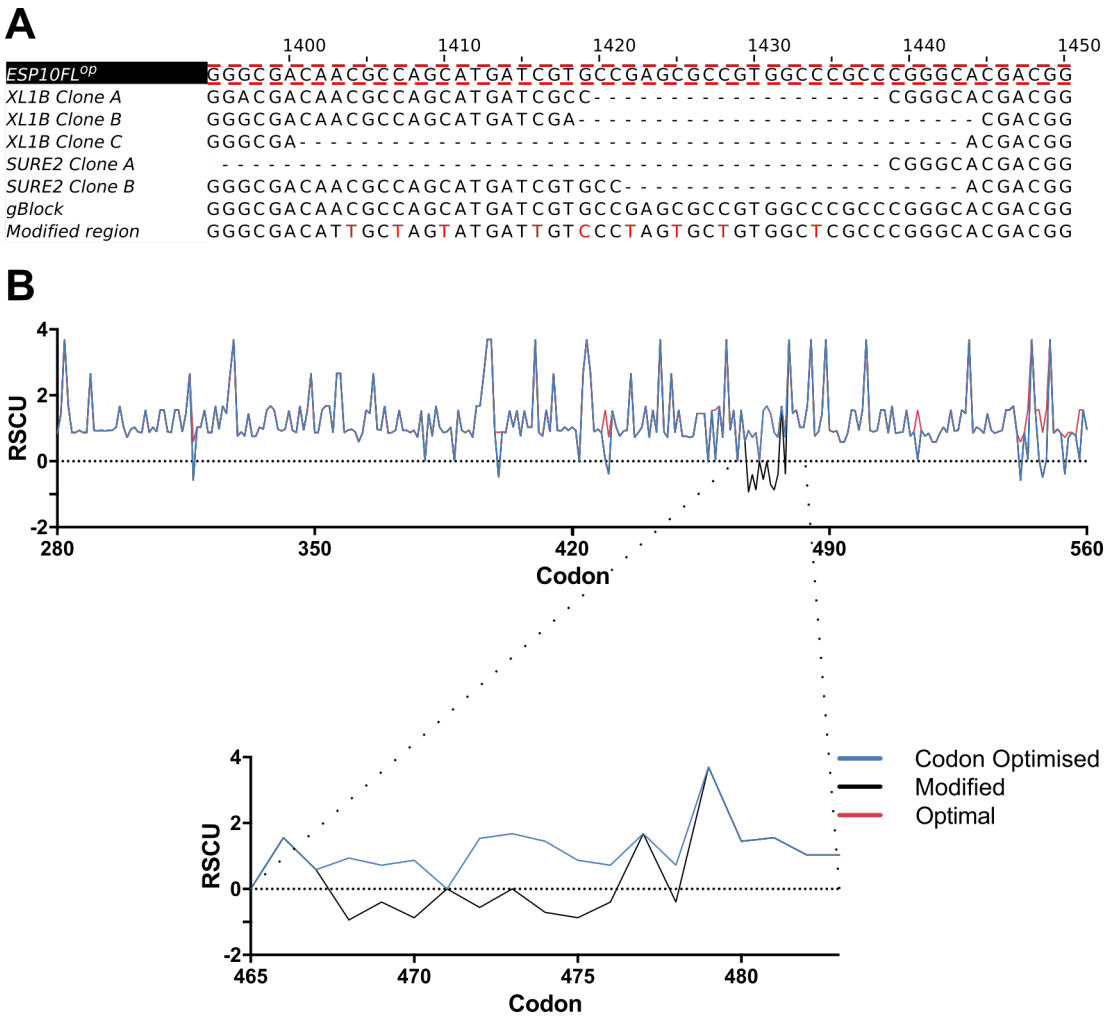
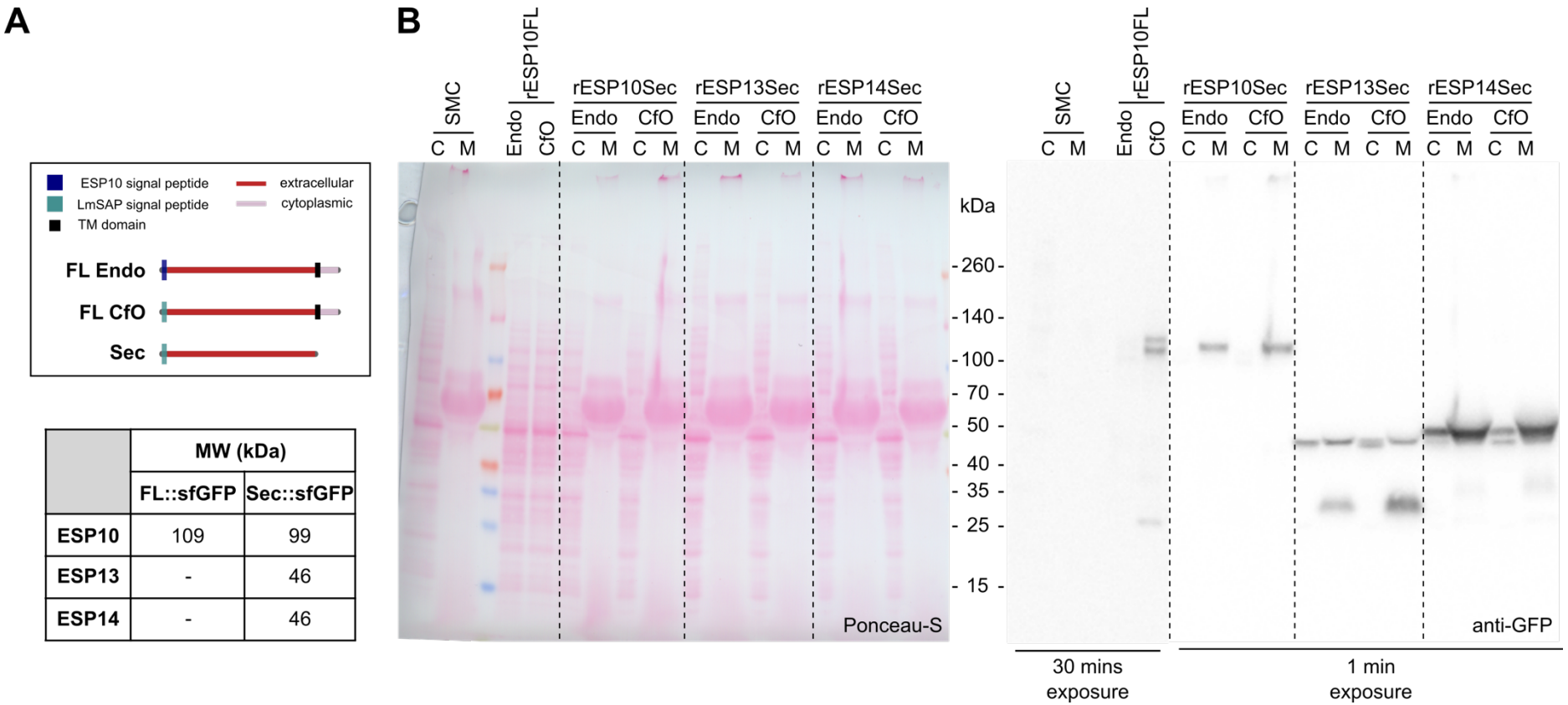


Figure 6-7 A 45 bp region inhibits cloning of *ESP10FL^{op}*. A Multiple DNA sequence alignments across a region of *ESP10FL^{op}* which hindered cloning. Sequences obtained from Sanger sequencing of the synthesised DNA fragment (gBlock), and of plasmid DNA isolated from bacterial clones. XL1B and SURE2 indicate different *E. coli* cloning strains. “Modified region” indicates an XL1B clone that retained the 45 bp region as a result of modification by PCR. Modified bases indicated in red. **B** Comparison of the RSCU values across the region in **A** according to the *C. fasciculata* ribosomal protein codon bias. The codon optimised sequence is shown in blue, the modified fragment sequence is shown in black and the optimal sequence is shown in red.



	MW (kDa)	
	FL::sfGFP	Sec::sfGFP
ESP10	109	99
ESP13	-	46
ESP14	-	46

Figure 6-8 Codon optimisation does not significantly increase recombinant protein expression levels in *C. fasciculata*. Ponceau-S stained membrane and anti-GFP immunoblot of SMC cells expressing recombinant *T. brucei* ESP10FL and Sec, ESP13Sec and ESP14Sec, fused to sfGFP^{op}. The proteins are expressed using either their endogenous codon sequence (Endo) or using a codon optimised sequence based on the *C. fasciculata* ribosomal codon bias (CfO). All recombinant proteins have the LmSAP signal peptide, except for ESP10FL (Endo) which has its endogenous signal peptide. **A** Top panel shows topology schematics of full length (FL) and secreted (Sec) proteins. Table contains predicted molecular weights of each of the fusion proteins. **B** Immunoblot, consisting of whole cell lysates (C) and acetone precipitated media

(M). The blot shows that for this set of proteins at least, codon optimisation does not significantly affect recombinant protein expression levels in *C. fasciculata* SMC cells.

6.2.5. Secretion of recombinant *T. brucei* proteins from *L. tarentolae*

In 2015, Rooney *et al* reported secretion of recombinant *T. brucei* surface proteins from *L. tarentolae* in quantities up to 10 mg/L. My attempts to secrete other *T. brucei* proteins from *C. fasciculata* did not reach these levels and were as much as 50-fold lower. As a control for secreted recombinant protein expression, a request was sent to Rooney *et al* for their expression vectors, which they kindly provided (pLEXSY_hyg2_ISG65 and pLEXSY_hyg2_LiTat1.3). The vectors required selection with hygromycin B. The T7TR strain in our lab already encodes *HYG*^{R} for background selection. Thus, to enable selection of these vectors in T7TR, the drug resistance was replaced with NEO^R, for selection with G418 (creating pLEXSY_neo2_ISG65 and pLEXSY_neo2_LiTat1.3). No further modifications were made to these two plasmids.

As mentioned in the Chapter 4 discussion, the vectors used by Rooney *et al* were designed to integrate into the rDNA locus, relying on PolI readthrough transcription. As such, these vectors were linearised and transfected into T7TR cells prior to clonal selection with G418. As the recombinant proteins encoded in these vectors contained no GFP fusion, 3 clones of each were instead screened for recombinant protein levels by anti-His_6 immunoblot (data not shown). The highest expressing clones then had their protein levels quantified through resolution by SDS-PAGE next to His_6-eGFP of known concentrations followed by an anti-His_6 immunoblot (Figure 6-9). Both rLiTat1.3Sec and rISG65Sec were expressed to levels equivalent to >12mg/L, agreeing with the levels reported by Rooney *et al.*

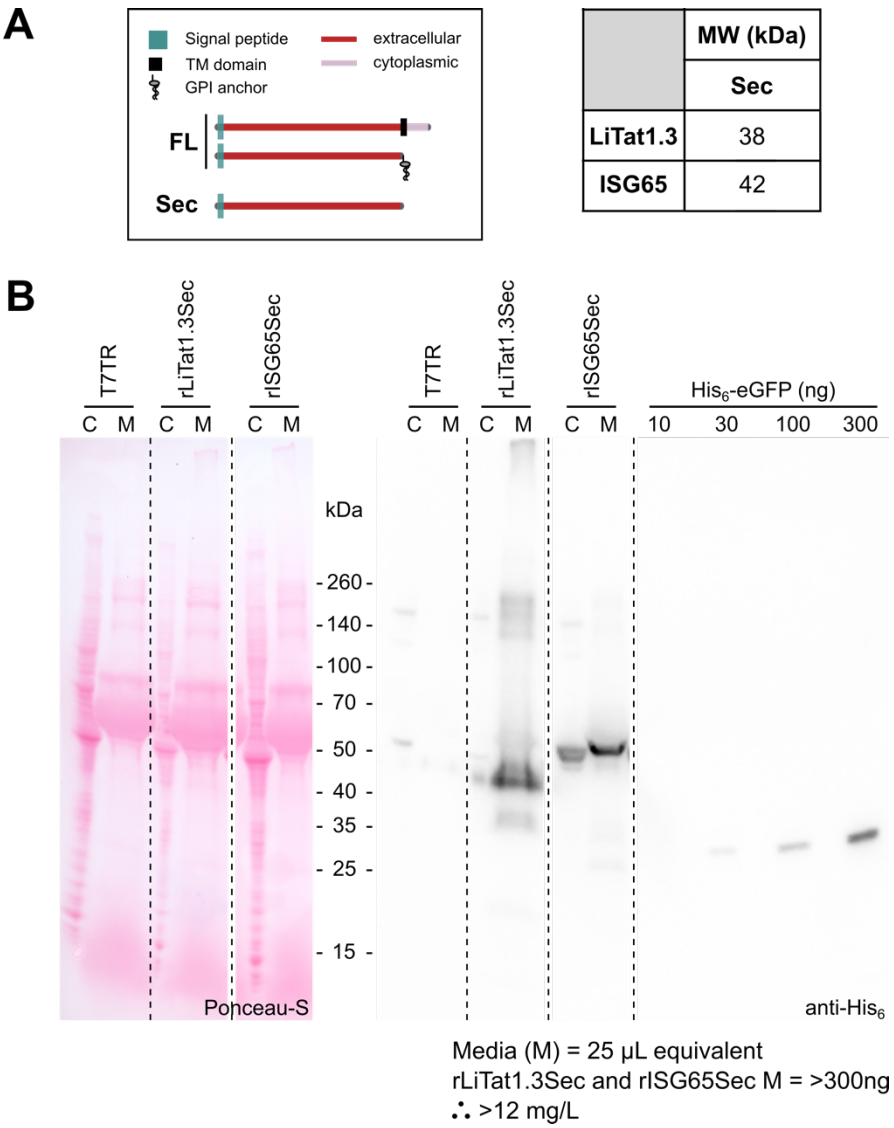


Figure 6-9 *L. tarentolae* secretes over 12 mg/L of rLiTat1.3Sec and rISG65Sec. Ponceau-S stained membrane and anti-His_6 immunoblot of T7TR cells expressing recombinant *T. brucei* surface proteins. **A** Left panel shows topology schematics of full length transmembrane and GPI-anchored proteins (FL) and secreted (Sec) proteins. Table contains predicted molecular weights of each of the secreted proteins. **B** Immunoblot consisting of whole cell lysates (C) and acetone precipitated media (M). Known concentrations of His_6-GFP were loaded for quantification of protein levels. rLiTat1.3 and rISG65 reached levels equivalent to >12mg/L.

The high yield of rVSG and rISG65 secreted by *L. tarentolae* could be protein specific. VSG and ISG65 are two major surface proteins of *T. brucei*; as such, they may have evolved to be suited to high expression, being easier to fold and process. Alternatively, the high yield could be due to the lack of a GFP fusion on these recombinant proteins, leaving them only with a small disordered His_6-tag as opposed to the large ordered barrel of GFP, potentially allowing easier

folding and processing. Thus, an experiment was designed to see if these expression levels could be matched in *C. fasciculata*, through the expression of rISG65Sec, with and without a sfGFP^{op} fusion. To enable this *sfGFP^{op}* was removed from pCExS (creating pCExS2), allowing secreted expression of recombinant proteins with no sfGFP^{op} fusion. *ISG65Sec* was cloned into both pCExS and pCExS2.

Rooney *et al* used an ISG65 (Tbg.972.2.1720) from *T. b. gambiense*, whereas in this project, the ISG65 comes from *T. b. brucei* (Tb427_020015800). These two are similar but with some differences (72% identities and 82% positives upon alignment). As such, the *T. b. gambiense* ISG65 ORF from pLEXSY_hyg2_ISG65 was also expressed in *C. fasciculata* (pCExS2), and the *T. b. brucei* ISG65 from pCExS_ISG65Sec was expressed in *L. tarentolae* (pLEXSY_neo2) as added controls. The full list of recombinant ISG65s being tested, along with expression plasmid and host organism used, is shown in Table 6-1.

Table 6-1 Combinations of different recombinant ISG65s with their expression plasmids for testing expression levels in *C. fasciculata* or *L. tarentolae*.

Recombinant protein	Expression Plasmid	Host organism
Tbb ISG65Sec-sfGFP^{op}	pCExS	*C. fasciculata*
Tbb ISG65Sec	pCExS2	*C. fasciculata*
	pLEXSY_neo2	*L. tarentolae*
Tbg ISG65Sec	pCExS2	*C. fasciculata*
	pLEXSY_neo2	*L. tarentolae*

pCEx constructs were transfected as circular DNA into SMC cells and populations subjected to selection with Hygromycin B. pLEXSY constructs were linearised, transfected into T7TR, and independent clones selected for using G418. 3 clones from pLEXSY transfections were screened for recombinant protein levels by immunoblotting with anti-His_6 antibodies. The highest expressing clones were taken forward for comparison to the pCEx transfected populations (Figure 6-10). In the *C. fasciculata* whole cell lysates, a band is observed at ~120 kDa as a result of non-specific binding (based on the fact that it appears in the parental SMC line). Removing the sfGFP^{op} from ISG65Sec did

lead to an increase in protein levels in *C. fasiculata*, similar to the levels seen in *L. tarentolae*. However, the majority of the signal was detected in whole cell lysates rather than in the medium. Whether ISG65 derived from *T. b. brucei* or *T. b. gambiense* made no difference to protein levels, although there was a difference in apparent molecular weight between these two. This result shows that removing sfGFP^{op} from rISG65Sec can increase its production in *C. fasiculata*, but that the cells fail to secrete the protein.

A

	MW (kDa)	
	Sec	**Sec::sfGFP**
Tbb_ISG65	42.3	70
Tbg_ISG65	41.8	-

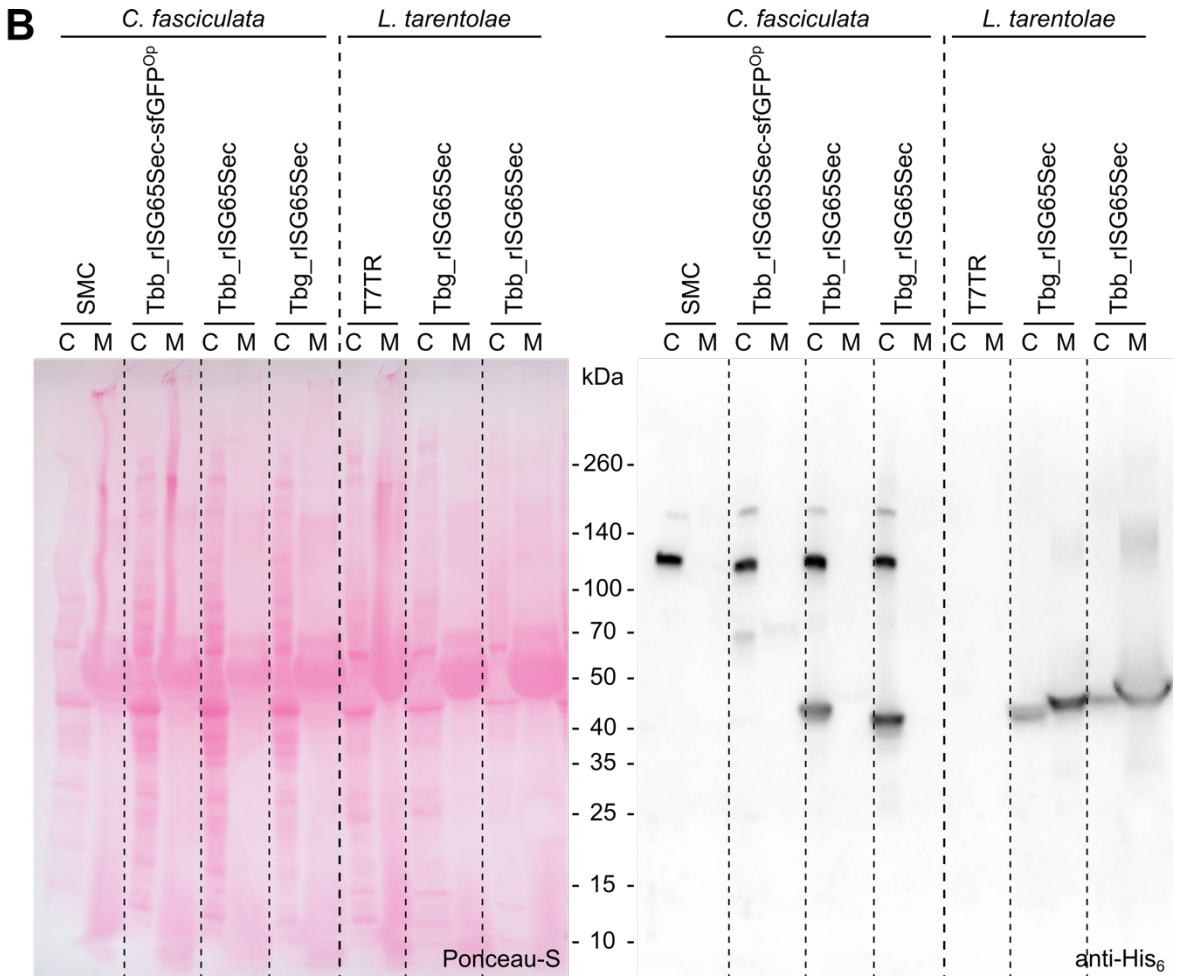


Figure 6-10 Removal of sfGFP from rISG65Sec increases protein levels but not secretion yield in *C. fascicualta.* Ponceau-S stained membrane and anti-His_6 immunoblot of SMC and T7TR cells expressing different forms of rISG65Sec. **A** Table indicates the different forms being expressed and their predicted molecular weights. **B** Immunoblot consists of whole cell lysates (C) and acetone precipitated media (M). Whole cell

lysates of *C. fasciculata* lines have a non-specific band at ~120 kDa. Removal of sfGFP increases rISG65Sec expression in SMCs to levels similar to those of T7TR cells, but protein is not detected in the culture supernatant.

Even when *C. fasciculata* express rISG65Sec-sfGFP^{op}, the majority of the signal is seen within the cellular fraction. In contrast, when secreting rESP10Sec-sfGFP^{op}, rESP13Sec-sfGFP^{op} or rESP14Sec-sfGFP^{op}, the majority of the protein is secreted and found in the medium fraction (Figure 6-8). Thus, the failure of *C. fasciculata* to secrete of rISG65Sec without sfGFP^{op} may not be down to the high levels of the protein produced, but instead be caused by a specific failure of *C. fasciculata* to efficiently process ISG65Sec. Based on this, more success might be had in attempting to use *C. fasciculata* to secrete rESP10Sec, rESP13Sec and rESP14Sec lacking their sfGFP^{op} fusions. Although it cannot be ruled out that the secretory pathway of *C. fasciculata* may simply have a low capacity unsuited to high-level protein secretion. If this is the case, *L. tarentolae* could be the better option for production of these proteins. Therefore, secretion of rESP10Sec, rESP13Sec and rESP14Sec lacking a GFP fusion was tested in both CExSy and LEXSY. For this purpose, rESP10Sec^{op}, rESP13Sec^{op} and rESP14Sec^{op} were cloned into pCExS2 and pLEXSY_neo2.

The pCExS2 constructs were transfected as episomes into SMC cells and populations selected with Hygromycin B. The pLEXSY constructs were linearised and transfected into T7TR, with clonal selection and screening for recombinant protein levels carried out as before. Clones and populations were then assessed by an anti-His$_6$ immunoblot (Figure 6-11). Once again, a non-specific band appears at ~120 kDa in the *C. fasciculata* cell lysates. Unlike rISG65Sec, removal of sfGFP^{op} from rESP10Sec, rESP13Sec and rESP14Sec did not increase protein amounts in SMCs, even causing a drop in the levels of ESP13Sec and ESP14Sec. Recombinant proteins of interest were expressed and secreted to much greater success in T7TR cells. rESP10Sec and rESP14Sec were secreted to particularly high levels, similar to those seen for rLiTat1.3Sec and rISG65Sec (>12 mg/L, Figure 6-10). rESP13Sec was produced to lower levels, yet its production still surpassed that achieved from SMCs.

In contrast to previous experiments, all of the recombinant proteins produced in SMCs showed a much stronger signal in the cellular fraction relative to the secreted media fraction (and relative to the His$_6$-eGFP ruler) than had been seen previously. A variable here compared to prior experiments, is that the blot

was probed with an anti-His_6 antibody as opposed to anti-GFP antibodies. There is evidence that different anti-His_6 antibodies can vary in specificity when detecting His-tags on different proteins, even once boiled in Laemmli and resolved by SDS-PAGE (Debeljak *et al.*, 2006). Debeljak *et al* probed immunoblots containing three different His-tagged protein – EPO, DHFR and SP56 (produced in CHO cells). They reported that a particular anti-His antibody detected all the tagged proteins, showing equal signal, whilst three other anti-His antibody detected DHFR and SP56 only, not detecting EPO at all. There was also variation in the relative DHFR and SP56 signals between the different antibodies. These results show that the context of a His_6-tag can have a large effect on detected signal. Thus, the irregularities observed in the SMC samples could be caused by the use of an anti-His_6 antibody for immunoblotting. This also suggests that the signal seen for the *L. tarentolae* samples may not be representative of the protein quantities. However, the immunoblot in Figure 6-9 (also probed with the same anti-His_6 antibody) did reveal that the signal detected for rVSGSec and rISG65Sec produced in *L. tarentolae* matched the protein amounts reported by Rooney *et al*. Thus, probing *L. tarentolae* samples with the anti-His_6 antibody used here may be more reliable.

As >1mg amounts of recombinant *T. brucei* surface proteins are required for vaccination experiments, the best option so far appears to be secretion from *L. tarentolae* cells. Despite this, *C. fasciculata* may still convey advantages relating to protein folding and processing that could be useful if a smaller amount of protein is required. It remains to characterise purification efficiency from both systems, and properties of recombinant proteins derived from each.

A

	MW (kDa)	
	Sec	**Sec::sfGFP**
ESP10	71	99
ESP13	18	46
ESP14	18	46

B

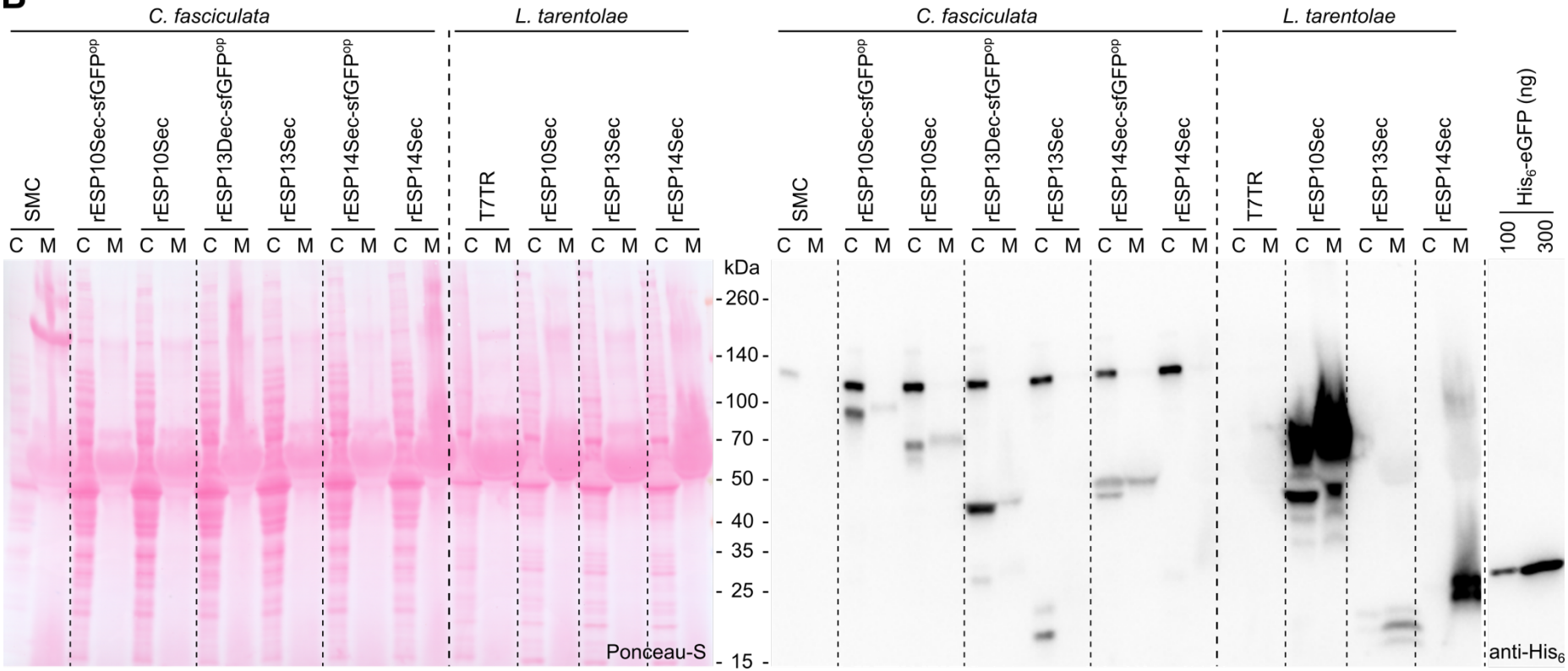


Figure 6-11 Recombinant *T. brucei* surface proteins are secreted to higher levels from *L. tarentolae* compared to *C. fasciculata*. Ponceau-S stained membrane and anti-His_6 immunoblot of SMC and T7TR cells expressing different recombinant *T. brucei* proteins in secreted (Sec) forms (no membrane association domains). **A** Table indicating the different forms being expressed and their predicted molecular weights **B** Immunoblot consist of whole cell lysates (C) and acetone precipitated media (M). Whole cell lysates of *C. fasciculata* lines have a non-specific band at ~120 kDa. Known concentrations of His_6-GFP were loaded for quantification of expression levels. rESP10Sec and rESP14Sec secreted from T7TR cells reached levels equivalent to >12mg/L.

6.3. Discussion

This chapter describes attempts to express *T. brucei* surface proteins in both CExSy and LEXSY. Expression of these proteins to high levels in CExSy proved challenging. Initial experiments suggested the lack of *T. brucei* signal peptide recognition by *C. fasciculata* to be a large contributing factor to low expression in these cells. Previous work showing *T. brucei* signal peptides not to be recognised by canine derived microsomes (Al-Qahtani *et al.*, 1998) can perhaps be explained by the unusual SRP in *T. brucei* (Lustig *et al.*, 2005). Other eukaryotic SRPs contain one RNA molecule and multiple associated proteins. The *T. brucei* SRP contains two RNA molecules and lacks proteins found in mammalian and yeast SRPs (SRP9 and 14 in mammals; Srp7p, Srp14p and Srp21p in yeast). These differences in SRP may dictate recognition of different signal peptides. However, this does not explain the inability of *C. fasciculata* to recognise and process *T. brucei* signal peptides, as *C. fasciculata* also lacks the same SRP proteins and has the second RNA molecule (based on orthologues found in the genome sequence).

The problem with *T. brucei* signal peptide recognition was overcome by testing different signal peptides from *L. mexicana* and *C. fasciculata*. Using signal peptides of a secreted acid phosphatase from either organism led to secretion into the culture medium. Use of a *C. fasciculata* GP63 signal peptide also led to secretion but to a lower amount. There is evidence that GPIAPs in *T. brucei* require post-translational translocation for entry into the ER (Goldshmidt *et al.*, 2008). If this is also the case for *C. fasciculata*, it might explain the lower efficiency of the GP63 signal peptide, as post-translational translocation could lead to build-up of overexpressed proteins in the cytoplasm, causing protein miss-folding, aggregation and degradation. Other steps could have been taken to positively impact secretion from these cells. Large throughput signal peptide screens have proven effective in other systems. For example, in yeast it has been shown that using a library of different endogenous signal peptides can cause differential secretion of β-galactosidase (LacA) (Mori *et al.*, 2015). Activity of the secreted enzyme was used as a read-out and varied from no activity to ~twice that of the secreted enzyme with its native signal peptide. With more time available, a screen of *C. fasciculata* signal peptides on

secretion efficiency could have been beneficial towards obtaining high level secretion from these cells.

An alternative approach is specific engineering of a single signal peptide. A study utilising LEXSY and the LmSAP to secrete antibody fragments, looked at modifying the signal peptide sequence and downstream restriction enzyme sites (and therefore downstream amino acids) in an attempt to improve secretion efficiency (Klatt and Konthur, 2012). Klatt and Konthur utilised SignalP 3.0 to predict the number and probability of likely cleavage sites within *in silico* DNA constructs. Their hypothesis was that a high predicted cleavage probability of a single cleavage site would improve secretion efficiency relative to prediction of multiple cleavage sites with lower probability. Four signal peptides were tested with different numbers of predicted cleavage sites and probabilities. Depending on the signal peptide used, the average secretion levels ranged from 0.04–4 mg/L, with the signal peptides containing cleavage sites of high predicted probability producing the highest levels of secretion. However, a large difference in secretion levels was seen (~3-fold) between the two best performing signal peptides in this study, both of which had similar predictions for cleavage probability. These two signal peptides only differed at the amino acid positioned -3 relative to the cleavage site (with the better performing of the two having alanine and the other having threonine). This is not unexpected, as it has been reported previously that small, non-charged amino acids are preferred in positions -1 and -3 relative to the signal peptide cleavage site and that alanine is more frequently used that threonine (von Heijne, 1983). It does highlight, however, that when using *in silico* analysis for signal peptide optimisation, both cleavage site probability and specific amino acids should be taken into account.

Whilst *in silico* analysis can prove effective, it does come with the caveat that attaching signal peptides to different proteins can alter cleavage predictions and lead to a sub-optimal signal peptide-protein combination. Especially as it has been shown that amino acids downstream of the cleavage site play a role in efficiency of cleavage (Güler-Gane *et al.*, 2016). As such, optimising the signal peptide cleavage site to be compatible with restriction sites for cloning, and with a specific recombinant protein, may not be a practical option, as further modification may be required for each different protein expressed. One option

would be to utilise Gibson assembly for cloning signal peptide-protein combinations, allowing easy modification of the sequence around the cleavage site without this being dictated by restriction sites.

Another aspect to consider when dealing with secretory expression is the 3' UTR. In yeast it has been reported that two 3' UTRs (from the small plasma membrane proteins *PmpI* and *PmpII*) play a role in targeting mRNAs, and their associated translating ribosomes and nascent proteins, towards the ER membrane (Chartron *et al.*, 2016; Loya *et al.*, 2008). In addition, these UTRs play a role in recruitment of SRP, as mRNA encoding GFP (no signal peptide) and containing either of these UTRs could be co-immunoprecipitated with one of the components of SRP (Chartron *et al.*, 2016). These UTRs have subsequently been used to improve secretion of recombinant proteins in yeast, leading to a 2.3-fold increase in specific activity of a secreted bacterial endoglucanase (Besada-Lombana and Da Silva, 2019). Whilst no such UTRs have been studied in trypanosomatids, a screen of 3' UTRs from different secreted proteins may prove effective in *C. fasciculata*.

The drawback to 3' UTR modification would be potentially lower expression levels due to the nature of trypanosomatid mRNA processing. However, tuning protein levels for optimal secretion is a strategy in itself. In yeast, overexpression of proteins does not always lead to an increase in secreted product, and can even cause lower overall secretion levels due to saturation of the secretory pathway, build-up of misfolded proteins, and induction of the unfolded-protein response (including ERAD) (Love *et al.*, 2012; Shusta *et al.*, 1998). Thus, using tuneable expression can increase yield (Shusta *et al.*, 1998). This evidence that secretion capacity effects secretion yield may also shed light on the lack of an effect conveyed by codon optimisation on the secretion of *T. brucei* proteins from *C. fasciculata*. If the secretion capacity for these proteins had already been saturated, then further codon optimisation would not lead to higher yield. No extra signal was detected in the cellular fractions of lines expressing the codon optimised proteins (Figure 6-8).

To overcome secretory bottlenecks, other strategies have been explored in yeast involving engineering strains in an attempt to improve secretion capacity (Besada-Lombana and Da Silva, 2019; Kim *et al.*, 2014; Shusta *et al.*,

1998). These approaches have included increasing ER size, reducing the level of ERAD and retro-translocation, overexpression of ER chaperones, removal of proteases, and increasing the capacity for forward transport of proteins from the ER. Whilst all of these are novel strategies with the potential to increase secretion capacity, they each could encompass an entire project on their own and were not explored here.

Despite the range of possible avenues for improving recombinant *T. brucei* protein secretion efficiency in *C. fasciculata*, currently *L. tarentolae* has proven to be a better choice for secretion of higher amounts of these proteins into the culture medium. It was pointed out in Chapter 4 that *L. tarentolae* lacks calreticulin, an otherwise universal trypanosomatid ER chaperone involved in the folding of glycosylated proteins (Raymond *et al.*, 2012). This was one of the factors that led to the suggestion that this system may not be the optimal choice for secretion of recombinant trypanosomatid proteins and thus the setup of CExSy. However, it might be that *L. tarentolae* has a more streamlined secretory pathway than other trypanosomatids, providing a greater capacity for protein folding, processing and downstream secretion.

C. fasciculata might still be suitable to applications other than vaccination studies that require lower amounts of protein (such as diagnostics, and structural studies by cryoEM; see Chapter 8). The results in this chapter suggest that *C. fasciculata* may correctly modify *T. brucei* surface proteins (Figure 6-2). As such, attempts were made to purify proteins from both *L. tarentolae* and *C. fasciculata* for further characterisation and comparison. This is described in the next chapter.

Chapter 7. Protein Purification

7.1. Introduction

L. tarentolae produces higher levels of recombinant *T. brucei* surface proteins than *C. fasciculata*. However, for applications that require limited amounts of protein, *C. fasciculata* may convey some advantages over *L. tarentolae* in terms of protein folding and processing; in particular regarding N-glycosylation. *L. tarentolae* has been shown to carry out paucimannose-type and complex-type (Breitling *et al.*, 2002; JenaBioscience), whilst *C. fasciculata* has been shown to carry out oligomannose-type (Parodi, 1993). It is unknown whether *C. fasciculata* can also generate the two former types of glycosylation. If it can, then use of CExSy could be useful for production all *T. brucei* proteins, albeit currently at low yield. If not, CExSY may still be useful for proteins with oligomannose-type glycosylation. This chapter describes attempts made to purify recombinant proteins from both LEXSY and CExSy for further characterisation that addresses the above question. ESP10, 13 and 14 were selected for direct comparison between the two systems. ISG65 was also used as a LEXSY expression positive control. For clarity, proteins produced in CExSy are referred to with a -Cf suffix, whilst proteins produced in LEXSY are referred to with an -Lt suffix (e.g. rESP10Sec-Cf and rESP10Sec-Lt). All CExSy recombinant proteins have a sfGFP^{op} fusion, whilst LEXSY recombinant proteins do not (unless stated).

7.2. Results

7.2.1. Selective protein precipitation by ammonium sulphate

For purification of ESP10, 13, 14 and ISG65 from *C. fasciculata* or *L. tarentolae*, a two-step approach was taken. Firstly, recombinant protein secreted into the culture medium were ‘salted out’ using ammonium sulphate. Subsequently, these were affinity purified via immobilized metal affinity chromatography (IMAC) using nickel-nitrilotriacetic acid (Ni-NTA) sepharose. This method couples crude purification from large volumes of tissue culture

supernatant to specific isolation of proteins containing a His_6 tag. This approach was particularly useful in this instance, as secreted recombinant proteins were present at low amount per litre of cell culture.

Salting-out is based upon the dependence of protein solubility on the salt concentration of a solution (Duong-Ly and Gabelli, 2014; Wingfield, 2001). At low concentrations of salt, anions and cations will interact with charged residues on protein surfaces, neutralising them and preventing aggregation. As the concentration of salt gets higher, the proteins will begin to minimise their hydrophobic surface area through tighter folding and aggregation, releasing bound water molecules and increasing the entropy of the system. This aggregation leads to protein precipitation. The percentage saturation of salt required for protein precipitation is protein-protein dependant, aiding in protein purification through selective precipitation. Ammonium sulphate is commonly used for this type of purification, as it is highly soluble, and both ions (NH_4^+ and SO_4^{2-}) are early members of the respective cation and anion Hoffmeister series, which dictate which ions are most proficient at salting-out proteins. In addition, salts that convey a strong salting-out effect tend to stabilise tertiary structure through the tighter folding of the precipitated proteins.

Initial tests of ammonium sulphate precipitation from culture medium were carried out on recombinant proteins secreted from *C. fasciculata*. Tests were carried out as follows. SMC cells expressing recombinant proteins were grown to $2x10^8$ cels/mL under agitation, incubated for a further 48 h, and then used as such: cells and medium were separated by centrifugation, and sampled for SDS-PAGE and immunoblot analysis (samples ‘C’, cells; sample ‘M’, medium) as described in previous Chapters. The medium was transferred to at 4ºC, and ammonium sulphate was added slowly to the desired percentage saturation with gentle stirring. Once fully dissolved, the solution was left stirring for 30 minutes at 4ºC. Precipitated proteins were harvested by centrifugation, resuspended in PBS and prepared for SDS-PAGE. Additional ammonium sulphate was then added to the cleared medium up to the next desired percentage saturation for a second precipitation step. And so forth.

Initial tests used SMCs secreting rESP10Sec-Cf. The culture medium was separated from cells and divided into two samples, and each was subjected to

three rounds of precipitation using 40, 50, and 60, or 30, 60 and 90% saturation of ammonium sulphate. Laemmli samples of equivalent proportions were resolved by SDS-PAGE and then probed with anti-GFP antibodies (Figure 7-1A). This revealed rESP10Sec-Cf to precipitate at 60% ammonium sulphate saturation. Most medium proteins precipitate at 50% or >60% saturation. Two bands were observed for secreted rESP10Sec-Cf rather than one as had been seen previously, likely due to incomplete denaturation of the protein.

The same approach was tested to precipitate rESP13Sec-Cf and rESP14Sec-Cf (Figure 7-1B and C). For both proteins, this was done with 40, 50, 60, 70 and 90% saturation sequentially. 40% was the lowest saturation used as the previous test showed that very few medium contaminants precipitate below 50% ammonium sulphate saturation. Both recombinant proteins largely precipitated at 60% saturation, much like rESP10Sec-Cf.

Elsewhere sfGFP has been used as a solubility enhancer for protein production in *E. coli* (Wu *et al.*, 2009). Thus, the similar precipitation results of rESP10Sec-Cf, rESP13Sec-Cf and rESP14Sec-Cf could be due to sfGFP^{op} conveying comparable solubility properties to each of the different recombinant proteins, particularly as signal resulting from cleaved sfGFP^{op} observed in the rESP13Sec-Cf samples appears in the same fractions as the intact protein (Figure 7-1B). This similar behaviour simplifies purification as the same protocol can be used for the precipitation of all three recombinant proteins – i.e. a first cut of 50% ammonium sulphate saturation for removal of contaminants, and a second cut of 60% for recombinant protein precipitation.

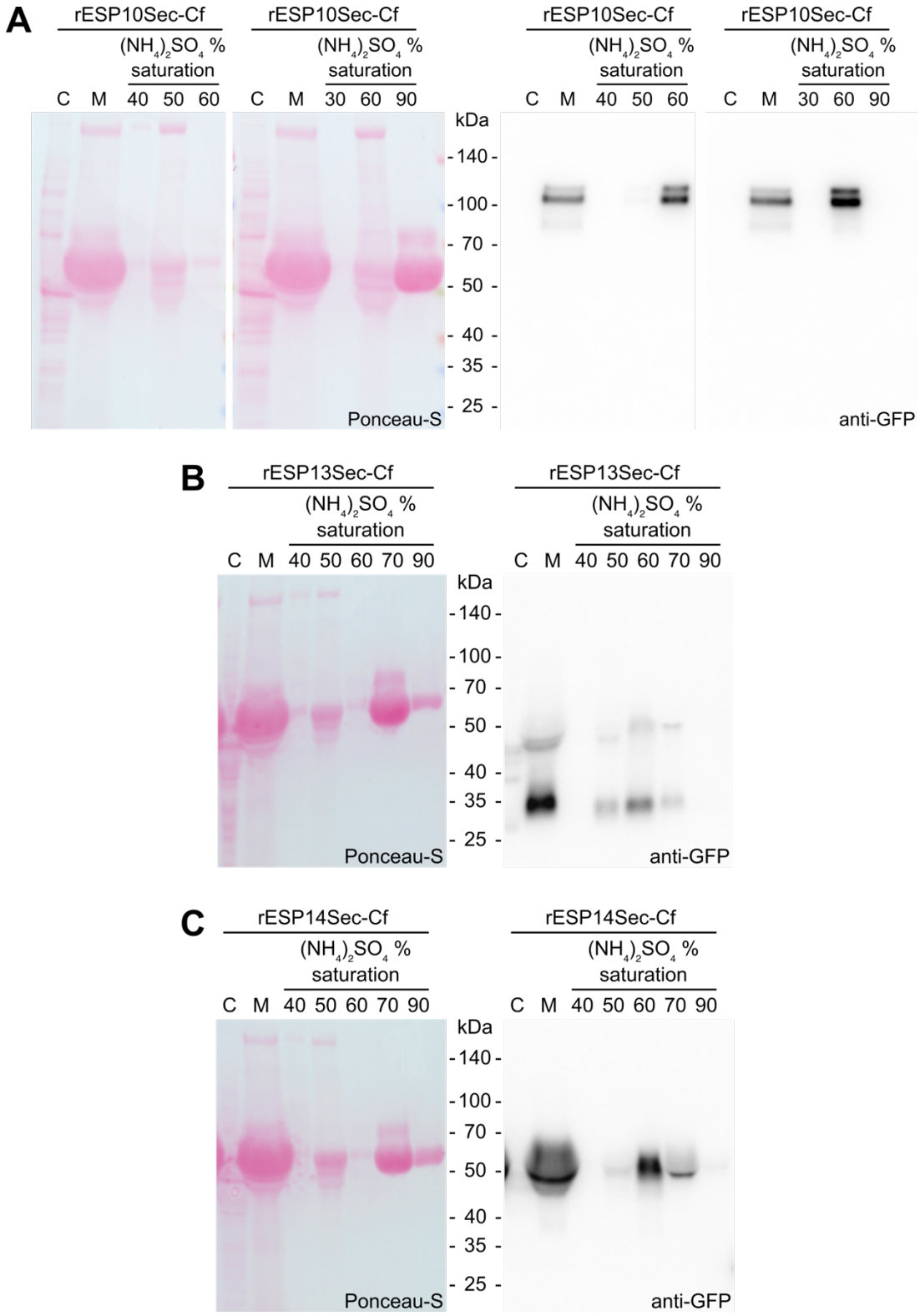


Figure 7-1 Recombinant proteins secreted from SMCs precipitate from culture medium at 60% ammonium sulphate saturation. Ponceau-S stained membranes and anti-GFP immunoblots of SMC cells secreting recombinant *T. brucei* proteins as sfGFP^{op} fusions – **A** rESP10Sec-Cf, **B** rESP13Sec-Cf, and **C** rESP14Sec-Cf. Samples include whole cell lysates (C), acetone-precipitated media (M), and ammonium sulphate-precipitated samples made through sequential saturation of culture medium with increasing amounts of ammonium sulphate. All three recombinant proteins largely

precipitate at 60% saturation (including sfGFP^{op} cleaved from rESP13Sec-Cf), whilst medium components precipitate at 50% and >60% saturation.

In *L. tarentolae*, the same candidate protein set above, plus the additional positive control rISG65Sec-Lt, are not expressed as fusions to sfGFP^{op} and may be post-translationally modified or glycosylated differently to proteins expressed in *C. fasciculata*. Like GFP, N-glycosylation can also impact on protein solubility, because N-glycans consist of hydrophilic molecules that interact with water (Varki, 2017). Thus, the ammonium sulphate precipitation procedure described above was also tested on recombinant proteins produced by *L. tarentolae*. Figure 7-2 shows the result of an initial test using 40, 50 and 60% saturation (70% and 90% were also tested but no further recombinant protein precipitated at these saturation percentages – data not shown). Here, recombinant proteins precipitated across multiple fractions. This indicates that the same proteins may be differentially folded or processed within each species, causing different solubilities and presenting a potential drawback of the LEXSY system. However, it cannot be ruled out that this differential precipitation would also have occurred to proteins produced in CExSy had they been expressed with no sfGFP^{op} fusion. Based on the result, purification of proteins secreted from *L. tarentolae* would be done using 60% ammonium sulphate saturation.

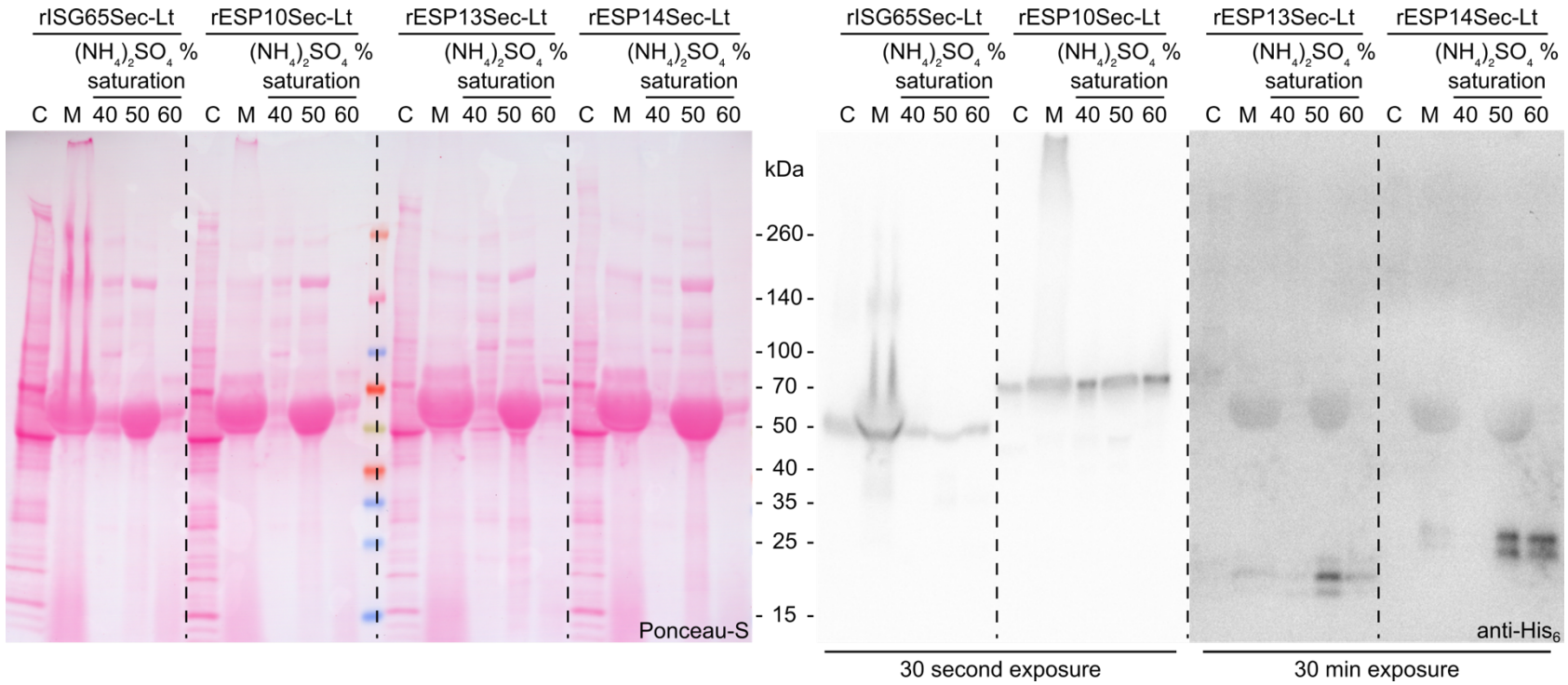


Figure 7-2 Recombinant proteins secreted from T7TR cells precipitate from culture medium across 40–60% ammonium sulphate saturation. Ponceau-S stained membrane and anti-His_6 immunoblot of T7TR cells secreting recombinant *T. brucei* proteins – rISG65Sec-Lt rESP10Sec-Lt, rESP13Sec-Lt and rESP14Sec-Lt. Samples include whole cell lysates (C), acetone-precipitated media (M), and ammonium sulphate-precipitated samples made through sequential saturation of culture medium with increasing amounts of ammonium sulphate. All four recombinant candidates precipitate across 40, 50 and 60% saturation.

7.2.2. Nickel-NTA purification

After ammonium sulphate precipitation, the next step was pulldown of recombinant proteins using Ni-NTA agarose. Initial attempts used small-scale, 50 mL cultures of *C. fasciculata* lines expressing the candidate protein set. Based on prior amount estimation by immunobloting, 50 mL cultures of SMCs secreting these recombinant proteins should produce approximately 10, 5 and 40 µg of rESP10Sec-Cf, rESP13Sec-Cf and rESP14Sec-Cf respectively.

Cells were grown to top density ($2x10^8$ cels/mL) under agitation and maintained as such for a further 48 h, and ammonium sulphate precipitation carried out (using 50% and 60% saturation sequentially) on culture medium separated from cells by centrifugation. Proteins precipitated at 60% ammonium sulphate saturation were resuspended in 1 mL of binding buffer and incubated with an excess Ni-NTA agarose (for ease of handling, 20 µL of resin was used for each sample, equivalent to >10x excess) for 2 h with agitation at 4^{o}C. Unbound material was collected by centrifugation and the resin washed three times with binding buffer containing 20 mM imidazole, then once with 40 mM imidazole, prior to elution with 250 mM imidazole. Laemmli samples were prepared at each step and equivalent amounts were resolved by SDS-PAGE and probed with anti-GFP antibodies in immunoblots.

Figure 7-3 shows that recombinant proteins did not extensively bind to Ni, being mostly present in the flow-through. It is possible that residual ammonium sulphate from the previous purification step could have inhibited binding to the resin, as ammonium salts have been used in IMAC to elute proteins (Kastner and Neubert, 1991). However, upon dialysis of the flow through samples against a large volume of binding buffer (to remove unwanted molecules that might be interfering with the pulldown), no improvement was seen in binding to Ni-NTA resin (Figure 7-4).

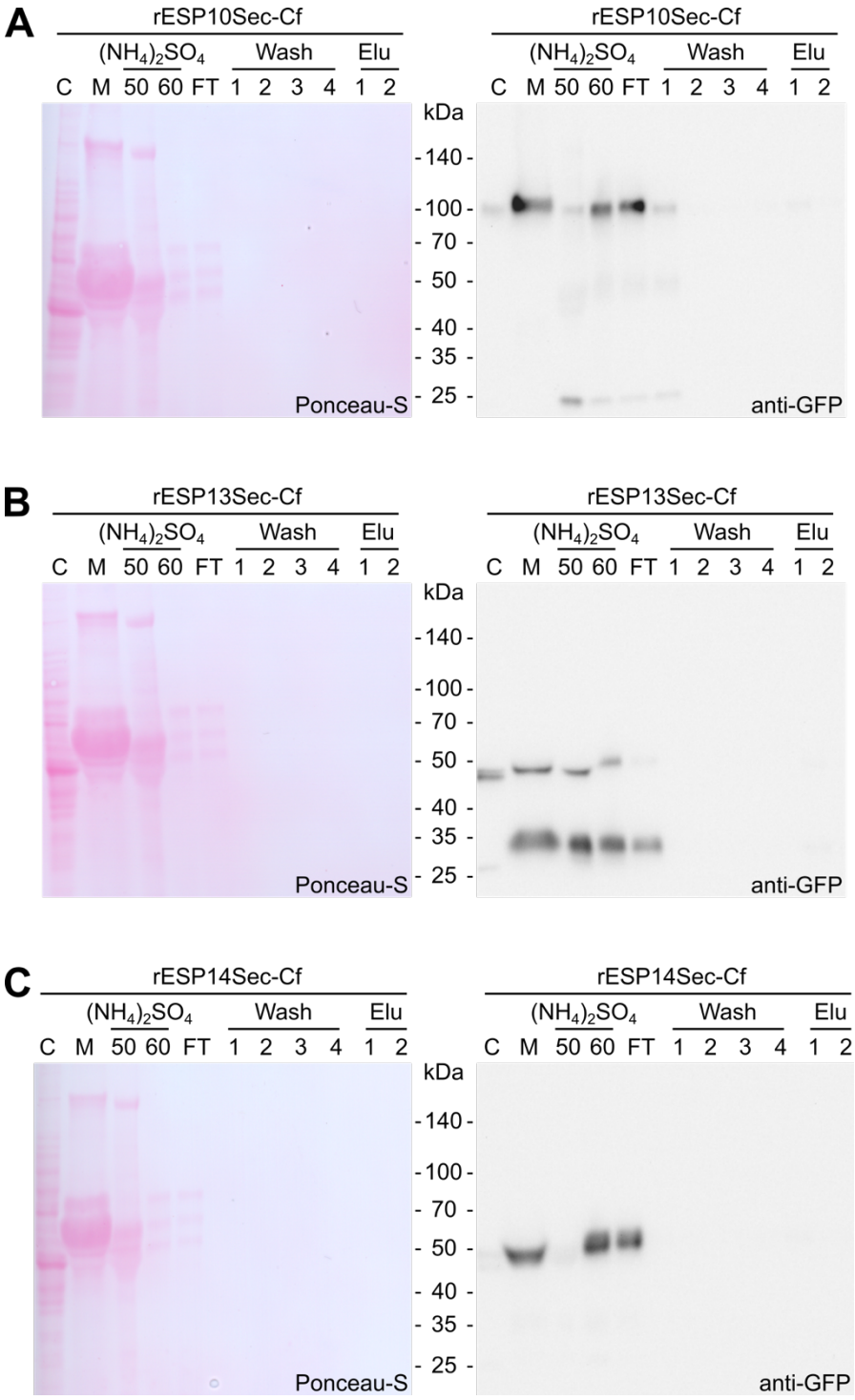


Figure 7-3 Recombinant proteins secreted from SMCs fail to bind to Ni-NTA resin. Ponceau-S stained membranes and anti-GFP immunoblots of attempts to purify recombinant *T. brucei* proteins, secreted from SMC cells, by IMAC – **A** rESP10Sec-Cf, **B** rESP13Sec-Cf, and **C** rESP14Sec-Cf. Samples include whole cell lysates (C), acetone precipitated media (M), ammonium sulphate-precipitated samples, imidazole washes and elution samples (Elu). The majority of the GFP signal (representing recombinant proteins) is seen in the flow-through.

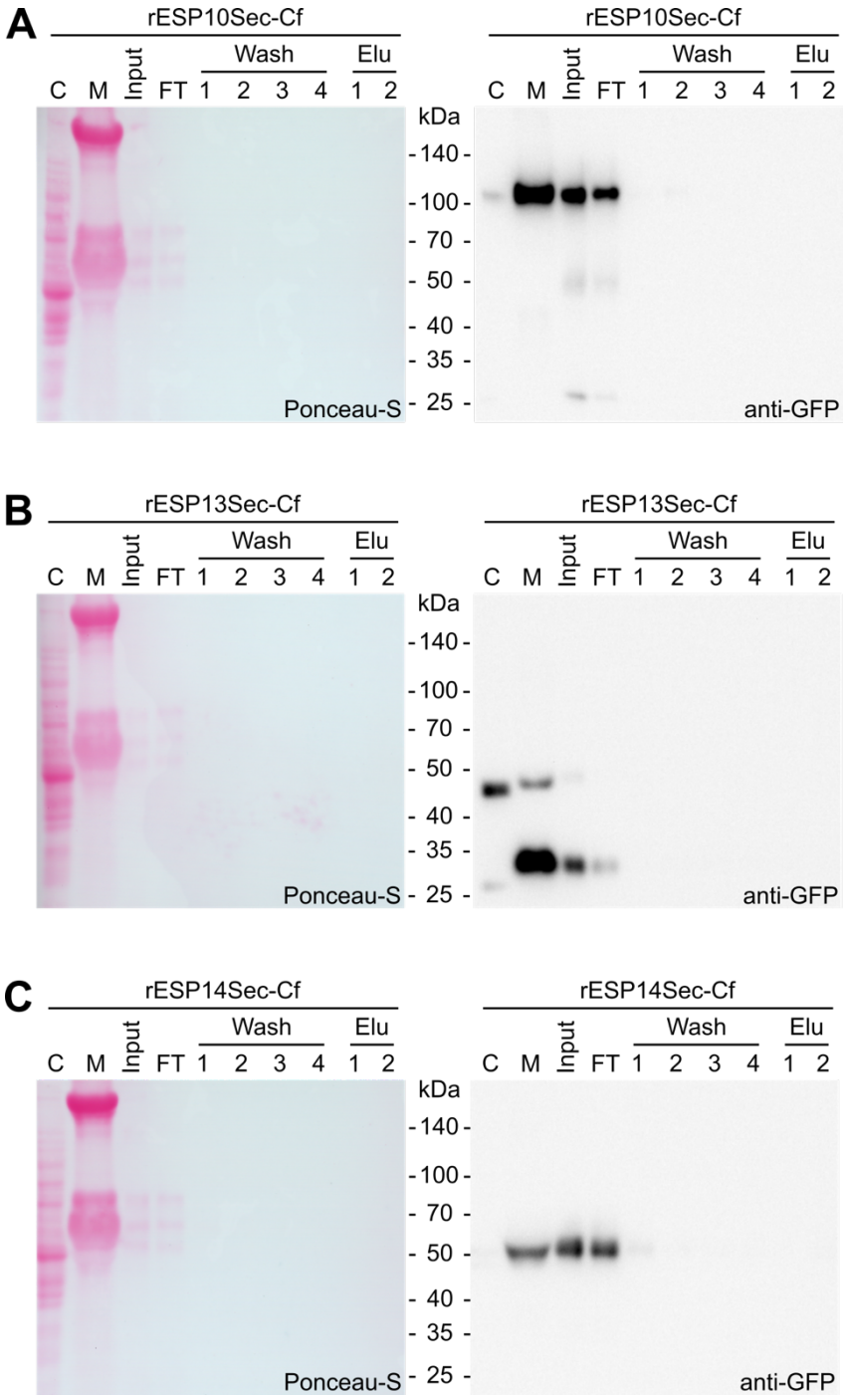


Figure 7-4 Sample dialysis prior to Ni-NTA column did not improve capture of recombinant proteins. Ponceau-S stained membranes and anti-GFP immunoblots of attempts to purify recombinant *T. brucei* proteins, secreted from SMC cells, by IMAC, as in Figure 7-3. The input samples consist of the FT samples from Figure 7-3 after dialysis. **A** rESP10Sec-Cf, **B** rESP13Sec-Cf, and **C** rESP14Sec-Cf.

Inability to capture recombinant proteins by Ni-NTA could be a result of loss of the His_6 tag. To test for its presence, immunoblots shown in Figure 7-3 and Figure 7-4 had their primary and secondary antibodies removed by hot

detergent treatment ('membrane stripping') and re-probed with an anti-His_6 monoclonal antibody. Figure 7-5 shows the original anti-GFP blots alongside the reprobes with anti-His_6. Blots on the left are from the first IMAC experiment (shown in Figure 7-3), whilst blots on the right are from the post-dialysis experiment (from Figure 7-4). As seen in Chapter 6 (Figure 6-10), probing with the anti-His_6 antibody resulted in a much stronger signal in the cell fraction relative to the medium fraction (the precise opposite is seen with the anti-GFP antibodies). Strikingly, despite GFP signal being detected in flow-through samples, His_6 signal is extremely weak or absent in the same lanes in reprobes. Yet, anti-His_6 immunoblots confirm the lack of binding to and elution from Ni columns. It appears that the His_6-tags may have been specifically removed from the recombinant protein, possibly by proteolysis. This could be specific to C-terminal His-tagging of $sfGFP^{op}$ (rather than specific to *C. fasciculata*'s secretory pathway), as in a previous purification attempt of rESP10FL-Lt (which had a $sfGFP^{op}$ fusion), the recombinant protein also failed to bind to the Ni-NTA resin (Figure 7-6A).

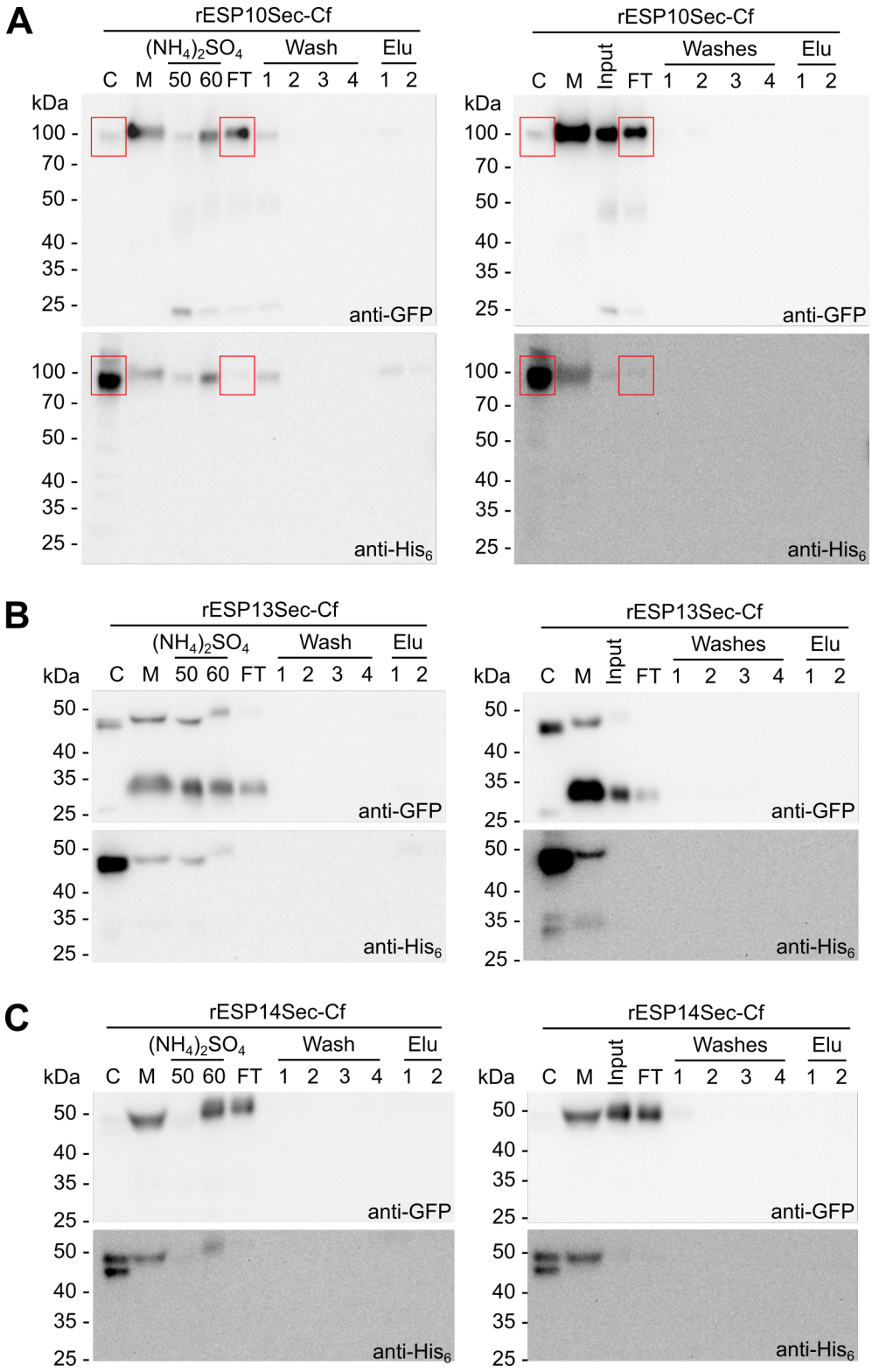
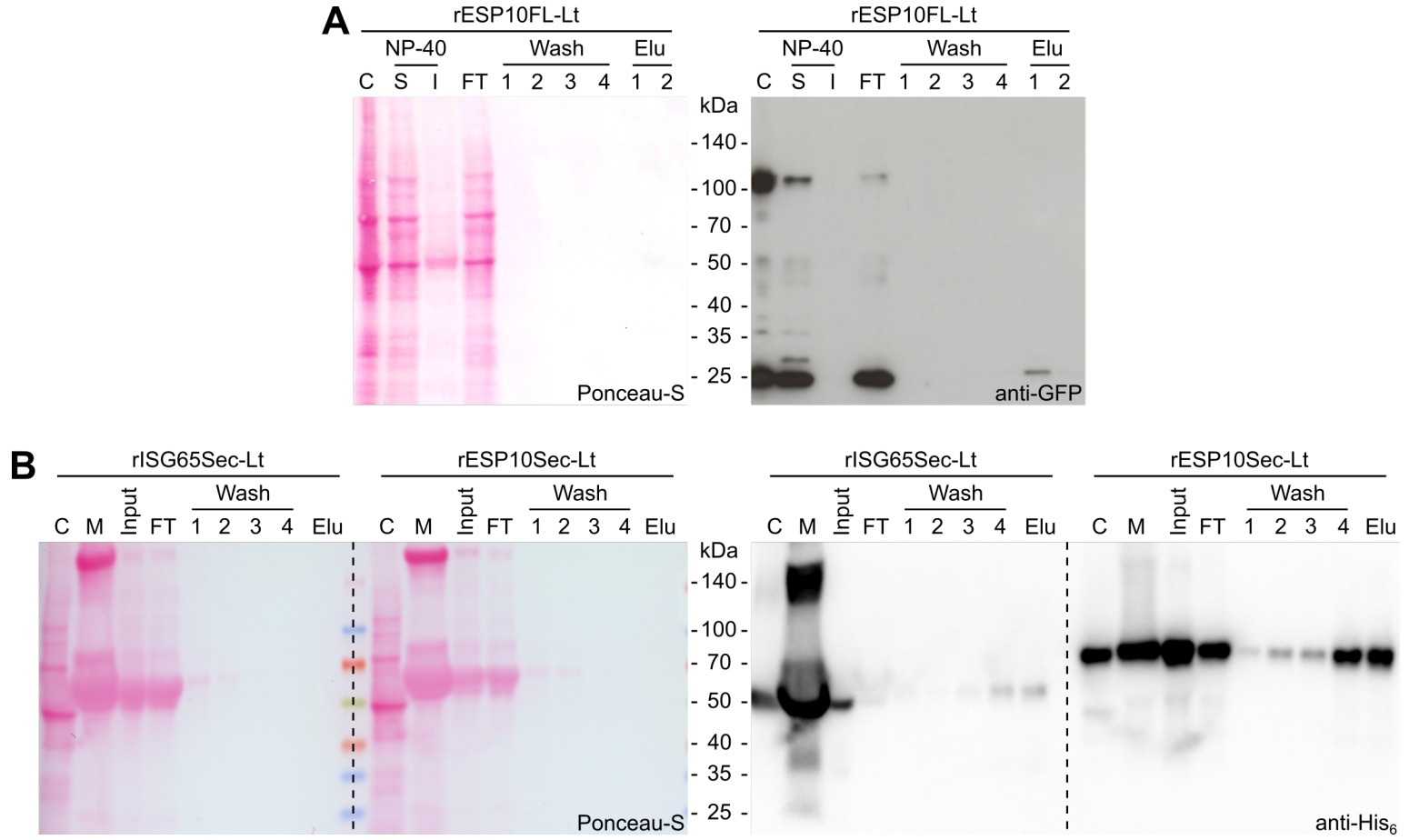


Figure 7-5 His_6-tags on SMC secreted recombinant proteins are either lost or concealed upon secretion into the medium. Anti-GFP immunoblots stripped of antibodies and re-probed with an anti-His_6 monoclonal antibody. Blots on the left originate from the first IMAC experiment (Figure 7-3), whilst blots on the right are from the post-dialysis experiment (Figure 7-4). Signal detected by an anti-His_6 antibody differs to that detected by anti-GFP antibodies (examples highlighted by red boxes in **A**), indicating loss or concealment of the His_6-tags. **A** rESP10Sec-Cf, **B** rESP13Sec-Cf, and **C** rESP14Sec-Cf.

Despite the results described above, rISG65Sec-Lt has been successfully purified from *L. tarentolae* using Ni-NTA IMAC (Rooney *et al.*, 2015). Additionally, as purification issues encountered with rESP10Sec-Cf, rESP13Sec-Cf, and rESP14Sec-Cf were thought to be due to the presence of the fluorescent protein tag, similar drawbacks were not expected to be encountered when attempting to purify rISG65Sec-Lt, rESP10Sec-Lt, rESP13Sec-Lt or rESP14Sec-Lt (which lack $sfGFP^{op}$). For the latter recombinant proteins, anti-His_6 immunoblotting showed the histidine tag to remain present; and protein levels in 50 mL of culture to be approximately 1, 2, 0.2 and 0.2 mg of rISG65Sec-Lt, rESP10Sec-Lt, rESP13Sec-Lt and rESP14Sec-Lt respectively.

Therefore, small-scale purification of these proteins from *L. tarentolae* was attempted, using the same procedure as for *C. fasciculata* recombinant proteins, except that leishmanial culture supernatant was precipitated at 60% ammonium sulphate saturation without an initial cut at 50% saturation. Following ammonium sulphate treatment of the medium, samples were dialysed against binding buffer for removal of unwanted, binding-interfering molecules. Equal proportions of each sample were prepared for SDS-PAGE and immunoblot analysis (Figure 7-6B and C). Prior to binding to Ni-NTA resin, most of the rISG65Sec-Lt sample was lost due to breaking of the centrifuge tube (explaining the significant protein reduction in the gel 'Input' lane). Monitoring recombinant protein behaviour across the purification procedure (via detection of the His_6 tag; Figure 7-6B and C) indicated that most rISG65Sec-Lt, rESP13Sec-Lt and rESP14Sec-Lt present in the input was captured by the Ni-NTA; whereas only approximately half of the rESP10Sec-Lt input amount was captured (the remaining unbound protein detected in the flow through). Notably, greater protein capture was seen for those of smaller molecular mass (between 20 and 50 kDa) than for ESP10 (>70 kDa). Irrespective of the recombinant protein mass, a significant amount was eluted from the Ni resin in the 4^{th} wash, which contains 40 mM imidazole.


A
rESP10FL-Lt
NP-40
Wash
Elu
C S I FT 1 2 3 4 1 2
kDa
140
100
70
50
40
35
25
Ponceau-S
anti-GFP
B
rISG65Sec-Lt
rESP10Sec-Lt
C M Input FT 1 2 3 4 Elu
Ponceau-S
anti-His6

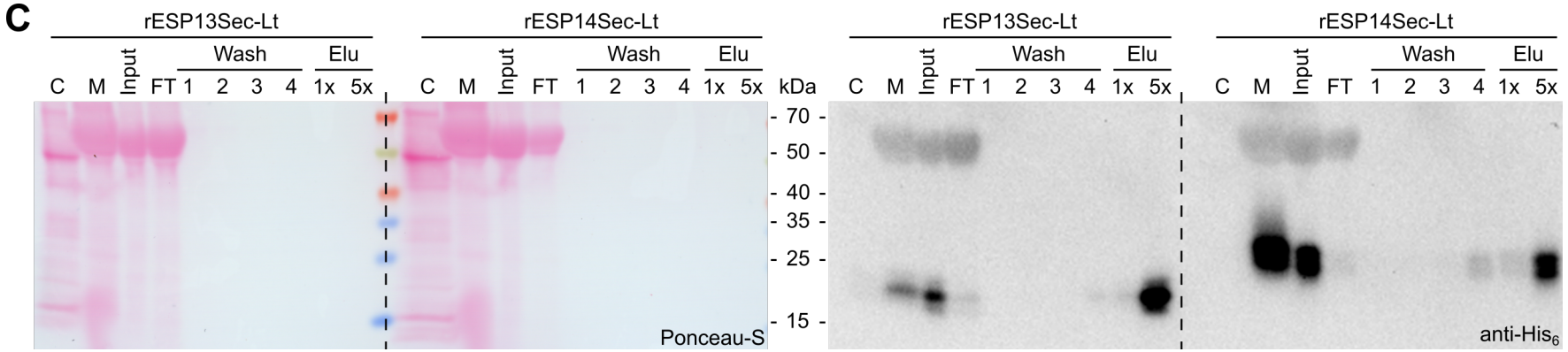


Figure 7-6 Recombinant proteins secreted from *L. tarentolae*, lacking sfGFP^{op}, can be pulled down by Ni-NTA agarose resin. Ponceau-S stained membranes and anti-GFP immunoblots of attempts to purify recombinant *T. brucei* proteins, produced in T7TR cells, by IMAC – **A** rESP10FL-Lt fused to sfGFP^{op}, **B** rISG65Sec-Lt and rESP10Sec-Lt, and **C** rESP13Sec-Lt and rESP14Sec-Lt. Elution samples were loaded either in equal proportion to the other samples or at 5x concentration. **A** includes soluble (S) and insoluble (I) samples after extraction of cellular proteins with NP-40. rESP10FL-Lt fused to sfGFP^{op} does not bind to the Ni-NTA resin, but all the other recombinant proteins do. The band above 50 kDa in **C** is likely non-specific binding revealed upon long exposure time.

Total amount of protein present in eluted fractions was measured by the Bicinchoninic Acid (BCA) method, and revealed 100, 60, 20 and 20 µg in rISG65Sec-Lt, rESP10Sec-Lt, rESP13Sec-Lt and rESP14Sec-Lt samples respectively. These amounts are particularly low compared to the estimated starting amounts, even taking into account protein loss during the purification procedure, as monitored by immunoblotting (Figure 7-6B and C). Low amount of rISG65Sec-Lt control can at least be explained in part due to sample loss caused by a broken tube during ammonium sulphate precipitation. The lower-than-expected amounts of the other recombinant proteins might result from the fact that previous estimation of protein levels was based on anti-His_6 immunoblot signal comparison to known amounts of recombinant His_6-eGFP. As discussed in Chapter 6, anti-histidine antibodies can vary in sensitivity when detecting His_6 tags on different proteins (Debeljak *et al.*, 2006). Thus, the starting levels of *L. tarentolae* recombinant proteins may have been much lower than initially estimated.

Despite these discrepancies, 1 µg of each elution sample (as determined by BCA protein quantification) was resolved by SDS-PAGE for visualisation by Coomassie Blue staining. Given the significant amount of recombinant protein detected in the 4th resin wash (Figure 7-6B and C), an equivalent amount of this sample was also analysed. Known amounts of BSA were prepared and included in the analysis for comparison of protein amount. Visualisation of the proteins eluted showed remarkable similarity between the different purifications, suggesting these proteins to be largely common contaminants, and not the recombinant proteins of interest (Figure 7-7A). The latter proteins, perhaps with the exception of rESP10Sec-Lt, were not present in sufficient amounts to be detected by Coomassie Blue. Yet, the 'BSA ruler' confirmed that each elution lane (E) indeed contained approximately 1 µg.

Coomassie dyes have a sensitivity limit of roughly 100 ng of protein. On the other hand, ruthenium-based fluorescent dyes are highly sensitive, easy to use and compatible with mass spectrometry techniques (the latter as the means to ultimately determine protein identify). Therefore, elution samples analysed by Coomassie stained were re-run in SDS-PAGE and detected by SYPRO Ruby staining (the theoretical detection limit of SYPRO Ruby can be as low as

1 ng/band) (Figure 7-7B). Once again, the lanes' banding pattern was very similar across different samples, indeed suggesting contamination. The two major bands (of ~68 and 210 kDa) visible in all lanes likely represent carried-over BSA from the FBS that supplements the culture medium. A band migrating above 70 kDa is suggestive of rESP10Sec-Lt and appears present at ~100 ng. But no other elution showed a band of the corresponding molecular mass of the respective recombinant protein.

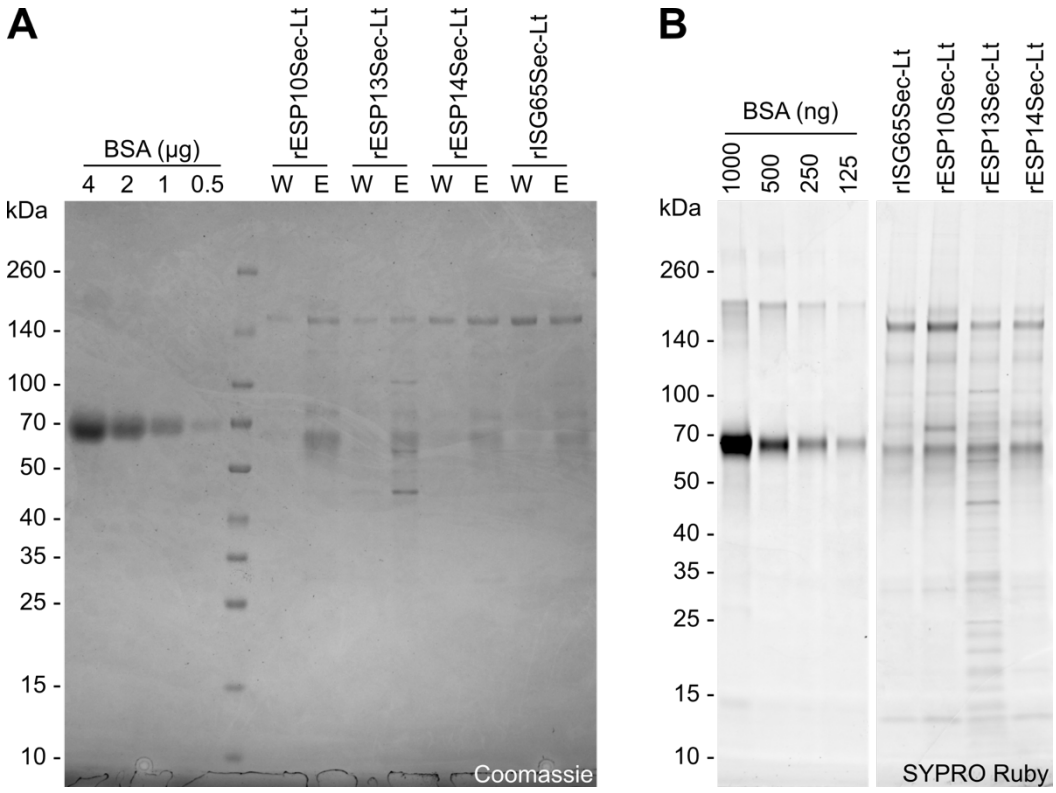


Figure 7-7 Unsuccessful purification of recombinant proteins expressed in *L. tarentolae*. A Coomassie- and **B** SYPRO Ruby-stained gels containing the 4th wash (W) and elution (E) from Ni-NTA pulldown of proteins secreted from *L. tarentolae*. Known amounts of BSA were loaded to confirm protein amounts. All elutions contained a large number of contaminant proteins and very little/no recombinant protein of interest.

It is clear that more work is required to optimise recombinant protein purification from *L. tarentolae*. The discrepancy in protein amounts assessed by different quantification/estimation methods is a concern, and this should be investigated further before a decision is made to abandon the LEXSY

expression system. Due to time constraints, nothing further has yet been done on this regard.

7.2.3. Anti-GFP nanobodies for pulldown of CExSy recombinant proteins

Camelids, such as llamas, produce an unusual subset of immunoglobulins which lack light chains and consist of only heavy chain homodimers (Hamers-Casterman *et al.*, 1993). In recent years, single domain-containing 'nanobodies' have been derived from these unusual immunoglobulins as alternatives to traditional antibodies. They are small, stable molecules that can be readily produced and purified recombinantly from bacteria (Harmsen and De Haard, 2007). This reduces the cost and batch-to-batch variability that can arise in the production of polyclonal antibodies. Colleagues at the Rockefeller University in the US have generated anti-GFP nanobodies with high epitope affinities (Fridy *et al.*, 2014). Two of these nanobodies, which bind to distinct epitopes on opposite sides of the GFP molecule, have previously been produced and purified in our lab by Sarah Whipple. These are termed LaG16 (Fridy *et al.*, 2014) and GFPBP, for GFP binding protein (commercially available as GFP-trap from ChromoTek) (Kirchhofer *et al.*, 2010). These two reagents could be a useful resource for purifying recombinant sfGFP^{op} fusions expressed in *C. fasciculata*, particularly in light of the inability to bind these proteins to a Ni-NTA resin.

40 µg of LaG16 or GFPBP nanobodies were individually coupled to Cyanogen bromide (CNBr)-activated sepharose. Uncoupled nanobodies were washed off, unused active sites in the sepharose were blocked (with 1 M Tris-HCl pH 8), and mock acid elutions performed (with 100 mM glycine, pH 2.7) to remove weakly associated nanobodies. As a test, 10 µg of recombinant His_6-eGFP was allowed to bind to each nanobody-sepharose overnight at 4^{o}C with agitation. After such a period, both LaG16 and GFPBP sepharose were visibly green whilst respective supernatants were clear. Resins were washed with TBS-T, and acid elution was used to release eGFP. Samples were taken at each step of the procedure, including from the sepharose post-elution, for both total protein quantification and analysis (Figure 7-8). One tenth of each sample

was resolved by SDS-PAGE and stained with Coomassie Blue (except for post-elution nanobody-sepharose samples, of which one fifth was analysed).

By visual inspection of the stained gel (Figure 7-8), both LaG16 and GFPBP nanobodies coupled efficiently with sepharose, with no unbound nanobody being detected. Most of the 10 µg of eGFP bound to the nanobody-sepharose, with only a small amount detectable in the supernatant of LaG16-sepharose, and none in the supernatant of GFPBP-sepharose. The elution was less efficient, releasing approximately 50% of the eGFP in each case.

Actual amounts of coupled nanobodies, and bound and released eGFP were quantified by BCA assay. >95% of both nanobodies coupled to the sepharose. 84% of the eGFP bound to the LaG16-sepharose, of which 60% eluted, giving a yield of 50%. 90% of the eGFP bound to the GFPBP-sepharose, of which 57% eluted, giving a yield of 51%.

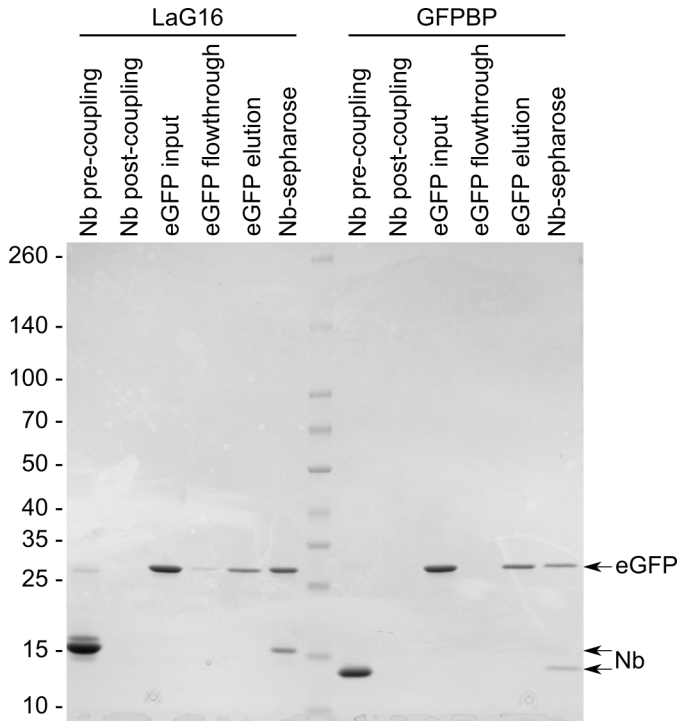


Figure 7-8 Anti-GFP nanobodies, coupled to CNBr-activated sepharose, pulldown eGFP. Coomassie stained gel containing samples of nanobody (Nb) pre- and post-coupling to sepharose, and subsequent eGFP binding and elution samples. Nanobodies consist of either LaG16 or GFPBP. Nb-sepharose samples are 2x concentrated relative to the other samples. Coupling of nanobodies and binding of eGFP were efficient. Acid elution of eGFP was less efficient, resulting in 50% yield from both nanobodies.

In addition to their avidity for GFP, the two tested nanobodies were selected because they bind to epitopes on opposite sides of the GFP barrel structure (Fridy *et al.*, 2014). As both nanobodies resulted in similar yields of eGFP upon pulldown, they were used together for the capture of recombinant proteins, in an attempt to increase the amount of bound protein. Thus, LaG16- and GFPBP-sepharose were mixed in equal amount and used to attempt purification of rESP10Sec-Cf and rESP14Sec-Cf from the flow-through samples from Figure 7-4. rESP13Sec-Cf was not attempted at this point due to the large loss of protein during the previous Ni-NTA purification experiment. Pulldown was attempted exactly as before, other than the use of mixed nanobodies together and two elution steps (to improve the release of bound protein). Each step of the procedure was sampled and analysed by SDS-PAGE and anti-GFP immunoblotting (Figure 7-9).

The nanobody-sepharose mix successfully bound and released rESP10Sec-Cf and rESP14Sec-Cf (Figure 7-9A). Comparing with known amounts of eGFP, input samples were estimated to be 3 µg and 4 µg rESP10Sec-Cf and rESP14Sec-Cf respectively. Elutions of rESP14Sec-Cf released approximately 2 µg of protein, giving ~50% yield (similar to that seen for eGFP). rESP10Sec-Cf elutions resulted in a lower, 23% yield, with approximately 700 ng being released.

To confirm protein quantities and assess purity, a quarter of the volume eluted from each purification (estimated to represent ~175 ng rESP10Sec-Cf and ~ 500 ng rESP14Sec-Cf) were resolved by SDS-PAGE next to known amounts of BSA for analysis by SYPRO Ruby staining (Figure 7-9B). The gel revealed nanobody-sepharose elutions to be reasonably pure. The major band visible in each lane matches the apparent molecular weight of rESP10Sec-Cf and rESP14Sec-Cf (Figure 7-9A). Minor bands of molecular mass similar to the BSA ruler were visible in both elutions, and interpreted as carry-over BSA from the cell culture medium. Purifying rESP14Sec-Cf also showed a weak band of ~ 37 kDa not previously present in Ni-based purifications (Figure 7-7B). The identity of the protein(s) in this band remains unknown. Quantities of rESP10Sec-Cf and rESP14Sec-Cf were estimated to be 560 and 660 ng (2.2 and 2.6 µg total) respectively (by comparison of BSA ruler and recombinant

protein immunoblot signal intensity). Based on these estimates, the pulldown yields were 73% for rESP10Sec-Cf and 65% for rESP14Sec-Cf.

Whilst elution conditions may still be optimised for improvement of final yield, and further chromatography steps could be added to improve purity (such as size exclusion), the results show anti-GFP nanobodies to be a viable step in the purification of recombinant proteins produced in *C. fasciculata*. The absolute amounts purified here are likely not representative of the attainable quantities (due to the loss of protein in previous IMAC purification attempts). It is believed that, were a fresh, 1 litre culture of *C. fasciculata* expressing *T. brucei* proteins of interest being used for purification by anti-GFP nanobodies, much greater amount of recombinant protein could be obtained. These experiments are planned for after PhD thesis submission.

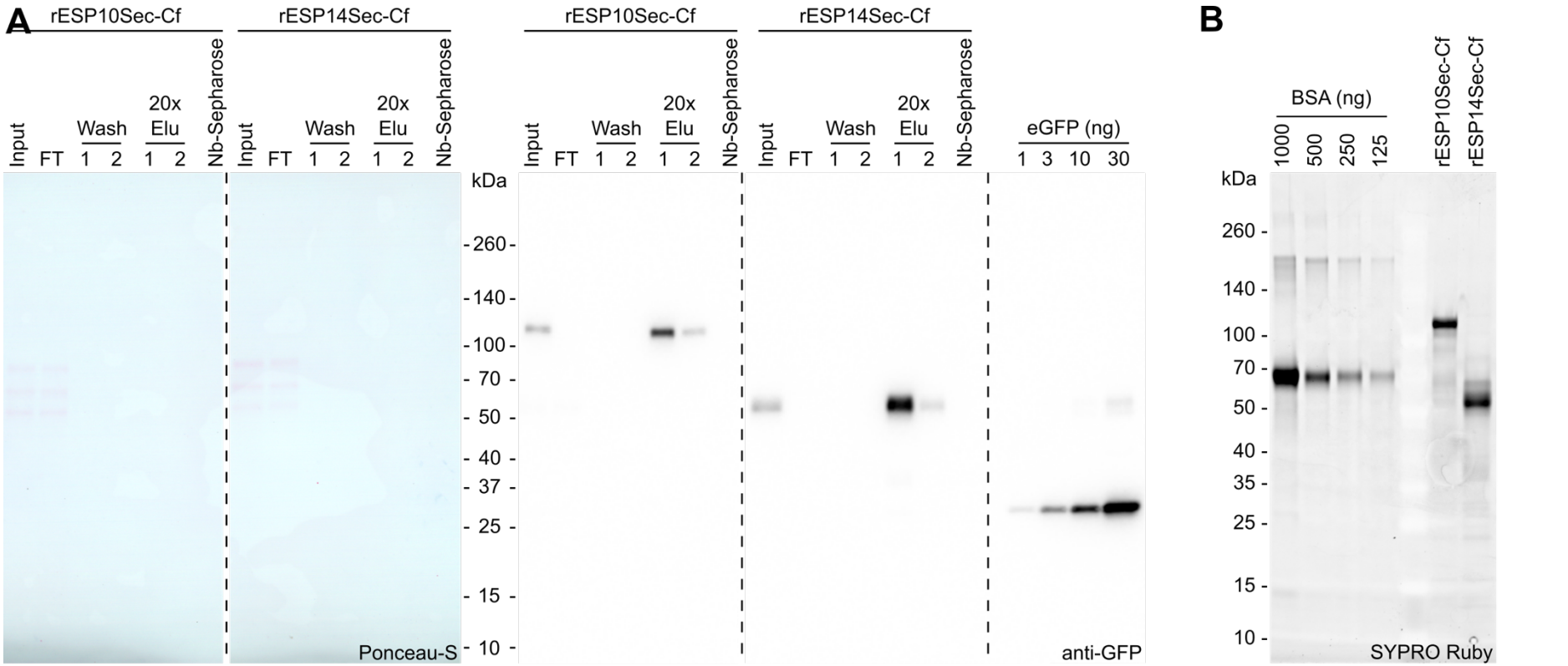


Figure 7-9 Anti-GFP nanobodies, coupled to CNBr-activated sepharose, successfully pulldown recombinant proteins fused to sfGFP^{op}. **A** Ponceau-S stained membrane and anti-GFP immunoblot of attempt to purify recombinant *T. brucei* proteins, produced in SMCs, by pulldown with anti-GFP nanobodies. Input, flow-through and wash lanes contain 1/2000th of total sample. Elution samples (Elu) contain 1/100th total sample. Based on the input and elution samples, rESP10Sec-Cf yield is 23% (700 ng total), whilst rESP14Sec-Cf yield is 55% (2200 ng total). **B** SYPRO ruby stained polyacrylamide gel containing purified rESP10Sec-Cf and rESP14Sec-Cf, alongside known amounts of BSA for quantification. Purified proteins amount to 2.2 µg (73% yield) and 2.6 µg (65% yield) respectively.

7.2.4. Glycosylation status of recombinant proteins produced in CExSy and LEXSY

As discussed in section 7.1, an objective of this Chapter was to assess the glycosylation status of multiple recombinant proteins produced in *C. fasciculata* and *L. tarentolae*. Different types of protein N-glycosylation can be distinguished through the use of glycosidases with different specificities. Subsequent shifts in molecular mass, resulting from removal of N-glycans, can be observed by gel electrophoresis. Two commonly used enzymes are Endoglycosidade H (Endo-H) and peptide-N-glycosidase F (PNGase-F). PNGase F is an amidase that cleaves all types of N-glycosylation, between the innermost GlcNAc residue of the N-glycan and the asparagine residue that releases the N-glycan. That creates an aspartic acid in place of the asparagine residue on the protein. PNGase F's only limitation is that it cannot cleave if there is an α1–3 fucose linked to the core GlcNAc residue (although such modification has not been reported in kinetoplastids). Endo-H, on the other hand, cleaves within the chitobiose core of high mannose, identifying only oligomannose-type N-glycosylation. Individually, both enzymes can help determine the general glycosylation pattern on glycoproteins (Figure 7-10).

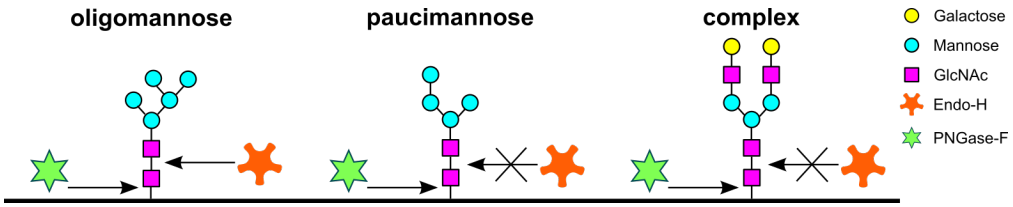


Figure 7-10 Substrate specificity of two commonly-used glycosidases. PNGase-F cleaves most types of N-glycans, unless the innermost GlcNAc residue is linked 1,3 to fucose. Endo-H only cleaves oligomannose N-glycans.

T. brucei can carry out oligomannose, paucimannose and complex-type N-glycosylation. *T. brucei* proteins ESP10, ESP13 and ESP14 have previously been tagged at their endogenous loci with sfGFP (Gadelha *et al.*, 2015) and glycosylation types elucidated using PNGase-F, Endo-H and anti-GFP immunoblotting (Sarah Whipple, personal communication). ESP10 and ESP13 contain a mix of N-glycan types, whilst ESP14 contains only paucimannose

and/or complex-type N-glycans (Table 7-1). Therefore, these three proteins present a varied test set to assess the abilities of *C. fasciculata* and *L. tarentolae* to carry out different types of N-glycosylation on recombinant *T. brucei* proteins. rISGSec-Lt was also included as a control, as it has been previously shown to be N-glycosylated when secreted from *L. tarentolae* (Rooney *et al.*, 2015).

Deglycosylation experiments were carried out in tandem on the set of proteins secreted from *C. fasciculata* (post ammonium sulphate precipitation but prior to nanobody pulldown), secreted from *L. tarentolae* (post IMAC puification), or on endogenous-locus sfGFP-tagged proteins in *T. brucei* cells. Samples were treated with Endo-H (E), PNGase-F (P) or with no enzyme (-), the latter to control for non-specific proteolysis. After deglycosylation, Laemmli samples were prepared and resolved by SDS-PAGE and electro-transferred to nitrocellulose for analysis with ant-GFP or anti-His_6 antibodies (Figure 7-11). The predicted MWs of the different proteins and their glycosylation status (based on Figure 7-11) are shown in Table 7-1.

Table 7-1 Summary of molecular weight (MW) and glycosylation status of *T. brucei* proteins expressed in different organisms. N-glycans are based on the results in Figure 7-11.

Protein	Predicted MW	Host organism	N-glycans	
			Oligomannose	Paucimannose / complex-type
ESP10FL::sfGFP	109	*T. brucei*	✓	✓
rESP10Sec::sfGFP^{op}	98	*C. fasciculata*	✓	✓
rESP10Sec	71	*L. tarentolae*	✗	✓
ESP13FL::sfGFP	55	*T. brucei*	✓	✓
rESP13Sec::sfGFP^{op}	46	*C. fasciculata*	✓	✓
rESP13Sec	19	*L. tarentolae*	✗	✓
ESP14FL::sfGFP	46	*T. brucei*	✗	✓
rESP14Sec::sfGFP^{op}	45	*C. fasciculata*	✓	✓
rESP14Sec	18	*L. tarentolae*	✗	✓
rISG65Sec	42	*L. tarentolae*	✗	✓

Results of the glycosidase treatments of *T. brucei* endogenous proteins (Figure 7-11) agree with past work by Sarah Whipple (not shown). Regarding the *C. fasciculata* recombinant proteins, rESP10Sec-Cf shows a similar pattern to its *T. brucei* counterpart upon deglycosylation, confirming that *C. fasciculata* can carry out oligomannose N-glycosylation, but that it can also conduct paucimannose and/or complex-type N-glycosylation.

In contrast, rESP13Sec-Cf and rESP14Sec-Cf differ to their *T. brucei* equivalents. ESP13FL-Tb contains a mix of glycan types between its putative glycosylation sites; whilst ESP14FL-Tb only contains oligomannose N-glycans. On the other hand, rESP13Sec-Cf and rESP14Sec-Cf show two bands upon Endo-H digest. The higher bands match those of the negative controls, whilst the lower bands match the shifts upon PNGase-F digestion, suggesting a mix of different N-glycosylation types within the same protein sequence.

These results show that *C. fasciculata* recognises *T. brucei* glycosylation sequons. This was very apparent in the glycosylation pattern observed for rESP10Sec-Cf. Yet, subtle differences between glycan synthesis and/or transfer onto amino acid acceptance sites may exist between these two closely-related organism, perhaps explaining the glycosylation status of ESP13 and 14 endogenously expressed by *T. brucei* and recombinantly expressed by *C. fasciculata* (Figure 7-11).

All four candidate proteins recombinantly expressed in *L. tarentolae* show a remarkable similar pattern of glycosylation (Figure 7-11). Endo-H treatment does not alter recombinant protein mass, whereas PNGase-F does, indicating the presence of paucimannose or complex-type glycosylation. This observation agrees with previous data on various other proteins produced in this organism (Breitling *et al.*, 2002; Rooney *et al.*, 2015). Additionally, both rESP13Sec-Lt and rESP14Sec-Lt migrate as a doublet prior to deglycosylation, and as a singlet, lower MW band when glycans are removed. This could reflect distinct sugar moieties being attached to the same protein during post-translational modification.

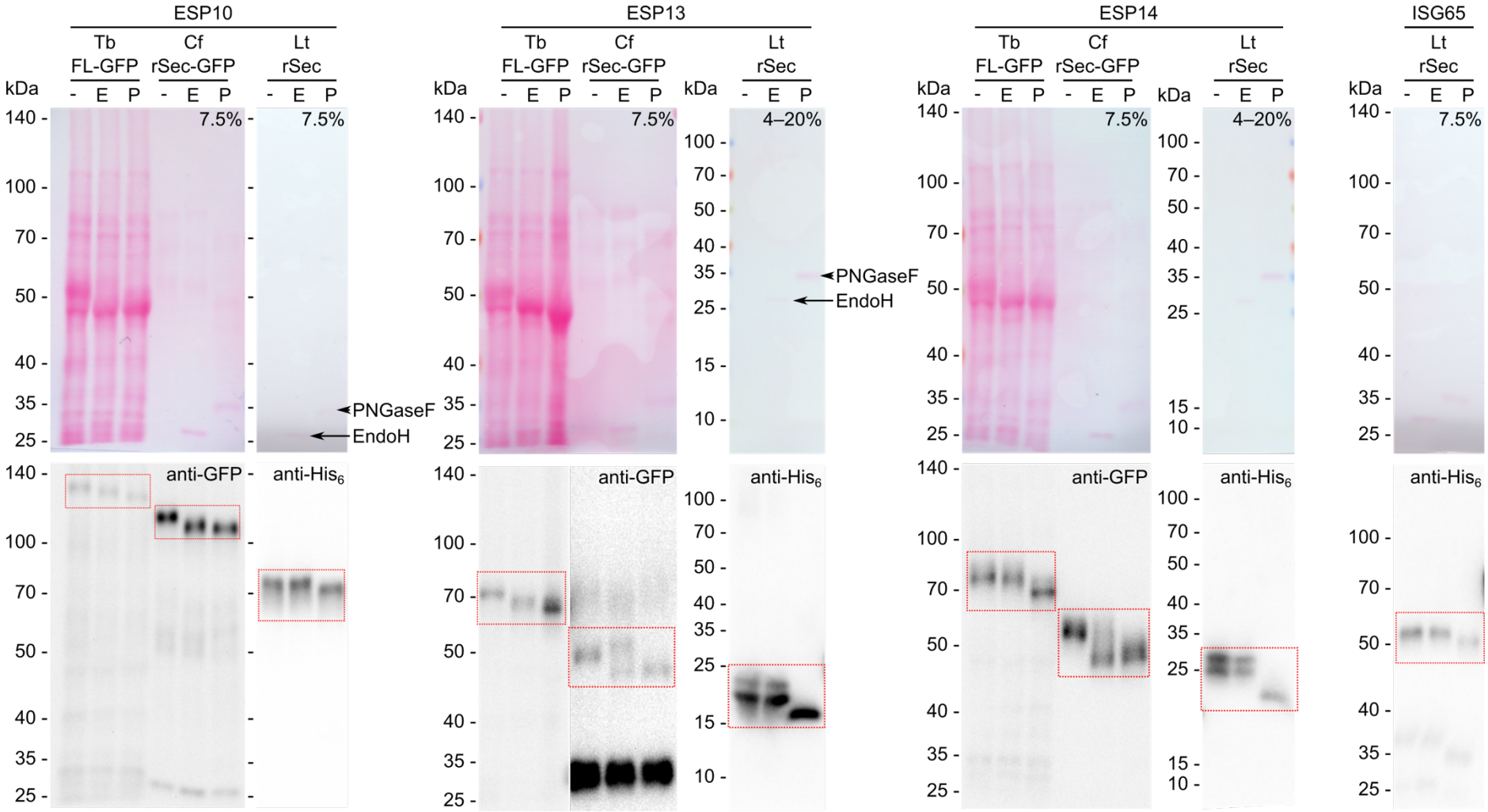

ESP10
Tb FL-GFP
Cf rSec-GFP
Lt rSec
- E P - E P - E P
kDa
7.5%
7.5%
PNGaseF
EndoH
anti-GFP
anti-His6
ESP13
Tb FL-GFP
Cf rSec-GFP
Lt rSec
7.5%
4–20%
PNGaseF
EndoH
anti-GFP
anti-His6
ESP14
Tb FL-GFP
Cf rSec-GFP
Lt rSec
7.5%
4–20%
anti-GFP
anti-His6
ISG65
Lt rSec
7.5%
anti-His6

Figure 7-11 *T. brucei*, *C. fasciculata* and *L. tarentolae* differ in their N-glycosylation of four *T. brucei* proteins. Ponceau-S stained membranes and anti-GFP/anti-His_6 immunoblots of untreated or deglycosylated *T. brucei* surface proteins produced in *T. brucei*, *C. fasciculata* or *L. tarentolae*. Percentage of gel acrylamide indicated in top right of Ponceau-S panels. Antibodies used in immunoblotting are indicated in top right of immunoblot panels. *T. brucei* samples are whole proteomes of sfGFP-endogenous-locus tagged cells. *C. fasciculata* and *L. tarentolae* samples are partially-purified secreted recombinant proteins. Samples include no glycosidase enzyme (-), Endo-H (E), or PNGas-F (P). Relevant bands are highlighted with red boxes. Results indicate that *C. fasciculata* glycosylate proteins in a similar manner to *T. brucei* but with differences in site recognition, whilst *L. tarentolae* only attaches paucimannose / complex-type N-glycans.

7.3. Discussion

This chapter describes initial attempts to purify recombinant *T. brucei* proteins produced in *L. tarentolae* and *C. fasciculata*. A difficulty when purifying secreted recombinant proteins is separating them away from intrinsic components of the culture medium, particularly evident here due to the use of FBS. A simpler way to reduce this contamination would be to grow cells in less complex medium and/or in the absence of FBS. Jena BioScience states in its LEXSY manual that *L. tarentolae* can be grown at the same rate and to the same top density achieved here without FBS supplementation. Yet, we were not able to reproduce this in our lab. However, recent communication with Dr. Andreas Licht (Head of the LEXSY Department at Jena Bioscience) has shed some light on the matter. He stated that the BHI source has a substantial effect on the growth of these organisms, with large variability between suppliers. The BHI powder used in this project was purchased from Sigma, not Jena, and could explain the difference in growth characteristics observed here. In addition, Dr. Licht suggested that increasing BHI concentration as much as 5-fold can improve growth. Alternatively, EX-CELL 405 Serum-Free Medium for insect cells could also be used. Given more time, this feedback could have been implemented, and *L. tarentolae* and *C. fasciculata* culturing optimised for growth without FBS (possibly simplifying medium composition and easing secreted protein purification).

Growth medium without FBS would likely improve capture of secreted His-tagged proteins by Ni-NTA. The amount of contaminating proteins seen in Figure 7-7 may have caused the low binding of recombinant proteins to Ni-NTA sepharose (Figure 7-6) due to competition for binding sites. Thus, growth without FBS might not only result in less complex protein mixture, but also lead to an increase in specific recombinant protein affinity for Ni-NTA. The amounts of recombinant proteins eluted from Ni-NTA were also concerning, especially considering the relative signal observed from anti-His_6 immunoblots (Figure 7-6). As discussed, this discrepancy could be due to the behaviour of the anti-His_6 antibody, leading to overestimation of starting protein concentrations. However, the rISG65Sec protein level reported by Rooney and colleagues suggest this may not be the case.

Recombinant protein purification using anti-GFP nanobodies showed some success, pulling down specific recombinant proteins with very low levels of contaminantion. Additionally, the quantities of purified recombinant proteins obtained more closely aligned with initial estimates based on anti-GFP immunoblots. Although *C. fasiculata* produces lower amounts of secreted recombinant protein, these proteins could be used for purposes that require <1 mg, such as structural or receptor-ligand interaction studies (these will be expanded upon more in the final discussion chapter). Moreover, the results presented in this Chapter showed that proteins produced in *C. fasciculata* carry similar glycosylation patterns to those of the endogenous *T. brucei* proteins, demonstrating the potential value of CExSy. A mix of N-glycan types decorating the same protein was at times observed. This could be due to differences in their OST enzymes. *T. brucei* contains three OSTs with different but overlapping substrate specificities (Izquierdo *et al.*, 2012; Jinnelov *et al.*, 2017). Based upon sequence homology, the *C. fasciculata* genome encodes four OST genes, but they differ around the residues thought to be involved in substrate specificities of the *T. brucei* OSTs.

An alternative explanation to the discrepancies seen between glycosylation patterns may be saturation of the OSTs by overexpression of proteins, leading to variable attachment of different glycosylation types. In *T. brucei*, the substrate specificities of different OSTs are similar enough that if synthesis of one type of N-glycan is inhibited, an alternative type can be attached in its place (Izquierdo *et al.*, 2012). Thus, if a particular OST enzyme is saturated, or if the concentration of a particular N-glycan precursor is limiting, then the cells may attach different N-glycan types to the same site, leading to a pattern similar to the one seen in *C. fasciculata*. Indeed, rESP10Sec-Cf, whose glycosylation did match that of the endogenous ESP10, is present at lower levels than rESP14Sec-Cf and rESP13Sec-Cf (based upon the amount of cleaved sfGFP^{op} seen in immunoblots of rESP13Sec-Cf). Yet, the explanation of differing OST specificities cannot be ruled out, and the rESP10Sec-Cf pattern may simply be a result of not identical but overlapping recognition sequons recognised by *C. fasciculata* and *T. brucei* OSTs.

In contrast to *T. brucei* and *C. fasciculata*, *L. tarentolae* was shown only to attach paucimannose or complex-type glycosylation to the recombinant proteins investigated in this Chapter, agreeing with previous reports (Breitling *et al.*, 2002; JenaBioscience). Paucimannose structures have been reported to appear as a result of saturation of modifying enzymes that produce the fully processed, complex-type N-glycans due to high production of recombinant proteins (Andreas Licht, oral presentation at the Recombinant Protein Production conference in Crete, 2019). Thus, for the proteins assessed so far, *L. tarentolae* does not appear to distinguish between different N-glycosylation sites, simply attaching the same N-glycans to different sites, with differences in glycan structure being dictated by protein level. Despite this, *L. tarentolae* does glycosylate recombinant *T. brucei* ESP14 (but not ESP10 or ESP13) with N-glycans similar to those used in *T. brucei* and, thus, may still be potentially a useful system for production of specific proteins for vaccination experiments.

The results of this chapter further support the utility of CExSy for the production of *T. brucei* surface glycoproteins when large quantities of protein are not required, or when large culture vessels (>5 litres) are available.The purity provided by nanobody purification of recombinant proteins from *C. fasciculata* should not be ignored. On the other hand, LEXSY, whilst initially showing higher levels of recombinant protein secretion, did not produce proteins of the same purity. Further attempts should be made to high-yield culture *L. tarentolae* without FBS for improvement of recombinant protein purification.

Chapter 8. Discussion

8.1. The GPI anchor as a protein sorting signal

In Chapter 3 of this thesis, sorting mechanisms by which GPIAPs are transported to their target membrane were explored, including the influence of GPISSs and GPI valence. The results obtained here suggest that neither of these factors is sufficient to direct the movement of a set of membrane components within the trypanosome cell. A mechanism based upon protein amount had also been proposed, but no correlation between surface protein level and cellular localisation was observed here.

A possible sorting mechanism not examined thus far is the interaction of GPIAPs with physical diffusion barriers. As discussed in Chapter 1, cytoskeletal structures associate with the FP membrane (e,g, collar, the hook complex and rootlet microtubules; Figure 1-16). These structures abut the cytoplasmic face of the plasma membrane, whereas GPIAPs interact exclusively with the extracellular leaflet. Indirect association via a transmembrane-domain protein could fasten GPIAPs to the underlying cytoskeleton, in a manner analogous to the p24 COPII adaptor proteins in the ER membrane (Kruzel *et al.*, 2017). Yet, this awaits investigation.

In neurons, protein composition differs between sematodendritic and axonal membranes. These membrane domains are separated by a region called the axonal initial segment (AIS; Figure 8-1) which acts as a diffusion barrier (Leterrier, 2018). The AIS is a structure that interacts with both microtubules and the plasma membrane. An integral component of the AIS is the protein ankyrinG (orange in Figure 8-1). AnkyrinG connects to spectrin proteins, which in turn connect to actin rings that are spaced along the axon. AnkyrinG contains both a membrane-binding domain and a domain for interaction with microtubule-associated proteins. Knockdown of ankyrinG dismantles the AIS and causes axons to take on the characteristics of dendrites (Hedstrom *et al.*, 2008).

The membrane binding domain of ankyrinG allows it to recruit ion channels and cell adhesion molecules to the AIS (Leterrier, 2018). It is this gathering of

transmembrane proteins that is thought to form a diffusion barrier within the AIS, acting as a thick “forest” that slows down two-dimensional diffusion of proteins through the region. The differing membrane compositions either side of the AIS appear to be maintained by vesicular trafficking of distinct cargo to these specific regions of the plasma membrane. Thus, whilst the AIS in neurons is an example of membrane-cytoskeletal association acting in the formation of diffusion barriers, it must differ from barriers in trypanosomes, where all membrane traffic is directed to the flagellar pocket. The parasite requires a selective diffusion barrier rather than an unbiased blockade.

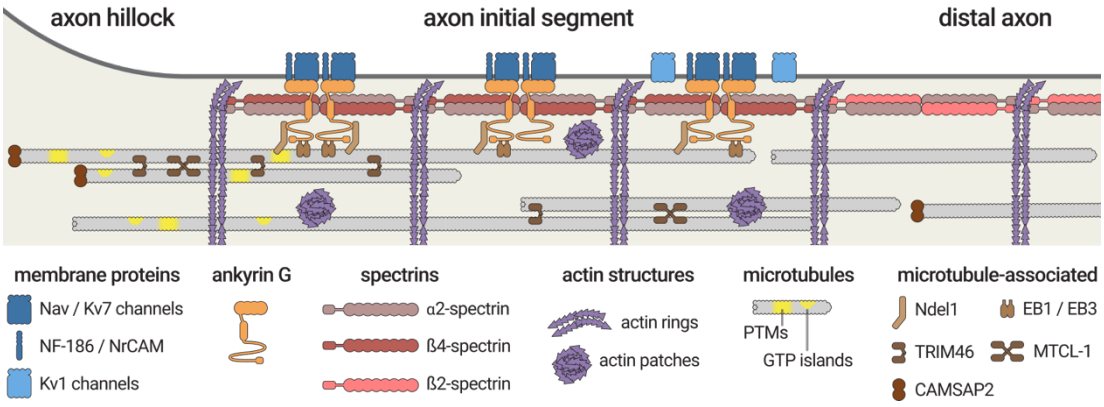


Figure 8-1 Molecular organisation of the AIS. Transmembrane proteins (blue) are anchored in the AIS by ankyrinG (orange). AnkyrinG is inserted into the α2/β4-spectrin complex which separates actin rings (purple). The distal axon spectrin complex, on the other hand, is formed by α2/β2-spectrin. AnkyrinG connects to microtubules via microtubule-binding proteins (light brown). Figure taken from (Leterrier, 2018).

Other examples of membrane diffusion barriers include structures formed by septins, first observed in yeast. There, a septin ring forms at the bud neck and prevents diffusion of certain membrane proteins between mother and daughter cells (Takizawa *et al.*, 2000). In mammalian sperm cells, septins form part of a ring structure known as the annulus, a region which separates two membrane domains of sperm tails. Knockout of septin genes causes posterior tail domain-specific proteins to redistribute to the whole sperm cell surface (Kwitny *et al.*, 2010). Septins are also found at the base of primary cilia in mammalian epithelial cells. Here, they help maintain asymmetric protein distribution between cilium and plasma membrane (Hu *et al.*, 2010). At the base of the

primary cilium, septin2 is required for the assembly of a 9-protein complex (Chih *et al.*, 2012). Disruption of this complex causes increased diffusion into the cilium membrane and inhibition of Hh signalling (which occurs at the cilium membrane). For example, in cells expressing a GPI-anchored GFP, knockdown of individual components of the septin complex led to a ~1.5-fold increase in the number of GFP-positive cilia (Chih *et al.*, 2012).

Regarding *T. brucei*, if protein complexes are involved in the retention or escape of proteins from the FP, the collar may act as a frame upon which such a complex could be constructed. Notably, freeze-fracture electron microscopy experiments showed clusters of intramembrane particles (IMPs) localised to the FP membrane region abutting the collar (Gadelha *et al.*, 2009). Further to this, collar involvement in diffusion barrier formation is supported by data from our lab showing loss of FP retention of four distinct transmembrane proteins upon knockdown of the collar component BILBO1 (Whipple, 2017). It would be interesting to explore whether the same knockdown would also lead to the redistribution of FP-resident GPIAPs.

The study by Albisetti and collegues which discovered a second component of the FPC (named FPC4), utilised BILBO1 as bait in a yeast two-hybrid screen (Albisetti *et al.*, 2017). In addition to FPC4, severel other proteins were identified as putative BILBO1 binders. Further characterisation of these proteins might be beneficial in the determination of putative membrane diffusion barriers. An alternative approach would be to use FP-localised GPIAPs as bait. This could be done in tandem with affinity purification of tagged proteins, to search for possible interacting proteins that might play a role in GPIAP sorting. For example, the *T. brucei* TfR has been shown to interact with other glycoproteins (Mehlert *et al.*, 2012). This was done using the lectin ricin, which binds to terminal β-galactose residues and, thus, identifies complex N-glycans. Ricin-binding glycoproteins were extracted from hypotonicaly-lysed cells (thus, depleted of their GPIAPs), before separation by SDS-PAGE and immunoblot using anti-TfR with or without pre-incubation with pure TfR. This revealed two bands of apparent molecular weights 55 and 97 kDa, providing an explantion for the previous suggestion that TfR is associated with pNAL structures (Nolan *et al.*, 1999), despite not having any directly attached (Mehlert *et al.*, 2012).

Whilst the posibility of a TfR-glycoprotein interaction playing a role in GPIAP sorting is purely speculative, it would be intriguing to know the identity of these potential protein interactors, and whether the interaction is maintained upon addition of a GPI anchor to ESAG7.

8.2. Recombinant expression of *T. brucei* BSF surface proteins

8.2.1. Vaccination

In 2012, the World Health Organisation (WHO) targeted human African trypanosomiasis (HAT, as well as leishmaniasis and Chagas' disease) for elimination by 2020 (WHO, 2012). Whilst the number of reported cases of sleeping sickness has dropped below 2000 (Barrett, 2018), disease control efforts cannot be relaxed, with an estimated >50 million people still at risk (Franco *et al.*, 2018). Without a viable vaccine, disease control relies on detection and treatment. However, in the face of lowering incidence, active screening programmes become more challenging and less viable. Thus, vaccination is still a desirable option for achieving disease elimination. Without it, eradication of parasite reservoirs in endemic areas will be extremely challenging. Previous vaccination studies have shown some promise, with some *T. brucei* antigens conveying partial protection in mice (see Chapter 4 for more details; also reviewed by La Greca and Magez, 2011). However, these were largely only successful with low parasite load, and none of these studies have led to further development of vaccines in the field.

The definition of a high-confidence surfeome by Gadelha and colleagues (2015) gives access to >100 validated *T. brucei* BSF surface proteins to test in vaccinology. Thus, an aim of this PhD was to produce recombinant *T. brucei* BSF surface proteins and test their applicability as vaccines. In attempting to do this, trypanosomatid-based expression systems (namely, the commercially available LEXSY, and the novel CExSy) were evaluated for their potential to correctly produce these proteins in large quantities. Neither system led to the purification of large amounts of recombinant protein, although there is still scope for further development of each system regarding optimal growth conditions, increasing protein yield and improving purification. These improvements were

discussed in previous chapters and will not be expanded upon here. As it stands, use of LEXSY or CExSy for the production of recombinant antigens for a vaccination study would require 5–10 litres of cell culture. This is obviously feasible in an academic lab setting. However, these quantities are only conducive with a low-throughput targeted approach. An initial aim of this project was to evaluate vaccine potential of multiple surfeome components at once. With these restrictions, it is even more important to pre-select the ideal candidates to test. The surfeome could be mined to identify more proteins with large predicted extracellular domains, coupled with endogenous-locus tagging to validate cell body/flagellum localisation and quantify protein levels. This could be further supplemented with analysis using epitope prediction algorithms. For example, the Immune Epitope Database (IEDB; www.iedb.org) is a comprehensive resource containing experimental data on >95% of known immunogenic epitopes studied in primates (including humans), mice and other animals (Vita *et al.*, 2019). It includes multiple epitope prediction tools for the analysis of potential B-cell and T-cell response.

As an alternative to the targeted approach above, large-scale ELISA screens could be carried out using small amounts of recombinant protein, prior to large-scale recombinant production of promising candidates for immunisation tests. Such screens require <1 µg of individual recombinant proteins. In a study by Elton and colleagues, this strategy was used to screen vaccine candidates for the apicomplexan parasite *Babesia microti* (aetiological agent of the tick born disease human babesiosis; Elton *et al.*, 2019). To select proteins, they mined the *B. microti* genome sequence for ORFs predicted to encode secreted or surface-localised proteins, excluding those smaller than 10 kDa, multipass transmembrane proteins lacking a large contiguous extracellular domain, and proteins not expressed in blood stages. This sieved 54 candidates, whose extracellular domains were recombinantly expressed in mammalian cells. Forty-one were successfully expressed, and screened (either in native or heat-denatured form) by ELISA using sera from experimentally-infected BALB/c mice. The screen identified 16 immunoreactive recombinant proteins. All 16 showed weaker or no signal upon denaturation, implying that correct folding had a large bearing on epitope recognition. Indeed, one of the identified proteins had

previously not shown immunoreactivity when expressed in *E. coli*, highlighting the need for careful choice of expression system.

8.2.2. Diagnostics

An incidental benefit of the large-scale screens discussed above is that they can also lead to the identification of surface antigens useful for disease diagnosis. In the absence of a vaccine, efficient diagnosis is essential for disease control, allowing administration of curative drugs, reducing both the burden on the patient and the risk of transmission. Diagnosis of HAT is challenging, as patients with early stage disease present with generic symptoms (e.g. fever, headache), and are commonly misdiagnosed with other diseases, including malaria (Bukachi *et al.*, 2017). As such, many patients are not diagnosed until the disease reaches second stage, in which the parasites have crossed the blood-brain barrier and begin to cause more extreme neuro-psychiatric and endocrinal problems. At this point, the disease is harder to treat and can result in coma and death.

Until recent years (even now, to a large extent) diagnosis relied on the Card Agglutination Test for Trypanosomiasis (CATT) and microscopic analysis of blood, lymph and cerebrospinal fluid (Bonnet *et al.*, 2015). CATT utilises a freeze-dried suspension of whole *T. b. gambiense* cells expressing the LiTat1.3 VSG variant. This test is specifically for the *T. b. gambiense* form of the disease (which accounts for 95% of HAT cases in West and Central Africa) and results in reported sensitivities ranging 63–100% (Lumbala *et al.*, 2018; Mitashi *et al.*, 2012). The large sensitivity range seen between different studies is likely caused by expression of different VSG variants in different parasite populations in different areas. Aside from sensitivity issues of diagnostic tests, HAT foci are often centred around remote rural areas that do not have access to diagnostic equipment required for the above tests (Bukachi *et al.*, 2017; Mitashi *et al.*, 2012). Hence, under-detection is often a problem, and the risk of outbreak is still present. Active screening programmes go some way to help with this, but these require trained mobile teams with specialised equipment and, as such, are difficult to implement, particularly as the number of cases continue to fall.

These current issues emphasise the need for alternative rapid diagnostic tests (RDTs) that are simple to implement.

Multiple new RDTs have recently been developed and are now available. Two of these RDTs (SD BIOLINE HAT and HAT Sero-*K*-SeT, manufactured by Alere and Coris BioConcept respectively) utilise native VSG antigens purified from *T. b. gambiense* (Lumbala *et al.*, 2018). Production of these native antigens requires injection of rats with human-infective trypanosomes and, as such, comes with risk and added animal use. To overcome this, a new generation of RDT (SD BIOLINE HAT 2.0) was developed using recombinant VSG (LiTat1.5) produced using a baculovirus expression system, and ISG65 expressed in *E. coli* (Sullivan *et al.*, 2013). As with CATT, these RDTs have been reported to vary in sensitivity from ~60% to >95% (Jamonneau *et al.*, 2015; Lumbala *et al.*, 2018). Again, this could be due to their use of variant antigens, reducing sensitivity when encountering different variants in the wild. Indeed, one study which compared the diagnostic potentials of CATT, SD BIOLINE HAT and SD BIOLINE HAT 2.0, found their sensitivities to be higher in passive rather than active screening (Lumbala *et al.*, 2018). Passive screening tends to find more cases in late-stage disease; hence, these patients may have encountered a larger range of VSG variants through antigenic switching and have a higher likelihood of showing positive with an RDT. The SD BIOLINE HAT 2.0 RDT was found to have the highest overall sensitivity of the three tests (71% *vs.* 63% for CATT and 59% for SD BIOLINE HAT; Lumbala *et al.*, 2018). This higher sensitivity may result from the use of ISG65 rather than VSG. However, as a multicopy gene, even ISG65 may be subject to variation in different isolates through recombination and/or differential regulation of individual gene copy expression levels.

Another potential drawback to SD BIOLINE HAT 2.0 is that the ISG65 used is recombinantly produced in bacteria and, hence, not glycosylated (Sullivan *et al.*, 2013). Additionally, the VSG used will differ in its glycosylation pattern to native protein, having been produced in insect cells (i.e. baculovirus expression system). Thus, these proteins may lack vital epitopes for host immune recognition. For example, one particular HIV-neutralising antibody is known to bind specifically to gp120 N-glycans rather than the underlying protein structure (Scanlan *et al.*, 2002). Another well-known example of glycan modulation of

immune response is the ABO blood group system, whereby different glycan types on the H antigen can lead to complement haemolysis upon transfusion of the incorrect blood type to a patient (Yamamoto *et al.*, 1990). The performance of HAT RDTs might, therefore, be improved through protein production using trypanosome expression systems, such as those explored in this thesis. This is supported by the work of Rooney and colleagues, who used LEXSY to produce *T. b. gambiense* LiTat 1.3, LiTat 1.5 and ISG65, and evaluating their diagnostic potential using sera from patients infected with *T. b. gambiense* or *T. b. rhodesiense* (*g*-HAT and *r*-HAT; Rooney *et al.*, 2015). The antigens showed >92% sensitivity to *g*-HAT sera and 37–73% to *r*-HAT sera.

While progress is being made, further investigation into potential protein candidates for RDTs will aid in more efficient diagnosis and disease control. Resources such as the surfeome will feed this research area with additional targets. Coupled with promising results from LEXSY-produced proteins (Rooney *et al.*, 2015), high-throughput screens using trypanosome-based expression systems could be a viable approach to tackling this issue.

8.2.3. Protein structure and function

Understanding *T. brucei* protein structure and how this relates to function can inform on the fundamental biology of the parasite and potentially contribute towards rational drug design. It can also inform decisions on vaccination and diagnostic candidates, providing detailed information on how a protein might stand in the context of the VSG coat, and whether it may extend past the VSG distal tip. For example, the solved crystal structures of HpHbR from both *T. brucei* and *T. congolense* suggest that HpHbR likely protrudes out of their respective VSG coats (Higgins *et al.*, 2013; Stødkilde *et al.*, 2014). This allows access to ligands, but also to immune surveillance, highlighted by *T. brucei* susceptibility to killing by an anti-HpHbR antibody-drug conjugate (MacGregor *et al.*, 2019).

To date, only a handful of *T. brucei* BSF surface protein structures have been solved (VSG, SRA and HpHbR, as previously discussed in Chapters 1, 6 and above). Protein crystallisation can be challenging; high concentrations of proteins are required for testing multiple different buffers and conditions. Crystal

formation can take months and may not occur at all. For efficient formation of well-ordered crystals, proteins must be purified to high homogeneity (McPherson and Gavira, 2014); thus, glycans are often removed to reduce heterogeneity. Due to these challenges and restrictions, an alternate technique has exploded in popularity – single-particle cryo-electron microscopy (cryo-EM). Cryo-EM relies on averaging the signals obtained from lots of single protein particles within 2D electron micrographs. In contrast to crystallography, cryo-EM needs very little protein (<20 µg/EM grid; Herzik *et al.*, 2019; Khoshouei *et al.*, 2017; Kim *et al.*, 2015; Liu *et al.*, 2019), and much less time and luck. It is also much more conducive with solving structures of N-glycosylated proteins (Lee *et al.*, 2016).

The increasing popularity of cryo-EM has led to a rapid improvement in technology and sample preparation methods. As recently as 2016, cryoEM had largely been used for proteins and complexes of >200 kDa, with the smallest solved structure reported being the 135 kDa ABC exporter TmrAB, to a resolution of 10 Å (Kim *et al.*, 2015). More recently, sub-100 kDa structures have been solved, including haemoglobin (64 kDa, 3.2 Å resolution; Khoshouei *et al.*, 2017), alcohol dehydrogenase (82 kDa, 2.7 Å; Herzik *et al.*, 2019) and even GFP (26 kDa, 3.8 Å; Liu *et al.*, 2019) through the use of a designed modular scaffold protein complex which could symmetrically bind 12 GFP molecules.

The only *T. brucei* protein structure analysed so far by single-particle cryo-EM is that of the ribosome (Hashem *et al.*, 2013). With the advances in cryo-EM discussed above, it becomes feasible that the structures of *T. brucei* surface proteins might be readily solved by this technique, allowing further exploration of the surface landscape of this important extracellular parasite. With the low amount of protein required for cryo-EM, there is more flexibility in choosing the recombinant expression system to produce proteins of interest. In light of this, CExSy (or LEXSY) could be useful tools in the production of such proteins for this purpose.

8.3. Concluding statement

This thesis has attempted to explore the surface of *T. brucei* through the study of surface protein localisation and recombinant protein production. The data presented here show that sorting of GPIAPs in *T. brucei* BSF is likely a multifactorial process regulated by different mechanisms, rather than by a single, universal system. In addition, a novel kinetoplastid expression system (CExSy) was generated and characterised, that produces sfGFP in the range of milligrams per litre of cell culture. This system, along with the commercially available LEXSY (also tested here), may prove useful tools for particular biochemical applications, whilst further improvements and characterisation could broaden their applicability.

V. References

Abdelwahab, N.Z., Crossman, A.T., Sullivan, L., Ferguson, M.A.J., Urbaniak, M.D., 2012. Inhibitors incorporating zinc-binding groups target the GlcNAc-PI de-N-acetylase in *Trypanosoma brucei*, the causative agent of African sleeping sickness. *Chem. Biol. Drug Des.* 79, 270–278.

Ahearn, I.M., Tsai, F.D., Court, H., Zhou, M., Jennings, B.C., Ahmed, M., Fehrenbacher, N., Linder, M.E., Philips, M.R., 2011a. FKBP12 binds to acylated H-Ras and promotes depalmitoylation. *Mol. Cell* 41, 173–185.

Ahearn, I.M., Haigis, K., Bar-Sagi, D., Philips, M.R., 2011b. Regulating the regulator: post-translational modification of RAS. *Nat. Rev. Mol. Cell Biol.* 13, 39–51.

Ahmad, M.F., Yadav, B., Kumar, P., Puri, A., Mazumder, M., Ali, A., Gourinath, S., Muthuswami, R., Komath, S.S., 2012. The GPI anchor signal sequence dictates the folding and functionality of the Als5 adhesin from *Candida albicans*. *PLoS One* 7, e35305.

Akopian, D., Shen, K., Zhang, X., Shan, S., 2013. Signal recognition particle: an essential protein-targeting machine. *Annu. Rev. Biochem.* 82, 693–721.

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, 521.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P., 2008a. Membrane structure, in: Anderson, M., Granum, S. (Eds.), *Molecular Biology of the Cell*. Garland Science, pp. 617–650.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P., 2008b. Intracellular compartments and protein sorting, in: Anderson, M., Granum, S. (Eds.), *Molecular Biology of the Cell*. Garland Science, pp. 695–748.

Albisetti, A., Florimond, C., Landrein, N., Vidilaseris, K., Eggenspieler, M., Lesigang, J., Dong, G., Robinson, D.R., Bonhivers, M., 2017. Interaction between the flagellar pocket collar and the hook complex via a novel microtubule-binding protein in *Trypanosoma brucei*. *PLOS Pathog.* 13, e1006710.

Alcolea, P.J., Alonso, A., García-Tabares, F., Toraño, A., Larraga, V., 2014. An Insight into the proteome of Crithidia fasciculata choanomastigotes as a

comparative approach to axenic growth, peanut lectin agglutination and differentiation of *Leishmania* spp. promastigotes. *PLoS One* 9, e113837.

Alsford, S. *et al.*, 2012. High-throughput decoding of antitrypanosomal drug efficacy and resistance. *Nature* 482, 232–236.

Alsford, S., Horn, D., 2011. Elongator protein 3b negatively regulates ribosomal DNA transcription in African trypanosomes. *Mol. Cell. Biol.* 31, 1822.

Ammar, Z., Plazolles, N., Baltz, T., Coustou, V., 2013. Identification of trans-sialidases as a common mediator of endothelial cell activation by African trypanosomes. *PLoS Pathog.* 9, e1003710.

André, J., Harrison, S., Towers, K., Qi, X., Vaughan, S., McKean, P.G., Ginger, M.L., 2013. The tubulin cofactor C family member TBCCD1 orchestrates cytoskeletal filament formation. *J. Cell Sci.* 126, 5350–5356.

Ashida, H., Hong, Y., Murakami, Y., Shishioh, N., Sugimoto, N., Kim, Y.U., Maeda, Y., Kinoshita, T., 2005. Mammalian PIG-X and yeast Pbn1p are the essential components of glycosylphosphatidylinositol-mannosyltransferase I. *Mol. Biol. Cell* 16, 1439–1448.

Aslett, M. *et al.*, 2010. TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Res.* 38, D457.

Atrih, A., Richardson, J.M., Prescott, A.R., Ferguson, M.A.J., 2005. *Trypanosoma brucei* glycoproteins contain novel giant poly-N-acetyllactosamine carbohydrate chains. *J. Biol. Chem.* 280, 865–871.

Baeshen, N.A., Baeshen, M.N., Sheikh, A., Bora, R.S., Ahmed, M.M.M., Ramadan, H.A.I., Saini, K.S., Redwan, E.M., 2014. Cell factories for insulin production. *Microb. Cell Fact.* 13, 141.

Barrett, M.P., 2018. The elimination of human African trypanosomiasis is in sight: Report from the third WHO stakeholders meeting on elimination of gambiense human African trypanosomiasis. *PLoS Negl. Trop. Dis.* 12, 10–13.

Bart, J.-M., Cordon-Obras, C., Vidal, I., Reed, J., Perez-Pastrana, E., Cuevas, L., Field, M.C., Carrington, M., Navarro, M., 2015. Localization of serum resistance-associated protein in *Trypanosoma brucei rhodesiense* and transgenic *Trypanosoma brucei brucei*. *Cell. Microbiol.* 17, 1523–1535.

Bartossek, T. *et al.*, 2017. Structural basis for the shielding function of the dynamic trypanosome variant surface glycoprotein coat. *Nat. Microbiol.* 2,

1523–1532.

Bate, C., Nolan, W., McHale-Owen, H., Williams, A., 2016. Sialic acid within the glycosylphosphatidylinositol anchor targets the cellular prion protein to synapses. *J. Biol. Chem.* 291, 17093–17101.

Batram, C., Jones, N.G., Janzen, C.J., Markert, S.M., Engstler, M., 2014. Expression site attenuation mechanistically links antigenic variation and development in *Trypanosoma brucei*. *Elife* 3, e02324.

Beneke, T., Madden, R., Makin, L., Valli, J., Sunter, J., Gluenz, E., 2017. A CRISPR Cas9 high-throughput genome editing toolkit for kinetoplastids. *R. Soc. Open Sci.* 4.

Bentley, S.J., Jamabo, M., Boshoff, A., 2019. The Hsp70/J-protein machinery of the African trypanosome, *Trypanosoma brucei*. *Cell Stress Chaperones* 24, 125–148.

Berriman, M. *et al.*, 2005. The genome of the African Trypanosome *Trypanosoma brucei*. *Science (80-.).* 309, 416–422.

Besada-Lombana, P.B., Da Silva, N.A., 2019. Engineering the early secretory pathway for increased protein secretion in *Saccharomyces cerevisiae*. *Metab. Eng.* 55, 142–151.

Biebinger, S., Clayton, C., 1996. A plasmid shuttle vector bearing an rRNA promoter is extrachromosomally maintained in *Crithidia fasciculata*. *Exp. Parasitol.* 83, 252–258.

Bindels, D.S. *et al.*, 2017. mScarlet: a bright monomeric red fluorescent protein for cellular imaging. *Nat. Methods* 14, 53–56.

Birch, J., Axford, D., Foadi, J., Meyer, A., Eckhardt, A., Thielmann, Y., Moraes, I., 2018. The fine art of integral membrane protein crystallisation. *Methods* 147, 150–162.

Black, S.J., Mansfield, J.M., 2016. Prospects for vaccination against pathogenic African trypanosomes. *Parasite Immunol.* 38, 735–743.

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

Boehm, C.M. *et al.*, 2017. The trypanosome exocyst: A conserved structure revealing a new role in endocytosis. *PLOS Pathog.* 13, e1006063.

Böhme, U., Cross, G.A.M., 2002. Mutational analysis of the variant surface

glycoprotein GPI-anchor signal sequence in *Trypanosoma brucei*. *J. Cell Sci.* 115, 805–816.

Bonhivers, M., Nowacki, S., Landrein, N., Robinson, D.R., 2008. Biogenesis of the trypanosome endo-exocytotic organelle is cytoskeleton mediated. *PLoS Biol.* 6, e105.

Bonnet, J., Boudot, C., Courtioux, B., 2015. Overview of the diagnostic methods used in the field for human African trypanosomiasis: What could change in the next years? *Biomed Res. Int.* 2015, 1–10.

Bonnon, C. *et al.*, 2010. Selective export of human GPI-anchored proteins from the endoplasmic reticulum. *J. Cell Sci.* 123, 1705–1715.

Bouvier, J., Etges, R.J., Bordiers, C., 1985. Identification and purification of membrane and soluble forms of the major surface protein of *Leishmania* promastigotes. *J. Biol. Chem.* 260, 15504–15509.

Breitling, R. *et al.*, 2002. Non-pathogenic trypanosomatid protozoa as a platform for protein research and production. *Protein Expr. Purif.* 25, 209–218.

Bretthauer, R.K., Castellino, F.J., 1999. Glycosylation of *Pichia pastoris*-derived proteins. *Biotechnol. Appl. Biochem* 30, 193–200.

Bukachi, S.A., Wandibba, S., Nyamongo, I.K., 2017. The socio-economic burden of human African trypanosomiasis and the coping strategies of households in the South Western Kenya foci. *PLoS Negl. Trop. Dis.* 11, e0006002.

Burkard, G.S., Jutzi, P., Roditi, I., 2011. Genome-wide RNAi screens in bloodstream form trypanosomes identify drug transporters. *Mol. Biochem. Parasitol.* 175, 91–94.

Burton, P., McBride, D.J., Wilkes, J.M., Barry, J.D., McCulloch, R., 2007. Ku heterodimer-independent end joining in *Trypanosoma brucei* cell extracts relies upon sequence microhomology. *Eukaryot. Cell* 6, 1773–1781.

Buxbaum, L.U., Milne, K.G., Werbovetz, K.A., Englund, P.T., 1996. Myristate exchange on the *Trypanosoma brucei* variant surface glycoprotein. *Proc. Natl. Acad. Sci. U. S. A.* 93, 1178–1183.

Buxbaum, L.U., Raper, J., Opperdoes, F.R., Englund, P.T., 1994. Myristate exchange. A second glycosyl phosphatidylinositol myristoylation reaction in African trypanosomes. *J. Biol. Chem.* 269, 30212–30220.

Campillo, N., Carrington, M., 2003. The origin of the serum resistance

associated (SRA) gene and a model of the structure of the SRA polypeptide from *Trypanosoma brucei rhodesiense*. *Mol. Biochem. Parasitol.* 127, 79–84.

Capes, A.S., Crossman, A., Urbaniak, M.D., Gilbert, S.H., Ferguson, M.A.J., Gilbert, I.H., 2014. Probing the substrate specificity of *Trypanosoma brucei* GlcNAc-PI de-N-acetylase with synthetic substrate analogues. *Org. Biomol. Chem.* 12, 1919–1934.

Capewell, P. *et al.*, 2013. The TgsGP gene is essential for resistance to human serum in *Trypanosoma brucei gambiense*. *PLoS Pathog.* 9, e1003686.

Caplan, A J, Cyr, D.M., Douglas, M.G., 1992. YDJ1p facilitates polypeptide translocation across different intracellular membranes by a conserved mechanism. *Cell* 71, 1143–1155.

Caplan, Avrom J, Tsai, J., Casey, P.J., Douglass, M.G., 1992. Farnesylation of YDJlp is required for function at elevated growth temperatures in *Saccharomyces cereuisiae*. *J. Biol. Chem.* 267, 18890–18895.

Caramelo, J.J., Parodi, A.J., 2015. A sweet code for glycoprotein folding. *FEBS Lett.* 589, 3379–3387.

Caras, I.W., Moran, P., 1994. The requirements for GPI-attachment are similar but not identical in mammalian cells and parasitic protozoa. *Brazilian J. Med. Biol. Res.* 27, 185–188.

Carrington, M., Miller, N., Blum, M., Roditi, I., Wiley, D., Turner, M., 1991. Variant specific glycoprotein of *Trypanosoma brucei* consists of two domains each having an independently conserved pattern of cysteine residues. *J. Mol. Biol.* 221, 823–835.

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

Carrió, M.M., Cubarsi, R., Villaverde, A., 2000. Fine architecture of bacterial inclusion bodies. *FEBS Lett.* 471, 7–11.

Castillon, G.A., Aguilera-Romero, A., Manzano-Lopez, J., Epstein, S., Kajiwara, K., Funato, K., Watanabe, R., Riezman, H., Muñiz, M., 2011. The yeast p24 complex regulates GPI-anchored protein transport and quality control by monitoring anchor remodeling. *Mol. Biol. Cell* 22, 2924–2936.

Chambers, A.C., Aksular, M., Graves, L.P., Irons, S.L., Possee, R.D., King, L.A.,

2018. Overview of the baculovirus expression system, in: *Current Protocols in Protein Science*. John Wiley & Sons, Inc., Hoboken, NJ, USA, pp. 5.4.1-5.4.6.

Chang, T., Milne, K.G., Güther, M.L.S., Smith, T.K., Ferguson, M.A.J., 2002. Cloning of *Trypanosoma brucei* and *Leishmania major* genes encoding the GlcNAc-phosphatidylinositol de-N-acetylase of glycosylphosphatidylinositol biosynthesis that is essential to the African sleeping sickness parasite. *J. Biol. Chem.* 277, 50176–50182.

Chartron, J.W., Hunt, K.C.L., Frydman, J., 2016. Cotranslational signal-independent SRP preloading during membrane targeting. *Nature* 536, 224–228.

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.

Cherepanova, N., Shrimal, S., Gilmore, R., 2016. N-linked glycosylation and homeostasis of the endoplasmic reticulum. *Curr. Opin. Cell Biol.* 41, 57–65.

Chih, B., Liu, P., Chinn, Y., Chalouni, C., Komuves, L.G., Hass, P.E., Sandoval, W., Peterson, A.S., 2012. A ciliopathy complex at the transition zone protects the cilia as a privileged membrane domain. *Nat. Cell Biol.* 14, 61–72.

Chung, K.T., Shen, Y., Hendershot, L.M., 2002. BAP, a mammalian BiP-associated protein, is a nucleotide exchange factor that regulates the ATPase activity of BiP. *J. Biol. Chem.* 277, 47557–47563.

Chung, W.-L., Leung, K.F., Carrington, M., Field, M.C., 2008. Ubiquitylation is required for degradation of transmembrane surface proteins in trypanosomes. *Traffic* 9, 1681–1697.

Ciulla, D.A., Jorgensen, M.T., Giner, J.-L., Callahan, B.P., 2018. Chemical bypass of general base catalysis in Hedgehog protein cholesterolysis using a hyper-nucleophilic substrate. *J. Am. Chem. Soc.* 140, 916–918.

Clayton, C., 2019. Regulation of gene expression in trypanosomatids: living with polycistronic transcription. *Open Biol.* 9, 190072.

Comini, M., Menge, U., Wissing, J., Flohé, L., 2005. Trypanothione synthesis in

crithidia revisited. *J. Biol. Chem.* 280, 6850–6860.

Conde, R., Cueva, R., Pablo, G., Polaina, J., Larriba, G., 2004. A search for hyperglycosylation signals in yeast glycoproteins. *J. Biol. Chem.* 279, 43789–43798.

Costantini, L.M., Fossati, M., Francolini, M., Snapp, E.L., 2012. Assessing the tendency of fluorescent proteins to oligomerize under physiologic conditions. *Traffic* 13, 643–649.

Cranfill, P.J. *et al.*, 2016. Quantitative assessment of fluorescent proteins. *Nat. Methods* 13, 557–562.

Croset, A. *et al.*, 2012. Differences in the glycosylation of recombinant proteins expressed in HEK and CHO cells. *J. Biotechnol.* 161, 336–348.

Cross, G.A.M., 2017. Tools for genetic analysis in *Trypanosoma brucei*. www.tryps.rockefeller.edu/trypsru2_genetics.html (accessed 7.17.19).

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

Cross, G.A.M., 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.

d'Avila-Levy, C.M., Almeida Dias, F., Melo, A.C.N., Martins, J.L., Carvalho Santos Lopes, A.H., Dos Santos, A.L.S., Vermelho, A.B., Branquinha, M.H., 2006. Insights into the role of gp63-like proteins in lower trypanosomatids. *FEMS Microbiol. Lett.* 254, 149–156.

D'Avila-Levy, C.M., Altoé, E.C.F., Uehara, L.A., Branquinha, M.H., Santos, A.L.S., 2014. GP63 function in the interaction of trypanosomatids with the invertebrate host: Facts and prospects, in: Santos, A.L.S., Branquinha, M.H., D'Avila-Levy, C.M., Kneipp, L.F., Sodré, C.L. (Eds.), *Proteins and Proteomics of Leishmania and Trypanosoma*. Springer Netherlands: Dordrecht, pp. 253–270.

Damerow, M.;, Rodrigues, J.A., Wu, D.;, Güther, M.L.S., Mehlert, A.;, Ferguson, M.A.J., 2014. Identification and functional characterization of a highly divergent N-acetylglucosaminyltransferase I (TbGnTI) in *Trypanosoma brucei*. *J. Biol. Chem.* 289, 9328–9339.

Damerow, M., Graalfs, F., Güther, M.L.S., Mehlert, A., Izquierdo, L., Ferguson,

M.A.J., 2016. A gene of the β3-glycosyltransferase family encodes N-acetylglucosaminyltransferase II function in *Trypanosoma brucei*. *J. Biol. Chem.* 291, 13834–13845.

Daniels, J.-P., Gull, K., Wickstead, B., William, S., 2012. The trypanosomatid-specific N terminus of RPA2 Is required for RNA polymerase I assembly, localization, and function. *Eukaryot. Cell* 5, 662–672.

Davis, T.W., Brown, J.D., Walter, P., 1996. Signal sequences specify the targeting route to the endoplasmic reticulum membrane. *J. Cell Biol.* 132, 269–278.

de Freitas Nascimento, J., Kelly, S., Sunter, J., Carrington, M., 2018. Codon choice directs constitutive mRNA levels in trypanosomes. *Elife* 7.

de Souza, W., de Carvalho, T.M.U., Barrias, E.S., de Souza, W., de Carvalho, T.M.U., Barrias, E.S., 2010. Review on Trypanosoma cruzi: Host cell interaction. *Int. J. Cell Biol.* 2010, 1–18.

Dean, S., Sunter, J., Wheeler, R.J., Hodkinson, I., Gluenz, E., Gull, K., 2015. A toolkit enabling efficient, scalable and reproducible gene tagging in trypanosomatids. *Open Biol.* 5, 140197.

Dean, S., Marchetti, R., Kirk, K., Matthews, K.R., 2009. A surface transporter family conveys the trypanosome differentiation signal. *Nature* 459, 213–217.

Debeljak, N., Feldman, L., Davis, K.L., Komel, R., Sytkowski, A.J., 2006. Variability in the immunodetection of His-tagged recombinant proteins. *Anal. Biochem.* 359, 216–223.

DeJesus, E., Kieft, R., Albright, B., Stephens, N.A., Hajduk, S.L., 2013. A single amino acid substitution in the group 1 *Trypanosoma brucei gambiense* haptoglobin-hemoglobin receptor abolishes TLF-1 binding. *PLoS Pathog.* 9, e1003317.

Delhi, P., Queiroz, R., Inchaustegui, D., Carrington, M., Clayton, C., 2011. Is there a classical nonsense-mediated decay pathway in trypanosomes? *PLoS One* 6, e25112.

Dempsey, W.L., Mansfield, J.M., 1983. Lymphocyte function in experimental African trypanosomiasis. V. Role of antibody and the mononuclear phagocyte system in variant-specific immunity. *J. Immunol.* 130, 405–411.

Deshaies, R.J., Koch, B.D., Werner-Washburne, M., Craig, E.A., Schekman, R.,

1988. A subfamily of stress proteins facilitates translocation of secretory and mitochondrial precursor polypeptides. *Nature* 332, 800–805.

Dhar, P., McAuley, J., 2019. The role of the cell surface mucin MUC1 as a barrier to infection and regulator of inflammation. *Front. Cell. Infect. Microbiol.* 9, 117.

Diffley, P., 1985. *Trypanosoma brucei*: Immunogenicity of the variant surface coat glycoprotein of virulent and avirulent subspecies. *Exp. Parasitol.* 59, 98–107.

Dingermann, T., Frank-Stoll, U., Werner, H., Wissmann, A., Hillen, W., Jacquet, M., Marschalek, R., 1992. RNA polymerase III catalysed transcription can be regulated in Saccharomyces cerevisiae by the bacterial tetracycline repressor-operator system. *EMBO J.* 11, 1487–1492.

Dudek, J., Pfeffer, S., Lee, P.-H., Jung, M., Cavalié, A., Helms, V., Förster, F., Zimmermann, R., 2015. Protein transport into the human endoplasmic reticulum. *J. Mol. Biol.* 427, 1159–1175.

Duffy, J., Patham, B., Mensa-Wilmot, K., 2010. Discovery of functional motifs in h-regions of trypanosome signal sequences. *Biochem. J.* 426, 410–415.

Duong-Ly, K.C., Gabelli, S.B., 2014. Salting out of proteins using ammonium sulfate precipitation, in: *Methods in Enzymology*. Elsevier Inc., pp. 85–94.

Dyrløv Bendtsen, J., Nielsen, H., von Heijne, G., Brunak, S., 2004. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* 340, 783–795.

Edkins, A.L., Ludewig, M.H., Blatch, G.L., 2004. A *Trypanosoma cruzi* heat shock protein 40 is able to stimulate the adenosine triphosphate hydrolysis activity of heat shock protein 70 and can substitute for a yeast heat shock protein 40. *Int. J. Biochem. Cell Biol.* 36, 1585–1598.

Eisenhaber, B., Bork, P., Eisenhaber, F., 1998. Sequence properties of GPI-anchored proteins near the ω-site: constraints for the polypeptide binding site of the putative transamidase. *Protein Eng.* 11, 1155–1161.

Eisenhaber, B., Eisenhaber, S., Kwang, T.Y., Grüber, G., Eisenhaber, F., 2014. Transamidase subunit GAA1/GPAA1 is a M28 family metallo-peptide-synthetase that catalyzes the peptide bond formation between the substrate protein's omega-site and the GPI lipid anchor's phosphoethanolamine. *Cell Cycle* 13, 1912–1917.

Eisenhaber, B., Sinha, S., Wong, W.-C., Eisenhaber, F., 2018. Function of a

membrane-embedded domain evolutionarily multiplied in the GPI lipid anchor pathway proteins PIG-B, PIG-M, PIG-U, PIG-W, PIG-V, and PIG-Z. *Cell Cycle* 1–7.

Elton, C.M., Rodriguez, M., Ben Mamoun, C., Lobo, C.A., Wright, G.J., 2019. A library of recombinant *Babesia microti* cell surface and secreted proteins for diagnostics discovery and reverse vaccinology. *Int. J. Parasitol.* 49, 115–125.

Engstler, M. *et al.*, 2007. Hydrodynamic flow-mediated protein sorting on the cell surface of trypanosomes. *Cell* 131, 505–515.

Engstler, M., Weise, F., Bopp, K., Grünfelder, C.G., Günzel, M., Heddergott, N., Overath, P., 2005. The membrane-bound histidine acid phosphatase TbMBAP1 is essential for endocytosis and membrane recycling in *Trypanosoma brucei*. *J. Cell Sci.* 118, 2105–2118.

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

Esson, H.J., Morriswood, B., Yavuz, S., Vidilaseris, K., Dong, G., Warren, G., 2012. Morphology of the trypanosome bilobe, a novel cytoskeletal structure. *Eukaryot. Cell* 11, 761–772.

Estévez, A.M., 2008. The RNA-binding protein TbDRBD3 regulates the stability of a specific subset of mRNAs in trypanosomes. *Nucleic Acids Res.* 36, 4573–4586.

Ferguson, M. a, Homans, S.W., Dwek, R. a, Rademacher, T.W., 1988. Glycosyl-phosphatidylinositol moiety that anchors *Trypanosoma brucei* variant surface glycoprotein to the membrane. *Science (80-.).* 239, 753–759.

Ferguson, M.A., Kinoshita, T., Hart, G.W., 2009. Glycosylphosphatidylinositol Anchors, in: Varki, A. et al. (Eds.), *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press.

Ferguson, M.A.J., Hart, G.W., Kinoshita, T., 2017. Glycosylphosphatidylinositol Anchors, in: Varki, A. et al. (Eds.), *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press.

Ferguson, M.A.J., Low, M.G., Cros, G.A.M., 1985. Glycosyl-sn-1,2-dimyristylphosphatidylinositol is covalently linked to *Trypanosoma brucei*

Variant Surface Glycoprotein. *J. Biol. Chem.* 260, 14547–14555.

Ferrante, A., Allison, A.C., 1983. Alternative pathway activation of complement by African trypanosomes lacking a glycoprotein coat. *Parasite Immunol.* 5, 491–498.

Field, M.C., Carrington, M., 2009. The trypanosome flagellar pocket. *Nat. Rev. Microbiol.* 7, 775–786.

Field, M.C., Menon, A.K., Cross, G.A., 1991. A glycosylphosphatidylinositol protein anchor from procyclic stage *Trypanosoma brucei*: lipid structure and biosynthesis. *EMBO J.* 10, 2731–2739.

Filosa, J.N. *et al.*, 2019. Dramatic changes in gene expression in different forms of *Crithidia fasciculata* reveal potential mechanisms for insect-specific adhesion in kinetoplastid parasites. *PLoS Negl. Trop. Dis.* 13, e0007570.

Finke, K., Plath, K., Panzner, S., Prehn, S., Rapoport, T.A., Hartmann, E., Sommer, T., 1996. A second trimeric complex containing homologs of the Sec61p complex functions in protein transport across the ER membrane of S. cerevisiae. *EMBO J.* 15, 1482–1494.

Florimond, C. *et al.*, 2015. BILBO1 is a scaffold protein of the flagellar pocket collar in the pathogen *Trypanosoma brucei*. *PLoS Pathog.* 11, e1004654.

Fraering, P., Imhof, I., Meyer, U., Strub, J.M., van Dorsselaer, A., Vionnet, C., Conzelmann, A., 2001. The GPI transamidase complex of *Saccharomyces cerevisiae* contains Gaa1p, Gpi8p, and Gpi16p. *Mol. Biol. Cell* 12, 3295–306.

Franco, J.R., Cecchi, G., Priotto, G., Paone, M., Diarra, A., Grout, L., Simarro, P.P., Zhao, W., Argaw, D., 2018. Monitoring the elimination of human African trypanosomiasis: Update to 2016. *PLoS Negl. Trop. Dis.* 12, e0006890.

Freymann, D., Down, J., Carrington, M., Roditi, I., Turner, M., Wiley, D., 1990. 2.9 A resolution structure of the N-terminal domain of a variant surface glycoprotein from *Trypanosoma brucei*. *J. Mol. Biol.* 216, 141–160.

Fridy, P.C. *et al.*, 2014. A robust pipeline for rapid production of versatile nanobody repertoires. *Nat. Methods* 11, 1253–1260.

Furger, A., Schurch, N., Kurath, U., Roditi, I., 1997. Elements in the 3' untranslated region of procyclin mRNA regulate expression in insect forms of *Trypanosoma brucei* by modulating RNA stability and translation. *Mol.*

Cell. Biol. 17, 4372–4380.

Furuse, M., 2010. Molecular basis of the core structure of tight junctions. *Cold Spring Harb. Perspect. Biol.* 2, a002907.

Gadelha, C., Zhang, W., Chamberlain, J.W., Chait, B.T., Wickstead, B., Field, M.C., 2015. Architecture of a host-parasite interface: Complex targeting mechanisms revealed through proteomics. *Mol. Cell. Proteomics* 14, 1911–1926.

Gadelha, C., Rothery, S., Morphew, M., McIntosh, J.R., Severs, N.J., Gull, K., 2009. Membrane domains and flagellar pocket boundaries are influenced by the cytoskeleton in African trypanosomes. *Proc. Natl. Acad. Sci. U. S. A.* 106, 17425–17430.

Gadelha, C., Holden, J.M., Allison, H.C., Field, M.C., 2011. Specializations in a successful parasite: what makes the bloodstream-form African trypanosome so deadly? *Mol. Biochem. Parasitol.* 179, 51–8.

Gadelha, C., Wickstead, B., de Souza, W., Gull, K., Cunha-e-Silva, N., 2005. Cryptic paraflagellar rod in endosymbiont-containing kinetoplastid protozoa. *Eukaryot. Cell* 4, 516–525.

Gatz, C., Quailt, P.H., 1988. TnlO-encoded tet repressor can regulate an operator-containing plant promoter (cauliflower mosaic virus 35S promoter/electroporation/transient chloramphenicol acetyltransferase assays). *Proc. Nati. Acad. Sci. USA* 85, 1394–1397.

Gibson, D.G., Young, L., Chuang, R.-Y., Venter, J.C., Hutchison, C.A., Smith, H.O., 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* 6, 343–345.

Goldshmidt, H., Sheiner, L., Butikofer, P., Roditi, I., Uliel, S., Gunzel, M., Engstler, M., Michaeli, S., 2008. Role of protein translocation pathways across the endoplasmic reticulum in *Trypanosoma brucei*. *J. Biol. Chem.* 283, 32085–32098.

Gomes, A.R., Byregowda, S.M., Belamaranahally, Veeregowda, M., Balamurugan, V., 2016. An Overview of heterologous expression host systems for the production of recombinant proteins. *Adv. Anim. Vet. Sci.* 4, 346–356.

Gottier, P., Gonzalez-Salgado, A., Menon, A.K., Liu, Y.-C., Acosta-Serrano, A., Bütikofer, P., 2017. RFT1 protein affects glycosylphosphatidylinositol (GPI)

anchor glycosylation. *J. Biol. Chem.* 292, 1103–1111.

Graf, F.E. *et al.*, 2013. Aquaporin 2 mutations in *Trypanosoma brucei* gambiense field isolates correlate with decreased susceptibility to pentamidine and melarsoprol. *PLoS Negl. Trop. Dis.* 7, e2475.

Grandgenett, P.M., Otsu, K., Wilson, H.R., Wilson, M.E., Donelson, J.E., 2007. A function for a specific zinc metalloprotease of African trypanosomes. *PLoS Pathog.* 3, 1432–1445.

Grondin, K., Kündig, C., Roy, G., Ouellette, M., 1998. Linear amplicons as precursors of amplified circles in methotrexate-resistant *Leishmania tarentolae*. *Nucleic Acids Res.* 26, 3372–3378.

Grove, D.E., Fan, C.-Y., Ren, H.Y., Cyr, D.M., 2011. The endoplasmic reticulum-associated Hsp40 DNAJB12 and Hsc70 cooperate to facilitate RMA1 E3-dependent degradation of nascent CFTRDeltaF508. *Mol. Biol. Cell* 22, 301–314.

Grunfelder, C.G., Engstler, M., Weise, F., Schwarz, H., Stierhof, Y.-D., Morgan, G.W., Field, M.C., Overath, P., 2003. Endocytosis of a glycosylphosphatidylinositol-anchored protein via clathrin-coated vesicles, sorting by default in endosomes, and exocytosis via RAB11-positive carriers. *Mol. Biol. Cell* 14, 2029–2040.

Grunfelder, 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* 3, 547–559.

Güler-Gane, G., Kidd, S., Sridharan, S., Vaughan, T.J., Wilkinson, T.C.I., Tigue, N.J., 2016. Overcoming the refractory expression of secreted recombinant proteins in mammalian cells through modification of the signal peptide and adjacent amino acids. *PLoS One* 11, e0155340.

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.

Güther, M.L., Ferguson, M.A., 1995. The role of inositol acylation and inositol deacylation in GPI biosynthesis in *Trypanosoma brucei*. *EMBO J.* 14, 3080–3093.

Güther, M.L., Leal, S., Morrice, N.A., Cross, G.A., Ferguson, M.A., 2001.

Purification, cloning and characterization of a GPI inositol deacylase from *Trypanosoma brucei*. *EMBO J.* 20, 4923–4934.

Guther, M.L., Masterson, W.J., Ferguson, M.A., 1994. The effects of phenylmethylsulfonyl fluoride on inositol-acylation and fatty acid remodeling in African trypanosomes. *J. Biol. Chem.* 269, 18694–18701.

Gutmann, D.A.P., Mizohata, E., Newstead, S., Ferrandon, S., Postis, V., Xia, X., Henderson, P.J.F., van Veen, H.W., Byrne, B., 2007. A high-throughput method for membrane protein solubility screening: the ultracentrifugation dispersity sedimentation assay. *Protein Sci.* 16, 1422–1428.

Hall, B., Allen, C.L., Goulding, D., Field, M.C., 2004. Both of the Rab5 subfamily small GTPases of *Trypanosoma brucei* are essential and required for endocytosis. *Mol. Biochem. Parasitol.* 138, 67–77.

Hamers-Casterman, C., Atarhouch, T., Muyldermans, S., Robinson, G., Hammers, C., Songa, E.B., Bendahman, N., Hammers, R., 1993. Naturally occurring antibodies devoid of light chains. *Nature* 363, 446–448.

Hamilton, S.R., Gerngross, T.U., 2007. Glycosylation engineering in yeast: the advent of fully humanized yeast. *Curr. Opin. Biotechnol.* 18, 387–392.

Harlow, E., Lane, D., 2014. Immunizing animals, in: Greenfield, E.A. (Ed.), *Antibodies : A Laboratory Manual*. Cold Spring Harbor Laboratory Press, pp. 107–200.

Harmsen, M.M., De Haard, H.J., 2007. Properties, production, and applications of camelid single-domain antibody fragments. *Appl. Microbiol. Biotechnol.* 77, 13–22.

Harrison, R.L., Jarvis, D.L., 2006. Protein N-Glycosylation in the baculovirus-insect cell expression system and engineering of insect cells to produce “mammalianized” recombinant glycoproteins. *Adv. Virus Res.* 68, 159–191.

Hashem, Y. *et al.*, 2013. High-resolution cryo-electron microscopy structure of the *Trypanosoma brucei* ribosome. *Nature* 494, 385–389.

Haßdenteufel, S., Johnson, N., Paton, A.W., Paton, J.C., High, S., Zimmermann, R., 2018. Chaperone-mediated Sec61 channel gating during ER import of small precursor proteins overcomes Sec61 inhibitor-reinforced energy barrier. *Cell Rep.* 23, 1373–1386.

Hedstrom, K.L., Ogawa, Y., Rasband, M.N., 2008. AnkyrinG is required for maintenance of the axon initial segment and neuronal polarity. *J. Cell Biol.*

183, 635–640.

Heins, L., Frohberg, C., Gatz, C., 1992. The Tn 10-encoded tet repressor blocks early but not late steps of assembly of the RNA polymerase II initiation complex in vivo. *MGG Mol. Gen. Genet.* 232, 328–331.

Helenius, J., Ng, D.T.W., Marolda, C.L., Walter, P., Valvano, M.A., Aebi, M., 2002. Translocation of lipid-linked oligosaccharides across the ER membrane requires Rft1 protein. *Nature* 415, 447–450.

Herzik, M.A., Wu, M., Lander, G.C., 2019. High-resolution structure determination of sub-100 kDa complexes using conventional cryo-EM. *Nat. Commun.* 10, 1032.

Higgins, M.K., Tkachenko, O., Brown, A., Reed, J., Raper, J., Carrington, M., 2013. Structure of the trypanosome haptoglobin-hemoglobin receptor and implications for nutrient uptake and innate immunity. *Proc. Natl. Acad. Sci. U. S. A.* 110, 1905–1910.

Hirata, T. *et al.*, 2018. Identification of a Golgi GPI-N-acetylgalactosamine transferase with tandem transmembrane regions in the catalytic domain. *Nat. Commun.* 9, 405.

Homans, S.W., Ferguson, M.A.J., Dwek, R.A., Rademacher, T.W., Anand, R., Williams, A.F., 1988. Complete structure of the glycosyl phosphatidylinositol membrane anchor of rat brain Thy-1 glycoprotein. *Nature* 334, 601–604.

Hong, Y., Nagamune, K., Morita, Y.S., Nakatani, F., Ashida, H., Maeda, Y., Kinoshita, T., 2006. Removal or maintenance of inositol-linked acyl chain in glycosylphosphatidylinositol is critical in trypanosome life cycle. *J. Biol. Chem.* 281, 11595–11602.

Hong, Y., Maeda, Y., Watanabe, R., Inoue, N., Ohishi, K., Kinoshita, T., 2000. Requirement of PIG-F and PIG-O for transferring phosphoethanolamine to the third mannose in glycosylphosphatidylinositol. *J. Biol. Chem.* 275, 20911–20919.

Hong, Y., Maeda, Y., Watanabe, R., Ohishi, K., Mishkind, M., Riezman, H., Kinoshita, T., 1999. Pig-n, a mammalian homologue of yeast Mcd4p, is involved in transferring phosphoethanolamine to the first mannose of the glycosylphosphatidylinositol. *J. Biol. Chem.* 274, 35099–35106.

Hong, Y., Kinoshita, T., 2009. Trypanosome glycosylphosphatidylinositol

biosynthesis. *Korean J. Parasitol.* 47, 197–204.

Hu, Q., Milenkovic, L., Jin, H., Scott, M.P., Nachury, M. V, Spiliotis, E.T., Nelson, W.J., 2010. A septin diffusion barrier at the base of the primary cilium maintains ciliary membrane protein distribution. *Science* 329, 436–439.

Ichimura, Y. *et al.*, 2000. A ubiquitin-like system mediates protein lipidation. *Nature* 408, 488–492.

Ihara, M. *et al.*, 2005. Cortical organization by the septin cytoskeleton is essential for structural and mechanical integrity of mammalian spermatozoa. *Dev. Cell* 8, 343–352.

Izquierdo, L., Schulz, B.L., Rodrigues, J.A., Güther, M.L.S., Procter, J.B., Barton, G.J., Aebi, M., Ferguson, M.A.J., 2009a. Distinct donor and acceptor specificities of *Trypanosoma brucei* oligosaccharyltransferases. *EMBO J.* 28, 2650–2661.

Izquierdo, L., Nakanishi, M., Mehlert, A., Machray, G., Barton, G.J., Ferguson, M.A.J., 2009b. Identification of a glycosylphosphatidylinositol anchor-modifying β1-3 N-acetylglucosaminyl transferase in *Trypanosoma brucei*. *Mol. Microbiol.* 71, 478–491.

Izquierdo, L., Atrih, A., Rodrigues, J.A., Jones, D.C., Ferguson, M.A.J., 2009c. *Trypanosoma brucei* UDP-glucose:glycoprotein glucosyltransferase has unusual substrate specificity and protects the parasite from stress. *Eukaryot. Cell* 8, 230–40.

Izquierdo, L., Mehlert, A., Ferguson, M.A.J., 2012. The lipid-linked oligosaccharide donor specificities of *Trypanosoma brucei* oligosaccharyltransferases. *Glycobiology* 22, 696–703.

Jackson, A.P., Vaughan, S., Gull, K., 2006. Evolution of tubulin gene arrays in Trypanosomatid parasites: genomic restructuring in *Leishmania*. *BMC Genomics* 7, 261.

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.

Jamonneau, V. *et al.*, 2015. Accuracy of individual rapid tests for serodiagnosis of *gambiense* sleeping sickness in West Africa. *PLoS Negl. Trop. Dis.* 9, e0003480.

Jaquenoud, M., Pagac, M., Signorell, A., Benghezal, M., Jelk, J., Bütikofer, P.,

Conzelmann, A., 2008. The Gup1 homologue of *Trypanosoma brucei* is a GPI glycosylphosphatidylinositol remodelase. *Mol. Microbiol.* 67, 202–212.

Jeacock, L., Faria, J., Horn, D., 2018. Codon usage bias controls mRNA and protein abundance in trypanosomatids. *Elife* 7, e32496.

Jelk, J. *et al.*, 2013. Glycoprotein biosynthesis in a eukaryote lacking the membrane protein Rft1. *J. Biol. Chem.* 288, 20616–20623.

JenaBioscience, n.d. LEXSY - Eukaryotic protein expression in *Leishmania tarentolae* - Jena Bioscience. www.jenabioscience.com/lexsy-expression (accessed 4.29.19).

Jensen, D. *et al.*, 2011. COPII-mediated vesicle formation at a glance. *J. Cell Sci.* 124, 1–4.

Jinnelov, A., Ali, L., Tinti, M., Güther, M.L.S., Ferguson, M.A.J., 2017. Single-subunit oligosaccharyltransferases of *Trypanosoma brucei* display different and predictable peptide acceptor specificities. *J. Biol. Chem.* 292, 20328–20341.

Johnson, N., Vilardi, F., Lang, S., Leznicki, P., Zimmermann, R., High, S., 2012. TRC40 can deliver short secretory proteins to the Sec61 translocon. *J. Cell Sci.* 125, 3612–3620.

Jones, D.C., Mehlert, A., Güther, M.L.S., Ferguson, M.A.J., 2005. Deletion of the glucosidase II gene in *Trypanosoma brucei* reveals novel N-glycosylation mechanisms in the biosynthesis of variant surface glycoprotein. *J. Biol. Chem.* 280, 35929–35942.

Kabeya, Y., Mizushima, N., Yamamoto, A., Oshitani-Okamoto, S., Ohsumi, Y., Yoshimori, T., 2004. LC3, GABARAP and GATE16 localize to autophagosomal membrane depending on form-II formation. *J. Cell Sci.* 117, 2805–2812.

Kane, S.P., 2019. Sample Size Calculator. ClinCalc. www.clincalc.com/stats/samplesize.aspx (accessed 7.17.19).

Kang, J.Y., Hong, Y., Ashida, H., Shishioh, N., Murakami, Y., Morita, Y.S., Maeda, Y., Kinoshita, T., 2005. PIG-V involved in transferring the second mannose in glycosylphosphatidylinositol. *J. Biol. Chem.* 280, 9489–9497.

Kang, X., Szallies, A., Rawer, M., Echner, H., Duszenko, M., 2002. GPI anchor transamidase of *Trypanosoma brucei*: in vitro assay of the recombinant protein and VSG anchor exchange. *J. Cell Sci.* 115, 2529–2539.

Kastner, M., Neubert, D., 1991. High-performance metal chelate affinity chromatography of cytochromes P-450 using Chelating Superose. *J. Chromatogr. A* 587, 43–54.

Kelly, E.E. *et al.*, 2012. The Rab family of proteins: 25 years on. *Biochem. Soc. Trans.* 40, 1337–1347.

Khan, A.H., Bayat, H., Rajabibazl, M., Sabri, S., Rahimpour, A., 2017. Humanizing glycosylation pathways in eukaryotic expression systems. *World J. Microbiol. Biotechnol.* 33.

Khoshouei, M., Radjainia, M., Baumeister, W., Danev, R., 2017. Cryo-EM structure of haemoglobin at 3.2 Å determined with the Volta phase plate. *Nat. Commun.* 8, 16099.

Kieft, R., Capewell, P., Turner, C.M.R., Veitch, N.J., MacLeod, A., Hajduk, S., 2010. Mechanism of *Trypanosoma brucei gambiense* (group 1) resistance to human trypanosome lytic factor. *Proc. Natl. Acad. Sci. U. S. A.* 107, 16137–16141.

Kim, H., Yoo, S.J., Kang, H.A., 2014. Yeast synthetic biology for the production of recombinant therapeutic proteins. *FEMS Yeast Res.* 15, 1–16.

Kim, J. *et al.*, 2015. Subnanometre-resolution electron cryomicroscopy structure of a heterodimeric ABC exporter. *Nature* 517, 396–400.

Kinoshita, T., Fujita, M., 2016. Biosynthesis of GPI-anchored proteins: special emphasis on GPI lipid remodeling. *J. Lipid Res.* 57, 6–24.

Kipandula, W., Smith, T.K., MacNeill, S.A., 2017. Tandem affinity purification of exosome and replication factor C complexes from the non-human infectious kinetoplastid parasite *Crithidia fasciculata*. *Mol. Biochem. Parasitol.* 217, 19–22.

Kirchhofer, A. *et al.*, 2010. Modulation of protein properties in living cells using nanobodies. *Nat. Struct. Mol. Biol.* 17, 133–138.

Klatt, S., Konthur, Z., 2012. Secretory signal peptide modification for optimized antibody-fragment expression-secretion in *Leishmania tarentolae*. *Microb. Cell Fact.* 11, 1.

Kořený, L., Lukeš, J., Oborník, M., 2010. Evolution of the haem synthetic pathway in kinetoplastid flagellates: An essential pathway that is not essential after all? *Int. J. Parasitol.* 40, 149–156.

Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L.L., 2001. Predicting

transmembrane protein topology with a hidden markov model: application to complete genomes. *J. Mol. Biol.* 305, 567–580.

Kruzel, E.K., Zimmett, G.P., Bangs, J.D., 2017. Life stage-specific cargo receptors facilitate glycosylphosphatidylinositol-anchored surface coat protein transport in *Trypanosoma brucei*. *mSphere* 2, e00282-17.

Kushnir, S., Cirstea, I.C., Basiliya, L., Lupilova, N., Breitling, R., Alexandrov, K., 2011. Artificial linear episome-based protein expression system for protozoon *Leishmania tarentolae*. *Mol. Biochem. Parasitol.* 176, 69–79.

Kushnir, S., Gase, K., Breitling, R., Alexandrova, K., 2005. Development of an inducible protein expression system based on the protozoan host *Leishmania tarentolae*. *Protein Expr. Purif.* 42, 37–46.

Kwitny, S., Klaus, A. V, Hunnicutt, G.R., 2010. The annulus of the mouse sperm tail is required to establish a membrane diffusion barrier that is engaged during the late steps of spermiogenesis. *Biol. Reprod.* 82, 669–678.

La Greca, F., Magez, S., 2011. Vaccination against trypanosomiasis. *Hum. Vaccin.* 7, 1225–1233.

Laffitte, M.-C.N., Leprohon, P., Hainse, M., Légaré, D., Masson, J.-Y., Ouellette, M., 2016. Chromosomal ranslocations in the parasite *Leishmania* by a MRE11/RAD50-independent microhomology-mediated end joining mechanism. *PLOS Genet.* 12, e1006117.

Lakkaraju, A.K.K., Thankappan, R., Mary, C., Garrison, J.L., Taunton, J., Strub, K., 2012. Efficient secretion of small proteins in mammalian cells relies on Sec62-dependent posttranslational translocation. *Mol. Biol. Cell* 23, 2712–2722.

Lança, A.S.C., de Sousa, K.P., Atouguia, J., Prazeres, D.M.F., Monteiro, G.A., Silva, M.S., 2011. *Trypanosoma brucei*: Immunisation with plasmid DNA encoding invariant surface glycoprotein gene is able to induce partial protection in experimental African trypanosomiasis. *Exp. Parasitol.* 127, 18–24.

Lane-Serff, H., MacGregor, P., Lowe, E.D., Carrington, M., Higgins, M.K., 2014. Structural basis for ligand and innate immunity factor uptake by the trypanosome haptoglobin-haemoglobin receptor. *Elife* 3, e05553.

Lee, J.H., Ozorowski, G., Ward, A.B., 2016. Cryo-EM structure of a native, fully glycosylated, cleaved HIV-1 envelope trimer. *Science* 351, 1043–1048.

Leterrier, C., 2018. The axon initial segment: An updated viewpoint. *J. Neurosci.* 38, 2135–2145.

Leung, K.F., Riley, F.S., Carrington, M., Field, M.C., 2011. Ubiquitylation and developmental regulation of invariant surface protein expression in trypanosomes. *Eukaryot. Cell* 10, 916–931.

Lin, D.T.S., Conibear, E., 2015. ABHD17 proteins are novel protein depalmitoylases that regulate N-Ras palmitate turnover and subcellular localization. *Elife* 4, e11306.

Linxweiler, M., Schick, B., Zimmermann, R., 2017. Let's talk about secs: Sec61, Sec62 and Sec63 in signal transduction, oncology and personalized medicine. *Signal Transduct. Target. Ther.* 2, 1–10.

Liu, Y., Huynh, D.T., Yeates, T.O., 2019. A 3.8 Å resolution cryo-EM structure of a small protein bound to an imaging scaffold. *Nat. Commun.* 10, 1864.

Longtine, M.S., Bi, E., 2003. Regulation of septin organization and function in yeast. *Trends Cell Biol.* 13, 403–409.

Loomes, K.M., Senior, H.E., West, P.M., Roberton, A.M., 1999. Functional protective role for mucin glycosylated repetitive domains. *Eur. J. Biochem.* 266, 105–111.

Love, K.R., Politano, T.J., Panagiotou, V., Jiang, B., Stadheim, T.A., Love, J.C., 2012. Systematic single-cell analysis of *Pichia pastoris* reveals secretory capacity limits productivity. *PLoS One* 7, e37915.

Low, P., Dallner, G., Mayor, S., Cohen, S., Chait, B.T., Menon, A.K., 1991. The mevalonate pathway in the bloodstream form of *Trypanosoma brucei*: Identification of dolichols containing 11 and 12 isoprene residues. *J. Biol. Chem.* 266, 19250–19257.

Loya, A., Pnueli, L., Yosefzon, Y., Wexler, Y., Ziv-Ukelson, M., Arava, Y., 2008. The 3'-UTR mediates the cellular localization of an mRNA encoding a short plasma membrane protein. *RNA* 14, 1352–1365.

Luca, V.C., Jude, K.M., Pierce, N.W., Nachury, M. V, Fischer, S., Garcia, K.C., 2015. Structural basis for Notch1 engagement of Delta-like 4. *Science* 347, 847–853.

Lukeš, J., Paris, Z., Regmi, S., Breitling, R., Mureev, S., Kushnir, S., Pyatkov, K., Jirků, M., Alexandrov, K.A., 2006. Translational initiation in *Leishmania tarentolae* and *Phytomonas serpens* (Kinetoplastida) is strongly influenced

by pre-ATG triplet and its 5 sequence context. *Mol. Biochem. Parasitol.* 148, 125–132.

Lumbala, C., Biéler, S., Kayembe, S., Makabuza, J., Ongarello, S., Ndung'u, J.M., 2018. Prospective evaluation of a rapid diagnostic test for *Trypanosoma brucei gambiense* infection developed using recombinant antigens. *PLoS Negl. Trop. Dis.* 12, e0006386.

Lustig, Y. *et al.*, 2007. Down-regulation of the trypanosomatid signal recognition particle affects the biogenesis of polytopic membrane proteins but not of signal peptide-containing proteins. *Eukaryot. Cell* 6, 1865–1875.

Lustig, Y., Goldshmidt, H., Uliel, S., Michaeli, S., 2005. The Trypanosoma brucei signal recognition particle lacks the Alu-domain-binding proteins: purification and functional analysis of its binding proteins by RNAi. *J. Cell Sci.* 118, 4551–4562.

Lyman, S.K., Schekman, R., 1997. Binding of secretory precursor polypeptides to a translocon subcomplex is regulated by BiP. *Cell* 88, 85–96.

Lyman, S.K., Schekman, R., Whitfield, K.M., Vogel, J.P., Rose, R.W., 1995. Interaction between BiP and Sec63p is required for the completion of protein translocation into the ER of *Saccharomyces cerevisiae*. *J. Cell Biol.* 131, 1163–1171.

MacGregor, P. *et al.*, 2019. A single dose of antibody-drug conjugate cures a stage 1 model of African trypanosomiasis. *PLoS Negl. Trop. Dis.* 13, e0007373.

Maeda, Y., Watanabe, R., Harris, C.L., Hong, Y., Ohishi, K., Kinoshita, K., Kinoshita, T., 2001. PIG-M transfers the first mannose to glycosylphosphatidylinositol on the lumenal side of the ER. *EMBO J.* 20, 250–261.

Manful, T., Fadda, A., Clayton, C., 2011. The role of the 5'-3' exoribonuclease XRNA in transcriptome-wide mRNA degradation. *RNA* 17, 2039–2047.

Manna, P.T., Boehm, C., Leung, K.F., Natesan, S.K., Field, M.C., 2014. Life and times: synthesis, trafficking, and evolution of VSG. *Trends Parasitol.* 30, 251–258.

Masuishi, Y., Kimura, Y., Arakawa, N., Hirano, H., 2016. Identification of glycosylphosphatidylinositol-anchored proteins and ω-sites using TiO2-based affinity purification followed by hydrogen fluoride treatment. *J.*

Proteomics 139, 77–83.

Matlack, K.E., Misselwitz, B., Plath, K., Rapoport, T.A., 1999. BiP acts as a molecular ratchet during posttranslational transport of prepro-alpha factor across the ER membrane. *Cell* 97, 553–564.

Mattanovich, D., Branduardi, P., Dato, L., Gasser, B., Sauer, M., Porro, D., 2012. Recombinant protein production in yeasts. Humana Press, Totowa, NJ, pp. 329–358.

Maxwell, S E, Ramalingam, S., Gerber, L.D., Brink, L., Udenfriend, S., 1995. An active carbonyl formed during glycosylphosphatidylinositol addition to a protein is evidence of catalysis by a transamidase. *J. Biol. Chem.* 270, 19576–19582.

Maxwell, S. E., Ramalingam, S., Gerber, L.D., Udenfriend, S., 1995. Cleavage without anchor addition accompanies the processing of a nascent protein to its glycosylphosphatidylinositol-anchored form. *Proc. Natl. Acad. Sci.* 92, 1550–1554.

Mayor, S., Menon, A.K., Cross, G.A.M., 1990. Glycolipid precursors for the membrane anchor of *Trypanosoma brucei* variant surface glycoproteins II. lipid structures of phosphatidylinositol-specific phospholipase C sensitive and resistant glycolipids. *J. Biol. Chem.* 265, 6174–6181.

Mazet, M. *et al.*, 2013. Revisiting the central metabolism of the bloodstream forms of *Trypanosoma brucei*: production of acetate in the mitochondrion is essential for parasite viability. *PLoS Negl. Trop. Dis.* 7, e2587.

McPherson, A., Gavira, J.A., 2014. Introduction to protein crystallization. *Acta Crystallogr. Sect. F, Struct. Biol. Commun.* 70, 2–20.

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

Mehlert, A., Ferguson, M.A.J., 2007. Structure of the glycosylphosphatidylinositol anchor of the *Trypanosoma brucei* transferrin receptor. *Mol. Biochem. Parasitol.* 151, 220–223.

Mehlert, A, Richardson, J.M., Ferguson, M.A., 1998. Structure of the glycosylphosphatidylinositol membrane anchor glycan of a class-2 variant surface glycoprotein from *Trypanosoma brucei*. *J. Mol. Biol.* 277, 379–392.

Mehlert, A., Wormald, M.R., Ferguson, M.A.J., 2012. Modeling of the N-

glycosylated transferrin receptor suggests how transferrin binding can occur within the surface coat of *Trypanosoma brucei*. *PLoS Pathog.* 8, e1002618.

Mehlert, Angela, Zitzmann, N., Richardson, J.M., Treumann, A., Ferguson, M.A.J., 1998. The glycosylation of the variant surface glycoproteins and procyclic acidic repetitive proteins of *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* 91, 145–152.

Mendonça-Previato, L., Penha, L., Garcez, T.C., Jones, C., Previato, J.O., 2013. Addition of α-O-GlcNAc to threonine residues define the post-translational modification of mucin-like molecules in *Trypanosoma cruzi*. *Glycoconj. J.* 30, 659–66.

Meyer, U., Benghezal, M., Imhof, I., Conzelmann, A., 2000. Active site determination of Gpi8p, a caspase-related enzyme required for glycosylphosphatidylinositol anchor addition to proteins. *Biochemistry* 39, 3461–3471.

Milne, K.G., Ferguson, M.A.J., Masterson, W.J., 1992. Inhibition of the GlcNAc transferase of the glycosylphosphatidylinositol anchor biosynthesis in African trypanosomes. *Eur. J. Biochem.* 208, 309–314.

Mitashi, P., Hasker, E., Lejon, V., Kande, V., Muyembe, J.-J., Lutumba, P., Boelaert, M., 2012. Human African trypanosomiasis diagnosis in first-line health services of endemic countries, a systematic review. *PLoS Negl. Trop. Dis.* 6, e1919.

Mittra, B., Ray, D.S., 2004. Presence of a poly(A) binding protein and two proteins with cell cycle-dependent phosphorylation in *Crithidia fasciculata* mRNA cycling sequence binding protein II. *Eukaryot. Cell* 3, 1185.

Miyagawa-Yamaguchi, A., Kotani, N., Honke, K., 2015. Each GPI-anchored protein species forms a specific lipid raft depending on its GPI attachment signal. *Glycoconj. J.* 32, 531–540.

Miyagawa-Yamaguchi, A., Kotani, N., Honke, K., 2014. Expressed glycosylphosphatidylinositol-anchored horseradish peroxidase identifies co-clustering molecules in individual lipid raft domains. *PLoS One* 9, e93054.

Monie, A., Hung, C.-F., Roden, R., Wu, T.-C., 2008. Cervarix: a vaccine for the prevention of HPV 16, 18-associated cervical cancer. *Biologics* 2, 97–105.

Montagna, G., Cremona, M.L., Paris, G., Amaya, M.F., Buschiazzo, A., Alzari, P.M., Frasch, A.C.C., 2002. The trans-sialidase from the african trypanosome *Trypanosoma brucei*. *Eur. J. Biochem.* 269, 2941–2950.

Moran, P., Caras, I.W., 1994. Requirements for glycosylphosphatidylinositol attachment are similar but not identical in mammalian cells and parasitic protozoa. *J. Cell Biol.* 125, 333–343.

Mori, A., Hara, S., Sugahara, T., Kojima, T., Iwasaki, Y., Kawarasaki, Y., Sahara, T., Ohgiya, S., Nakano, H., 2015. Signal peptide optimization tool for the secretion of recombinant protein from *Saccharomyces cerevisiae*. *J. Biosci. Bioeng.* 120, 518–525.

Morriswood, B., Schmidt, K., 2015. A MORN repeat protein facilitates protein entry into the flagellar pocket of *Trypanosoma brucei*. *Eukaryot. Cell* 14, 1081–1093.

Mugo, E., Clayton, C., 2017. Expression of the RNA-binding protein RBP10 promotes the bloodstream-form differentiation state in *Trypanosoma brucei*. *PLoS Pathog.* 13, e1006560.

Munday, J.C., Settimo, L., de Koning, H.P., 2015. Transport proteins determine drug sensitivity and resistance in a protozoan parasite, *Trypanosoma brucei*. *Front. Pharmacol.* 6, 32.

Mussmann, R., Engstler, M., Gerrits, H., Kieft, R., Toaldo, C.B., Onderwater, J., Koerten, H., van Luenen, H.G.A.M., Borst, P., 2004. Factors affecting the level and localization of the transferrin receptor in *Trypanosoma brucei*. *J. Biol. Chem.* 279, 40690–40698.

Mussmann, R., Janssen, H., Calafat, J., Engstler, M., Ansorge, I., Clayton, C., Borst, P., 2003. The expression level determines the surface distribution of the transferrin receptor in *Trypanosoma brucei*. *Mol. Microbiol.* 47, 23–35.

Nagamune, K., Ohishi, K., Ashida, H., Hong, Y., Hino, J., Kangawa, K., Inoue, N., Maeda, Y., Kinoshita, T., 2003. GPI transamidase of *Trypanosoma brucei* has two previously uncharacterized (trypanosomatid transamidase 1 and 2) and three common subunits. *Proc. Natl. Acad. Sci. U. S. A.* 100, 10682–10687.

Nagamune, K. *et al.*, 2000. Critical roles of glycosylphosphatidylinositol for *Trypanosoma brucei*. *Proc. Natl. Acad. Sci. U. S. A.* 97, 10336–10341.

Nakada, C. *et al.*, 2003. Accumulation of anchored proteins forms membrane

diffusion barriers during neuronal polarization. *Nat. Cell Biol.* 5, 626–632.

Nakanishi, M., Karasudani, M., Shiraishi, T., Hashida, K., Hino, M., Ferguson, M.A.J., Nomoto, H., 2014. TbGT8 is a bifunctional glycosyltransferase that elaborates N-linked glycans on a protein phosphatase AcP115 and a GPI-anchor modifying glycan in *Trypanosoma brucei*. *Parasitol. Int.* 63, 513–518.

Nishimura, A., Linder, M.E., 2013. Identification of a novel prenyl and palmitoyl modification at the CaaX motif of Cdc42 that regulates RhoGDI binding. *Mol. Cell. Biol.* 33, 1417–1429.

Nolan, D.P., Geuskens, M., Pays, E., 1999. N-linked glycans containing linear poly-N-acetyllactosamine as sorting signals in endocytosis in *Trypanosoma brucei*. *Curr. Biol.* 9, 1169–1172.

Nothaft, H., Szymanski, C.M., 2013. Bacterial protein N-glycosylation: new perspectives and applications. *J. Biol. Chem.* 288, 6912–6920.

Ohishi, K., Nagamune, K., Maeda, Y., Kinoshita, T., 2003. Two subunits of glycosylphosphatidylinositol transamidase, GPI8 and PIG-T, form a functionally important intermolecular disulfide bridge. *J. Biol. Chem.* 278, 13959–13967.

Overath, P., Engstler, M., 2004. Endocytosis, membrane recycling and sorting of GPI-anchored proteins: *Trypanosoma brucei* as a model system. *Mol. Microbiol.* 53, 735–744.

Pal, A., Hall, B.S., Nesbeth, D.N., Field, H.I., Field, M.C., 2002. Differential endocytic functions of *Trypanosoma brucei* Rab5 isoforms reveal a glycosylphosphatidylinositol-specific endosomal pathway. *J. Biol. Chem.* 277, 9529–9539.

Paladino, S., Lebreton, S., Tivodar, S., Campana, V., Tempre, R., Zurzolo, C., 2008. Different GPI-attachment signals affect the oligomerisation of GPI-anchored proteins and their apical sorting. *J. Cell Sci.* 121, 4001–4007.

Papadopoulou, B., Dumas, C., 1997. Parameters controlling the rate of gene targeting frequency in the protozoan parasite *Leishmania*. *Nucleic Acids Res.* 25, 4278–4286.

Parodi, A.J., 1993. N -Glycosylation in trypanosomatid protozoa. *Glycobiology* 3, 193–199.

Parodit, A.J., Quesada-Allue, L.A., 1982. Protein glycosylation in Trypanosoma

cruzi: Characterisation of dolichol-bound monosaccharides and oligosaccharides synthesised “in vivo.” *J. Biol. Chem.* 257, 7637–7640.

Pechmann, S., Chartron, J.W., Frydman, J., 2014. Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP in vivo. *Nat. Struct. Mol. Biol.* 21, 1100–1105.

Pechmann, S., Frydman, J., 2013. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat. Struct. Mol. Biol.* 20, 237–243.

Pédelacq, J.-D., Cabantous, S., Tran, T., Terwilliger, T.C., Waldo, G.S., 2006. Engineering and characterization of a superfolder green fluorescent protein. *Nat. Biotechnol.* 24, 79–88.

Pérez-Morga, D. *et al.*, 2005. Apolipoprotein L-I promotes trypanosome lysis by forming pores in lysosomal membranes. *Science* 309, 469–472.

Phillips, G.J., Silhavy, T.J., 1992. The *E. coli* ffh gene is necessary for viability and efficient protein export. *Nature* 359, 744–746.

Pierleoni, A. *et al.*, 2008. PredGPI: a GPI-anchor predictor. *BMC Bioinformatics* 9, 392.

Pinger, J. *et al.*, 2018. African trypanosomes evade immune clearance by O-glycosylation of the VSG surface coat. *Nat. Microbiol.* 3, 932–938.

Pion, C. *et al.*, 2014. Characterization and immunogenicity in mice of recombinant influenza haemagglutinins produced in *Leishmania tarentolae*. *Vaccine* 32, 5570–5576.

Poon, S.K., Peacock, L., Gibson, W., Gull, K., Kelly, S., 2012. A modular and optimized single marker system for generating *Trypanosoma brucei* cell lines expressing T7 RNA polymerase and the tetracycline repressor. *Open Biol.* 2, 110037.

Porter, J.A., Young, K.E., Beachy, P.A., 1996. Cholesterol modification of hedgehog signaling proteins in animal development. *Science* 274, 255–259.

Puig, B. *et al.*, 2019. GPI-anchor signal sequence influences PrPC sorting, shedding and signalling, and impacts on different pathomechanistic aspects of prion disease in mice. *PLoS Pathog.* 15, e1007520.

Quax, T.E.F., Claassens, N.J., Söll, D., van der Oost, J., 2015. Codon bias as a means to fine-tune gene expression. *Mol. Cell* 59, 149–161.

Quesada-Allue, L.A., Parodi, A.J., 1983. Novel mannose carrier in the trypanosomatid Crithidia fasciculata behaving as a short alpha-saturated polyprenyl phosphate. *Biochem. J.* 212, 123–128.

Radwanska, M., Magez, S., Dumont, N., Pays, A., Nolan, D., Pays, E., 2000. Antibodies raised against the flagellar pocket fraction of *Trypanosoma brucei* preferentially recognize HSP60 in cDNA expression library. *Parasite Immunol.* 22, 639–650.

Raymond, F. *et al.*, 2012. Genome sequencing of the lizard parasite *Leishmania tarentolae* reveals loss of genes associated to the intracellular stage of human pathogenic species. *Nucleic Acids Res.* 40, 1131–1147.

Requena, J.M., Soto, M., Quijada, L., Carrillo, G., Alonso, C., 1997. A region containing repeated elements is associated with transcriptional termination of *Leishmania infantum* ribosomal RNA genes. *Mol. Biochem. Parasitol.* 84, 101–110.

Resh, M.D., 2016. Fatty acylation of proteins: the long and the short of it. *Prog. Lipid Res.* 63, 120.

Resh, M.D., 2013. Covalent lipid modifications of proteins. *Curr. Biol.* 23, R431-435.

Rojas, F. *et al.*, 2019. Oligopeptide signaling through TbGPR89 drives trypanosome quorum sensing. *Cell* 176, 306-317.e16.

Rooney, B., Piening, T., Büscher, P., Rogé, S., Smales, C.M., 2015. Expression of *Trypanosoma brucei gambiense* antigens in *Leishmania tarentolae*. Potential for Use in rapid serodiagnostic tests (RDTs). *PLoS Negl. Trop. Dis.* 9, e0004271.

Rosano, G.L., Ceccarelli, E.A., 2014. Recombinant protein expression in *Escherichia coli*: advances and challenges. *Front. Microbiol.* 5, 172.

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.

Salmon, D., Geuskens, M., Hanocq, F., Hanocq-Quertier, J., Nolan, D., Ruben, L., Pays, E., 1994. A novel heterodimeric transferrin receptor encoded by a pair of VSG expression site-associated genes in *T. brucei*. *Cell* 78, 75–

86.

Sanchez-Garcia, L., Martín, L., Mangues, R., Ferrer-Miralles, N., Vázquez, E., Villaverde, A., 2016. Recombinant pharmaceuticals from microbial cells: a 2015 update. *Microb. Cell Fact.* 15, 33.

Sanyal, S., Frank, C.G., Menon, A.K., 2008. Distinct flippases translocate glycerophospholipids and oligosaccharide diphosphate dolichols across the endoplasmic reticulum. *Biochemistry* 47, 7937–7946.

Scanlan, C.N. *et al.*, 2002. The broadly neutralizing anti-human immunodeficiency virus type 1 antibody 2G12 recognizes a cluster of α1->2 mannose residues on the outer face of gp120. *J. Virol.* 76, 7306–7321.

Schnare, M.N., Collings, J.C., Spencer, D.F., Gray, M.W., 2000. The 28S-18S rDNA intergenic spacer from *Crithidia fasciculata*: repeated sequences, length heterogeneity, putative processing sites and potential interactions between U3 small nucleolar RNA and the ribosomal RNA precursor. *Nucleic Acids Res.* 28, 3452–3461.

Schneider, C.A., Rasband, W.S., Eliceiri, K.W., 2012. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* 9, 671–675.

Schneider, P., Treumann, A., Milne, K.G., Mcconville, M.J., Zitzmann, N., Ferguson, M.A.J., 1996. Structural studies on a lipoarabinogalactan of *Crithidia fasciculata*, Biochem. J.

Schwartz, K.J., Peck, R.F., Tazeh, N.N., Bangs, J.D., 2005. GPI valence and the fate of secretory membrane proteins in African trypanosomes. *J. Cell Sci.* 118, 5499–5511.

Schwede, A., Macleod, O.J.S., MacGregor, P., Carrington, M., 2015. How does the VSG coat of bloodstream form African trypanosomes interact with external proteins? *PLOS Pathog.* 11, e1005259.

Sevova, E.S., Bangs, J.D., 2009. Streamlined architecture and glycosylphosphatidylinositol-dependent trafficking in the early secretory pathway of African trypanosomes. *Mol. Biol. Cell* 20, 4739–4750.

Shao, S., Hegde, R.S., 2011. A Calmodulin-dependent translocation pathway for small secretory proteins. *Cell* 147, 1576–1588.

Sharma, D.K., Smith, T.K., Weller, C.T., Crossman, A., Brimacombe, J.S., Ferguson, M.A., 1999. Differences between the trypanosomal and human GlcNAc-PI de-N-acetylases of glycosylphosphatidylinositol membrane

anchor biosynthesis. *Glycobiology* 9, 415–422.

Sharp, P.M., Li, W.H., 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.* 24, 28–38.

Shcherbakova, A., Preller, M., Taft, M.H., Pujols, J., Ventura, S., Tiemann, B., Buettner, F.F.R., Bakker, H., 2019. C-mannosylation supports folding and enhances stability of thrombospondin repeats. *Elife* 8.

Shimogawa, M.M., Saada, E.A., Vashisht, A.A., Barshop, W.D., Wohlschlegel, J.A., Hill, K.L., 2015. Cell surface proteomics provides insight into stage-specific remodeling of the host-parasite interface in Trypanosoma brucei. *Mol. Cell. Proteomics* 14, 1977–1988.

Shishioh, N., Hong, Y., Ohishi, K., Ashida, H., Maeda, Y., Kinoshita, T., 2005. GPI7 is the second partner of PIG-F and involved in modification of glycosylphosphatidylinositol. *J. Biol. Chem.* 280, 9728–9734.

Shusta, E. V, Raines', R.T., Pliickthun, A., Wittrup, K.D., 1998. Increasing the secretory capacity of *Saccharomyces cerevisiae* for production of single-chain antibody fragments. *Nat. Biotechnol.* 16, 773–777.

Siegel, T.N., Tan, K.S.W., Cross, G.A.M., 2005. Systematic study of sequence motifs for RNA trans splicing in *Trypanosoma brucei*. *Mol. Cell. Biol.* 25, 9586–9594.

Siegel, V., Walter, P., 1988. Each of the activities of signal recognition particle (SRP) is contained within a distinct domain: analysis of biochemical mutants of SRP. *Cell* 52, 39–49.

Smith, T.K., Crossman, A., Borissow, C.N., Paterson, M.J., Dix, A., Brimacombe, J.S., Ferguson, M.A., 2001. Specificity of GlcNAc-PI de-N-acetylase of GPI biosynthesis and synthesis of parasite-specific suicide substrate inhibitors. *EMBO J.* 20, 3322–3332.

Smith, T.K., Bütikofer, P., 2010. Lipid metabolism in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* 172, 66–79.

Smith, T.K., Crossman, A., Brimacombe, J.S., Ferguson, M.A.J., 2004. Chemical validation of GPI biosynthesis as a drug target against African sleeping sickness. *EMBO J.* 23, 4701–4708.

Smith, T.K., Paterson, M.J., Crossman, A., Brimacombe, J.S., Ferguson, M.A.J., 2000. Parasite-specific inhibition of the glycosylphosphatidylinositol biosynthetic pathway by stereoisomeric substrate analogues. *Biochemistry*

39, 11801–11807.

Smith, T.K., Sharma, D.K., Crossman, A., Brimacombe, J.S., Ferguson, M.A., 1999. Selective inhibitors of the glycosylphosphatidylinositol biosynthetic pathway of *Trypanosoma brucei*. *EMBO J.* 18, 5922–5930.

Sonnhammer, E.L., von Heijne, G., Krogh, A., 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proceedings. Int. Conf. Intell. Syst. Mol. Biol.* 6, 175–182.

Sou, Y., Tanida, I., Komatsu, M., Ueno, T., Kominami, E., 2006. Phosphatidylserine in addition to phosphatidylethanolamine Is an in vitro target of the mammalian Atg8 modifiers, LC3, GABARAP, and GATE-16. *J. Biol. Chem.* 281, 3017–3024.

Stahl, N., Baldwin, M., Hecker, R., Pan, K.-M., Burlingame, A., Prusiner, S., 1992. Glycosylinositol phospholipid anchors of the scrapie and cellular prion proteins contain sialic acid. *Biochemistry* 31, 5043–5053.

Stanley, P., Taniguchi, N., Aebi, M., 2017. N-Glycans, in: Varki, A. et al. (Eds.), *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press.

Steen, P. Van den, Rudd, P.M., Dwek, R.A., Opdenakker, G., 1998. Concepts and principles of O-Linked glycosylation. *Crit. Rev. Biochem. Mol. Biol.* 33, 151–208.

Steentoft, C. *et al.*, 2013. Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. *EMBO J.* 32, 1478–1488.

Stetsenko, A., Guskov, A., Stetsenko, A., Guskov, A., 2017. An overview of the top ten detergents used for membrane protein crystallization. *Crystals* 7, 197.

Stødkilde, K., Torvund-Jensen, M., Moestrup, S.K., Andersen, C.B.F., 2014. Structural basis for trypanosomal haem acquisition and susceptibility to the host innate immune system. *Nat. Commun.* 5, 5487.

Stokes, M.J., Murakami, Y., Maeda, Y., Kinoshita, T., Morita, Y.S., 2014. New insights into the functions of PIGF, a protein involved in the ethanolamine phosphate transfer steps of glycosylphosphatidylinositol biosynthesis. *Biochem. J.* 463, 249–256.

Sullivan, L., Wall, S.J., Carrington, M., Ferguson, M.A.J., 2013. Proteomic selection of immunodiagnostic antigens for human African trypanosomiasis and generation of a prototype lateral flow immunodiagnostic device. *PLoS*

Negl. Trop. Dis. 7, e2087.

Sunter, J., Webb, H., Carrington, M., 2013. Determinants of GPI-PLC localisation to the flagellum and access to GPI-anchored substrates in trypanosomes. *PLoS Pathog.* 9, e1003566.

Sunter, J.D., 2016. A vanillic acid inducible expression system for *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* 207, 45–48.

Takada, R., Satomi, Y., Kurata, T., Ueno, N., Norioka, S., Kondoh, H., Takao, T., Takada, S., 2006. Monounsaturated fatty acid modification of Wnt protein: its role in Wnt secretion. *Dev. Cell* 11, 791–801.

Takeuchi, H., Kantharia, J., Sethi, M.K., Bakker, H., Haltiwanger, R.S., 2012. Site-specific O-glucosylation of the epidermal growth factor-like (EGF) repeats of notch: efficiency of glycosylation is affected by proper folding and amino acid sequence of individual EGF repeats. *J. Biol. Chem.* 287, 33934–33944.

Takizawa, P.A., DeRisi, J.L., Wilhelm, J.E., Vale, R.D., 2000. Plasma Membrane Compartmentalization in Yeast by Messenger RNA Transport and a Septin Diffusion Barrier. *Science (80-.).* 290, 341–344.

Tetaud, E., Lecuix, I., Sheldrake, T., Fairlamb, A.H., 2002. A new expression vector for *Crithidia fasciculata* and *Leishmania* 120, 195–204.

Tiengwe, C. *et al.*, 2017. Controlling transferrin receptor trafficking with GPI-valence in bloodstream stage African trypanosomes. *PLOS Pathog.* 13, e1006366.

Tiengwe, C., Muratore, K.A., Bangs, J.D., 2016. Surface proteins, ERAD and antigenic variation in *Trypanosoma brucei*. *Cell. Microbiol.* 18, 1673–1688.

Triggs, V.P., Bangs, J.D., 2003. Glycosylphosphatidylinositol-dependent protein trafficking in bloodstream stage *Trypanosoma brucei*. *Eukaryot. Cell* 2, 76–83.

Ubeda, J.-M. *et al.*, 2014. Genome-wide stochastic adaptive DNA amplification at direct and inverted DNA repeats in the parasite *Leishmania*. *PLoS Biol.* 12, e1001868.

Umaer, K., Bush, P.J., Bangs, J.D., 2018. Rab11 mediates selective recycling and endocytic trafficking in *Trypanosoma brucei*. *Traffic* 19, 406–420.

Urbaniak, M.D., Capes, A.S., Crossman, A., O'Neill, S., Thompson, S., Gilbert, I.H., Ferguson, M.A.J., 2014. Fragment screening reveals salicylic

hydroxamic acid as an inhibitor of *Trypanosoma brucei* GPI GlcNAc-PI de-N-acetylase. *Carbohydr. Res.* 387, 54–58.

Urbaniak, M.D., Crossman, A., Chang, T., Smith, T.K., van Aalten, D.M.F., Ferguson, M.A.J., 2005. The N-acetyl-D-glucosaminylphosphatidylinositol De-N-acetylase of glycosylphosphatidylinositol biosynthesis is a zinc metalloenzyme. *J. Biol. Chem.* 280, 22831–22838.

Uzureau, P. *et al.*, 2013. Mechanism of *Trypanosoma brucei gambiense* resistance to human serum. *Nature* 501, 430–434.

Vainauskas, S., Menon, A.K., 2003. A conserved proline in the last transmembrane segment of Gaa1 is required for glycosylphosphatidylinositol (GPI) recognition by GPI transamidase. *J. Biol. Chem.* 279, 6540–6545.

Valente, M., Castillo-Acosta, V.M., Vidal, A.E., González-Pacanowska, D., 2019. Overview of the role of kinetoplastid surface carbohydrates in infection and host cell invasion: Prospects for therapeutic intervention. *Parasitology*.

Vanhollebeke, B., De Muylder, G., Nielsen, M.J., Pays, A., Tebabi, P., Dieu, M., Raes, M., Moestrup, S.K., Pays, E., 2008. A haptoglobin-hemoglobin receptor conveys innate immunity to *Trypanosoma brucei* in humans. *Science* 320, 677–681.

Varki, A., 2017. Biological roles of glycans. *Glycobiology* 27, 3–49.

Varma, Y., Hendrickson, T., 2010. Methods to study GPI anchoring of proteins. *Chembiochem* 11, 623–636.

Vickerman, K., 1969. On The Surface Coat and Flagellar Adhesion in Trypanosomes. *J. Cell Sci.* 5, 163–193.

Vieira Gomes, A. *et al.*, 2018. Comparison of yeasts as hosts for recombinant protein production. *Microorganisms* 6, 38.

Vigueira, P.A., Paul, K.S., 2011. Requirement for acetyl-CoA carboxylase in *Trypanosoma brucei* is dependent upon the growth environment. *Mol. Microbiol.* 80, 117–132.

Vita, R., Mahajan, S., Overton, J.A., Dhanda, S.K., Martini, S., Cantrell, J.R., Wheeler, D.K., Sette, A., Peters, B., 2019. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* 47, D339–D343.

von Heijne, G., 1985. Signal sequences. *J. Mol. Biol.* 184, 99–105.

von Heijne, G., 1983. Patterns of amino acids near signal-sequence cleavage sites. *Eur. J. Biochem.* 133, 17–21.

Wang, J., Böhme, U., Cross, G.A.., 2003. Structural features affecting variant surface glycoprotein expression in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* 128, 135–145.

Wang, X., Gu, J., Ihara, H., Miyoshi, E., Honke, K., Taniguchi, N., 2006. Core fucosylation regulates epidermal growth factor receptor-mediated intracellular signaling. *J. Biol. Chem.* 281, 2572–2577.

Wang, X. *et al.*, 2005. Dysregulation of TGF-beta1 receptor activation leads to abnormal lung development and emphysema-like phenotype in core fucose-deficient mice. *Proc. Natl. Acad. Sci. U. S. A.* 102, 15791–15796.

Webb, H., Carnall, N., Vanhamme, L., Rolin, S., Van Den Abbeele, J., Welburn, S., Pays, E., Carrington, M., 1997. The GPI-phospholipase C of *Trypanosoma brucei* is nonessential but influences parasitemia in mice. *J. Cell Biol.* 139, 103–114.

Weinhold, B. *et al.*, 2012. Deficits in sialylation impair podocyte maturation. *J. Am. Soc. Nephrol.* 23, 1319–1328.

Weitzmann, A., Volkmer, J., Zimmermann, R., 2006. The nucleotide exchange factor activity of Grp170 may explain the non-lethal phenotype of loss of Sil1 function in man and mouse. *FEBS Lett.* 580, 5237–5240.

Whipple, S., 2017. Dissection of flagellar pocket function in *Trypanosoma brucei*. PhD thesis. University of Nottingham.

WHO, 2012. Accelerating work to overcome the global impact of neglected tropical diseases – a roadmap for implementation. Geneva, Switzerland.

Wiggins, C.A., Munro, S., 1998. Activity of the yeast MNN1 alpha-1,3-mannosyltransferase requires a motif conserved in many other families of glycosyltransferases. *Proc. Natl. Acad. Sci. U. S. A.* 95, 7945–7950.

Wingfield, P.T., 2001. Protein precipitation using ammonium sulfate. *Curr. Protoc. Protein Sci.* Appendix 3.

Wirtz, E., Clayton, C., 1995. Inducible gene expression in Trypanosomes mediated by a prokaryotic repressor. *Science (80-.).* 268, 1179–1183.

Wirtz, E., Hoek, M., Cross, G.A.M., 1998. Regulated processive transcription of chromatin by T7 RNA polymerase in *Trypanosoma brucei*. *Nucleic Acids Res.* 26, 4626–4634.

Wirtz, E., Leal, S., Ochatt, C., Cross, G.A., 1999. A tightly regulated inducible expression system for conditional gene knock-outs and dominant-negative genetics in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* 99, 89–101.

Wu, B., Guo, W., 2015. The Exocyst at a Glance. *J. Cell Sci.* 128, 2957–2964.

Wu, X., Wu, D., Lu, Z., Chen, W., Hu, X., Ding, Y., 2009. A novel method for high-level production of TEV protease by superfolder GFP tag. *J. Biomed. Biotechnol.* 2009, 591923.

Xiao, X. *et al.*, 2017. Cholesterol modification of smoothened Is required for Hedgehog signaling. *Mol. Cell* 66, 154-162.e10.

Yamamoto, F., Clausen, H., White, T., Marken, J., Hakomori, S., 1990. Molecular genetic basis of the histo-blood group ABO system. *Nature* 345, 229–233.

Yao, C., 2010. Major surface protease of trypanosomatids: one size fits all? *Infect. Immun.* 78, 22–31.

Zhou, Q. *et al.*, 2010. A comparative proteomic analysis reveals a new bi-lobe protein required for bi-lobe duplication and cell division in *Trypanosoma brucei*. *PLoS One* 5, e9660.

Zhou, Q., Qiu, H., 2019. The mechanistic impact of N-glycosylation on stability, pharmacokinetics, and immunogenicity of therapeutic proteins. *J. Pharm. Sci.* 108, 1366–1377.

Zhu, Y., Fraering, P., Vionnet, C., Conzelmann, A., 2005. Gpi17p does not stably interact with other subunits of glycosylphosphatidylinositol transamidase in *Saccharomyces cerevisiae*. *Biochim. Biophys. Acta* 1735, 79–88.

Ziegelbauer, K., Overath, P., 1993. Organization of two invariant surface glycoproteins in the surface coat of *Trypanosoma brucei*. *Infect. Immun.* 61, 4540–4545.

Ziegelbauer, K., Overath, P., 1992. Identification of invariant surface glycoproteins in the bloodstream stage of *Trypanosoma brucei*. *J. Biol. Chem.* 267, 10791–10796.

Zimmermann, R., Eyrisch, S., Ahmad, M., Helms, V., 2011. Protein translocation across the ER membrane. *Biochim. Biophys. Acta - Biomembr.* 1808, 912–924.

Zitzmann, N., Mehlert, A., Carrouée, S., Rudd, P.M., Ferguson, M.A., 2000.

Protein structure controls the processing of the N-linked oligosaccharides and glycosylphosphatidylinositol glycans of variant surface glycoproteins expressed in bloodstream form *Trypanosoma brucei*. *Glycobiology* 10, 243–249.

Zoll, S., Lane-Serff, H., Mehmood, S., Schneider, J., Robinson, C. V., Carrington, M., Higgins, M.K., 2018. The structure of serum resistance-associated protein and its implications for human African trypanosomiasis. *Nat. Microbiol.* 3, 295–301.

VI. Appendix

Appendix 1. pCExC DNA sequence

```
   1 GAATTCTAAT ACGACTCACT ATAGGGATAT CAAGCTTAGC AAAACGAAGA CATGCGGCGT
  61 GGATCACCAG TACTCTCTCT TTCTCTTTCT TTTTCTCACT GCCTCCGCTC TCTCTCTCTC
 121 CCTCTTTTTC TCACCTCTTC CTCTCTCTAC CCACACCAAC GCGCACTGCC CCCCATGCTG
 181 ACGCCGCAGA ATCACGTTCA CGTGGTCTAT ACCAAACAAA ACAACACCCC ACCAACAGCG
 241 TACCGCTTTA CAACCATAGA TCTACCATGC TCGAGGGCGC TAGCGAGAAC CTGTACTTCC
 301 AGGGCACTAG TCGCAAGGGC GAGGAGCTGT TCACGGGCGT GGTGCCGATC CTGGTGGAGC
 361 TGGACGGCGA CGTGAACGGC CACAAGTTCA GCGTGCGCGG CGAGGGCGAG GGCGACGCGA
 421 CGAACGGCAA GCTGACGCTG AAGTTCATCT GCACGACGGG CAAGCTGCCG GTGCCGTGGC
 481 CGACGCTGGT GACGACGCTG ACGTACGGCG TGCAGTGCTT CGCGCGCTAC CCGGACCACA
 541 TGAAGCAGCA CGACTTCTTC AAGAGCGCGA TGCCGGAGGG CTACGTGCAG GAGCGCACGA
 601 TCAGCTTCAA GGACGACGGC ACGTACAAGA CGCGCGCGGA GGTGAAGTTC GAGGGCGACA
 661 CGCTGGTGAA CCGCATCGAG CTGAAGGGCA TCGACTTCAA GGAGGACGGC AACATCCTGG
 721 GCCACAAGCT GGAGTACAAC TTCAACAGCC ACAACGTGTA CATCACGGCG GACAAGCAGA
 781 AGAACGGCAT CAAGGCGAAC TTCAAGATCC GCCACAACGT GGAGGACGGC AGCGTGCAGC
 841 TGGCGGACCA CTACCAGCAG AACACGCCGA TCGGCGACGG CCCGGTGCTG CTGCCGGACA
 901 ACCACTACCT GAGCACGCAG AGCGTGCTGA GCAAGGACCC GAACGAGAAG CGCGACCACA
 961 TGGTGCTGCT GGAGTTCGTG ACGGCGGCGG GCATCACGCA CGGCATGGAC GAGCTGTACA
1021 AGGGATCCCA CCACCATCAC CACCACTAAG CGGCCGTCAT ATGGCGATGG TTCGACAGGT
1081 CCGCAGCATT TTTCTCTCTT TCCCTTTTTC CGCACCGAAT GAAACGCAGG GAGGCGCTGG
1141 GGAAAGGAGA GAGGAGATGG CGCACGTCTT CGCCACGACT TGTGCCGCGT TGAAGCCAAT
1201 TCGTTTGTTT CTTTAGCTTT CCCTTTTCTT TTTTCTCATT TCCATCTCAA ACAAAACAAA
1261 CAAAAAAGCA AACATCTATT GTAGAAGCGA GAGGGCAGTT GACGCTGTTT GTCTTCGCAA
1321 AAAAGAATAG ATTGGGACAG GGGCATTGCA GACGGCGCCG CAGCAACCTG GCTGTGACGA
1381 CAGGTTGTGG TTGTTGTTGT TGTCCTTGAC GCGTGTGCGT TGGTGTGTGT GTGTGCCCGC
1441 GCTTCCTCTT TCTCTGGGTT GTGGATCGTT CCCTTCCTTT TCGTGACTAC GTCCGCCAGG
1501 TTAGCGGCAG GCGATGAGAT GGAGAAGAGC CTTTTTGGCA TTTCTCATTC TCCGTCGCTG
1561 TTGCTGGTGT TGTCATGCGG TAAAAGACAA TTATCAAGAT GGTCGTCTCC AGCCCTCCTT
1621 TTTCAGTCAA CCGGCTGGTG TCGTTGCCGT CAAAAATATA TGCTGCCACT TCGTGTAGAG
1681 GCGAAATTCT GTGACAGTGC CTTTTACCAC ATTTCTCTCC CCTTCTCTCT CTCTCCCCTC
1741 GCCGTTTTTC TCCACAAAAA AGAAAGTCAT CAGCGAAGGG CCCACACTGT GTAGTGCATC
1801 GACGCACGCT ACGGCTCATT TTCATTGCGG TGTCACACCG CCGTCTCCCA TTCTTCCTCT
1861 CGCACAATGC ATGGCCCGTT TGTTATCTAT GCAGTATTCT GCAGGCTGCT AACAAAGCCC
1921 GAAAGGAAGC TGAGTTGGCT GCTGCCACCG CTGAGCAATA ACTAGCATAA CCCCTTGGGG
1981 CCTCTAAACG GGTCTTGAGG GGTTTTTTGC TGAAAGGAGA ACTATATCCG GATAGGGCTG
2041 CTAACAAAGC CCGAAAGGAA GCTGAGTTGG CTGCTGCCAC CGCTGAGCAA TAACTAGCAT
2101 AACCCCTTGG GGCCTCTAAA CGGGTCTTGA GGGGTTTTTT GCTGAAAGGA GAACTATATC
2161 CGGATTGGGC CTCTAGACAC AGCAGAGAGA GAAGCCCCCA CCCAGCCCTC ACACAGGCGT
2221 GTCGTGAATC ATCTCCCTTT CTCGCCTTTC ATGTGTGGCT TCAAGGAGCC CAGGCGCGCC
2281 AAAAGAGGAG TCCGGGTGTG AGAGGAGCTG CCTGGCTACC AGGATGAGTG CCCGTGCCTG
2341 CGCCGCTCTG TGGCCGCTGC ACACACTCAG AGGTGGCGGT GTGCGTGCTG TCACTCTTTA
2401 CTACCCCCGT TTCGGTTCGC GGGCGGGTGT CGAGGAGAGC GCTGCCCCTC TCCTTCCCTC
2461 TCCCCATCCG CAAGAAAGCA AGTGGCCAGT GCGATACACG AGGCCAGCCA ACTCTACTTA
2521 GTGTTCCTGC AGGAAGCGGA TGGCGAGCAC GTGGCGCGCG GCCGCCCCCT GCGAGGAGCA
2581 GCTGCAGCAG CATCCCCCTA TAACCCAGAG AAGAAGAGCA AAGGTGAGCG CCACGCTGCG
2641 CCGCGCCAAT ACGCCCTACA GAGAGAAGAT AAATAAAAAG GAACACAGCT GCGAAAGACT
2701 ACACGACTTA TCAAGAAGGC CGTTCGTGGG AGCGAGATGG CAAAGAGGTG TCGCTTTCCC
2761 CCGTGATGGA ATCCGCGAAG GCGGGCAGGG AGAAAGGCGA GGACAGCTCC CCCATGGACC
2821 GGCCCGGACT TCCCCACTCC CCCATGCACC ACACCACACA CCACCCCGCC CACCGAGGCT
2881 TAACTTTGCA AAGGAGACGA GACAAAGTGT TTTTCGGAGC AAAAGGACAA CGGGGAAAGT
2941 AAACAGGTAC CAATCCGGAT ATAGTTCTCC TTTCAGCAAA AAACCCCTCA AGACCCGTTT
3001 AGAGGCCCCA AGGGGTTATG CTAGTTATTG CTCAGCGGTG GCAGCAGCCA ACTCAGCTTC
3061 CTTTCGGGCT TTGTTAGCAG CCCTATCCGG ATATAGTTCT CCTTTCAGCA AAAAACCCCT
3121 CAAGACCCGT TTAGAGGCCC CAAGGGGTTA TGCTAGTTAT TGCTCAGCGG TGGCAGCAGC
3181 CAACTCAGCT TCCTTTCGGG CTTTGTTAGC AGCCTGCAGA ATACTGCATA GATAACAAAC
3241 GGGCCCGGAC GAACCGGAGC TCGCAGAACT GTGGCACGCA TACATCCCCA AAGCAACACG
3301 CTCCCGTTTT TCGCTTCTTC TTTGTTCTTT GTATAAAGGC ACGCATCAGC GCCAAAGCGT
```

```
3361 GTTTGAGAGC GTACAGCGGC CACGACGTGG ATCAGGAAGA AAGGAGCCGT GGGTGCGTTG
3421 CTGACCTCTT ATTCTTGTCT TTTTTGTATT TTTGTTTTTG TTGATTAGTT TTGAAGGCAC
3481 AGATCGTGTA TGCGTGCTTT TAAAATCTGT GTATGTGCGC AGCGGCACGT GAGTAAAGAA
3541 GCGCACTCGG ACACACCCAC CGACCTGCGT GACGGACCCG AAGATACCGG GAAAAGAAAA
3601 GATGTTTGGA GTCAAATAAA TAAAAAAGCC CACTCCAACA CTCAAATCAC CCTTTTCGCC
3661 TCGAATTGAT TAAAATGATG CGGAAGGCGT CACCACTGGA AAACAAACAC GCACGATATG
3721 GAGTAGCGGA CAAATGGCAT CAGCACAGAA ACAAATATGT ATATAGTAAA ACGTACGGCG
3781 GTCGCAGGTG ACACCCACTG CGACGCTGCG GGGCAAAAAG AGCAGGCGCA ATCATGTTAG
3841 AAAGAAAAAA CACAGAAAAA GGGGAGCGAG GGAAGGAGGA GGGAAGCCGA GAGGACACAC
3901 CGCCGTGCAC ACCGGTCAAC AGGATGAGAA GCAGCGAGAA AGACATCTTC CTCGCTTTGG
3961 TATTTACTTC CCCTCTTCCC TCTTCTCCTG AATGTCTCTG GCGCTCTCAC TGCACACGTG
4021 GGAGGTGGTC AGCACAGTGC CGCAAAATAG AAGCGCCTGA GGCATCGAGG GAGTACTTAC
4081 GGAGAGGAAA GGGGGACGGA AGTGCGACTC ATCACCGCTG CCGTAGGCAG GAGACCCAAA
4141 TCGGCTAAGA AAGAAGAAGA AAAATAACGG TGATAACAAG ACAAAACGGA AGGAAGAAAA
4201 ACGGAGAGTA AAAAGAGGAA GAGGATAGCA CCGACATTAG AGCAAACAAA AAGAGGAGTT
4261 GGCGACGTGG CGGGAACAGT AAAACAGAGA AGGGCAGAGA GGGGCGCTCC GCCACCACCC
4321 ATCTCTGCCT CTTTTTCTCG ATACTCCGAC CTATGCGTCC GCCTTATACT TCCCTCTTTC
4381 TCCGCCTGCT GGATCCCTAT TCCTTTGCCC TCGGACGAGT GCTGAGGCGT CGGTTTCCAC
4441 TATCGGCGAG TACTTCTACA CAGCCATCGG TCCAGACGGC CGCGCTTCTG CGGGCGATTT
4501 GTGTACGCCC GACAGTCCCG GCTCCGGATC GGACGATTGC GTCGCATCGA CCCTGCGCCC
4561 AAGCTGCATC ATCGAAATTG CCGTCAACCA AGCTCTGATA GAGTTGGTCA AGACCAATGC
4621 GGAGCATATA CGCCCGGAGC CGCGGCGATC CTGCAAGCTC CGGATGCCTC CGCTCGAAGT
4681 AGCGCGTCTG CTGCTCCATA CAAGCCAACC ACGGCCTCCA GAAGAAGATG TTGGCGACCT
4741 CGTATTGGGA ATCCCCGAAC ATCGCCTCGC TCCAGTCAAT GACCGCTGTT ATGCGGCCAT
4801 TGTCCGTCAG GACATTGTTG GAGCCGAAAT CCGCGTGCAC GAGGTGCCGG ACTTCGGGGC
4861 AGTCCTCGGC CCAAAGCATC AGCTCATCGA GAGCCTGCGC GACGGACGCA CTGACGGTGT
4921 CGTCCATCAC AGTTTGCCAG TGATACACAT GGGGATCAGC AATCGCGCAA ATGAAATCAC
4981 GCCATGTAGT GTATTGACCG ATTCCTTGCG GTCCGAATGG GCCGAACCCG CTCGTCTGGC
5041 TAAGATCGGC CGCAGCGATC GCATCCATGG CCTCCGCGAC CGGCTGAAGA ACAGCGGGCA
5101 GTTCGGTTTC AGGCAGGTCT TGCAACGTGA CACCCTGTGC ACGGCGGGAG ATGCAATAGG
5161 TCAGGCTCTC GCTGAACTCC CCAATGTCAA GCACTTCCGG AATCGGGAGC GCGGCCGATG
5221 CAAAGTGCCG ATAAACATAA CGATCTTTGT AGAAACCATC GGCGCAGCTA TTTACCCGCA
5281 GGACATATCC ACGCCCTCCT ACATCGAAGC TGAAAGCACG AGATTCTTCG CCCTCCGAGA
5341 GCTGCATCAG GTCGGAGACG CTGTCGAACT TTCCGATCAG AAACTTCTCG ACAGACGTCG
5401 CGGTGAGTTC AGGCTTTTTC ATGTCGACGC TTGACAAGTG GAAGATAGTT GGACAAGTAG
5461 AGGTGTAAGG TCGAAGTGGA CAGAGCGGAA TGTGACGTTG ACTGATGAGT AATACTCGCG
5521 TTTAATATGA GTGTGTTATC GTGAAAGAAA AGACAGCTGT TTGCTCGAAA AGGACACGCG
5581 CCATGTACAC GCAGGACACA ACAACAACCG GAGGAAGAAG GCAGAGAAGT GGTGAGTGGA
5641 TTAGTCAAAG AGCGCAGCCT GAGGCGATAG CTGGCAAAGG CGGTGATAAT TAGACAAGGC
5701 AACGCGACGG TTTCAGTGAT ATTTACAGTG CTTCAGTCAC GAAGCAGTGG CGGTCTTACT
5761 GGGGAAGGTA TCCAGAAGAG CCTACAAAAC CGATATGCTA CCGACGTGTT GGATTGACGG
5821 TTTAGAAAAT GCAACACGCC ATCACTGTGA GGGACTTAGA GTCCTTAGCC AGCAACCTCG
5881 CGCTCTTGCT TTAGAGAAAG CACTGGCCTC TTTGTCACAC ATGGGCACAC ACACCCACAC
5941 ACAAACACAC ACACATAAGC GCTATTCTGG TCTTGTCGGT AAACTCTAGC CTAGGCCCTA
6001 TAGTGAGACG TATTAGAGCT CCAGCTTTTG TTCCCTTTAG TGAGGGTTAA TTCCGAGCTT
6061 GGCGTAATCA TGGTCATAGC TGTTTCCTGT GTGAAATTGT TATCCGCTCA CAATTCCACA
6121 CAACATACGA GCCGGAAGCA TAAAGTGTAA AGCCTGGGGT GCCTAATGAG TGAGCTAACT
6181 CACATTAATT GCGTTGCGCT CACTGCCCGC TTTCCAGTCG GGAAACCTGT CGTGCCAGCT
6241 GCATTAATGA ATCGGCCAAC GCGCGGGGAG AGGCGGTTTG CGTATTGGGC GCTCTTCCGC
6301 TTCCTCGCTC ACTGACTCGC TGCGCTCGGT CGTTCGGCTG CGGCGAGCGG TATCAGCTCA
6361 CTCAAAGGCG GTAATACGGT TATCCACAGA ATCAGGGGAT AACGCAGGAA AGAACATGTG
6421 AGCAAAAGGC CAGCAAAAGG CCAGGAACCG TAAAAAGGCC GCGTTGCTGG CGTTTTTCCA
6481 TAGGCTCCGC CCCCCTGACG AGCATCACAA AAATCGACGC TCAAGTCAGA GGTGGCGAAA
6541 CCCGACAGGA CTATAAAGAT ACCAGGCGTT TCCCCCTGGA AGCTCCCTCG TGCGCTCTCC
6601 TGTTCCGACC CTGCCGCTTA CCGGATACCT GTCCGCCTTT CTCCCTTCGG GAAGCGTGGC
6661 GCTTTCTCAT AGCTCACGCT GTAGGTATCT CAGTTCGGTG TAGGTCGTTC GCTCCAAGCT
6721 GGGCTGTGTG CACGAACCCC CCGTTCAGCC CGACCGCTGC GCCTTATCCG GTAACTATCG
6781 TCTTGAGTCC AACCCGGTAA GACACGACTT ATCGCCACTG GCAGCAGCCA CTGGTAACAG
6841 GATTAGCAGA GCGAGGTATG TAGGCGGTGC TACAGAGTTC TTGAAGTGGT GGCCTAACTA
6901 CGGCTACACT AGAAGGACAG TATTTGGTAT CTGCGCTCTG CTGAAGCCAG TTACCTTCGG
```

```
6961 AAAAAGAGTT GGTAGCTCTT GATCCGGCAA ACAAACCACC GCTGGTAGCG GTGGTTTTTT
7021 TGTTTGCAAG CAGCAGATTA CGCGCAGAAA AAAAGGATCT CAAGAAGATC CTTTGATCTT
7081 TTCTACGGGG TCTGACGCTC AGTGGAACGA AAACTCACGT TAAGGGATTT TGGTCATGAG
7141 ATTATCAAAA AGGATCTTCA CCTAGATCCT TTTAAATTAA AAATGAAGTT TTAAATCAAT
7201 CTAAAGTATA TATGAGTAAA CTTGGTCTGA CAGTTACCAA TGCTTAATCA GTGAGGCACC
7261 TATCTCAGCG ATCTGTCTAT TTCGTTCATC CATAGTTGCC TGACTCCCCG TCGTGTAGAT
7321 AACTACGATA CGGGAGGGCT TACCATCTGG CCCCAGTGCT GCAATGATAC CGCGAGACCC
7381 ACGCTCACCG GCTCCAGATT TATCAGCAAT AAACCAGCCA GCCGGAAGGG CCGAGCGCAG
7441 AAGTGGTCCT GCAACTTTAT CCGCCTCCAT CCAGTCTATT AATTGTTGCC GGGAAGCTAG
7501 AGTAAGTAGT TCGCCAGTTA ATAGTTTGCG CAACGTTGTT GCCATTGCTA CAGGCATCGT
7561 GGTGTCACGC TCGTCGTTTG GTATGGCTTC ATTCAGCTCC GGTTCCCAAC GATCAAGGCG
7621 AGTTACATGA TCCCCCATGT TGTGCAAAAA AGCGGTTAGC TCCTTCGGTC CTCCGATCGT
7681 TGTCAGAAGT AAGTTGGCCG CAGTGTTATC ACTCATGGTT ATGGCAGCAC TGCATAATTC
7741 TCTTACTGTC ATGCCATCCG TAAGATGCTT TTCTGTGACT GGTGAGTACT CAACCAAGTC
7801 ATTCTGAGAA TAGTGTATGC GGCGACCGAG TTGCTCTTGC CCGGCGTCAA TACGGGATAA
7861 TACCGCGCCA CATAGCAGAA CTTTAAAAGT GCTCATCATT GGAAAACGTT CTTCGGGGCG
7921 AAAACTCTCA AGGATCTTAC CGCTGTTGAG ATCCAGTTCG ATGTAACCCA CTCGTGCACC
7981 CAACTGATCT TCAGCATCTT TTACTTTCAC CAGCGTTTCT GGGTGAGCAA AAACAGGAAG
8041 GCAAAATGCC GCAAAAAAGG GAATAAGGGC GACACGGAAA TGTTGAATAC TCATACTCTT
8101 CCTTTTTCAA TATTATTGAA GCATTTATCA GGGTTATTGT CTCATGAGCG GATACATATT
8161 TGAATGTATT TAGAAAAATA AACAAATAGG GGTTCCGCGC ACATTTCCCC GAAAAGTGCC
8221 ACCTAAATTG TAAGCGTTAA TATTTTGTTA AAATTCGCGT TAAATTTTTG TTAAATCAGC
8281 TCATTTTTTA ACCAATAGGC CGAAATCGGC AAAATCCCTT ATAAATCAAA AGAATAGACC
8341 GAGATAGGGT TGAGTGTTGT TCCAGTTTGG AACAAGAGTC CACTATTAAA GAACGTGGAC
8401 TCCAACGTCA AAGGGCGAAA AACCGTCTAT CAGGGCGATG GCCCACTACG TGAACCATCA
8461 CCCTAATCAA GTTTTTTGGG GTCGAGGTGC CGTAAAGCAC TAAATCGGAA CCCTAAAGGG
8521 AGCCCCCGAT CTAGAGCTTG ACGGGGAAAG CCATC
```

Appendix 2. pCExT DNA sequence

```
   1 GAATTCTCCC TATCAGTGAT AGAGAATTCT AATACGACTC ACTATAGGGA TATCTCCCTA
  61 TCAGTGATAG AGATATCCCT ATCAGTGATA GAGAAGCTTA GCAAAACGAA GACATGCGGC
 121 GTGGATCACC AGTACTCTCT CTTTCTCTTT CTTTTTCTCA CTGCCTCCGC TCTCTCTCTC
 181 TCCCTCTTTT TCTCACCTCT TCCTCTCTCT ACCCACACCA ACGCGCACTG CCCCCCATGC
 241 TGACGCCGCA GAATCACGTT CACGTGGTCT ATACCAAACA AAACAACACC CCACCAACAG
 301 CGTACCGCTT TACAACCATA GATCTACCAT GCTCGAGGGC GCTAGCGAGA ACCTGTACTT
 361 CCAGGGCACT AGTCGCAAGG GCGAGGAGCT GTTCACGGGC GTGGTGCCGA TCCTGGTGGA
 421 GCTGGACGGC GACGTGAACG GCCACAAGTT CAGCGTGCGC GGCGAGGGCG AGGGCGACGC
 481 GACGAACGGC AAGCTGACGC TGAAGTTCAT CTGCACGACG GGCAAGCTGC CGGTGCCGTG
 541 GCCGACGCTG GTGACGACGC TGACGTACGG CGTGCAGTGC TTCGCGCGCT ACCCGGACCA
 601 CATGAAGCAG CACGACTTCT TCAAGAGCGC GATGCCGGAG GGCTACGTGC AGGAGCGCAC
 661 GATCAGCTTC AAGGACGACG GCACGTACAA GACGCGCGCG GAGGTGAAGT TCGAGGGCGA
 721 CACGCTGGTG AACCGCATCG AGCTGAAGGG CATCGACTTC AAGGAGGACG GCAACATCCT
 781 GGGCCACAAG CTGGAGTACA ACTTCAACAG CCACAACGTG TACATCACGG CGGACAAGCA
 841 GAAGAACGGC ATCAAGGCGA ACTTCAAGAT CCGCCACAAC GTGGAGGACG GCAGCGTGCA
 901 GCTGGCGGAC CACTACCAGC AGAACACGCC GATCGGCGAC GGCCCGGTGC TGCTGCCGGA
 961 CAACCACTAC CTGAGCACGC AGAGCGTGCT GAGCAAGGAC CCGAACGAGA AGCGCGACCA
1021 CATGGTGCTG CTGGAGTTCG TGACGGCGGC GGGCATCACG CACGGCATGG ACGAGCTGTA
1081 CAAGGGATCC CACCACCATC ACCACCACTA AGCGGCCGTC ATATGGCGAT GGTTCGACAG
1141 GTCCGCAGCA TTTTTCTCTC TTTCCCTTTT TCCGCACCGA ATGAAACGCA GGGAGGCGCT
1201 GGGGAAAGGA GAGAGGAGAT GGCGCACGTC TTCGCCACGA CTTGTGCCGC GTTGAAGCCA
1261 ATTCGTTTGT TTCTTTAGCT TTCCCTTTTC TTTTTTCTCA TTTCCATCTC AAACAAAACA
1321 AACAAAAAAG CAAACATCTA TTGTAGAAGC GAGAGGGCAG TTGACGCTGT TTGTCTTCGC
1381 AAAAAAGAAT AGATTGGGAC AGGGGCATTG CAGACGGCGC CGCAGCAACC TGGCTGTGAC
1441 GACAGGTTGT GGTTGTTGTT GTTGTCCTTG ACGCGTGTGC GTTGGTGTGT GTGTGTGCCC
1501 GCGCTTCCTC TTTCTCTGGG TTGTGGATCG TTCCCTTCCT TTTCGTGACT ACGTCCGCCA
1561 GGTTAGCGGC AGGCGATGAG ATGGAGAAGA GCCTTTTTGG CATTTCTCAT TCTCCGTCGC
1621 TGTTGCTGGT GTTGTCATGC GGTAAAAGAC AATTATCAAG ATGGTCGTCT CCAGCCCTCC
1681 TTTTTCAGTC AACCGGCTGG TGTCGTTGCC GTCAAAAATA TATGCTGCCA CTTCGTGTAG
1741 AGGCGAAATT CTGTGACAGT GCCTTTTACC ACATTTCTCT CCCCTTCTCT CTCTCTCCCC
1801 TCGCCGTTTT TCTCCACAAA AAAGAAAGTC ATCAGCGAAG GGCCCACACT GTGTAGTGCA
1861 TCGACGCACG CTACGGCTCA TTTTCATTGC GGTGTCACAC CGCCGTCTCC CATTCTTCCT
1921 CTCGCACAAT GCATGGCCCG TTTGTTATCT ATGCAGTATT CTGCAGGCTG CTAACAAAGC
1981 CCGAAAGGAA GCTGAGTTGG CTGCTGCCAC CGCTGAGCAA TAACTAGCAT AACCCCTTGG
2041 GGCCTCTAAA CGGGTCTTGA GGGGTTTTTT GCTGAAAGGA GAACTATATC CGGATAGGGC
2101 TGCTAACAAA GCCCGAAAGG AAGCTGAGTT GGCTGCTGCC ACCGCTGAGC AATAACTAGC
2161 ATAACCCCTT GGGGCCTCTA AACGGGTCTT GAGGGGTTTT TTGCTGAAAG GAGAACTATA
2221 TCCGGATTGG GCCTCTAGAC ACAGCAGAGA GAGAAGCCCC CACCCAGCCC TCACACAGGC
2281 GTGTCGTGAA TCATCTCCCT TTCTCGCCTT TCATGTGTGG CTTCAAGGAG CCCAGGCGCG
2341 CCAAAAGAGG AGTCCGGGTG TGAGAGGAGC TGCCTGGCTA CCAGGATGAG TGCCCGTGCC
2401 TGCGCCGCTC TGTGGCCGCT GCACACACTC AGAGGTGGCG GTGTGCGTGC TGTCACTCTT
2461 TACTACCCCC GTTTCGGTTC GCGGGCGGGT GTCGAGGAGA GCGCTGCCCC TCTCCTTCCC
2521 TCTCCCCATC CGCAAGAAAG CAAGTGGCCA GTGCGATACA CGAGGCCAGC CAACTCTACT
2581 TAGTGTTCCT GCAGGAAGCG GATGGCGAGC ACGTGGCGCG CGGCCGCCCC CTGCGAGGAG
2641 CAGCTGCAGC AGCATCCCCC TATAACCCAG AGAAGAAGAG CAAAGGTGAG CGCCACGCTG
2701 CGCCGCGCCA ATACGCCCTA CAGAGAGAAG ATAAATAAAA AGGAACACAG CTGCGAAAGA
2761 CTACACGACT TATCAAGAAG GCCGTTCGTG GGAGCGAGAT GGCAAAGAGG TGTCGCTTTC
2821 CCCCGTGATG GAATCCGCGA AGGCGGGCAG GGAGAAAGGC GAGGACAGCT CCCCCATGGA
2881 CCGGCCCGGA CTTCCCCACT CCCCCATGCA CCACACCACA CACCACCCCG CCCACCGAGG
2941 CTTAACTTTG CAAAGGAGAC GAGACAAAGT GTTTTTCGGA GCAAAAGGAC AACGGGGAAA
3001 GTAAACAGGT ACCAATCCGG ATATAGTTCT CCTTTCAGCA AAAAACCCCT CAAGACCCGT
3061 TTAGAGGCCC CAAGGGGTTA TGCTAGTTAT TGCTCAGCGG TGGCAGCAGC CAACTCAGCT
3121 TCCTTTCGGG CTTTGTTAGC AGCCCTATCC GGATATAGTT CTCCTTTCAG CAAAAAACCC
3181 CTCAAGACCC GTTTAGAGGC CCCAAGGGGT TATGCTAGTT ATTGCTCAGC GGTGGCAGCA
3241 GCCAACTCAG CTTCCTTTCG GGCTTTGTTA GCAGCCTGCA GAATACTGCA TAGATAACAA
3301 ACGGGCCCGG ACGAACCGGA GCTCGCAGAA CTGTGGCACG CATACATCCC CAAAGCAACA
3361 CGCTCCCGTT TTTCGCTTCT TCTTTGTTCT TTGTATAAAG GCACGCATCA GCGCCAAAGC
3421 GTGTTTGAGA GCGTACAGCG GCCACGACGT GGATCAGGAA GAAAGGAGCC GTGGGTGCGT
```

```
3481 TGCTGACCTC TTATTCTTGT CTTTTTTGTA TTTTTGTTTT TGTTGATTAG TTTTGAAGGC
3541 ACAGATCGTG TATGCGTGCT TTTAAAATCT GTGTATGTGC GCAGCGGCAC GTGAGTAAAG
3601 AAGCGCACTC GGACACACCC ACCGACCTGC GTGACGGACC CGAAGATACC GGGAAAAGAA
3661 AAGATGTTTG GAGTCAAATA AATAAAAAAG CCCACTCCAA CACTCAAATC ACCCTTTTCG
3721 CCTCGAATTG ATTAAAATGA TGCGGAAGGC GTCACCACTG GAAAACAAAC ACGCACGATA
3781 TGGAGTAGCG GACAAATGGC ATCAGCACAG AAACAAATAT GTATATAGTA AAACGTACGG
3841 CGGTCGCAGG TGACACCCAC TGCGACGCTG CGGGGCAAAA AGAGCAGGCG CAATCATGTT
3901 AGAAAGAAAA AACACAGAAA AAGGGGAGCG AGGGAAGGAG GAGGGAAGCC GAGAGGACAC
3961 ACCGCCGTGC ACACCGGTCA ACAGGATGAG AAGCAGCGAG AAAGACATCT TCCTCGCTTT
4021 GGTATTTACT TCCCCTCTTC CCTCTTCTCC TGAATGTCTC TGGCGCTCTC ACTGCACACG
4081 TGGGAGGTGG TCAGCACAGT GCCGCAAAAT AGAAGCGCCT GAGGCATCGA GGGAGTACTT
4141 ACGGAGAGGA AAGGGGGACG GAAGTGCGAC TCATCACCGC TGCCGTAGGC AGGAGACCCA
4201 AATCGGCTAA GAAAGAAGAA GAAAAATAAC GGTGATAACA AGACAAAACG GAAGGAAGAA
4261 AAACGGAGAG TAAAAAGAGG AAGAGGATAG CACCGACATT AGAGCAAACA AAAAGAGGAG
4321 TTGGCGACGT GGCGGGAACA GTAAAACAGA GAAGGGCAGA GAGGGGCGCT CCGCCACCAC
4381 CCATCTCTGC CTCTTTTTCT CGATACTCCG ACCTATGCGT CCGCCTTATA CTTCCCTCTT
4441 TCTCCGCCTG CTGGATCCCT ATTCCTTTGC CCTCGGACGA GTGCTGAGGC GTCGGTTTCC
4501 ACTATCGGCG AGTACTTCTA CACAGCCATC GGTCCAGACG GCCGCGCTTC TGCGGGCGAT
4561 TTGTGTACGC CCGACAGTCC CGGCTCCGGA TCGGACGATT GCGTCGCATC GACCCTGCGC
4621 CCAAGCTGCA TCATCGAAAT TGCCGTCAAC CAAGCTCTGA TAGAGTTGGT CAAGACCAAT
4681 GCGGAGCATA TACGCCCGGA GCCGCGGCGA TCCTGCAAGC TCCGGATGCC TCCGCTCGAA
4741 GTAGCGCGTC TGCTGCTCCA TACAAGCCAA CCACGGCCTC CAGAAGAAGA TGTTGGCGAC
4801 CTCGTATTGG GAATCCCCGA ACATCGCCTC GCTCCAGTCA ATGACCGCTG TTATGCGGCC
4861 ATTGTCCGTC AGGACATTGT TGGAGCCGAA ATCCGCGTGC ACGAGGTGCC GGACTTCGGG
4921 GCAGTCCTCG GCCCAAAGCA TCAGCTCATC GAGAGCCTGC GCGACGGACG CACTGACGGT
4981 GTCGTCCATC ACAGTTTGCC AGTGATACAC ATGGGGATCA GCAATCGCGC AAATGAAATC
5041 ACGCCATGTA GTGTATTGAC CGATTCCTTG CGGTCCGAAT GGGCCGAACC CGCTCGTCTG
5101 GCTAAGATCG GCCGCAGCGA TCGCATCCAT GGCCTCCGCG ACCGGCTGAA GAACAGCGGG
5161 CAGTTCGGTT TCAGGCAGGT CTTGCAACGT GACACCCTGT GCACGGCGGG AGATGCAATA
5221 GGTCAGGCTC TCGCTGAACT CCCCAATGTC AAGCACTTCC GGAATCGGGA GCGCGGCCGA
5281 TGCAAAGTGC CGATAAACAT AACGATCTTT GTAGAAACCA TCGGCGCAGC TATTTACCCG
5341 CAGGACATAT CCACGCCCTC CTACATCGAA GCTGAAAGCA CGAGATTCTT CGCCCTCCGA
5401 GAGCTGCATC AGGTCGGAGA CGCTGTCGAA CTTTCCGATC AGAAACTTCT CGACAGACGT
5461 CGCGGTGAGT TCAGGCTTTT TCATGTCGAC GCTTGACAAG TGGAAGATAG TTGGACAAGT
5521 AGAGGTGTAA GGTCGAAGTG GACAGAGCGG AATGTGACGT TGACTGATGA GTAATACTCG
5581 CGTTTAATAT GAGTGTGTTA TCGTGAAAGA AAAGACAGCT GTTTGCTCGA AAAGGACACG
5641 CGCCATGTAC ACGCAGGACA CAACAACAAC CGGAGGAAGA AGGCAGAGAA GTGGTGAGTG
5701 GATTAGTCAA AGAGCGCAGC CTGAGGCGAT AGCTGGCAAA GGCGGTGATA ATTAGACAAG
5761 GCAACGCGAC GGTTTCAGTG ATATTTACAG TGCTTCAGTC ACGAAGCAGT GGCGGTCTTA
5821 CTGGGGAAGG TATCCAGAAG AGCCTACAAA ACCGATATGC TACCGACGTG TTGGATTGAC
5881 GGTTTAGAAA ATGCAACACG CCATCACTGT GAGGGACTTA GAGTCCTTAG CCAGCAACCT
5941 CGCGCTCTTG CTTTAGAGAA AGCACTGGCC TCTTTGTCAC ACATGGGCAC ACACACCCAC
6001 ACACAAACAC ACACACATAA GCGCTATTCT GGTCTTGTCG GTAAACTCTA GCCTAGGCCC
6061 TATAGTGAGA CGTATTAGAG CTCCAGCTTT TGTTCCCTTT AGTGAGGGTT AATTCCGAGC
6121 TTGGCGTAAT CATGGTCATA GCTGTTTCCT GTGTGAAATT GTTATCCGCT CACAATTCCA
6181 CACAACATAC GAGCCGGAAG CATAAAGTGT AAAGCCTGGG GTGCCTAATG AGTGAGCTAA
6241 CTCACATTAA TTGCGTTGCG CTCACTGCCC GCTTTCCAGT CGGGAAACCT GTCGTGCCAG
6301 CTGCATTAAT GAATCGGCCA ACGCGCGGGG AGAGGCGGTT TGCGTATTGG GCGCTCTTCC
6361 GCTTCCTCGC TCACTGACTC GCTGCGCTCG GTCGTTCGGC TGCGGCGAGC GGTATCAGCT
6421 CACTCAAAGG CGGTAATACG GTTATCCACA GAATCAGGGG ATAACGCAGG AAAGAACATG
6481 TGAGCAAAAG GCCAGCAAAA GGCCAGGAAC CGTAAAAAGG CCGCGTTGCT GGCGTTTTTC
6541 CATAGGCTCC GCCCCCCTGA CGAGCATCAC AAAAATCGAC GCTCAAGTCA GAGGTGGCGA
6601 AACCCGACAG GACTATAAAG ATACCAGGCG TTTCCCCCTG GAAGCTCCCT CGTGCGCTCT
6661 CCTGTTCCGA CCCTGCCGCT TACCGGATAC CTGTCCGCCT TTCTCCCTTC GGGAAGCGTG
6721 GCGCTTTCTC ATAGCTCACG CTGTAGGTAT CTCAGTTCGG TGTAGGTCGT TCGCTCCAAG
6781 CTGGGCTGTG TGCACGAACC CCCCGTTCAG CCCGACCGCT GCGCCTTATC CGGTAACTAT
6841 CGTCTTGAGT CCAACCCGGT AAGACACGAC TTATCGCCAC TGGCAGCAGC CACTGGTAAC
6901 AGGATTAGCA GAGCGAGGTA TGTAGGCGGT GCTACAGAGT TCTTGAAGTG GTGGCCTAAC
6961 TACGGCTACA CTAGAAGGAC AGTATTTGGT ATCTGCGCTC TGCTGAAGCC AGTTACCTTC
7021 GGAAAAAGAG TTGGTAGCTC TTGATCCGGC AAACAAACCA CCGCTGGTAG CGGTGGTTTT
```

7081 TTTGTTTGCA AGCAGCAGAT TACGCGCAGA AAAAAAGGAT CTCAAGAAGA TCCTTTGATC
7141 TTTTCTACGG GGTCTGACGC TCAGTGGAAC GAAAACTCAC GTTAAGGGAT TTTGGTCATG
7201 AGATTATCAA AAAGGATCTT CACCTAGATC CTTTTAAATT AAAAATGAAG TTTTAAATCA
7261 ATCTAAAGTA TATATGAGTA AACTTGGTCT GACAGTTACC AATGCTTAAT CAGTGAGGCA
7321 CCTATCTCAG CGATCTGTCT ATTTCGTTCA TCCATAGTTG CCTGACTCCC CGTCGTGTAG
7381 ATAACTACGA TACGGGAGGG CTTACCATCT GGCCCCAGTG CTGCAATGAT ACCGCGAGAC
7441 CCACGCTCAC CGGCTCCAGA TTTATCAGCA ATAAACCAGC CAGCCGGAAG GGCCGAGCGC
7501 AGAAGTGGTC CTGCAACTTT ATCCGCCTCC ATCCAGTCTA TTAATTGTTG CCGGGAAGCT
7561 AGAGTAAGTA GTTCGCCAGT TAATAGTTTG CGCAACGTTG TTGCCATTGC TACAGGCATC
7621 GTGGTGTCAC GCTCGTCGTT TGGTATGGCT TCATTCAGCT CCGGTTCCCA ACGATCAAGG
7681 CGAGTTACAT GATCCCCCAT GTTGTGCAAA AAAGCGGTTA GCTCCTTCGG TCCTCCGATC
7741 GTTGTCAGAA GTAAGTTGGC CGCAGTGTTA TCACTCATGG TTATGGCAGC ACTGCATAAT
7801 TCTCTTACTG TCATGCCATC CGTAAGATGC TTTTCTGTGA CTGGTGAGTA CTCAACCAAG
7861 TCATTCTGAG AATAGTGTAT GCGGCGACCG AGTTGCTCTT GCCCGGCGTC AATACGGGAT
7921 AATACCGCGC CACATAGCAG AACTTTAAAA GTGCTCATCA TTGGAAAACG TTCTTCGGGG
7981 CGAAAACTCT CAAGGATCTT ACCGCTGTTG AGATCCAGTT CGATGTAACC CACTCGTGCA
8041 CCCAACTGAT CTTCAGCATC TTTTACTTTC ACCAGCGTTT CTGGGTGAGC AAAAACAGGA
8101 AGGCAAAATG CCGCAAAAAA GGGAATAAGG GCGACACGGA AATGTTGAAT ACTCATACTC
8161 TTCCTTTTTC AATATTATTG AAGCATTTAT CAGGGTTATT GTCTCATGAG CGGATACATA
8221 TTTGAATGTA TTTAGAAAAA TAAACAAATA GGGGTTCCGC GCACATTTCC CCGAAAAGTG
8281 CCACCTAAAT TGTAAGCGTT AATATTTTGT TAAAATTCGC GTTAAATTTT TGTTAAATCA
8341 GCTCATTTTT TAACCAATAG GCCGAAATCG GCAAAATCCC TTATAAATCA AAAGAATAGA
8401 CCGAGATAGG GTTGAGTGTT GTTCCAGTTT GGAACAAGAG TCCACTATTA AAGAACGTGG
8461 ACTCCAACGT CAAAGGGCGA AAAACCGTCT ATCAGGGCGA TGGCCCACTA CGTGAACCAT
8521 CACCCTAATC AAGTTTTTTG GGGTCGAGGT GCCGTAAAGC ACTAAATCGG AACCCTAAAG
8581 GGAGCCCCCG ATCTAGAGCT TGACGGGGAA AGCCATC

Appendix 3. pCExS DNA sequence

```
   1 GAATTCTAAT ACGACTCACT ATAGGGATAT CAAGCTTAGC AAAACGAAGA CATGCGGCGT
  61 GGATCACCAG TACTCTCTCT TTCTCTTTCT TTTTCTCACT GCCTCCGCTC TCTCTCTCTC
 121 CCTCTTTTTC TCACCTCTTC CTCTCTCTAC CCACACCAAC GCGCACTGCC CCCCATGCTG
 181 ACGCCGCAGA ATCACGTTCA CGTGGTCTAT ACCAAACAAA ACAACACCCC ACCAACAGCG
 241 TACCGCTTTA CAACCATAGA TCTGCCATGG CCAGCCGCCT CGTCCGTGTG CTGGCCGCCG
 301 CCATGCTGGT TGCAGCGGCC GTGTCGGTGG ACGCTGGCCT CGAGGGCGCT AGCGAGAACC
 361 TGTACTTCCA GGGCACTAGT CGCAAGGGCG AGGAGCTGTT CACGGGCGTG GTGCCGATCC
 421 TGGTGGAGCT GGACGGCGAC GTGAACGGCC ACAAGTTCAG CGTGCGCGGC GAGGGCGAGG
 481 GCGACGCGAC GAACGGCAAG CTGACGCTGA AGTTCATCTG CACGACGGGC AAGCTGCCGG
 541 TGCCGTGGCC GACGCTGGTG ACGACGCTGA CGTACGGCGT GCAGTGCTTC GCGCGCTACC
 601 CGGACCACAT GAAGCAGCAC GACTTCTTCA AGAGCGCGAT GCCGGAGGGC TACGTGCAGG
 661 AGCGCACGAT CAGCTTCAAG GACGACGGCA CGTACAAGAC GCGCGCGGAG GTGAAGTTCG
 721 AGGGCGACAC GCTGGTGAAC CGCATCGAGC TGAAGGGCAT CGACTTCAAG GAGGACGGCA
 781 ACATCCTGGG CCACAAGCTG GAGTACAACT TCAACAGCCA CAACGTGTAC ATCACGGCGG
 841 ACAAGCAGAA GAACGGCATC AAGGCGAACT TCAAGATCCG CCACAACGTG GAGGACGGCA
 901 GCGTGCAGCT GGCGGACCAC TACCAGCAGA ACACGCCGAT CGGCGACGGC CCGGTGCTGC
 961 TGCCGGACAA CCACTACCTG AGCACGCAGA GCGTGCTGAG CAAGGACCCG AACGAGAAGC
1021 GCGACCACAT GGTGCTGCTG GAGTTCGTGA CGGCGGCGGG CATCACGCAC GGCATGGACG
1081 AGCTGTACAA GGGATCCCAC CACCATCACC ACCACTAAGC GGCCGTCATA TGGCGATGGT
1141 TCGACAGGTC CGCAGCATTT TTCTCTCTTT CCCTTTTTCC GCACCGAATG AAACGCAGGG
1201 AGGCGCTGGG GAAAGGAGAG AGGAGATGGC GCACGTCTTC GCCACGACTT GTGCCGCGTT
1261 GAAGCCAATT CGTTTGTTTC TTTAGCTTTC CCTTTTCTTT TTTCTCATTT CCATCTCAAA
1321 CAAAACAAAC AAAAAAGCAA ACATCTATTG TAGAAGCGAG AGGGCAGTTG ACGCTGTTTG
1381 TCTTCGCAAA AAAGAATAGA TTGGGACAGG GGCATTGCAG ACGGCGCCGC AGCAACCTGG
1441 CTGTGACGAC AGGTTGTGGT TGTTGTTGTT GTCCTTGACG CGTGTGCGTT GGTGTGTGTG
1501 TGTGCCCGCG CTTCCTCTTT CTCTGGGTTG TGGATCGTTC CCTTCCTTTT CGTGACTACG
1561 TCCGCCAGGT TAGCGGCAGG CGATGAGATG GAGAAGAGCC TTTTTGGCAT TTCTCATTCT
1621 CCGTCGCTGT TGCTGGTGTT GTCATGCGGT AAAAGACAAT TATCAAGATG GTCGTCTCCA
1681 GCCCTCCTTT TTCAGTCAAC CGGCTGGTGT CGTTGCCGTC AAAAATATAT GCTGCCACTT
1741 CGTGTAGAGG CGAAATTCTG TGACAGTGCC TTTTACCACA TTTCTCTCCC CTTCTCTCTC
1801 TCTCCCCTCG CCGTTTTTCT CCACAAAAAA GAAAGTCATC AGCGAAGGGC CCACACTGTG
1861 TAGTGCATCG ACGCACGCTA CGGCTCATTT TCATTGCGGT GTCACACCGC CGTCTCCCAT
1921 TCTTCCTCTC GCACAATGCA TGGCCCGTTT GTTATCTATG CAGTATTCTG CAGGCTGCTA
1981 ACAAAGCCCG AAAGGAAGCT GAGTTGGCTG CTGCCACCGC TGAGCAATAA CTAGCATAAC
2041 CCCTTGGGGC CTCTAAACGG GTCTTGAGGG GTTTTTTGCT GAAAGGAGAA CTATATCCGG
2101 ATAGGGCTGC TAACAAAGCC CGAAAGGAAG CTGAGTTGGC TGCTGCCACC GCTGAGCAAT
2161 AACTAGCATA ACCCCTTGGG GCCTCTAAAC GGGTCTTGAG GGGTTTTTTG CTGAAAGGAG
2221 AACTATATCC GGATTGGGCC TCTAGACACA GCAGAGAGAG AAGCCCCCAC CCAGCCCTCA
2281 CACAGGCGTG TCGTGAATCA TCTCCCTTTC TCGCCTTTCA TGTGTGGCTT CAAGGAGCCC
2341 AGGCGCGCCA AAAGAGGAGT CCGGGTGTGA GAGGAGCTGC CTGGCTACCA GGATGAGTGC
2401 CCGTGCCTGC GCCGCTCTGT GGCCGCTGCA CACACTCAGA GGTGGCGGTG TGCGTGCTGT
2461 CACTCTTTAC TACCCCCGTT TCGGTTCGCG GGCGGGTGTC GAGGAGAGCG CTGCCCCTCT
2521 CCTTCCCTCT CCCCATCCGC AAGAAAGCAA GTGGCCAGTG CGATACACGA GGCCAGCCAA
2581 CTCTACTTAG TGTTCCTGCA GGAAGCGGAT GGCGAGCACG TGGCGCGCGG CCGCCCCCTG
2641 CGAGGAGCAG CTGCAGCAGC ATCCCCCTAT AACCCAGAGA AGAAGAGCAA AGGTGAGCGC
2701 CACGCTGCGC CGCGCCAATA CGCCCTACAG AGAGAAGATA AATAAAAAGG AACACAGCTG
2761 CGAAAGACTA CACGACTTAT CAAGAAGGCC GTTCGTGGGA GCGAGATGGC AAAGAGGTGT
2821 CGCTTTCCCC CGTGATGGAA TCCGCGAAGG CGGGCAGGGA GAAAGGCGAG GACAGCTCCC
2881 CCATGGACCG GCCCGGACTT CCCCACTCCC CCATGCACCA CACCACACAC CACCCCGCCC
2941 ACCGAGGCTT AACTTTGCAA AGGAGACGAG ACAAAGTGTT TTTCGGAGCA AAAGGACAAC
3001 GGGGAAAGTA AACAGGTACC AATCCGGATA TAGTTCTCCT TTCAGCAAAA AACCCCTCAA
3061 GACCCGTTTA GAGGCCCCAA GGGGTTATGC TAGTTATTGC TCAGCGGTGG CAGCAGCCAA
3121 CTCAGCTTCC TTTCGGGCTT TGTTAGCAGC CCTATCCGGA TATAGTTCTC CTTTCAGCAA
3181 AAAACCCCTC AAGACCCGTT TAGAGGCCCC AAGGGGTTAT GCTAGTTATT GCTCAGCGGT
3241 GGCAGCAGCC AACTCAGCTT CCTTTCGGGC TTTGTTAGCA GCCTGCAGAA TACTGCATAG
3301 ATAACAAACG GGCCCGGACG AACCGGAGCT CGCAGAACTG TGGCACGCAT ACATCCCCAA
3361 AGCAACACGC TCCCGTTTTT CGCTTCTTCT TTGTTCTTTG TATAAAGGCA CGCATCAGCG
3421 CCAAAGCGTG TTTGAGAGCG TACAGCGGCC ACGACGTGGA TCAGGAAGAA AGGAGCCGTG
```

```
3481 GGTGCGTTGC TGACCTCTTA TTCTTGTCTT TTTTGTATTT TTGTTTTTGT TGATTAGTTT
3541 TGAAGGCACA GATCGTGTAT GCGTGCTTTT AAAATCTGTG TATGTGCGCA GCGGCACGTG
3601 AGTAAAGAAG CGCACTCGGA CACACCCACC GACCTGCGTG ACGGACCCGA AGATACCGGG
3661 AAAAGAAAAG ATGTTTGGAG TCAAATAAAT AAAAAAGCCC ACTCCAACAC TCAAATCACC
3721 CTTTTCGCCT CGAATTGATT AAAATGATGC GGAAGGCGTC ACCACTGGAA AACAAACACG
3781 CACGATATGG AGTAGCGGAC AAATGGCATC AGCACAGAAA CAAATATGTA TATAGTAAAA
3841 CGTACGGCGG TCGCAGGTGA CACCCACTGC GACGCTGCGG GGCAAAAAGA GCAGGCGCAA
3901 TCATGTTAGA AAGAAAAAAC ACAGAAAAAG GGGAGCGAGG GAAGGAGGAG GGAAGCCGAG
3961 AGGACACACC GCCGTGCACA CCGGTCAACA GGATGAGAAG CAGCGAGAAA GACATCTTCC
4021 TCGCTTTGGT ATTTACTTCC CCTCTTCCCT CTTCTCCTGA ATGTCTCTGG CGCTCTCACT
4081 GCACACGTGG GAGGTGGTCA GCACAGTGCC GCAAAATAGA AGCGCCTGAG GCATCGAGGG
4141 AGTACTTACG GAGAGGAAAG GGGGACGGAA GTGCGACTCA TCACCGCTGC CGTAGGCAGG
4201 AGACCCAAAT CGGCTAAGAA AGAAGAAGAA AAATAACGGT GATAACAAGA CAAAACGGAA
4261 GGAAGAAAAA CGGAGAGTAA AAAGAGGAAG AGGATAGCAC CGACATTAGA GCAAACAAAA
4321 AGAGGAGTTG GCGACGTGGC GGGAACAGTA AAACAGAGAA GGGCAGAGAG GGGCGCTCCG
4381 CCACCACCCA TCTCTGCCTC TTTTTCTCGA TACTCCGACC TATGCGTCCG CCTTATACTT
4441 CCCTCTTTCT CCGCCTGCTG GATCCCTATT CCTTTGCCCT CGGACGAGTG CTGAGGCGTC
4501 GGTTTCCACT ATCGGCGAGT ACTTCTACAC AGCCATCGGT CCAGACGGCC GCGCTTCTGC
4561 GGGCGATTTG TGTACGCCCG ACAGTCCCGG CTCCGGATCG GACGATTGCG TCGCATCGAC
4621 CCTGCGCCCA AGCTGCATCA TCGAAATTGC CGTCAACCAA GCTCTGATAG AGTTGGTCAA
4681 GACCAATGCG GAGCATATAC GCCCGGAGCC GCGGCGATCC TGCAAGCTCC GGATGCCTCC
4741 GCTCGAAGTA GCGCGTCTGC TGCTCCATAC AAGCCAACCA CGGCCTCCAG AAGAAGATGT
4801 TGGCGACCTC GTATTGGGAA TCCCCGAACA TCGCCTCGCT CCAGTCAATG ACCGCTGTTA
4861 TGCGGCCATT GTCCGTCAGG ACATTGTTGG AGCCGAAATC CGCGTGCACG AGGTGCCGGA
4921 CTTCGGGGCA GTCCTCGGCC CAAAGCATCA GCTCATCGAG AGCCTGCGCG ACGGACGCAC
4981 TGACGGTGTC GTCCATCACA GTTTGCCAGT GATACACATG GGGATCAGCA ATCGCGCAAA
5041 TGAAATCACG CCATGTAGTG TATTGACCGA TTCCTTGCGG TCCGAATGGG CCGAACCCGC
5101 TCGTCTGGCT AAGATCGGCC GCAGCGATCG CATCCATGGC CTCCGCGACC GGCTGAAGAA
5161 CAGCGGGCAG TTCGGTTTCA GGCAGGTCTT GCAACGTGAC ACCCTGTGCA CGGCGGGAGA
5221 TGCAATAGGT CAGGCTCTCG CTGAACTCCC CAATGTCAAG CACTTCCGGA ATCGGGAGCG
5281 CGGCCGATGC AAAGTGCCGA TAAACATAAC GATCTTTGTA GAAACCATCG GCGCAGCTAT
5341 TTACCCGCAG GACATATCCA CGCCCTCCTA CATCGAAGCT GAAAGCACGA GATTCTTCGC
5401 CCTCCGAGAG CTGCATCAGG TCGGAGACGC TGTCGAACTT TCCGATCAGA AACTTCTCGA
5461 CAGACGTCGC GGTGAGTTCA GGCTTTTTCA TGTCGACGCT TGACAAGTGG AAGATAGTTG
5521 GACAAGTAGA GGTGTAAGGT CGAAGTGGAC AGAGCGGAAT GTGACGTTGA CTGATGAGTA
5581 ATACTCGCGT TTAATATGAG TGTGTTATCG TGAAAGAAAA GACAGCTGTT TGCTCGAAAA
5641 GGACACGCGC CATGTACACG CAGGACACAA CAACAACCGG AGGAAGAAGG CAGAGAAGTG
5701 GTGAGTGGAT TAGTCAAAGA GCGCAGCCTG AGGCGATAGC TGGCAAAGGC GGTGATAATT
5761 AGACAAGGCA ACGCGACGGT TTCAGTGATA TTTACAGTGC TTCAGTCACG AAGCAGTGGC
5821 GGTCTTACTG GGGAAGGTAT CCAGAAGAGC CTACAAAACC GATATGCTAC CGACGTGTTG
5881 GATTGACGGT TTAGAAAATG CAACACGCCA TCACTGTGAG GGACTTAGAG TCCTTAGCCA
5941 GCAACCTCGC GCTCTTGCTT TAGAGAAAGC ACTGGCCTCT TTGTCACACA TGGGCACACA
6001 CACCCACACA CAAACACACA CACATAAGCG CTATTCTGGT CTTGTCGGTA AACTCTAGCC
6061 TAGGCCCTAT AGTGAGACGT ATTAGAGCTC CAGCTTTTGT TCCCTTTAGT GAGGGTTAAT
6121 TCCGAGCTTG GCGTAATCAT GGTCATAGCT GTTTCCTGTG TGAAATTGTT ATCCGCTCAC
6181 AATTCCACAC AACATACGAG CCGGAAGCAT AAAGTGTAAA GCCTGGGGTG CCTAATGAGT
6241 GAGCTAACTC ACATTAATTG CGTTGCGCTC ACTGCCCGCT TTCCAGTCGG GAAACCTGTC
6301 GTGCCAGCTG CATTAATGAA TCGGCCAACG CGCGGGGAGA GGCGGTTTGC GTATTGGGCG
6361 CTCTTCCGCT TCCTCGCTCA CTGACTCGCT GCGCTCGGTC GTTCGGCTGC GGCGAGCGGT
6421 ATCAGCTCAC TCAAAGGCGG TAATACGGTT ATCCACAGAA TCAGGGGATA ACGCAGGAAA
6481 GAACATGTGA GCAAAAGGCC AGCAAAAGGC CAGGAACCGT AAAAAGGCCG CGTTGCTGGC
6541 GTTTTTCCAT AGGCTCCGCC CCCCTGACGA GCATCACAAA AATCGACGCT CAAGTCAGAG
6601 GTGGCGAAAC CCGACAGGAC TATAAAGATA CCAGGCGTTT CCCCCTGGAA GCTCCCTCGT
6661 GCGCTCTCCT GTTCCGACCC TGCCGCTTAC CGGATACCTG TCCGCCTTTC TCCCTTCGGG
6721 AAGCGTGGCG CTTTCTCATA GCTCACGCTG TAGGTATCTC AGTTCGGTGT AGGTCGTTCG
6781 CTCCAAGCTG GGCTGTGTGC ACGAACCCCC CGTTCAGCCC GACCGCTGCG CCTTATCCGG
6841 TAACTATCGT CTTGAGTCCA ACCCGGTAAG ACACGACTTA TCGCCACTGG CAGCAGCCAC
6901 TGGTAACAGG ATTAGCAGAG CGAGGTATGT AGGCGGTGCT ACAGAGTTCT TGAAGTGGTG
6961 GCCTAACTAC GGCTACACTA GAAGGACAGT ATTTGGTATC TGCGCTCTGC TGAAGCCAGT
7021 TACCTTCGGA AAAAGAGTTG GTAGCTCTTG ATCCGGCAAA CAAACCACCG CTGGTAGCGG
```

```
7081 TGGTTTTTTT GTTTGCAAGC AGCAGATTAC GCGCAGAAAA AAAGGATCTC AAGAAGATCC
7141 TTTGATCTTT TCTACGGGGT CTGACGCTCA GTGGAACGAA AACTCACGTT AAGGGATTTT
7201 GGTCATGAGA TTATCAAAAA GGATCTTCAC CTAGATCCTT TTAAATTAAA AATGAAGTTT
7261 TAAATCAATC TAAAGTATAT ATGAGTAAAC TTGGTCTGAC AGTTACCAAT GCTTAATCAG
7321 TGAGGCACCT ATCTCAGCGA TCTGTCTATT TCGTTCATCC ATAGTTGCCT GACTCCCCGT
7381 CGTGTAGATA ACTACGATAC GGGAGGGCTT ACCATCTGGC CCCAGTGCTG CAATGATACC
7441 GCGAGACCCA CGCTCACCGG CTCCAGATTT ATCAGCAATA AACCAGCCAG CCGGAAGGGC
7501 CGAGCGCAGA AGTGGTCCTG CAACTTTATC CGCCTCCATC CAGTCTATTA ATTGTTGCCG
7561 GGAAGCTAGA GTAAGTAGTT CGCCAGTTAA TAGTTTGCGC AACGTTGTTG CCATTGCTAC
7621 AGGCATCGTG GTGTCACGCT CGTCGTTTGG TATGGCTTCA TTCAGCTCCG GTTCCCAACG
7681 ATCAAGGCGA GTTACATGAT CCCCCATGTT GTGCAAAAAA GCGGTTAGCT CCTTCGGTCC
7741 TCCGATCGTT GTCAGAAGTA AGTTGGCCGC AGTGTTATCA CTCATGGTTA TGGCAGCACT
7801 GCATAATTCT CTTACTGTCA TGCCATCCGT AAGATGCTTT TCTGTGACTG GTGAGTACTC
7861 AACCAAGTCA TTCTGAGAAT AGTGTATGCG GCGACCGAGT TGCTCTTGCC CGGCGTCAAT
7921 ACGGGATAAT ACCGCGCCAC ATAGCAGAAC TTTAAAAGTG CTCATCATTG GAAAACGTTC
7981 TTCGGGGCGA AAACTCTCAA GGATCTTACC GCTGTTGAGA TCCAGTTCGA TGTAACCCAC
8041 TCGTGCACCC AACTGATCTT CAGCATCTTT TACTTTCACC AGCGTTTCTG GGTGAGCAAA
8101 AACAGGAAGG CAAAATGCCG CAAAAAAGGG AATAAGGGCG ACACGGAAAT GTTGAATACT
8161 CATACTCTTC CTTTTTCAAT ATTATTGAAG CATTTATCAG GGTTATTGTC TCATGAGCGG
8221 ATACATATTT GAATGTATTT AGAAAAATAA ACAAATAGGG GTTCCGCGCA CATTTCCCCG
8281 AAAAGTGCCA CCTAAATTGT AAGCGTTAAT ATTTTGTTAA AATTCGCGTT AAATTTTTGT
8341 TAAATCAGCT CATTTTTTAA CCAATAGGCC GAAATCGGCA AAATCCCTTA TAAATCAAAA
8401 GAATAGACCG AGATAGGGTT GAGTGTTGTT CCAGTTTGGA ACAAGAGTCC ACTATTAAAG
8461 AACGTGGACT CCAACGTCAA AGGGCGAAAA ACCGTCTATC AGGGCGATGG CCCACTACGT
8521 GAACCATCAC CCTAATCAAG TTTTTTGGGG TCGAGGTGCC GTAAAGCACT AAATCGGAAC
8581 CCTAAAGGGA GCCCCCGATC TAGAGCTTGA CGGGGAAAGC CATC
```

Appendix 4. pCExP DNA sequence

```
   1 GAATTCATTT AAATCCTGCG CGTGCTGTGT ACTTCTGCGA CGCCACACTC TGTGTGTGCG
  61 TCTGCGGGAG TGCCACCCCC ACACTCACAC AAACACACAC AACCCTGTTG CTGTGTGTTA
 121 TATGTGTGCG TGTGTGCGTG GCTTCAAACT TGTTTGTTAA CACACCCCTA GCACACACAA
 181 TACGCGAGAA CACCCAAAGG GATACATATC CTGTCAGTGC GCTGCGCTGG GCTTACCGTA
 241 CGTGCCGAAC ATGTTTTGTG TGTGCTGCCG TTCTTTTACG TGTTGCAGCA GAAACTCGGT
 301 TGGAATGGAG TCGGCACTTG ATTGATGCTG CACTTCGAGG AATTTCAGCC AAACAAGAGG
 361 TCAACAACCA AGCCACGTTT GTGAGTGCGG GTGTGTGTCA AAACACACCC ACACCCACAT
 421 GCAAAGCTTA GCAAAACGAA GACATGCGGC GTGGATCACC AGTACTCTCT CTTTCTCTTT
 481 CTTTTTCTCA CTGCCTCCGC TCTCTCTCTC TCCCTCTTTT TCTCACCTCT TCCTCTCTCT
 541 ACCCACACCA ACGCGCACTG CCCCCCATGC TGACGCCGCA GAATCACGTT CACGTGGTCT
 601 ATACCAAACA AAACAACACC CCACCAACAG CGTACCGCTT TACAACCATA GATCTACCAT
 661 GCTCGAGGGC GCTAGCGAGA ACCTGTACTT CCAGGGCACT AGTCGCAAGG GCGAGGAGCT
 721 GTTCACGGGC GTGGTGCCGA TCCTGGTGGA GCTGGACGGC GACGTGAACG GCCACAAGTT
 781 CAGCGTGCGC GGCGAGGGCG AGGGCGACGC GACGAACGGC AAGCTGACGC TGAAGTTCAT
 841 CTGCACGACG GGCAAGCTGC CGGTGCCGTG GCCGACGCTG GTGACGACGC TGACGTACGG
 901 CGTGCAGTGC TTCGCGCGCT ACCCGGACCA CATGAAGCAG CACGACTTCT TCAAGAGCGC
 961 GATGCCGGAG GGCTACGTGC AGGAGCGCAC GATCAGCTTC AAGGACGACG GCACGTACAA
1021 GACGCGCGCG GAGGTGAAGT TCGAGGGCGA CACGCTGGTG AACCGCATCG AGCTGAAGGG
1081 CATCGACTTC AAGGAGGACG GCAACATCCT GGGCCACAAG CTGGAGTACA ACTTCAACAG
1141 CCACAACGTG TACATCACGG CGGACAAGCA GAAGAACGGC ATCAAGGCGA ACTTCAAGAT
1201 CCGCCACAAC GTGGAGGACG GCAGCGTGCA GCTGGCGGAC CACTACCAGC AGAACACGCC
1261 GATCGGCGAC GGCCCGGTGC TGCTGCCGGA CAACCACTAC CTGAGCACGC AGAGCGTGCT
1321 GAGCAAGGAC CCGAACGAGA AGCGCGACCA CATGGTGCTG CTGGAGTTCG TGACGGCGGC
1381 GGGCATCACG CACGGCATGG ACGAGCTGTA CAAGGGATCC CACCACCATC ACCACCACTA
1441 AGCGGCCGTC ATATGGCGAT GGTTCGACAG GTCCGCAGCA TTTTTCTCTC TTTCCCTTTT
1501 TCCGCACCGA ATGAAACGCA GGGAGGCGCT GGGGAAAGGA GAGAGGAGAT GGCGCACGTC
1561 TTCGCCACGA CTTGTGCCGC GTTGAAGCCA ATTCGTTTGT TTCTTTAGCT TTCCCTTTTC
1621 TTTTTTCTCA TTTCCATCTC AAACAAAACA AACAAAAAAG CAAACATCTA TTGTAGAAGC
1681 GAGAGGGCAG TTGACGCTGT TTGTCTTCGC AAAAAAGAAT AGATTGGGAC AGGGGCATTG
1741 CAGACGGCGC CGCAGCAACC TGGCTGTGAC GACAGGTTGT GGTTGTTGTT GTTGTCCTTG
1801 ACGCGTGTGC GTTGGTGTGT GTGTGTGCCC GCGCTTCCTC TTTCTCTGGG TTGTGGATCG
1861 TTCCCTTCCT TTTCGTGACT ACGTCCGCCA GGTTAGCGGC AGGCGATGAG ATGGAGAAGA
1921 GCCTTTTTGG CATTTCTCAT TCTCCGTCGC TGTTGCTGGT GTTGTCATGC GGTAAAAGAC
1981 AATTATCAAG ATGGTCGTCT CCAGCCCTCC TTTTTCAGTC AACCGGCTGG TGTCGTTGCC
2041 GTCAAAAATA TATGCTGCCA CTTCGTGTAG AGGCGAAATT CTGTGACAGT GCCTTTTACC
2101 ACATTTCTCT CCCCTTCTCT CTCTCTCCCC TCGCCGTTTT TCTCCACAAA AAAGAAAGTC
2161 ATCAGCGAAG GGCCCACACT GTGTAGTGCA TCGACGCACG CTACGGCTCA TTTTCATTGC
2221 GGTGTCACAC CGCCGTCTCC CATTCTTCCT CTCGCACAAT GCATACCGAC AAGACCAGAA
2281 TAGCGCTTAT GTGTGTGTGT TTGTGTGTGG GTGTGTGTGC CCATGTGTGA CAAAGAGGCC
2341 AGTGCTTTCT CTAAAGCAAG AGCGCGAGGT TGCTGGCTAA GGACTCTAAG TCCCTCACAG
2401 TGATGGCGTG TTGCATTTTC TAAACCGTCA ATCCAACACG TCGGTAGCAT ATCGGTTTTG
2461 TAGGCTCTTC TGGATACCTT CCCCAGTAAG ACCGCCACTG CTTCGTGACT GAAGCACTGT
2521 AAATATCACT GAAACCGTCG CGTTGCCTTG TCTAATTATC ACCGCCTTTG CCAGCTATCG
2581 CCTCAGGCTG CGCTCTTTGA CTAATCCACT CACCACTTCT CTGCCTTCTT CCTCCGGTTG
2641 TTGTTGTGTC CTGCGTGTAC ATGGCGCGTG TCCTTTTCGA GCAAACAGCT GTCTTTTCTT
2701 TCACGATAAC ACACTCATAT TAAACGCGAG TATTACTCAT CAGTCAACGT CACATTCCGC
2761 TCTGTCCACT TCGACCTTAC ACCTCTACTT GTCCAACTAT CTTCCACTTG TCAAGCGTCG
2821 ACATGAAAAA GCCTGAACTC ACCGCGACGT CTGTCGAGAA GTTTCTGATC GGAAAGTTCG
2881 ACAGCGTCTC CGACCTGATG CAGCTCTCGG AGGGCGAAGA ATCTCGTGCT TTCAGCTTCG
2941 ATGTAGGAGG GCGTGGATAT GTCCTGCGGG TAAATAGCTG CGCCGATGGT TTCTACAAAG
3001 ATCGTTATGT TTATCGGCAC TTTGCATCGG CCGCGCTCCC GATTCCGGAA GTGCTTGACA
3061 TTGGGGAGTT CAGCGAGAGC CTGACCTATT GCATCTCCCG CCGTGCACAG GGTGTCACGT
3121 TGCAAGACCT GCCTGAAACC GAACTGCCCG CTGTTCTTCA GCCGGTCGCG GAGGCCATGG
3181 ATGCGATCGC TGCGGCCGAT CTTAGCCAGA CGAGCGGGTT CGGCCCATTC GGACCGCAAG
3241 GAATCGGTCA ATACACTACA TGGCGTGATT TCATTTGCGC GATTGCTGAT CCCCATGTGT
3301 ATCACTGGCA AACTGTGATG GACGACACCG TCAGTGCGTC CGTCGCGCAG GCTCTCGATG
3361 AGCTGATGCT TTGGGCCGAG GACTGCCCCG AAGTCCGGCA CCTCGTGCAC GCGGATTTCG
3421 GCTCCAACAA TGTCCTGACG GACAATGGCC GCATAACAGC GGTCATTGAC TGGAGCGAGG
```

3481 CGATGTTCGG GGATTCCCAA TACGAGGTCG CCAACATCTT CTTCTGGAGG CCGTGGTTGG
3541 CTTGTATGGA GCAGCAGACG CGCTACTTCG AGCGGAGGCA TCCGGAGCTT GCAGGATCGC
3601 CGCGGCTCCG GGCGTATATG CTCCGCATTG GTCTTGACCA ACTCTATCAG AGCTTGGTTG
3661 ACGGCAATTT CGATGATGCA GCTTGGGCGC AGGGTCGATG CGACGCAATC GTCCGATCCG
3721 GAGCCGGGAC TGTCGGGCGT ACACAAATCG CCCGCAGAAG CGCGGCCGTC TGGACCGATG
3781 GCTGTGTAGA AGTACTCGCC GATAGTGGAA ACCGACGCCT CAGCACTCGT CCGAGGGCAA
3841 AGGAATAGGG ATCCAGCAGG CGGAGAAAGA GGGAAGTATA AGGCGGACGC ATAGGTCGGA
3901 GTATCGAGAA AAAGAGGCAG AGATGGGTGG TGGCGGAGCG CCCCTCTCTG CCCTTCTCTG
3961 TTTTACTGTT CCCGCCACGT CGCCAACTCC TCTTTTTGTT TGCTCTAATG TCGGTGCTAT
4021 CCTCTTCCTC TTTTTACTCT CCGTTTTTCT TCCTTCCGTT TTGTCTTGTT ATCACCGTTA
4081 TTTTTCTTCT TCTTTCTTAG CCGATTTGGG TCTCCTGCCT ACGGCAGCGG TGATGAGTCG
4141 CACTTCCGTC CCCCTTTCCT CTCCGTAAGT ACTCCCTCGA TGCCTCAGGC GCTTCTATTT
4201 TGCGGCACTG TGCTGACCAC CTCCCACGTG TGCAGTGAGA GCGCCAGAGA CATTCAGGAG
4261 AAGAGGGAAG AGGGGAAGTA AATACCAAAG CGAGGAAGAT GTCTTTCTCG CTGCTTCTCA
4321 TCCTGTTGAC CGGTGTGCAC GGCGGTGTGT CCTCTCGGCT TCCCTCCTCC TTCCCTCGCT
4381 CCCCTTTTTC TGTGTTTTTT CTTTCTAACA TGATTGCGCC TGCTCTTTTT GCCCCGCAGC
4441 GTCGCAGTGG GTGTCACCTG CGACCGCCGT ACGTTTTACT ATATACATAT TTGTTTCTGT
4501 GCTGATGCCA TTTGTCCGCT ACTCCATATC GTGCGTGTTT GTTTTCCAGT GGTGACGCCT
4561 TCCGCATCAT TTTAATCAAT TCGAGGCGAA AAGGGTGATT TGAGTGTTGG AGTGGGCTTT
4621 TTTATTTATT TGACTCCAAA CATCTTTTCT TTTCCCGGTA TCTTCGGGTC CGTCACGCAG
4681 GTCGGTGGGT GTGTCCGAGT GCGCTTCTTT ACTCACGTGC CGCTGCGCAC ATACACAGAT
4741 TTTAAAAGCA CGCATACACG ATCTGTGCCT TCAAAACTAA TCAACAAAAA CAAAAATACA
4801 AAAAAGACAA GAATAAGAGG TCAGCAACGC ACCCACGGCT CCTTTCTTCC TGATCCACGT
4861 CGTGGCCGCT GTACGCTCTC AAACACGCTT TGGCGCTGAT GCGTGCCTTT ATACAAAGAA
4921 CAAAGAAGAA GCGAAAAACG GGAGCGTGTT GCTTTGGGGA TGTATGCGTG CCACAGTTCT
4981 GCTCTAGAAA TTCTATGGAC GATTGTGGGC ATATTTTACG CACATCGCCC GCGCAGGCGC
5041 TGTTTTGCTA AAACTCGTGT CTGAGACAAG CAGCCAGCTG GTTCTACCGA GCCAGGTGGC
5101 GGGCAAGTCC CAGACACACA CCCAGGGACT TTTCTCTCTC TCTCTTCGCA TGGGCAACCT
5161 AGCGGAGAGC GAGGGGGGAA GTCACCGCAT TACAACCTAG GAAGGCGAGT CGAAACGGTG
5221 CGTGGATGCC GCGTTTGTGC CAACACAAGC TAAGGTCACA CCACGAACGC ACCGAGCCAG
5281 ACAAACCATA CCCCTATTCA CACAACTGTA TGTGACCTCA CACTCACACA CACAATTTAT
5341 TTGTGCGTGT GGCGTGCGTG TGTCACCACT GCGGTTTTGC CCCCAGTCTT ATTACTGCTT
5401 CTTTGTACCC CTGCTTCTTG ATTTAAATGA GCTCCAGCTT TTGTTCCCTT TAGTGAGGGT
5461 TAATTCCGAG CTTGGCGTAA TCATGGTCAT AGCTGTTTCC TGTGTGAAAT TGTTATCCGC
5521 TCACAATTCC ACACAACATA CGAGCCGGAA GCATAAAGTG TAAAGCCTGG GGTGCCTAAT
5581 GAGTGAGCTA ACTCACATTA ATTGCGTTGC GCTCACTGCC CGCTTTCCAG TCGGGAAACC
5641 TGTCGTGCCA GCTGCATTAA TGAATCGGCC AACGCGCGGG GAGAGGCGGT TTGCGTATTG
5701 GGCGCTCTTC CGCTTCCTCG CTCACTGACT CGCTGCGCTC GGTCGTTCGG CTGCGGCGAG
5761 CGGTATCAGC TCACTCAAAG GCGGTAATAC GGTTATCCAC AGAATCAGGG GATAACGCAG
5821 GAAAGAACAT GTGAGCAAAA GGCCAGCAAA AGGCCAGGAA CCGTAAAAAG GCCGCGTTGC
5881 TGGCGTTTTT CCATAGGCTC CGCCCCCCTG ACGAGCATCA CAAAAATCGA CGCTCAAGTC
5941 AGAGGTGGCG AAACCCGACA GGACTATAAA GATACCAGGC GTTTCCCCCT GGAAGCTCCC
6001 TCGTGCGCTC TCCTGTTCCG ACCCTGCCGC TTACCGGATA CCTGTCCGCC TTTCTCCCTT
6061 CGGGAAGCGT GGCGCTTTCT CATAGCTCAC GCTGTAGGTA TCTCAGTTCG GTGTAGGTCG
6121 TTCGCTCCAA GCTGGGCTGT GTGCACGAAC CCCCCGTTCA GCCCGACCGC TGCGCCTTAT
6181 CCGGTAACTA TCGTCTTGAG TCCAACCCGG TAAGACACGA CTTATCGCCA CTGGCAGCAG
6241 CCACTGGTAA CAGGATTAGC AGAGCGAGGT ATGTAGGCGG TGCTACAGAG TTCTTGAAGT
6301 GGTGGCCTAA CTACGGCTAC ACTAGAAGGA CAGTATTTGG TATCTGCGCT CTGCTGAAGC
6361 CAGTTACCTT CGGAAAAAGA GTTGGTAGCT CTTGATCCGG CAAACAAACC ACCGCTGGTA
6421 GCGGTGGTTT TTTTGTTTGC AAGCAGCAGA TTACGCGCAG AAAAAAAGGA TCTCAAGAAG
6481 ATCCTTTGAT CTTTTCTACG GGGTCTGACG CTCAGTGGAA CGAAAACTCA CGTTAAGGGA
6541 TTTTGGTCAT GAGATTATCA AAAAGGATCT TCACCTAGAT CCTTTTAAAT TAAAAATGAA
6601 GTTTTAAATC AATCTAAAGT ATATATGAGT AAACTTGGTC TGACAGTTAC CAATGCTTAA
6661 TCAGTGAGGC ACCTATCTCA GCGATCTGTC TATTTCGTTC ATCCATAGTT GCCTGACTCC
6721 CCGTCGTGTA GATAACTACG ATACGGGAGG GCTTACCATC TGGCCCCAGT GCTGCAATGA
6781 TACCGCGAGA CCCACGCTCA CCGGCTCCAG ATTTATCAGC AATAAACCAG CCAGCCGGAA
6841 GGGCCGAGCG CAGAAGTGGT CCTGCAACTT TATCCGCCTC CATCCAGTCT ATTAATTGTT
6901 GCCGGGAAGC TAGAGTAAGT AGTTCGCCAG TTAATAGTTT GCGCAACGTT GTTGCCATTG
6961 CTACAGGCAT CGTGGTGTCA CGCTCGTCGT TTGGTATGGC TTCATTCAGC TCCGGTTCCC
7021 AACGATCAAG GCGAGTTACA TGATCCCCCA TGTTGTGCAA AAAAGCGGTT AGCTCCTTCG

```
7081 GTCCTCCGAT CGTTGTCAGA AGTAAGTTGG CCGCAGTGTT ATCACTCATG GTTATGGCAG
7141 CACTGCATAA TTCTCTTACT GTCATGCCAT CCGTAAGATG CTTTTCTGTG ACTGGTGAGT
7201 ACTCAACCAA GTCATTCTGA GAATAGTGTA TGCGGCGACC GAGTTGCTCT TGCCCGGCGT
7261 CAATACGGGA TAATACCGCG CCACATAGCA GAACTTTAAA AGTGCTCATC ATTGGAAAAC
7321 GTTCTTCGGG GCGAAAACTC TCAAGGATCT TACCGCTGTT GAGATCCAGT TCGATGTAAC
7381 CCACTCGTGC ACCCAACTGA TCTTCAGCAT CTTTTACTTT CACCAGCGTT TCTGGGTGAG
7441 CAAAAACAGG AAGGCAAAAT GCCGCAAAAA AGGGAATAAG GGCGACACGG AAATGTTGAA
7501 TACTCATACT CTTCCTTTTT CAATATTATT GAAGCATTTA TCAGGGTTAT TGTCTCATGA
7561 GCGGATACAT ATTTGAATGT ATTTAGAAAA ATAAACAAAT AGGGGTTCCG CGCACATTTC
7621 CCCGAAAAGT GCCACCTAAA TTGTAAGCGT TAATATTTTG TTAAAATTCG CGTTAAATTT
7681 TTGTTAAATC AGCTCATTTT TTAACCAATA GGCCGAAATC GGCAAAATCC CTTATAAATC
7741 AAAAGAATAG ACCGAGATAG GGTTGAGTGT TGTTCCAGTT TGGAACAAGA GTCCACTATT
7801 AAAGAACGTG GACTCCAACG TCAAAGGGCG AAAAACCGTC TATCAGGGCG ATGGCCCACT
7861 ACGTGAACCA TCACCCTAAT CAAGTTTTTT GGGGTCGAGG TGCCGTAAAG CACTAAATCG
7921 GAACCCTAAA GGGAGCCCCC GATCTAGAGC TTGACGGGGA AAGCCATC
```

Appendix 5. PIPS reflective statement

Note to examiners:

This statement is included as an appendix to the thesis in order that the thesis accurately captures the PhD training experienced by the candidate as a BBSRC Doctoral Training Partnership student.

The Professional Internship for PhD Students is a compulsory 3-month placement which must be undertaken by DTP students. It is usually centred on a specific project and must not be related to the PhD project. This reflective statement is designed to capture the skills development which has taken place during the student's placement and the impact on their career plans it has had.

PIPS Reflective Statement

Between October and December 2018, I undertook a PIP placement in London at a medical communications agency – DDB Remedy. Here I worked in the copy team, where I participated in client projects, developing my writing and editing skills. I also contributed to pitches for new business through idea generation and researching background information on different diseases.

The placement at DDB taught me about the inner workings of a copy department, providing insight into a career in medical communications. I found the placement enjoyable, and it helped me develop useful skills that were beneficial to the remainder of my PhD and will continue to be in my future career. Whilst working at DDB did not provide me with a definitive career path, I would consider a role in scientific writing in the future.