THE UNIVERSITY *of* EDINBURGH

Translational control and the escape from translational arrest in stumpy form *Trypanosoma brucei*

Stephanie Monk

Thesis submitted to the University of Edinburgh for the degree of Doctor of Philosophy

2012

Table of contents

List of Figures

List of Tables

Abstract

The transmission of *Trypanosoma brucei*, the causative agent of human African trypanosomiasis, depends upon the development in the bloodstream of 'stumpy forms' from non-transmissible 'slender forms'. In stumpy forms many mRNAs are downregulated and translation is generally repressed. However, a small subset of genes escape this repression and are upregulated, presumably as an adaptation for transmission. To understand the basic of this, regulatory sequences within the 3'UTR of a major stumpy-enriched transcript (an ESAG9 gene) have been characterised. This identified a signal responsible for gene silencing in slender forms and gene activation when cells develop to stumpy forms.

An investigation was made of upstream open reading frames (uORFs) as a mechanism for the control of stumpy form gene expression. No evidence was found of uORF control, but one gene investigated was found to produce two transcripts through *trans*-splicing at different sites. These transcripts, which were found to exhibit some differential abundance between life-cycle stages, would generate a long and short form (from an internal ATG) of the encoded protein. Both are predicted to contain a UBA/TS-N (ubiquitin associated) domain, however, the longer form of the protein is also predicted to contain a transmembrane helix and cleavable signal peptide, suggesting a different localisation. However, ectopic expression of either protein form with a Ty epitope tag resulted in the same protein localisation.

Additionally, the transcripts of two translational protein homologues, TbeIF4E4 and TbeIF6, were identified as upregulated in stumpy forms. Radiolabelled-methionine experiments and polysome analysis showed that overexpression or RNAi-mediated ablation of TbeIF6 resulted in a decrease in protein synthesis and decrease in translation. Unlike its archaeal homologue, TbeIF6 protein was not induced by cold-shock treatment.

Finally, to identify which transcripts escape translational repression in stumpy forms an analysis was made of polysome-associated transcripts by RNA-sequencing. This identified potentially interesting genes for further investigation, and showed that

many procyclic-enriched transcripts were also enriched in stumpy form polysome-associated RNA, confirming these cells as preadapted for transmission. Together, this work has characterised a 3'UTR regulatory element in a stumpy-enriched transcript, examined alternative trans-splicing of another transcript, investigated two translational protein homologues and identified transcripts that escape translational repression in the transmissible life-cycle stage of *T. brucei*.

Declaration

I declare that all of the work presented in this thesis is my own, unless otherwise acknowledged. The work presented in this thesis has not been submitted for any other degree or professional qualification.

Stephanie Monk
May 2012

Acknowledgements

There are many people who I would like to thank for their help and support throughout my PhD. It has been greatly appreciated and this would not have been possible without them.

Firstly, I would like to thank Keith for all his support and guidance throughout my PhD. Especially for always giving me his time to answer my questions, for giving me interesting and exciting projects, and I am very grateful for his belief and investment in me.

I would also like to thank the other members of the Matthews lab, both past and present, for all their invaluable help and advice. Peg for help with northern blots and sucrose gradients; Melanie for help with northern blots and sucrose gradients also, as well as sharing thesis writing frustrations; Katelyn for help with bioinformatics; Irene for help with western blots; Paul for analysis of the Illumina digital tag sequencing data; and many others for help. Special thanks go to Paula for all her help and support, and for being a truly great friend. Thank you to you all also for making the lab such an enjoyable place to have spent the last four years.

Also, thanks to Dr Terry Smith for performing the radiolabelling experiments and welcoming us to St. Andrews; Pankaj Barua for the Perl script; Martin Waterfall for his help and expertise with the flow cytometry; and Keith Matthews, Deborah Hall and Julie Wilson for performing all the animal work.

Thank you too to everyone who helped make my time in Edinburgh so enjoyable; Paula, Yvonne, Ruth and Antoine for many fun lunchtime chats, and to my friends outside the lab as well.

Finally, I would like to thank my family, especially my Mum, for their continual love and support, and for always believing in me.

Abbreviations

8-pCPT-2'-O-Me-cAMP	8- (4- Chlorophenylthio)- 2'- O- methyladenosine- 3', 5'- cyclic monophosphate
ACO	Aconitase
BLAST	Basic Local Alignment Search Tool
bp	Base pair
BSF	Bloodstream form
CA	*Cis*-aconitate
cAMP	Cyclic adenosine monophosphate
CAT	Chloroamphenical acetyl transferase
cDNA	Complementary DNA
CNS	Central nervous system
COX	Cytochrome oxidase
DALY	Disability-adjusted life year
DAPI	4,6-diamidine-2,phenylindole
DHS	Deoxyhypusine synthase
DIG	Digoxygenin
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
EATRO	East African Trypanosomiasis Research Organisation
eIF	Eukaryotic translation initiation factor
ELISA	Enzyme-linked immunosorbent assay
EP	EP-procyclin
ES	Expression site
ESAG	Expression site associated gene
ESB	Expression site body
EtBr	Ethidium bromide
FCS	Foetal calf serum
FITC	Fluorescein isothiocyanate
GDP	Guanosine diphoshate
GO	Gene ontology
GRESAG	Genes related to ESAG
GTP	Guanosine triphosphate
HAT	Human African trypanosomiasis
HRP	Horse radish peroxidase
IgG	Immunoglobulin G
IRE	Iron response element
IRP	Iron regulatory protein
kDNA	Kinetoplast DNA
LB	Luria Bertani
MAP	Mitogen-activated protein
MAPK	MAP kinase
Met	Methionine
miRNA	microRNA
mRNA	Messenger RNA
nt	Nucleotide
ORF	Open reading frame

PAD	Protein(s) associated with differentiation
PAS	Polyadenylation site
P-body	Processing body
PBS	Phosphate-buffered saline
PCF	Procyclic form
PCR	Polymerase chain reaction
PDB	Protein database
PIC	Preinitiation complex
Pol	RNA polymerase
PSG	Phosphate-Saline-Glucose
RACE	Rapid Amplification of cDNA ends
RF	Release factor
RISC	RNA-induced silencing complex
RNA	Ribonucleic acid
RNAi	RNA interference
rRNA	Ribosomal RNA
RSAT	Regulatory sequence analysis tool
SAS	Splice acceptor site
SDS-PAGE	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
SIF	Stumpy induction factor
SL	Splice leader
TAE	Tris-acetate-EDTA
TC	Ternary complex
TCA	Trichloroacetic acid
Tet	Tetracycline
UBA/TS-N	Ubiquitin-associated
uORF	Upstream open reading frame
UTR	Untranslated region
VSG	Variable surface glycoprotein
ZFP	Zinc finger protein

Chapter 1

Introduction

1.1 Trypanosomiasis

Trypanosoma brucei species are unicellular protozoan parasites belonging to Phylum Euglenozoa, Order Kinetoplastida, and Family Trypanosomatidae (Vickerman, 1994). Trypanosomes are extremely ancient protists, likely diverged from the eukaryotic lineage soon after the early eukaryote cell had acquired the mitochondrion through endosymbiosis (Gray, 1992).

The trypanosomatids include, among others, *T. brucei*, *T. cruzi* and *Leishmania* species, which also comprise the group termed the 'TriTryps'. All of these cause neglected tropical disease; human African trypanosomiasis (HAT) and Nagana in the case of *T. brucei*, Chagas disease, also known as American trypanosomiasis, in the case of *T. cruzi*, and the spectrum of Leishmaniasis disease in the case of the *Leishmania* species.

Chagas disease is found mainly in Latin America and human infection principally occurs by transmission of *T. cruzi* through the faeces of triatomine bugs. It is estimated that 10 million people worldwide are infected with *T. cruzi* (http://www.who.int/mediacentre/factsheets/fs340/en/index.html). Leishmaniasis is a spectrum of diseases covering cutaneous, mucocutaneous, and the most serious, visceral forms of the disease. *Leishmania* parasites are transmitted to human hosts by the bite of a sandfly, which then proliferate in the phagolysosome of macrophages after phagocytosis.

The studies presented in this thesis concern *T. brucei*, the causative agent of HAT and Nagana in cattle.

1.1.1 Impact of African trypanosomiasis

HAT, also known as African sleeping sickness, is caused by infection with the *T. brucei* subspecies, *T. b. gambiense* and *T. b. rhodesiense*. Recent estimates puts the prevalence of the disease between 50 000-70 000 cases (Brun *et al.*, 2009), and disease is limited to sub-Saharan Africa, due to the distribution of the parasite vector, the tsetse fly (see Figure 1.1).

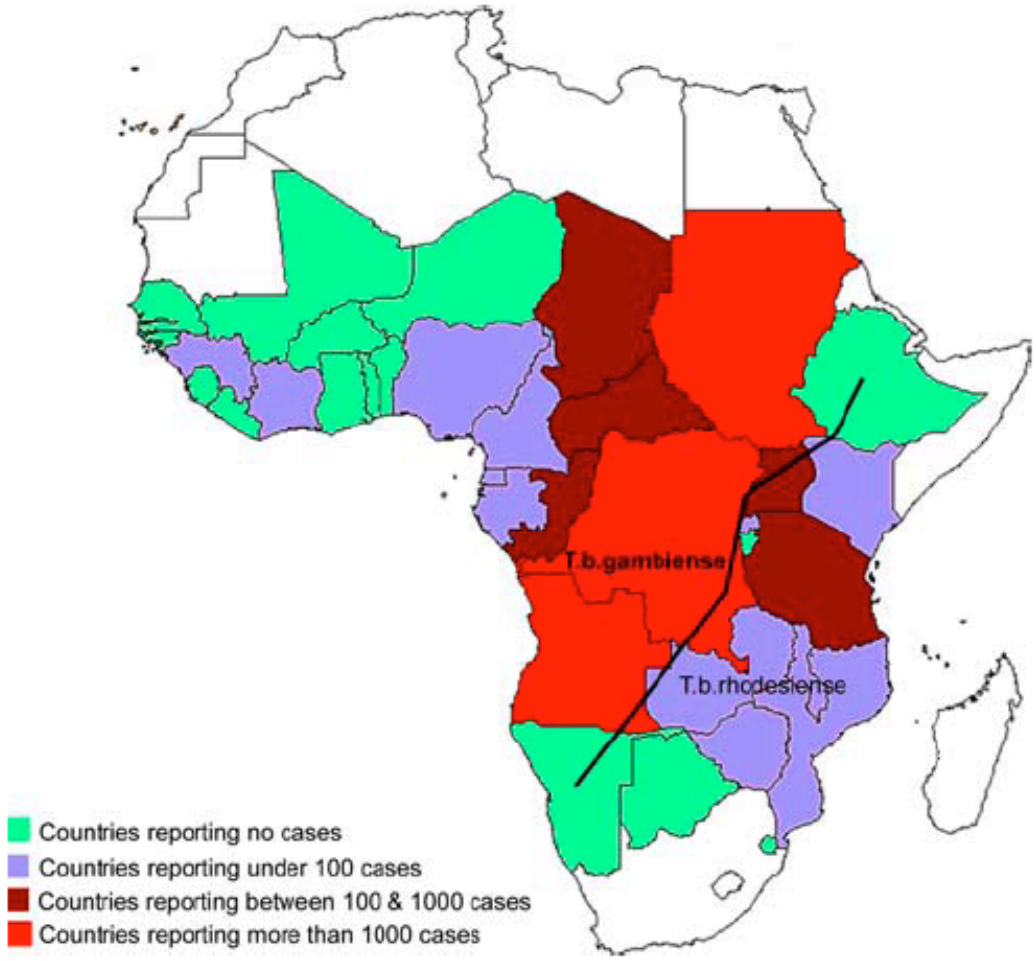


Figure 1.1 Map of Africa with the epidemiological status of HAT in countries where the disease is considered endemic. Figure taken from Simarro *et al*. (2008).

Infection with *T. b. gambiense* is found in west and central Africa, where it causes a chronic infection accounting for more than 90% of disease cases, whereas *T. b. rhodesiense* infection is found in eastern and southern Africa, causing an acute infection accounting for less than 10% of cases (Simarro *et al.*, 2008).

The disease mostly affects rural areas, where it causes a great economic burden. A measure of disease burden, disability-adjusted life year (DALY) was estimated in 2002 at 1.54 million DALYs lost to HAT (Fevre *et al.*, 2008). The individual household economic burden of a case of HAT is high, including the costs of care at home and at hospitals, of seeking diagnosis, medical fees, and transport costs, as well as income lost by the infected individual and their carers (reviewed in (Boelaert *et*

al., 2010). This is also the case even when diagnostics and drugs are provided free of charge and causes a delay in the uptake of treatment. Also having a great economic impact is the infection of cattle with the subspecies *T. b. brucei*, which is non-pathogenic to man, but limits agricultural production, food resources and the economic growth of the affected countries (Simarro *et al.*, 2008).

1.1.2 Human African Trypanosomiasis (HAT) disease

HAT is characterised by two stages of disease progression. At the initial, haemalymphatic, stage parasites are present in the bloodstream and lymphatic system. The symptoms are wide ranging and include intermittent fever, splenomegaly, malaise, adenopathies, cardiovascular manifestations, and pruritus (Vincendeau and Bouteille, 2006). One specific symptom that is characteristic of infection is swollen cervical lymph nodes.

The second, meningoencephalitic, stage occurs when the parasite enters the central nervous system (CNS) by crossing the blood-brain barrier. At this stage neurological symptoms develop gradually, including the characteristic alteration of the circadian sleep/wake cycle, resulting in daytime somnolence and nocturnal insomnia, that gives the disease its name, sleeping sickness. The general signs of the first stage remain and other symptoms can include sensory disturbances, psychiatric disorders such as confusion or aggression, and tremor (Vincendeau and Bouteille, 2006). Ultimately, if left untreated, the infection results in coma and death.

Chronic infection, caused by *T. b. gambiense* infection, results in an infection without the appearance of symptoms for months or years, often until the advanced disease stage, in comparison to the acute infection, caused by *T. b. rhodesiense*, where symptoms of the first stage are apparent after a few weeks or months (Simarro *et al.*, 2008).

Within disease control programmes, a three step method is used for diagnosis, based on screening, diagnostic confirmation, and staging (Brun *et al.*, 2009). For the screening of populations at risk of *T. b. gambiense* infection, the card agglutination

test for trypansomiasis (CATT) is used widely. CATT is a serological test that is fast, practical and reported to be 87-98% sensitive and 93-95% specific (Brun *et al.*, 2009). However, due to incomplete sensitivity, microscopic examination of blood or lymph samples to detect the presence of parasites is required, especially for those individuals with suspected infections but negative CATT results. Staging of the disease is then required, because HAT treatment differs for the first and second stages. Lumber puncture is required to assess cerebrospinal fluid for the presence of more than 5 white blood cells/µl, parasites, or elevated protein levels (>370 mg/L), which confirms disease progression to the second stage (Brun *et al.*, 2009).

1.1.3 HAT treatment

There are four drugs registered to treat HAT, these are suramin, pentamidine, melarsoprol, and eflornithine. Selection of drug for treatment is based on the disease stage and subspecies of trypanosome responsible for infection (see Table 1.1), each drug suffers from difficulty in administration and often toxicity. Indeed, melarsoprol, an arsenical, results in the death of up to 5% of those treated with the drug, due to the related toxicity (Barrett, 2010).

Table 1.1 The four drug treatments in use to treat HAT and their main adverse reactions (table taken from (Brun *et al.*, 2009).

	Stage	Route of application	Dosing	Main adverse drug reactions
Trypanosoma brucei gambiense				
Pentamidine*	First	Intramuscular	4 mg/kg bodyweight at 24 h intervals for 7 days	Hypoglycaemia, injection site pain, diarrhoea, nausea, vomiting
Eflornithine	Second	Intravenous Infusion of >30 min	100 mg/kg bodyweight at 6 h intervals for 14 days	Diarrhoea, nausea, vomiting, convulsions; anaemia, leucopenia, and thrombocytopenia
Melarsoprol†	Second	Intravenous	2·2 mg/kg bodyweight at 24 h intervals for 10 days	Encephalopathic syndromes, skin reactions (pruritus, maculopapular eruptions), peripheral motoric (palsy) or sensorial (paraesthesia) neuropathies, thrombophlebitis
Trypanosoma brucei rhodesiense				
Suramin*	First	Intravenous	Test dose of 4–5 mg/kg bodyweight at day 1, then five injections of 20 mg/kg bodyweight every 7 days (eg, day 3, 10, 17, 24, 31); maximum dose per injection 1 g	Hypersensitivity reactions (acute, late); albuminuria, cylinduria, haematuria, peripheral neuropathy
Melarsoprol*	Second	Intravenous	Three series of 3·6, 3·6, 3·6 mg/kg bodyweight, the series spaced by intervals of 7 days; maximum dose per day 180 mg	Encephalopathic syndromes, skin reactions (pruritus, maculopapular eruptions), peripheral motoric (palsy) or sensorial (paraesthesia) neuropathies, thrombophlebitis

*Endemic countries: according to national legislature or guidelines. †Only where eflornithine is not available or where melarsoprol is first-line treatment according to national guidelines.

The administration and toxicity drug problems, together with evidence of drug resistance, means that there is a desperate need for the development of new drugs.

However, the recent entry of fexinidazole and other compounds into preclinical or clinical trials offers hope for new HAT treatments (Barrett, 2010). Additionally, the recent development of an genome-scale RNA interference (RNAi) target sequencing (RIT-seq) screen approach to identifying the molecular components contributing to current drug efficacy and resistance should provide a powerful tool to assist the rational design of new drugs and combat of drug resistance (Alsford *et al.*, 2012).

1.2 Trypanosomes

As well as causing disease of medical importance, trypanosomes are interesting to study due to unique aspects of their cell biology. For example, RNA editing was first discovered in trypanosomes (Benne *et al.*, 1986), where uridines are inserted and deleted in mitochondrial transcripts. Subsequently, different types of RNA editing, such as adenosine-to-inosine RNA editing, have been discovered in various other organisms. Additionally, trypanosomes use RNA polymerase I (Pol I) for transcription of some protein-coding genes and polycistronic gene expression, these features being unique or unusual for a eukaryote. Linked to this, trypanosomes greatly rely on the post-transcriptional regulation of gene expression (discussed in Section 1.7) making these organisms particular suitable to study this type of gene regulation. Being an evolutionarily early-branching eukaryote, unicellular, with a published genome sequence, availability of RNAi as a tool to study gene function, and easy to culture *in vitro*, also make *T. brucei* a particularly useful system for scientific study.

1.2.1 Lifecycle of T. brucei

T. brucei exhibits a digenetic lifecycle, taking place in the mammalian host and the arthropod vector, the tsetse fly. The parasite alternates between proliferative and non-proliferative states and undergoes great morphological and metabolic changes (Vickerman, 1985), in addition to major changes in gene expression (Clayton, 2002), in order to adapt to the different environments. The developmental cycle of *T. brucei* is depicted schematically in Figure 1.2.

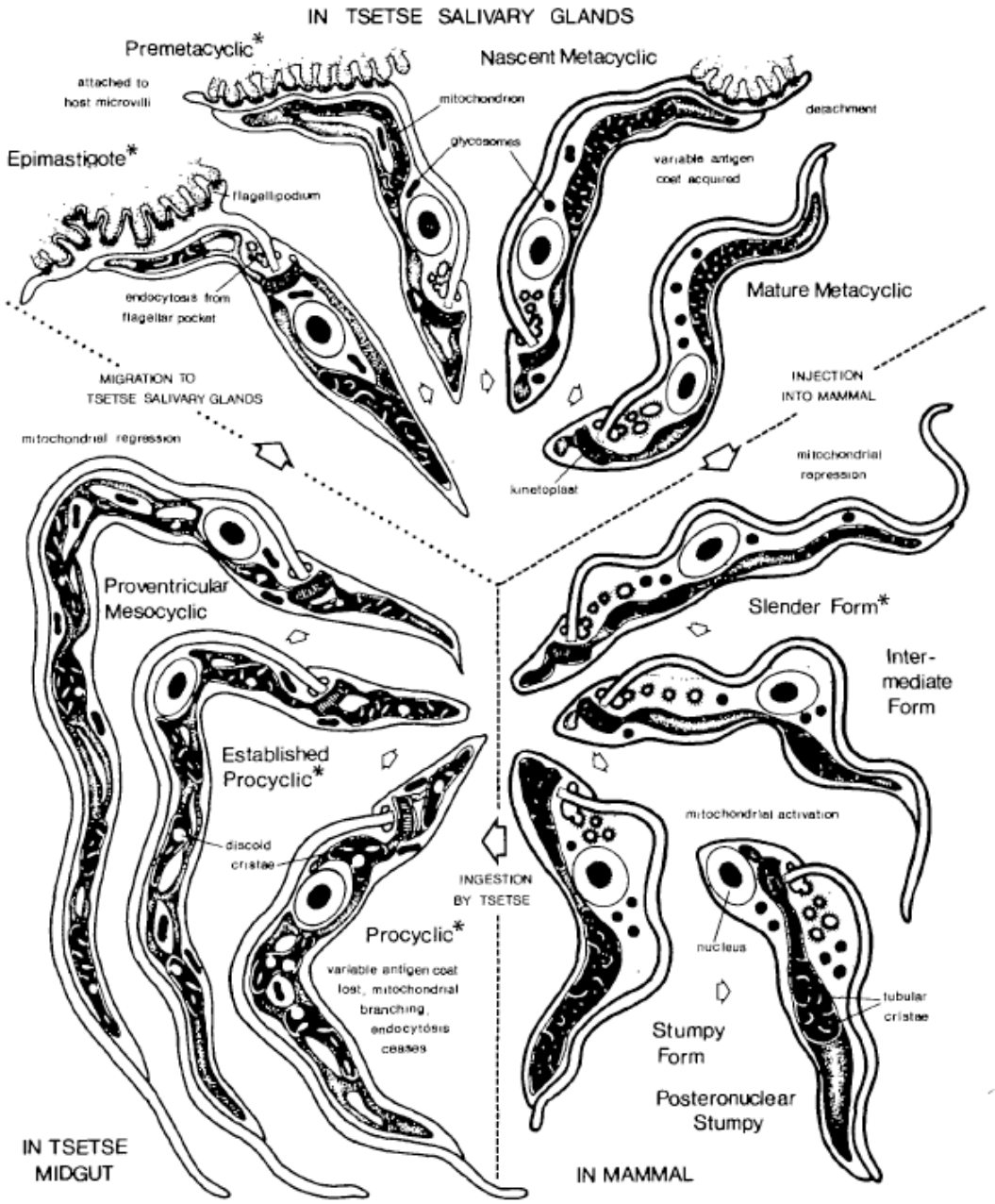


Figure 1.2 Schematic diagram depicting the developmental cycle of *T. brucei* in the mammalian host and tsetse fly vector, with changes in cell surface, mitochondrion, glycosomes, receptor mediated endocytosis, and relative size of developmental stage shown. However, it should be noted that this schematic diagram is missing detail on the long and short epimastigote forms that are present in the tsetse fly (Van Den Abbeele *et al.*, 1999). Asterisks indicate proliferative stages. Figure taken from Vickerman (1985).

Mammalian infection occurs through the bite of a tsetse fly (*Glossina* species) harbouring *T. brucei* parasites during a bloodmeal, and deposition of infective metacyclic form trypanosomes in the dermal connective tissue from the fly's salivary

glands (Vickerman, 1985). From here trypanosomes enter the lymphatic and blood systems and, as mentioned above, at a later stage of the disease they cross the blood-brain barrier to enter the CNS. Trypanosomes transform into slender forms that express variant surface glycoproteins (VSG), a glycoprotein that coats the surface of the mammalian form trypanosomes, and also the metacyclic form in preparation for transmission to the mammalian host. Slender forms are the proliferative stage in the mammal, and multiplication occurs by binary fission. These forms possess a simple mitochondrion that is devoid of cristae as the parasite uses the abundant blood glucose as an energy source for glycolysis (Ryley, 1962).

At the peak of parasitaemia, a proportion of the slender form population irreversibly differentiate into stumpy forms in response to a density-dependent parasite-derived signal, via an intermediate form. Stumpy forms, which have a shortened, broader cellular morphology, are non-proliferative, cell cycle arrested in G1 or G0 (Shapiro *et al.*, 1984), show increased mitochondrial activity (Priest and Hajduk, 1994) and are preadapted for transmission to the tsetse fly. Stumpy forms represent an important stage in the parasite lifecycle as their division-arrest enables the control of parasite numbers, therefore prolonging host survival and transmission probability, and also allows the coordination of morphological changes with cell cycle re-entry upon tsetse transmission (Matthews, 2005).

Upon another tsetse fly bloodmeal, bloodstream trypanosomes are ingested by the tsetse fly. Here, while slender forms die or differentiate into stumpy forms in the anterior midgut, stumpy forms differentiate into procyclic forms (Vickerman, 1985). The VSG coat is lost and replaced with the insect form surface coat, procyclin (Roditi *et al.*, 1989). The procyclic form is again proliferative and changed morphologically to become long and thinner. Additionally, the mitochondrion develops cristae and possesses functional cytochrome-mediated electron transport and all citric acid cycle enzymes are present as the preferred energy source shifts to proline (Priest and Hajduk, 1994). To complete the lifecycle back to the infective metacyclic form, the parasites must migrate to the salivary glands and pass through a number of different developmental stages (see Figure 1.2). These developmental

stages also include long and short epimastigote forms (Van Den Abbeele *et al.*, 1999) which are not depicted in Figure 1.2. Additionally, although genetic exchange has long been demonstrated to take place in *T. brucei* (Jenni *et al.*, 1986), a recent report now provides evidence for meiosis occurring in the tsetse salivary gland within a newly described meiotic developmental lifecycle stage (Peacock *et al.*, 2011).

The phenomenon of antigenic variation in bloodstream form trypanosomes is particularly important to the trypanosome lifecycle, as this process enables the parasites to escape the host immune system and thus survive in the mammalian bloodstream. Antigenic variation is achieved through the expression of a dense variant surface glycoprotein (VSG) coat on the trypanosome surface membrane, the molecular identity of which they change periodically to avoid clearance of the parasite by host antibody-mediated killing (reviewed in (McCulloch, 2004). To obtain the expression of a single VSG protein at any one time in a trypanosome cell, only one of about 20 telomeric VSG expression sites (ES) is transcribed at a time from a subnuclear body, the expression site body (ESB) (Navarro and Gull, 2001); reviewed in (McCulloch, 2004). Within the bloodstream expression sites (BES), expression site-associated genes (ESAGs) are co-transcribed with the *VSG* (Hertz-Fowler *et al*., 2008). The control of transcriptional switching between VSG genes is probably by epigenetic regulation, in part by a histone methyltransferase, DOTB1 (Figueiredo *et al.*, 2008).

1.2.2 Cell biology

The trypanosome cell has a strictly organised cell structure containing precisely positioned single copy organelles, these being the kinetoplast, mitochondrion, nucleus, flagellum and flagellar pocket (Matthews, 2005). The cell biology exhibits many developmental changes throughout the trypanosome lifecycle (reviewed in (Matthews, 2005). The cell structure is defined by the microtubule cytoskeleton, producing an elongated cell shape (see Figure 1.3). The flagellum is responsible for the motility of the trypanosome and extends along the extracellular surface posterior to anterior, originating intracellularly from a basal body that is connected to the

kinetoplast (reviewed in (Vaughan and Gull, 2003). The site where the flagellum originates from, termed the flagellar pocket, is a plasma membrane invagination and the only site of exo- and endocytosis (Overath and Engstler, 2004).

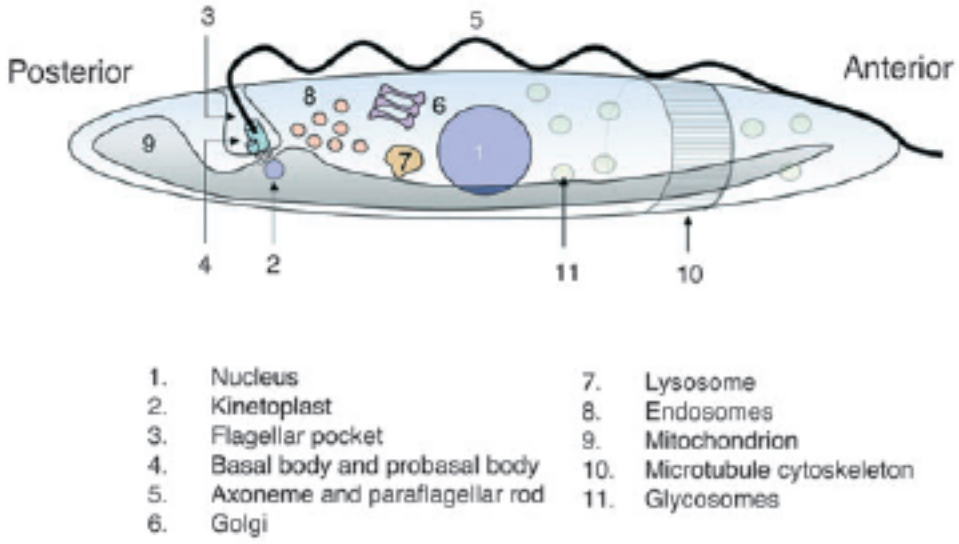


Figure 1.3 *T. brucei* cell ultrastructure. A simplified representation of the major trypanosome cell structural feature locations. Figure taken from Matthews, 2005.

Trypanosomes possess two DNA-containing structures, the kinetoplast, which holds the mitochondrial DNA and the nucleus, which holds the nuclear DNA. During the cell cycle both must be replicated and thus the cell cycle is tractable by the number of these two structures present in the cell (Sherwin and Gull, 1989). Cells in G1 or G0 have one kinetoplast and one nucleus (1K1N), cells in G2/M possess a 2K1N configuration, and post-mitotic cells have a 2K2N configuration (Sherwin and Gull, 1989). Thus, analysis of cells using a DNA stain, such as 4,6-diamidino-2-phenylindole (DAPI), allows the cell cycle stage to be determined. During the cell cycle, each organelle must be replicated as well and this is also performed in a well-ordered manner (Woodward and Gull, 1990; Bangs, 2011). For example, the emergence from the flagellar pocket of the short daughter flagellum during flagellum replication is one of the earliest events in cell duplication. In the procyclic form, the growth of the daughter flagellum tracks along the length of the old flagellum guided by a special connective structure, the flagellar connector, at the distal tip of the daughter flagellum (Moreira-Leite *et al.*, 2001).

1.3 Trypanosome differentiation

1.3.1 Slender to stumpy differentiation

As mentioned above, slender forms differentiate into stumpy forms in response to a density-dependent parasite-derived signal. This signal is termed stumpy induction factor (SIF) and is currently unidentified. Although the identity and signalling pathway are unknown, some details have been determined.

It has been shown that slender to stumpy differentiation can proceed independently of host cues and involves a density dependent mechanism, and additionally that SIF is of low relative molecular mass and released by trypanosomes (Vassella *et al.*, 1997). The cyclic AMP (cAMP) signalling pathway has been implicated in the transduction of the SIF signal, due to the findings that treatment with a membrane-permeable cAMP derivative or the phosphodiesterase inhibitor, etazolate, results in induction of stumpy morphology and growth arrest similarly to SIF, and also that SIF results in an increase in intracellular cAMP (Vassella *et al.*, 1997).

The membrane permeable cAMP analog used in that study, 8-(4-chlorophenylthio)-cAMP (pCPTcAMP), was also shown to cause the differentiation to stumpy forms of monomorphic culture-adapted bloodstream forms (Breidbach *et al.*, 2002). Monomorphic strains are so called because they have lost the ability to differentiate to stumpy forms in the host, as opposed to pleomorphic strains that have not (Matthews *et al.*, 2004). This study therefore indicates that in monomorphic strains this inability may result from a failure in the SIF signalling pathway rather than in the differentiation process itself (Breidbach *et al.*, 2002).

Additionally, it has been shown that slender forms differentiate into 'stumpy-like' forms through treatment with a hydrolysable cAMP analog, 8-pCPT-2'-O-Me-cAMP, which also has an antiproliferative effect, neither of which were observed with hydrolysis-resistant, membrane permeable cAMP analogs (Laxman *et al.*, 2006). This suggested that it is the metabolic products derived from the membrane permeable cAMP analogs that cause differentiation to stumpy-like forms rather than

the analogs themselves, and therefore, transformation may not be mediated directly by cAMP (Laxman *et al.*, 2006).

Other chemical treatments can also generate stumpy-like forms. For example, the transformation of slender forms to intermediate and stumpy forms has been reported with difluoromethylornithine (DFMO) treatment, which causes depletion of intracellular polyamines (Giffin and McCann, 1989). Also, the cysteine proteinase inhibitor, Z-Phe-Ala-CHN_2, caused the induction of stumpy-like forms, which was proposed to be triggered by prevention of cell cycle progression (Scory *et al.*, 2007). Finally, treatment with troglitazone treatment reportedly caused a similar differentiation to stumpy-like forms, although cell cycle arrest in G1/G0 was not observed (Denninger *et al.*, 2007).

Signalling by two protein kinases has also been implicated in slender to stumpy differentiation. Specifically, depletion of the MAP kinase 5 (MAPK5) resulted in increased sensitivity to SIF (Domenicali Pfister *et al.*, 2006) and deletion of the ZFK kinase also resulted in an increased rate of slender to stumpy form differentiation (Vassella *et al.*, 2001).

1.3.2 Stumpy to procyclic differentiation

It has been shown that monomorphic slender forms can be induced to differentiate to procyclic forms *in vitro* when stimulated with two triggers, a decrease in temperature from 37°C to 27°C and the addition of *cis*-aconitate at a high concentration (Czichos *et al.*, 1986). Subsequently, it was determined that a further cold shock of 20°C confers increased sensitivity to micromolar concentrations of *cis*-aconitate and citrate that is more similar to the concentration likely encountered by the trypanosome in the tsetse bloodmeal environment (Engstler and Boshart, 2004). It was also proposed that the cold shock induces the expression of a cell-surface receptor for citrate/*cis*-aconitate, this being dependent on previous cell cycle arrest and differentiation to stumpy form cells (Engstler and Boshart, 2004).

Recently, the PAD (proteins associated with differentiation) family of carboxylate-transporter proteins were shown to be required for perception of the citrate/*cis*-aconitate differentiation signal (Dean *et al.*, 2009). One member of the family, PAD1, was enriched in stumpy forms, and another, PAD2, was enriched in procyclic forms with cold shock causing upregulation and relocation of the protein to the cell surface in stumpy forms (Dean *et al.*, 2009), consistent with the predicted properties of the putative cell-surface receptor for citrate/*cis*-aconitate (Engstler and Boshart, 2004).

Also important in the signalling pathway are two phosphatases, TbPTP1 and TbPIP39. TbPTP1 acts as a molecular brake in bloodsteam forms, inhibiting differentiation to procyclic forms (Szoor *et al.*, 2006). Downstream of this, TbPIP39 is negatively regulated by TbPTP1, but becomes phosphorylated and activated during differentiation, enhancing stumpy to procyclic form differentiation (Szoor *et al.*, 2010). Additionally, many changes in gene expression also accompany development to procyclic forms, which are discussed in more detail in Section 1.7.2.

1.4 Gene expression in stumpy forms

Many trypanosome genes have been found to exhibit differential gene expression in different lifecycle stages, these genes being identified individually or by transcriptomic analysis. Of particular importance to this project are genes that have been identified as enriched in stumpy forms. A number of microarray analyses (Kabani *et al.*, 2009; Jensen *et al.*, 2009; Queiroz *et al.*, 2009), as well as an RNA sequencing study (Nilsson *et al.*, 2010), have been performed that included RNA extracted from stumpy forms and identified stumpy-enriched, as well as other lifecycle stage-enriched, transcripts. These types of analyses have resulted in the identification of many genes that would be interesting for further study. This is particularly true for the stumpy form lifecycle stage, because as discussed below, these cells show translational repression, and so any transcripts upregulated here are likely to be important for stumpy form biology.

1.4.1 Translational repression in stumpy forms

Between the well-studied stages, the proliferative slender and procyclic forms and the non-proliferative stumpy form, there are major changes in protein translation. Stumpy forms show a dramatically reduced level of protein synthesis with respect to slender or procyclic forms, as revealed by analysis of newly synthesised proteins with radiolabelled methionine incorporation (Brecht and Parsons, 1998). A decrease in protein synthesis was also observed in cells induced to stumpy-like forms with difluoromethylornithine treatment (Bitonti *et al.*, 1988). Brecht and Parsons (1998) showed that there was no difference in the ability of transcripts from the different lifecycle stages to be translated *in vitro* using a rabbit reticulocyte translation system. Therefore, the transcripts themselves were not responsible for the decreased protein synthesis, although, importantly the authors noted that there may be changes in the interaction of the mRNA cap with the trypanosome cap binding protein in stumpy forms that were undetectable in the *in vitro* system due to the unique trypanosome cap 4 structure (Bangs *et al.*, 1992). The decrease in protein synthesis in stumpy forms was also not attributable to differing poly(A) tail lengths, as these showed no significant variation in length between the lifecycle stages. Through the analysis of polysome profiles, a well-established method for determining global translation, it was observed that stumpy forms exhibited a greatly reduced level of polysome-associated transcripts (Brecht and Parsons, 1998). Thus, the reduced protein synthesis in this lifecycle stage is due to a repression of translation.

While there is a decrease in the levels of protein synthesis in stumpy forms compared to slender forms, two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) of ^{35}S-labelled proteins showed that the majority of those synthesised are the same, with only some being stage-specific (Shapiro and Kimmel, 1987). As stumpy forms are preadapted for transmission to the tsetse fly, presumably they must express proteins that enable this preadaptation. Therefore, although translation is repressed in stumpy forms, a small subset of transcripts must escape this repression to generate proteins enriched in stumpy forms.

1.4.2 Stumpy form enriched proteins

Some important genes have been found to exhibit stumpy form specific expression. For example, the PAD1 protein mentioned above, is specifically expressed in stumpy forms and is also enriched in this lifecycle stage at the level of the mRNA (Dean *et al.*, 2009).

Additionally, the expression-site-associated gene (ESAG) 9 family mRNAs are strongly enriched in stumpy forms (Barnwell *et al.*, 2010; Jensen *et al.*, 2009); see Figure 1.4). These proteins were found to be secreted by stumpy form trypanosomes and thus are implicated in infection chronicity or parasite transmission (Barnwell *et al.*, 2010). In that study, one member of this family was determined to be present in two expression sites of *T. brucei* EATRO 2340 and was also named 'ESAG9-EQ' because it was the gene most similar to ESAG9 genes in *T. equiperdum* (a trypanosome species which parasitizes horses) (Florent *et al.*, 1991).

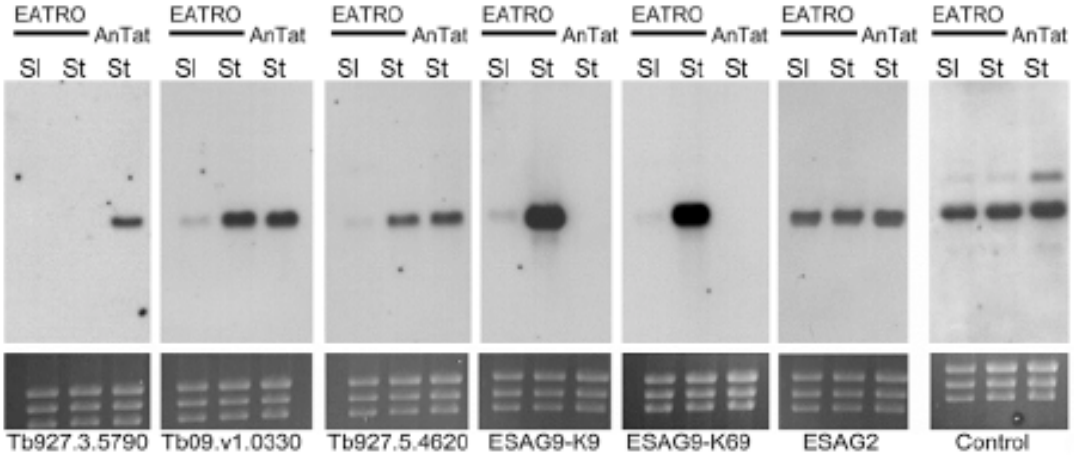


Figure 1.4 Expression of five *ESAG9* genes in monomorphic slender (EATRO Sl) and pleomorphic stumpy forms of *T. brucei* EATRO 2340 GUTat 7.2 (EATRO St) and pleomorphic stumpy forms of *T. b. brucei* AnTat1.1 (AnTat St). As controls, an ESAG2-specific probe and a constitutively expressed transcript (Tb11.02.3640) were hybridised with the same RNA samples. Ethidium bromide staining of rRNA indicates lane loading. Tb927.5.4620 is also named 'ESAG9-EQ'. Figure taken from Barnwell *et al*. (2010).

Another protein exhibiting stumpy form enriched expression is the homologue of aconitase, TbACO. The expression of TbACO was lowest in slender forms, 8-fold

enriched in stumpy forms, and increased 4-fold further in procyclic forms (Saas *et al.*, 2000). However, the *TbACO* mRNA showed only minor lifecycle changes in abundance, indicating regulation through translational or post-translational mechanisms (Saas *et al.*, 2000). Aconitase is an enzyme that catalyses the conversion of citrate to isocitrate via the intermediate of *cis*-aconitate, which as citrate and *cis*-aconitate can trigger differentiation to the procyclic form, makes this protein of interest.

1.5 Eukaryotic translation

Translation of eukaryotic mRNAs is believed to occur through the scanning mechanism of translational initiation. This mechanism proposes that the small ribosomal subunit (40S) is recruited to the mRNA at the 5' end and then migrates linearly along the mRNA until an AUG codon is encountered (reviewed in (Kozak, 2002). However, two secondary mechanisms allow initiation at downstream AUG codons in some mRNAs, reinitiation and context-dependent leaky scanning, and additionally internal ribosome entry sites (IRES). These mechanisms are discussed below in Section 1.6.

Eukaryotic translation initiation (see Figure 1.5; reviewed in (Sonenberg and Hinnebusch, 2009) begins with recruitment to the mRNA of the preassembled 43S preinitiation complex (PIC). The PIC comprises of the 40S ribosomal subunit, Met-tRNAi (methionyl tRNA specialised for initiation), eukaryotic initiation factor (eIF) 2, eIF1, eIF1A, eIF3 and eIF5 (Pestova *et al.*, 1998). The eIF2 and Met-tRNAi are preassembled prior to PIC formation in the ternary complex (TC), where eIF2 is in its GTP-bound form (Choi *et al.*, 1998). Loading on to the ribosome is mediated by the eIF4F complex, which consists of eIF4E, the cap binding factor, and eIF4A, an RNA helicase, both bound to the scaffolding protein, eIF4G (Gingras *et al.*, 1999). Thus, eIF4E is responsible for 5' cap recognition and through it's binding to eIF4G, which binds to eIF3, the recruitment of the PIC, and eIF4A is presumed to unwind secondary structure present in the mRNA to assist scanning (LeFebvre *et al.*, 2006; Marsden *et al.*, 2006). Scanning downstream proceeds until an AUG codon is placed

in the P (peptidyl) decoding ribosome site, which base-pairs with the Met-tRNAi anticodon. This results in GTP hydrolysis in the ternary complex stimulated by eIF5B, and eIF2-GDP is released along with other initiation factors, allowing the large ribosomal subunit (60S) to join, thus forming the 80S ribosome (Pestova *et al.*, 2000; Lee *et al.*, 2002). Protein synthesis therefore begins when the appropriate aminoacyl-tRNA enters into the A (aminoacyl) site and the first peptide bond is synthesised (reviewed in (Leung *et al.*, 2011).

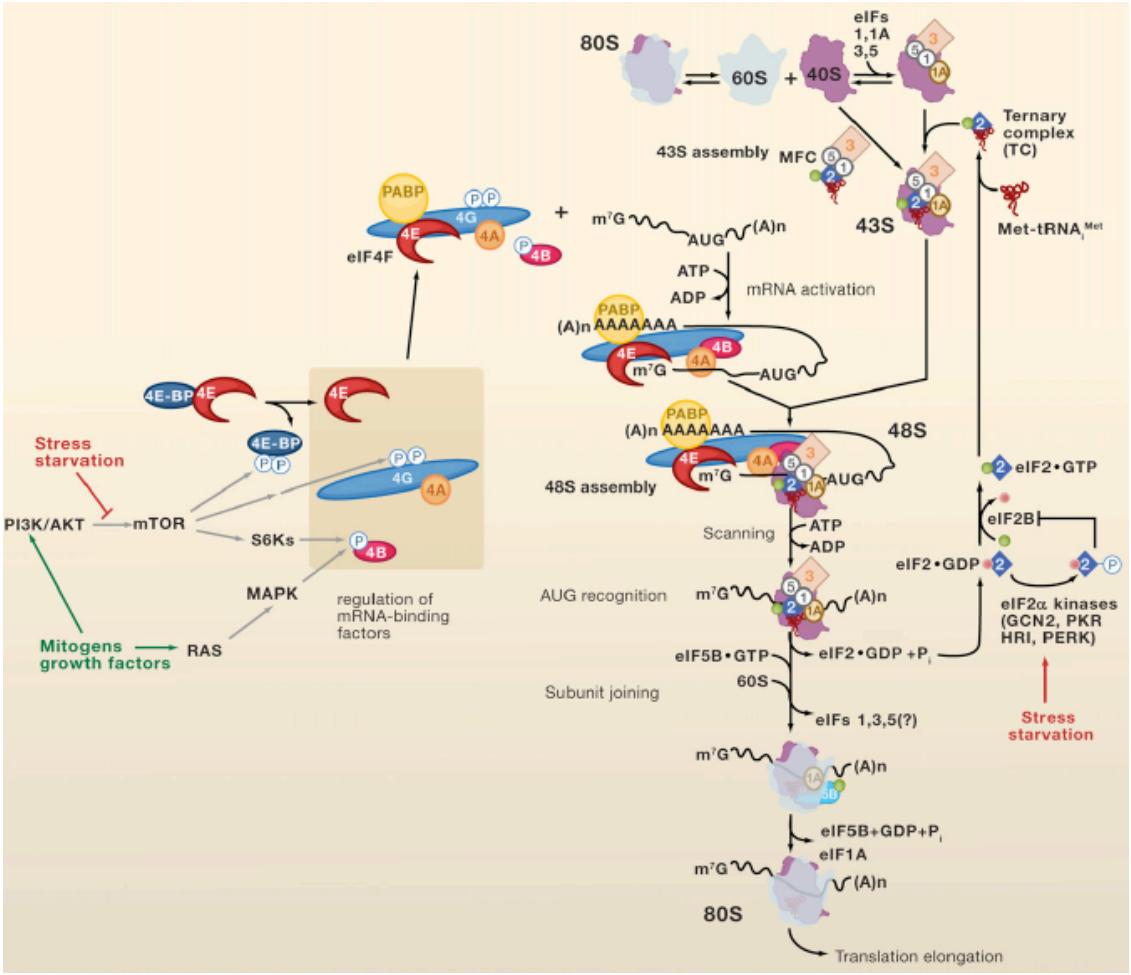


Figure 1.5 Schematic diagram of eukaryotic cap-dependent translation initiation and its regulation by eIF2α kinases and other signalling pathways. Figure taken from Sonenberg and Hinnebusch (2009).

Translation progresses with polypeptide chain elongation through the entry into the A site of an aminoacyl-tRNA associated with a GTP-bound elongation factor (EF), until correct base-pairing to the codon at the A site triggers GTP hydrolysis, peptide

bond formation and translocation of the ribosome to the following codon (reviewed in (Leung *et al.*, 2011). This process repeats until a stop codon is encountered in the A site, which is recognised by release factors (RFs), causing hydrolysis of the polypeptidyl-tRNA, releasing the nascent protein chain and subsequently dissociation of the ribosome (reviewed in (Leung *et al.*, 2011).

A single mRNA can have multiple associated ribosomes engaged in translation; this structure is referred to as a polyribosome or polysome (Warner *et al.*, 1963). Circularisation of the mRNA can be mediated by interaction of eIF4G with poly(A) binding protein (PABP) which links the cap and poly(A) tail of the mRNA (Wells *et al.*, 1998). This could allow reinitiation of dissociated ribosomes on the mRNA cap post-termination (Rajkowitsch *et al.*, 2004).

The translation initiation factors, eIF4E and eIF2, as well as eIF5A and eIF6, are discussed in further detail below, as they have particular relevance to this PhD project.

1.5.1 eIF4E

The structure of eIF4E resembles a cupped hand, consisting of a curved, antiparallel 8-stranded β sheet backed by 3 long α helices (Marcotrigiano *et al.*, 1997). Cap binding occurs through the concave basal surface, where conserved tryptophan side-chains are important for recognition (Marcotrigiano *et al.*, 1997).

eIF4E represents a molecule thought to be important for determining translation rates due to its critical role in mRNA cap recognition and recruitment of the ribosome through its interaction with eIF4G and because it is believed to be present in limiting concentrations (Duncan *et al.*, 1987). Consequently, eIF4E activity is regulated through multiple mechanisms (reviewed in(Raught and Gingras, 1999).

Firstly, eIF4E expression is regulated through transcriptional control by the transcription factor, c-Myc, and p53 (Zhu *et al.*, 2005). Also, phosphorylation status of eIF4E appears to correlate with translation rate, although this is not always

observed (reviewed in(Raught and Gingras, 1999). eIF4E is phosphorylated at Ser209 by mitogen-activating protein kinase (MAPK)-interacting serine/threonine kinase (Mnk) 1 and 2, substrates of extracellular signal-related kinase (ERK) or p38 mitogen-activated protein (MAP) kinases (Pyronnet *et al.*, 1999).

Additionally, the 4E-BP (eIF4E-binding proteins) family of repressor proteins regulate eIF4E, through binding to eIF4E and preventing formation of a functional eIF4F complex by inhibiting binding to eIF4G (Haghighat *et al.*, 1995). Phosphorylation also regulates the activity of the 4E-BPs, as phosphorylation results in dissociation of the protein from eIF4E (Pause *et al.*, 1994). Another level of control results from the eIF4E homologous protein (4EHP), which shows homology to eIF4E and binds specifically to capped RNA (Rom *et al.*, 1998), but does not interact with eIF4G (Joshi *et al.*, 2004), again preventing formation of a functional eIF4F complex.

1.5.2 eIF2

Downregulation of general protein synthesis in response to starvation, stress, or viral infections in mammalian cells is mediated by phosphorylation of the α subunit of eIF2 (Hershey, 1991). Following translation initiation, exchange of GDP for GTP in the inactive eIF2-GDP complex to reform the active ternary complex is catalysed by eIF2B. Phosphorylation of Ser-51 of eIF2α inhibits the recycling of eIF2 by binding to eIF2B and sequestering it, therefore, preventing translation initiation (Rowlands *et al.*, 1988).

A number of kinases have been found to phosphorylate eIF2α in mammalian cells. These include protein kinase R, which is activated upon double-stranded RNA binding (Kostura and Mathews, 1989), and protein kinase R-like ER kinase (PERK), which is activated by the ER stress of accumulation of unfolded proteins (Harding *et al.*, 1999). Although general protein synthesis is repressed by eIF2α phosphorylation, a subset of mRNAs are still actively translated, this being mediated by upstream open reading frame (uORF) mechanisms (discussed below in Section 1.5.5).

1.5.3 eIF5A

eIF5A is a protein shown to promote translation elongation in yeast (Saini *et al.*, 2009). This is the only known protein to contain the unusual amino acid hypusine (N^{ε}-(4-amino-2-hydroxybutyl)lysine) in eukaryotes and archaea (Cooper *et al.*, 1983).

Hypusine is synthesised only as a posttranslational modification of one specific lysine residue of eIF5A through the activities of two enzymes in a two-step process (Murphey and Gerner, 1987; Park *et al.*, 1982; Park *et al.*, 2006). The first enzymatic reaction involves the transfer of the aminobutyl moiety of spermidine to form an intermediate, deoxyhypusine, catalysed by deoxyhypusine synthase (DHS). Deoxyhypusine hydroxylase (DOHH) then hydroxylates this intermediate to form hypusine. The presence of hypusine is essential for eIF5A function, as it is required for the association of eIF5A with ribosomes (Jao and Chen, 2006).

1.5.4 eIF6

eIF6 is a protein that functions as an anti-association factor, as it prevents association of the 40S and 60S subunits by binding to 60S, thus preventing 80S formation (Russell and Spremulli, 1980). Although termed a translation initiation factor, the protein does not function as a true initiation factor, as it is not required for yeast translation in *in vitro* assays (Si and Maitra, 1999). Rather, its function in translation appears to be mediated by its anti-association activity preventing translation. Nonetheless, a role for eIF6 has been reported in yeast in 60S biogenesis, because depletion of eIF6 in this organism resulted in impaired processing of pre-rRNA precursors that are the constituents of 60S subunits (Basu *et al.*, 2001).

eIF6 is found in both Archaea and Eukarya, but not in Eubacteria, where transcription and translation are coupled (Miluzio *et al.*, 2009). Yeast and human eIF6 proteins share 72% identity (Si *et al.*, 1997). The structures have been determined for archaeal (*Methanococcus jannaschii*) aIF6 and *Saccharomyces cerevisiae* eIF6 by X-ray crystallography and demonstrate that IF6 possesses a

phylogenetically conserved unique protein structure termed 'pentein' (Groft *et al.*, 2000). This pentein structure is a pentameric fold consisting of five quasi identical α/β subdomains arrayed around a five-fold axis of pseudosymmetry. Phylogenetic analysis of known eIF6 and aIF6 sequences suggested an origin from a common ancestor before divergence of the archaeal and eukaryotic kingdoms (Groft *et al.*, 2000).

Recently, the crystal structure of the 60S subunit from *Tetrahymena thermophila* in complex with eIF6 has been determined (Klinge *et al.*, 2011). This structure demonstrated that eIF6 binds through the larger of its flat surfaces to the ribosomal protein, RPL23, of the 60S subunit, where it would prevent 40S binding. In yeast, release of eIF6 binding to the 60S is promoted by elongation factor-like 1 (Efl1p) (Senger *et al.*, 2001), and the homologue of Schwachman-Bodian-Diamond syndrome (SBDS) protein, Sdo1 (Menne *et al.*, 2007). In mammals, eIF6 interacts with RACK1 (receptor for activated protein kinase C [PKC] 1) in the cytoplasm and phosphorylation of eIF6 by PKC is believed to mediate dissociation (Ceci *et al.*, 2003).

Investigation of eIF6 heterozygous mice revealed that mammalian eIF6 is required for efficient translation initiation while ribosome biogenesis was unaffected, and additionally that eIF6 regulates translation in response to extracellular signals (Gandin *et al.*, 2008). In yeast, eIF6 is essential for cell growth (Si and Maitra, 1999), and mice heterozygous for eIF6 exhibited decreased growth, while eIF6 knockout was lethal (Gandin *et al.*, 2008), demonstrating the essential role of this protein.

eIF6 protein is found in both the cytoplasm, where its anti-association activity prevents translation, and in the nucleolus, where it is involved in 60S biogenesis (reviewed in (Miluzio *et al.*, 2009) (see Figure 1.6). Localisation is controlled by the phosphorylation status of eIF6 at Ser-174 and Ser-175, where nuclear export is dependent on phosphorylation by a nuclear isoform of casein kinase 1 (CK1) (Biswas *et al.*, 2011).

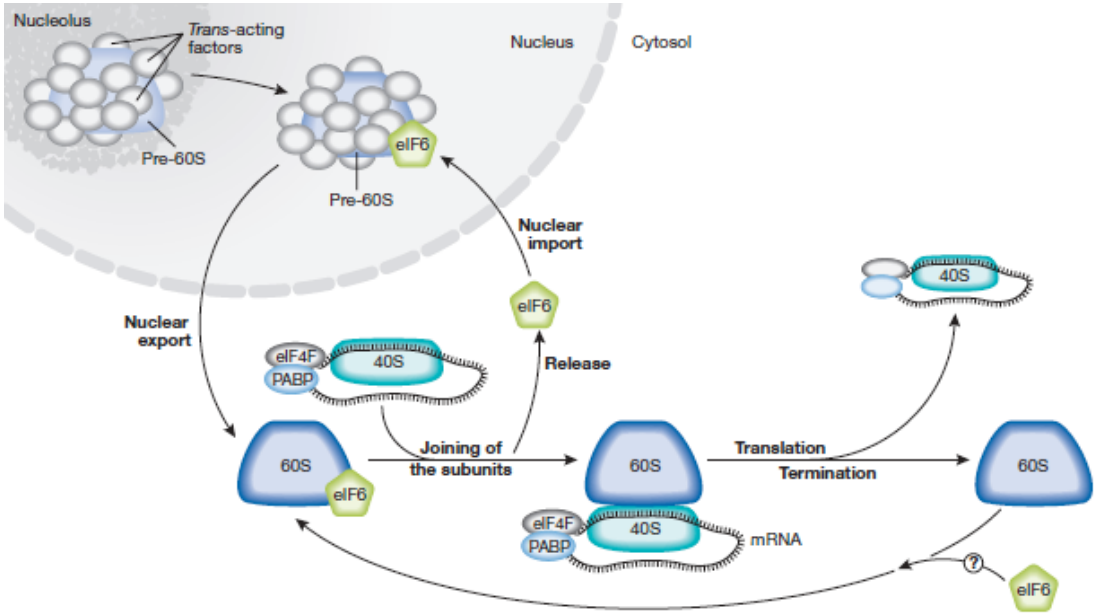


Figure 1.6 Schematic diagram depicting eIF6 in ribosome biogenesis in the nucleus and translational control in the cytosol. eIF: eukaryotic initiation factor; PABP: poly(A)-binding protein. Figure taken from Miluzio *et al*. (2009).

eIF6 has also been implicated in controlling the expression of specific genes. In *Xenopus laevis*, eIF6 is believed to be involved in a mechanism acting on the specific translation of transcripts regulating cell survival, upstream of B-cell lymphoma gene-2 (Bcl-2)/Bcl-2-associated X protein (Bax) (De Marco *et al.*, 2010). eIF6 has also been reported to be involved microRNA (miRNA) silencing and RNAi. Human eIF6 was found to associate with the RNA-induced silencing complex (RISC), and depletion of eIF6 in human or *Caenorhabditis elegans* decreases miRNA-mediated silencing (Chendrimada *et al.*, 2007). In *Fenneropenaeus chinensis* and *Marsupenaeus japonicus*, eIF6 interacts with TRBP (trans-activation response RNA-binding protein), a component of RISC (Wang *et al.*, 2009; Wang *et al.*, 2012). Knockdown of eIF6 or TRBP in *M. japonicus* impaired the double-stranded RNA (dsRNA)-induced RNAi pathway (Wang *et al.*, 2012). Additionally, eIF6 may selectively regulate the protein synthesis of β-catenin mRNA (Ji *et al.*, 2008).

In Archaea, aIF6 has been characterised in *Sulfolobus solfataricus*, where it also shows anti-association activity by binding to 50S ribosomal subunits and preventing formation of 70S ribosomes (Benelli *et al.*, 2009). Although a role for mammalian eIF6 in G1/S phase progression through the cell cycle was identified (Gandin *et al.*,

2008), aIF6 was found to uniformly expressed in the cell cycle. However, aIF6 protein was found to be upregulated following both cold- and heat shock, indicating that it may negatively regulate protein synthesis under unfavourable environmental conditions (Benelli *et al.*, 2009).

1.5.5 Control of the scanning mechanism of translation initiation

A number of mechanisms have been identified that affect the scanning mechanism of translation initiation; these are GC-rich 5'UTRs, leaky scanning, upstream open reading frames (uORFs) and IRES. GC-rich 5'UTRs are highly structured which prevents the ribosome from traversing the 5'UTR, therefore inhibiting translation initiation (reviewed in Kozak, 2002).

Recognition of an AUG start codon is enhanced by the context of the surrounding sequence. In mammals, the optimal context is GCCRCCaugG, where R (position -3) represents a purine and is the most important position functionally (Kozak, 1986) and the G in position +4 contributes strongly too (Kozak, 1997). If the first AUG is in a strong context, translation initiates there, however, if it is in a weak context, leaky scanning can result whereby scanning continues to a downstream AUG for initiation and can therefore result in production of two proteins from one transcript (reviewed in Kozak, 2002).

An uORF is a small ORF present at the 5' end of a transcript upstream from the main ORF. The presence of an uORF can affect the translation of the main ORF through two mechanisms. The scanning ribosome can initiate translation on the AUG of the uORF, causing translation of the uORF peptide which results in reinitiation of scanning, believed to be due to the 40S ribosome remaining bound to the mRNA (Kozak, 2002). Alternatively, it can inhibit translation of the main ORF by translation of the uORF peptide causing stalling of the ribosome. For example, the arginine attenuator peptide encoded by an uORF present in yeast *CPA1* mRNA causes ribosome stalling on the uORF termination codon (Gaba *et al.*, 2005). However, for reinitiation to occur the 40S subunit must reacquire the ternary complex and therefore the length between the stop codon of the uORF and the AUG

of the main ORF is an important determinant of whether this occurs (reviewed in (Hinnebusch, 1997). This is the mechanism by which phosphorylation of eIF2α is thought to allow the expression of certain transcripts while causing global translation inhibition, i.e. through reduction of the pool of ternary complexes, thus controlling the reinitiation ability of ribosomes on the main ORF or an additional downstream uORF. The perhaps best characterised example of this is the translation of *GCN4* in yeast (reviewed in Hinnebusch, 1997) and it has also been demonstrated for GADD34 (growth arrest DNA-inducible gene 34) in mammals (Lee *et al.*, 2009).

An IRES is a structural element that can cause the initiation of translation by direct recruitment of ribosomes to an AUG present in an internal site of the mRNA and is therefore independent of the cap binding complex, eIF4F. The first example of IRES was idenitifed in picornaviral mRNAs (Jang *et al.*, 1988).

1.6 Gene expression in *Trypanosoma brucei*

1.6.1 Genome organisation

The genome of *T. brucei* has been sequenced and published (Berriman *et al.*, 2005), in addition to those of *T. cruzi* (El-Sayed *et al.*, 2005) and *L. major* (Ivens *et al.*, 2005). The *T. brucei* genome is organised into 11 megabase-sized chromosomes and an unspecified number of small and intermediate-sized chromosomes that encode sequences with similarity to megabase chromosome subtelomeric regions. The genes on the megabase chromosomes are arranged in long, nonoverlapping gene clusters which are transcribed as polycistrons.

The polycistronic transcription units (PTUs) are marked at the boundaries by four histone variants and composed of arrays of sometimes more than 100 genes, but with no evidence of functionally related gene clusters as seen for prokaryotic operons (Siegel *et al.*, 2009). Transcription is carried out by RNA polymerase II (Pol II), except for the major surface glycoprotein genes (the procyclins and VSGs), which are transcribed by Pol I (Gunzl *et al.*, 2003). For Pol II-expressed genes, no promoters have been identified, with the exception of the splice leader (SL) RNA

(Gilinger and Bellofatto, 2001), whereas Pol I promoters have been identified. Instead, probable Pol II transcription start sites show strong enrichment of histones H4K10ac, H2AZ, H2BV, and the bromodomain factor BDF3 (Siegel *et al.*, 2009). Thus, while Pol II-expressed genes are likely regulated at the post-transcriptional level, Pol I-expressed genes are subject to different regulatory mechanisms for gene expression at the transcriptional level.

1.6.2 Trans-splicing and polyadenylation

Monocistronic mRNAs are obtained from polycistronic RNA precursors by the two processes of *trans*-splicing, at the 5' end, and cleavage/polyadenylation of the 3' end. The *trans*-splicing reaction involves the addition of a 39 nucleotide (nt) splice leader RNA to the 5' splice acceptor site (SAS) (Agabian, 1990). The splice leader RNA also possesses a modified 5' mRNA cap structure, termed cap 4 and in some aspects is unique to trypanosomatids, which generates a functional mRNA (Bangs *et al.*, 1992). Hypermethylated cap 4 also has methylation of the first four nucleotide positions, this hypermethylation is required for efficient translation rates (Zamudio *et al.*, 2009). Similarly to *cis*-splicing, the *trans*-splicing acceptor site is an AG dinuceotide downstream of a polypyrimidine tract (Huang and Van der Ploeg, 1991; Matthews *et al.*, 1994). The polyadenylation site (PAS) has no known consensus signal, but instead is linked to *trans*-splicing of the downstream gene, occurring 100-400nt upstream from the polypyrimidine tract (Matthews *et al.*, 1994; Benz *et al.*, 2005).

Trans-splicing, therefore, overcomes the restriction that only the 5' open reading frame would be translated in the polycistron by the scanning mechanism of initiation, through the separation of individual mRNAs. *Cis*-splicing is extremely rare, but does occur in *T. brucei* for at least two genes, poly(A)-polymerase (Mair *et al.*, 2000) and a DNA/RNA helicase (Berriman *et al.*, 2005). Identification of introns in other *T. brucei* genes appears unlikely as an RNA-seq study did not identify any new examples, although, the authors reported that the methodology did not provide a reliable way to identify introns (Siegel *et al.*, 2010).

1.6.3 Identification of splicing and polyadenylation sites

An analysis of known SAS and PAS sites in *T. brucei* cDNAs found that in general trans-splicing occurs at the first AG dinucleotide following a polypyrimidine tract of 8-25nt, resulting in a median 5'UTR length of 68nt and mean length of 129nt (Benz *et al.*, 2005). Additionally, polyadenylation occurs between 80 and 140nt downstream of the polypyrimidine tract, at a position of one or more A residues, resulting in a median 3'UTR length of 348nt and mean length of 587nt. These data were then used to generate a predictive algorithm for processing sites for most protein coding genes in the *T. brucei* genome (Benz *et al.*, 2005).

Since that study, four studies have been recently published which used Illumina/Solexa high-throughput RNA sequencing to determine SAS and PAS sites in *T brucei* on a genome wide scale (Siegel *et al.*, 2010; Nilsson *et al.*, 2010; Kolev *et al.*, 2010; Veitch *et al.*, 2010). These studies varied in the *T. brucei* strains analysed and used varying approaches to sequencing. For example, Siegel *et al*. (2010) used *in vitro* cultured *T. b. brucei* slender monomorphic bloodstream and procyclic forms with RNA-seq, Nilsson *et al*. (2010) used *T. b. brucei* slender and stumpy (monomorphic and pleomorphic) bloodstream forms, and procyclic forms with a novel approach termed spliced leader trapping (SLT), Kolev *et al*. (2010) used *T. b. rhodesiense* procyclic forms, and Veitch *et al*. (2010) used *T. b. gambiense* slender bloodstream and procyclic forms with a digital gene expression approach (which relies on the restriction enzyme, *Nla*III).

Siegel *et al*. (2010) identified the SAS for 6959 genes and the PAS for 5948 genes, noting that most genes have between one and three alternative SAS and PAS are more dispersed. Additionally, the paper observed that for 488 genes the SAS was identified downstream from the originally assigned ATG start codon, meaning that a subsequent in-frame ATG must be used for the true coding sequence. Nilsson *et al*. (2010) identified the SAS for 85% of the annotated protein-coding genes, with more than 2500 alternative splicing events being identified that appear to be lifecycle stage-regulated in many cases. The data from all the studies confirmed that the major splice acceptor dinucleotide is AG and that the SAS is preceded 14-43nt upstream by

a polypyrimidine tract (Siegel *et al.*, 2011). The observed median 3'UTR length was reported as 400nt and 388nt (Siegel *et al.*, 2010; Kolev *et al.*, 2010). The high abundance of alternative SAS and alternative PAS site usage found in these studies was surprising and the authors noted that this could represent a means for generating the same protein but with differing properties (for example, the inclusion/exclusion of targeting signals for different cellular localisations) and a means for regulating gene expression (Siegel *et al.*, 2011), which has also been previously proposed (Helm *et al.*, 2008).

1.7 Regulation of gene expression in *T. brucei*

Due to the transcription of almost all trypanosome protein-coding genes by Pol II in polycistronic transcription units and the observation that co-transcribed genes do not share common patterns of regulation, regulation of gene expression is believed to occur at the post-transcriptional level (Clayton & Shapira, 2007). Post-transcriptional control of gene expression can be made at various levels in eukaryotes, such as the regulation of splicing and polyadenylation, mRNA stability and degradation, RNA interference, translation, and protein stability. Studies of the mechanisms of trypanosome post-transcriptional control have so far focused more on the characterisation of 3'UTR elements, RNA binding proteins and mRNA degradation and these appear to be particularly important.

1.7.1 Characterisation of 3'UTR elements

The most well characterised trypanosome 3' untranslated region (3'UTR) is that of procyclin. Procyclin is the major surface antigen expressed in the lifecycle stages present in the tsetse fly and this stage-specific expression is conferred by the 3'UTR (Berberof *et al.*, 1995). There were three elements in the 3'UTR identified to be responsible for this control; two positive elements, one within the first 40nt of the 3'UTR and also a conserved 16-mer (loops I and III), which prevent transcript degradation, and a negative element (loop II) (Furger *et al.*, 1997). The structure of the 3'UTR of EP1 procyclin was determined experimentally to be composed of three domains, two of which form stable structures corresponding to loops I and III

(Drozdz and Clayton, 1999). The 3'UTR of the GPEET procyclin also confers regulation by glycerol and hypoxia (Vassella *et al.*, 2000).

Additionally, investigation has been made of the 3'UTRs of the nuclear-encoded mitochondrial cytochrome oxidase (COX) complex components. These components are developmentally regulated, being enriched in the procyclic form at the mRNA level (Mayho *et al.*, 2006). The *COX* mRNAs exhibit differential mRNA stability and for at least three subunits the 3'UTRs possess regulatory regions contributing to translational control (Mayho *et al.*, 2006). This study also identified a 26-mer sequence in one of the 3'UTRs that is overrepresented in procyclic-enriched transcripts.

Other 3'UTRs conferring developmental gene expression control (reviewed in (Clayton, 2002) include an amino acid transporter (AATP11) that is enriched in procyclic forms (Robles and Clayton, 2008), phosphoglycerate kinase transcripts (Blattner and Clayton, 1995), and *VSG* mRNA that is expressed specifically in bloodstream forms (Berberof *et al.*, 1995).

1.7.2 RNA-binding proteins

Due to their reliance on post-transcriptional control of gene expression, kinetoplastid genomes show an enrichment in genes encoding RNA-binding proteins compared to unicellular organisms that rely on transcriptional control (Ivens *et al.*, 2005). In *T. brucei*, a particularly well characterised set of RNA-binding proteins are the three zinc finger protein (ZFP) proteins, one of which, TbZFP3, enhances differentiation to procyclic forms and shows a polysome association that is developmentally regulated (Paterou *et al.*, 2006). This protein shows specific association with the *EP1* and *GPEET procyclin* isoform transcripts and therefore provided the missing link between the regulatory regions defined in these mRNAs (see above) and translation of these transcripts (Walrad *et al.*, 2009).

Other examples in *T. brucei* include the RNA helicase, DHH1, which is important for the correct expression of many developmentally regulated genes (Kramer *et al.*,

2010); the RNA-binding protein PUF (Pumilio/Fem-3) 9, which stabilises a set of transcripts during S-phase (Archer *et al.*, 2009); RNA-binding proteins ALBA3 and 4, implicated in differentiation of mesocyclic to epimastigote insect forms (Subota *et al.*, 2011); and the RNA-binding protein, TbDRBD3, which binds to and promotes the stability of a subset of developmentally regulated mRNAs encoding membrane proteins (Estevez, 2008). Therefore, in trypanosomes RNA-binding proteins can regulate the expression of specific sets of mRNAs during particular cellular processes.

1.7.3 mRNA degradation

The decay of mRNAs is achieved by removal of the poly(A) tail by the PAN2/PAN3 complex, CCR4/CAF1/NOT complex, or poly(A)-specific ribonuclease (PARN), and then 3'-5' exonucleolytic digestion or decapping and 5'-3' exonucleolytic digestion (reviewed in (Parker and Song, 2004). In *T. brucei*, a complex containing CAF1 and NOT proteins was identified and CAF1 was shown to possess deadenylation activity with its depletion resulting in delayed deadenylation and degradation of mRNAs (Schwede *et al.*, 2008). Additionally, *T. brucei* PARN-1 has been shown to be a functional deadenylase and responsible for the degradation of a subset of mRNAs (Utter *et al.*, 2011). In yeast, the exosome is complex composed 3'-5' riboexonucleases responsible for most 3'-5' exonucleolytic digestion and *T. brucei* possesses a simplified exosome complex, being composed of fewer proteins (Estevez *et al.*, 2001). Depletion of the major 5'-3' exoribonuclease, XRNA, resulted in the stabilization of most unstable mRNAs (Manful *et al.*, 2011). Taken together, these results suggest that *T. brucei* possesses the normal mechanisms for mRNA degradation.

1.7.4 RNAi

In eukaryotes, RNA interference (RNAi) and related small-RNA-mediated pathways play an important role in the silencing of gene expression, initiated by small RNAs (~20-30nt) with sequences complementary to regions of the transcripts they regulate (reviewed in (Jinek and Doudna, 2009). *T. brucei* possesses the RNAi pathway, being one of the first organisms where it was discovered (Ngo *et al.*, 1998), and

homologues of RNAi proteins, namely two Dicer-like enzymes (TbDCL1 and TbDCL2) (Patrick *et al.*, 2009) and an Argonaute (AGO) protein (TbACO1) (Durand-Dubief and Bastin, 2003; Atayde *et al.*, 2011). TbDCL2 is nuclear and involved in the endogenous RNAi response to retroposons and repeats, whereas TbDCL1 is mostly cytoplasmic and proposed to be involved in processing any cytoplasmic dsRNA (Patrick *et al.*, 2009).

RNAi in *T. brucei* is involved in the suppression of two classes of retroposon, INGI (Kimmel *et al.*, 1987) and SLACS (Spliced Leader Associated Conserved Sequence) (Aksoy *et al.*, 1990) transcripts (Shi *et al.*, 2004). Prolonged culture of *T. brucei* deficient in RNAi results in genomic rearrangements of SLACS, supporting the hypothesis that this pathway serves to control the spread of retroposons in the genome (Patrick *et al.*, 2008). The RNAi pathway has not been implicated the control of expression of any other transcripts.

1.7.5 Translation

To date, there have been few studies examining components of the translational machinery in *T. brucei*. The proteins studied thus far are members of eIF4F cap-binding complex (Freire *et al.*, 2011; Dhalia *et al.*, 2006) and three potential eIF2α kinases (TbeIF2K1-3; Moraes *et al.*, 2007).

T. brucei contains four homologues of eIF4E, TbeIF4E1-4 (Clayton and Shapira, 2007), which have been investigated to some extent. One study has reported that TbeIF4E1 and 2 localise to the nucleus and cytoplasm, while TbeIF4E3 and 4 are cytoplasmic, but all are constitutively expressed across bloodstream and procyclic stages (Freire *et al.*, 2011). Additionally, through RNAi-mediated ablation, TbeIF4E3 or TbeIF4E1 and 2 together were found to be essential for procyclic form viability, whereas TbeIF4E1, 3 and 4 were found to be essential in the bloodstream form. Knockdown of TbeIF4E3 alone, or TbeIF4E1 and 2, or TbeIF4E1 and 4 resulted in translation inhibition as assessed by radiolabelled methionine incorporation. TbeIF4E1, 2 and 4 were found to bind a cap analogue, but not TbeIF4E3, and TbeIF4E3 and 4 were found to interact with *T. brucei* eIF4G

homologues (Freire *et al.*, 2011). However, the characterisation of each of the homologues in this paper are limited by two main criticisms, firstly, the cap-binding efficiency was assessed through binding to m7 cap, and not the significantly different trypanosome cap 4, meaning that this cap-binding analysis had limited biological relevance. Secondly, a direct role in translation initiation was not demonstrated, as the authors did not perform sucrose density polysome analysis to identify where the proteins migrate using their antibodies to the homologues or the effects on translation of TbeIF4E RNAi knockdown.

Leishmania major also possesses four homologues of eIF4E, LmEIF4E1-4, which are well conserved with the *T. brucei* homologues (Dhalia *et al.*, 2005) and better characterised. LmEIF4E1 and 4 bind m^7GTP and cap 4, and interact with mammalian 4EBP1, LmEIF4E1 is more abundant in the amastigote lifecycle stage, LmEIF4E2 binds mainly to cap 4 and comigrates with polysomal fractions on sucrose gradients, whereas LmEIF4E3 bound mainly to m^7GTP and comigrated with 80S ribosomes (Yoffe *et al.*, 2006). Additionally, LmEIF4E4 was found to interact with an eIF4G homologue, LeishIF4G-3 (Yoffe *et al.*, 2009; Zinoviev *et al.*, 2011). Interestingly, in *Leishmania* a novel non-conserved 4E-Interacting Protein has been identified which binds to LmEIF4E1 in promastigotes but not amastigotes, and so was proposed to be involved in a switch between cap-dependent and independent translation (Zinoviev *et al.*, 2011).

Two homologues of eIF4A, a member of the cap-binding complex along with eIF4E and eIF4G, have also been studied. TbEIF4AI is a very abundant and predominately cytoplasmic protein, whose depletion dramatically reduced protein synthesis and prevented cell proliferation, and TbEIF4AIII is a nuclear protein, whose depletion also stopped cell proliferation but with slower kinetics (Dhalia *et al.*, 2006).

As noted previously, the presence of multiple copies of eIF4F subunits in the typanosomatids may reflect their association with different life-cycle stages or the translation of specific mRNA subsets (Dhalia *et al.*, 2005). This could make these

proteins particularly interesting to study in the context of life-cycle stage dependent translation.

Three potential eIF2α kinases were identified in *T. brucei*, and one, TbeIF2K2, was characterised to be a transmembrane glycoprotein expressed in bloodstream and procyclic forms (Moraes *et al.*, 2007). This protein was capable of phosphorylating yeast and mammalian eIF2α at Ser-51 and localised mainly to the membrane of the flagellar pocket, which is the site of exo- and endocytosis, suggesting that it may have a role in a sensing mechanism (Moraes *et al.*, 2007).

Interestingly, it was recently shown that in *T. cruzi*, nutritional stress results in increased levels of Tc-eIF2α phosphorylation and inhibition of global translation initiation, and also that Tc-eIF2α phosphorylation is required for parasite differentiation to infective forms (Tonelli *et al.*, 2011). In *Leishmania*, eIF2α phosphorylation has also been associated with decreased protein synthesis and differentiation (Lahav *et al.*, 2011). Phosphorylation of eIF2α and inhibition of translation have been shown to control sporozoite latency in *Plasmodium*, through the action of an eIF2α kinase, IK2 (Zhang *et al.*, 2010), possibly with a Puf protein, Puf2 (Muller *et al.*, 2011). Additionally, eIF2α phosphorylation by another eIF2α kinase, PfeIK1, is involved in the stress response to amino acid starvation (Fennell *et al.*, 2009) In *T. brucei*, however, eIF2α phosphorylation is not associated with stress response of heat shock (Kramer *et al.*, 2008).

Therefore, studies so far indicate that *T. brucei* appears to contain homologues of the main translational protein machinery.

1.8 Project aims

To date, gene expression studies in trypanosomes have focussed entirely on the proliferative life-cycle stages in mammalian hosts (bloodstream slender forms) or the tsetse mid-gut (procyclic forms). Nothing is known about gene expression control in non-proliferative stumpy forms. However, here, translational control is likely to be

very important because most mRNAs are translationally repressed in this quiescent cell-type, though a small subset of mRNAs are known to be highly expressed, likely to assist survival of the stumpy form, or to preadapt this parasite for transmission to the Tsetse mid-gut. Such transcripts, therefore, provide an excellent model for the escape of translational repression, although the regulatory mechanisms are unknown.

Therefore, the question to be addressed in this work was how are translation and the escape from translational arrest in stumpy form *T. brucei* controlled? The specific aims of this work were to:

- Determine what signals control the expression, or escape from translational arrest, of genes expressed in stumpy forms
- Investigate if mechanisms such as uORFs are involved in the control of the escape from repression, or repression, of genes expressed in stumpy forms
- Investigate if specific translational proteins are involved in translational control in stumpy forms
- Identify which transcripts escape translational repression in stumpy forms

Chapter 2

Materials and Methods

2.1 Bioinformatic analysis

2.1.1 Basic bioinformatic tools

The sequenced Trypanosomatid genomes and associated gene entry annotations were accessed through the TriTryp database (http://tritrypdb.org/tritrypdb/).

Multi-alignment of DNA sequences or protein sequences were performed online using the ClustalW2 programme (Multiple Sequence Alignment tool, version 2.1) from the EMBL-EBI web server (http://www.ebi.ac.uk/Tools/msa/clustalw2/) with the default settings (Larkin *et al.*, 2007).

Protein-protein BLAST (Basic Local Alignment Search Tool; (Altschul *et al.*, 1990) searches were performed online using the EMBL-EBI web server (http://blast.ncbi.nlm.nih.gov/Blast.cgi) with the default settings, or accessed through the TriTryp database web server (http://tritrypdb.org/tritrypdb/) for the interrogation of *T. brucei* annotated genes alone.

For prediction of RNA secondary structure the *S*rna application module of the freely available web-based software package *S*fold (Ding *et al.*, 2004) was used with the default settings (http://sfold.wadsworth.org/cgi-bin/index.pl).

Annotated chromosome sequences of the *T. brucei* TREU 927/4 genome were viewed using Artemis software available from the Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/resources/software/artemis/). Predicted splice sites (Benz *et al.*, 2005) were included by using the 'Read an Entry' option in Artemis.

Protein structure homology modelling was performed using SwissModel (Arnold *et al.*, 2006; Kiefer *et al.*, 2009) using the 'Automated Mode' (http://swissmodel.expasy.org/). Models were then viewed using the Swiss-PdbViewer application (available from http://spdbv.vital-it.ch/).

2.1.2 Identification of potential uORFs

For computational identification of potential uORFs present in the *T. brucei* TREU 927/4 genome, firstly, for all annotated genes in the published genome the 100 nucleotide of 5' intergenic sequence immediately preceding the start codon of each annotated gene, was obtained from the database (http://www.genedb.org/genedb/tryp/) using the 'list download' facility. These sequences were then translated into amino acid sequence using the ExPASy tool, http://searchlauncher.bcm.tmc.edu/seq-util/Options/sixframe.html, or http://biotools.umassmed.edu/cgi-bin/biobin/transeq. A PERL script was then used to identify within these sequences of a methionine codon followed by a stop codon with at least two amino acid residues in between, this being designated as a potential uORF sequence (PERL script written and executed by Pankaj Barua, University of Edinburgh).

2.1.3 Basic GO term analysis

A Gene Ontology (GO) term search was performed with a SQL query (Appendix 2; written by Christiane Hertz-Fowler, Wellcome Trust Sanger Institute, Hinxton) containing the IDs of the genes of interest using http://www.berkeleybop.org/goose. The frequency of each GO term returned for a group of genes was determined using the 'pivot table' function in Microsoft Office Excel and used for comparison of terms returned for different groups of genes.

2.1.4 GO Toolbox analysis

The GO-Stats online tool (http://genome.crg.es/GOToolBox/) (Martin *et al.*, 2004) was used to find enriched and depleted GO terms associated with the gene lists under investigation compared to those associated with the whole *T. brucei* genome. For each gene list a GTB file was generated using the Create Dataset tool (http://genome.crg.es/cgi-bin/GOToolBox/gotbx-dataset-form.cgi). Within this tool the 'Trypanosoma brucei GeneDB version' was selected as the corresponding annotation set, all three (biological process, molecular function, cellular component) ontologies were selected, and the other settings remained at default. The GTB file

was then used as input in the GO-Stats tool, 'hypergeometric' selected as the statistical test and 'Bonferroni' selected as the correction.

2.1.5 Regulatory Sequence Analysis Tools (RSAT) analysis

To detect possible regulatory sequences in a set of genes the online Regulatory Sequence Analysis Tools (RSAT; http://rsat.scmbb.ulb.ac.be/rsat/) was used, as these tools detect over-represented oligonucleotides in sets of DNA sequences. The 'sequence retrieval' tool at TriTrypDB (http://tritrypdb.org/tritrypdb/) was used to obtain the specified amount of genomic sequence upstream/downstream of the translation start/stop codon of each gene ID in the gene set under investigation. The retrieved sequences were used in the 'oligo-analysis (words)' RSAT tool, with the parameters remaining at the default settings, except that the oligomer length chosen was 8 and Markov model selected. The output from this was then transferred to the 'string-based pattern matching (dna-pattern)' tool as input, and again the parameters remained at their default settings, except that 'matching statistics' was also selected as output returned. Again this output was used as input in the 'feature map' tool, the parameters again remained at default.

For investigation with the ESAG9-EQ element sequence (see Section 3.9.3), the RSAT DNA-pattern matching tool was used. The default settings were used, except the 'direct only' option for strand searched was selected, 'start' as the origin, allowing 2 substitutions in the pattern, and also 'match scores' and 'matching statistics' were selected in the output options.

2.1.6 Additional protein prediction programmes

The Pfam 24.0 database (http://pfam.sanger.ac.uk) sequence search was used with the protein sequence of interest using the default parameters (including the default value of 1.0 for the E-value) to identify potential protein domains within a sequence (Finn *et al.*, 2010). The SignalP 3.0 server (www.cbs.dtu.dk/services/SignalP) was used to predict the presence and location of signal peptide cleavage sites within amino acid sequences used as input with the default settings, 'Eukaryotes' selected as the organism group and 'Both' selected as the method (Bendtsen *et al.*, 2004b).

The PSORT server (http://psort.hgc.jp/) was used to assess iPSORT, WoLF PSORT, and PSORT II programmes (Nakai and Horton, 1999). For iPSORT prediction, 'non-plant protein' was selected and for WoLF PSORT prediction, 'animal' was selected as the organism type. The SecretomeP 2.0 server (http://www.cbs.dtu.dk/services/SecretomeP/) was used with the 'Mammalian' setting to predict non-classical protein secretion (Bendtsen *et al.*, 2004a). The TMHMM server v. 2.0 (http://www.cbs.dtu.dk/services/TMHMM/) was used to predict transmembrane helices (Krogh *et al.*, 2001). The TargetP (http://www.cbs.dtu.dk/services/TargetP/) programme (Emanuelsson *et al.*, 2000) was used with 'Non-plant' selected as the organism group, 'Perform cleavage site predictions' and the default cutoff selected.

2.2 Trypanosomes

The monomorphic trypanosome strains used in this work were *T. b. brucei* Lister 427 (bloodstream form), *T. b. brucei* 427-449 (bloodstream and procyclic forms), and *T. b. brucei* 'single marker' cell line (bloodstream form). The pleomorphic trypanosome strains used in this work were *T. b. brucei* AnTat 1.1 and *T. b. brucei* AnTat 1.1 90:13.

The *T. b. brucei* Lister 427 cell line was the most commonly used cell line in this study and details of its lineage can be found on the website of Prof. George Cross (http://tryps.rockefeller.edu/DocumentsGlobal/lineage_Lister427.pdf). The *T. b. brucei* 427-449 cell line is the Lister 427 cell line that has been previously engineered to possess stable integration of the pHD449 construct (Biebinger *et al.*, 1997). Therefore, these cells express the tetracycline (Tet) repressor and were used for inducible, ectopic protein expression. The *T. b. brucei* Lister 427 'single marker' bloodstream cell line (from Dr Achim Schnaufer, University of Edinburgh) expresses the Tet repressor (TetR) and T7 RNA polymerase (Wirtz *et al.*, 1999) and was, therefore, used for inducible RNAi-mediated transcript ablation and ectopic protein expression.

The *T. b. brucei* AnTat 1.1 cell line was used for murine infections to obtain slender, intermediate and stumpy form populations to obtain RNA samples for Illumina digital tag sequencing. The *T. b. brucei* AnTat 1.1 90:13 cell line (Engstler and Boshart, 2004) was used for transfection of pleomorphic cells with CAT reporter gene constructs (see Section 3.9) and RNAi constructs (see Section 4.9.2).

2.2.1 Laboratory culture

Monomorphic bloodstream form *T. brucei* were maintained at logarithmic concentrations at 37°C with 5% CO_2 in HMI-9 medium (Hirumi and Hirumi, 1989) containing 20% foetal calf serum (FCS; Fetal Bovine Serum Gold, PAA) and penicillin-streptomycin (100 Units/ml Penicillin, 100 µg/ml Streptomycin; Invitrogen) using vented flasks. Procyclic form *T. brucei* were kept at 27°C in SDM-79 medium (Cross and Manning, 1973) containing ~15% FCS, 0.1% hemin and penicillin-streptomycin (100 Units/ml Penicillin, 100 µg/ml Streptomycin; Invitrogen) using non-vented flasks. The additions to both media were sterilised by passing through a single-use syringe filter (0.20 µm; Sartorius Stedim Biotech) into the medium.

When cultured *in vitro* following transfection (see Section 2.2.4), pleomorphic bloodstream form *T. b. brucei* cells were kept in HMI-9 in monomorphic bloodstream form conditions. The medium used was always pre-warmed to 37°C, the time out of 37°C incubation was minimised and the cell density was maintained below 1 x 10^6 cells/ml were possible.

For storage of trypanosome cells, 10 ml of the cell culture in exponential-phase growth was pelleted at 2000 rpm for 10 minutes and resuspended in 1 ml of the appropriate medium with 7% glycerol. This was then transferred into a cyrovial tube and put at -80°C. For longer term storage the cryovial was transferred into liquid nitrogen storage. To thaw the cells, the cryovial was placed briefly in a 42°C water bath until thawed, then the culture transferred into 9 ml of medium appropriate for the life-cycle stage and 24 hours later any necessary selective drugs for the cell line were added.

Unless otherwise stated, cell concentration was determined by a 1:40 dilution of the cell culture in Coulter Isoton II Diluent (Beckman Coulter) using a Z2 Coulter Particle Count and Size Analyzer (Beckman Coulter) measuring the range of 2.5-7 microns.

The single marker cell line was maintained with 2.5 µg/ml G418 Disulphate (Melford); the 427-449 cell line was maintained with 0.5 µg/ml phleomycin (Invitrogen). Induction of RNAi or ectopic protein expression from transfectant cell lines with pQuadra, pALC14 or pHD451-Ty plasmids was performed using 10 µg/ml or 1 µg/ml Tetracycline (Sigma; dissolved in ethanol for a stock solution).

2.2.2 Purification of trypanosomes from murine blood

The mice used in this study were the MF1 strain and male. Each mouse was infected with trypanosomes in a maximum volume of 0.2 ml by intraperitonal (IP) injection. If infected for the harvest of a stumpy form population, the mouse was injected (IP) with 250 µl cyclophosphamide (25 mg/ml) the day prior to infection. If necessary, the infection was monitored by microscopic examination of a small volume of blood obtained by tail-snip of the infected mouse, then the blood from trypanosome-infected mouse was collected by heart puncture. The blood was allowed to pass through a diethylaminoethyl (DEAE) cellulose (DE-52) anion exchange column (Lanham and Godfrey, 1970) equilibrated with PSG (pH7.8, see Appendix 1). Note that all animal work was performed by Prof. Keith Matthews, Deborah Hall or Julie Wilson in accordance with their Home Office licences.

2.2.3 Transfection of monomorphic trypanosomes

Prior to transfection, 15 µg of the construct was linearised in a 50 µl *NotI* (Promega) restriction enzyme digest incubated overnight at 37°C. Linearisation was confirmed by comparison of 1 µl of the digest to uncut vector on a 1% w/v agarose gel made with TAE (Tris-acetate EDTA) buffer. The remaining digest solution was purified using Nucleospin Extract II (Macherey-Nagel) according to the manufacturer's instructions, except that the elution step was performed four times, each with 50 µl

of elution buffer. The DNA was ethanol precipitated by adding 20 µl of 3M sodium acetate and 550 µl 100% ethanol, briefly vortexing the solution and incubating at -80°C overnight. Then, the solution was centrifuged at 14000 rpm for 10 minutes at 4°C and the pellet was allowed to air-dry before resuspension with 5 µl 1mM Tris-HCl pH8, 0.1mM EDTA.

For transfection, 4 x 10^7 logarithmically-growing trypanosomes were pelleted at 2000 rpm for 10 minutes and resuspended in 100 µl transfection buffer (either AMAXA Nucleofector solution [Lonza] or Ingenio Electroporation Solution [Mirus]) with the purified DNA. The resuspended cells were added to an AMAXA/Mirus cuvette and transfected using an AMAXA Nucleofector machine (Lonza) with programme X-001 or Z-001 (Burkard *et al.*, 2007; Schumann Burkard *et al.*, 2011). Following transfection, cells were allowed to recover in HMI-9 overnight, before selection of transfectants 24 hours post-transfection using serial dilutions (neat, 1:10, 1:100, 1:1000) of the cells in HMI-9 containing the selective drug. As a control, the parental cell line was also used for serial dilutions with HMI-9 containing the selective drug to ensure lethality of the drug for cells not possessing the transfected construct.

For transfections with the CAT449 plasmid, transfectants were selected and maintained with 0.5 or 2.5 µg/ml phleomycin (Invitrogen); 2 µg/ml hygromycin B (InvivoGen) for the pHD451-Ty plasmid; 0.5 µg/ml puromycin (Calbiochem) for the pALC14 plasmid; and 2.5 µg/ml phleomycin for the pQuadra plasmid.

2.2.4 Transfection of pleomorphic trypanosomes

Preparation of *Not*I-digested DNA for transfection was performed as for transfection of monomorphic cells (section 2.2.3). Mice were infected with the AnTat1.1 90:13 cell line and during slender parasitaemia the blood was collected by heart puncture (as described in Section 2.2.2).

The blood sample was transferred into 25 ml of HMI-9 medium (pre-warmed to 37°C, as was all the HMI-9 medium used for pleomorphic cell culture). These flasks

were left to stand upright at 37°C for a few hours to allow the red blood cells to settle. The media containing the trypanosome cells was removed, without disturbing the settled red blood cells, transferred into a flask of 25 ml of HMI-9, and another 25 ml of HMI-9 added to the settled blood cells. Again, these flasks were left to stand upright at 37°C. The following day, without disturbing the settled red blood cells, the media containing the trypanosome cells (about 75 ml) was transferred from both flasks into a fresh flask. The cell concentration was determined using a haemocytometer and the culture divided for the number of transfections to be performed. Cells were pelleted at 2000 rpm for 8 minutes and the supernatant discarded. The cells were resuspended with 80 µl AMAXA Nucleofector solution (Lonza) with the linearised DNA to be transfected. A negative control using sterile, distilled water instead of DNA was also performed in parallel. The resuspended cells were transferred to a cuvette (Lonza) and transfected using an AMAXA Nucleofector machine (Lonza) with programme X-001. The cells were then transferred into 25 ml of HMI-9 medium and 37°C.

A few hours following transfection, the cells were transferred into 24-well culture plates (Greiner Bio-one), in four different cell concentrations (1 ml of culture in each well). The following day, 1 ml of HMI-9 medium containing the selective drug at double the final concentration was added to each well. A range of different drug concentrations was applied for each cell concentration. For selection of transfectant cells with phleomycin, the drug concentration range was 1-2 µg/ml. As a control, cells in HMI-9 medium lacking the selective drug were also included in a culture plate. Further dilutions of cells in selective media were made later if necessary. Following the death of all the control cells not transfected with DNA, each well containing surviving transfectant cells was diluted into a larger culture volume (5 ml) with selective drug. Two surviving cultures were used to infect a mouse (as described in Section 2.2.2). For this, cells were pelleted from a small volume of the cell culture (typically 1 ml) by centrifugation at ~9000 rpm for ~4 minutes using a microfuge, and resuspended in about 200 µl HMI-9 medium for injection.

2.2.5 Differentiation to 'stumpy-like' form using 8-pCPT-2'-O-Me-cAMP

8- (4- Chlorophenylthio)- 2'- O- methyladenosine- 3', 5'- cyclic monophosphate (8-pCPT-2'-O-Me-cAMP; Biolog Life Science Institute) dissolved in DMSO (Dimethyl sulfoxide) to a concentration of 10 mM (briefly vortexed to ensure all the compound had dissolved) was added to a 25 ml cell culture (~5 $x10^5$ cells/ml concentration) to give a final concentration of 10 μM. As a control, to an identical cell culture the same volume of DMSO alone was added. Following incubation at 37°C for 48 hours, 5 ml of each culture was used to generate cell lysate samples for use in a CAT ELISA assay (as described in Section 2.6) and the remaining culture used for RNA extraction and then Northern blot analysis (as described in Section 2.4).

2.2.6 Differentiation to procyclic form using cis-aconitate

Monomorphic bloodstream form trypanosomes were induced to differentiate to procyclic forms *in vitro* in HMI-9 medium using 6 mM *cis*-aconitate (Sigma) and incubation at 27°C in 24-well culture plates (Greiner Bio-one). In parallel, a control was performed of cells untreated with *cis*-aconitate, this being performed for cells induced and uninduced for RNAi or ectopic protein expression, with each treatment condition performed in triplicate. Differentiation was assessed 72 hours post-induction for cells stained positive for EP-procyclin (see Section 2.8.3).

2.2.7 Cold shock treatment

The *T. b. brucei* Lister 427 cell line was used and cell lysate samples taken for Western blot analysis as described in Section 2.7. A cell lysate sample was taken at 0 hours, then the cells split into three groups and transferred to three temperatures (37°C, 27°C and 20°C). Cell lysate samples were taken at various time-points up to 5 hours after transfer (see Section 5.6.2).

2.3 Sub-cloning

2.3.1 Oligonucleotides

All oligonucleotides were obtained from Sigma-Aldrich and resuspended with distilled water. A list of the names and sequences of oligonucleotides (primers) used in this study can be found in Table 2.1.

Table 2.1 List of oligonucleotides (primers) used in this study.

Number	Name	Sequence (5'-3') [a]
1	M13 Forward	GTAAAACGACGGCCAGT
2	M13 Reverse	CAGGAAACAGCTATGAC
3	Aconitase 5'UTR F2	gtgctcgagtggggtcttggtccag
4	Aconitase 5'UTR R2	cccaagcttacacaagactccctcttc
5	Aconitase 3'UTR F2	ggcggatccggaacaccgaaacctg
6	Aconitase 3'UTR R2	ggggaagaccttgctggtttgctttttcggg
7	Aconitase 3'UTR 400nt F	cgggcggatcctaactacatatttggtacatg
8	Aconitase 3'UTR 300nt F	gcgcggatccagctgacttggtgcc
9	Aconitase 3'UTR 200nt F	ggcggatccgcgacggcttaacgag
10	Aconitase 3'UTR 100nt F	gggggcggatccggtgacaaaaacctttg
11	Aconitase 3'UTR 850nt R	ccgaagaccttgctgGTTCCCTCTACCTTATGC
12	ESAG9-EQ 3'UTR F2	cccggatccaaatcattaccgcaattcc
13	ESAG9-EQ 3'UTR R2	cccgaagaccttgctgttatcacgttccc
14	ESAG9-EQ 3'UTR 350nt F	ggccggatccaacaccgcatgatg
15	ESAG9-EQ 3'UTR 300nt F	ggcccggatcctgctttcaatgaag
16	ESAG9-EQ 3'UTR 250nt F	ggcgcggatccagaggatcgaataac
17	E3d R	ggggTCTAGAgttccatccacatgatcttcattg
18	E3d F	ggggTCTAGAactaatgtaatcataagtagcgtacatgag
19	ADAPT dT	GGCCACGCGTCGACTAGTACTTTTTTTTTTTTTT
20	AUAP	GGCCACGCGTCGACTAGTAC
21	CAT 3	cCGTGGCCAATATGGACAACTTC
22	CAT 4	GGTGCTGATGCCGCTGGCG
23	Aldolase 1.1 F	gggGGATCCTGCCCATTTAGTTAG
24	Aldolase 1.1 R	gcTCTAGACGAGACAAGGGAAAAGC
25	Aldolase 1.2 F	ggggTCTAGATGTCTTTTCCGTGGAAAGGTTC
26	Aldolase 1.2 R	GAAGACcttgctgGGGCACGAATCCCC
27	E3eI F	ggatctagagttgtggagagac
28	E3eI R	gggtctagagttattcgatcctc
29	E3eI F2	ctagggttgtggagagacaagataagaggatcgaataact
30	E3eI R2	ctagagttattcgatcctcttatcttgtctctccacaacg
31	E3eI GeneArt F	CGTCAAGGCCACGTGTC
32	E3eI GeneArt R	GAGGCCAGTCTTGTGCTC
33	Tb09.211.2300 5'UTR F	ccctcgagcgatgtgacttaatg
34	Tb09.211.2300 5'UTR R2	cccaagcttcaccacaaccacac
35	Tb11.47.0019 5'UTR F	ccctcgagtgttgttgttgttg
36	Tb11.47.0019 5'UTR R2	cccaagcttcacacgcaaagg
37	ESAG9-EQ 5'UTR F2	ccctcgaggtttacaatccaac
38	ESAG9-EQ 5'UTR R2	cccaagctttttcccccgac
39	Tb09m F	ggtggtgtgatatatacgtgtttggttggtg
40	Tb09m R	caccaaccaaacacgtatatatcacaccacc
41	Tb11m F	ctcatccgccgtcacgaggatagaaagg
42	Tb11m R	cctttctatcctcgtgacggcggatgag
43	EQm F	gtggtattttgtacgatttgttagtttacgtcgggggg
44	EQm R	cccccgacgtaaactaacaaatcgtacaaaataccac
45	Splice leader	CTATTATTAGAACAGTTTCTGTACTATATTG
46	Tb09 a2	cCAAAACTGTAGGAAATGGTATCG
47	Tb09 b2	CAGCATTTCTAAGCATGGGATC

48	Tb11 a2	ccCTCTAGCCGAGATGTTG
49	Tb11 b2	cGCTGACCCGAGATCATC
50	Tb09 a	gccTTGCCGAACACACTAACCAA
51	Tb09 b	gccCAGGTGCCATATATCGAAAC
52	Tb11 a	ccgAACAACAAAAGGCCTTCCCTC
53	Tb11 b	gggTCAACTGTCATCTCCGCAAC
54	as077	GCATAATGCCACTGCGCTTT
55	as076	GAAGTTCGCTCCATGAACCT
56	Tb09 a F	gggTACCAATGTGATGGTTGGTTAGTGTGTTCGGCAA
57	Tb09 a R	gggTACCATAGAGTTGGGCGAAAAGAAGTAAGGGAACA
58	Tb11 a F	gggTACCAATGTGATGGAGGGAAGGCCTTTTGTTGTT
59	Tb11 a R	gggTACCATAGAGTTGGACGGCACGTTGGAGGTAATA
60	Tb09 b F	gggTACCAATGTGATGGCGGTGAGGCTCAGAAGGTAG
61	Tb09 b R	gggTACCATAGAGTTGGGTGAGTTCTTCCTGAAGCGG
62	Tb11 b F	gggTACCAATGTGATGGGTTGCGGAGATGACAGTTGA
63	Tb11 b R	gggTACCATAGAGTTGGTGATCAGGGTCGATTTGTGA
64	Tb11 long 451 F3	cCTCGAGATGTGTGTCTGTTTGTGTTGTC
65	Tb11 451 R	ccGGATCCAGAGCTATCTTTGTTCACAATG
66	Tb11 short 451 F	cccAAGCTTATGAATGTTTTTGATTCCAACAGC
67	Tb11a 1.1 F	gggaagcttAGGGAAGGCCTTTTGTTGTT
68	Tb11a 1.1 R	gggtctagaACGGCACGTTGGAGGTAATA
69	Tb11a 1.2 F	gggctcgagACGGCACGTTGGAGGTAATA
70	Tb11a 1.2 R	gggggatccAGGGAAGGCCTTTTGTTGTT
71	Tb11b 1.1 F	gggaagcttGTTGCGGAGATGACAGTTGA
72	Tb11b 1.1 R	gggtctagaTGATCAGGGTCGATTTGTGA
73	Tb11b 1.2 F	gggctcgagTGATCAGGGTCGATTTGTGA
74	Tb11b 1.2 R	gggggatccGTTGCGGAGATGACAGTTGA
75	TbeIF4E1 F	gggTACCAATGTGATGGaaagctggtgggaagttgtg
76	TbeIF4E1 R	gggTACCATAGAGTTGGgctgtgaagatgggtttcgt
77	TbeIF4E2 F	ggcTACCAATGTGATGGGCAGCAACTGCTACACCAAA
78	TbeIF4E2 R	gggTACCATAGAGTTGGCTTAACTTCCAGTCTCGCCG
79	TbeIF4E3 F	gggTACCAATGTGATGGGTTGGGGATGTTGAGTGCTT
80	TbeIF4E3 R	gggTACCATAGAGTTGGTCGACGGTTTTCCTTTCATC
81	TbeIF4E4 F	gggTACCAATGTGATGGAGCAACAGCAAACAGTGGTG
82	TbeIF4E4 R	gggTACCATAGAGTTGGAGGGTTTCGTCATGTTGGAG
83	TbK1 F	ggcTACCAATGTGATGGATCGTCCGGTACTACGATGC
84	TbK1 R	gggTACCATAGAGTTGGATTTTCGTTGCCACCTTGTC
85	TbK2 F	gggTACCAATGTGATGGCCTCGTTGGCTTTGGTATGT
86	TbK2 R	gggTACCATAGAGTTGGGAAAGAGTTGTGGCATGGGT
87	TbK3 F	ggcTACCAATGTGATGGTGGTGTTTTGTACGCCGTTA
88	TbK3 R	gggTACCATAGAGTTGGTGACCTTGTGTCGTCAGAGC
89	1770 pQ F	gggTACCAATGTGATGGTCTGACGATGTGGGTGTGTT
90	1770 pQ R	ggcTACCATAGAGTTGGTCTGCAATGGATGTACGGAA
91	4900 pQ F	ggcTACCAATGTGATGGCGATGGAACATGACATACGC
92	4900 pQ R	gggTACCATAGAGTTGGGTTTCAGTCCCTCTTCGCTG
93	INGI F	AAGGCCGCAGATTTACCACAG
94	INGI R	AGCGACGTTTCTTCGCAGTC
95	1770 451 F	GGCAAGCTTATGACACTGCGTACGCGCTTCGAGAGCTC
96	1770 451 R	GGGGATCCTGCCAGTTCATCCACCAACGTATCGCGC

[a] Restriction enzyme site sequences are underlined

2.3.2 Plasmids

The *T. brucei* expression vector, pHD449 (Biebinger *et al.*, 1997), was previously modified through replacement of the tetracycline resistance cassette present in this plasmid with the chloramphenicol acetyltransferase (CAT) coding region, to create the plasmid, CAT449 (Mayho *et al*., 2006). This plasmid contains the phleomycin resistance gene (*BLE*). The pQuadra vector system (Inoue *et al.*, 2005) was used for RNAi-mediated transcript ablation and is outlined in the schematic diagram shown in Appendix 2. Alternatively, pALC14 (Bochud-Allemann and Schneider, 2002) was used for RNAi-mediated transcript ablation. For ectopic overexpression of protein with a Ty epitope tag at the C terminus, the pHD451 vector that had been previously modified to include a Ty epitope tag sequence downstream of insertion site was used (Hendriks *et al.*, 2001). Each of these plasmids can be linearised with a *Not*I restriction enzyme digestion reaction (see Section 2.3.7).

The commercially available plasmid, pGEM-T Easy (Promega), was used for ligation with each PCR amplicon and following small-scale plasmid preparation (see Section 2.3.10) was sequenced (The Genepool, University of Edinburgh) with M13 forward and reverse primers (see Table 2.1) to establish insertion copy number, orientation and sequence.

2.3.4 Genomic DNA extraction

Genomic DNA was extracted from trypanosomes using DNeasy Blood & Tissue kit (Qiagen) according to the manufacturer's instructions for Animal Blood (Spin-Column Protocol), except that a lower amount of RNase was added (about 15 µl 10 mg/ml RNase per sample).

2.3.5 PCR amplification

DNA sequences were amplified in a polymerase chain reaction (PCR) using either GoTaq Hot Start polymerase (Promega) or the Expand High Fidelity PCR system (Roche). In both cases, the total reaction volume was 50 µl, containing 2 µl of each of a forward primer (25 µM), reverse primer (25 µM; see Table 2.1 for sequences), and genomic DNA as template. When using the GoTaq Hot Start polymerase,

reactions were made according to the manufacturer's instructions, containing 10u/µl GoTaq Hot Start polymerase (Promega) and 3 µl 25mM $MgCl_2$ (Promega). When using the Expand High Fidelity polymerase, the reaction was made according to the manufacturer's instructions using the Expand High Fidelity buffer containing $MgCl_2$ (Roche). A reaction containing water instead of genomic DNA (water control) and another reaction containing forward and reverse primers previously determined to result in a product (positive control) were performed in parallel with the experimental reaction(s). PCR cycles started with a denaturing step of 95°C for 5 minutes, then 30 cycles of 95°C for 30 seconds (20 seconds for the Expand High Fidelity PCR system), 45 seconds (30 seconds for the Expand High Fidelity PCR system) at the annealing temperature, and 72°C for 2 minutes, followed by a single step of 72°C for 4 minutes.

The product from each PCR reaction was analysed on an agarose gel (typically 1 or 1.2% w/v) made with TAE buffer containing ethidium bromide (0.4 µg/ml), after electrophoresis at ~120 volts. If products of various molecular weights were observed, the product of the correct molecular weight was excised from the agarose gel and purified using the Nucleospin Extract II kit (Macherey-Nagel) according to the manufacturer's instructions. Each PCR amplicon was then ligated (using either the PCR reaction directly or purified gel product) into the pGEM-T Easy plasmid (see Sections 2.3.2 and 2.3.9).

2.3.6 Oligonucleotide annealing reaction

Both oligonucleotides (each 100µM) to be annealed were mixed with distilled water and annealing buffer (final concentration: 10mM Tris, pH7.5-8.0, 50mM NaCl, 1mM EDTA) to obtain a final concentration of 25µM of each oligonucleotide. This was incubated at 90-95°C for 4 minutes, then allowed to slowly cool to room temperature over a period of about 45-60 minutes and transferred to 4°C.

2.3.7 Restriction enzyme digest

Plasmids were digested in double-enzyme digest reactions using 0.2u of each restriction enzyme per µl of reaction, compatible buffer, plasmid and BSA (New

England Biolabs). All restriction enzymes were obtained from Promega, except for *Bbs*I, which was obtained from New England Biolabs. Reactions were incubated at 37°C, usually overnight but at least for a minimum of 4 hours. The products of the digest were examined on an agarose gel as above (Section 2.3.5), with the product for ligation being excised and purified using the Nucleospin Extract II kit (Macherey-Nagel) according to the manufacturer's instructions.

2.3.8 CIAP treatment

The pQuadra3 plasmid required Calf Intestinal Alkaline Phosphatase (CIAP) treatment following digestion before use in a ligation reaction. The DNA was added to a reaction containing 0.01u of CIAP (Promega) per µl of reaction performed according to the manufacturer's instructions, except that the Stop buffer was not used and instead the DNA was gel-purified using the Nucleospin Extract II kit (Macherey-Nagel).

2.3.9 Ligation

Ligation of digested and purified DNA for insertion and vector was performed in a 10 µl reaction with T4 DNA ligase (Promega) according to the manufacturer's instructions, using 1µl of vector, 7.5 µl of DNA for insertion, and buffer (T4 DNA ligase 10X buffer or 2X Rapid Ligation buffer, both Promega). Except ligation using the pQuadra vector system contained 0.5 µl of pQuadra3 plasmid, 2 µl of pQuadra1 plasmid and 5.5 µl of DNA for insertion. A negative control reaction was performed in parallel, which included distilled water instead of the insert DNA in the reaction. Reactions were incubated either overnight at 4°C or at room temperature for a minimum of 4 hours.

2.3.10 Transformation of bacteria and small-scale plasmid preparation

Transformation of *Escherichia coli* XL1-blue bacteria (genotype: *Sup*E44, hsdR17, *rec*A1, *end*A1, *gyr*A46, *thi*, *rel*A1, *lac*, F' [*pro*AB^{+}, *lac*I^{q}, *lac*ZDM15, Tn10(*tet*^{r})]) was performed with each ligation reaction. 100 µl of competent bacterial cells were incubated with 5 µl of ligation reaction on ice for 30 minutes. The cells were then heat-shocked at 42°C for 30 seconds. Either 100-250 µl of LB (Luria Bertani)-broth

was added to the cells and they were allowed to recover by incubation with shaking at 37°C for up to 45 minutes and then plated out on LB-Agar plates containing 100 µg/ml amplicilin, or they were plated out immediately following the heat-shock step. The plates were incubated at 37°C overnight.

Individual colonies were selected, transferred into 3 ml LB-broth with 100 µg/ml amplicilin, and incubated overnight at 37°C with shaking. 1.5 ml of the cell culture was pelleted at 13 000 rpm for 5 minutes and resuspended with 100 µl of Solution I (see Appendix 1). 200 µl of Solution II (see Appendix 1) was added, mixed by inversion several times and incubated for 4-5 minutes at room temperature. Then, 150 µl of ice-cold Solution III (see Appendix 1) was added, again mixed by inversion and incubated on ice for 5 minutes. The sample was centrifuged at 13 000 rpm for 15 minutes and to the supernatant 900 µl of ice-cold 100% ethanol was added, the solution mixed by inversion several times and incubated on ice for a minimum of 20 minutes. Following this, the sample was centrifuged at 13 000 rpm for 10 minutes and then the pellet was washed with 70% ethanol. After removal of the supernatant, the pellet was allowed to air-dry and then resuspended with 40 µl distilled water and 1 µl ribonuclease A (10 mg/ml; Sigma).

2.3.11 Large-scale plasmid preparation

For plasmid amplification, 200 µl of the bacterial culture used for the small-scale plasmid preparation was added to 200 ml LB-broth containing 100 µg/ml Amplicilin and incubated overnight at 37°C with shaking. DNA extraction and preparation was performed using the Qiagen Plasmid Midi kit according to the manufacturer's instructions. The purified plasmid pellet was resuspended with 200 µl TE buffer.

2.3.12 Site-directed mutagenesis

To perform site-directed mutagenesis of an uORF start codon from ATG to ACG, the QuikChange XL Site-Directed Mutagenesis Kit (Stratagene) was used according to the manufacturer's instructions. Briefly, the CAT449 plasmid containing the sequence to be mutagenised was used in a mutant strand synthesis reaction with the cognate two complimentary oligonucleotides containing the appropriate T to C

mutation. The primers that were used are numbers 39-44 in Table 2.1, which had been obtained from Sigma Aldrich as purifications by polyacrylamide gel electrophoresis (PAGE). The amplification product was then digested with *Dpn*I restriction enzyme. Following digestion, 2 μl was used in a transformation of supplied XL-10-Gold ultracompetent cells, as per the manufacturer's instructions, except that LB-broth was used instead of NZY^+ broth and LB-Agar plates containing 100 μg/ml amplicilin were used. Resulting colonies were screened to confirm the desired mutation following small-scale plasmid preparation, as described above (Section 2.3.10), and sequencing (The GenePool, University of Edinburgh) using primer numbers 33 and 34, or 35 and 36, or 37 and 38 (see Table 2.1), as appropriate.

2.4 Northern blot analysis

All apparatus used for electrophoresis, blotting, hybridisation and detection were cleaned thoroughly by washing with detergent, rinsing with ultra pure water, rinsing with 70% ethanol, and rinsing with ultra pure water.

2.4.1 RNA extraction

RNA extraction was performed using the Qiagen RNeasy kit according to the manufacturer's instructions. Logarithmically growing trypanosomes were pelleted at 2000 rpm for 10 minutes, resuspended with 600 μl RLT buffer (Qiagen) containing β-mercaptoethanol (10 μl β-mercaptoethanol for 1 ml RLT buffer) and stored overnight at -80°C. The sample was thawed and vortexed for 1 minute, then RNA was purified using the manufacturer's instructions for 'animal cell spin' RNA purification. Elution of RNA was performed with 30 or 40 μl nuclease-free water. RNA concentration was determined using a NanoDrop Spectrophotometer ND-1000.

For the bloodstream form monomorphic cell total RNA sample for use with Illumina digital tag RNA sequencing, trypanosomes were washed twice with filter-sterilised PSG prior to RNA extraction as described above. This was identical for the stumpy form total RNA sample for use with Illumina digital tag RNA sequencing, except that the trypanosomes were only washed once following purification of the cells

from murine blood (as described in Section 2.2.2). Also, for the Illumina digital tag RNA sequencing samples, the optional on-column DNase-treatment (RNase-Free DNase Set, Qiagen) was included in the Qiagen RNeasy kit protocol for 'animal cell spin' RNA purification.

2.4.2 Electrophoresis of RNA samples

An RNA gel was made as follows, firstly, 2.4 g of agarose was completely dissolved in 160 ml distilled water, then 20 ml of 10X MOPS buffer (see Appendix 1) and 6ml of 37% formaldehyde were added. The solution was gently mixed and then allowed to set in a gel tray.

Equivalent amounts of RNA were loaded per lane (in the range of 0.5-2.5 μg per lane, in a maximum volume of 8 μl), each with 9 μl of formamide, 3 μl of 37% formaldehyde, 2 μl of 10X MOPS buffer and 2 μl of RNA loading buffer (see Appendix 1). Immediately prior to loading, the samples were heated at 65°C for 5 minutes. The gel was electrophoresed at 150V for 90 minutes in MOPS buffer (1X concentration). The gel was then stained with 25 μl ethidium bromide (10 mg/ml) in 200 ml 1X MOPS for 15 minutes with gentle agitation at room temperature. To destain the gel, three washes were performed with distilled water, for a total time of about 1 hour 45 minutes. The UV transilluminator setting of a GBOX (Syngene) was used to visualise the rRNA bands and a photograph was taken, so later the amount of rRNA in each lane could be quantified using GeneTools software (Syngene).

2.4.3 Blotting

RNA was transferred on to a nylon membrane using the upward capillary transfer method. A raised platform was created in a plastic container containing 10X SSC (see Appendix 1), on top of this platform a piece of chromatography paper (pre-wet in 10X SSC) was placed with two ends of the paper submersed in the 10X SSC liquid below. The gel containing the electrophoresed RNA was then placed on top (so that the open-wells of the gel contacted the filter paper). Saran wrap was used to cover this apparatus and the gel to prevent evaporation of the liquid, and a window slightly smaller than the size of the gel cut out of the saran wrap on the top of the gel.

Positively-charged nylon membrane (Roche) (pre-wet in 10X SSC) was placed on top of the gel, followed by two layers of chromatography paper (pre-wet in 2X SSC) slightly larger than the size of the membrane. Air bubbles were then removed from the layers by rolling a 10 ml pipette over the top while applying mild downward pressure. Finally, several layers of paper tissue were placed on top, followed by a Perspex plate, then a weight of about 500 g, all placed to ensure equal distribution of weight downwards on to the gel. The transfer was allowed to take place overnight. The following day RNA was cross-linked on to the membrane at 0.12 joules using UV-light (UV Crosslinker, Uvitec). The cross-linked membrane was kept dry and stored at room temperature in the dark until prehybridisation and the remainder of the Northern blot analysis protocol.

2.4.4 Riboprobe generation

Primers to amplify around a 200-400 nucleotide region of sequence specific to the gene of interest were designed using the freely available RNAit programme (http://trypanofan.path.cam.ac.uk/software/RNAit.html). These were used in a standard PCR with genomic DNA and the product of the correct size ligated into pGEM-T easy plasmid for sequencing as outlined in Section 2.3.5 above. The small-scale preparation of the plasmid that contained insert of the correct sequence was used as template in a standard PCR (two 50 μl reactions; see Section 2.3.5) with M13 forward and reverse primers (2 μl of each at 10 μM concentration per 50 μl reaction; Table 2.1). 5 μl of the product of this reaction was electrophoresed in a standard agarose gel (see Section 2.3.5). The remainder of the reaction was purified using the Nucleospin Extract II kit (Macherey-Nagel), or if multiple products were observed, the remainder of the reaction was used to gel-excise the product of the correct size, and this was purified using the same kit (see Section 2.3.5).

1 μg of the purified PCR product was used in a 20 μl reaction with either SP6 or T7 polymerase (Roche), depending of the orientation of the insert in the pGEM-T easy plasmid, to generate the anti-sense sequence, using a DIG RNA labeling kit (Roche) according to the manufacturer's instructions and including the optional DNase treatment step. After stopping the reaction by the addition of 2 μl of 0.2M EDTA

(pH8.0), 2.5 μl of 4M lithium chloride and 75 μl of cold 100% ethanol was added. This was mixed and incubated at -80°C for a minimum time period of overnight, then centrifuged for 15 minutes at 12 000 g at 4°C. The pellet was washed with 70% ethanol, then air-dried for 10 minutes and resuspended with 50 μl of distilled water and 1 μl of RNase inhibitor (Promega). The riboprobe was stored at -80°C and checked before use by electrophoresis on a standard agarose gel (see Section 2.3.5).

2.4.5 Riboprobe hybridisation and detection

The membrane was incubated for 1 hour at the hybridisation temperature (68°C) with 10 ml of hybridisation buffer (see Appendix 1). 2 μl of a riboprobe was added to 100 μl of hybridisation buffer and denatured by boiling for about 8 minutes. This was added to 7 ml of hybridisation buffer (pre-warmed to the hybridisation temperature) and added to the membrane after removal of the solution not containing the riboprobe. The membrane was incubated overnight at the hybridisation temperature. The following day, the membrane was washed for around 30 minutes with 2X SSC/0.1% SDS twice, and then once with 0.5X SSC/0.1% SDS for around 45 minutes, with each wash being at the hybridisation temperature. Then, the following steps were performed at room temperature with gentle agitation, using freshly cleaned boxes at each step; the membrane was blocked for 1 hour with 50 ml Maleic acid buffer (see Appendix 1) containing 1% DIG block solution (see Appendix 1), incubated for 30 minutes with 50 ml Maleic acid buffer containing 1% DIG block solution and 2 μl anti-DIG antibody (Anti-Digoxigenin-AP Fab fragments, Roche), then washed for 10 minutes with 50 ml Maleic acid buffer containing 0.3% Tween-20 (Fisher Scientific) three times, and finally incubated for 2 minutes in detection buffer (100 mM Tris-HCl pH 9.5, 100 mM NaCl).

The membrane was incubated for 2 minutes in Bag W with an appropriate amount of CDP-*star* (25 mM Roche; 1:100 dilution with detection buffer) for the membrane size (usually 0.5 ml or 1 ml). CDP-*star* solution was removed by applying slight pressure across the outside of the Bag W with tissue. The Bag W was then heat-sealed and incubated for 15 minutes at 37°C, before being cut to remove the sealing and exposed either to X-ray film or using a GBOX (Syngene; for signal

quantification by normalisation to the rRNA, see Section 2.4.2). In both cases the Bag W was taped down around the cut edges to prevent drying of the membrane. Following exposure, the Bag W edges were re-sealed and the membrane was stored at -20°C. If the membrane was re-exposed using the alternative exposure method the membrane was re-processed from the step of incubation in detection buffer.

2.5 Splice leader addition site and polyadenylation site mapping

Throughout the following procedure autoclaved eppendorf tubes, filter pipette tips and nuclease-free water were used where appropriate.

RNA was prepared from the various cell lines as described in 2.4.1 and used in a cDNA reaction as follows. 2 µg of RNA was mixed with nuclease-free water to give a volume of 22 µl, to this the following was added in the given order: 2 µl of primer, 8 µl of AMV RT 5X buffer (Promega), 2 µl of RNase inhibitor (RNasin, Promega), 4 µl of 2.5mM dNTPs, 2 µl of AMV reverse transcriptase (Promega), and DTT (0.1M, 1 µl per 100 µl). The primer used was either random hexamers (Promega) for 5'RACE, or an oligo-dT 3' primer containing an adapter sequence (primer 19, see Table 2.1) for 3'RACE. The reaction was incubated at 37°C for 10 minutes, then 42°C for 1 hour. An identical reaction was also performed in parallel for each RNA sample, termed an 'RT minus' control reaction that contained nuclease-free water instead of reverse transcriptase, to control for the amplification of contaminating genomic DNA in the subsequent PCR reactions. Additionally, another control reaction was performed in parallel, which was identical except for that nuclease-free water was used instead of an RNA sample.

The first PCR was a standard reaction, performed using GoTaq Hot Start polymerase (Promega) and 50°C annealing temperature (see Section 2.3.5) with 2 µl of each cDNA reaction or control cDNA reaction. A second (nested) PCR was then performed as described for the first PCR but with 2 µl of each first PCR instead of cDNA reaction. Also performed in parallel with each PCR was a negative control

reaction, which was identical except for the addition of water instead of cDNA template.

See Sections 3.4.1 and 4.4.3 for the primers used in mapping PAS and SAS. A schematic diagram of the procedure used to map SASs is shown in Figure 4.5 (A).

2.6 CAT ELISA

2.6.1 Cell lysis

For the chloramphenicol acetyltransferase (CAT) enzyme-linked immunosorbant assay (ELISA) logarithmically growing cells were used at a concentration of 1-4 x 10^6. For this protocol, a CAT ELISA kit (Roche) was used according to the manufacturer's instructions. Where possible, samples from cell lines possessing different CAT constructs to be compared were processed together. To allow normalisation of values to cell number at the end of the assay, cell cultures were counted directly before the following procedure. 5ml of the culture was pelleted and washed 3 times with chilled phosphate buffered saline (PBS); centrifugations were performed at 2000 rpm for 10 minutes at 4°C and cell pellets resuspended inbetween centrifugations. To lyse the cells, 1 ml of CAT ELISA lysis buffer (Roche) was added, the pellet resuspended and incubated for 25 minutes at room temperature. Following the incubation, lysed cells were centrifuged as above and two 500 µl aliquots of supernatant were snap-frozen in liquid nitrogen and stored at -80°C for later use in the ELISA.

2.6.2 ELISA

Cell lysate samples were thawed on ice and from each 500 µl aliquot two samples were used for replicates. Standard dilutions were prepared according to the manufacturer's instructions, each in duplicate and the results used to generate a calibration curve. Any dilutions of samples were made with sample buffer. Of each standard dilution or cell lysate sample, 200 µl was added to a well of an anti-CAT antibody-coated ELISA microplate and incubated at 37°C for one hour. Following incubation, each well was washed with 250 µl washing buffer 5 times. Then, 200 µl

of anti-CAT antibody conjugated to digoxigenin solution was added and the microplate incubated and washed as above. Next, 200 µl of anti-DIG-peroxidase antibody solution was added and again the microplate was incubated and washed as above. Finally, 200 µl of peroxidise substrate was added and the absorbance of each sample at 405 nm was measured using a BioTek ELx808 ELISA microplate reader.

2.6.3 CAT ELISA data analysis

The absorbance readings were recorded with Gen5 software (BioTek) and this was also used to generate a calibration curve. The r^2 value obtained was greater than 0.99 in each CAT ELISA assay, except for one instance when it was 0.977. The background absorbance value was determined as the mean of duplicate samples of sample buffer alone. This value was subtracted from the experimental absorbance values and the mean of duplicate samples calculated. This value was then divided by the cell concentration measured when the samples were generated (see Section 2.6.1).

2.7 Western blot analysis

2.7.1 Protein sample preparation

Trypanosome cultures were counted and then centrifuged at 2000 rpm for 10 minutes. The pellet was washed with 5 ml PBS and then resuspended with boiling Laemmli buffer (62.5 mM Tris-HCl pH6.8, 2% SDS, 10% glycerol, bromophenol blue [2 mg in 100ml], 5% β-mercaptoethanol) to give 0.4 x 10^6 cells/ml. Protein samples were then boiled for 3 minutes and stored at -20°C.

2.7.2 Electrophoresis of protein samples

Different percentage SDS-PAGE gels were made according to the component concentrations in Appendix 1. First a running gel was allowed to set between 1.5 mm glass plates, to about 1-1.5 cm from the top, with 70% ethanol added above the running gel solution. The 70% ethanol was removed, a stacking gel solution added above the running gel and a 10- or 15-well 1.5 mm comb inserted into the top of the space between the plates before allowing the stacking gel to set. Protein samples

were boiled for 5 minutes prior to loading on the SDS-PAGE gel. The samples were electrophoresed in running buffer (25 mM Tris-HCl pH8.3, 192 mM glycine, 0.1% SDS) at 150V until the dye-front started to run out of the base of the gel.

2.7.3 Semi-dry transfer of proteins

Electrophoresed proteins were transferred on to PVDF membrane (Immobilon-P, Millipore) using the semi-dry transfer method. In a Semi-dry Transfer Cell (Bio-Rad) a piece of extra thick blot paper (Bio-Rad) wet in semi-dry transfer buffer (see Appendix 1) was placed. On top of this the PVDF membrane was added (prior to this the membrane had been placed in methanol for about 15 seconds, then in distilled water for about 2 minutes and then semi-dry transfer buffer for about 5 minutes), and then the SDS-PAGE gel (which had been rinsed with distilled water). Finally, another piece of extra thick blot paper wet in semi-dry transfer buffer was placed on top and any air bubbles between the layers removed by rolling across the top of the blot paper using 10 ml pipette and applying slight downward pressure. The transfer was performed at 0.14 Amp (constant) for 90 minutes. Visualisation of transferred proteins was achieved by brief incubation of the membrane with Ponceau stain (0.4% Ponceau S [Sigma], 3% trichloroacetic acid [TCA]), then background staining removed by brief washes with distilled water. An image of the bands visible on the stained membrane was obtained with an Epson Perfection V200 Photo scanner. The membrane was then fully destained with PBS containing 0.1% Tween-20 (Fisher Scientific).

For membranes for use with the eIF6 antibody (BD Biosciences), the procedure outlined above was followed, except for the following modifications: the semi-dry transfer was at 1.2 mAmp/cm^2 for 1 hour and the Ponceau stain was removed entirely with distilled water.

2.7.4 Antibody hybridisation and detection

Typically, membranes were blocked by incubation with PBS containing 0.1% Tween-20 (Fisher Scientific) and 5% w/v non-fat dry milk (Marvel) for either 1 hour at room temperature or overnight at 4°C. The membrane was then incubated with the

primary antibody diluted in PBS containing 0.1% Tween 20 and 5% dry milk (Marvel) for 1 hour at room temperature with agitation. Following this the membrane was washed with PBS containing 0.1% Tween-20 for 30 minutes with agitation, changing the buffer every 10 minutes. Next, the membrane was incubated with the secondary antibody diluted in PBS containing 0.1% Tween 20 and 5% dry milk (Marvel) for 1 hour at room temperature with agitation. Again, the membrane was then washed with PBS containing 0.1% Tween 20 for 30 minutes with agitation, changing the buffer every 10 minutes. The ECL Plus Western Blotting Detection System (GE Healthcare) was used according to the manufacturer's instructions to detect bound antibody. Fluorescent signal was detected using a GBOX (SynGene) and GeneSnap software (SynGene) or by exposure to X-ray film.

For detection of eIF6 protein with the eIF6 antibody (BD Biosciences), the following protocol was followed. The membrane was blocked by incubation with blocking buffer (10 mM Tris pH7.5, 100 mM NaCl, 0.1% Tween 20, 5% w/v non-fat dry milk [Marvel]) for either 1 hour at room temperature or overnight at 4°C. The membrane was then incubated with the primary antibody diluted in blocking buffer for 1 hour at room temperature with agitation. Following this the membrane was washed with wash buffer (10 mM Tris pH7.5, 100 mM NaCl, 0.1% Tween 20) for 30 minutes with agitation, changing the buffer every 5-10 minutes. Next, the membrane was incubated with the secondary antibody diluted in blocking buffer for 1 hour at room temperature with agitation. Again, the membrane was then washed with wash buffer for 30 minutes with agitation, changing the buffer every 5-10 minutes. The ECL Plus Western Blotting Detection System (GE Healthcare) was then used to detect bound antibody as outlined above.

The dilutions of the antibodies used in this study are given in Table 2.2.

Table 2.2. Dilution of antibodies used in Western blot analysis, immunofluorescence microscopy and fluorescence-activated cell sorting (FACS).

Antibody	Dilution
Western blot analysis	
BB2	1:5
α-eIF6 (BD Bioscience)	1:1000
α-mouse IgG-HRP (Sigma)	1:5000
α-rabbit IgG-HRP (Sigma)	1:5000
Immunofluorescence microscopy	
BB2	1:50
α-mouse-FITC (Sigma)	1:200
Fluorescence-activated cell sorting (FACS)	
α-EP-procyclin	1:500
α-mouse IgG-FITC (Sigma)	1:500

2.8 Immunofluorescence

2.8.1 Methanol-fixation of cells and DAPI staining for microscopy

Typically, 1 ml of cell culture was centrifuged at 7000 rpm for 3 minutes using a microfuge. The supernatant was removed to leave about 40 µl and the cells resuspended in this remaining solution. About 20 µl of this cell solution was spread on to a glass microscope slide using the side of a pipette tip and allowed to air-dry, before the slide was immersed in ice-cold methanol for storage at -20°C.

The methanol-fixed cell slides were rehydrated in PBS for 10 minutes. Then the slide was incubated for 2 minutes in a humidity chamber with 50 µl of DAPI (1 µg/ml). The slide was washed once in PBS for 5 minutes and a coverslip was fixed to each slide with 50 µl of PDA (p-phenylenediamine; 10 mg/ml) and mowiol (see Appendix 1) mix (50 µl PDA and 500 µl mowiol). This was left to dry at room temperature for around 1-2 hours, then transferred to 4°C for storage in the dark. Slides were analysed on an Axioskop 2 plus microscope (Zeiss) and micrographs captured using QCapture software (QImaging).

2.8.2 Paraformaldehyde-fixation of cells and antibody staining for microscopy

For each slide, 1 x 10^7 cells were centrifuged at 1000 rpm for 10 minutes, the pellet washed with 5 ml vPBS (see Appendix 1), and centrifuged again at 1000 rpm for 10 minutes. The pellet was resuspended with 3 ml vPBS and 1 ml of 8% paraformaldehyde and incubated for 10 minutes at 4°C. Cells were pelleted at 2000 rpm for 10 minutes, washed with 5 ml vPBS and re-pelleted. The cells were resuspended in about 150 µl vPBS, spread on a polysine microscope slide, and left for 20 minutes in a humidity chamber. The liquid was poured off the slide, 200 µl of PBS with 0.1% Triton X-100 (Sigma) added and incubated for 2 minutes. After being washing twice for 5 minutes in PBS, the slide was then incubated for 10 minutes in a humidity chamber with 150 µl PBS containing 2% BSA (Bovine Serum Albumin; PAA). Then, stained with 150 µl of primary antibody diluted in PBS with 2% BSA for 60 minutes in a humidity chamber. Following three washes of 10 minutes in PBS, the slide was stained with 150 µl of secondary antibody diluted in PBS with 2% BSA for 60 minutes in a humidity chamber, and again washed three times for 10 minutes with PBS. The slide was then stained with 50 µl DAPI (1 µg/ml) for 2 minutes in a humidity chamber, washed for 5 minutes with PBS and a coverslip fixed with 50 µl of PDA and mowiol as above (see Section 2.8.1). Slides were analysed on an Axioskop 2 plus microscope (Zeiss), micrographs captured using QCapture software (QImaging) and images processed using ImageJ (http://imagej.nih.gov/ij) and Adobe Photoshop.

For the dilutions of the antibodies used in this study see Table 2.2.

2.8.3 Fluorescence-activated cell sorting

Cells were induced to differentiate as described in Section 2.2.6. As a positive control for procyclin expression, 1 ml of an *in vitro* procyclic form cell culture was used. Centrifugations in the following procedure were performed at 2000 g for 5-10 minutes. For each cell culture 1 ml was transferred to a FACS tube (5ml Polystyrene Round-Bottom Tube; BD Falcon) and the cells washed twice with PBS. The cells were then resuspended with 500 µl of fix solution (PBS with 2% formaldehyde, 0.05% gluteraldehyde) and incubated overnight at 4°C. Again, the cells were washed

twice with PBS, resuspended with 200 µl of blocking solution (PBS with 2% BSA) and incubated for 30 minutes at 4°C. After a wash step with PBS, the cells were resuspended with 200 µl of primary antibody (diluted in blocking solution; for antibody dilution see Table 2.2) and incubated at 4°C for hour. This wash and incubation step was repeated with the secondary antibody (diluted in blocking solution). Finally, the cells were washed with PBS and resuspended with ~500 µl of PBS and stored overnight at 4°C until the remainder of the protocol.

About 15-30 minutes prior to cell analysis, 10 µl of DAPI (1 µg/ml) was added to the cell sample. Cell analysis was performed using a BD FACS LSR II flow cytometer (BD Biosciences) and BD FACSDiva software. Data analysis was performed using FlowJo software (with assistance from Martin Waterfall, University of Edinburgh; see Section 5.7).

2.9 Polysome analysis

2.9.1 Harvest and lysis of cells

For the Illumina digital tag sequencing experiment:

For the 427 bloodstream monomorphic cell samples, cells were grown in 20 flasks of 50 ml culture volumes to a cell density of 1-2 x10^6 cells/ml. The cells were incubated with cycloheximide (100 µg/ml) for 10 minutes at 37°C. Then, pelleted by centrifugation at 2500 rpm for 10 minutes and washed twice in ice-cold PSG containing 100 µg/ml cycloheximide. Groups of 5 flasks were pooled after the first pelleting step to give 4 samples for polysome analysis.

For the stumpy cell sample, AnTat1.1 cells were purified from the blood of five mice 6 days post-infection as described in Section 2.2.2, cell density was determined and then the cells were incubated with cycloheximide (100 µg/ml) for 10 minutes at 37°C. The cells were then pelleted by centrifugation at 2500 rpm for 10 minutes and washed once with ice-cold PSG containing 100 µg/ml cycloheximide.

For the RNAi/ectopic protein expression experiments:

At 24 or 48 hours prior to cell harvest, tetracycline (1 or 10 µg/ml) was added to 250 ml of cells at a concentration to provide a cell concentration of ~1-3 x10^6 cells/ml at the time of cell harvest. As a control, an identical group of cells were established that were untreated with tetracycline (uninduced control), which was processed in parallel with the induced cells. At the time of cell harvest, cells were incubated with cycloheximide (100 µg/ml) for 10-15 minutes at 37°C. The cells were pelleted by centrifugation at 2500 rpm for 10 minutes at 4°C and washed twice in ice-cold PBS containing 100 µg/ml cycloheximide.

For both types of experiment:
Each cell pellet was then resuspended with 750 µl of polysome buffer (120mM NaCl, 40mM TrisHCl pH7.0, 2mM $MgCl_2$) containing 100 µg/ml cycloheximide, 0.75 µl RNase inhibitor (RNasin, Promega), protease inhibitor cocktail (1 tablet per 15 ml polysome buffer; Roche) and 0.1mM DTT (DL-Dithiothreitol; Sigma). The samples were then snap-frozen in liquid nitrogen and stored at -80°C for later use.

On the day of sucrose gradient preparation and fractionation, cell samples were lysed (unthawed) through sonication for ~3-6 minutes in a water bath with sonication (Bandelin Sonorex) with brief periods of vortexing. 5 µl of sample was removed to confirm cell lysis by microscopic examination. Cell lysates were then centrifuged at 12 000 rpm for 4 minutes at 4°C and the supernatant transferred to a fresh eppendorf tube containing 1 µl of RNase inhibitor (RNasin).

2.9.2 Sucrose gradient preparation

50% and 15% sucrose solutions (w/v; sucrose in polysome buffer) were sterilized through a single-use syringe filter (0.20 µm; Sartorius Stedim Biotech) and completed with 100 µg/ml cycloheximide, 0.5 µl/ml RNase inhibitor (RNasin, Promega), protease inhibitor cocktail (1 tablet per 50 ml polysome buffer; Roche) and 0.1mM DTT. Gradients were poured using a Hoefer SG 50 Gradient Maker (Amersham Biosciences), magnetic stirrer and MasterFlex C/L peristaltic pump (Cole-Parmer) into Polyallomer centrifuge tubes (14 x 95 mm; Beckman) and stored at 4°C (not longer than a couple of hours) until ultra-centrifugation and fractionation.

2.9.3 Ultra-centrifugation and fractionation

Cell lysate was layered on to the top of a 15-50% sucrose gradient. Samples were centrifuged in a pre-cooled Beckman L-60 Ultracentrifuge at 40 000 rpm for 2 hours at 4°C, using a SW40 rotor. The gradient centrifuge tube was pierced with a Brandel Tube Piercer and Fluorinert FC-77 (Sigma-Aldrich) pumped through as the high-density chase solution with a Brandel Syringe Pump. The gradient was then separated into fractions using a Foxy R1 machine (Teledyne Isco) and the optical density read at 254 nm throughout fractionation on a UA-6 Absorbance Detector (Teledyne Isco). One fraction was taken for every 13 drops and the sensitivity typically set to 1 or 2. PeakTrak software (Teledyne Isco, Inc.) was used to record the absorbance output.

2.9.4 RNA and protein extraction from gradient fractions

RNA extraction was made using the Qiagen RNeasy kit. Following fractionation, 3.5 ml of Buffer RLT with β-mercaptoethanol (10 μl β-mercaptoethanol per 1 ml RLT buffer) was added to each polysome fraction and transferred to -80°C overnight. After thawing, 2.5 ml of 100% ethanol was added and mixed. RNA was extracted from each fraction sample using a separate RNeasy column and according to the manufacturer's instructions for the 'RNA cleanup' protocol, including the on-column DNase treatment (RNase-Free DNase Set, Qiagen). RNA was eluted from each column with 30 or 40 μl nuclease-free water and stored at -80°C. RNA concentration was determined using a NanoDrop Spectrophotometer ND-1000 and for the Illumina digital tag sequencing experiment, 10% of each sample was examined on an RNA gel (see Section 2.4.2).

2.10 [^{35}S]-methionine labelling

The radiolabelled methionine experiments were performed by Dr Terry Smith (University of St. Andrews) similarly to previously reported (Smith *et al.*, 2009). Cells were pelleted, washed with methionine-free minimal essential media and resuspended with 2 ml of the same media to a concentration of 1 x 10^7 cells/ml. Cells

were labelled for one hour at 37°C with 10 mCi [^{35}S]-methionine in an orbital shaker water bath.

For use with SDS-PAGE, a volume of 1 ml was pelleted, washed with trypanosome dilution buffer (TDB; 25 mM KCl, 400 mM NaCl, 5 mM $MgSO_4$, 100 mM Na_2HPO_4, NaH_2PO_4, 100 mM glucose), resuspended with pre-heated SDS-PAGE sample buffer and boiled for 5 minutes. A 10% SDS-PAGE gel and Coomassie blue staining was used to visualise the proteins. The gel was then destained, soaked in En^{3}hance (NEN) for 30 minutes, washed twice with water, soaked in 10% glycerol, dried and exposed to XAR-5 film at -70°C.

For radiolabel incorporation quantification, the remaining 1 ml was split into triplicate samples, these being centrifuged, washed with TDB, then resuspended with TCA. The protein pellet was washed and reconstituted with 1% SDS, the incorporation quantified scintillation counting.

2.11 Illumina digital tag RNA sequencing

To determine the quantitative expression levels of mRNA, Illumina digital tag RNA sequencing was used. For this, RNA samples were sequenced by Eurofins MWG Operon from short-insert (17+4 bp) DNA libraries using Illumina HiSeq 2000 sequencing. Eurofins MWG Operon also prepared the library prior to sequencing using the workflow shown in the schematic diagram in Figure 2.1. Following binding of the short-insert cDNA library to the flow cell, the Cluster Station isothermally amplifies the cDNA constructs by bridge PCR to create clonal clusters of ~1000 copies each. These template clusters are then directly sequenced using sequencing by synthesis technology and four fluorescently labelled, reversibly terminated nucleotides. This sequencing used the single read module with 1x 50 bp read length.

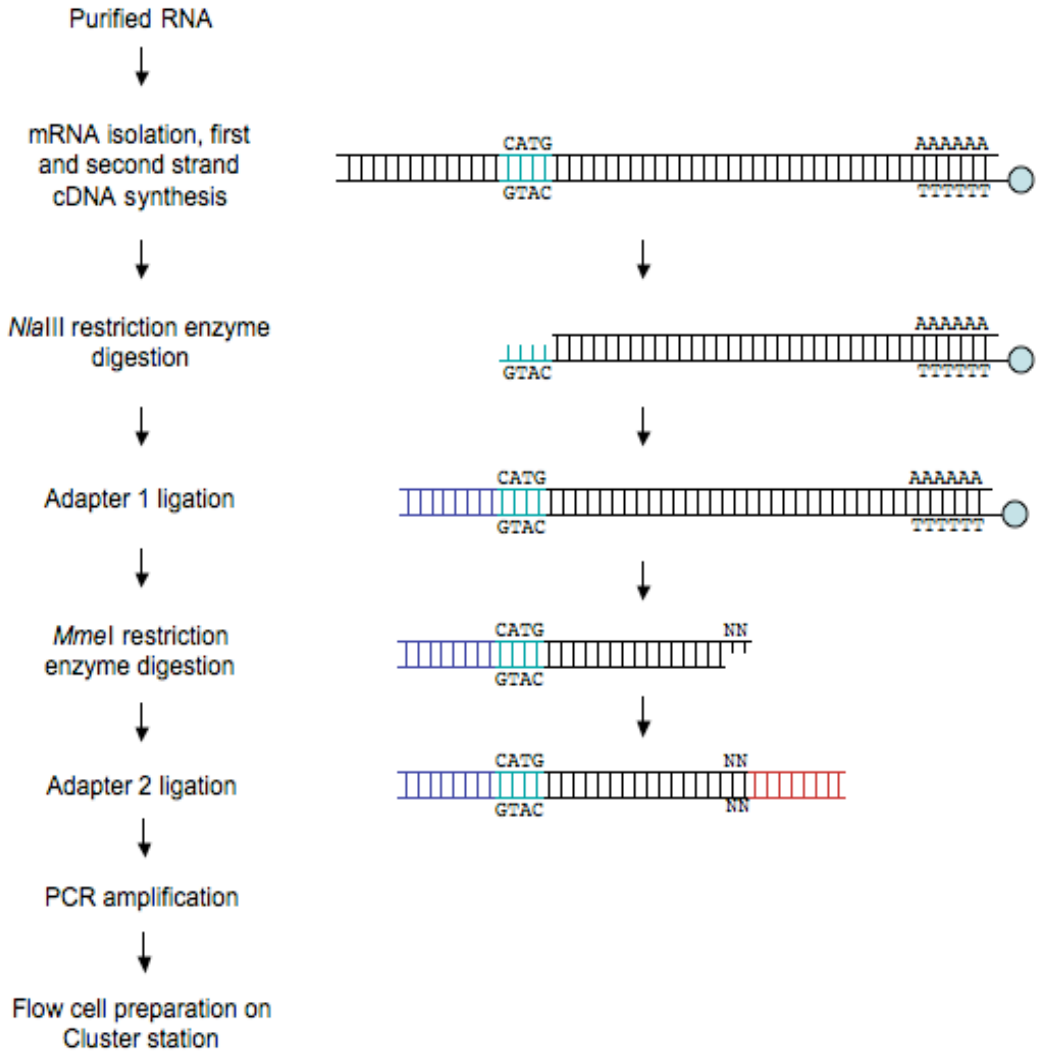


Figure 2.1 Schematic diagram outlining the workflow of library preparation for Illumina digital tag sequencing by Eurofins MWG Operon. The isolation of mRNAs was achieved with oligo (dT) magnetic beads (blue circle), the first strand of oligo (dT) bound cDNA was synthesised from the mRNA by reverse transcription, the mRNA degraded by RNase, then the second strand cDNA was synthesised. The cDNA was cleaved at *Nla*III restriction enzyme site sequences (CATG; shown in turquoise) through digestion with *Nla*III. Adapter 1 (shown in dark blue) was ligated to the digested end. The Adapter 1 contains sequence that together with the start of the *Nla*III sequence forms an *Mme*I restriction enzyme site sequence. *Mme*I cuts 20/18nt downstream from the *Mme*I site, so digestion with *Mme*I, therefore, creates a 17nt sequence flanked by Adapter 1 and the *Nla*III restriction site sequence. 3' fragments were removed by precipitation of the magnetic beads. An Adapter 2 sequence (shown in red) was then ligated to the site of *Mme*I cleavage, resulting in a 17 bp tag sequence of the original mRNA molecule. The Adapters 1 and 2 contain sequence used for linear PCR amplification (15 cycles) of each 17 bp tag sequence. Each of the amplified cDNA constructs were gel purified and then bound to the flow cell on the Ilumina Cluster Station via complementary sequences present in the Adapter 1 and in the oligos attached to the flow cell surface.

Chapter 3

Results

Functional analysis of UTRs from stumpy form enriched genes

To date, nothing is known about the mechanisms of control of gene expression in stumpy form *T. brucei*. However, the control of gene expression is important in this life-cycle stage since there is a repression of gene expression for most transcripts, except a small subset of genes that escape this repression. Regulation of gene expression in trypanosomatids is believed to operate at the post-transcriptional level and several examples where 3'UTRs are important to life-cycle stage-dependent control of expression have been identified (reviewed in Clayton, 2002 and see Section 1.6). To investigate if signals present within the UTRs are important for control of the expression of stumpy-enriched genes, two genes that exhibit increased expression at this lifecycle stage where chosen for UTR analysis. These two genes were ESAG9-EQ and aconitase. Analysis was made of the expression level of a reporter gene under the control of the different UTR sequences to assess any gene regulation conferred by those sequences. The aim of these experiments was to ask:

1. Is each UTR responsible for repression of gene expression in slender forms, and if so, are there specific sequences within the UTR that contribute to this?

2. Is each UTR responsible for alleviation of gene expression repression in stumpy forms, and again, if so, are there specific sequences within the UTR that contribute to this?

3.1 CAT reporter gene assay

A CAT reporter assay approach was used to assess the contributions of a given 5' or 3'UTR to the control of gene expression. This approach has previously been used successfully to map 3'UTR control elements of COX genes (Mayho *et al*., 2006). The approach uses the expression vector, CAT449, which is illustrated in Figure 3.1. The CAT449 plasmid is a modified pHD449 plasmid (Biebinger *et al*., 1997), such that the Tetracycline repressor sequence present in pHD449 is replaced with the chloramphenicol acetyltransferase (CAT) coding region (Mayho *et al*., 2006), which functions as the reporter gene in this assay.

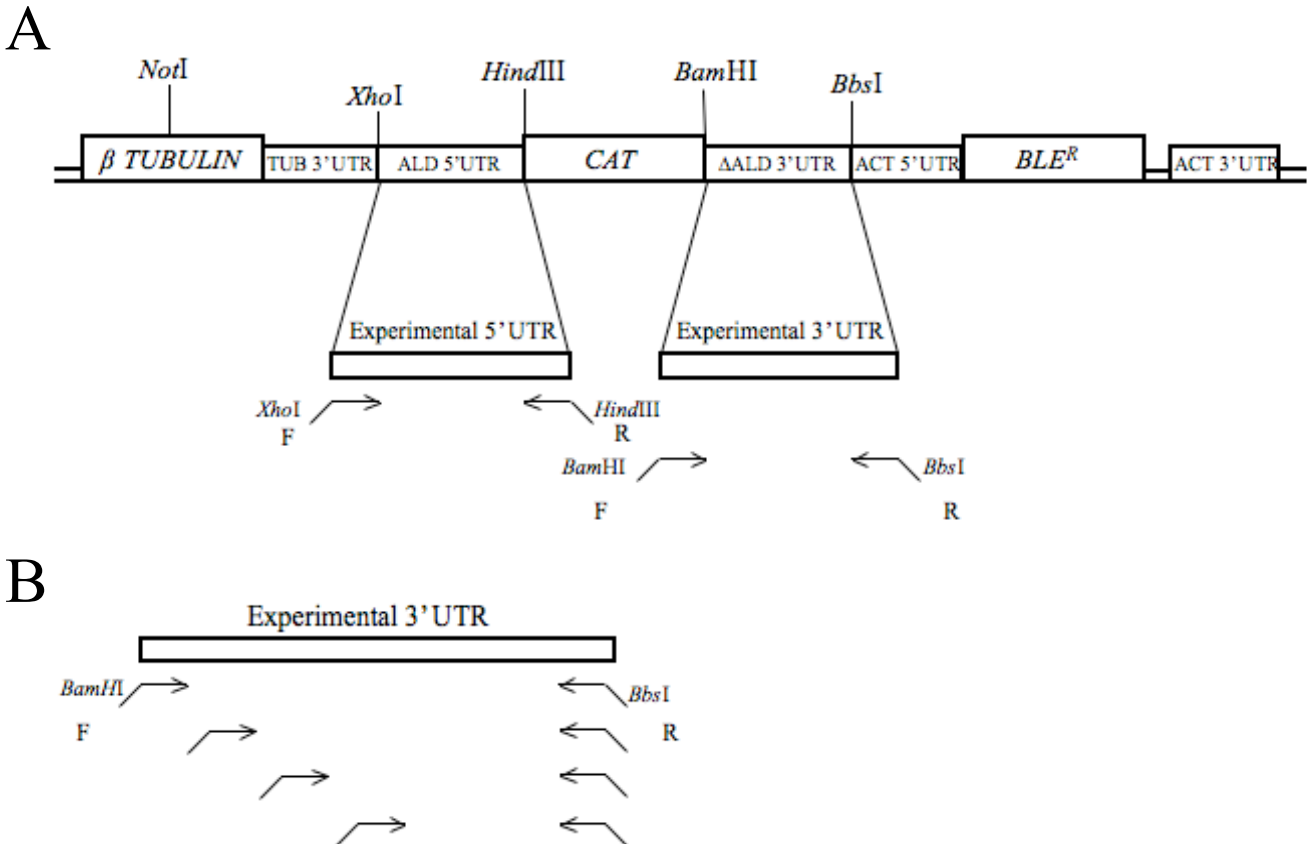


Figure 3.1 Schematic diagram of the sub-cloning strategy for 5'/3'UTR analysis with CAT reporter gene assay. A. Diagram of the region of CAT449 vector with the chloramphenicol acetyl transferase (CAT), phleomycin resistance (BLE^R), and β-tubulin genes and flanking 5' and 3'UTRs (TUB: tubulin, ACT: actin). Aldolase (ALD) 5'UTR and truncated 3'UTR (ΔALD) flank the CAT gene, the ΔALD 3'UTR conferring constitutive expression; replacement of these with an experimental UTR used the *Xho*I and *Hind*III (5'UTR) or *Bam*HI and *Bbs*I (3'UTR) restriction enzyme sites. B. Representation of a deletion series for an experimental 3'UTR. Forward primers were designed to create ~50nt sequential deletions along the UTR from the 5' end. Bent arrows indicate where PCR primers were designed to amplify from, containing the restriction enzyme site sequence indicated next to the arrow. F = forward, R = reverse. Drawing not to scale. Adapted from Mayho *et al*. (2006).

Hence, as with pHD449, digestion with *Not*I results in linearisation of the CAT449 vector in the tubulin targeting sequence and, consequently, integration into the *T. brucei* genome is through the β-tubulin locus. The β-tubulin locus consists of a repetitive array of β-tubulin and α-tubulin genes on chromosome 1. After plasmid integration, the CAT reporter gene is consequently transcribed by read-through Pol II transcription of this locus. This is potentially important, because the same machinery is being used to transcribe the CAT reporter gene as the endogenous gene whose regulation is under investigation. As a result, any regulatory elements present in the sequence under examination should be able to function effectively in the reporter gene assay. Although one gene investigated, ESAG9-EQ, may also be transcribed by Pol I as it is present in expression sites (Barnwell *et al.*, 2010).

3.1.1 Generation of CAT449-based constructs

Flanking the CAT gene in the CAT449 vector are the aldolase 5' UTR and truncated aldolase 3'UTRs. The truncated aldolase 3'UTR provides constitutive CAT gene expression (Mayho *et al*., 2006), since it is not developmentally regulated (Biebinger *et al*., 1997). Investigation of experimental 5'UTRs was made by replacement of the aldolase 5'UTR through *Xho*I and *Hind*III restriction enzyme sites, and of experimental 3'UTRs by replacement of the truncated aldolase 3'UTR through *Bam*HI and *Bbs*I restriction sites (see Figure 3.1 A). In each case, UTR replacement was achieved by amplification of the relevant DNA sequence by PCR using primers containing the appropriate restriction enzyme site sequences for insertion. Each amplicon was initially sub-cloned into the pGEM T-easy vector (see Sections 2.3.9 and 2.3.10). After sequence verification, this was followed by digestion from pGEM T-easy with those restriction enzymes and ligation into similarly digested, purified CAT449 plasmid (see Sections 2.3.7 and 2.3.9). Each plasmid construct was amplified by large-scale plasmid preparation and, following linearisation with *Not*I, transfected into the *T. b. brucei* monomorphic bloodstream form strain Lister 427 (see Sections 2.3.11 and 2.2.3). Like pHD449, CAT449 contains the phleomycin resistance gene (BLE^R), so phleomycin (0.5 – 2.5 μg /ml) was used to select for successful transfectants. Selection with phleomycin was made on serial dilutions of

transfected cells, to allow clonal selection, with a control of drug-treated parental (non-transfected) cells being included to confirm drug sensitivity (see Section 2.2.3). Only after all control cells were dead from the drug action were transfectant clones isolated.

3.1.2 Determining CAT mRNA and protein levels

Assessment of CAT protein levels in successful transfectant cell lines was made through CAT ELISA assays using a CAT ELISA kit (Roche) according to the manufacturer's instructions (see Section 2.6). Each sample analysed in the assay was present in duplicate on the CAT ELISA plate from which the mean was calculated. Experimental replicates refer to the repetition of the whole CAT ELISA assay from a repeated generation of samples. In each assay, a sample from the cell line obtained from transfection with the unmodified CAT449 plasmid (referred to as the 'unmodified vector' construct) was included to allow comparison of results from assays performed at different times. As such, in each assay the value obtained for the unmodified vector cell line was arbitrarily adjusted to 100%, and the other values adjusted to the percentage of this value.

Assessment of *CAT* mRNA levels in successful transfectant cell lines was made through northern blot analysis (see Section 2.4). To detect *CAT* mRNA, a digoxygenin-labelled probe designed against the CAT coding region was used from laboratory stocks. Again, an RNA sample from the cell line obtained from transfection with unmodified CAT449 plasmid was included in each northern blot to allow comparison of results between various blots. As previously seen (Mayho *et al*., 2006), the use of the *CAT* probe with RNA samples obtained from cells transfected with the CAT449 vector results in the detection of three bands. The smallest of these is the *CAT* mRNA; the larger two are presumed to be processing intermediates of the *CAT* mRNA (see Figure 3.3 for an example).

As mentioned above, the truncated aldolase 3'UTR in CAT449 gives constitutive CAT reporter gene expression. Thus any experimental 5' or 3'UTR which causes repression of gene expression in slender forms would cause a decrease in CAT

expression relative to the unmodified CAT449 vector. Therefore, if the 5' and/or 3'UTRs of ESAG9-EQ or aconitase contribute to gene regulation, it would be expected that insertion of that sequence would result in a decrease in CAT expression, as expression of both of these genes is repressed in slender forms *in vivo*.

For any UTR that conferred repression of gene expression in slender forms, subsequent analysis involved creation of a deletion series to delineate any regions of that UTR that were responsible for the repression. To produce a deletion series, PCR reactions were carried out with primers to create ~100nt or ~50nt sequential deletions. Therefore, for a 5'UTR deletion series, the same forward primer would be used as for the 'full-length' construct but with reverse primers complementary to sequence ~100/50nt sequentially along the length of the 5'UTR. Similarly, for a 3'UTR deletion series, the same reverse primer would be used as for the 'full-length' construct but with forward primers complementary to sequence ~100/50nt sequentially along the length of the 3'UTR. Thus, deletions were made from the 3' end of a 5'UTR and from the 5' end of a 3'UTR, to maintain the regulatory SAS/PAS signals present respectively. A schematic of a 3'UTR deletion series is shown in Figure 3.1 (B).

3.2 Is the aconitase 5'UTR or 3'UTR responsible for control of gene expression?

It has been shown previously that the *T. brucei* aconitase (TbACO) protein is enriched in stumpy forms compared to slender forms (Saas *et al*., 2000). However, only minor changes in the transcript abundance between the lifecycle stages were observed, which suggests that expression of the TbACO protein is regulated at the translational or post-translation level (Saas *et al*., 2000). To investigate if either the 5' or 3'UTR of the transcript were responsible for this regulation, the 5' or 3'UTR present in the CAT449 vector were replaced, as described above, with 500nt of sequence either upstream or downstream of the TbACO gene (Tb927.10.14000), respectively.

3.2.1 Analysis of the 5' & 3' UTR

For TbACO the nearest upstream gene annotated in the *T. brucei* TREU 927 genome sequence is Tb927.10.14010 and the nearest downstream gene annotated is Tb927.10.13990. The intergenic region between the stop codon of Tb927.10.14010 and the start codon of TbACO is 1982 bp. The intergenic region between the stop codon of TbACO and the start codon of Tb927.10.13990 is 2649 bp. These intergenic regions are shown in Figure 3.2.

A

TAGgtaacatggcggtatcggcgacaaggtcgttttctacaatgcattactacctccacactgtaaac
cacactttacgcggtgtccatttttctattcaaactctgtgcttcctttggtatgtagtcactggccc
cacgaaattccttttctgcgcttcttacactgatcagtagagggataagtgggatttgcagaaagaat
ttggtgacgcaagtgctccaccaggagcgcggggagcgggaaggaggcacagggaggggggaatatgga
tgttttgaagtcgtcatgtacagtatgctttttcgtgatgaaaaaatttacttgcaccacaagacat
tgcagggatggtacacggcgattgcattacagtgcacatctctttttttctttttttgaactgctgttc
ccccttttacatcattttcggacttttttattttgatgataaatgaaggatctgcctgcacgtgtgcga
tggggatatcaatatagaggaaacaattgactttgttccatttttcacgtgacccagcgacgtgcagc
tggggaagtcgtctcagactcgcgggctgtgctgtttcccgtttgaatgtgatgtcgcagtcgtcata
tatgtgagtatgatgctgtttttattgacgtgcacctggtttccatagtttccgctttcagcttcaca
gggaccaaagaataatgttccccacaagcacaaataaagatagtgtacattacatacggtcacgatgt
atatcatccgtgatgtcatggtttccgtagttttccgctgggattaaacctgtttgatgtttcctcgg
ctggttcaccatatgtgtggatgcgcttgcccgtctccgcggtggtaaacgcattgtggtggaagtct
cggcactatatggcaagggacacaagccagaagtgtgaaccgcattcgggagcctctgccgctggaca
gcagcagtggaagcacaaaatattaaagttgcggtgagtcagcaagcacaactcctggcggggtgagg
ctgctgttggattagatttgttcttttagccctcagtgcggagaccccaccgcgttgcgtcacgtgtg
tccctccgatcccccgtttgcgttggcatccagcttgtgcagagcagctgtagttactcctgctattt
ttttttttgtggtgtgactgctctaagtttttcacccttggagtggacggttacgtttatgggatg
ttaactttgcgtccgcaggccacgaagatgccaacggggatcgtatgttaggaaggggagtgaggact
ggcagccgctgttaagaggtgtggtgtggtggcgggtccctctttcttttgttattggaggtcagtct
ctcttactctatcgatagtagcatagcctgatgtgcgtagaacaggggatcgtgacgactgtcaatcg
5'UTR(s)
atgactcgctcagttgacggtgatccgctgttcttttggtggaaagtgcgtgcgtggggtcttggtcc
agccttcaattttcgccctgtcgcaccgttccgtgtacctctcttaggtttgttatttgtggatggct
atcgaggatcatatacccacacacacacacacaatatgtgcaatatggttgttgccaggaaagcgcgt
ctatgtgaaatgaaacctacagaagtggctgaacgtttgtgtggacaacgaacggtttcttgctttct
actcttctttcgcagttgacggagcatccacgggccacccgaaaggtgggcacacagtcagtatgagg
tgcctaccaaacattttgcaggtcaaacaaagtttcttacttgtagccgcattgggcgggggttttct
gtggcattctattttacatgcctttgtgtttatacttttgcattagtaaggttataaaaataaggtt
tacatcgactacactgccaaaataataggaaaggtatatactgcacaagcgaaagggtaagaagaggg
5'UTR(e)
agtcttgtgtATG

Figure 3.2 continues on the following page

B

3'UTR(s)

TAGcggaacaccgaaacctgttcctcagagaataagcctgtgaattagtgaggggcgtagcaagggat

Possible IRE

400nt(s)

atctttgtttaccgcttccgtattcatcgttattgctaactacatatttggtacatggttgggtgtgt

300nt(s)

ggggcggctactaccgccacatcgtgagacgaggtagaagcgccgtttaactgtctccccctttggcca

gctgacttggtgccatgccgcacatttgttccaacgagggttgaagaggatgcaaaaaaaaaaaaaga

200nt(s)

ggtaaaaacgaggttcgacgacgctattatggcgacggcttaacgagatcagaacgttattgttattg

100nt(s)

ttatttatttatttatttttgaggaagcaaaacgggaaaaggggatctatggatattaatggtgaca

aaaacctttgtcggtgaacgtggacgatgcgtaaaacaatggaaggctctggatcggtcgcaacaccc

3'UTR(e)

ctcgccccctcccccgaaaaagcaaactttagcgggcaacgaacaaaaaaaaaaaagagccaggagag

gggtagagagggagtggttggatgacttttagcttccgctaaacatgtgaaggtcgggtattttgtcc

aatttgaagcgcttttgcattcatttacgccgcataggtgctaattcgcttatttacttatttattaa

cttcactcatacacgcctgagagcagggcaaaaggaagggaggtgtgaaaagacgtaatgcaaggtct

taaaattatgatattgggaaagatgcatattagtgcatcctatttttattttagtatttgtttgatttc

850nt(e)

catgtgaggaaggcaaaagaggggagtgcaaaaaaaaaaaacggcataaggtagagggaactatgtgt

gatgaggagtgtggagtac…(2149bp)…ATG

Figure 3.2 Intergenic regions for Tb927.10.14000 (TbACO). A. The genomic sequence from the stop codon of the nearest upstream annotated gene, Tb927.10.14010, to the start codon of TbACO. B. The genomic sequence from the stop codon of TbACO to the start codon of the nearest downstream annotated gene, Tb927.10.13990. Because this intergenic region is 2649bp in length, only the first 900nt of the sequence is shown. The possible iron-response element (IRE) sequence is indicated by the line. Sequences are in the 5' to 3' direction. Red font: stop codon, red highlight: start codon, grey highlight: start (s) or end (e) of each sequence used to generate the construct named above that residue, ag: SAS (Saas *et al*., 2000), a: PAS (Siegel *et al*., 2010).

Ideally, the whole of each intergenic region would be used for analysis in the CAT449 vector to ensure that all endogenous regulatory signals were included. However, due to the extended length of each of the upstream and downstream intergenic regions a shorter region was chosen for analysis. Primers 3 and 4 (see Table 2.1) were used to amplify the 500nt region upstream of the TbACO start codon and primers 5 and 6 were used to amplify the 500nt region downstream of the TbACO stop codon (Figure 3.2) using *T. b. brucei* Lister 427-449 genomic DNA. These were used to replace the 5' or 3'UTR of the CAT gene in CAT449 to generate the '5'UTR' and '3'UTR' constructs respectively.

The 5' splice acceptor site (SAS) for TbACO has been previously determined to be at position -100 relative to the first ATG by 5'RACE analysis (Saas *et al*., 2000; marked on sequence in Figure 3.2 A). 3'RACE analysis in the same study revealed two polyadenylation sites (PAS) for TbACO, as PCR products obtained from this were 500 and 810 bp. Therefore, 500nt was chosen as the length for analysis of the 5' and 3'UTRs as this length of sequence should contain the signals for splice-leader addition and polyadenylation respectively.

As described above, stable transfectant cell lines generated with the '5'UTR' and '3'UTR' constructs (Figure 3.3 A) in *T. b. brucei* monomorphic bloodstream form strain Lister 427 were selected with 0.5 μg/ml phleomycin. These were then used to assess *CAT* mRNA abundance by northern blot analysis and CAT protein abundance through CAT ELISA assays. For each construct, two biological replicates (independently derived clones) were examined in CAT ELISA assays. The mean of the biological replicates for each construct is shown in Figure 3.3 (B). Three biological replicates for each construct were examined by northern blot analysis and this result is shown in Figure 3.3 (C).

As shown in Figure 3.3 (B) the 5'UTR and 3'UTR constructs resulted in vastly different levels of CAT protein expression. For the 5'UTR construct, the CAT protein level was only reduced to 80% (±8.8%) of CAT expression seen for the unmodified vector. However, replacement of the truncated aldolase 3'UTR in the 3'UTR construct resulted in a dramatic reduction in CAT protein expression to 4.7% (±0.75%) of CAT expression with the unmodified vector.

In contrast, the *CAT* mRNA expression pattern (Figure 3.3 C) appeared to be different to the CAT protein expression pattern for the 5'UTR and 3'UTR construct clones. The 5'UTR clones show a high *CAT* mRNA level with respect to the unmodified vector clone, similar to the CAT protein level seen for this construct. Similarly, the 3'UTR clones showed abundant *CAT* mRNA with respect to the unmodified vector clone, contrasting with the CAT protein level for the 3'UTR construct.

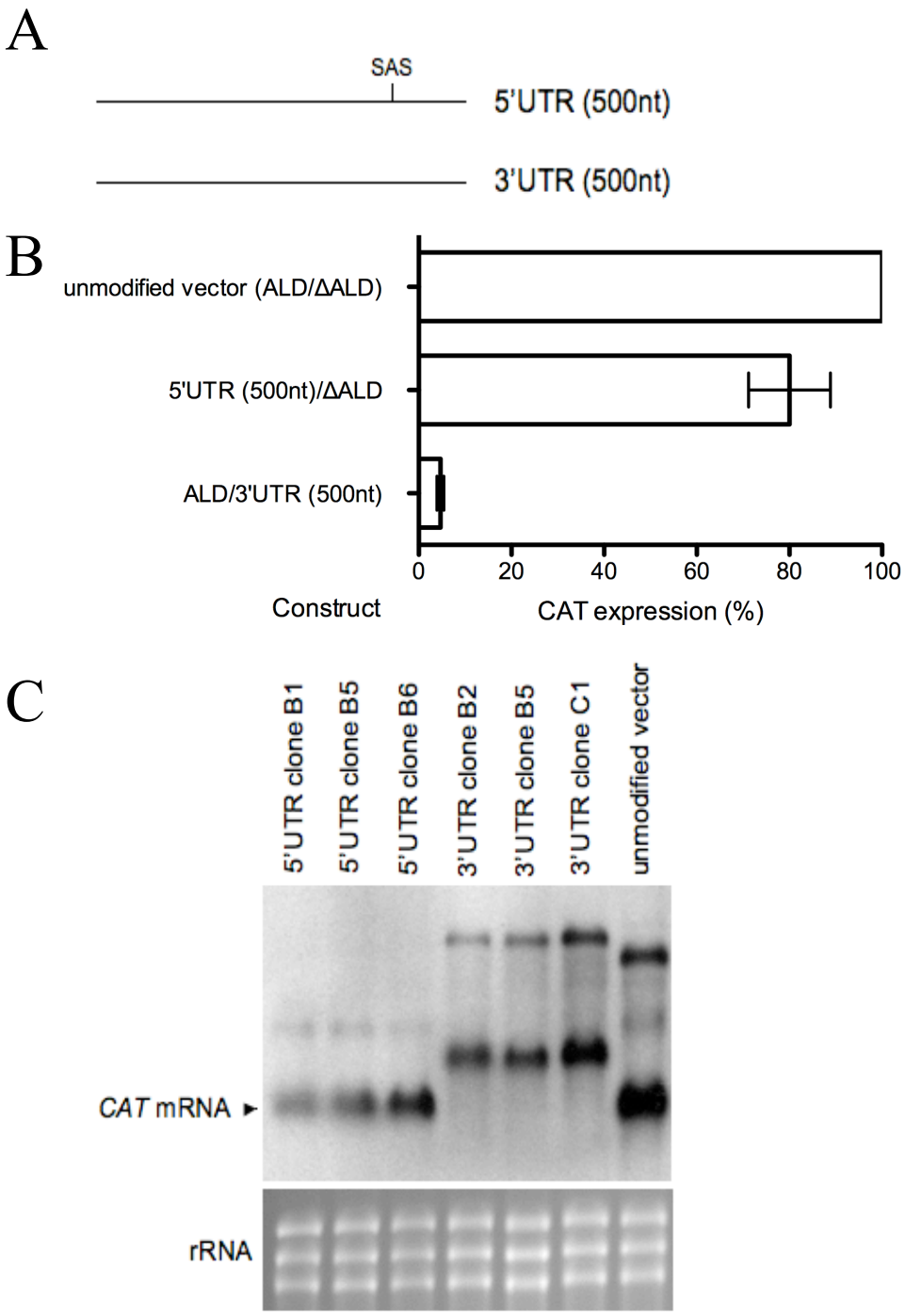


Figure 3.3 Analysis of TbACO 5' and 3'UTR using a CAT reporter gene assay. A. Schematic of sequences analysed with the '5'UTR' and '3'UTR' constructs, with the locations of known *trans*-splicing processing. SAS: 5' splice-acceptor site. B. CAT ELISA assay results from two independently derived clones for each of the 5'UTR and 3'UTR constructs, and one clone of the unmodified vector. Two experimental replicates were performed for each clone. The mean ± standard error for biological replicates is shown. C. Northern blot analysis using a digoxygenin-labelled RNA probe against the CAT coding region with RNA samples from three independently derived clones for each of the 5'UTR or 3'UTR constructs and one clone of unmodified vector as a control. Ethidium bromide staining of rRNA is shown in the lower panel as the loading control.

Also apparent in Figure 3.3 (C) is a size difference between the *CAT* mRNA derived from the 5'UTR clones and in the 3'UTR clones. This is likely due to the locations of mRNA processing sites for the *CAT* transcript in each case. The predicted SAS location would generate a transcript with a short 5'UTR for the 5'UTR construct, of a similar length to the aldolase 5'UTR from the unmodified vector. In contrast, the PAS used in the 3'UTR construct was expected to produce a longer 3'UTR than that produced with the truncated aldolase 3'UTR present in the 5'UTR and unmodified vector constructs. Therefore, the *CAT* mRNA would run at a higher molecular weight in the 3'UTR construct RNA samples, as was observed in Figure 3.3.

The relatively abundant *CAT* mRNA derived from both the 5'UTR and 3'UTR constructs was expected because, as mentioned above, endogenous TbACO shows little developmental regulation of RNA levels between life-cycle stages, with instead regulation being made at the protein levels (Saas *et al*., 2000).

This supports what was observed for the 3'UTR construct clones: low CAT protein expression but high CAT mRNA expression. Therefore, it would appear that there are regulatory elements present within the 500nt downstream sequence for TbACO that cause repression of gene expression in slender forms. Thus, the 3'UTR was selected for further analysis.

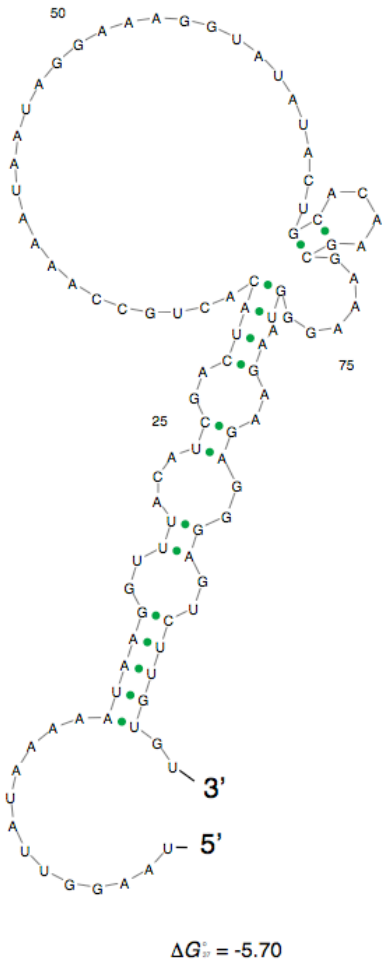
3.2.2 3'UTR deletion series

From the data presented in Figure 3.3 it would appear that the majority of repression of TbACO expression in bloodstream forms is mediated through the 3'UTR, although some minor contribution may be made by the 5'UTR. Therefore, the next step was to generate a deletion series of the 3'UTR sequence to further define the regulatory regions responsible for this repression. Initially, constructs of sequential 100nt deletions were designed to generate the constructs '400nt', '300nt', '200nt' and '100nt' (see Figure 3.2). Primers 7, 8, 9 and 10 were used in PCR reactions each with the primer 6 (see Table 2.1) for the amplification and cloning of these constructs.

Additionally, to further define any regulatory regions, the predicted RNA secondary structures of the sequences used for analysis of the 5' and 3'UTR were analysed. For this the freely available web-based software package *S*fold (Ding *et al.*, 2004) was used (see Section 2.1.1). *S*fold is a programme that uses an algorithm that produces a statistical sample of RNA secondary structures according to the Boltzmann distribution for a given sequence. The output predictions from the *S*fold '*S*rna' programme gives centroid structures that are each produced from a cluster of similar predicted structures. The ensemble centroid structure is considered as the single structure that best represents the central tendency of the set of the entire collection of structures sampled (Ding *et al*., 2005). The ensemble centroid structures for the sequences used in the 5'UTR (from the mapped SAS, Saas *et al*., 2000) and 3'UTR constructs are given in Figure 3.4 A and B respectively.

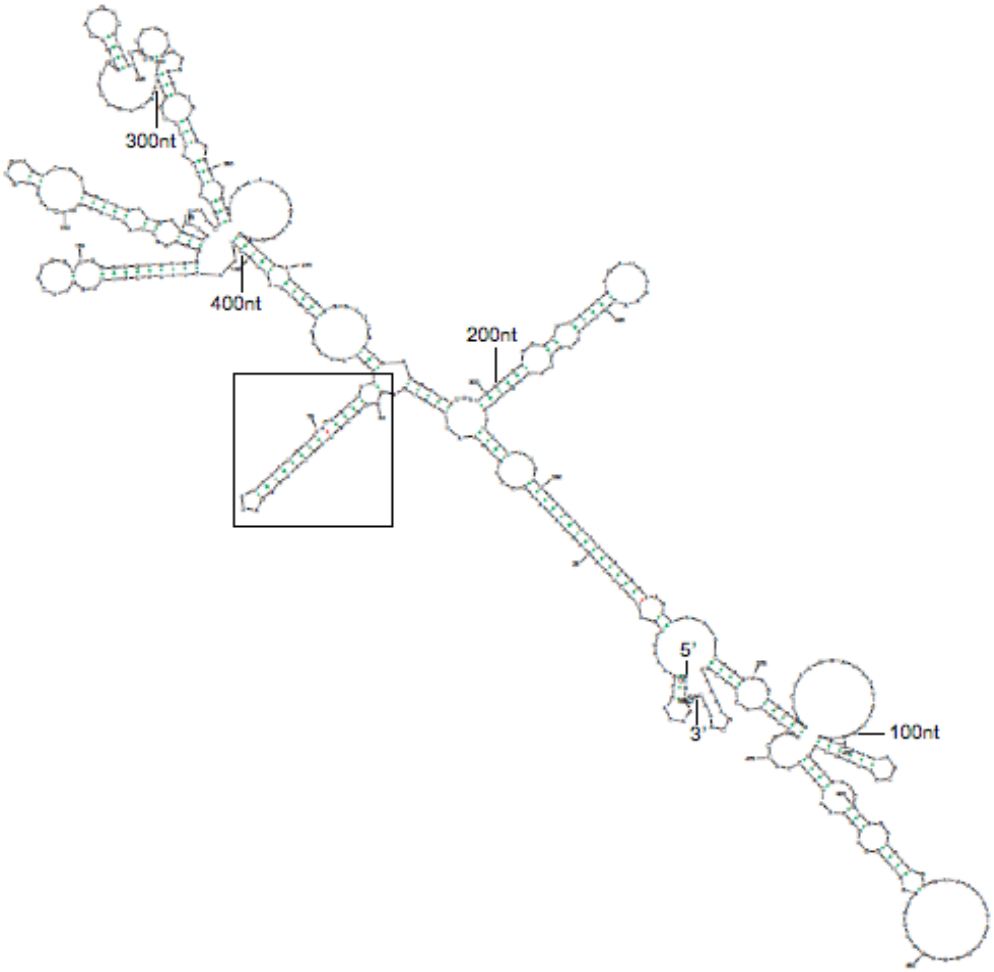
As the 3'UTR sequence used for the RNA secondary structure prediction with *S*fold was 500nt and so relatively long, the predicted folding consisted of a number of smaller structures. One potentially interesting predicted structure was observed in the *S*fold for the 3'UTR sequence, this being the structure outlined in a box in Figure 3.4 B (an enlargement of the structure is shown in Figure 3.4 C). This is because it appears to bear some similarity to an iron response element (IRE), which is a stem-loop structure. In metazoans, iron regulatory proteins, IRP1 and IRP2, regulate iron homeostasis post-transcriptionally by binding to IREs present in the mRNAs of genes involved in iron metabolism. Binding of an IRP to an IRE present in a 5'UTR represses translation of that transcript and binding to an IRE present in a 3'UTR can suppress mRNA degradation (reviewed in Volz, 2008). A schematic of an IRE secondary structure and the three-dimensional structure of an IRE bound by an IRP is shown in Figure 3.4 (D; taken from Volz, 2008). By comparison of Figure 3.4 C and D, it can be observed that the possible IRE in the TbACO 3'UTR shares similarity to the conserved canonical IRE stem-loop structure, in that it possesses a

A



ΔG°_{37} = -5.70

B



ΔG°_{37} = -126.80

Figure 3.4 continues on the following page

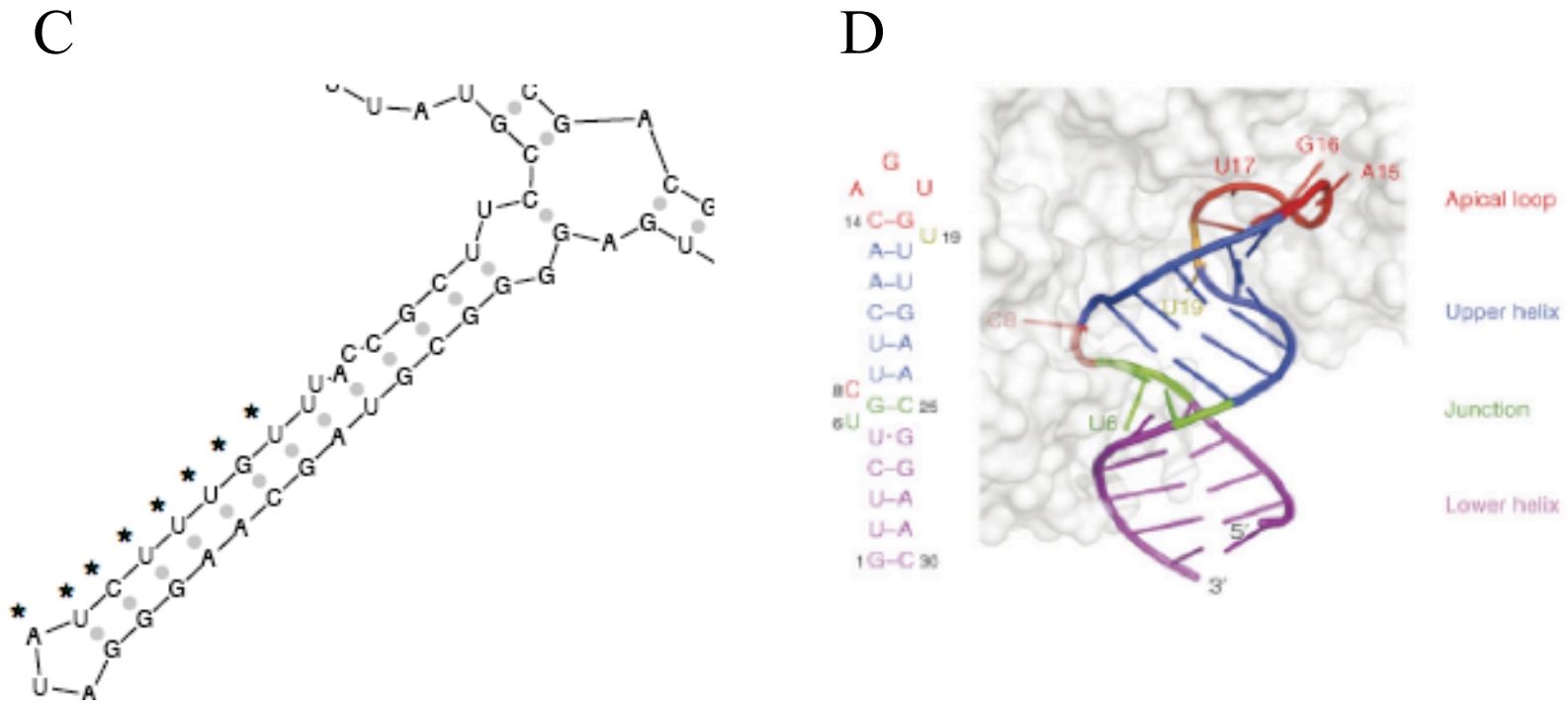


Figure 3.4 Predicted RNA secondary structures for TbACO 5' and 3'UTRs. A. *S*fold predicted secondary RNA structure for TbACO upstream sequence used in 5'UTR construct from the mapped splice addition site (Saas *et al*., 2000). B. *S*fold predicted secondary RNA structure for TbACO downstream sequence used in 3'UTR construct. The 5' and 3' ends are indicated as well as the start of each deletion series sequence (see Figure 3.2). C. Enlargement of boxed region from B. showing the possible iron response element (IRE) structure. Asterisks indicate sequence with similarity to human mAco IRE sequence given in Volz (2008). D. Figure from Volz (2008) showing schematic of secondary structure of an IRE and three-dimensional structure of the IRE bound by an iron regulatory protein (IRP).

triplet of unpaired nucleotides (AUA) at the apical loop and two un-paired nucleotides (C and U) along the length of the stem. However, the structure is lacking the un-paired nucleotide below the apical loop, the two un-paired nucleotides are located further from the apical loop, and the highly conserved signature sequence of the apical loop (CAGUGX, where X = U, C, or A; Volz, 2008) is different (GAUAU-). Yet it was also noticed that part of the sequence of the possible IRE in the TbACO 3'UTR is similar to part of the primary sequence of a human mitochondrial aconitase IRE (A-UcUUUGU; Volz, 2008); this is indicated in Figure 3.4 (C) with asterisks.

Although the similarity between this sequence and the canonical IRE sequence is low, the possibility that this predicted structure functions as an IRE provided a lead for further investigation. As such, the 400nt deletion construct discussed above would be particularly interesting because this involves deletion of this potential IRE sequence, along with other sequence (see Figure 3.4 B).

3.2.3 TbACO PAS identification

During the course of this work a study was published that, through high-throughput RNA sequencing of *in vitro* cultured bloodstream form and procyclic form *T. brucei*, identified the SAS and PAS for ~7000 and ~6000 genes respectively (Siegel *et al*. 2010). From the results of that study, two PASs were identified for TbACO, these being at positions +516nt and +843nt relative to the TbACO stop codon (see Figure 3.2 B), this being approximately consistent with the earlier study by Saas *et al*. (2000). As noted earlier, neither of these PAS locations are within the sequence used for the 3'UTR construct, since that sequence stops at +500nt, just short of the first PAS identified. Thus there could be another explanation for why low CAT protein expression was seen using the 3'UTR construct; the use of an inappropriate PAS would cause the *CAT* RNA to not be expressed highly as the signals needed for appropriate polyadenylation of *CAT* RNA would be absent. Therefore, it could be that another site is chosen for polyadenylation that is not optimal and affects the translation efficiency of the transcript, resulting in lower CAT protein expression. Consequently, it was necessary to repeat the analysis of the 3'UTR by including

850nt of downstream sequence instead of 500nt. This length would ensure correct processing of the *CAT* transcript by including both endogenous PAS sites at +516nt and +843nt and thus any repression of gene expression would be due to regulatory regions within the sequence investigated. To make this construct, named '850nt', primer 11 was used in a PCR reaction with primer 5.

Unfortunately, at this point we became aware that another laboratory was also investigating the regulation of TbACO gene expression through analysis of the 5' and 3'UTR in a similar manner. Therefore, this work on the TbACO 5' and 3'UTR was ceased. Consequently, the creation of the 3'UTR deletion series constructs and 850nt construct were not completed and the effects of these sequences on gene expression were not analysed. Instead, the 5' and 3'UTRs of another gene upregulated in stumpy forms, ESAG9-EQ, was analysed for control of gene expression in this study. The results for this will be presented in the remainder of this chapter.

3.3 Is the ESAG9-EQ 3'UTR responsible for the control of gene expression?

ESAG9-EQ (Tb927.5.4620) is so-called because it was the *T. brucei* ESAG9 sequence that was found to be most similar to the *T. equiperdum* ESAG9 sequences (Barnwell *et al*. 2010; Florent *et al*. 1991). ESAG9 proteins are a family whose mRNAs are strongly developmentally regulated, being highly enriched in stumpy forms compared to slender forms (Barnwell *et al*. 2010). To investigate if either the 5' or 3'UTR of the transcript were responsible for this developmental regulation the 5' or 3'UTR sequences present in the CAT449 vector were replaced, as described earlier (Section 3.1.1), with sequence either upstream or downstream of the ESAG9-EQ gene, respectively. The results of the analysis of the ESAG9-EQ 5'UTR are given in Chapter 4, whereas the results of analysis of the ESAG9-EQ 3'UTR are given in the rest of this chapter.

3.3.1 Analysis of the ESAG9-EQ full-length 3'UTR and 3'UTR deletion series

For ESAG9-EQ the nearest downstream gene annotated in the *T. b. brucei* TREU 927 genome sequence is Tb927.5.4630. The intergenic region between the stop codon of ESAG9-EQ and the start codon of Tb927.5.4630 is 3057 bp, which is shown in Figure 3.5. One polyadenylation site for ESAG9-EQ has been previously mapped from cDNA to locate at position +217 bp relative to the ESAG9-EQ stop codon (Keith Matthews, unpublished observations) (see Figure 3.5). To include any regulatory sequences up to, and beyond, the polyadenylation site, 400nt of the downstream sequence was used for analysis. Therefore, for cloning of the 3'UTR of ESAG9-EQ, primers 12 and 13 were used to amplify by PCR 400nt of sequence downstream of the ESAG9-EQ stop codon, this creating the 'full-length' 3'UTR construct. Initially, PCRs were performed using genomic DNA from *T. b. brucei* strain Lister 427-449 as template. However, these reactions were unsuccessful, possibly due to the absence of ESAG9-EQ in the ESAG9 gene repertoire of this strain, or strain differences in the sequence regions where primers were designed to anneal. However, success was achieved by using genomic DNA from *T. b. brucei* EATRO 2340 as template instead.

A

```
3'UTR(s)                                                    350nt(s)
TAGaaatcattaccgcaattccatggaatactgtgaaaacgtctattgagattaacaccgcatgatga
g
                                  300nt(s)

____
gaggaagagttgaaagttggaaactagcttgcaactgctttcaatgaagatcatgtggatggaacgtt
g
 element    250nt(s)
tggagagacaagataagaggatcgaataacactaatgtaatcataagtagcgtacatgagtaaaacca
g

taactggaatggaggaaaagcagaataaagttgttagcgcaagaagcaaaaaatgttcacattaagaa
c

gacacgaggcaaggtgagaaacaggcgacaaccttaacaatattattattaattaaacaaggaaattt
a
                                                          3'UTR(e)
acttcaaaagtattaaggacatcgtagcattgaatttttttttttagggaacgtgataa…(2647bp)…
ATG
```

Figure 3.5 continues on the following page

B

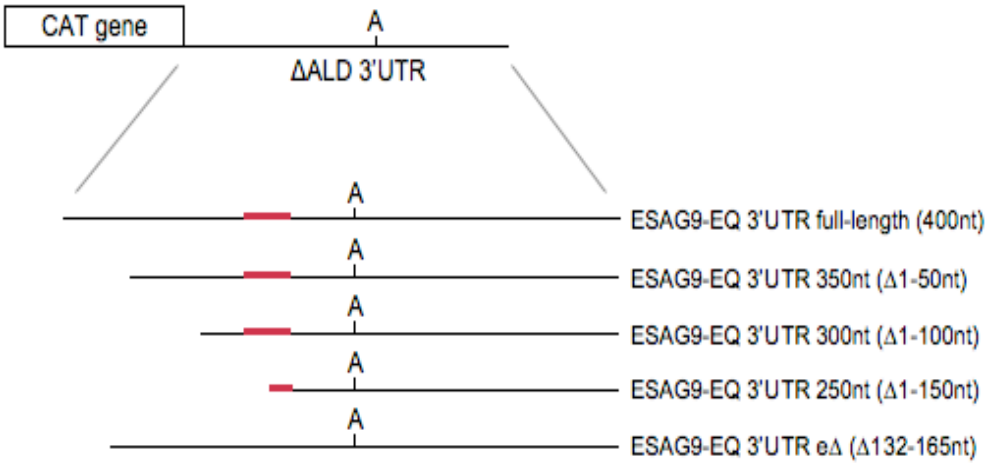


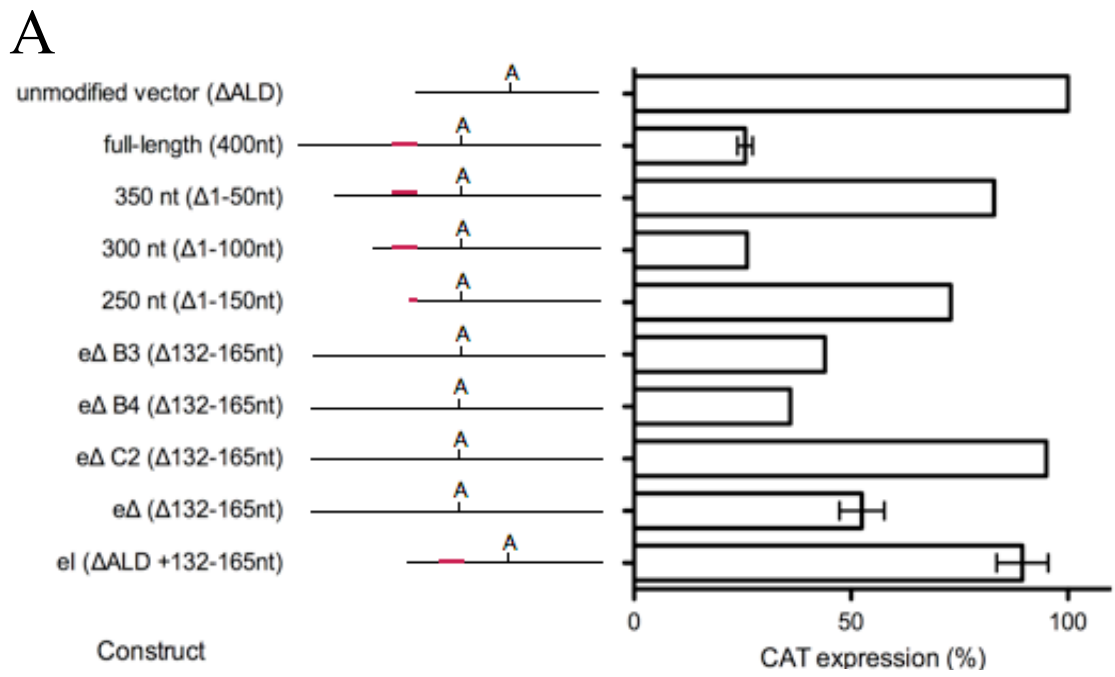
Figure 3.5 Downstream intergenic region for Tb927.5.4620 (ESAG9-EQ). A. The genomic sequence from the stop codon of the ESAG9-EQ gene to the start codon of the nearest downstream annotated gene, Tb927.5.4630, is shown. Because this intergenic region is 3057 bp in length, only the first 400nt of the sequence is shown. The regulatory element sequence is indicated by the line. The sequence is shown in the 5' to 3' direction. Red font: stop codon, red highlight: start codon, grey highlight: start (s) or end (e) of each sequence used to generate the construct named above that residue, a: polyadenylation site (PAS) mapped previously from cDNA (+217 bp relative to ESAG9-EQ stop codon), pink highlight: PASs from 5'RACE of e clones B3, B4, and C2 (see section 3.4.1). B. Schematic diagram showing replacement of the truncated aldolase 3'UTR (ΔALD 3'UTR) in the CAT449 plasmid with each of the ESAG9-EQ 3'UTR sequences used for analysis. 'A': PAS; red bar: element sequence location.

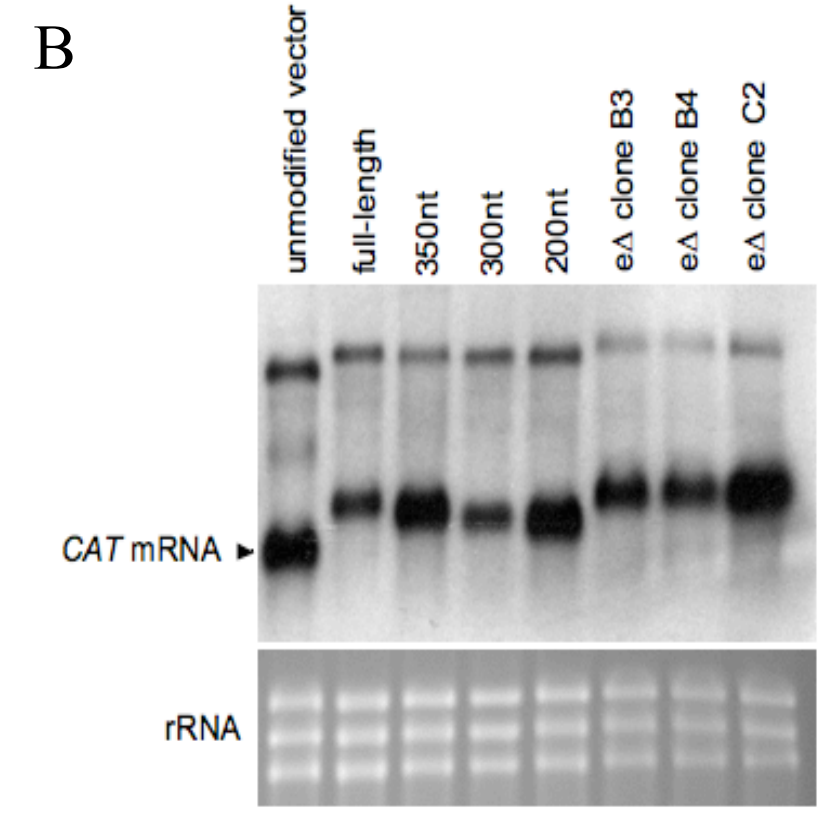
As for TbACO, a deletion series was constructed by substituting the forward primer of the PCR reaction with one complementary to sequence either about 50nt, 100nt, or 150nt downstream of the ESAG9-EQ stop codon (primers 14, 15, and 16, see Table 2.1 and Figure 3.1 B) these being used with primer 13, to generate the sequences used for the 3'UTR constructs '350nt', '300nt', and '250nt', respectively (see Figure 3.5). As the mapped PAS was located at +217 bp, all the deletion series constructs were expected to maintain the PAS signal present in the full-length 3'UTR construct. As before, stable transfectant cell lines were generated using *T. b. brucei* monomorphic bloodstream form strain Lister 427, these being selected with 0.5 μg/ml phleomycin. Thereafter, clonal selection for each construct transfection was as

described above (Section 3.1), although, initially only one clone (biological replicate) was analysed for each of the full-length, 350nt, 300nt, and 250nt constructs. Data for the *CAT* mRNA and protein derived from the resulting clones is shown in Figure 3.6 (A and B).

In comparison to the unmodified vector, the full-length ESAG9-EQ 3'UTR construct conferred repression of CAT expression at the level of protein (Figure 3.6 A; 26% ±1.7% of CAT protein expression of the parental clone) and at the level of mRNA (Figure 3.6 B; 27% the level of *CAT* mRNA expression of the unmodified vector). The data presented in Figure 3.6 (A) for the full-length construct is the mean of 5 biological replicates (independently derived clones). Because expression of ESAG9-EQ mRNA is repressed in slender forms (Barnwell *et al*., 2010, see Figure 1.4), a repression of CAT expression was expected from any sequence that contributes to developmental expression. Thus, from Figure 3.6, it can be concluded that the ESAG9-EQ 3'UTR contributes to the repression of gene expression in slender forms. Because the level of both *CAT* RNA and CAT protein were low, this demonstrates that the control of ESAG9-EQ expression in slender forms is likely to be generated at the level of mRNA abundance, contrasting with TbACO where translational control appears important.

Interestingly, investigation of the deletion series constructs revealed that the repression seen with the intact 3'UTR was then lost with the 350nt and 250nt constructs, where CAT protein expression increased to 83% and 73% respectively. However, repression was retained with the 300nt construct, since CAT protein expression remained at 26% of that seen for the unmodified vector (Figure 3.6 A). Figure 3.6 (B) shows that for each deletion construct construct, the RNA expression level matched the pattern of protein expression, such that *CAT* mRNA expression increased with the 350nt deletion construct (75% the level *CAT* mRNA of the unmodified vector), decreased with the 300nt construct (53%) and increased again with the 250nt construct (151%). Therefore, this supports the level of RNA as being important for the regulation of ESAG9-EQ expression in slender forms.





Normalisation (%): 100 27 75 53 151 94 66 135

Figure 3.6 Analysis of the ESAG9-EQ 3'UTR using a CAT reporter gene assay. A. CAT protein expression from the constructs shown as a percentage of expression from one clone of unmodified vector (control). The number of biological replicates (independently derived clones) used was: Full-length: 5, 350nt, 300nt, 250nt, eΔ B3/B4/C2: 1 each, eΔ: 6, eI: 4; each with 1-2 experimental replicates. Mean ±

standard error of biological replicates is shown on the graph. At the left of the graph are schematic representations of the sequence analysed in each construct, with the polyadenylation site location indicated with 'A' and the regulatory element location indicated by the red line. B. Northern blot analysis using a digoxygenin-labelled RNA probe against the CAT coding region, with RNA from the unmodified vector, 1 biological replicate of each of the full-length, 350nt, 300nt, and 250nt constructs, and 3 biological replicates of the eΔ construct. The lower panel shows the ethidium bromide-staining of the rRNA present in each sample analysed to indicate the loading of each lane. Shown below are the values of mRNA quantitation following normalisation to rRNA as a percentage of that obtained for the unmodified vector.

By northern blot analysis (Figure 3.6 B) size differences were observed in the *CAT* mRNAs derived from the full-length, 350nt, 300nt and 250nt constructs in comparison to the unmodified vector and each other. As for the TbACO 5' and 3'UTR constructs, this is likely to be due to differences in UTR length between the constructs, with each of the ESAG9-EQ 3'UTR-based mRNAs using the same PAS in each construct.

3.3.2 Possible regulatory regions within the ESAG9-EQ 3'UTR

On initial examination, these results could indicate that there is a region that negatively regulates expression that is lost with the first 50nt deletion (350nt), a region that positively regulates expression lost with the next 50nt deletion (300nt), and again a region that negatively regulates expression that is lost with the final 50nt deletion (250nt). However, it is also possible that there may be a structure or structures present within the 3'UTR that are responsible for the results observed. Therefore, to investigate the potential RNA structures of the full-length and deletion series 3'UTRs, the *S*fold programme, *S*rna, was used to analyse each 3'UTR construct up to the mapped ESAG9-EQ PAS.

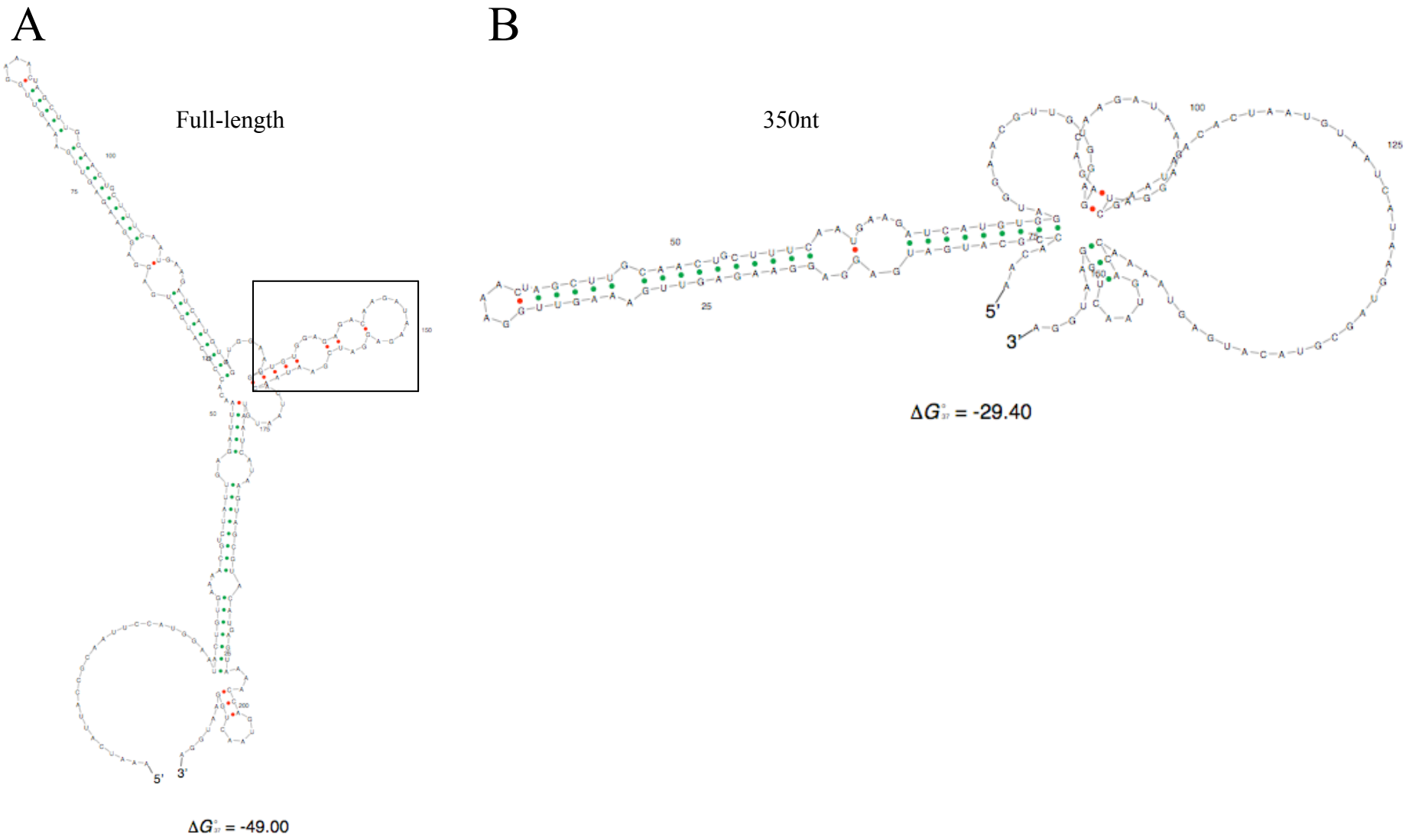


Figure 3.7 continues on the following page

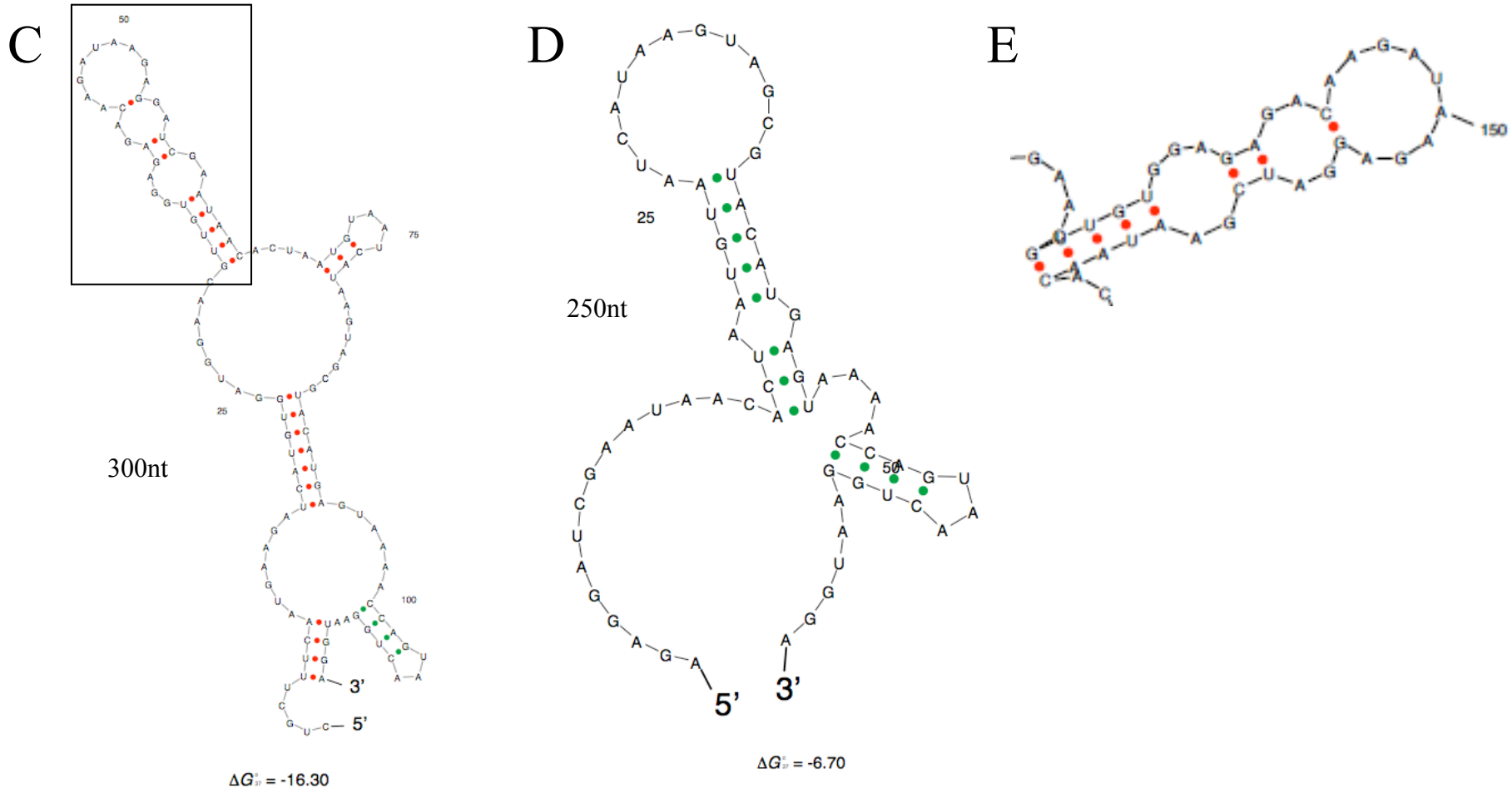


Figure 3.7 RNA secondary structure predictions for ESAG9-EQ 3'UTR sequences. The *S*fold (http://sfold.wadsworth.org/cgi-bin/index.pl) software package *S*rna was used with the sequences used for construct generation extending to the mapped PAS. Shown are the Ensemble Centroid structures generated. A. Full-length 3'UTR construct sequence. B. 350nt 3'UTR construct sequence. C. 300nt 3'UTR construct sequence. D. 250nt 3'UTR construct sequence. E. Enlargement of boxed region in A. Boxes indicate the regulatory element structure.

The *S*fold ensemble centroid RNA structure prediction for each construct is provided in Figure 3.7. Interestingly, the two constructs that confer repression of CAT gene expression, the full length 3'UTR and 300nt 3'UTR, are predicted to possess a stem-loop structure that is absent from the other two ensemble centroid structures, either because it is not predicted to form (350nt 3'UTR construct) or because some of the sequence forming the structure has been deleted (250nt 3'UTR construct). The structure is indicated with boxes in Figure 3.7 (A and C) and an enlargement of the structure is shown in Figure 3.7 (E). This structure, therefore, represents a candidate for the regulatory element responsible for repression of gene expression caused by the ESAG9-EQ 3'UTR, and was termed the ESAG9-EQ 'element'. Consequently, the next step was to investigate this element in more detail.

3.4 Element deletion

To assess whether the element structure identified from the *S*fold analysis was responsible for the repression of CAT gene expression observed with the constructs in which it is predicted to form, the sequence for the element structure was deleted from the full-length 3'UTR sequence. Hence, if this structure was responsible for repression of gene expression, its deletion from the construct would result in an increase in CAT gene expression, due to an alleviation of the repressive signal.

The sub-cloning strategy for this deletion construct replaced the element sequence with an *Xba*I restriction site (see Figure 3.8). To achieve this, two PCR reactions were performed using the PCR product that was used to create the full-length construct as template; one reaction with primers 17 and 12 to amplify the region of the full-length 3'UTR sequence upstream of the element sequence, the other with primers 18 and 13 to amplify the region of the full-length 3'UTR sequence downstream of the element sequence. The two PCR amplicons were used in a single ligation reaction with *Bam*HI- and *Bbs*I-digested CAT449 vector. The resulting construct was termed 'element deletion' or 'eΔ'. For this construct, three independently derived clones (biological replicates) were analysed instead of one as for the previous 3'UTR constructs.

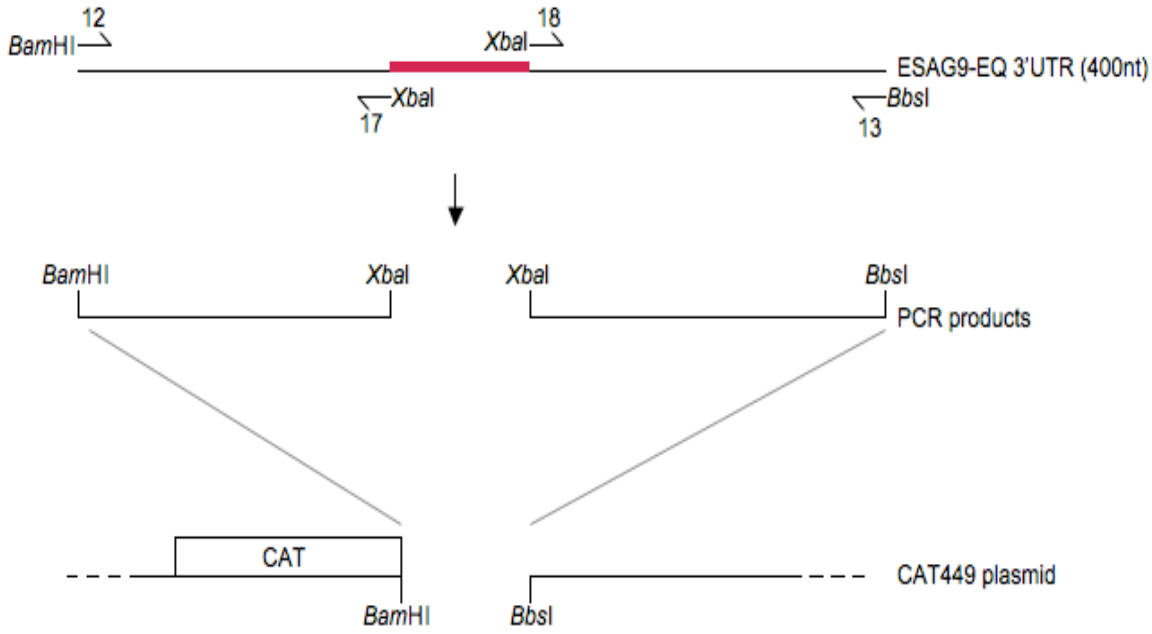


Figure 3.8 Schematic diagram showing the sub-cloning strategy used for generating the ESAG9-EQ element deletion (eΔ) 3'UTR construct. Two PCR reactions were performed using the full-length ESAG9-EQ 3'UTR as template and primers (indicated with bent arrows; see Table 2.1 for the sequences) numbers 12 and 17, and 13 and 18, for two PCRs. Primer 12 contained the *Bam*HI restriction enzyme site sequence; primer 13 contained the *Bbs*I sequence; primers 17 and 18 contained the *Xba*I sequence. These reactions amplified the sequence upstream and downstream of the ESAG9-EQ element sequence (indicated by the red box). Then the PCR products were cloned into the pGEM T-easy plasmid, digested with the appropriate restriction enzymes, and purified (see Section 2.3). These were then used in a ligation reaction with CAT449 plasmid that had been digested with *Bam*HI and *Bbs*I and purified (see Section 2.3.9), in order to replace the truncated aldolase 3'UTR present in this vector, as for the other constructs (see Figure 3.1).

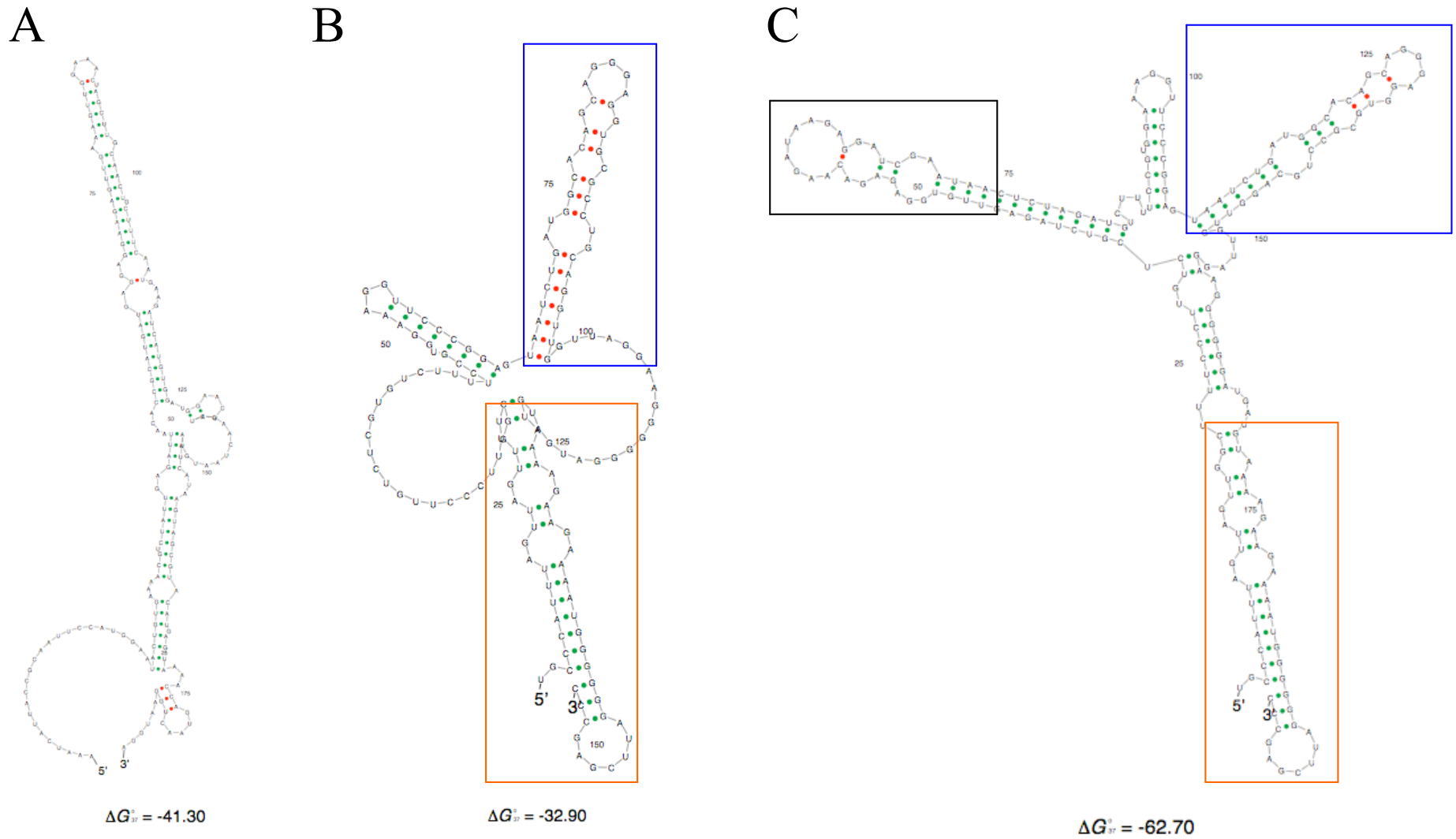


Figure 3.9 continues on the following page

Figure 3.9 RNA secondary structure predictions for ESAG9-EQ element deletion and element insertion 3'UTR sequences. The *S*fold (http://sfold.wadsworth.org/cgi-bin/index.pl) software package *S*rna was used with the sequences used for construct generation. Shown are the Ensemble Centroid structures generated. A. Element deletion (eΔ) construct sequence extending to the mapped PAS. B. Truncated aldolase 3'UTR sequence. C. Element insertion (eI) construct sequence. Black box indicates the regulatory element structure, blue and orange boxes indicate structures conserved between predicted folding of the element insertion and truncated aldolase 3'UTR sequences.

The sequence used for the eΔ 3'UTR sequence was also analysed by *S*fold to verify that the predicted folding of this deleted 3'UTR would not differ significantly from that predicted for the full-length 3'UTR. The ensemble centroid structure obtained for the eΔ 3'UTR sequence is shown in Figure 3.9 (A). In comparison to the ensemble centroid structure obtained for the full-length 3'UTR sequence (see Figure 3.7 A), it is clear that the predicted folding of both sequences is very similar, with only minor alterations to the folding evident around the region of the element deletion. This meant that any change in CAT expression levels between the full-length and eΔ 3'UTR constructs should be due to the deletion of the element sequence, or its predicted structure, and not due to any other changes in the predicted structure of the 3'UTR.

In comparison to the full-length construct, two of the three clones analysed of the eΔ construct gave some alleviation of the repression of *CAT* mRNA and protein expression (Figure 3.6 A and B, clones B3 and B4; 44% and 36% of the unmodified vector CAT protein expression, 94% and 66% of the unmodified vector *CAT* mRNA expression, respectively). The third clone, however, exhibited a much greater *CAT* mRNA and protein expression (Figure 3.6 A and B, clone C2; 95% of the unmodified vector CAT protein expression, 135% of the unmodified vector *CAT* mRNA expression). The reason for this difference between the clones is unclear. One possible explanation is that either clones B3 and B4 or clone C2 possessed an incorrect 3'UTR sequence. This could be possible if the full-length 3'UTR amplicon contaminated the ligation reaction for the eΔ CAT449 construct. In this case, the DNA used for transfection would accidentally contain both the full-length and eΔ CAT449 constructs. This is a possibility, since the *CAT* mRNA under the control of either of these UTRs would run at approximately the same molecular weight, as is observed with *CAT* mRNA from the full-length and eΔ construct clones (Figure 3.6 B). Therefore, it was considered important to confirm the CAT 3'UTR sequence present in each transfectant line, this being carried out by 3'RACE for each of the three eΔ clones.

3.4.1 Investigation of possible causes of the aberrant result for clone C2

3'RACE involved generation of cDNA from each RNA sample using an oligo-dT 3' primer containing an adapter sequence (primer 19, see Table 2.1 and Section 2.5; Mayho *et al*., 2006). This cDNA was then used as template in a PCR reaction with a CAT gene-specific primer (primer 21) and a primer consisting of the adapter sequence (primer 20). A second PCR amplification was performed with another CAT gene-specific primer (primer 22), complementary to CAT gene sequence downstream from primer 21, and primer 20 again. This second amplification therefore generated a nested PCR product to increase specificity. The main PCR product from each reaction was then gel-excised and ligated into pGEM T-Easy vector. Following transformation and small-scale plasmid preparation, two bacterial colonies were sequenced with M13 forward and reverse primers (primers 1 and 2), for each of the B3, B4 and C2 clones.

The sequences returned from 3'RACE showed that all three clones contained the correct, deleted, 3'UTR sequence downstream of the CAT gene sequence. Additionally, the 3'RACE sequencing also provided the PAS location for each clone, these being shown in Figure 3.10 (A) as filled circles. There were found to be three different PAS locations and each being also indicated on the sequence in Figure 3.5. As can be seen, the PASs for each clone were located close to each other, although for each there was minor variation between results from the two bacterial colonies from the same clone. One PAS was located further downstream than the rest; this was one of the sequences obtained from clone B3. However, this does not explain the difference in CAT gene expression because the clone showing more abundant expression was clone C2, the PASs for which matched those from clone B3 and clone B4.

As the 3'UTR sequences were confirmed as correct for each eΔ clone and there was no significant difference in PAS location between the clones, another explanation may be that this was just an erroneous result that is not repeatable. To investigate this result, RNA and protein samples were collected again from each clone after the cell lines were revived from storage at -80°C. CAT ELISA assay and northern blot

analysis was repeated with these new samples. However, the results obtained were the same as that presented in Figure 3.6 (data not shown), indicating that the original results were not erroneous.

A

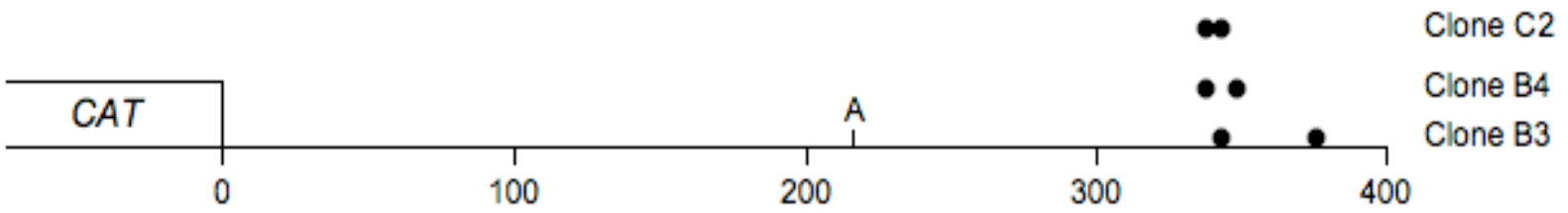


Figure 3.10 continues on the following page

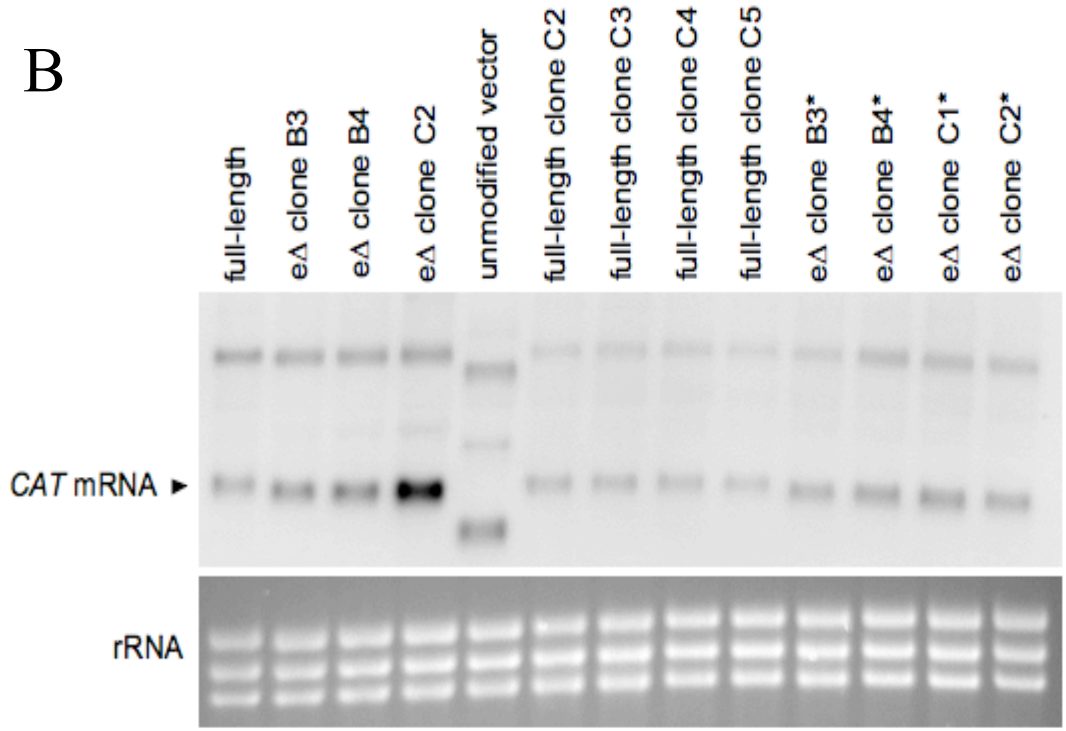


Figure 3.10 Investigation of the aberrant result obtained for the element deletion (eΔ) construct clone C2. A. Polyadenylation site (PAS) locations identified by 3'RACE in the 3'UTR of the three element deletion (eΔ) construct clones (B3, B4 and C2). PAS locations identified by 3'RACE are indicated by filled circles; these locations are also shown on the sequence in Figure 3.5. The PAS location previously mapped for ESAG9-EQ cDNA (Keith Matthews, unpublished observations; see Section 3.3.1) is indicated by 'A'. B. Northern blot analysis using a digoxygenin-labelled RNA probe against the CAT coding region, with RNA from the unmodified vector, full-length and eΔ (clones B3, B4 and C2) clones as shown in Figure 3.6 (B), as well as four additional clones (C2, C3, C4 and C5) of the full-length construct and four additional clones (B3*, B4*, C1* and C2*) of the eΔ construct. The lower panel shows the ethidium bromide-staining of the rRNA present in each sample analysed to indicate the loading of each lane.

3.4.2 Additional full-length and eΔ 3'UTR construct clones

Finally, it was hypothesised that the unusual expression level observed with the eΔ clone C2 may have been caused by integration of the construct into a different location of the β-tubulin locus, an alternative locus, or by multiple integration of the construct into the β-tubulin locus. To investigate whether the result for clone C2 was indeed an odd result, the full-length and eΔ 3'UTR constructs were re-transfected to generate 4 additional and independently derived clones for each construct. CAT mRNA and protein expression was assessed again for these new clones; i.e. full-length clones C2, C3, C4 and C5, and eΔ clones B3*, B4*, C1*, and C2*. All the new full-length clones gave similar CAT protein expression to that obtained from the first clone analysed; therefore, the mean of all 5 of these biological replicates is shown in Figure 3.6 (A). Similarly, all the new eΔ clones gave similar CAT protein expression to that obtained for clones B3 and B4, indicating that the result for clone C2 was an odd result. Hence, the mean of the 6 biological replicates (i.e. excluding the data obtained from clone C2) is presented in Figure 3.6 (A) as eΔ (53% ±5.2% of the CAT protein expression from unmodified vector). The result of the northern blot analysis with RNA from the original and new full-length and eΔ clones is provided in Figure 3.10 (B). As can be seen, the *CAT* mRNA levels are similar for the new clones as the original clones for each construct, except clone C2 which has a much greater *CAT* mRNA expression (262% of the level of *CAT* mRNA expression of the unmodified vector), further confirming this as an odd cell line. It is possible that this is due to integration location or multiple integration of the construct into the β-tubulin locus. However, further experiments, such as southern blot analysis of genomic DNA obtained from each clone, would be needed to confirm this.

From the data obtained from multiple clones of the full-length and eΔ 3'UTR constructs (presented in Figure 3.6), it would appear that deletion of the element sequence, and therefore, predicted structure, causes some alleviation of CAT gene expression at the RNA level, therefore, suggesting that it contributes to the repression of gene expression in slender forms. However, comparison of the data obtained from the full-length, 250nt and eΔ 3'UTR constructs in Figure 3.6 (A), shows that the 250nt construct alleviates repression more than the eΔ 3'UTR

construct, indicating that the element is not fully responsible repression. Instead, additional regulatory region(s) must contribute.

3.5 Insertion of the element into an unrelated 3'UTR

To confirm the negative regulatory function of the element deleted in the eΔ 3'UTR construct, the element was inserted into an unrelated 3'UTR sequence. Specifically, the stem-loop element sequence was inserted into the truncated aldolase 3'UTR sequence that is present in the CAT449 vector, to generate an 'element insertion' or 'eI' 3'UTR construct.

A number of potential different placements of the element sequence within the truncated aldolase 3'UTR sequence were analysed using *S*fold to identify the location that resulted in the least disruption to the structures already present. This was carried out to minimise the chance that any change in CAT expression was caused by a change in overall structure, rather than insertion of the element. Figure 3.9 (B) shows the ensemble centroid structure predicted for the truncated aldolase 3'UTR. As can be seen there are three prominent stem-loop structures predicted and two loops of sequence without predicted folding. Thus, it was thought that these sequences without predicted structure would be less likely to contribute to gene regulation and so insertion of the element sequence in various positions along these sequences was tested. Figure 3.9 (C) shows the ensemble centroid structure predicted for the location chosen and, so, the eI 3'UTR sequence. This maintains two of the three prominent stem-loop structures predicted in the truncated aldolase 3'UTR (blue and orange boxed regions in Figure 3.9 B and C), however, the third predicted prominent stem-loop was not predicted to form. The element structure was also predicted to form (indicated with a black box in Figure 3.9 C). However, in order to obtain this construct it was necessary to include a palindromic sequence at the 5' and 3' ends of the stem-loop element sequence to assist its predicted folding. For this, the *Xba*I restriction site sequence was chosen to maintain consistency with the eΔ 3'UTR construct.

3.5.1 Sub-cloning of the eI 3'UTR construct

Sub-cloning of the eI sequence to create this construct proved difficult. The initial sub-cloning strategy involved three PCR reactions using the unmodified vector or full-length 3'UTR construct DNA as template; one reaction with primers 23 and 24 (see Table 2.1) to amplify the region of the truncated aldolase 3'UTR sequence upstream of the site for element insertion ('aldolase 1.1'), another with primers 25 and 26 to amplify the region of the truncated aldolase 3'UTR sequence downstream of the site for element insertion ('aldolase 1.2'), the third with primers 27 and 28 to amplify the element sequence. Primers 24, 25, 27 and 28 contained the *Xba*I restriction site sequence, primer 23 contained the *Bam*HI restriction site sequence, and primer 26 contained the *Bbs*I restriction site sequence. Thus, after insertion of each PCR product into pGEM T-easy vector, digestion of the sequences with these three restriction enzymes, and then ligation of the three digested sequences in one ligation reaction with *Bam*HI- and *Bbs*I-digested CAT449 vector, the eI construct would be generated. However, it became clear that this strategy would not work, because the element sequence PCR product was very short in length and one ligation reaction with all three digested sequences would not be successful. Moreover, ligation of two digested products into pGEM T-easy vector, then digestion of this for ligation of the third digested product would not work either due to the presence of two *Xba*I restriction sites in the eI sequence.

Therefore, a second sub-cloning strategy was attempted. This used the aldolase 1.1 and 1.2 pGEM T-easy constructs generated above, but instead of generation of the element sequence through PCR, an approach with complementary oligonucleotide sequences was used. To circumvent the problem of two *Xba*I restriction sites mentioned above, one of the *Xba*I restriction sites was changed so that one nucleotide of the sequence was mutated to another nucleotide. Thus, the sequence would not be recognised for digestion by *Xba*I but would still assist the folding of the element structure. After analysis using *S*fold of different mutations of the last nucleotide in the upstream *Xba*I restriction site in the eI sequence, the mutation found to least affect the predicted folding was A-to-G. The complementary oligonucleotide sequences designed were primers 29 and 30 (see Table 2.1). The

strategy involved annealing of these oligonucleotides together in an annealing reaction (see Section 2.3.6) to generate the element sequence with overhangs compatible for insertion into *Xba*I digested sequences (see Figure 3.11). Then, a ligation reaction would be performed with the annealed oligonucleotides and *Xba*I-digested aldolase 1.1 pGEM T-easy vector. This construct would then be digested with *Xba*I and *Pst*I and ligated with *Xba*I- and *Pst*I-digested aldolase 1.2 pGEM T-easy vector. However, despite numerous attempts, ligation of annealed oligonucleotides into *Xba*I-digested aldolase 1.1 pGEM T-easy vector was always unsuccessful. This was hampered by the inability to confirm successful annealing of the complementary oligonucleotides, due to the short length of their sequence and the presence of overhangs preventing the ligation into pGEM T-easy vector for sequencing. It is likely that annealing was unsuccessful due to the relatively extended length of the oligonucleotides and the element sequence, which may have caused the formation of hairpins in the oligonucleotides or annealing of the oligonucleotides so that the correct overhangs for insertion into *Xba*I digested sequences were not present. Therefore, this cloning strategy was abandoned.

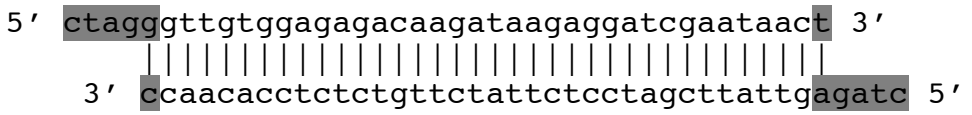


Figure 3.11 Annealed complementary oligonucleotides for an eI 3'UTR sub-cloning strategy. The sequence is the identified ESAG9-EQ element sequence with nucleotides added at both 5' and 3' ends to create overhangs compatible for insertion of the annealed oligonucleotides into *Xba*I-digested DNA (grey highlight).

A strategy involving mutagenesis of the truncated aldolase 3'UTR was then considered. This would consist of mutagenesis to insert the element sequence with flanking *Xba*I restriction enzyme sites into the truncated aldolase 3'UTR using the QuikChange XL Site-Directed Mutagenesis kit (Stratagene). However, due to the extended length of the sequence for the insert and the restrictions on design of the primers for successful mutagenesis, primers could not be designed according to the

guidelines for mutagenesis, even using only a third of the insert sequence at a time. Therefore, this approach was not initiated.

Finally, the eI 3'UTR sequence, with flanking *Bam*HI and *Bbs*I restriction enzyme sites for ligation into *Bam*HI- and *Bbs*I-digested CAT449 vector, was synthesised by a commercial company, GENEART AG (Germany; GENEART services presently being offered through www.invitrogen.com), in their pMA-T vector. However, digestion of this construct with *Bam*HI and *Bbs*I produced insufficient amount of digested insert DNA for gel-purification and use in a ligation reaction. Attempts to transform *E. coli* XL-blue bacteria with the plasmid were unsuccessful, as DNA obtained from transformed bacterial colonies ran on agarose gels at a much lower molecular weight. Possibly this was due to incompatibility of the plasmid and the bacterial strain. Consequently, a PCR reaction was performed with primers 31 and 32, using the pMA-T plasmid sequence as template, to amplify the region of the plasmid containing the eI sequence. The eI sequence was verified after insertion into pGEM T-easy and following *Bam*HI and *Bbs*I digestion was inserted into *Bam*HI- and *Bbs*I-digested CAT449 vector. As for the previous constructs analysed, *T. b. brucei* monomorphic bloodstream form strain Lister 427 were then transfected with the construct, although the selection of stable transfectant clones used 2.5 μg/ml phleomycin. Four independently derived clones (biological replicates) of the eI construct were analysed.

3.5.2 Element insertion causes a small decrease in CAT reporter gene expression

The four eI construct clones analysed each resulted in very limited repression of CAT protein expression in comparison to the unmodified vector (Figure 3.6 A), giving a mean CAT protein expression of 90% ±6.0 of the expression from the unmodified vector. The *CAT* mRNA expression for each of the four eI construct clones is shown in Figure 3.12. The *CAT* mRNA expression level was also reduced with this construct, as the mean level of CAT mRNA expression for the clones was 67% of that of the unmodified vector (52%, 49%, 51% and 114% for clones D1, D4, C5 and C6, respectively).

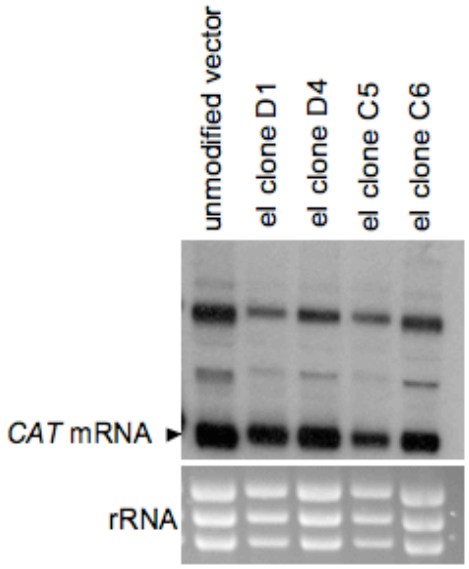


Figure 3.12 Analysis of the *CAT* mRNA level obtained with the element insertion (eI) construct in a CAT reporter gene assay. Northern blot analysis using a digoxygenin-labelled RNA probe against the CAT coding region, with RNA from the unmodified vector and 4 biological replicates of the eI construct (clones D1, D4, C5, and C6). The lower panel shows the ethidium bromide-staining of the rRNA present in each sample analysed to indicate the loading of each lane.

As described above, the only difference between the 3'UTR downstream of the CAT gene in the eI construct and the unmodified vector is the insertion of the regulatory element sequence. Therefore, it is likely that the element sequence is responsible for this repression. The repression seen, however, was very minor and much less than seen with the full-length 3'UTR construct or as great as the extent of repression release seen with the eΔ 3'UTR construct compared to the full-length 3'UTR construct. This could be due to an inability of the element sequence to form the predicted structure (see Figure 3.7 E). Another possibility is that additional sequence of the ESAG9-EQ 3'UTR surrounding the element sequence is needed for the element's negative regulatory function or that regulatory proteins that bind the element sequence are unable to due to the secondary structure present in this 3'UTR. However, the small repression of gene expression seen with the eI construct does support the negative regulatory function of the element sequence, even out of the ESAG9-EQ 3'UTR sequence context. Out of context, however, its contribution was only minor.

3.6 Additional 3'UTR deletion series cell lines

With the odd result obtained from one clone of the eΔ construct (clone C2, see Section 3.4), the analysis of only one clone expressing CAT under the control of the 350nt, 300nt, and 250nt 3'UTRs could be criticised, as it unknown whether any of these represented aberrant results too. Therefore, an analysis of more clones for each of these constructs was needed. To achieve this, each construct was re-transfected, again into the *T. b. brucei* monomorphic bloodstream form strain Lister 427, but selection of stable transfectant clones was with 2.5 μg/ml (instead of with 0.5 μg/ml previously) phleomycin. Two independently derived clones (biological replicates) of each of the 350nt, 300nt, and 250nt 3'UTR constructs were analysed. Hence, with the previous clones analysed (Section 3.3.1), a total of three biological replicates were analysed for each of these constructs.

3.6.1 Additional 350nt construct clones

The CAT protein expression obtained from the additional clones is shown in Figure 3.13 (A) as a percentage of that obtained from the unmodified vector. The result presented in this graph for the 350nt construct is the mean of the two additional clones (D5 and C2) because, as can be seen when comparing Figure 3.6 (A) and Figure 3.13 (A), these two clones resulted in a much lower level of CAT protein expression than the original 350nt construct clone (clones of the D5 and C2 had a CAT protein expression level of 47% ±2.5% versus the original clone that had a level of 83%). The *CAT* mRNA expression for these two additional clones is shown in Figure 3.13 (B) and appears to be variable between the two clones. However, when the levels of *CAT* mRNA were quantified using the rRNA level, both clones were found to have a similar expression level, this being 48% and 42% of the expression from the unmodified vector (for clone D5 and C2, respectively). Therefore, both clones D5 and C2 had a lower *CAT* mRNA signal than that of the unmodified vector, in contrast to the original clone, which had a *CAT* mRNA expression level of 75%. Therefore, unfortunately it appears that the original 350nt construct clone may not have been representative, similarly to the eΔ construct clone

C2. It was for this reason that the original clone was excluded from the data presented in Figure 3.13 (A).

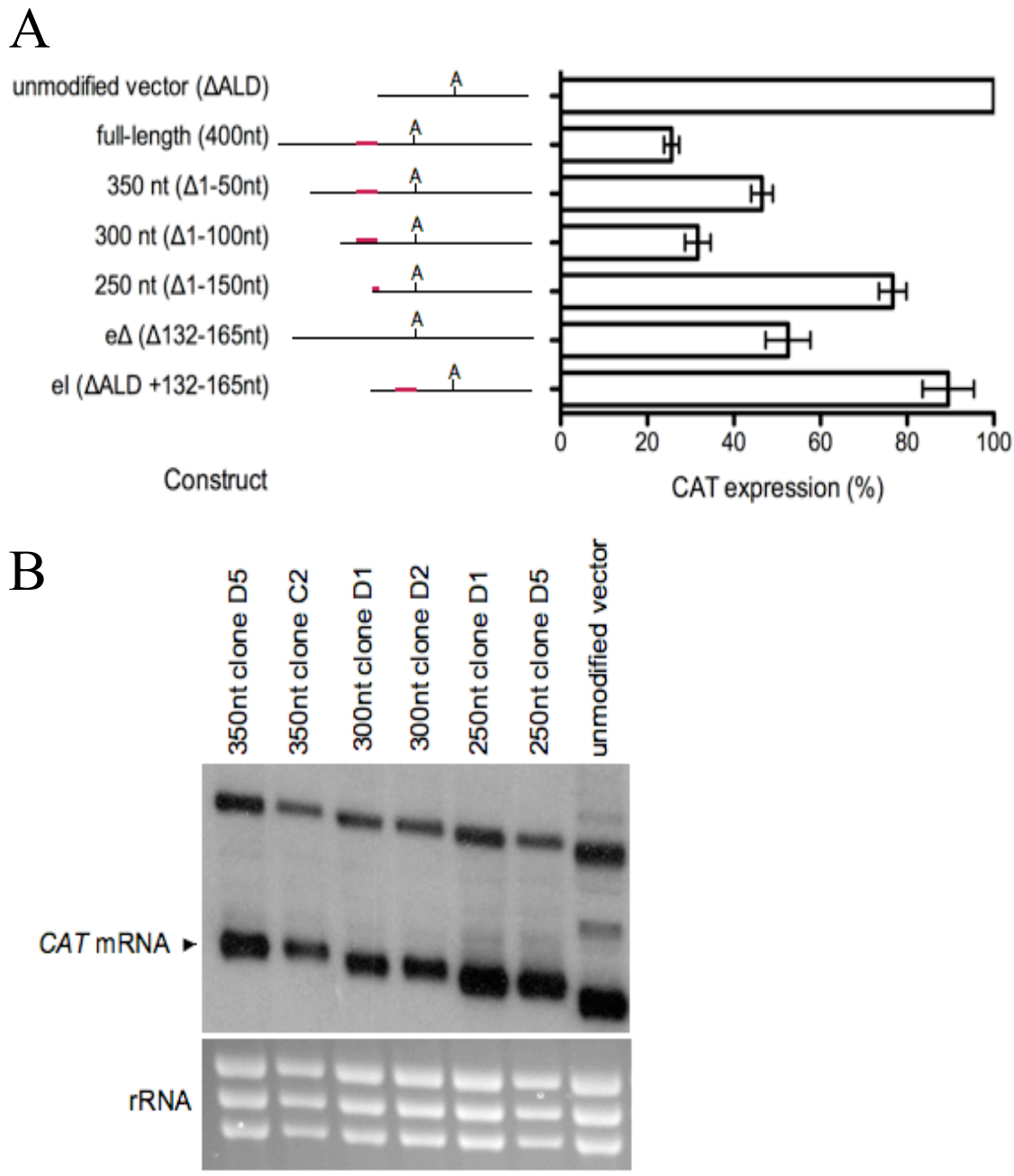


Figure 3.13 Additional 350nt, 300nt and 250nt construct clone analysis using a CAT reporter gene assay. A. CAT protein expression from the constructs shown as a percentage of expression from one unmodified vector clone (control). Final results incorporating data obtained from all clones, excluding any odd results. Number of biological replicates (independently derived clones) used: Full-length: 5, 350nt: 2, 300nt: 3, 250nt: 3, eΔ: 6, eI: 4; each with 1-2 experimental replicates. Mean ±

standard error of biological replicates is shown on the graph. At the left of the graph are schematic representations of the sequence analysed in each construct, with the polyadenylation site location indicated with 'A' and regulatory element location indicated by the red line. B. Northern blot analysis using a digoxygenin-labelled RNA probe against the CAT coding region of RNA samples from the additional clones of the 350nt, 300nt, and 250nt constructs and one clone of unmodified vector as a control. Shown in the lower panel is the ethidium bromide-staining of the rRNA present in each sample to indicate the loading of each lane.

3.6.2 Additional 300nt and 250nt construct clones

Fortunately, the additional 300nt and 250nt construct clones gave similar CAT expression results to their respective original clones. Therefore, the data presented in Figure 3.13 (A) for each of these constructs is the mean of the three clones (biological replicates). The mean values are therefore, for the 300nt construct clones, 32% of the CAT protein expression obtained from the unmodified vector (previously 26% for the original clone alone), and for the 250nt construct clones, 77% (previously 73% for the original clone alone). Hence, none of the clones for either of these two constructs appeared to have produced an aberrant level of CAT protein expression and so the data for all these clones was included in the data presented in Figure 3.13.

Analysis of the *CAT* mRNA levels for these additional clones is also shown in Figure 3.13 (B). For each of the 300nt and 250nt constructs, the individual clones produced a similar level of *CAT* mRNA expression to each other. These levels are similar to the CAT protein expression level obtained for each construct and similar to that observed for the original clones of each construct (see Figure 3.6 B).

The discovery that the original 350nt clone was generating an aberrant expression profile was disappointing since from the data presented in Figure 3.13 it now appeared that the loss of the predicted element structure in this construct was not responsible for the entire alleviation of repression seen with the 250nt construct. Instead, the predicted element structure may be responsible for some of this alleviation and possibly there is another negative regulatory region that is lost in the deletion to form the 250nt construct.

3.7 Is the ESAG9-EQ 3'UTR responsible for gene expression control in 'stumpy-like' forms?

Next, the question of whether the ESAG9-EQ 3'UTR is responsible for control of gene expression, i.e. release of repression, in stumpy forms was addressed. For this, the same CAT reporter gene assay approach could be used as to dissect the control of gene expression in slender forms. However, the cells used for the CAT reporter gene assays above were monomorphic and so these cells have lost the ability to differentiate to stumpy forms. The ideal experiment would be to analyse the same constructs in pleomorphic bloodstream form cells, where true slender and stumpy forms could be analysed. However, this would involve re-transfection of each construct and so re-generation of the cell lines, a huge task, given the inefficient transfection of pleomorphic trypanosomes. Therefore, experiments were performed on the existing monomorphic cell lines analysed above, which were driven to embark on stumpy formation by chemical induction.

3.7.1 Generation of monomorphic 'stumpy-like' forms

Treatment with a compound, 8-pCPT-2'-O-Me-cAMP, has been previously reported to turn monomorphic bloodstream form *T. brucei* into 'stumpy-like' forms (Laxman *et al*., 2006). Laxman *et al*. (2006) showed that 8-pCPT-2'-O-Me-cAMP, a hydrolysable cAMP analogue, has an anti-proliferative effect in bloodstream form *T. brucei* cells and that treated cells also have stumpy morphology and NADH diaphorase activity (a marker of intermediate and stumpy forms). Therefore, treatment with this compound was used to assess the control of CAT reporter gene expression in ESAG9-EQ 3'UTR based construct cell lines when 'stumpy-like' to give an idea of expression in stumpy forms.

These experiments involved treatment of each cell line with 10 μM 8-pCPT-2'-O-Me-cAMP (dissolved in DMSO) and, as a control, treatment with DMSO alone (see Section 2.2.3). CAT ELISA assay and northern blot analysis were performed in the same way described above in Section 3.1.2, using samples taken from both the treated and control cells 48 hours post-addition. The CAT protein expression results

are presented as the fold change in CAT protein levels between the treated, stumpy-like, cells and the control cells. Moreover, for each construct there were 2 or 3 biological replicates. For each cell line there were two experimental replicates, meaning that the CAT ELISA assays and northern blot analyses were performed with samples taken from two separate 8-pCPT-2'-O-Me-cAMP treatment experiments. This is with the exception of the unmodified vector clone, which had one biological replicate and four experimental replicates, this line being included in each 8-pCPT-2'-O-Me-cAMP treatment experiment as a control.

In each experiment involving treatment with 8-pCPT-2'-O-Me-cAMP, the cell number was determined (see Section 2.2.1) immediately prior to sample generation for CAT ELISA assay and northern blot analysis at 48 hours post-induction. This was to enable the values obtained from the CAT ELISA assay to be adjusted for the cell number for each sample. Also, this allowed confirmation of the anti-proliferative effect of the compound on trypanosomes.

3.7.2 The ESAG9-EQ 3'UTR confers an alleviation of protein repression in stumpy-like forms

8-pCPT-2'-O-Me-cAMP was used to treat the cell lines discussed in Sections 3.3, 3.4, 3.5 and 3.6 to assess the contribution of the various 3'UTRs to control of CAT gene expression upon stumpy-like cell formation. The results for the various 3'UTR constructs are shown in Figure 3.14.

The change in CAT protein levels in cells treated with 10 μM 8-pCPT-2'-O-Me-cAMP compared to control cells is shown in Figure 3.14 (A). The data presented separately for each construct is the mean ± standard error of the biological replicates (unmodified vector and 350nt constructs: $n = 1$, full-length and eΔ constructs: $n = 3$, all other constructs: $n = 2$).

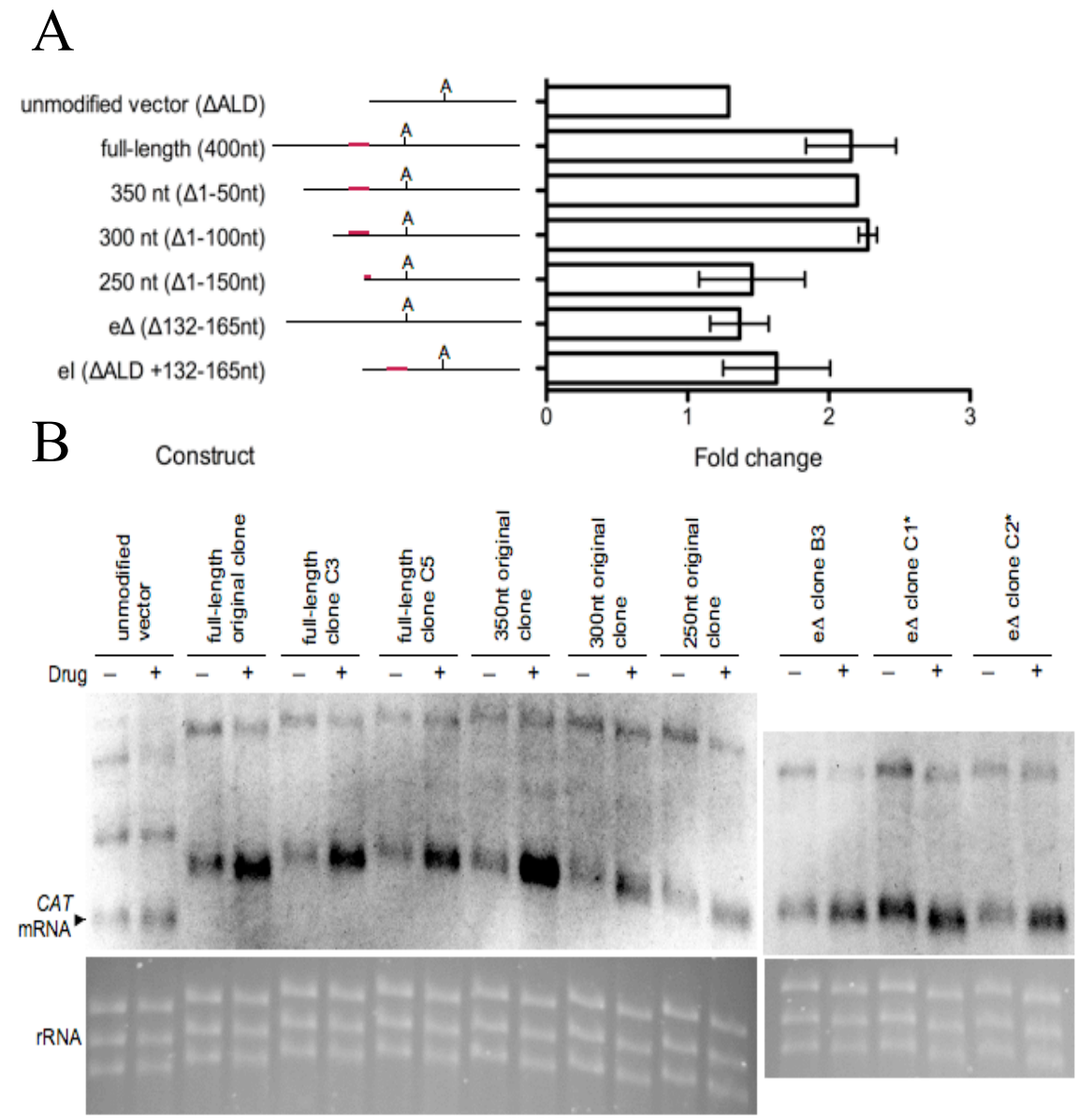


Figure 3.14 continues on the following page

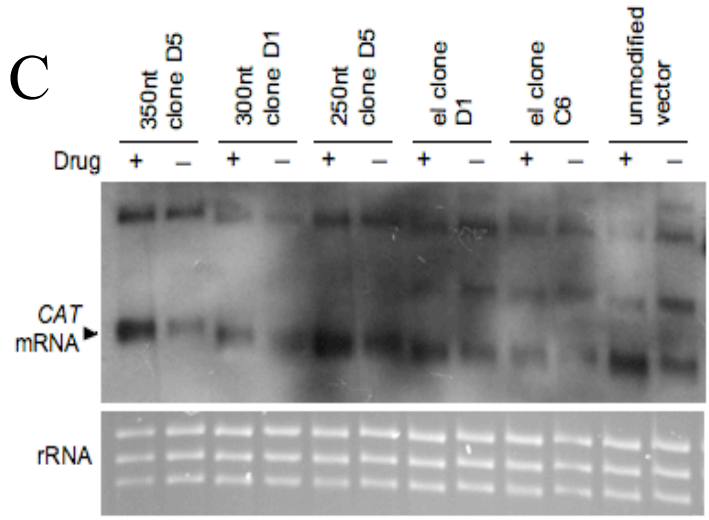


Figure 3.14 Response of CAT reporter gene expression to 8-pCPT-2'-O-Me-cAMP treatment in ESAG9-EQ 3'UTR based constructs. A. Fold change in CAT protein expression levels between cells treated with 10 µM 8-pCPT-2'-O-Me-cAMP and control cells. Shown is the mean ± standard error of biological replicates; unmodified vector and 350nt: $n = 1$, full-length and eΔ: $n = 3$; other constructs: $n = 2$. At the left of the graph are schematic representations of the sequence analysed in each construct, with the polyadenylation site location indicated with 'A' and the regulatory element location indicated by the red line. B. and C. Northern blot analyses of RNA samples from two different 8-pCPT-2'-O-Me-cAMP treatment experiments using a digoxygenin-labelled RNA probe against the CAT coding region, each with the unmodified vector clone as a control. Samples are either from cells treated with 10 µM 8-pCPT-2'-O-Me-cAMP (+ drug) or control cells (- drug). Lower panels show the ethidium bromide-staining of the rRNA present in each sample analysed to indicate the loading of each lane. Shown below are the mRNA quantifications following normalisation to rRNA as a percentage of that obtained for the unmodified vector.

As can be seen in Figure 3.14 (A), there was a small response of the unmodified vector to treatment with the compound (1.3 fold change), although the mRNA levels generated a lower fold-change (1.13 fold-change, Figure 3.14 B). This is unexpected because the truncated aldolase 3'UTR was not expected to be developmentally regulated (see Section 3.1.1). Nonetheless, as this construct was included as a control in all of the experiments performed, the values present in induced and untreated samples provided a suitable normalisation control.

The full-length (400nt) ESAG9-EQ 3'UTR caused an increase in CAT protein expression in response to treatment with this compound by 2.2-fold compared to DMSO treatment. This was a larger increase in the fold change than was observed with the unmodified vector. This is significant because in stumpy forms there is a general repression of translation (Section 1.4.1), whereas for this construct there was increased expression, demonstrating that it was not translationally repressed. This increased expression supported a role for the ESAG9-EQ 3'UTR sequence in the release of gene expression repression in stumpy forms. It also implied that there are regulatory signal(s) within the ESAG9-EQ 3'UTR able to respond upon differentiation to stumpy-like forms.

For the 350nt construct, the data presented in Figure 3.14 (A) is unfortunately only the result obtained for one clone. This is because initial experiments using this compound included the original clone for this construct, which later was discovered to give an aberrant level of CAT protein expression (see Section 3.6.1). The mRNA abundance obtained from this original clone is shown in Figure 3.14 (B), demonstrating that this clone also gives an aberrant result in this drug treatment analysis (5-fold increased mRNA level with drug treatment, see Section 3.7.3 also). Therefore, the results obtained for this clone in stumpy-like forms were not included in the data presented in Figure 3.14 (A). The results obtained from the one clone used for this analysis (clone D5) show that there was no difference in CAT protein expression upon treatment compared to the full-length construct; there was still a fold change of 2.2. Therefore, deletion of the first 50nt of ESAG9-EQ 3'UTR sequence did not affect the response to 8-pCPT-2'-O-Me-cAMP treatment observed with the full-length 3'UTR.

Deletion of the next 50nt of the ESAG9-EQ 3'UTR sequence with the 300nt construct also did not affect the response to 8-pCPT-2'-O-Me-cAMP treatment (2.3 fold change). However, further deletion of the sequence by 50nt more (i.e. the 250nt construct) did affect the response to 8-pCPT-2'-O-Me-cAMP treatment as the fold change decreased to 1.5-fold. This was almost a return to the value obtained for the unmodified vector, indicating a responsive element had been deleted.

Results obtained with the eΔ construct, where only the sequence of the predicted structural element is deleted (see Section 3.4), showed a fold change of 1.4, similar to that of the 250nt construct. Therefore, this suggests that it is the region containing the element sequence that is responsible for the response to 8-pCPT-2'-O-Me-cAMP treatment in the ESAG9-EQ 3'UTR sequence, because its deletion either within the eΔ construct or within the 250nt construct eliminated the response.

To analyse if this element could also respond out of context, the eI construct, where the element sequence is inserted into the truncated aldolase 3'UTR (see Section 3.5), was analysed. This demonstrated a slight increase in the fold change upon treatment, matching the unmodified vector (1.6-fold). Similar to the results obtained from analysis of CAT reporter gene expression in slender monomorphic bloodstream form cells, this supported a small contribution of this sequence to gene expression in stumpy forms (see Section 3.5). Therefore, context is likely important for the functionality of this sequence element.

3.7.3 Response of each construct at the RNA level in stumpy-like forms

Analysis of the *CAT* mRNA levels in response to 8-pCPT-2'-O-Me-cAMP treatment for each of the cell lines is presented in Figure 3.14 (B and C). As mentioned above, treatment with 8-pCPT-2'-O-Me-cAMP resulted in an anti-proliferative effect as previously reported (Laxman *et al*., 2006). This generated low RNA yields for samples from the cells treated with this compound. Consequently, lower amounts of RNA than usual (typically 0.5 μg per sample) were analysed in northern blot analyses in these experiments. As shown in Figure 3.14 (B), there was a slight response of the unmodified vector to the treatment, this producing a small increase in *CAT* mRNA expression levels, similarly to that observed for the CAT protein expression for this clone (Figure 3.14 A). This was more pronounced in the experiments with the second group of cell lines analysed (Figure 3.14 C), as was the CAT protein expression response. The reason for this is unknown. However, when compared to the *CAT* mRNA expression levels for the full-length ESAG9-EQ clones (Figure 3.14 B), it is apparent that there was a much stronger response to treatment in

the test constructs bearing the ESAG9-EQ 3'UTR (1.93 fold-change, Figure 3.14 B). This matched the CAT protein expression response discovered for this construct, confirming that the ESAG9-EQ 3'UTR contributes to alleviation of gene expression repression in stumpy forms, likely through mRNA control.

Similarly, the 3'UTR deletion series also respond to 8-pCPT-2'-O-Me-cAMP treatment with an elevated *CAT* mRNA expression level (Figure 3.14 B and C). This was greatest for the 350nt original clone (Figure 3.14 B), although this was the clone that produced an aberrant level of CAT expression and thus should be ignored from this analysis, as discussed above. Each of the remaining deletion series constructs generated an increase in *CAT* mRNA expression in stumpy-like forms, with these responses being comparable to each other (300nt fold-change: 2.72, 250nt fold change: 2.52). Therefore, the increased CAT protein expression observed for these constructs in stumpy-like forms (Figure 3.14 A) may be due to increased *CAT* mRNA expression.

Interestingly, for the 250nt construct the fold-change of CAT protein was less (1.5-fold, Figure 3.14 A) than the fold-change in mRNA quantification (2.5-fold, Figure 3.14 B) between drug-treated and untreated cells, which indicates that this construct is subject to some translational repression. However, this interpretation is limited by the quantification of only the mRNA abundance obtained from one clone, indeed comparison of the mRNA levels obtained from another clone (Figure 4.13 C) suggests that this may yield a greater fold-change from quantification than the original clone. However, the dark area covering this section of the Northern blot would prevent meaningful quantification. Clearly further Northern blot analyses and quantifications would be required before any firm conclusions could be made, however, this was hampered by the low amount of RNA obtained from the drug-treated cells.

The eΔ 3'UTR clones also exhibited an elevated *CAT* mRNA expression in response to 8-pCPT-2'-O-Me-cAMP treatment (1.47 mean fold-change, Figure 3.14 B). Along with the *CAT* mRNA expression data obtained for the deletion series constructs this

suggests that the element sequence is not responsible for the alleviation of mRNA expression repression. Similarly, the eI 3'UTR construct did not generate a response to 8-pCPT-2'-O-Me-cAMP treatment by greatly increased *CAT* mRNA levels (Figure 3.14 C). Although there did appear to be some variation in the response between clones (e.g. clone D1 appears to respond more than clone C6), the response was of a similar level to that observed with the unmodified vector.

From the results presented in Figure 3.14 (B and C) it can be concluded that the ESAG9-EQ 3'UTR is responsible for alleviation of RNA expression in stumpy forms. Also, we can conclude that the sequences responsible for this are present in all the ESAG9-EQ 3'UTR constructs analysed. Potentially, therefore, the sequence responsible must lie more than 250nt from the stop codon, beyond the region analysed in the deletion constructs used here.

3.8 Analysis of additional unmodified vector cell lines

The earlier results indicated that it was important to analyse multiple clones for a given construct to avoid individual cell artefacts. However, in all the earlier experiments it was the same unmodified vector clone that all experimental 3'UTR constructs were compared to. Therefore, further cell lines were generated to confirm the earlier analyses. Hence, linearised unmodified vector plasmid was re-transfected as described before (Section 3.6) to obtain four additional, independently derived clones (biological replicates), which were analysed alongside the original clone, meaning that in total five unmodified vector clones were analysed.

3.8.1 Analysis of additional unmodified vector clones in slender monomorphic bloodstream form cells

CAT mRNA and CAT protein expression was assessed in the slender monomorphic bloodstream form cells of the additional unmodified vector clones. The CAT protein expression results are presented as the percentage expression obtained for the original clone in Figure 3.15 (A). The data shown is the mean of two experimental replicates for each clone. Three out of the four additional clones, clones D2, D6 and

C5, show a similar level of CAT protein expression to the original clone (99%, 111%, and 127%, respectively), although clone C5 is a little greater than the others.

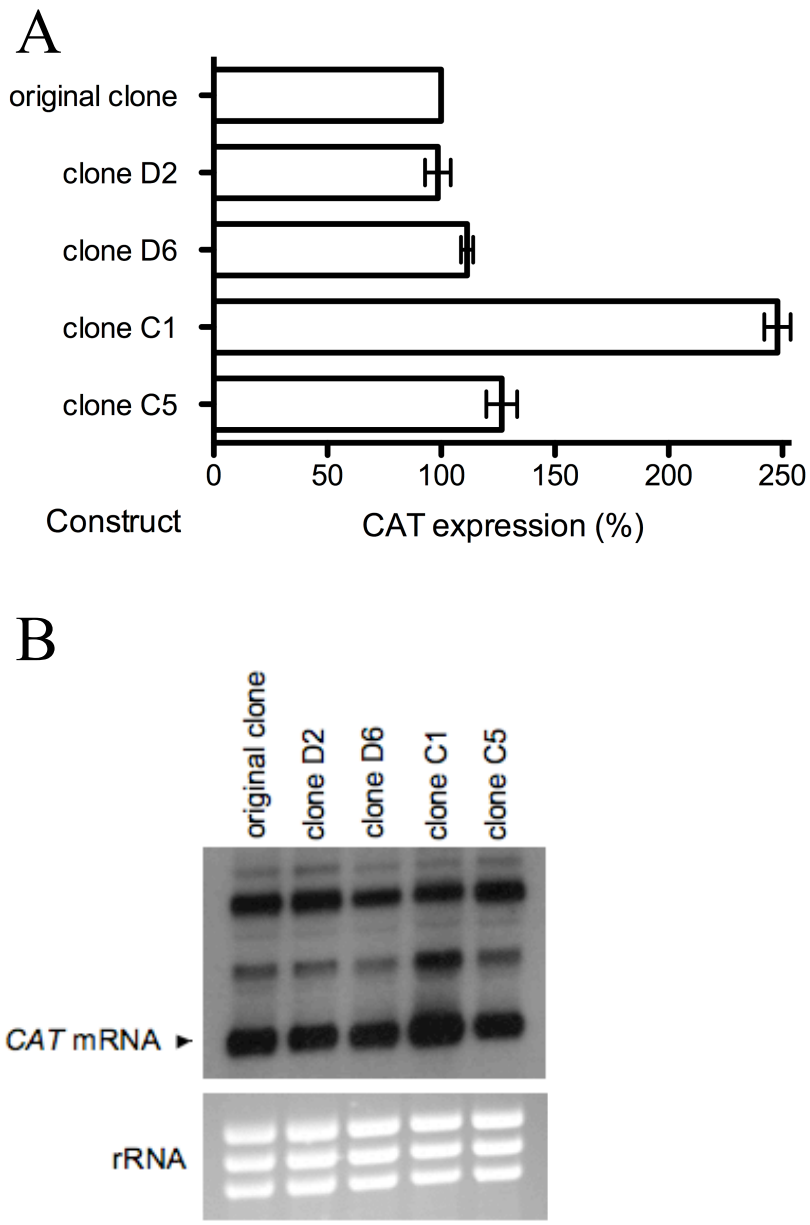


Figure 3.15 Analysis of additional unmodified vector clones using a CAT reporter gene assay. A. CAT protein expression from the additional unmodified vector clones (clones D2, D6, C1 and C5) shown as a percentage of expression from the original unmodified vector clone (control). The mean ± standard error of two experimental replicates is shown on the graph. B. Northern blot analysis of the same clones as in A. using a digoxygenin-labelled RNA probe against the CAT coding region. The lower panel shows the ethidium bromide-staining of the rRNA present in each sample analysed to indicate the loading of each lane.

This, however, is probably not too dissimilar to the variability seen within different clones of the same construct for other constructs. It is also apparent here that clone C1 gave an aberrant expression profile, similar to eΔ clone C2 and the 350nt original clone. The expression level for this clone was 248% of the expression seen with the original clone.

Analysis of the reporter gene mRNA levels revealed similar levels to that of the protein. The northern blot analysis for these clones is shown in Figure 3.15 (B). As can be seen, clones D2, D6 and C5, show a similar level of *CAT* mRNA expression to the original clone (90%, 88%, and 141% of the expression of the unmodified vector, respectively). In contrast, clone C1 exhibited a much greater level of *CAT* mRNA (230% of the expression of the unmodified vector), similar to the CAT protein expression level for this clone.

From the data presented in Figure 3.15, it appeared as if the CAT reporter gene expression levels obtained from the original clone of the unmodified vector were consistent with other clones (D2, D6 and C5). Therefore, although the clones of the other constructs were only compared to this one clone, it seems as if this was an adequate control.

3.8.2 Analysis of additional unmodified vector clones in 'stumpy-like' form cells

It was also important to verify the response of the control construct when treated with 8-pCPT-2'-O-Me-cAMP. Therefore, the additional clones of the unmodified vector were tested in this assay alongside the original clone, with clone C1 being excluded. The assay was performed as described in Section 3.7 and the results are shown in Figure 3.16.

The change in CAT protein levels detected in samples from cells treated with 10 μM 8-pCPT-2'-O-Me-cAMP is shown in Figure 3.16 (A). As the data is presented separately for each clone, the data shown is the mean ± standard error of the experimental replicates (n = 2, except for the original clone where n = 4). Unfortunately, there was some variability in the results obtained from the original

clone in two of the CAT ELISA assays performed, resulting in a larger standard error than for the other clones. Nonetheless, the results demonstrated that between the different clones there was a similar level of change in CAT protein expression upon treatment in the original, D2, D6 and C5 clones (1.3, 1.2, 1.5, 1.3 fold changes, respectively).

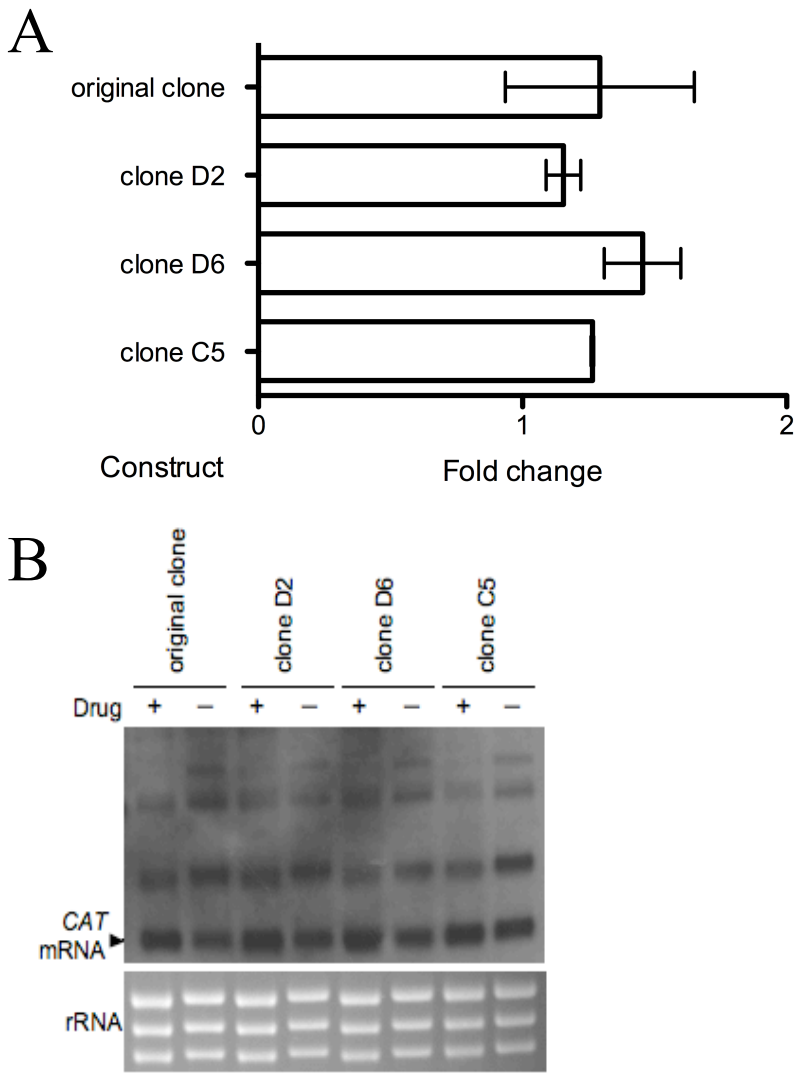


Figure 3.16 Response of CAT reporter gene expression to 8-pCPT-2'-O-Me-cAMP treatment in the unmodified vector clones. A. Fold change in CAT protein expression levels between cells treated 10 µM 8-pCPT-2'-O-Me-cAMP and control cells for each of the unmodified vector clones (the original clone and clones D2, D6 and C2). Shown is the mean ± standard error of experimental replicates; n = 2, except original clone n = 4. B. Northern blot analysis using a digoxygenin-labelled RNA probe against the CAT coding region with RNA samples from each clone shown in A. either treated with 10 µM 8-pCPT-2'-O-Me-cAMP (+ drug) or control untreated

cells (- drug). The lower panel shows the ethidium bromide-staining of the rRNA present in each sample analysed to indicate the loading of each lane.

CAT mRNA expression levels in each of the clones with and without 10 μM 8-pCPT-2'-O-Me-cAMP is shown in Figure 3.16 (B). There was a slight increase in *CAT* mRNA expression in cells treated with the compound, similar to that seen for the CAT protein levels in Figure 3.16 (A). However, this appeared to be similar for all the clones. It was also similar to that seen with the original clone in Figure 3.14 (C).

Therefore, from the data presented in Figure 3.16 it appears to be acceptable that only one clone, the original clone, was used as a control in the 8-pCPT-2'-O-Me-cAMP treatment experiments described in Section 3.7, because similar results were obtained for several additional unmodified vector clones.

3.9 Is the ESAG9-EQ 3'UTR responsible for gene expression control in stumpy forms?

The data obtained from the 8-pCPT-2'-O-Me-cAMP treatment experiments described in Section 3.7 suggested that the ESAG9-EQ 3'UTR was important for the control of gene expression upon development to 'stumpy-like' forms, and in particular that the element sequence identified was responsible for this response. However, this work was performed with monomorphic bloodstream form cells and so unable to differentiate to true stumpy forms. Consequently, it remained important to assess CAT gene expression from the same constructs in true slender and stumpy form cells to determine whether the regulatory element is also responsible for the control of gene expression in these life-cycle stages. Therefore, to further investigate this, four of the constructs were selected for transfection into pleomorphic trypanosomes (the AnTat 1.1 90:13 strain; see Section 2.2.4). These being the full-length ESAG9-EQ 3'UTR, element deletion (eΔ), and element insertion (eI) constructs, with two independently derived cell lines being obtained for each construct. Additionally, two independently derived cell lines of the unmodified

vector were used as a control (these two cell lines were generated by Dr. Paula MacGregor, University of Edinburgh).

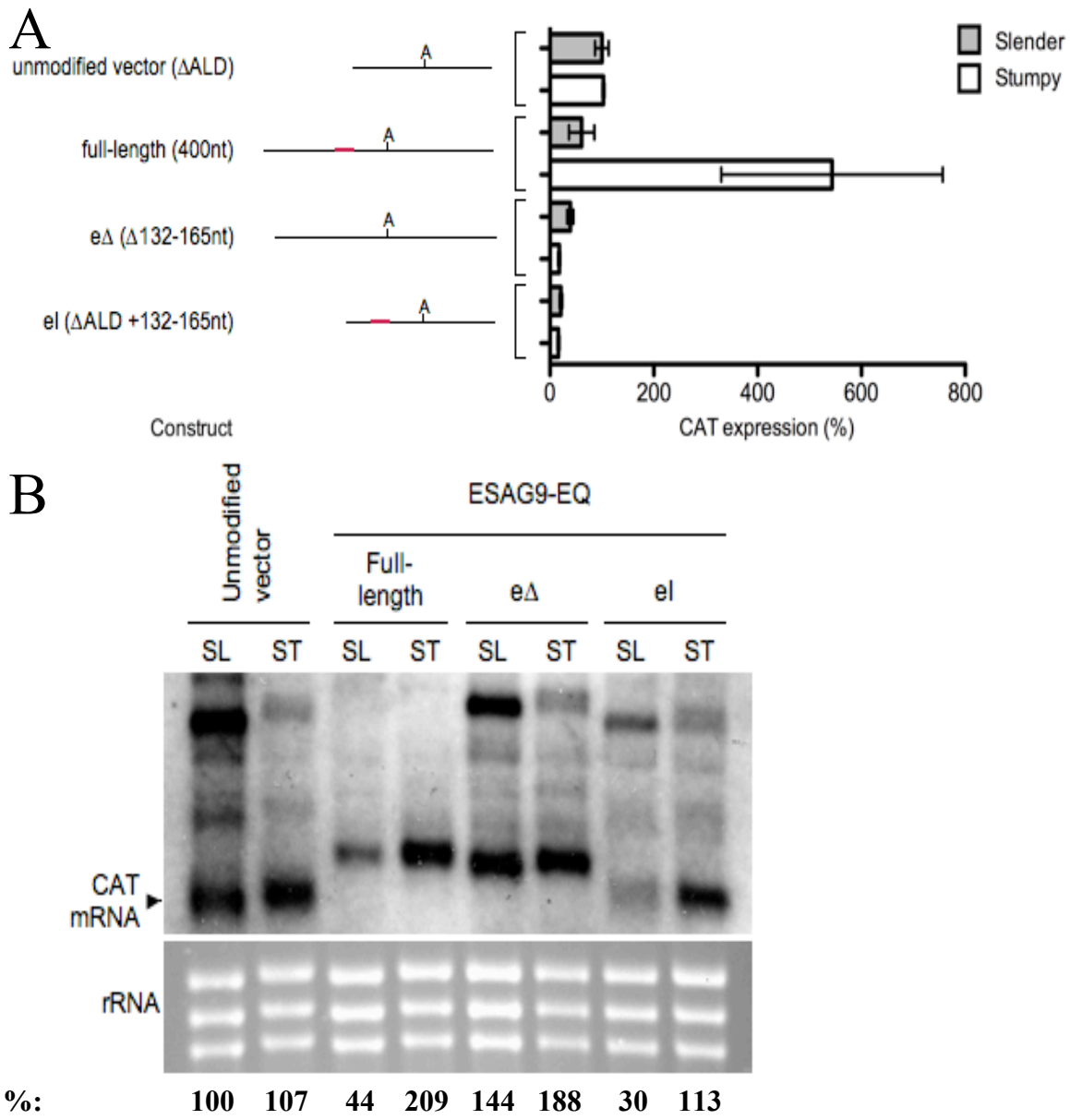


Figure 3.17 Analysis of the ESAG9-EQ 3'UTR and associated element in true slender and stumpy forms using a CAT reporter gene assay. A. CAT protein expression from the constructs in slender form (grey bars) and stumpy form (white bars) cells, shown as a percentage of expression from cell lines of the unmodified vector slender form cells. The data shown is the mean ± standard error of two biological replicates (cell lines) for each construct. Therefore, because $n = 2$ for each, the standard error bars represent the data range. To the left of the graph are schematic representations of the sequence analysed in each construct, with the polyadenylation site location indicated with 'A' and regulatory element location indicated by the red line. B. Northern blot analysis using a digoxygenin-labelled RNA probe against the

CAT coding region of RNA samples from one cell line of each of the unmodified vector, full-length ESAG9-EQ 3'UTR, element deletion (eΔ) and element insertion (eI) constructs in slender (SL) and stumpy (ST) forms. Shown in the lower panel is the ethidium bromide-staining of the rRNA present to indicate the loading of each lane. Shown below are the mRNA quantifications following normalisation to rRNA, as a percentage of that obtained for the unmodified vector.

The level of CAT protein expression obtained in true slender and stumpy forms for each construct are shown in Figure 3.17 (A). For the unmodified vector, there was no difference in the level of CAT protein expression observed upon differentiation to the stumpy form. In contrast, there was a very large increase in CAT protein expression for the full-length ESAG9-EQ 3'UTR construct in response to differentiation to the stumpy form. This was not observed for the element deletion construct upon stumpy form differentiation, therefore, demonstrating that it was the element sequence that was entirely responsible for the observed response to differentiation of the full-length construct. However, this element was not successful in functioning out of context in the element insertion construct, because no increase in CAT protein expression was found upon stumpy form differentiation in these cells.

The *CAT* mRNA expression levels in slender and stumpy form cells for one of the cell lines obtained for each construct are shown in Figure 3.17 (B). For the unmodified vector, there was little change in the level of *CAT* mRNA upon stumpy form differentiation, similar to the CAT protein levels obtained for this construct (see Figure 3.17 A). Again, in contrast to this, the *CAT* mRNA level increases upon differentiation for the full-length ESAG9-EQ 3'UTR construct. For the element deletion construct, the level of *CAT* mRNA was similar in slender and stumpy forms, with both having a high level of expression. Like the full-length ESAG9-EQ 3'UTR, the *CAT* mRNA level increased in stumpy form differentiation with the element insertion construct. However, this was not reflected by an increased level of CAT protein in stumpy form cells for this construct (see Figure 3.17 A). The possible reasons for this inability of the element sequence to function out of context are discussed in Section 3.5.2.

Also observable in Figure 3.14 (B) are differences in the abundance of *CAT* mRNA processing intermediates. The greatest molecular weight band appears to be more abundant in slender forms than in stumpy forms, except for the full-length 3'UTR construct, where the higher molecular weight band is not visible. The reasons for these observations are not clear,

The data presented in Figure 3.17 confirm that the full-length ESAG9-EQ 3'UTR is important for the control of gene expression upon development to the stumpy form, and that the identified element sequence was responsible for this response. Comparison of the *CAT* mRNA levels in slender forms for the full-length and eI constructs with the eΔ construct revealed that the presence of the element sequence appeared to result in the destabilisation of the *CAT* mRNA in slender forms (Figure 3.17 B). However, comparison of the CAT protein expression levels for the full-length construct with the eΔ construct demonstrated that the presence of the element sequence was required for increased translation in stumpy forms. Therefore, in slender forms the identified element sequence appeared to destabilise the mRNA, repressing gene expression, and in stumpy forms this sequence increased translation, thereby releasing the repression on gene expression.

3.10 Is the ESAG9-EQ 3'UTR element found in other ESAG9 3'UTRs?

Because all ESAG9 genes investigated have demonstrated the same stumpy form transcript enrichment (Barnwell *et al*., 2010), the conservation/presence of a similar predicted structure to the element identified here for ESAG9-EQ, in the 3'UTRs of other ESAG9 genes was investigated by a bioinformatic study.

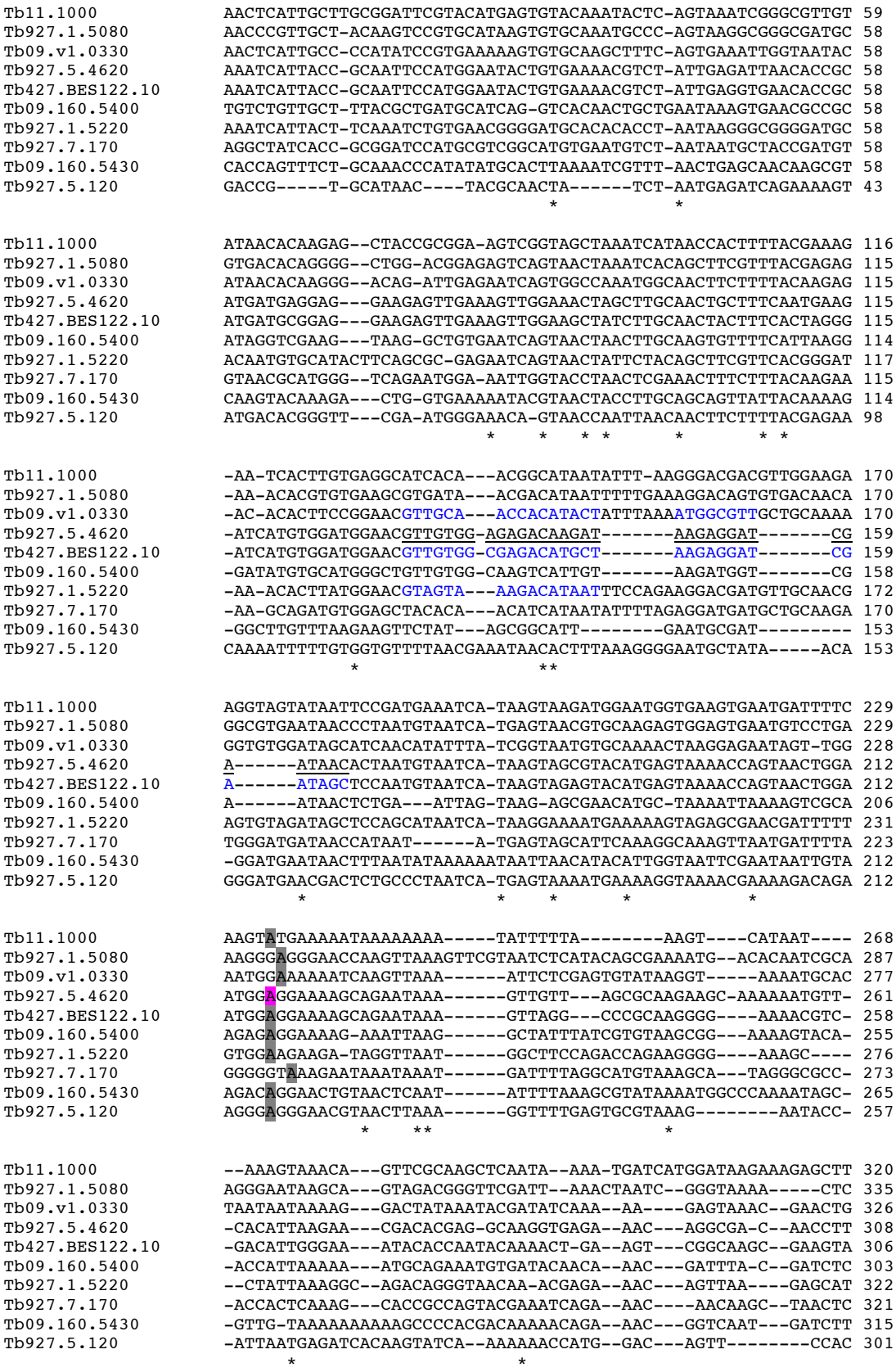


Figure 3.18 continues on the following page

```
Tb11.1000           AACTTTAATGGTATTAAGGAAGCGGTGATTTATGGAGTT-----TTTAAAAGTGCAG-TG 374
Tb927.1.5080        AGGTTCAATACTATTGGGAAGTTTGTGGTTGATGAACTT-----TTCAAAAGTGTAA-TG 389
Tb09.v1.0330        AAA---AATGTTACTGCAAA---TGAGGTAAGAAAATTTAA--CTTCATTGGTATTA-AG 377
Tb927.5.4620        AAC---AATATTATTATTAA---TTAAACAAGGAAATTTAA--CTTCAAAAGTATTA-AG 359
Tb427.BES122.10     AAA---AATGCGGTCATGAA---CGCAACCGAGCAACTTAC--TTTCAAGAGTATTG-TG 357
Tb09.160.5400       AAA---TAATTGATTTTAAA---TAGGAAAAGGAAATATGA--GGTTAATACTATCA-AA 354
Tb927.1.5220        CAC---AAC---ATCTTTGA---TGAGGCAGGGAAAATAAA--ATTCGATTGTGTTG-AA 370
Tb927.7.170         AAGAAAAATGTGATAGTAAA---TGGGTTAAGGAAACCTAA--CTTCAGTGGTACTG-GG 375
Tb09.160.5430       ACG---AATGTTTTTATTAA----------AATGAACTTAA--CCTTAATGGAACTG-AG 359
Tb927.5.120         AGG---ACCGTTGATAAGGC----------AAAGAATCTAAAATTTTAACGGTGTTGGAG 348
                                                   *        *

Tb11.1000           AAAGATGCCGTGTGGCGAAAGTTAGAGCCCT--------------CGAAAAG-TGC--C- 416
Tb927.1.5080        GACGTTGCTGCATATCAAAAGTTAACCTTAT--------------TGCGAGGATGC--T- 432
Tb09.v1.0330        AAAG-TAACACAC------AGTTGAATTTTT--------------CAAAAATGTAA--TA 414
Tb927.5.4620        GACATCGTAGCATTGAA----TTTTTTTTTT--------AG------------------- 388
Tb427.BES122.10     AAGACAGTAGAATTTAAAATATTTTTTCTTT--------AGT---TAAATGGACAAAGTA 406
Tb09.160.5400       AA-ACCGCAATGAGGAG-------TTTCTTT--------AA-----AGATGTATAA--TA 391
Tb927.1.5220        GAGTCCAT----TCCCG-----TTTTTCTTTTGCATTCAAAA---AAAAAACATAA--TA 416
Tb927.7.170         GAAGGTGTAAC----CGAAGATTTTTTTTTTTA----ATGTA---AGGAAATTTGCTGCA 424
Tb09.160.5430       AAATCCGCATCACATGAAGTATTC-------------AAAAA--AAAAAGGACCACTTTG 404
Tb927.5.120         AGATCTTCTTCA-GTGAACTATTTTTTTTTGCGTTCTAAAAATTATAAAGAGCTGCCCCG 407


Tb11.1000           ----------GAAGGCA-----------------------GTCGTAAAGATGAA---CTT 440
Tb927.1.5080        ----------GATGGCA-----------------------GTCGTGAGGGCAAG---CTG 456
Tb09.v1.0330        ----------GAAGTTA-----------------------GT-GTGAGGATTAG---CTT 437
Tb927.5.4620        ---------------------------------GG--AACGTGATAA------------- 400
Tb427.BES122.10     ----------AGAACCACGTTGCATGCAGGGGGGA--AACTTGATACCAGAGAG---ACT 451
Tb09.160.5400       ----------AGAGTCACTGAACACGCAAG---GGTTAACATACTGCAGAGCGG---CTT 435
Tb927.1.5220        ----------GGAATCGCTTATCACTTGGG---GG---ATACATTATGAAGAGA---GTA 457
Tb927.7.170         TATCC-----AAAGTCAGCG--------------------CTTTTGCAAAGGGA---GTG 456
Tb09.160.5430       GTTCCGAATCGGAAAC-CGTCACAGCGAAGAT--GTAAATGGGTTAAGGAGGAGGGTCTT 461
Tb927.5.120         -TGCCTAGAAGAAGGCACATTTCTAAGAGAGC--GAAAATGGTTTAA--AGCAGGA-CTT 461
                                                    *

Tb11.1000           AACGGAATCGTTAATTTAAATTTTGGTGACGAT-ATTAATCGTTTTGTTAACAATACATA 499
Tb927.1.5080        GATGAAATCGTAAATTTTAATTGTGGGGA------------ATTTAG---AGACCACCT- 500
Tb09.v1.0330        AATAGCATTGCAAATTTAAATCTTTCTAAAGAA-GCTAACTATCTCTTTAAGCATGTGTG 496
Tb927.5.4620        ------------------------------------------------------------
Tb427.BES122.10     AACGAAATCGTAGCGTCACGTTATGGGGGGGGGCACTGAATACTTTGAC----------- 500
Tb09.160.5400       AACGGAAACGTAGGCTTGCGCTGCGTGGAAGAG-AGTGATTGTTTCAGCAGTCAAATGTG 494
Tb927.1.5220        AATGTAGGTTACGAAGAATACTATGGAGATCACAAGCTAATAT----------------- 500
Tb927.7.170         AGGATAGTCGTAAAGGTAAGCTTAACGGG-----ATCGTAAATTTAAAT----------- 500
Tb09.160.5430       AGTGAAATTTTAAATTTAAATTGTGGGGAAGAA-ACTGAG-------------------- 500
Tb927.5.120         ATTTAGATCATTAACTTAAA---TGAGGAAAAC-GCAGACGAT----------------- 500


Tb11.1000           A----- 500
Tb927.1.5080        ------
Tb09.v1.0330        AAAT-- 500
Tb927.5.4620        ------
Tb427.BES122.10     ------
Tb09.160.5400       AGATTA 500
Tb927.1.5220        ------
Tb927.7.170         ------
Tb09.160.5430       ------
Tb927.5.120         ------
```

Figure 3.18 ClustalW2 alignment of sequences downstream of ESAG9 genes. For each ESAG9 gene, except Tb927.5.4620 (ESAG9-EQ) where 400nt was used, 500nt of sequence downstream of the coding region stop codon was aligned using the ClustalW2 Multiple Sequence Alignment tool (http://www.ebi.ac.uk/Tools/msa/clustalw2/) for DNA sequences and using the default parameters. Asterisks indicate positions of a single, fully conserved residue; pink highlight: mapped polyadenylation site (see Section 3.3.1); grey highlight: nearest adenosine nucleotide to mapped polyadenylation site for ESAG9-EQ in the alignment; the ESAG9-EQ element sequence is underlined; blue font: residues marked with asterisks in Figure 3.19.

Table 3.1 ClustalW2 scores from the alignment of sequences downstream of ESAG9 genes shown in Figure 3.18.

	Tb11.1000	Tb927.7.170	Tb927.1.5080	Tb09.160.5430	Tb927.5.4620	Tb09.v1.0330	Tb927.1.5220	Tb927.5.120	Tb09.160.5400	Tb427.BES122.10
Tb11.1000		61	63	53	59	59	56	53	54	56
Tb927.7.170			58	55	64	56	56	58	56	56
Tb927.1.5080				56	60	60	58	58	53	55
Tb09.160.5430					61	53	46	60	56	54
Tb927.5.4620						62	60	54	66	80
Tb09.v1.0330							56	54	54	57
Tb927.1.5220								58	56	58
Tb927.5.120									53	51
Tb09.160.5400										59
Tb427.BES122.10										

3.10.1 Sequence similarity for sequences downstream of ESAG9 genes

Firstly, to discover how similar the downstream sequence of each identified ESAG9 gene might be, these sequences were aligned to find regions of sequence conservation. Two genes that are annotated as ESAG9 pseudogenes (Tb927.3.5790 and Tb927.5.100) were excluded from this analysis, because being pseudogenes they may not be functional and so not subject to the same regulation. In each case, 500nt of sequence downstream of the stop codon for the coding region of each ESAG9 gene was downloaded using the Sequence Retrieval tool in the TriTryp database (http://tritrypdb.org/tritrypdb/), except for Tb927.5.4620 (ESAG9-EQ), where 400nt was used. These sequences were used as input in the ClustalW2 Multiple Sequence Alignment tool (http://www.ebi.ac.uk/Tools/msa/clustalw2/; see Section 2.1.1). The output is shown in Figure 3.18 and Table 3.1.

As can be seen in Figure 3.18 the sequences show a limited degree of similarity, there being few residues that were fully conserved between the downstream sequences (marked with asterisks in Figure 3.18), although some degree of sequence similarity can be observed between the sequences in the alignment. The most similar sequences to each other were ESAG9-EQ and Tb427.BES122.10, which had a score of 80 (see Table 3.1). The sequences least similar to each other were Tb09.160.5430 and Tb927.1.5220, with a score of 46 (Table 3.1). The sequence of the ESAG9-EQ element is shown underlined in Figure 3.18. The residues of this sequence did appear to show some conservation again, but it was not highly conserved or more conserved than other regions of sequence within the downstream sequences. It was however well conserved in Tb427.BES122.10, which is not surprising as ESAG9-EQ and Tb427.BES122.10 were the sequences most similar to each other, and also Tb09.160.5400.

3.10.2 Investigation of predicted RNA structure similarity

Although there was not a high degree of conservation of the sequence of the ESAG9-EQ element, it was possible that this region of each 3'UTR may still be able to form a similar structure. Therefore, the presence of a similar structure to that in the ESAG9-EQ 3'UTR was investigated in each ESAG9 downstream sequence using *S*fold. Because the downstream sequences used above were relatively long, the PAS location was probably within each 500nt sequence. However, because the presence of extra sequence that is not part of the actual 3'UTR could affect the RNA structure predicted, each of the 500nt downstream sequences were also shortened to an adenosine residue that may represent the PAS location (based on its location in the ESAG9-EQ 3'UTR), in an attempt to improve accuracy.

In order to achieve this the mapped PAS for ESAG9-EQ was used, which is shown in pink highlight on Figure 3.18. For each ESAG9 downstream sequence, the nearest adenosine nucleotide in the alignment to the ESAG9-EQ PAS was used to truncate the sequence to this residue. The residue chosen for each sequence is shown in Figure 3.18 in grey highlight. These truncated sequences were used as input for the *S*fold programme, *S*rna, as was used in the Sections above.

Each of the centroid structures returned as output for each sequence were analysed for any structure that resembled the predicted ESAG9-EQ 3'UTR element structure. Three of the ESAG9 gene downstream sequences analysed possessed a structure within the predicted secondary structure that had resemblance and also contained sequence that aligned with the element by ClustalW2 alignment (see Figure 3.18). These genes were Tb427.BES.122.10, Tb09.v1.0330, and Tb927.1.5220.

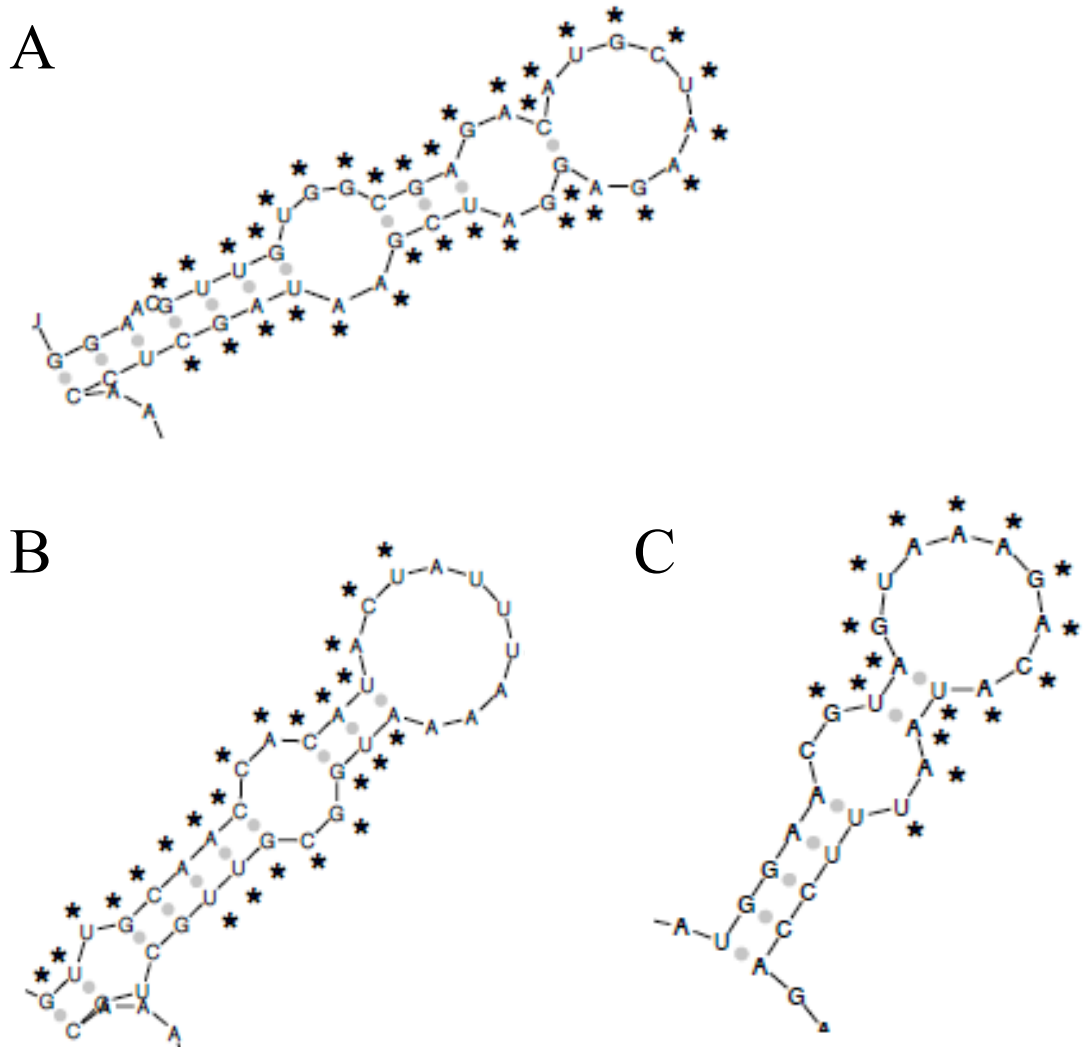


Figure 3.19 Predicted RNA structures of sequences downstream of ESAG9 genes bearing resemblance to that of the ESAG9-EQ element. Section of secondary RNA structure predicted by the *S*fold programme, *S*rna, for Tb427.BES122.10 (A), Tb09.v1.0330 (B), and Tb927.1.5220 (C). Asterisks indicate residues that aligned with the ESAG9-EQ element sequence by ClustalW2 alignment (see Figure 3.18).

The most similar structure to that of the ESAG9-EQ element was found in the centroid of cluster 1 predicted folding for Tb427.BES.122.10, which is shown in Figure 3.19 (A). Comparison to the predicted structure of the element (see Figure 3.7 E) reveals that the folding is almost identical, the differences being that there is an extra base pair (G and C) in between two of the loops in the Tb427.BES.122.10 structure and also two unpaired nucleotides lower down the stem. Tb427.BES.122.10 is the ESAG9 gene that had the highest similarity in downstream sequence to ESAG9-EQ, and the greatest similarity in the region of the element sequence (see Section 3.9.1). The residues that aligned in the region of the element sequence by ClustalW2 alignment are indicated in Figure 3.19 by asterisks. As can be seen, for the Tb427.BES.122.10 structure, the residues that make up this structure are the residues that aligned to the ESAG9-EQ element (Figure 3.19 A).

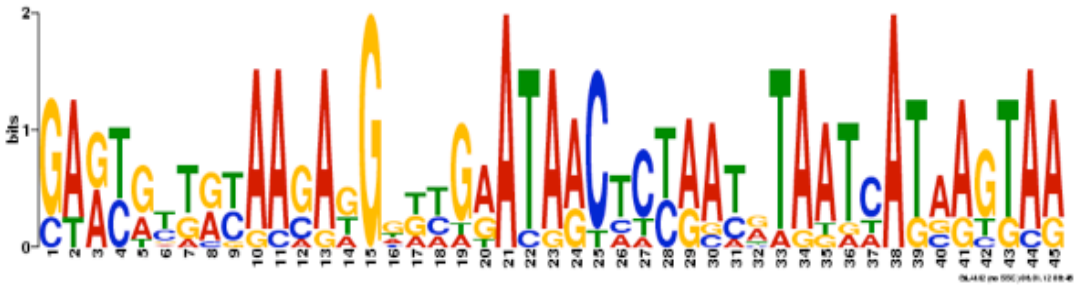
Tb09.v1.0330 and Tb927.1.5220 also had stem-loop structures that were predicted to form with residues that aligned to the ESAG9-EQ element, but were much less similar in their predicted folding (Figure 3.19 B and C). For Tb09.v1.0330, the structure found was in the centroid of cluster 2 and also the mfold predicted structures; for Tb927.1.5220, the structure found was in the centroid of cluster 6 predicted structure. The aligned residues that are indicated with asterisks in Figure 3.19 are also indicated in the alignment in Figure 3.18 in blue font. For both, the residues that aligned were present at the beginning of the EQ element sequence and for Tb927.1.5220 this did not correspond to the start of the predicted structure.

So, while it appears that the predicted structure found in Tb427.BES122.10 sequence is very similar to ESAG9-EQ, the structures found in Tb09.v1.0330 and Tb927.1.5220 could be only similar by coincidence. This is probably more likely for Tb927.1.5220, due to the residues of the alignment corresponding to a different region of the predicted structure than the ESAG9-EQ element predicted structure.

3.10.3 Other bioinfomatic analyses

To determine if there was conservation of the element sequence in the rest of the sequences outside of the region aligned with the element, the RNA Regulatory Sequence Tool (RSAT) DNA-pattern matching tool was used (see Section 2.1.5). The sequence of the element was used as the query pattern and the same 500nt of sequences downstream of the ESAG9 genes as sequences for analysis (but, again, a 400nt region was used for ESAG9-EQ). This resulted in only the element sequence in the ESAG9-EQ sequence being identified. The analysis was repeated with several variations on the query pattern, such as shortening the sequence at both ends and allowing purine/pyrimidine residues to be matched in the identification. Some searches resulted in identification of part of the region of Tb427.BES122.10 that aligned with the element in Figure 3.18 and therefore part of the structure predicted in Figure 3.19 (A), with a score of 0.88.

The MEME suite (http://meme.sdsc.edu/meme/intro.html) is a collection of tools that can be used for motif-based sequence analysis, of which GLAM2 is one. GLAM2 (http://meme.sdsc.edu/meme/cgi-bin/glam2.cgi) is a tool that can be used for identification of motifs in groups of related DNA sequences, using gapped alignments. So this was also used to determine if the element sequence would be identified as a conserved motif in the downstream ESAG9 sequences allowing for any insertions or deletions that may be present. The same downstream sequences of ESAG9 genes, truncated to the possible polyadenylation sites described in Section 3.9.2, were used as input with the default settings. This resulted in the identification of a number of different conserved sequences, one of which consisted of part of the element sequence. The output for this sequence is shown in Figure 3.20. However, this was not a particularly high scoring motif (score: 197.696; the sixth highest), and so the significance of this result is unclear.



	Start		End	Strand	Marginal score
		**********.*********			
Tb927.1.5080	158	cagtgtgacaacaggcgtgaataaccctaatgtaat.catgagtaa	202	+	39.5
Tb09.160.543	141	ca.ttgaatgcgatggatgaataactttaatataaaaaataattaa	185	+	10.1
Tb09.v1.0330	159	gt.tgctgcaaaaggtgtggatagcatcaacatatt.tatcggtaa	202	+	26.1
Tb927.1.5220	161	ga.tgttgcaacgagtgtagatagctccagcataat.cataaggaa	204	+	29.8
Tb927.5.120	141	gaatgctataacagggatgaacgactctgccctaat.catgagtaa	185	+	28.0
Tb427.BES122	142	ga.catgctaagaggatcgaatagctccaatgtaat.cataagtag	185	+	40.8
Tb11.1000	159	ga.cgttggaagaaggtagtataattccgatgaaat.cataagtaa	202	+	32.2
Tb927.7.170	159	ga.tgctgcaagatgggatgataaccataatatgag.tagcattca	202	+	17.7
Tb09.160.540	141	gt.cattgtaagatggtcgaataactctgaattagt.aagagcgaa	184	+	22.6
Tb927.5.4620	142	ga.caagataagaggatcgaataacactaatgtaat.cataagtag	185	+	40.5

Figure 3.20 Conservation of part of the element sequence in ESAG9 downstream sequences identified using GLAM2. Shown is the motif identified that consists of part of the ESAG9-EQ (Tb927.5.4620) element sequence. The upper graph shows aligned letters, the height of which refers to the frequency of the residue in that motif position in the sequences. Shown below is the motif in each ESAG9 downstream sequence, with the start and end nucleotide number, the strand the motif was found on, and a marginal score (a greater marginal score indicates better matches to the motif). Deletions in the alignments are shown with black dots. The part of the element sequence found in the motif is underlined.

The bioinformatic analyses described above indicate that there is some conservation of the sequence, and to a lesser extent the predicted structure, of the ESAG9-EQ element within downstream sequences for other ESAG9 genes, but that this conservation is quite low. However, the lack of conservation of the predicted element structure may also be due to inaccurate folding predictions due to the unknown PAS locations for the other ESAG9 genes. The low conservation is surprising because the ESAG9 family of genes show the same developmental regulation, so it would be expected that they would share the same regulatory signals.

3.11 Discussion

Previous studies to identify *cis*-acting sequences important to the control of *T. brucei* gene expression have focussed on the experimentally tractable, proliferative life-cycle stages; these being the slender bloodstream and procyclic form stages (see Section 1.7). Using two genes previously identified as upregulated in stumpy forms, aconitase (TbACO; Saas *et al*., 2000) and ESAG9-EQ (Barnwell *et al*., 2010), the work presented in this chapter aimed to investigate if the 5' or 3'UTR belonging to each of these genes was responsible for the control of their expression in stumpy forms.

3.11.1 Aconitase 5' and 3'UTRs

Regulation of the expression of the TbACO gene is believed to operate at the level of translation or post-translation (Saas *et al*., 2000). This is because the level of *TbACO* mRNA remains relatively constant throughout the lifecycle, however, the TbACO protein is enriched in stumpy forms compared to slender forms (Saas *et al*., 2000). Interestingly, in CAT reporter gene assays, the TbACO 3'UTR conferred gene repression in slender forms at the protein level, but not at the mRNA level (Figure 3.3). In contrast, the TbACO 5'UTR only resulted in a small reduction in the level of CAT protein (Figure 3.3 B). This, therefore, suggested that the TbACO 3'UTR is responsible for the repression of TbACO protein expression in slender forms.

Unfortunately, it was subsequently discovered that the TbACO 3'UTR sequence analysed in the CAT reporter assay did not possess either of the PASs identified for TbACO, these being located at positions +516 and +843 relative to the TbACO stop codon (Siegel *et al*., 2010). Thus, it is unclear whether the repression of gene expression observed for this construct was due to the presence of regulatory signals or the absence of a PAS. It would, therefore, be important to examine in the CAT reporter assay the TbACO 3'UTR sequence possessing both PASs, i.e. through examination of 850nt of downstream sequence.

The programme, *S*fold, was used to predict the secondary RNA structure for the 5' and 3'UTR sequences analysed and there was a structure observed in the prediction for the 3'UTR that exhibited some resemblance to an IRE. This is particularly interesting because usually an IRE situated in a 3'UTR prevents mRNA degradation, and, consequently, promotes translation (reviewed in Volz, 2008). Therefore, if this structure were responsible for the repression of gene expression in slender forms, this would be unusual. It was not investigated here whether the 3'UTR is responsible for the alleviation of gene repression in stumpy forms, but if so, then it is possible that this IRE-like structure could be responsible.

Also interestingly, TbACO accounts for the total aconitase activity in *T. brucei* and shows mitochondrial and cytoplasmic localisation (Saas *et al*., 2000). This contrasts with mammalian cells, which possess two aconitase enzymes, one mitochondrial and one cytoplasmic, which is also known as IRP1 (reviewed in Volz, 2008),. TbACO exhibits a high degree of conservation of the IRP RNA binding sequence, and therefore, it is possible that TbACO has a gene regulatory function (Saas *et al*., 2000). If the IRE-like structure does contribute to the regulation of TbACO gene expression, then this could represent a self-regulatory mechanism. Therefore, if further analysis of the TbACO 3'UTR was made using CAT reporter gene assays, the IRE-like sequence would be particularly interesting to investigate.

3.11.2 ESAG9-EQ 3'UTR

The ESAG9 gene family mRNAs are strongly stumpy-enriched (Barnwell *et al*., 2010). The 3'UTR of one member of this family, ESAG9-EQ, was investigated for its contribution to the control of gene expression using CAT reporter assays. The 'full length' 3'UTR was found to confer reporter gene repression in slender forms at the levels of mRNA and protein (Figure 3.6). However, the reporter gene expression was not completely repressed in slender forms, as there was some *CAT* mRNA and protein present in these cells (Figure 3.6). Although, the reason for this is not clear, it could be that there are additional sequences required for complete repression that were not present in the sequence analysed. Alternatively, it could be due to the analysis being performed in monomorphic cells, which do not represent true slender

forms because of their culture adaptation. However, following transfection into pleomorphic cells, analysis of the full-length 3'UTR construct in true slender cells also found that there was not complete repression of the reporter gene expression (Figure 3.17 A). Although the level of CAT protein expression in slender cells with this construct was somewhat reduced compared to slender cells with the unmodified vector (Figure 3.17 A). Nonetheless, it does appear that the ESAG9-EQ 3'UTR makes a significant contribution to gene repression in slender forms.

Initially, from the 3'UTR deletion series it appeared that the presence of the predicted structure of the element sequence (Figure 3.7) was responsible for gene silencing, as this correlated with gene repression (Figure 3.6). However, after analysing additional clones of the deletion series, it was apparent that the original 350nt construct clone produced an aberrant result and that this initial conclusion was incorrect (Figure 3.13). Nonetheless, from the new data (Figure 3.13 A), it appeared that the 33nt element sequence could still be a significant contributor to slender form repression, as the sequence is present in the full-length, 350nt, and 300nt constructs which show repression and only 15nt of the sequence is present in the 250nt construct which shows alleviation of repression. This was supported by the reporter gene expression observed for the element deletion (eΔ) construct, which exhibited alleviation of gene repression (Figure 3.13). However, a greater alleviation of gene repression was obtained with the 250nt construct than the eΔ construct, indicating that although the element sequence was a significant contributor to gene repression in slender forms, it was not entirely responsible. Additionally, analysis of reporter gene expression in true slender and stumpy forms later demonstrated that although the element sequence appeared to result in mRNA destabilisation in slender forms, deletion of this sequence did not result in increased protein levels (Figure 3.17).

It is possible that, as the structures obtained for the 3'UTR sequences used in the constructs are only predictions, the element structure does not actually form *in vivo*. If the two loops that comprise the top of the element structure do not form, they would create one larger loop, which was seen in the 350nt *S*fold ensemble centroid structure for this region. Alternatively, it is possible that the element structure does

form in the 350nt 3'UTR sequence because, although it was not predicted in the ensemble centroid structure for this sequence, it was predicted to form in one of the two *S*fold cluster centroid structures (data not shown). Either of these possibilities could explain why there was repression of gene expression with the 350nt construct when the element structure was not predicted to form. It would therefore be interesting to determine the structure of these 3'UTR sequences experimentally by, for example, RNase enzymatic digestion, to establish whether it is the element sequence or structure that is important for the control of gene expression.

Following chemical induction of development to stumpy-like forms, it appeared that the full-length 3'UTR was also responsible for alleviation of repression in these cells (Figure 3.14). It also appeared that the presence of the element sequence was responsible for this, because the alleviation was observed with the 350nt and 300nt constructs which possess this sequence as well, but alleviation was lost with the 250nt construct, which only possesses 15nt of the element sequence, and with the eΔ construct (Figure 3.14). This therefore, also indicates that it is the element sequence that is responsible, because the element structure was not predicted to form in the 350nt sequence (Figure 3.7). However, as noted above, it is possible that this structure does not form *in vivo* in the other sequences, or alternatively, that it may be present in the 350nt sequence as well. The element sequence, therefore, represents the first identification of an mRNA sequence responsive to a chemical signal believed to intersect with the stumpy-induction signalling pathway. Thus, it was essential to next determine the importance of this regulatory sequence in true slender and stumpy forms, and so, in response to density-dependent development and SIF signalling, following pleomorphic transfection of these constructs.

From analysis of reporter gene expression in stumpy-like forms it appeared that mRNA control was strongly contributing to increased expression in this development (Figure 3.14), in contrast to the results obtained for true slender and stumpy forms (discussed below). The reasons for this are unclear but could be attributable to differences in the cell types analysed, i.e. stumpy-like vs stumpy forms, given the incompleteness of stumpy form differentiation in the stumpy-like cells. Regardless,

the true slender and stumpy forms represent the most biologically relevant analysis and hence provide the most meaningful results.

Analysis of a selection of the CAT reporter gene constructs in true slender and stumpy forms showed that the full-length ESAG9-EQ 3'UTR sequence was important for the control of gene expression upon development to the stumpy form, and that the element sequence was responsible for this response (Figure 3.17 A). Therefore, this element sequence also represents the first identification of an mRNA sequence responsive to density-dependent development and SIF signalling. Also shown in this analysis in true slender and stumpy forms was that the element sequence appeared to result in mRNA destabilisation in slender forms (Figure 3.17 B), but that the presence of this sequence was required for increased translation in stumpy forms (Figure 3.17 A). These data lead to the proposal of the model shown in Figure 3.21 to describe the regulation of ESAG9-EQ expression in slender and stumpy forms. However, it remains unclear if the element sequence or predicted structure is important for this regulation.

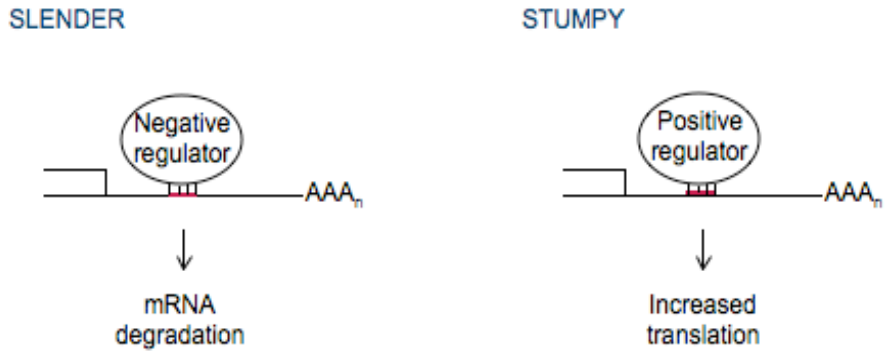


Figure 3.21 Proposed model for the mechanisms of regulation of ESAG9-EQ expression in slender and stumpy forms. In slender forms (left hand side), a negative regulator binds to the element sequence (depicted by the red bar) in the ESAG9-EQ 3'UTR and causes mRNA degradation of the transcript, repressing the expression of ESAG9-EQ. In stumpy forms (right hand side), a positive regulator binds to the element sequence and increases translation of the transcript, resulting in the release of gene expression repression in this life-cycle stage. AAA_n: poly(A) tail.

Nonetheless, whether the element sequence forms the predicted structure or whether the two loops comprising the top of the predicted structure do not form, it is possible that RNA-binding proteins bind to this sequence. In slender forms, this would result in repression of gene expression and in stumpy forms this would promote gene expression. Hence, it would be particularly interesting to identify any interacting proteins, specifically any that promote gene expression in stumpy forms. In particular, because the protein is responsive to SIF signalling in true stumpy forms this identification could provide a way to dissect the SIF signalling pathway from one of the end-points of the pathway and so from the opposite direction to signal transduction.

A number of different approaches could be employed to identify any proteins interacting with the element sequence. For example, one approach would be to use a method similar to the RaPID (RNA-binding protein purification and identification) technique, whereby the bacteriophage MS2 coat protein, fused to green fluorescent protein (GFP) and streptavidin-binding protein (SBP), binds to an experimental mRNA engineered to contain MS2 loops. Thus, mRNA-protein complex purification can be achieved using streptavidin-conjugated beads (Slobodin and Gerst, 2010). Subsequently, identification of interacting proteins is made using mass-spectrometry. A second approach would be to use a yeast three-hybrid (Y3H) system, which identifies RNA-protein interactions via binding of an RNA to two hybrid proteins resulting in activation of reporter gene transcription (SenGupta *et al.*, 1996). Additionally, another approach would be to use a recently described method consisting of attaching the experimental RNA sequence to a tRNA scaffold fused to a Streptavidin aptamer, and identifying bound proteins by mass spectrometry (Iioka *et al.*, 2011).

It was also investigated if the element sequence and predicted structure was found in other ESAG9 3'UTRs. The predicted structure was not detected by the methodology employed, except for in the 3'UTR of a very closely related ESAG9, Tb427.BES122.10 (Figure 3.19). However, the element sequence was more highly conserved among the ESAG9 3'UTRs (Figure 3.18) than the predicted element

structure. This could be important if the predicted element structure does not fold as predicted, where the sequence of the middle region of the element could be more critical.

Chapter 4

Results

Investigation of upstream open reading frames and different *trans*-splicing site usage in stumpy form enriched genes

In many eukaryotes, upstream open reading frames (uORFs) are used as a mechanism to regulate gene expression. For certain transcripts, such as *GCN4* in *Saccharomyces cerevisiae*, uORFs can be used to escape the global translational repression imposed by phosphorylation of eIF2α (reviewed in (Hinnebusch, 1997). An uORF is a short ORF upstream from the main ORF, which can regulate translation of the main ORF either by causing translational repression through ribosome stalling or by enhancing translation through promoting ribosome reinitiation on the main ORF start codon. Additionally, uORFs have been found to cause destabilisation of the mRNA by ribosome termination (reviewed in (Vilela and McCarthy, 2003). Regulation of gene expression by an uORF has been demonstrated in another protozoan parasite, *Plasmodium falciparum*. The gene encoding one variant of the variant surface antigen, VAR2CSA, which is implicated in placental malaria, has a small uORF that represses its translation (Amulic *et al.*, 2009).

The aim of this part of the project was to assess if uORFs are used by *T. brucei* as a mechanism for controlling the expression of cohorts of co-regulated genes in stumpy forms. Namely, are uORFs used either as a mechanism for the escape from translational repression of stumpy form enriched genes, or as a mechanism of translational repression for repressed genes in stumpy forms.

4.1 Genome-wide bioinformatic prediction of upstream open reading frames in *T. brucei*

To investigate if uORFs are used as a mechanism to control the expression of stumpy-enriched transcripts, the presence of potential uORFs was first assessed globally in *T. brucei* genes by a bioinformatic genome-wide screen. This screen was used to predict the abundance of uORFs in sequences upstream of every annotated gene in *T. brucei*. The dataset generated from this analysis could then be used in cross-reference to genes identified as stumpy-enriched to identify if there was an overrepresentation of genes with potential uORFs expressed in stumpy forms. If so, then uORFs may represent a mechanism for gene expression control in stumpy forms.

The bioinformatic uORF genome-wide screen firstly involved retrieval of DNA sequence upstream of each annotated gene in the *T. b. brucei* TREU 927 genome (http://www.genedb.org/Homepage/Tbruceibrucei927), using the Sequence Retrieval tool (now available through http://tritrypdb.org/tritrypdb/). Then, the amino acid sequence was deduced for each sequence by translation of each of the three forward reading frames using an online tool available at http://searchlauncher.bcm.tmc.edu/seq-util/Options/sixframe.html. A Perl script (written and executed by Pankaj Barua, University of Edinburgh) was then used with the deduced amino acid sequences to identify all incidences of a methionine followed by a stop codon with 2 or more amino acids in between, for each amino acid sequence. These identified sequences may therefore represent an uORF.

Table 4.1. Percentage of genes found to possess potential uORFs in the genome-wide and other dataset analyses by bioinformatic prediction.

Dataset	**Criteria** [a]	**Percentage** [b]
T. brucei TREU 927 genome [c]	100nt upstream sequence	25.38
	within predicted 5'UTR	nd
Stumpy-form upregulated genes [d]	100nt upstream sequence	30 (12/40)
	within predicted 5'UTR	20 (8/40)
Heat-shock regulated genes [e]	100nt upstream sequence	19.67 (12/61)
	within predicted 5'UTR	11.48 (7/61)

[a] The prediction was for 100nt of upstream sequence for each gene and then further analysed for whether the identified potential uORF(s) location was within the upstream sequence up to the predicted splice site (Benz *et al*., 2005) for each gene (predicted 5'UTR)
[b] nd: not determined; numbers in parentheses refer to the number of genes identified to possess a potential uORF(s) in the total number of genes in the dataset analysed
[c] *T. brucei*, strain TREU 927 genome (downloaded from www.genedb.org; Berriman *et al*., 2005)
[d] Kabani *et al*. (2009), Additional file 5 (genes exhibiting a -1,-1,-1,-1,-1 trinary code profile: genes elevated in stumpy-forms versus all other samples)
[e] Kramer *et al*. (2008), characterised genes in Table S1

Initially, 200nt of upstream sequence for each gene was used for uORF identification. However, 62.3% of genes were found to possess uORF sequence(s) by this analysis. As this number is relatively large it was believed that this was probably

due to the result of false uORF assignments because the sequence length used was too long. Therefore, 100nt of upstream sequence for each gene was used instead. This should result in a more accurate prediction of uORF abundance, because this value is closer to the median length of *T. brucei* 5'UTRs, which was predicted to be 68nt (Benz *et al*., 2005).

Using this approach, 25.38% of *T. brucei* genes were found to possess uORF sequence(s) (Table 4.1). The maximum number of uORFs found for one gene was six, which was only the case for one gene, Tb09.244.0340. The number of uORFs for each gene ranged from 1 to 4, with a median number of 1 and mean number of 1.2.

Interestingly, it was also noted that a low frequency of small sets of genes possessed predicted uORF(s) of identical amino acid sequence. In most of these sets the genes had the same gene product annotations and often were present on the same chromosome as each other. Thus, it appeared that these genes were probably the result of gene duplications where upstream sequence was also involved in the duplication event. This could mean that the predicted uORF is functional and important for control of gene expression so that the uORF sequence is conserved, or it may have no significance. However, further analysis of the upstream sequences for some of these sets of genes revealed that the whole 100nt upstream sequence was very well conserved between genes, suggesting that the uORFs within them were probably unimportant.

4.1.1 GO term enrichment

Next, an investigation was made to determine if there was any enrichment in different groups of genes in the dataset of predicted uORF-associated sequences. For this the Gene Ontology (GO) term annotations associated with each gene were analysed. GO terms are used to classify genes by their molecular function, involvement in a biological process, or cellular localisation. Therefore, a comparison of these associated annotations between the dataset of genes predicted to possess an uORF and the dataset of genes that are predicted to not, should reveal any enrichment in different classes of genes.

For each set of genes, the associated GO terms for each gene were obtained from the SQL environment of the Gene Ontology database (http://www.berkeleybop.org/goose) using the SQL query provided in Appendix 2. The frequency of each GO term was then compared between the results for the two datasets. Shown in Table 4.2 are the GO terms with a fold enrichment of 2 or greater in the dataset of genes predicted to possess an uORF.

Table 4.2 Gene Ontology (GO) annotations enriched by greater than 2-fold in genes predicted to possess potential uORF(s) compared to genes not predicted to possess potential uORF(s).

GO ID	GO description	Fold enrichment [a]
GO:0007009	plasma membrane organization	(6)
GO:0005885	Arp2/3 protein complex	5
GO:0008194	UDP-glycosyltransferase activity	4.4
GO:0000315	organellar large ribosomal subunit	(4)
GO:0006686	sphingomyelin biosynthetic process	(4)
GO:0047493	ceramide cholinephosphotransferase activity	(4)
GO:0004519	endonuclease activity	3
GO:0004185	serine-type carboxypeptidase activity	(3)
GO:0019889	pteridine metabolic process	(3)
GO:0006506	GPI anchor biosynthetic process	2.3
GO:0030629	U6 snRNA 3'-end binding	2
GO:0004497	monooxygenase activity	2
GO:0005507	copper ion binding	2
GO:0006725	cellular aromatic compound metabolic process	2
GO:0006816	calcium ion transport	2
GO:0006952	defense response	2
GO:0008094	DNA-dependent ATPase activity	2
GO:0015630	microtubule cytoskeleton	2
GO:0016462	pyrophosphatase activity	2
GO:0000113	nucleotide-excision repair factor 4 complex	(2)
GO:0004998	transferrin receptor activity	(2)
GO:0006259	DNA metabolic process	(2)
GO:0008324	cation transmembrane transporter activity	(2)
GO:0009103	lipopolysaccharide biosynthetic process	(2)
GO:0030036	actin cytoskeleton organization	(2)

[a] Numbers in parentheses are the frequency of the GO term in the genes predicted to possess potential uORF(s) dataset because the frequency of the GO term in the genes not predicted to possess potential uORF(s) dataset was zero.

As can be seen, a number of different GO terms were found to be enriched in the dataset of genes predicted to possess an uORF. These appear to be largely unrelated and involve many different cellular processes. For example, there are terms associated with the plasma membrane, metabolic processes, DNA-related activities, the cytoskeleton, and possible cell signalling activities (for example, 'pyrophosphatase activity'). These enriched terms therefore, appear to be general and unspecific, indicating that the genes possessing an uORF sequence are probably not enriched for any specific functional groups.

4.2 Is there an enrichment of potential uORFs in genes enriched in stumpy forms?

To investigate if there was an enrichment of genes with predicted uORFs in stumpy-enriched transcripts, a subset of genes identified as upregulated in stumpy forms by microarray analysis (Kabani *et al*., 2009) was investigated. This set of genes was found to be elevated in stumpy forms compared to all other samples investigated in that study (Additional file 5, genes exhibiting -1, -1, -1, -1, -1 trinary code; Kabani *et al*., 2009). This subset of genes was found to have an incidence of genes with predicted uORFs of 30% (12/40 genes), versus the overall genome frequency of 25.38%. Therefore, this set of stumpy form upregulated genes showed a small enrichment in potential uORFs.

Further analysis was made of the 12 stumpy form upregulated genes with predicted uORFs to determine whether the predicted uORF was likely to be within the 5'UTR of the gene. This further analysis was not performed on all genes with predicted uORFs found in the genome-wide study due to the large size of that dataset. To determine whether each predicted uORF was likely to be within the 5'UTR, the predicted 5' splice sites predicted from a study by Benz *et al*. (2005) were used. That study used data obtained from known *T. brucei* 5' splice sites to predict the splice site for each annotated gene in the *T. brucei* genome (Benz *et al*., 2005). To ascertain whether the uORF was within the predicted 5'UTR, each gene was viewed using Artemis software together with predicted 5' splice sites and uORF sequences.

This resulted in the discovery that the predicted uORFs were not within the predicted 5'UTR for 4 of these genes, leading to a decrease in the percentage of genes with predicted uORFs in this gene set to 20% (8/40 genes; Table 4.1). Therefore, the potential enrichment of genes with predicted uORFs in stumpy form upregulated genes compared to the genome-wide analysis appears to not be a true enrichment. However, this analysis of uORF location was not performed on the genome-wide analysis dataset, so it is unknown whether that percentage would also decrease through a similar analysis and how greatly. Of the 8 genes with predicted uORFs within the predicted 5'UTR, 4 are annotated as 'hypothetical protein, unlikely', 2 as 'hypothetical protein, conserved', 1 as 'varient surface glycoprotein (VSG)-related, putative', and 1 as 'ATP-dependent DEAD/H RNA helicase, putative' (Table 4.3).

Table 4.3 Genes identified as upregulated in stumpy forms or regulated during heat shock that have predicted uORFs within the predicted 5'UTR.

Gene ID	Gene Product	No. of uORFs
Upregulated in stumpy-forms [a]		
Tb927.1.390	hypothetical protein, unlikely	1
Tb09.160.4070	hypothetical protein, unlikely	1
Tb927.1.4510	hypothetical protein, unlikely	4
Tb09.211.1970	hypothetical protein, unlikely	1
Tb927.3.2540	variant surface glycoprotein (VSG)-related, putative	4
Tb11.47.0019	hypothetical protein, conserved	1
Tb09.211.2300	ATP-dependent DEAD/H RNA helicase, putative	1
Tb10.6k15.0300	hypothetical protein, conserved	3
Tb927.5.120	expression site-associated gene 9 (ESAG9) protein, putative	4
Tb927.7.170	expression site-associated gene 9 (ESAG9) protein, putative	1
Tb927.5.4620	expression site-associated gene 9 (ESAG9) protein, putative	1
Regulated in heat shock [b]		
Tb927.4.3110	tRNA-dihydrouridine synthase 4, putative	1
Tb927.3.5150	exonuclease, putative	1
Tb11.01.4750	elongation factor 1 gamma, putative	1
Tb927.4.2310	asparaginyl-tRNA synthetase, putative	1
Tb11.42.0003	t-complex protein 1 subunit beta	1
Tb11.02.1930	ATP-dependent RNA helicase	1
Tb10.70.6180	ATP-dependent DEAD/H RNA helicase	1

[a] Either from microarray analysis (Kabani *et al*., 2009) or annotated ESAG9 genes
[b] From Kramer *et al*. (2008; characterised genes in Table S1)

4.2.1 Predicted uORF incidence in other stumpy-enriched genes

Additional genes known to be upregulated in stumpy forms were also investigated for predicted uORFs and whether these were located within the predicted 5'UTR for each gene. One group of genes investigated were the PAD array of genes, which are upregulated in stumpy forms. None of the genes of this array were predicted to possess uORFs in the genome-wide prediction, and none had an uORF sequence within the predicted 5'UTR either. Another group of genes investigated were the ESAG9 family of genes, which are also enriched in stumpy forms. Of the 9 ESAG9 genes analysed, 7 were found to possess a uORF by the genome-wide analysis. When investigation was made as to whether these were within the predicted 5'UTR however, it was discovered that this was the case for only 3 of the genes (Tb927.5.120, Tb927.7.170, and Tb927.5.4620; Table 4.3). Interestingly, Tb927.5.4620 is ESAG9-EQ, whose 3'UTR was investigated for contributions to gene expression control in this study (see Chapter 3). The analysis of the 5'UTR is presented in this chapter due to the presence of an uORF within the predicted 5'UTR for this gene.

4.2.2 Investigation of predicted uORF incidence in another set of co-regulated genes

As described above, there was a slight enrichment in the number of genes predicted to possess an uORF in the genes upregulated in stumpy forms compared to the genome-wide analysis. To further investigate if this represented a true enrichment, another set of co-regulated genes was analysed for uORF enrichment in the same way. This set of genes was the gene set identified as regulated during heat shock in the study by Kramer *et al*. (2008). The study involved an investigation of the global steady state levels of mRNAs following heat shock through microarray analysis, to identify transcripts upregulated or downregulated due to this treatment (Kramer *et al*., 2008). The set of genes that were identified (Table S1, Kramer *et al*., 2008) were used for uORF analysis.

The percentage of genes predicted to possess an uORF in the genome-wide prediction was 19.67% for the heatshock dataset (12 of 61 genes; Table 4.1). When the uORF localisation within the predicted 5'UTR was considered, this percentage decreased to 11.48%, as the predicted uORFs for 5 of the 12 genes were found not to be present within the predicted 5'UTR (Table 4.1). These percentages are lower than those obtained for the genome-wide analysis and for the stumpy form upregulated gene set. Therefore, there may be an under-representation of uORFs in the genes regulated during heat shock.

Consequently, the small enrichment seen with the stumpy form upregulated gene set may be a true enrichment, even though this decreased with further analysis of uORF location. Additionally, when a set of genes found to be enriched in stumpy forms from a preliminary Illumina digital tag RNA sequencing experiment (Keith Matthews, unpublished data; the top 100 most highly enriched) was later analysed for potential uORF frequency within 100nt of upstream sequence, the incidence was found to be 18% of genes. Suggesting that there is not an enrichment of potential uORFs in stumpy-enriched genes compared to the genome-wide analysis. Nonetheless, following the original bioinformatic predictions, the next step was to investigate if the uORFs identified could be functional in stumpy forms and therefore, contribute to control of gene expression.

4.3 Do the 5'UTRs of stumpy-enriched genes with potential uORFs control gene expression?

The aim of this project part was to assess if uORFs are used by *T. brucei* as a mechanism for controlling gene expression in stumpy forms, either for the escape from translational repression of stumpy form enriched genes, or as a mechanism of translational repression for genes repressed in stumpy forms. To address this question the 5'UTRs of three stumpy-enriched genes which potentially possess an uORF were investigated. These were Tb09.211.2300 and Tb11.47.0019, found to be upregulated in stumpy forms by microarray analysis (Kabani *et al*. 2009) and ESAG9-EQ, which was found to be highly enriched in stumpy forms (Barnwell *et*

al., 2010), all of which were identified to possess an uORF sequence in their predicted 5'UTR (Section 4.2). Tb09.211.2300 was chosen because it is annotated as a putative ATP-dependent DEAD/H RNA helicase and so may play a role in gene expression control in stumpy forms. Tb11.47.0019 was chosen because although annotated as a conserved hypothetical protein, it possesses a predicted ubiquitin associated (UBA/TS-N) domain, and so may play a role in protein degradation in stumpy forms. Therefore, both looked potentially interesting for further study. ESAG9-EQ was chosen to complement the study of the 3'UTR for this gene (see Chapter 3).

To assess if each potential uORF was responsible for control of gene expression, first the contribution of each 5'UTR to the control of gene expression was analysed. For this, the same CAT reporter gene approach was used as was used to assess the TbACO 5'UTR (see Section 3.2). Therefore, the same expression vector, CAT449 (described in Section 3.1), containing the CAT reporter gene was used and the aldolase 5'UTR sequence was replaced with each experimental 5'UTR sequence. If the experimental 5'UTR was found to repress CAT reporter gene expression, then the potential uORF present in the 5'UTR sequence may be responsible for this effect.

4.3.1 Generation of 5'UTR constructs

To discover the 5'UTR sequences to be used for investigation, the annotated *T. brucei* TREU 927 genome sequence was viewed with Artemis software to identify the intergenic region upstream of each gene. For Tb09.211.2300 the nearest upstream annotated gene in the genome sequence is Tb09.211.2310. The intergenic region between the stop codon of Tb09.211.2310 and the annotated start codon of Tb09.211.2300 is 394 bp and is shown in Figure 4.1 (A). The genomic sequence of each chromosome can also be viewed with Artemis software together with the predicted splice addition sites (SAS) from Benz *et al*. (2005). For Tb09.211.2300, this revealed that the nearest predicted SAS was at position -125 relative to the annotated start codon of Tb09.211.2300 (see Figure 4.1 A). As can be seen in Figure 4.1 (A) the potential uORF is therefore within the predicted 5'UTR sequence. As the upstream intergenic region for Tb09.211.2300 is relatively short in length, this entire

intergenic sequence was used to generate the 'Tb09 5'UTR' construct. To amplify this region by PCR the primers 33 and 34 (see Table 2.1) were used.

A

Tb09 5'UTR(s)
TAGcgatgtgacttaatgtgccgtccaattctacctggttacaattttcatatattatatgc
ctttgcatcataggacgtcgctccttttttttttgcaatttagctttgcatggtaaatagtt
acactattttcttacgtccgcgtgtgtgtcgcccctattacccgatccttgctcgttttctc
tctgtctttatttgtttctgttgtttgacttcctgcctgttgggacctcggagcacgttgca
acttcgtgatttttctcgatgctaggagcttctttaaatttcttacagaggtggtgtgatat
uORF
atatgtgtttggttggtgtattgtactctttgcataaagtttctattagcttcatgtgagcg
Tb09 5'UTR(e)
tcgtactgtgggtgtggttgtggtgATGGTCGTTGCAGTGGGGTGGAAAATTCCCGAGTTGG
TTAGTGTGTTCGGCAAGCGAAAATTTACTTGTGAAGATGTACCTTCGTTTTTCTCTTTGTTT
TTTTTTCCTTCGCATTATCGTTTCGTCTGTCGTGTACTTGCCTTCTCTCCTC**TGCATTTGGT**
TCACCTTTATTCTCTGGTTATAAACGATACCATTTCCTACAGTTTTGTATATTCTGTGTTCC
CTTACTTCTTTTCGCTTTTTTTTTTTGATTTATATCTCGTAGGGATTCCTGCTGTGGCGATG

B

Tb11 5'UTR(s)
TAGtgttgttgttgttgctgttgtttttcttagttattaggggcgaatgtttctcatccgcc
uORF
Tb11 5'UTR(e)
gtcatgaggatagaaaggtatttatttttatagttaaggcctcctttgcgtgtgATGTGTGT
CTGTTTGTGTTGTCAAAGCTTTATCGGCAAATGGGTGGTGGTCTACCCGAGGGAAGGCCTTT
TGTTGTTGTTGTTGTTGTCTTATTGGGGAAGAAGTCAGCGCATAGAATTTTTTTTAAAA
GAAAAAATGGTGGCAAATTTGGGTAATATTCTTTTTCTTTCTATTTCTTTAAAAAGAAAAAA
ACGGTATTATAACAACATCTCGGCTAGAGTATATATATATATGTATATTACCTCCAACGTGC
CG**TCACGTTTTGTGCTTCTTAATCTATATATTTATCTCTCGGTACTGTTTTTTATGCCTTAC**
CCGACAAGTTACCATATTGCCCTTTGTCAAGCAACAAAACGCAAAGAACGTGGTTCGTTGAG
AATG

C

EQ 5'UTR(s)
TAA…(930bp)…gtttacaatccaacatatatatatataggtcgatccccctttgtctctt
uORF
EQ 5'UTR(e)
agtggtattttgtatgatttgttagtttacgtcgggggaaaaATG

Figure 4.1 Upstream intergenic regions of Tb09.211.2300, Tb11.47.0019 and ESAG9-EQ. A. and B. Upstream intergenic regions for Tb09.211.2300 (A) and Tb11.47.0019 (B). Shown are the nucleotide sequences from the stop codon of the nearest upstream annotated genes (Tb09.211.2310 and Tb11.47.0020 respectively) to the internal start codon (Siegel *et al*., 2010) of Tb09.211.2300 or Tb11.47.0019. The original start codon annotation is also indicated. C. Upstream intergenic region for Tb927.5.4620 (ESAG9-EQ). Shown is the genomic sequence from the stop codon of the nearest upstream annotated gene, Tb927.5.4610, to the annotated start codon for the coding sequence of ESAG9-EQ. Because this intergenic region is 1023 bp in

length, only the sequence immediately upstream, which was used to generate the 'EQ 5'UTR' construct (93nt), is shown. All sequences are in the 5' to 3' direction and from the *T. brucei* TREU 927 genome sequence. The locations of potential upstream open reading frames (uORFs) are indicated with the line above each sequence. Lower and upper cases denote the 5' intergenic region and nucleotide sequence of the original start codon annotation, respectively; red highlight: stop codon; red font: the original annotated and internal (also underlined) start codons; grey highlight: start (s) or end (e) of each sequence used to generate the construct named above that residue; ag: predicted SAS (Benz *et al*., 2005); AG: SASs found by RNA-sequencing (Siegel *et al*., 2010); blue font: 5'UTR identified by RNA-sequencing (Siegel *et al*., 2010); bold font: sequence used for the microarray oligonucelotide in the study by Kabani *et al*. (2009).

For Tb11.47.0019 the nearest upstream annotated gene in the genome sequence is Tb11.47.0020. The intergenic region between the stop codon of Tb11.47.0020 and the annotated start codon of Tb11.47.0019 is 113 bp and is shown in Figure 4.1 (B). Analysis of the predicted splice addition sites from Benz *et al*. (2005), as for Tb09.211.2300, revealed that the nearest predicted SAS was at position -84 relative to the annotated start codon of Tb11.47.0019 (see Figure 4.1 B). Therefore, again the potential uORF is within the predicted 5'UTR sequence (see Figure 4.1 B). As the upstream intergenic region for Tb11.47.0019 was again short in length, this entire intergenic sequence was used to generate the 'Tb11 5'UTR' construct. To amplify this region by PCR the primers 35 and 36 (see Table 2.1) were used.

For ESAG9-EQ (Tb927.5.4620) the nearest upstream annotated gene in the genome sequence is Tb927.5.4610. The intergenic region between the stop codon of Tb927.5.4610 and the annotated start codon of ESAG9-EQ is 1023 bp and is shown in Figure 4.1 (C). The nearest predicted splice addition site from Benz *et al*. (2005) was at position -42 relative to the annotated start codon of ESAG9-EQ. Again the potential uORF was within the predicted 5'UTR sequence (see Figure 4.1 C). As the upstream intergenic region for ESAG9-EQ was relatively long, a shorter region (93 bp) that included the predicted SAS, its associated polypyrimidine tract, and the potential uORF was used to generate the 'EQ 5'UTR' construct. To amplify this region (the start and end of which is shown on the sequence in Figure 4.1 C in grey highlight) by PCR the primers 37 and 38 (see Table 2.1) were used.

Replacement of the aldolase 5'UTR in CAT449 used the *Xho*I and *Hind*III restriction enzyme sites as shown in Figure 3.1 (A). Construct generation was as described in Section 3.1. Therefore, as before stable transfectant cell lines were generated using *T. b. brucei* monomorphic bloodstream form strain Lister 427 and selected with 0.5 μg/ml phleomycin. For each construct, two biological replicates (independently derived clones) were examined in CAT ELISA assays, the results being shown in Figure 4.2 (A). Three biological replicates for each construct were examined by Northern blot analysis and these results shown in Figure 4.2 (B).

4.3.2 Analysis of the 5'UTR constructs with the CAT reporter gene assay

In comparison to the unmodified vector, all three constructs with the Tb09 5'UTR, Tb11 5'UTR, or EQ 5'UTR gave a strong repression of CAT protein expression (Figure 4.2 A), where CAT protein was almost undetectable. As a percentage of CAT protein expression obtained for the unmodified vector these values were 0.91% (±0.39%), 1.3% (±0.77%), and 0.69% (±0.65%), for the Tb09 5'UTR, Tb11 5'UTR, and EQ 5'UTR constructs respectively.

Analysis of *CAT* mRNA levels by northern blot analysis revealed that these constructs also resulted in almost undetectable *CAT* mRNA levels in comparison to the unmodified vector (Figure 4.2 B). Even overnight exposures produced only very faint signals. However, for the EQ 5'UTR and Tb11 5'UTR constructs there was a strong signal obtained for a higher molecular weight band, that appeared in both cases to be a processing intermediate, similar in size to that seen for the unmodified vector (Figure 4.2 B). This was unlikely to be due to a shift in *CAT* mRNA size; the experimental 5'UTRs were not long enough to create the size shift observed.

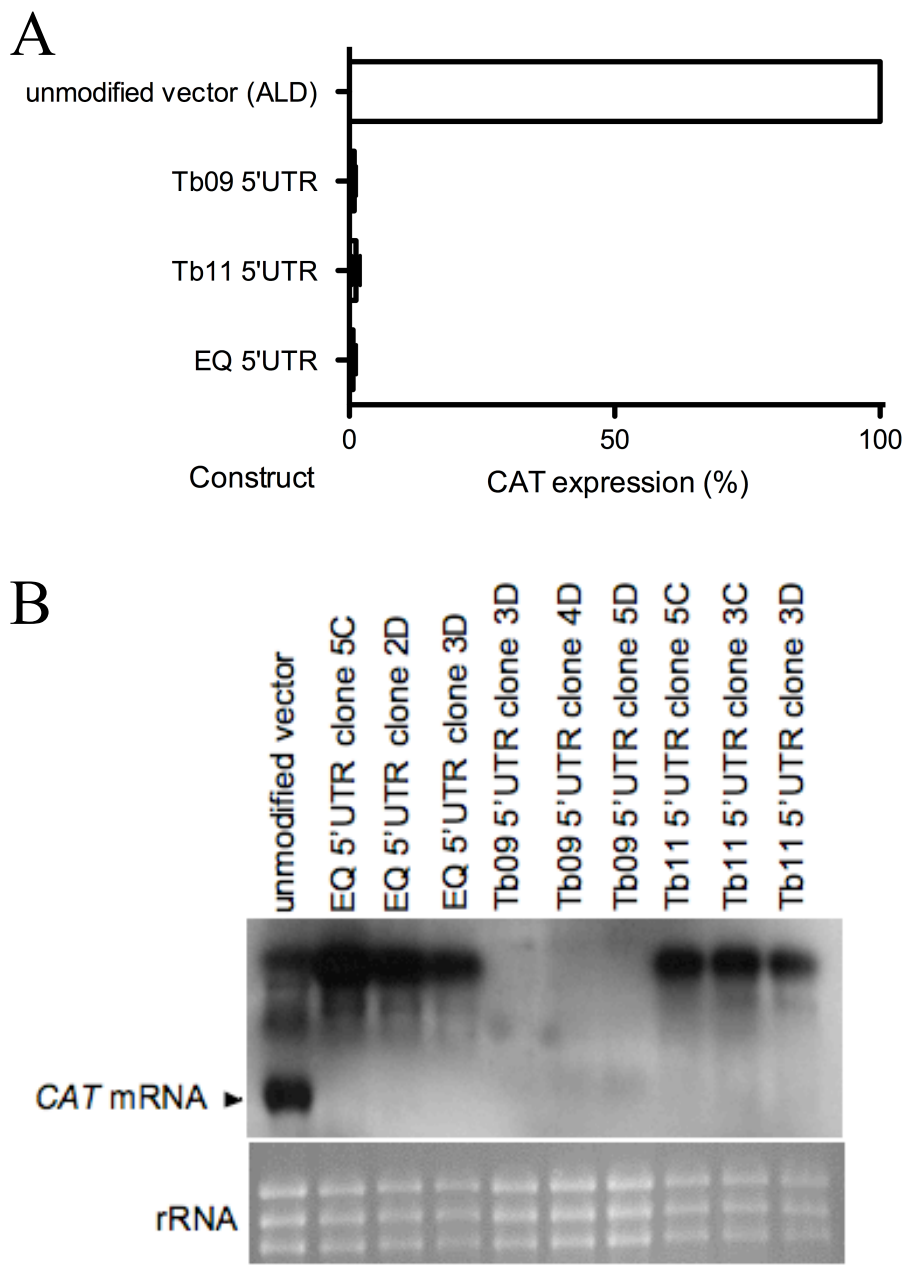


Figure 4.2 Analysis of Tb09.211.2300, Tb11.47.0019 and ESAG9-EQ 5'UTRs using a CAT reporter gene assay. A. CAT protein expression from the constructs indicated as a percentage of expression from one cell line of unmodified vector (control). For each of Tb09 5'UTR, Tb11 5'UTR and EQ 5'UTR the number of biological replicates was two, each having two experimental replicates. The mean ± standard error of biological replicates is shown on the graph. B. Northern blot analysis using a digoxygenin-labelled RNA probe against the CAT coding region. RNA samples: 3 biological replicates (independently derived clones) of each of Tb09 5'UTR, Tb11 5'UTR, and EQ 5'UTR; one clone of unmodified vector as a control. The panel below the blot shows the ethidium bromide staining of the rRNA present in each sample analysed to indicate loading of each lane.

From the data presented in Figure 4.2, it appears that each 5'UTR caused repression of gene expression in slender forms, at the levels of RNA and protein. Therefore, it was possible that the potential uORF in each 5'UTR sequence was responsible for this repression. If this were the case, the low *CAT* mRNA levels would suggest that this occurs through termination of translation of the uORF peptide stimulating mRNA decay. However, it could also be possible that these constructs were generating unstable *CAT* mRNA that was degraded due to the usage of incorrect splice addition sites upstream, within the vector sequence. Therefore, an investigation of the contribution of the potential uORF in each construct to the control of CAT gene expression was made.

4.4 Is the potential uORF responsible for repression in slender forms?

A functional analysis of the potential uORF in each 5'UTR sequence was made using the CAT reporter gene assay. The same constructs were used as used in the analysis in Section 4.3, except that the uORF was mutated. This mutagenesis was carried out using the QuikChange XL Site-Directed Mutagenesis kit from Stratagene (see Section 2.3.12) and involved introducing a single base pair mutation in the start codon for each potential uORF to change the sequence from ATG to ACG. As the start site for translation is therefore destroyed, a scanning ribosome will no longer recognise a start codon in this position, so will not be affected by an uORF sequence and should recognise the start codon of the downstream ORF. This approach has been used previously in a similar manner, to assess uORF function in malaria parasites by Amulic *et al*. (2009).

Primers for mutagenesis using QuikChange XL Site-Directed Mutagenesis kit (Stratagene; see Section 2.3.12) were designed according to the manufacturer's guidelines. For mutation of the Tb09 5'UTR, primers 39 and 40 were used; for mutation of the Tb11 5'UTR, primers 41 and 42 were used; for mutation of EQ 5'UTR, primers 43 and 44 were used (see Table 2.1). After ensuring mutagenesis of the potential uORF start codon from ATG to ACG was successful by DNA sequencing, each of the resulting constructs, named 'Tb09m', 'Tb11m', and 'EQm' 5'UTRs were used to generate stable transfectant cell lines. Again, *T. b. brucei*

monomorphic bloodstream form strain Lister 427 were used and selected with 0.5 μg/ml phleomycin. For each construct, two biological replicates (independently derived clones) were examined using CAT ELISA assays; the mean of the biological replicates is shown in Figure 4.3 (A). Three biological replicates for each construct were examined by Northern blot analysis and these results shown in Figure 4.2 (B and C).

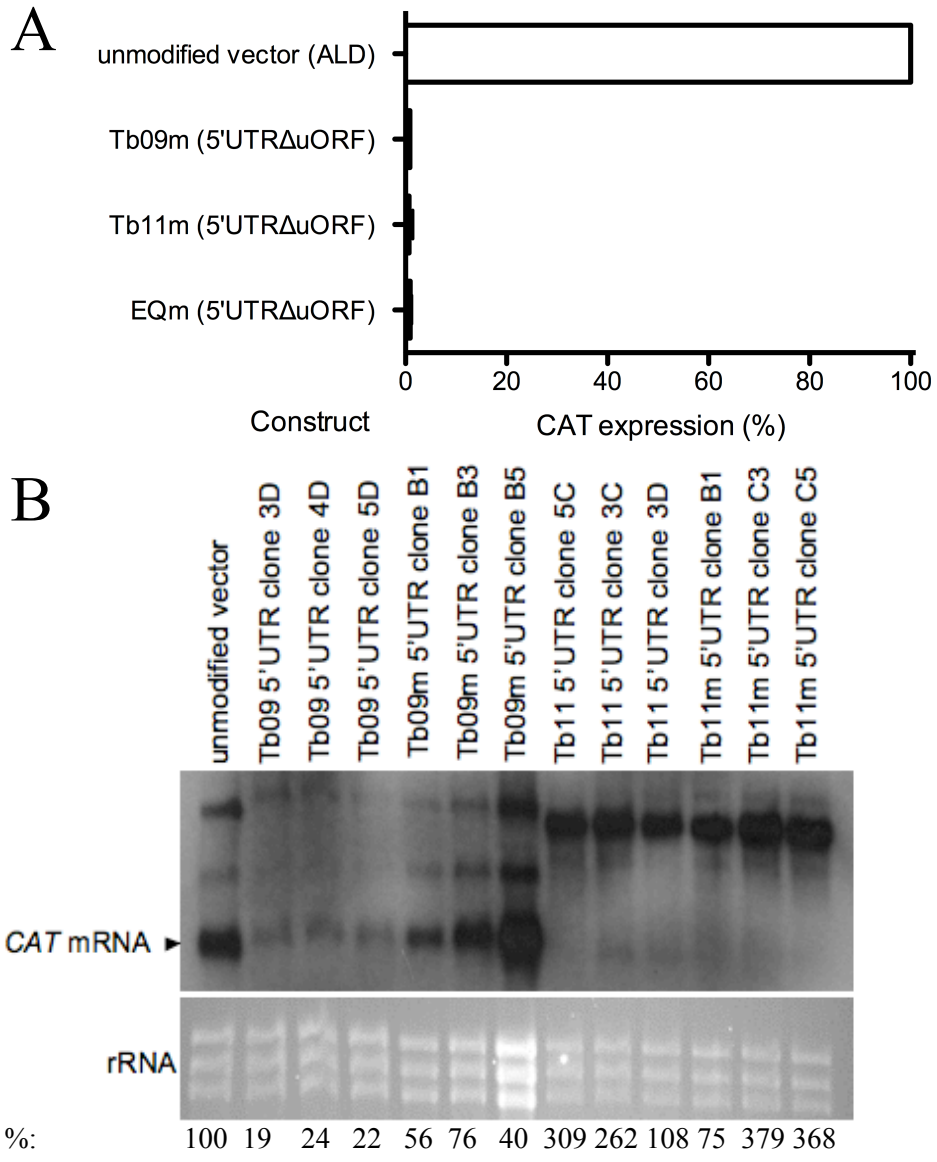


Figure 4.3 continues on the following page

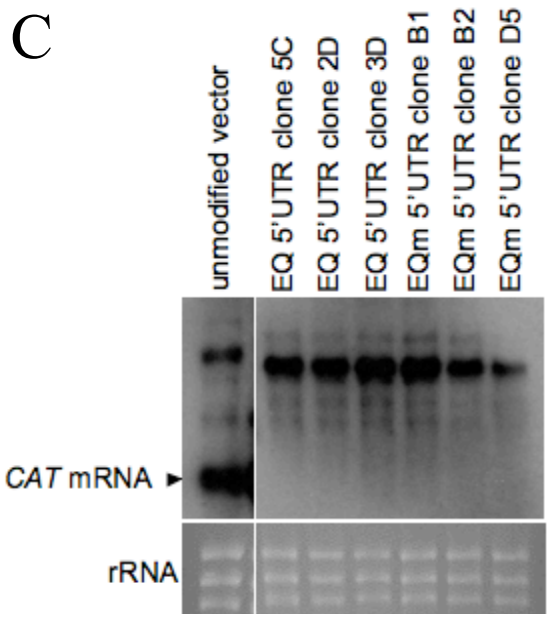


Figure 4.3 Analysis of Tb09.211.2300, Tb11.47.0019 and ESAG9-EQ 5'UTRs with mutated uORF start codon using a CAT reporter gene assay. The start codon of the uORF present in each 5'UTR was mutated from ATG to ACG using site-directed mutagenesis (see Section 2.3.12). A. CAT protein expression from the constructs indicated as a percentage of expression from one cell line of unmodified vector (control). For each of Tb09m, Tb11m and EQm the number of biological replicates was two, each having two experimental replicates. The mean ± standard error of biological replicates is shown on the graph. B. and C. Northern blot analyses using a digoxygenin-labelled RNA probe against the CAT coding region, each with an RNA sample of one clone of unmodified vector as a control. RNA samples: 3 biological replicates (independently derived clones) of each of Tb09 5'UTR, Tb09m 5'UTR, Tb11 5'UTR, and Tb11m 5'UTR (B); and of EQ 5'UTR and EQm 5'UTR (C). The panel below each blot shows the ethidium bromide staining of the rRNA present in each sample analysed to indicate loading of each lane. Shown below are the mRNA quantifications following normalisation to rRNA, as a percentage of that obtained for the unmodified vector.

As can be seen in Figure 4.3 (A), the loss of the potential uORF by mutation does not cause a release of repression of CAT protein expression. The levels of CAT protein expression from each of the three constructs as a percentage of the expression obtained for the unmodified vector were 0.88% (±0.12%), 0.72% (±0.69%), and 0.94% (±0.29%) for Tb09m, Tb11m, and EQm 5'UTR constructs respectively. This was very similar to the CAT protein expression seen with the non-mutated 5'UTR constructs described in Section 4.3 (see Figure 4.2).

The results of the *CAT* mRNA expression for these constructs and the previous non-mutated 5'UTR constructs are shown in Figure 4.3 (B and C). With this northern blot analysis however, the levels of *CAT* mRNA from the Tb09 5'UTR clones appear to be marginally greater than previously (see Figure 4.2 B). The reasons for this are unclear. For Tb11 5'UTR and EQ 5'UTR, mutation of the uORF with the Tb11m and EQm constructs resulted in no detectable change in *CAT* mRNA levels. However, there was an increase in *CAT* mRNA levels with mutation of the uORF in the Tb09 5'UTR sequence, notwithstanding the overloading of Tb09m clone B5. Nonetheless, as mentioned above this did not result in an increase in the level of CAT protein expression.

From the data presented in Figure 4.3, it appears that the potential uORF in each 5'UTR does not contribute to the repression of gene expression seen in slender forms. This is because there was no increase in CAT protein expression levels upon mutation, and therefore loss, of the potential uORF. Even though mutation of the uORF did result in an increase in *CAT* mRNA expression levels for the Tb09 5'UTR, this did not correlate with an increase in CAT protein levels. If the uORF in this 5'UTR were responsible for repression of gene expression by causing degradation of the transcript, it would be expected that loss of the uORF would result in increased RNA and therefore, protein levels. Since this was not observed, it appears that either there are other elements that are responsible for the repression of mRNA expression, or that the presence of these 5'UTR sequences in the vector somehow causes the generation of *CAT* transcripts that are rapidly degraded in a manner where the uORF is not functional.

4.5 Is gene expression repression released in 'stumpy-like' forms?

Despite the conclusion that the potential uORFs were not responsible for gene repression in slender forms (except possibly in the Tb09 5'UTR at the RNA level) it remained possible that they could be responsible for the release of gene repression in stumpy forms. Therefore, first it was necessary to determine if each 5'UTR sequence could drive gene expression in stumpy forms. If so, further analysis could be made

by analysis of the mutated uORF constructs in stumpy forms. As with the analysis made for the ESAG9-EQ 3'UTR (Chapter 3), the monomorphic cell lines analysed for CAT expression in slender forms discussed above were, therefore, treated with 8-pCPT-2'-O-Me-cAMP to assess CAT expression in 'stumpy-like' forms.

The experiments were performed as described in Section 3.7, with comparisons being made between treated and control cells. For each construct analysed, there were two biological replicates assessed. For each of these two experimental replicates were carried out, meaning that the CAT ELISA assays and northern blot analyses were performed with samples taken from two separate 8-pCPT-2'-O-Me-cAMP treatment experiments.

The CAT protein expression levels obtained for cells treated with 8-pCPT-2'-O-Me-cAMP or control treated for each construct, are shown in Figure 4.4 (A). As can be seen, treatment with the compound resulted in a similar very low, almost undetectable, level of CAT protein expression, to the untreated cells. Note that the data are presented as a percentage of the expression level obtained for the unmodified vector control cells because calculations of the fold-change between treated and untreated cells appears misleading. This was due to the very low levels of CAT protein expressed from the constructs being tested, which meant that the variation between samples could produce misleading fold changes.

The levels of *CAT* mRNA expression in the cells treated, or not, with 8-pCPT-2'-O-Me-cAMP for each construct, is shown in Figure 4.4 (B). A quantification of the amount of the predominant band present for each sample was performed, but here the values are a percentage of the untreated cell sample instead of the untreated unmodified vector sample, this being due to the absence of a similar size band to the predominant band observed in lanes 7-14 in this sample. This quantification indicates that drug treatment reduced the mRNA abundance of the Tb11.47.0019 and ESAG9-EQ 5'UTR construct transcripts (the situation for the Tb09.211.2300 5'UTR construct is unclear due to wide clonal variability).

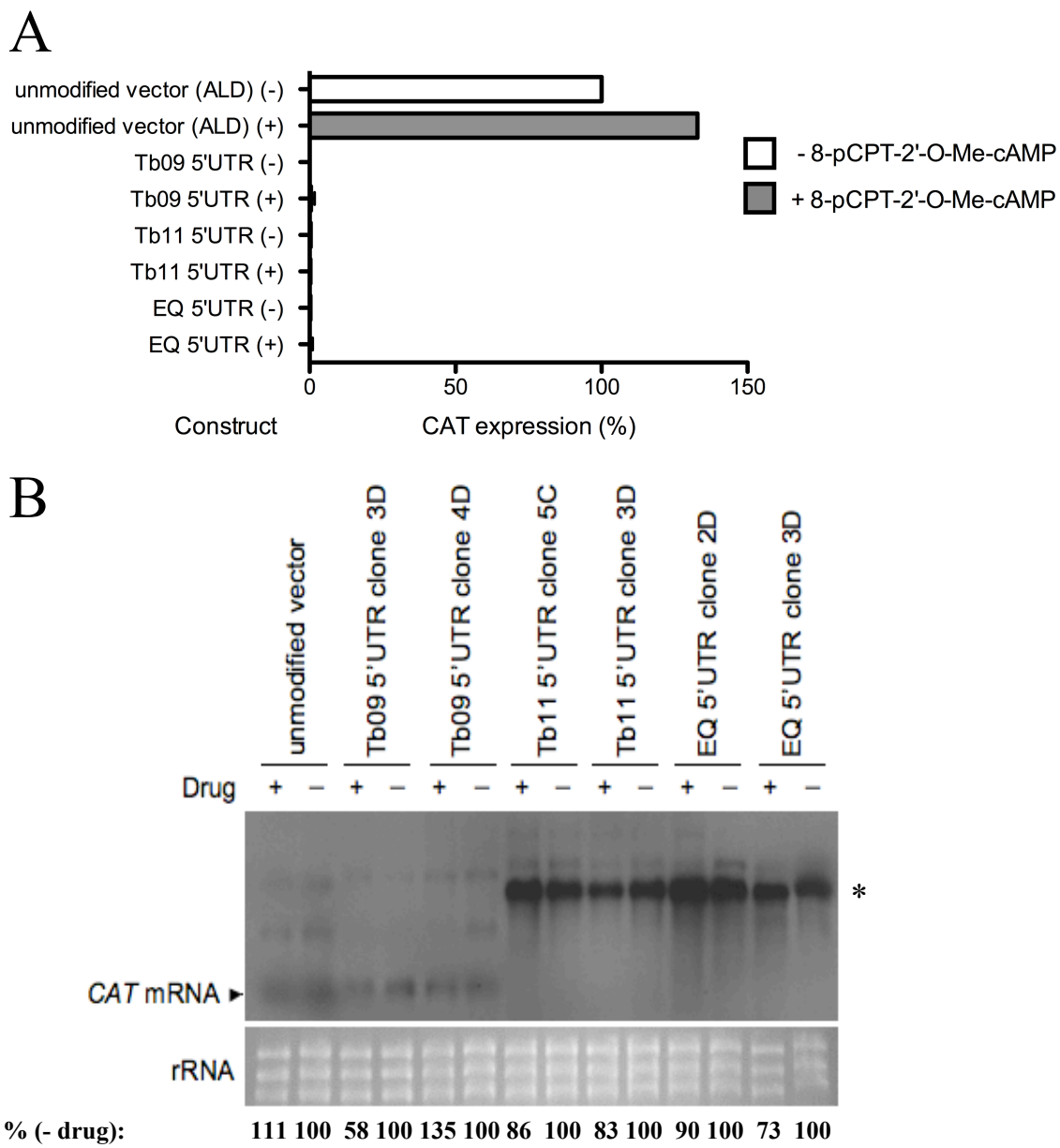


Figure 4.4 Response of CAT reporter gene expression to 8-pCPT-2'-O-Me-cAMP treatment in 5'UTR containing uORF based constructs. A. CAT protein expression for each construct, shown as a percentage of expression from the unmodified vector control treated cells, for cells treated with 10 µM 8-pCPT-2'-O-Me-cAMP and control treated cells. Shown is the mean ± standard error of two biological replicates; except the unmodified vector where $n = 1$. B. Northern blot analysis of RNA samples from one 8-pCPT-2'-O-Me-cAMP treatment experiment using a digoxygenin-labelled RNA probe against the CAT coding region, and the unmodified vector clone as a control. Samples are either from cells treated with 10 µM 8-pCPT-2'-O-Me-cAMP (+ drug) or control treated cells (- drug). Lower panel shows the ethidium bromide-staining of the rRNA present in each sample analysed to indicate loading of each lane. Below is quantification of the mRNA band indicated by the arrowhead (first 6 lanes) or the asterisk (rest of lanes) as a percentage of that of the untreated cell sample.

Because the endogenous transcripts are upregulated in stumpy cells, an increased level of expression would be expected if the 5'UTRs were responsible for alleviation of repression in stumpy forms, not the decrease observed in the mRNA quantification. The data shown in Figure 4.4 indicate that the 5'UTRs of Tb09.211.2300, Tb11.47.0019, or ESAG9-EQ are not responsible for a release of gene repression between slender and stumpy forms. However, this experiment does have the caveat that the reporter gene expression was assessed in stumpy-like cells and not true stumpy forms, which may produce a different result.

4.6 Internal splice site identification and usage for both Tb09.211.2300 and Tb11.47.0019 genes

Following the above analysis of the 5'UTR constructs and potential uORFs, it was discovered that both the Tb09.211.2300 and Tb11.47.0019 genes were members of a group of genes prediced to use an internal start codon (Siegel *et al*., 2010). As mentioned in Section 3.2.3 the Siegel *et al*. (2010) study used high-throughput RNA-sequencing to identify 5'-splice addition sites (SAS) for ~7000 genes, using *in vitro* cultured bloodstream and procyclic forms. From this, it was found that for 488 genes the identified SAS was downstream of the originally assigned gene start codon, meaning that a subsequent in-frame start codon must be used instead. Both Tb09.211.2300 and Tb11.47.0019 genes were found to be included in these 488 genes. These new start codon assignments and 5'UTRs identified are described below and shown in the sequences in Figure 4.1 (A and B).

4.6.1 SAS positions identified by Siegel et al. (2010)

The study by Siegel *et al*. (2010) found that the Tb09.211.2300 gene possessed one SAS only. This was located at position +261 relative to the original start codon annotation, generating a 5'UTR length of 17 bp downstream of this SAS (see Figure 4.1 A). This meant that the coding sequence length for Tb09.211.2300 would be 282 bp shorter than the original annotation and would generate a protein reduced by 94 amino acids in length. It also meant that the Tb09.211.2300 transcript would not contain the sequence with the potential uORF sequence (see Figure 4.1 A).

The study by Siegel *et al*. (2010) found that the Tb11.47.0019 gene possessed two SASs; located at positions +322 and +345 relative to the original start codon annotation (see Figures 4.1 B and 4.5). The 5'UTR lengths found were 55 and 32 bp respectively, but both would mean the same internal start codon was used (see Figure 4.1 B). This meant that the coding sequence length for Tb11.47.0019 would be 381 bp shorter than the original annotation and would result in a protein reduced by 127 amino acids in length. Again, it also meant that the Tb11.47.0019 transcript would not contain the potential uORF sequence (Figure 4.1 B).

4.6.2 SAS positions identified by Nilsson et al. (2010)

Another study also determined the 5' splice sites of 85% of annotated *T. brucei* annotated genes, but used a spliced leader trapping (SLT) methodology rather than RNA-sequencing (Nilsson *et al*., 2010). As well as identifying the splice site, this methodology was also able to determine the gene expression profile for slender forms, stumpy forms and procyclic forms. The SAS locations for Tb09.211.2300 and Tb11.47.0019 were therefore determined from the dataset obtained from this study, through http://splicer.unibe.ch/.

For the SAS location in Tb09.211.2300, only one SAS was found, which was present in each life-cycle stage and was the same as that found in the Siegel *et al*. (2010) study (see Figure 4.1 A). This therefore confirms the usage of an upstream splice site and an internal ATG used for this gene, and consequently, the absence of the uORF from the transcript, regardless of the lifecycle stage.

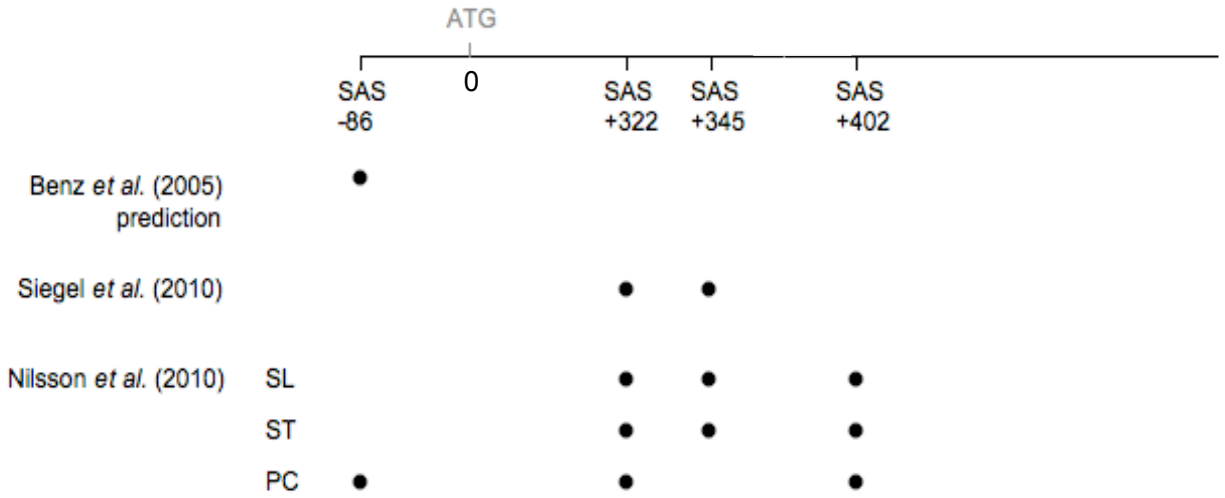


Figure 4.5 Schematic diagram showing the SAS locations identified for Tb11.47.0019. The SAS locations are indicated with 'SAS' and the position relative to the annotated ATG start codon (position 0). The black filled circles indicate the identification (or prediction from Benz *et al*., 2005) of the use of the above SAS from studies by Siegel *et al*. (2010) and Nilsson *et al*. (2010). SL: slender form; ST: stumpy form; PC: procyclic form. Drawing is not to scale.

For Tb11.47.0019, a number of different SAS locations were identified (see Figure 4.5). Each of the three life-cycle stages investigated possessed a SAS that was the same as that identified by Siegel *et al*. (2010) at position +322. Slender forms and stumpy forms were also found to possess a SAS at position +345 that was identified by Siegel *et al*. (2010) as well. Additionally, all three life-cycle stages were found to possess a minor SAS further downstream at position +402 relative to the original start codon annotation. The nearest downstream start codon in-frame with the stop codon of the original gene annotation, is almost immediately after the splice site, which would produce a protein of 352 amino acids. Following that, the next nearest in-frame start codon would produce a protein of 332 amino acids. Another minor SAS was found in procyclic forms only and is the SAS predicted by the Benz *et al*. (2005) study.

4.6.3 Determination of 5'UTRs using 5'RACE

In order to determine the 5'UTR present in the transcripts of the Tb09.211.2300 and Tb11.47.0019 genes in slender, stumpy and procyclic forms, 5'RACE was used. This was performed as described in Section 2.5 and a schematic diagram is shown in Figure 4.6 (A). Firstly, this involved generation of cDNA from RNA samples from bloodstream form Lister 427, AnTat1.1 stumpy forms and procyclic form Lister 427-449 cells using random hexamers as primers. To control for any amplification from contaminating genomic DNA, an identical reaction was set up for each RNA sample used, without reverse trancriptase. Additionally, as a negative control, an identical cDNA reaction was also performed (with reverse transcriptase present) but with the exception that water was added to the reaction instead of any RNA sample. The cDNA samples were then used as a template in PCRs, which were likewise used as template for a second nested PCR. This nested PCR approach was used because initial attempts using a single PCR reaction were unsuccessful for both genes.

For each of the Tb09.211.2300 and Tb11.47.0019 genes, two sets of primers were used; one set to amplify only the transcript that would be generated from a *trans*-splicing reaction at the predicted SAS (this region was named 'a') and one set to

amplify any transcripts that would be generated from downstream *trans*-splicing reactions, like that found by Siegel *et al*. (2010; this was named 'b'). The reason for these different reactions was that the PCR amplification would favour the amplification of shorter sequences, potentially creating bias in the analysis.

The first PCR used the splice leader primer (primer 45; Table 2.1) with either primer 46, 47, 48 or 49; the second PCR again used the splice leader primer (primer 45) with either primer 50, 51, 52 or 53. As a positive control, a reaction was included that used primer 54 in the first PCR reaction to generate a product of ~340 bp, then primer 55 in the second (nested) PCR reaction to generate a product of ~290 bp; these were primer sequences used successfully to map the SAS for TbREL2 (Tb927.1.3030) by Dr Achim Schnaufer (University of Edinburgh) using the same methodology. As a negative control, a PCR reaction with no template was included in each set of PCRs and contained the same primers as used for the positive control.

Shown in Figure 4.6 (B) are the products from the second PCRs. The major products of each reaction shown indicated with asterisks were excised from the gel and purified for ligation into pGEM T-easy vector (see Section 2.3.9). Following bacterial transformation and small-scale plasmid preparation (see Section 2.3.10), DNA from two bacterial colonies obtained for each amplicon where sequenced with M13 forward and reverse primers (primers 1 and 2). The major band for Tb11 'a' from slender and procyclic forms was not sequenced due to these products producing a much weaker signal in the initial experiment replicate.

For the Tb09.211.2300 gene, if there was a transcript present generated from the predicted SAS (Benz *et al*., 2005), a product of ~447 bp should be produced in the PCR using primers for the 'a' region and a product of ~785 bp using primers for the 'b' region. If there was a transcript present that was generated from the SAS identified at position +261, a product of ~126 bp should be produced in the PCR using primers for the 'b' region. From the sizes of main PCR product shown in Figure 4.6 (B), it would appear that only the transcript generated from the SAS identified at position +261 for this gene was present in each life-cycle stage.

A

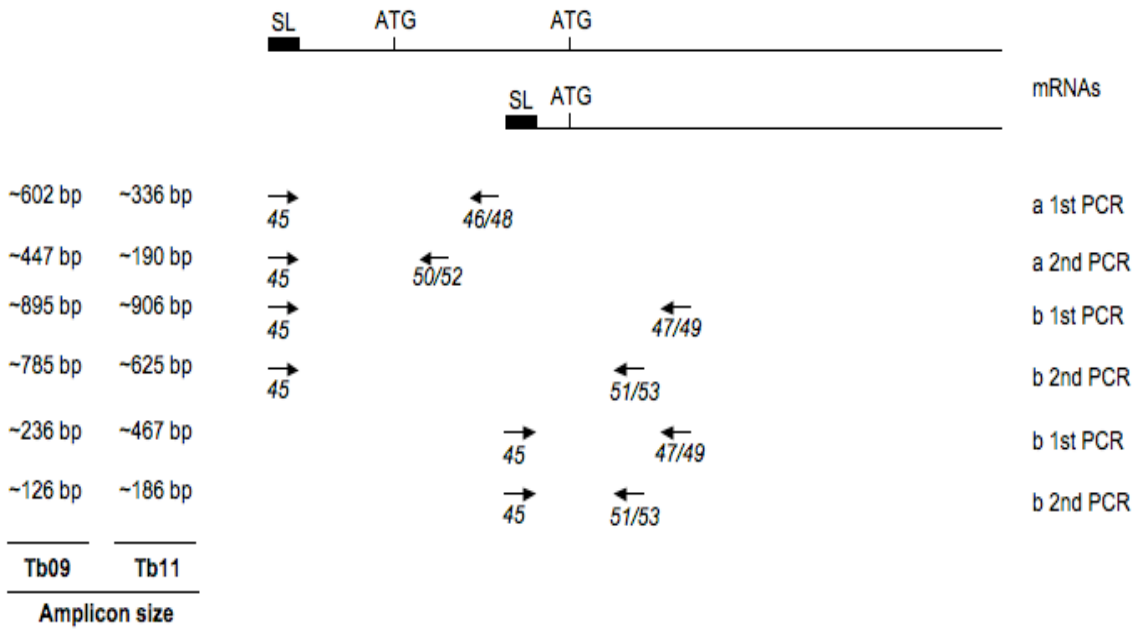


Figure 4.6 continues on the following page

B

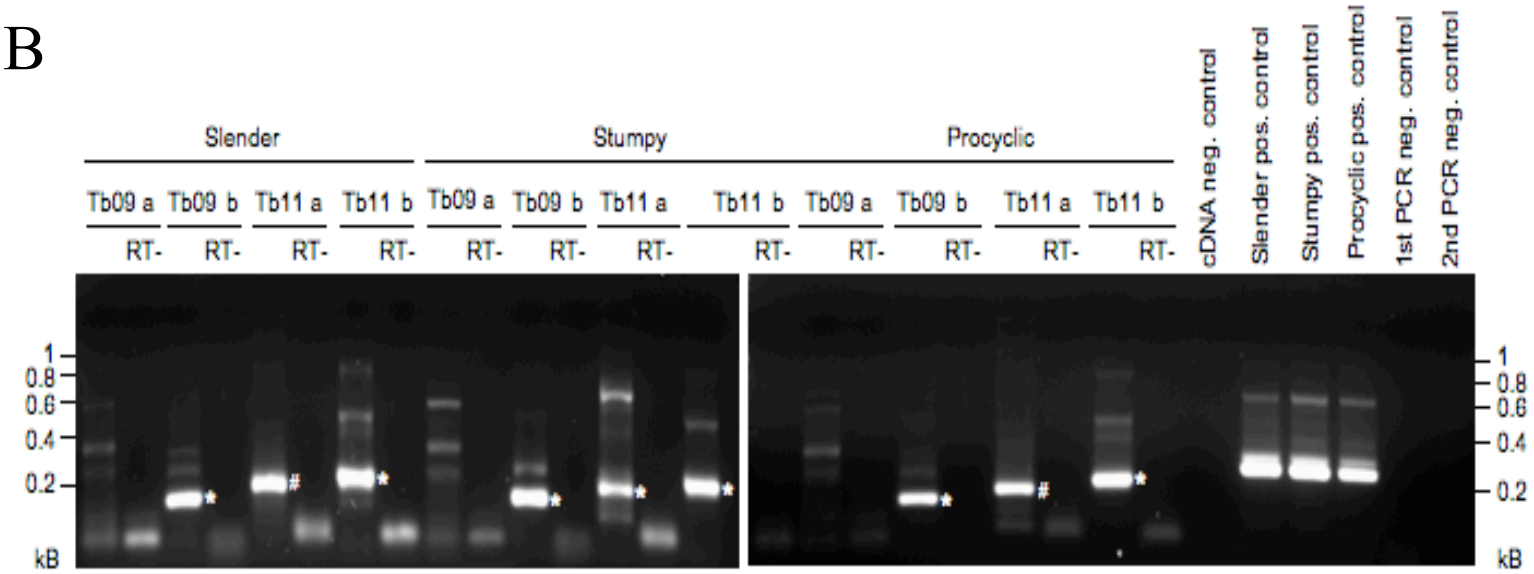


Figure 4.6 continues on the following page

Figure 4.6 Results of 5'RACE for Tb09.211.2300 ('Tb09') and Tb11.47.0019 ('Tb11'). A. Schematic diagram of PCR reactions using primers to amplify the 'a' and 'b' regions of each gene following generation of cDNA from RNA from each life cycle stage with random hexamers. Each PCR reaction used the spliced leader primer (primer number 45) as the forward primer and for each region two PCRs were performed, the second performed to form a nested PCR. Primer numbers are shown in italics and the sequences are given in Table 2.1. The expected PCR amplicon sizes are shown at the side of the diagram. B. Product from each of the second PCRs shown schematically in A. analysed on 1.5% w/v agarose gels. RT-: identical cDNA reaction performed with water instead of reverse transcriptase, neg. control: negative (water) control, pos. control: positive control using primers known to amplify Tb927.1.3030 (primer sequences provided by Dr Achim Schnaufer, University of Edinburgh) by 5'RACE. Asterisks mark bands used for sequencing, #s mark bands that were not sequenced. DNA sizes are shown in kilobases (kB). C. Schematic diagram showing the results of PCR product sequencing. In the grid on the right, positive identification of the 'a' or 'b' region 5'UTR depicted on the left, is indicated with a filled black circle. Filled grey circles indicate that a PCR product of the same size was identified for that sample, but that the 5'UTR sequence was not confirmed by sequencing. SL: splice leader; BSF: slender bloodstream form; ST: stumpy form; PCF: procyclic form.

For the Tb11.47.0019 gene, if there was a transcript present that is generated from the predicted SAS (at position -86; Benz *et al*., 2005), a product of ~190 bp should be produced in the PCR using primers for the 'a' region and a product of ~625 bp using primers for the 'b' region. If there was a transcript present that was generated from the SASs identified at positions +322/+345, a product of ~186/~163 bp should be produced in the PCR using primers for the 'b' region. From the sizes of the main PCR products (indicated in Figure 4.6 B with asterisks), it would appear that two transcripts are generated for this gene in each life-cycle stage; one from the predicted SAS and one from the SASs identified at positions +322/+345.

The results of sequencing of the main PCR products are given in Figure 4.6 (C). The identification of either the 'a' or 'b' region 5'UTR in each sample is indicated with a filled circle in the grid in Figure 4.6 (C). The sequencing confirmed that in each life-cycle stage investigated the 5'UTR identified by Siegel *et al*. (2010) was the sequence generated from the main PCR product for Tb09.211.2300 transcripts, as suggested by the product size. For Tb11.47.0019 transcripts, the sequencing confirmed the use of the predicted SAS (Benz *et al*., 2005) in each life-cycle stage, and the 5'UTR sequence generated by usage of the SASs identified at position +322/+345 (Siegel *et al*., 2010), as suggested by the PCR product sizes. As mentioned above, the main product for Tb11 'a' PCRs for slender forms and procyclic forms were not sequenced due to their much lower abundance in the first experimental replicate. Therefore, initially there was believed to be developmental regulation of alternative splice site usage for Tb11.47.0019, such that a longer transcript was generated with greater abundance in stumpy forms.

4.7 Is there a developmental difference in transcript abundance?

To further investigate the SAS location usage in the Tb09.211.2300 and Tb11.47.0019 genes in slender, stumpy and procyclic forms, northern blot analysis was used to analyse transcript abundance. Digoxygenin-labelled probes were generated against the 'a' and 'b' regions of each gene for this, which are shown in the schematic diagram in Figure 4.7 (A).

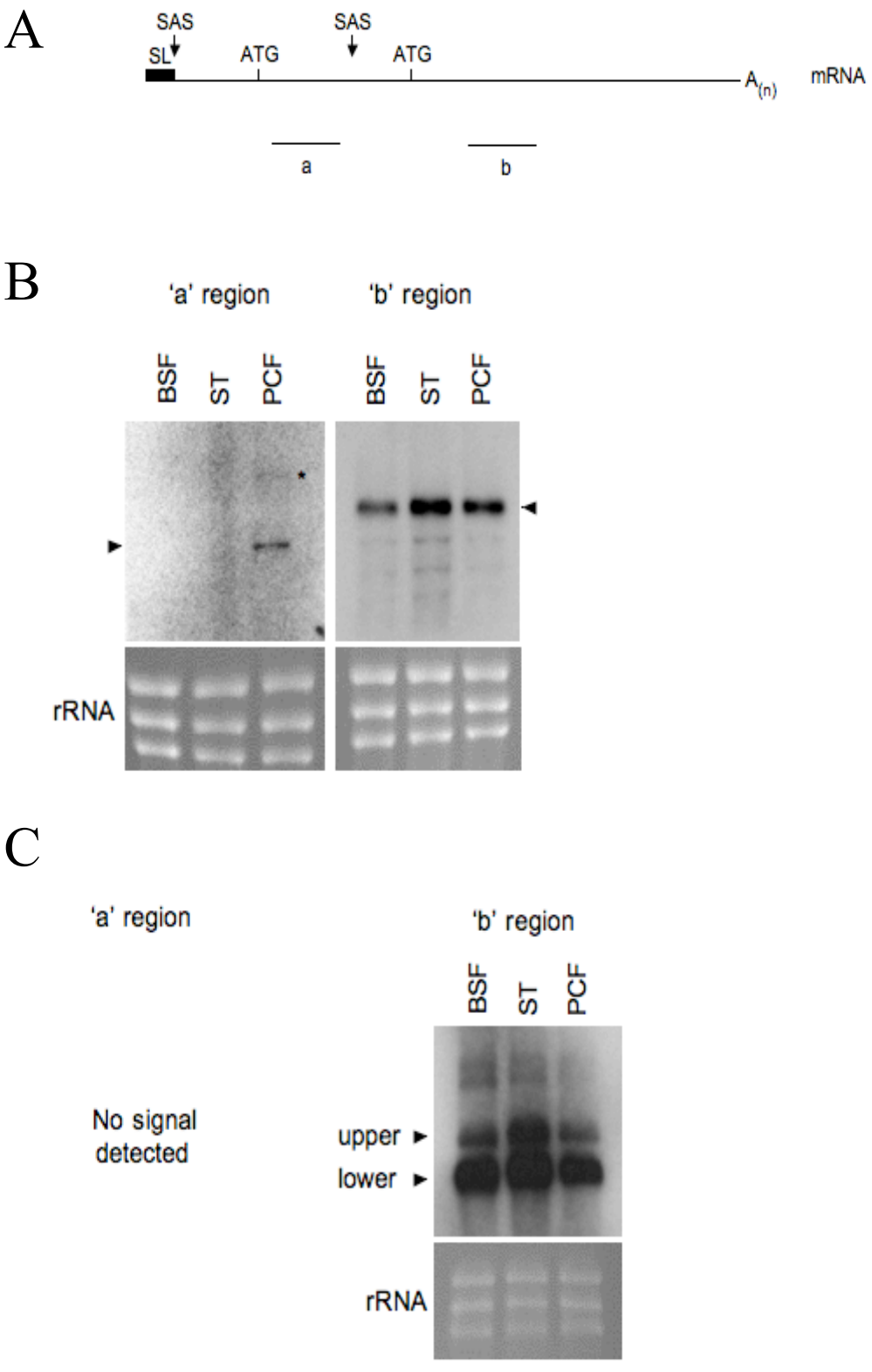


Figure 4.7 continues on the following page

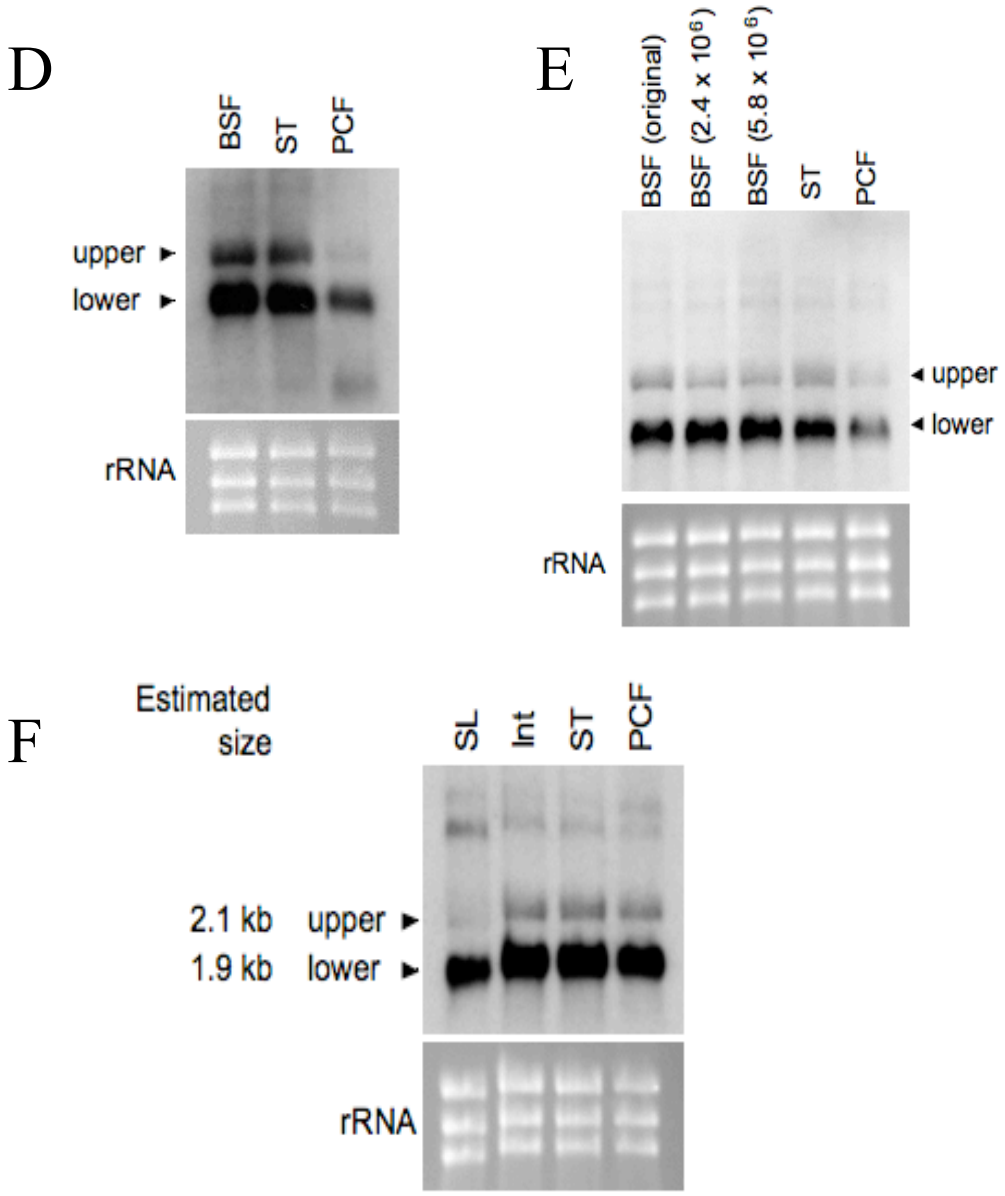


Figure 4.7 Detection of Tb09.211.2300 and Tb11.47.0019 transcripts by northern blot analyses. Digoxygenin-labelled probes were generated to detect the 'a' and 'b' regions of each gene transcript as shown in the schematic diagram in A. SL: splice leader; SAS: splice leader addition site; ATG: start codon. Northern blot analysis using the probes to the 'a' and 'b' regions of Tb09.211.2300 (B), and the probes to the 'a' (C only) and 'b' regions of Tb11.47.0019 (C, D, E, & F). C-E: used RNA samples from monomorphic slender bloodstream forms; F: RNA samples from pleomorphic slender forms, intermediate forms, and stumpy forms. BSF: bloodstream form; ST: stumpy form; PCF: procyclic form; SL: slender form; Int: intermediate form. Arrowheads in B mark bands detected, asterisk marks band of very feint signal. Arrowheads in C, D, E, & F mark upper and lower bands detected with Tb11.47.0019 'b' region probe. Estimated molecular sizes of both transcripts is shown in kb (F). The panel below each blot shows the ethidium bromide staining of the rRNA present in each sample analysed to indicate loading of each lane.

The probe to the 'a' region of each gene was designed against sequence of the gene that was downstream of the annotated start codon but upstream of the SAS identified by Siegel *et al*. (2010), Nilsson *et al*. (2010), and by 5'RACE above. To generate this for Tb09.211.2300, primers 56 and 57 (see Table 2.1) were used and for Tb11.47.0019 primers 58 and 59 were used in PCRs with *T. b. brucei* strain Lister 427-449 genomic DNA being the template for the reaction.

The probe to the 'b' region of each gene was designed against sequence of the gene that was downstream of the SAS identified by Siegel *et al*. (2010), Nilsson *et al*. (2010), and by 5'RACE above, and downstream of the next in-frame start codon following this SAS location. To generate this for Tb09.211.2300, primers 60 and 61 (see Table 2.1) were used and for Tb11.47.0019 primers 62 and 63 were used in PCRs as described above.

Each of the primers used to generate the probes mentioned in this section also contained *Bst*XI restriction enzyme sites compatible for sub-cloning in the pQuadra set of plasmids for RNAi, however the pALC14 plasmid was used for RNAi instead (see Section 4.10). After sub-cloning into pGEM T-easy plasmid, the sequence of each amplicon was verified. These plasmids were then used to generate the probes as described in Section 2.4.4.

4.7.1 Detection of Tb09.211.2300 transcripts

The results of northern blot analysis with the probes to the 'a' and 'b' regions of Tb09.211.2300 described above are shown in Figure 4.7 (B). As can be seen, the probe to the 'b' region detected one main band, which is indicated with an arrowhead at the right side of the blot. The band detected by the probe is of stronger signal intensity in the RNA sample from stumpy form cells, demonstrating that this transcript is upregulated in stumpy forms, matching the analyses by Kabani *et al*. (2009). There were three additional minor bands detected, but these appear to be cross-reactivity with the triplet of rRNA bands.

The probe to the ‘a’ region detected a faint band in the RNA sample from procyclic form cells, indicated with the arrowhead at the left of the blot. However, this appeared to be a similar size to one of the minor bands detected with the ‘b’ region probe that were thought to be the result of cross-reactivity. Another, even more faint, band was detected, which is indicated in Figure 5.7 (B) with an asterisk. This could represent a transcript produced by usage of the Benz *et al*. (2005) predicted SAS. However, the signal was so weak that it might also represent unspecific binding of the probe.

4.7.2 Detection of Tb11.47.0019 transcripts

The results of the first northern blot analysis with the probes to the ‘a’ and ‘b’ regions of Tb09.211.2300 described above are shown in Figure 4.7 (C). No signal was detected with the probe to the ‘a’ region. This was true for subsequent attempts involving northern blot analysis with a lower stringency hybridisation temperature (60°C, instead of 68°C) and a probe generated from the other strand of DNA template in case the amplicon was present in the opposite orientation to that suggested by sequencing of the vector. However, the probe to the ‘b’ region detected two main bands, which are indicated in Figure 4.7 (C) by the arrowheads and termed the upper and lower bands. It was speculated that the upper band may represent the transcript produced from usage of the predicted SAS and the lower band may represent the transcript produced from usage of the SAS at position +322. The signal of both bands, but especially the upper band, was of a greater intensity in the stumpy form RNA sample, suggesting upregulation of these transcripts in stumpy forms. This supported the initial 5’RACE analysis where there appeared to be a greater abundance of the long transcript form in stumpy forms.

However, when the northern blot analysis was repeated with the ‘b’ region probe a slightly different profile of expression was observed, this is shown in Figure 4.7 (D). Here the expression of both transcript forms appeared to be similar in both the slender and stumpy forms, but downregulated in procyclic forms. Two different slender form RNA samples were used for the blots shown in Figure 4.7 (C and D) and the cell concentration had been different at the time of sample generation.

Therefore, whether this difference in transcript abundance was caused by different cell density in the different samples was investigated. Shown in Figure 4.7 (E) is northern blot analysis performed with RNA samples generated from slender cells at different cell concentrations (2.4 $x10^6$ cells/ml and 5.8 $x10^6$ cells/ml) as well as a slender cell RNA sample used previously. Apparently, the cell concentration did not affect the abundance of either of the transcript forms.

Due to the variation of the signal intensities of both bands in the northern blot analyses shown above, RNA samples were generated from pleomorphic slender, intermediate and stumpy forms. This was performed because the previous slender form samples had been obtained from monomorphic slender forms and it was considered that an assessment of levels in true slender forms might be more informative. The northern blot analysis using these RNA samples is shown in Figure 4.7 (F). As can be seen, the abundance of both bands is greatly reduced in the slender form RNA sample. Intermediate and stumpy forms show increased abundance of the lower and upper bands compared to slender and procyclic forms. Therefore, the true slender forms appear to exhibit a lower abundance of the transcript forms than monomorphic slender forms.

These RNA samples and Northern blot analysis was also used to estimate the molecular size of both transcripts, which was 2.1 kb for the upper band and 1.9 kb for the lower band (Figure 4.7 F). These estimates are smaller than the predicted transcript sizes using the 5'UTR and ORF lengths, and largest 3'UTR length from the study by Siegel *et al*. (2010) combined, which are 1823nt and 1414nt for the long and short form, respectively. The significance of these estimated sizes is therefore unclear.

Because of the variation in transcript abundance observed between these northern blot analyses it is not possible to confidently conclude that there is developmental regulation of different splice site usage. However, taken all together they do seem to indicate that there might be upregulation of both transcripts in stumpy forms.

4.8 Potential functional consequences of alternative splice site usage for Tb11.47.0019

From the results of the 5'RACE experiments (shown in Figure 4.6; and supported by northern blot analyses, shown in Figure 4.7) the generation of two transcripts were found to be produced from the Tb11.47.0019 gene, resulting in the potential usage of two different start codons. To assess the possible functional consequences of this, the amino acid sequences of both long and short forms of the predicted proteins were used in bioinformatic analyses (see Section 2.1).

To identify any regions of the peptide sequences that may show homology to known conserved protein domains, the Pfam 24.0 database (http://pfam.sanger.ac.uk) was searched with the long (487 amino acids) and short (360 amino acids) form Tb11.47.0019 protein sequences. This returned one significant Pfam-A match and no Pfam-B matches for both forms. The significant result was a match to the UBA/TS-N domain (Ubiquitin associated/TS-N domain) at the C-terminal of the protein sequences.

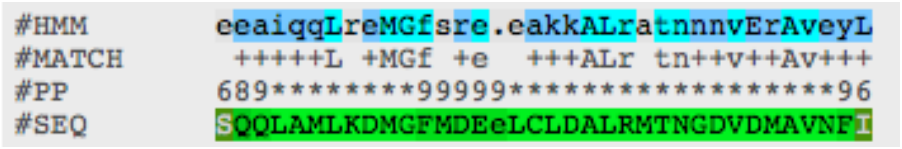


Figure 4.8 Tb11.47.0019 long and short protein sequences contain a predicted UBA/TS-N domain. The Pfam 24.0 database (http://pfam.sanger.ac.uk) sequence searches predicted that the Tb11.47.0019 protein sequence contains a UBA/TS-N domain (ubiquitin associated/TS-N domain), which is present in both the long (487 amino acids) and short (360 amino acids) form protein sequences at amino acids 446-481 (E-value: 2.4e-09) and 319-354 (E-value: 1.6e-09) respectively. #HMM: consensus of the HMM (hidden Markov model); #MATCH: the match between the query sequence and the HMM; #PP: posterior probability, or the degree of confidence in each individual aligned residue; #SEQ: query sequence, coloured according to the posterior probability (a colour scale of 0-100% where 0% is red and 100% is green).

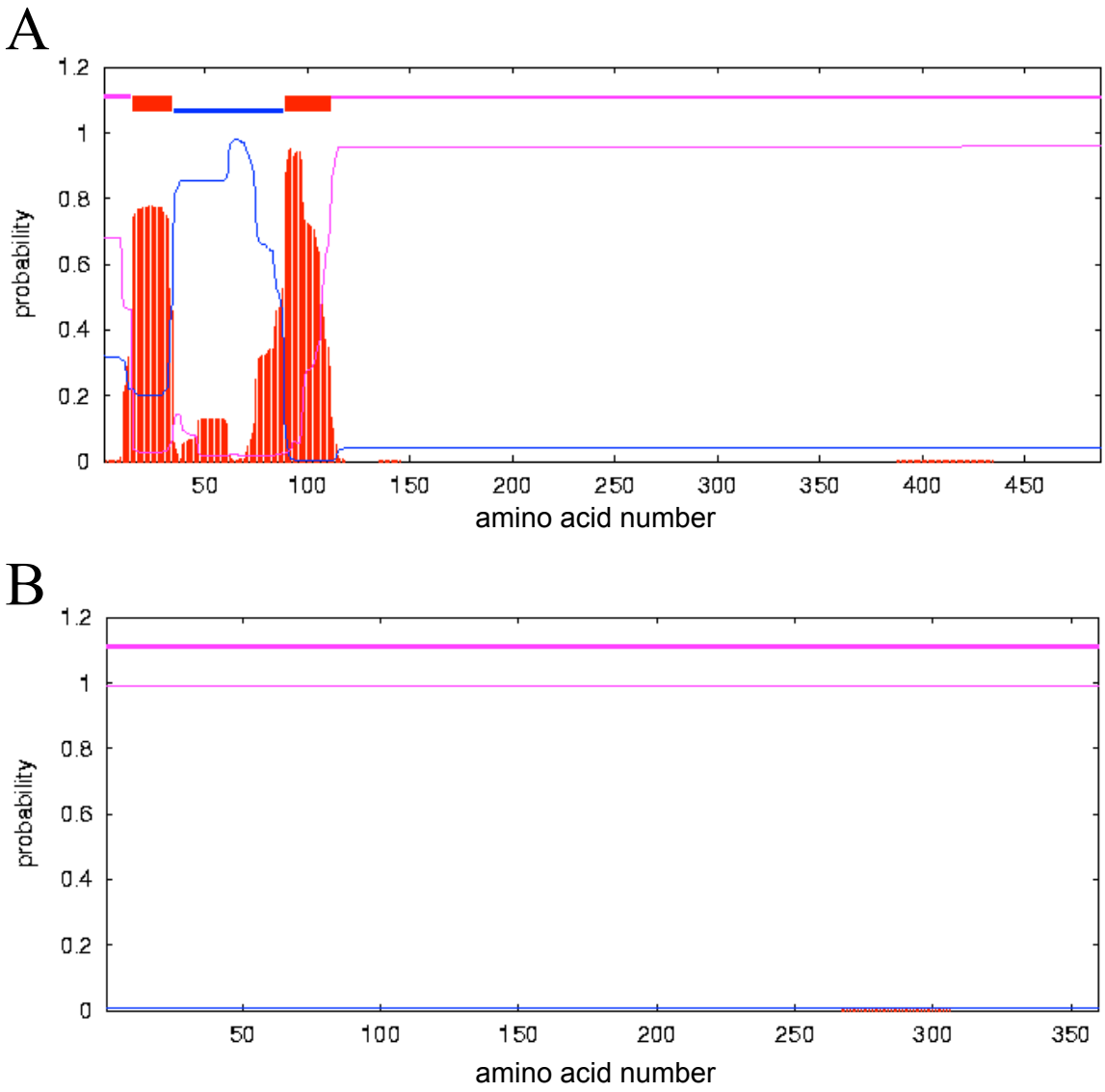


Figure 4.9 Prediction of transmembrane helix occurrence in long and short Tb11.47.0019 protein sequences. The graphical plot of posterior probabilities from the output generated by the transmembrane helice prediction tool, TMHMM2.0 (http://www.cbs.dtu.dk/services/TMHMM/), using either the long (A) or short (B) peptide sequences for Tb11.47.0019 as input; red: transmembrane; blue: inside; pink: outside; at the top of each plot (between 1 and 1.2) the N-best prediction is shown.

Figure 4.8 shows the output returned from the searches (both long and short sequences produced the same result). For the long protein form the E-value of the match was 2.4e-09 and for the short form was 1.6e-09, despite the two sequences only differing in the presence or absence of 127 amino acids at the N-terminal.

However, the posterior probability for each of the individual residues of the predicted domain are the same for both and are very high for most residues except the residues at each end of the potential domain.

Additionally, a programme for the prediction of protein transmembrane helices that is based on a hidden Markov model, TMHMM2.0 (http://www.cbs.dtu.dk/services/TMHMM/), was used. The results for both the long and short protein sequences are shown in Figure 4.9 (A and B respectively). For the long protein sequence it was predicted that there are two transmembrane helices near the N-terminus, with the N- and C-termini protein regions present on the non-cytoplasmic side of the membrane. However, this programme should not be used to predict protein location. This programme also noted that there is a possible N-terminal signal sequence and that predicted N-terminal region transmembrane helices often turn out to be signal peptides. For the short protein sequence, no transmembrane helices were predicted to be present, highlighting a potentially important difference between the forms.

To investigate which known proteins each sequence might show similarity to, the protein-protein BLAST (blastp) programme (http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins) was used with the default settings and each amino acid sequence as input. The top 20 protein sequences identified with the highest percentage identity match to either the long or short protein form sequences are shown in Table 4.4. The most similar proteins identified for both protein sequences were the same hypothetical proteins within the 'TriTryps' (*T. brucei*, *T. cruzi*, and *Leishmania* sp.s). The other proteins identified were mostly hypothetical proteins in other species or proteins that from the gene annotations appear to have a role with ubiquitin-related processes. This is supportive of a role in ubiquitin-related processes for both the long and short protein forms, and is linked to the identification of the potential UBA/TS-N domain in each protein sequence.

Table 4.4 Proteins (top 20) with the highest percentage identity match to the long or short protein form sequences using protein-protein BLAST (blastp; http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins) analysis.

Accession	Description	E value	% identity
Long protein form			
XP_828221.1	hypothetical protein [Trypanosoma brucei TREU927]	0	100
CBH17021.1	hypothetical protein, conserved [Trypanosoma brucei gambiense DAL972]	0	98.7
XP_810564.1	hypothetical protein [Trypanosoma cruzi strain CL Brener]	7.00E-73	43.19
XP_815088.1	hypothetical protein [Trypanosoma cruzi strain CL Brener]	1.00E-72	43.42
EFZ26585.1	hypothetical protein TCSYLVIO_7236 [Trypanosoma cruzi]	3.00E-72	42.45
XP_001565839.1	hypothetical protein [Leishmania braziliensis MHOM/BR/75/M2904]	1.00E-31	26.65
CAM45357.2	conserved hypothetical protein [Leishmania braziliensis MHOM/BR/75/M2904]	1.00E-29	26.47
CBZ35216.1	unnamed protein product [Leishmania donovani BPK282A1]	4.00E-28	32.33
XP_001467363.1	hypothetical protein [Leishmania infantum JPCM5]	1.00E-27	47.06
XP_848026.1	hypothetical protein [Leishmania major strain Friedlin]	2.00E-27	31.85
CBZ28167.1	conserved hypothetical protein [Leishmania mexicana MHOM/GT/2001/U1103]	1.00E-26	47.06
XP_003148386.1	ubiquitin family protein [Loa loa]	1.00E-12	29.85
XP_001902395.1	Ubiquitin family protein [Brugia malayi]	7.00E-11	33.75
EGD72894.1	hypothetical protein PTSG_04623 [Salpingoeca sp. ATCC 50818]	2.00E-09	29.7
NP_740884.1	hypothetical protein F15C11.2 [Caenorhabditis elegans]	4.00E-09	48.65
NP_740885.1	hypothetical protein F15C11.2 [Caenorhabditis elegans]	5.00E-09	26.02
NP_740883.1	hypothetical protein F15C11.2 [Caenorhabditis elegans]	6.00E-09	25.5
AAF80171.1	A1U [Homo sapiens]	9.00E-09	24.41
XP_002715368.1	PREDICTED: ataxin-1 ubiquitin-like interacting protein [Oryctolagus cuniculus]	1.00E-08	29.63
EFB18473.1	hypothetical protein PANDA_016682 [Ailuropoda melanoleuca]	1.00E-08	28.65
Short protein form			
XP_828221.1	hypothetical protein [Trypanosoma brucei TREU927]	0	100
CBH17021.1	hypothetical protein, conserved [Trypanosoma brucei gambiense DAL972]	0	99.17
XP_810564.1	hypothetical protein [Trypanosoma cruzi strain CL Brener]	5.00E-73	43.19
XP_815088.1	hypothetical protein [Trypanosoma cruzi strain CL Brener]	6.00E-73	43.42
EFZ26585.1	hypothetical protein TCSYLVIO_7236 [Trypanosoma cruzi]	1.00E-72	41.41
XP_001565839.1	hypothetical protein [Leishmania braziliensis MHOM/BR/75/M2904]	3.00E-32	26.65
CAM45357.2	conserved hypothetical protein [Leishmania braziliensis MHOM/BR/75/M2904]	5.00E-30	25.74
CBZ35216.1	unnamed protein product [Leishmania donovani BPK282A1]	7.00E-29	31.58
XP_001467363.1	hypothetical protein [Leishmania infantum JPCM5]	2.00E-28	47.06
XP_848026.1	hypothetical protein [Leishmania major strain Friedlin]	4.00E-28	31.2
CBZ28167.1	conserved hypothetical protein [Leishmania mexicana MHOM/GT/2001/U1103]	2.00E-27	47.06
XP_003148386.1	ubiquitin family protein [Loa loa]	2.00E-13	29.85
XP_001902395.1	Ubiquitin family protein [Brugia malayi]	3.00E-11	33.75
EGD72894.1	hypothetical protein PTSG_04623 [Salpingoeca sp. ATCC 50818]	1.00E-09	29.7
AAF80171.1	A1U [Homo sapiens]	4.00E-09	50
XP_001163250.1	PREDICTED: ubiquilin-4 isoform 3 [Pan troglodytes]	4.00E-09	26.02
NP_740883.1	hypothetical protein F15C11.2 [Caenorhabditis elegans]	4.00E-09	25.44
NP_740884.1	hypothetical protein F15C11.2 [Caenorhabditis elegans]	4.00E-09	24.41
BAG64068.1	unnamed protein product [Homo sapiens]	5.00E-09	25
XP_003355179.1	PREDICTED: ubiquilin-4 isoform 2 [Sus scrofa]	5.00E-09	37.25

```
XP_828221.1      MCVCLCCQSFIGKWVVVYPREGLLLLLLLSYWGRSQRIEFFLKEKMVANLGNILFLSIS 60
CBH17021.1       ------------------------------------------------------------
XP_810564.1      ------------------------------------------------------------


XP_828221.1      LKRKKRYYNNISARVYIYMYITSNVPSRFVLLNLYIYLSVLFFMPYPTSYHIALCQATKR 120
CBH17021.1       -------------------------------------------MPYPTSYHIALCQATKR 17
XP_810564.1      ------------------------------------------------------------


XP_828221.1      KERGSLRMNVFDSNSMLCQVVSNITDPVELFIEPEMTVGEVRALVAEMTVDSNDWNVRLT 180
CBH17021.1       KERGWLGMNVFDSNSMLCQVVSNITDPVELFIEPEMTVGEVRALVAEMTVDSNDWNVRLT 77
XP_810564.1      ---------------MLFQIVSNVADKVDLFIEDEMTVGEVRVLIGQM-IDADEKTIRLT 44
                               ** *:***::* *:**** ********.*:.:* :*::: .:***


XP_828221.1      FQGHTLEDDEETWSSVTQQYPVAQGVAPKLFSHASKKR-------EEKARRPVSLPLME- 232
CBH17021.1       FQGHTLEDDEETWSSVTQQYPVAQGVAPKLFSHASKKR-------EEKARPPVSLPLME- 129
XP_810564.1      FQGHVLEDDKATWELILRMYPLAEDTPRKLFAHVSERLRTGSASMEALRAALVSASGQED 104
                 ****.****: **. : : **:*:... ***:*.*::        *      ** .  *


XP_828221.1      EQKDRQQAKLMEPIIDTLVNDPNFSEMMISGQPSLRRLIDANPEVGRLFRNPETLKSLIM 292
CBH17021.1       EQKDRQQAKLMEPIIDTLVNDPNFSEMMISGQPSLRRLIDANPEVGRLFRNPETLKSLIM 189
XP_810564.1      EKAARVQARLMEPVIDLMVKDPSMLEMMLSSHPGLKQMLEKHPELKSHVCDPETLKTLMM 164
                 *:  * **:****:** :*:**.: ***:*.:*.*::::: :**:   . :*****:*:*


XP_828221.1      SQIDPDQRRAMNRNVQLQLAQISAIPGGEQLLERYTSGLMDDMDLNLDLVRNTKIKVPED 352
CBH17021.1       SQIDPDQRRAMNRNVQLQLAQISAIPGGEQLLERYTSGLMDDMDLNLDLVRNTKTKVPED 249
XP_810564.1      SQIDPDQRRSLNRGMGLQLAQLSALPGGEQLFEHLTSEYLNDMAA-LTETNVSSSSDNVD 223
                 *********::**.: *****:**:******:*: **  ::**   *  .. :. .   *


XP_828221.1      EKDDNSESVRGVNSEALPNPWSQSTNTAD--------SGNTN----TFS-----NAAMNL 395
CBH17021.1       EKDDNSESVRGVNSEALPNPWSQSTNTAD--------SGNTN----TFS-----NAAMNL 292
XP_810564.1      EAQARPDPTKEANSEALPNPWERQSSSLQGASGLNSMNGNTNNLFSAFGGLGAFGGSMPF 283
                 * : ..:..: .*********.:.:.: :        .****    :*.     ..:* :


XP_828221.1      FGIPGNVGHPSLGN-LFGSGLGGP--FPFMSPGISIPQTGVTASQRGDDAT-RWSQQLAM 451
CBH17021.1       FGIPGNVGHPSLGN-LFGSGLGGP--FPFMSPGISIPQTGVTASQRGDDAT-RWSQQLAM 348
XP_810564.1      AGLPGNIASLGHTNPVTAASFLPPRMLPSSHEGNHVPTTDPTHTREASANTGEWASQLAT 343
                  *:***:.  .  * : .:.:  *  :*    *  :* *. * ::...  * .*:.***


XP_828221.1      LKDMGFMDEELCLDALRMTNGDVDMAVNFIVNKDSS 487
CBH17021.1       LKDMGFMDEELCLEALRMTNGDVDMAVNFIVNKDSS 384
XP_810564.1      LREMGFDNEALCLEALRASNGDIDGAVNYIADKDK- 378
                 *::*** :* ***:*** :***:* ***:*.:**.
```

Figure 4.10 ClustalW2 (http://www.ebi.ac.uk/Tools/msa/clustalw2/) multiple alignment using the default settings with the protein sequence for Tb11.47.0019 (accession: XP_828221.1) and two protein sequences in *T. brucei gambiense* (accession: CBH17021.1) and *T. cruzi* (accession: XP_001565839.1) found to be similar by BLASTP analysis. Asterisk (*): single, fully conserved residue; colon (:): conservation between groups of strongly similar properties; period (.): conservation between groups of weakly similar properties. The potential UBA/TS-N domain identified by Pfam search (Figure 4.8) is shown underlined; the downstream methionine in the sequence produced from usage of the downstream SAS is shown in grey highlight.

Following this, the annotated protein sequence for Tb11.47.0019 (the long protein form, accession: XP_828221.1) was used in a multiple alignment with the two protein sequences showing the greatest sequence identity, hypothetical proteins from *T. brucei gambiense* (accession: CBH17021.1) and *T. cruzi* (accession: XP_001565839.1). This was performed with the ClustalW2 programme

(http://www.ebi.ac.uk/Tools/msa/clustalw2/) with the default settings and the alignment is shown in Figure 4.10. These protein sequences demonstrate a reasonably high degree of similarity in protein sequence. Some regions of sequence show a higher degree of conservation, such as the sequence identified as a potential UBA/TS-N domain in the Pfam searches (see Figure 4.8; underlined in Figure 4.10).

Interestingly, neither of the other protein sequences used for alignment was found to possess similarity to most of the N-terminal region of Tb11.47.0019 that is missing in the short form of the protein (the methionine residue produced from usage of the downstream SAS is shown in grey highlight in Figure 4.10). This was also found to be true in alignments using additional protein sequences identified by blastp analysis to show similarity (accessions: XP_001565839.1, AAF80171.1, XP_002810087.1, XP_001902395.1; data not shown). This could suggest that the N-terminal extension in Tb11.47.0019 is specific to this *T. b. brucei* gene. The locations of the domains predicted by the bioinformatic analyses for each protein sequence is shown in a schematic diagram in Figure 4.11.

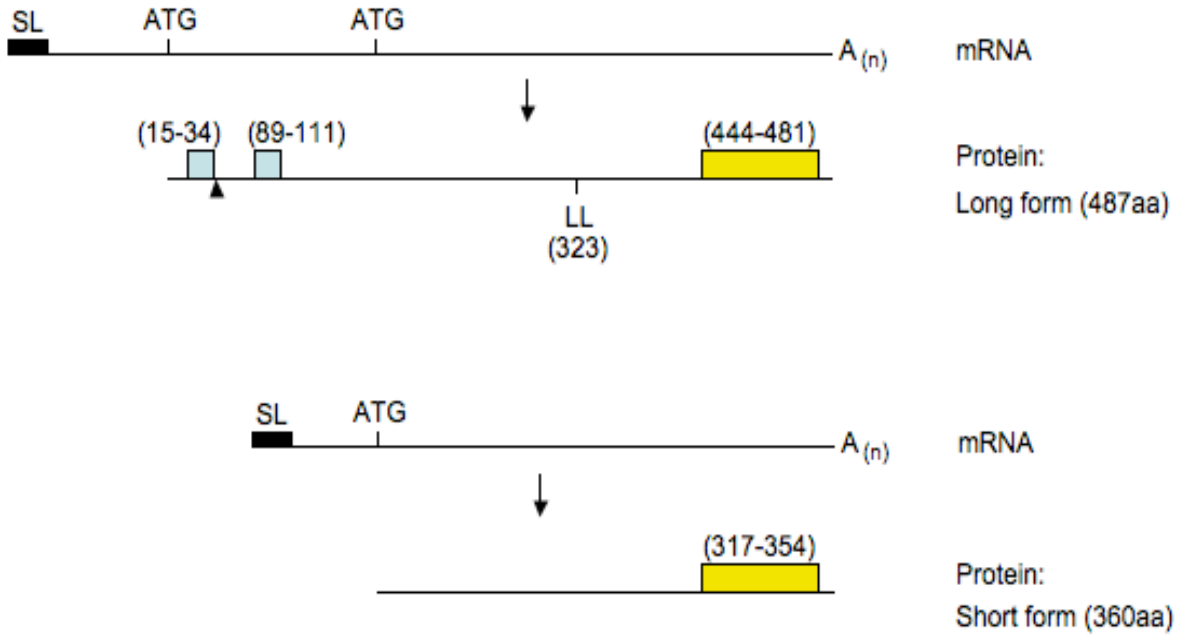


Figure 4.11 Schematic diagram of the generation of long and short protein forms from two SAS locations identified for Tb11.47.0019 with the locations of regions predicted by bioinformatic analyses. The arrowhead marks the position of putative signal peptide cleavage site in between amino acids 36 and 37 (probability: 0.798; SignalP). Blue boxes indicate locations of transmembrane helices (although the one nearest the N-terminal is likely to be a signal peptide); yellow boxes indicate location of the UBA/TS-N domain; dileucine motif is marked by 'LL'; numbers in parentheses refer to the amino acid numbers of the start-end of the domain. SL: splice leader; ATG: start codon. Drawing is not to scale.

The presence or absence of predicted transmembrane helices in the long or short protein forms identified above suggest that these proteins may have different cellular localisations. Therefore, additional bioinformatic analyses were also performed using each of the long and short form protein sequences to attempt to predict the cellular localisation for each protein. The tools used were the SignalP (www.cbs.dtu.dk/services/SignalP), iPSORT, WoLF PSORT, PSORTII (each through http://psort.hgc.jp/), SecretomeP (http://www.cbs.dtu.dk/services/SecretomeP/), and TargetP (http://www.cbs.dtu.dk/services/TargetP/) programmes, the results of which are summarised in Table 4.5.

As can be seen, very different predictions were made for each of the protein sequences. The long protein form was predicted to contain a cleavage signal peptide sequence, whereas the short protein form was not (SignalP, iPSORT and WoLF PSORT predictions). The long protein form was predicted to possess one transmembrane helix in addition to the signal peptide (PSORTII) as suggested by the TMHMM2.0 prediction above. This meant that the long protein sequence was predicted to have a membrane topology of Type 1a, where the N-terminus is present in the exoplasmic space and the C-terminus is cytosolic.

Neither protein was predicted to be mitochondrial by iPSORT. The long protein form was predicted to be either present in the plasma membrane, endoplasmic reticulum, or extracellular environment by WoLF PSORT analysis; cytoplasmic rather than nuclear, and present in the Golgi, plasma membrane, or endoplasmic reticulum by PSORTII; and to be part of the secretory pathway (SP), rather than to have a mitochondrial targeting peptide (mTP) or other location by TargetP, with high reliability (reliability class, RC, of 1; scale of 1 to 5, where 1 is the strongest prediction). SecretomeP predicted that this protein possibly enters the classical secretory pathway, due to the presence of the signal peptide predicted by SignalP and an NN-score of greater than 0.5.

Table 4.5 Summary of bioinformatic predictions[a] for the long and short peptide sequences of Tb11.47.0019.

Bioinformatic software programme	Long peptide sequence	Short peptide sequence
Pfam	UBA/TS-N domain (2.4e-09)	UBA/TS-N domain (1.6e-09)
SignalP	Signal peptide: YES	Signal peptide: NO
	Prediction: Signal peptide	Prediction: Non-secretory protein
	Signal peptide probability: 0.842	Signal peptide probability: 0.001
	Max. cleavage site probability: 0.798 between positions 36 & 37	Max. cleavage site probability: 0.001
iPSORT	Signal peptide: Yes (1.61333)	Signal peptide: No (0.60667)
	Mitochondrial: No (0.03333)	Mitochondrial: No (-0.06667)
WoLF PSORT	Plasma membrane: 18.0	Cytoskeleton: 19.0
	Endoplasmic reticulum: 8.0	Cytosol: 7.0
	Extracellular: 4.0	Cytosol/nuclear: 7.0
		Nuclear: 5.0
PSORTII	Cleavable signal peptide (1 to 36)	No N-terminal signal peptide
	Number of TMS(s) predicted: 1	Number of TMS(s) predicted: 0
	Membrane topology: Type 1a (cytoplasmic tail 106 to 487)	
	Dileucine motif in the tail: LL at 323	Dileucine motif in the tail: none
	Cytoplasmic/nuclear discrimination: cytoplasmic (reliability: 55.5)	Cytoplasmic/nuclear discrimination: nuclear (reliability: 70.6)
	k-NN prediction:	k-NN prediction:
	33.3%: Golgi	69.6%: nuclear
	33.3%: plasma membrane	21.7%: mitochondrial
	33.3%: endoplasmic reticulum	4.3%: cytoplasmic
		4.3%: cytoskeletal
Secretome P	NN-score: 0.837	NN-score: 0.384
TargetP	mTP: 0.014	mTP: 0.063
	SP: 0.957	SP: 0.136
	other: 0.114	other: 0.863
	RC: 1	RC: 2

[a] The bioinformatic tools listed were accessed through the following URLs: http://pfam.sanger.ac.uk; www.cbs.dtu.dk/services/SignalP; http://psort.hgc.jp/; http://www.cbs.dtu.dk/services/SecretomeP/; http://www.cbs.dtu.dk/services/TargetP/. UBA/TS-N: ubiquitin associated domain; TMS: transmembrane sequence; NN-score: a value greater than 0.5 predicts that the protein enters the secretory pathway; mTP: mitochondrial targeting peptide; SP: secretory pathway; other: other location; RC: reliability class, scale of 1 to 5 where 1 is the strongest prediction.

The short protein form was predicted to be either present in the cytoskeleton, cytosol, or have a nuclear location by WoLF PSORT analysis; nuclear rather than cytoplasmic, and to be either nuclear, mitochondrial, cytoplasmic, or cytoskeletal by PSORTII; and to have another location other than secretory or mitochondrial by

TargetP (RC: 2). SecretomeP predicted that this protein would not enter the secretory pathway, due to an NN-score of less than 0.5.

Interestingly, a dileucine (LL) motif was identified in the long protein form sequence but not in the short protein form sequence (PSORTII), despite the sequence being present in this sequence too. This is interesting because these motifs exhibit a wide range of functions in endocytosis and membrane receptor/protein trafficking (reviewed by (Pandey, 2010). The dileucine residues of the identified motif are preceded by a polar residue (Q) and a negatively charged amino acid (E), which is characteristic of these motifs (reviewed by Pandey, 2010). However, why this motif was not detected in the short protein form, as the motif sequence is also present here, is unknown.

These bioinformatic prediction tools, however, are not optimised for prediction with trypanosome protein sequences and so these predictions could be incorrect. Nonetheless, it does appear that the consensus of the predictions is that these proteins would have different cellular localisations to each other.

4.9 Do ectopically expressed long and short Tb11.47.0019 peptides have a different cellular localisation?

Due to the potential for different cellular localisations predicted for each of the long and short protein forms of Tb11.47.0019, the location of each of these proteins was investigated. This involved expression of either the long or short as an ectopic tagged copy.

4.9.1 Generation of long and short tagged-peptide expressing cell lines

For expression of either protein form, the plasmid, pHD451 was used. Sub-cloning of the DNA sequences for the protein forms is shown in the schematic diagram in Figure 4.12 (A). Either sequence was inserted into a region immediately preceding the sequence encoding a Ty epitope tag followed by a stop codon. Upstream of this region in the plasmid is the tetracycline repressor sequence containing an EP-

procyclin promoter. Therefore, expression of the protein with a Ty epitope tag is inducible with addition of tetracycline. The plasmid also contains a ribosomal spacer sequence, such that linearization of the plasmid with *Not*I restriction enzyme digestion allows integration into the *T. brucei* genome in this region. The plasmid possesses a hygromycin resistance gene to allow selection of cells that have successful integration of the plasmid.

For sub-cloning of the long protein sequence, the *Xho*I and *Bam*HI restriction enzyme sites of pHD451 were used, as this sequence contains the *Hind*III restriction site sequence. Primers 64 and 65 were used to amplify the sequence of Tb11.47.0019 from the annotated start codon to the nucleotide preceding the annotated stop codon (excluding the stop codon sequence allowed incorporation of the Ty epitope tag at the C-terminus). There were difficulties in obtaining the long sequence for Tb11.47.0019 as PCR amplifications always resulted in amplicons containing a G residue at the same position as a C residue in the amplicon for the short protein sequence and the *T. brucei* TREU 927 genome annotation. This would mean that the encoded protein would have a glutamic acid residue here instead of an aspartic acid residue as in the short protein form. Even though this is a conservative amino acid substitution, because these constructs were to be used for assessment of cellular localisation it was important that the only difference in the sequences was the N-terminal extension in the long protein form. However, repeated PCRs produced the same outcome and it was discovered from genome analysis that the Tb11.47.0019 gene is G/C polymorphic at this residue (Dr Paul Capewell, University of Edinburgh). Of 16 sequencing reads for the *T. brucei* TREU 927 genome, seven contained C at this position and nine contained G; in contrast the strain STIB247 was homozygous for G and the STIB386 strain was heterozygous for G/C. Therefore, it was assumed that this residue difference would not affect the function or localisation of the encoded proteins.

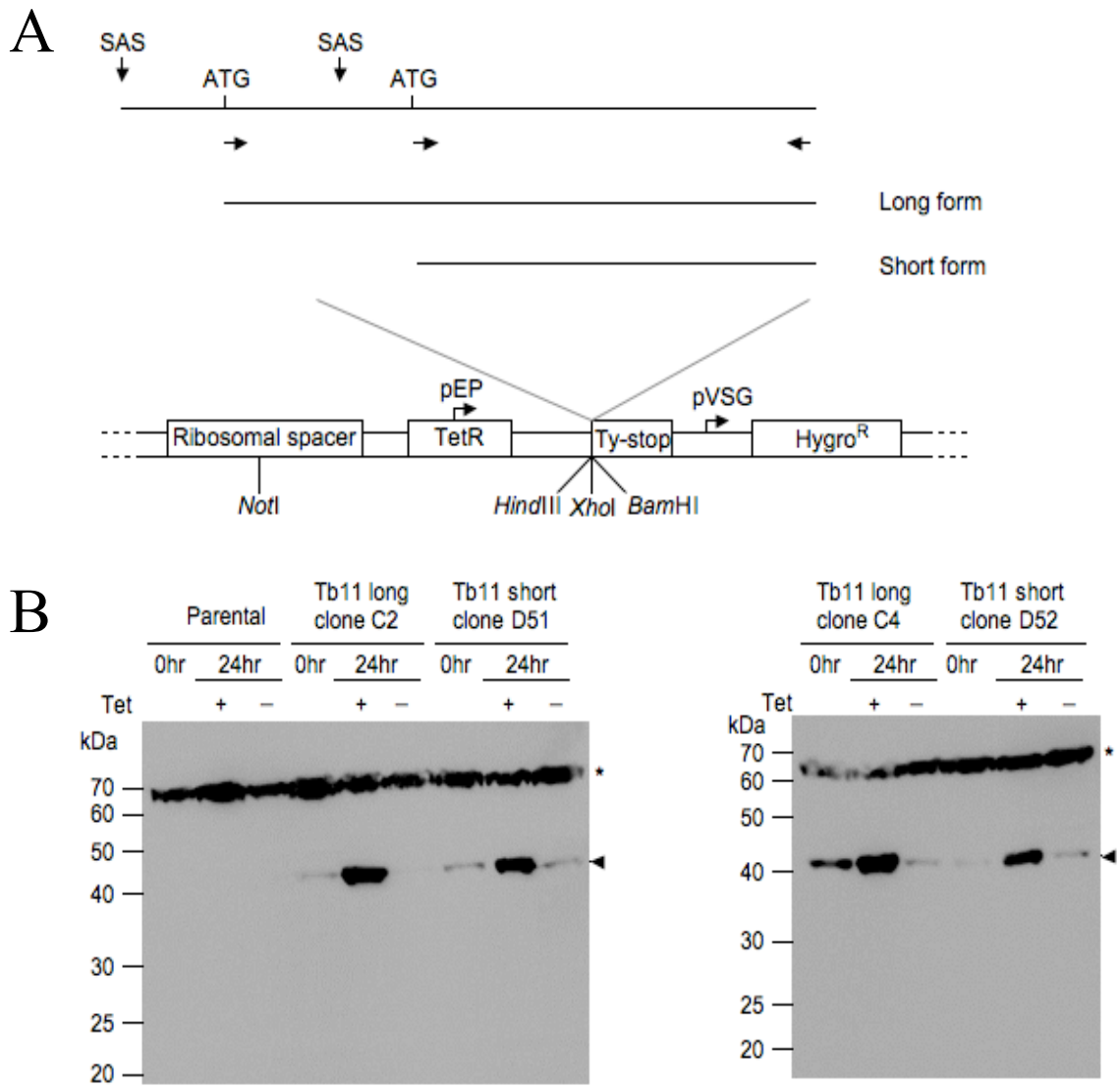


Figure 4.12 Ectopic overexpression of each of the long and short peptide forms of Tb11.47.0019 with a Ty epitope tag. A. Schematic diagram of construct generation using the expression vector, pHD451-Ty. Either the long or short form sequence was amplified by PCR (arrows depict primer binding regions) for ligation into pHD451-Ty through the *Xho*I/*Bam*HI or *Hind*III/*Bam*HI restriction enzyme sites (long or short form, respectively). Shown at the bottom is a representation of the pHD451-Ty vector showing the ribosomal spacer, tetracycline repressor (TetR), Ty epitope with stop codon (Ty-stop), hygromycin resistance gene (Hygro^R), EP-procyclin promoter (pEP), VSG promoter (pVSG), and restriction enzyme site locations. B. Western blot analyses using a primary antibody that recognises the Ty epitope (BB2) and a secondary antibody of anti-mouse IgG with a horse radish peroxidase (HRP) conjugate. Cell lysate samples were either from the parental cell line, or cells ectopically expressing the long form of Tb11 (clones C2 and C4) or ectopically expressing the short form of Tb11 (clones D51 and D52) at the times indicated (hr: hours after induction). The arrowhead marks the ectopically expressed peptide; the asterisk marks a cross-reactive band. Protein size is indicated in kiloDalton (kDa).

For sub-cloning of the short protein sequence, the *Hind*III and *Bam*HI restriction enzyme sites of pHD451 were used. Primers 66 and 65 were used to amplify the sequence from the start codon at position +378nt relative to the annotated start codon, to the nucleotide preceding the annotated stop codon, as for the long sequence above. Both of the pHD451 plasmids containing either the sequence for the long or short protein sequences were then transfected into *T. b. brucei* bloodstream form strain Lister 427-449 (meaning that these cells express the tetracycline repressor protein from transfection with pHD449 plasmid). Successful transfectants were selected with 2 μg/ml hygromycin B and two independently derived clones were assessed for each transfected construct.

4.9.2 Inducible ectopic expression of long and short form proteins

The cell lines obtained were tested for the ectopic overexpression of the tagged-proteins. Cell lysate samples were made from each of the cell lines (and the parental 427 449 cell line, as a control), then they were either induced for overexpression with tetracycline or uninduced and further cell lysate samples were made 24 hours later. These cell lysate samples were then probed by western blot analysis (see Section 2.7) with a primary antibody that recognises the Ty epitope tag (BB2) and a secondary antibody of anti-mouse IgG with a horse radish peroxidase (HRP) conjugate. The results of these western blot analyses are shown in Figure 4.12 (B).

All the cell lines tested inducibly overexpressed an ectopic tagged-protein copy (marked by arrowheads in Figure 4.12 B). Although, for each of the cell lines there was expression of the tagged proteins in the absence of induction (see the 0hr and – tet lanes), this was at a much lower expression level than when induced with tetracycline (see + tet lanes). As expected, there was no tagged protein expressed in the parental cell line. These antibodies also detected a higher molecular weight band of ~60-70 kDa. This however, is a cross-reactive band, being present in all the cell lysate samples tested. Thus, it serves as a useful loading control.

Interestingly, it was observed from the western blot analyses that the ectopically expressed proteins were present at the same molecular weight, despite transfection

with the sequences encoding the long or short protein forms. The length of the long protein sequence is 487 amino acids (excluding the length of the Ty epitope tag) and the short protein sequence is 360 amino acids long (again excluding the Ty epitope tag). The predicted protein molecular weight on the TriTryp database is 54 925 Da for the annotated Tb11.47.0019 gene (the long protein form). For the short protein form the molecular weight was determined as 39 870 Da (http://web.expasy.org/compute_pi/). The observed band is slightly larger than 40 kDa and so is consistent with this being the short protein form. The slightly larger weight than estimated is contributed by the Ty epitope tag (10 amino acids; estimated molecular weight of 1 167 Da).

4.9.3 Assessment of cellular localisation for long and short tagged-peptides

When cell lysate samples were taken at 24 hours after induction with tetracycline (or from uninduced controls) for western blot analyses above, cells were also fixed for protein localisation analysis. For this the cells were fixed in paraformaldehyde on polylysine slides and stained with the BB2 antibody, then anti-mouse IgG labelled with FITC, and finally the DNA stain, DAPI (see Section 2.8.2). The slides were then analysed by fluorescence microscopy, representative results of one experimental replicate of this immunofluorescence analysis being shown in Figure 4.13.

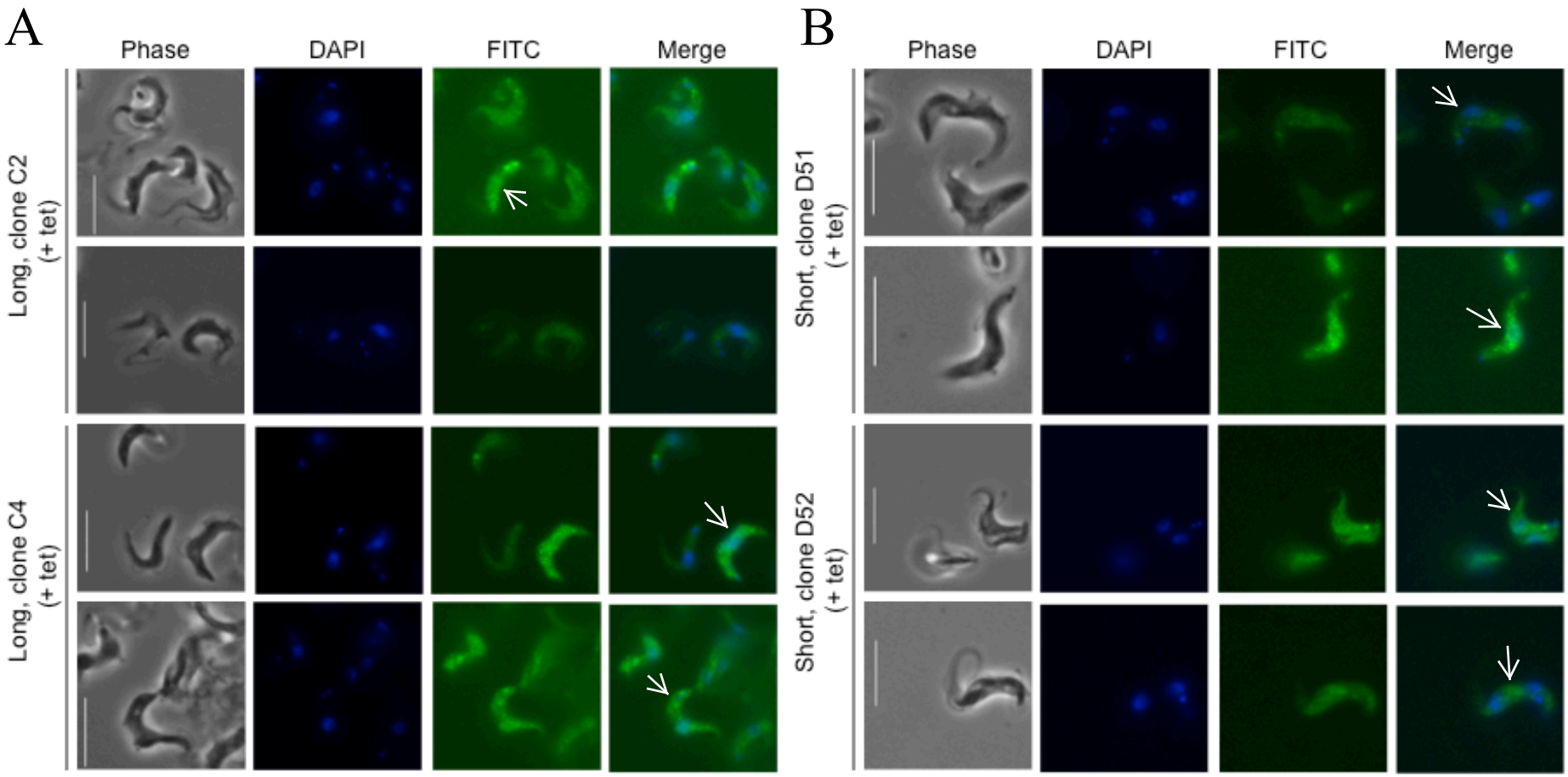


Figure 4.13 continues on the following page

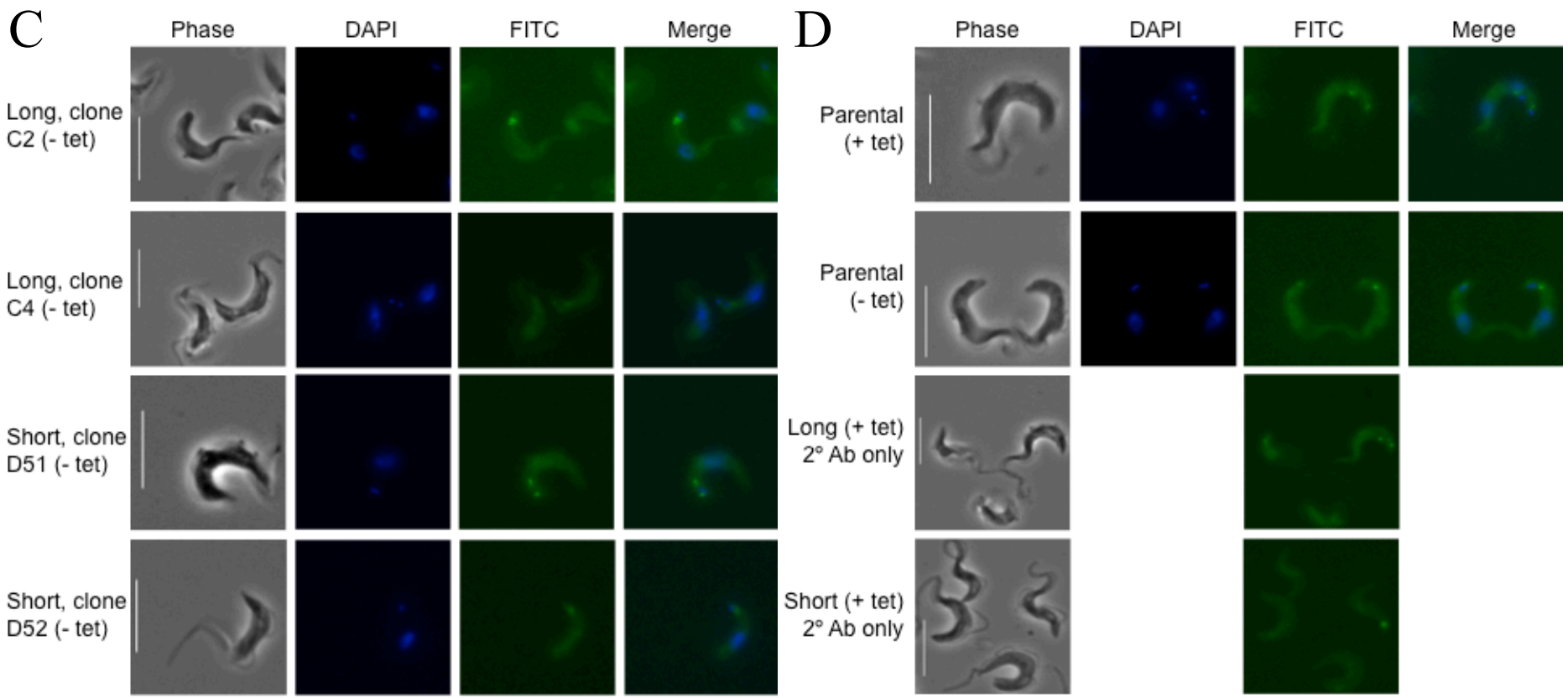


Figure 4.13 Immunofluorescence analysis. Cells were fixed in paraformaldehyde and stained with a primary antibody that recognises the Ty epitope tag (BB2), a secondary antibody (anti-mouse IgG labelled with FITC) and the DNA stain, DAPI. Micrographs show representative cells from two clones induced to ectopically express the long peptide form of Tb11.47.0019 (+ tet; A), two clones induced to ectopically express the short peptide form of Tb11.47.0019 (+ tet; B), or these same cells uninduced (- tet; C). Arrows indicate cells exhibiting more intense levels of FITC staining. Controls (D): the parental cell line treated or untreated with tetracycline; and cells induced to ectopically express the long or short peptide forms, but stained with secondary (2°) antibody only. Scale bar: 10 μm.

Figure 4.13 (A) shows four sets of micrographs obtained from cells induced to express the long protein form. In these cell populations there was only a low frequency of cells with bright FITC staining, most had a low level of staining, which through comparison with the controls was shown to be background staining (D). These cells with lower levels of FITC staining can be seen in these micrographs. For the cells that had more intense levels of FITC staining (indicated with arrows in Figure 4.13 A), the staining was observed to produce a punctate pattern throughout the cytoplasm. With the use of these antibodies often one or two intense spots of FITC staining were observed at the posterior end in close proximity to the DAPI-staining of the kinetoplast. However, this was not specific to the expression of either protein form as the control cells also exhibited similar staining (see below). The punctate staining pattern did not appear to be excluded from the area of DAPI staining of the nucleus. Hence, either the tagged protein is present in both the cytoplasm and nucleus, or it is present on/near the cell surface, which would mean that it would appear to be present in the cytoplasm and nucleus by this methodology. Investigation of cellular localisation with these antibodies but using confocal microscopy could address this issue.

Figure 4.13 (B) shows four sets of micrographs obtained from cells induced to express the short protein form. As can be seen, with comparison to the micrographs shown in Figure 4.13 (A), there was an identical pattern of FITC staining in this group of cells as in the cells induced to express the long protein form.

Shown in Figure 4.13 (C) are four sets of micrographs obtained from one of each of the cell lines shown in A and B, but not induced for expression of the tagged protein copy. These cells exhibited a low level of fluorescent staining, consistent with the low level of tagged protein expression in these uninduced cells (Figure 4.12 B). However, this staining was only slightly more intense than that observed as the background level of staining in the parental cell line, with or without tetracycline (Figure 4.13 D).

Also shown in Figure 4.13 (D) is the non-specific FITC staining of cells induced to express either tagged copy when the secondary antibody was used alone. Under these conditions, the same low level of background fluorescence staining is observed similar to that observed when both antibodies were used, and often with the same one or two bright fluorescent spots of signal at the posterior cell end mentioned above. This indicates that this signal is the result of unspecific binding of the secondary antibody. Although, the parental cell line and analysis with secondary antibody only produced a low level of background fluorescence staining in the cells, the same punctate pattern of more intense fluorescence staining seen for the induced cells was not observed.

In conclusion, both the cells induced to express tagged-copies of the long and short protein forms produced the same pattern of protein expression in the analysis shown in Figure 4.13. This is consistent with the observation that in both groups of cells the signal detected by western blot analysis of the tagged protein copies was of the same molecular size (Figure 4.12 B), and so these cells appear to express a tagged protein of the same molecular weight and localisation.

4.10 RNAi-mediated ablation of the long Tb11.47.0019 transcript or both transcripts

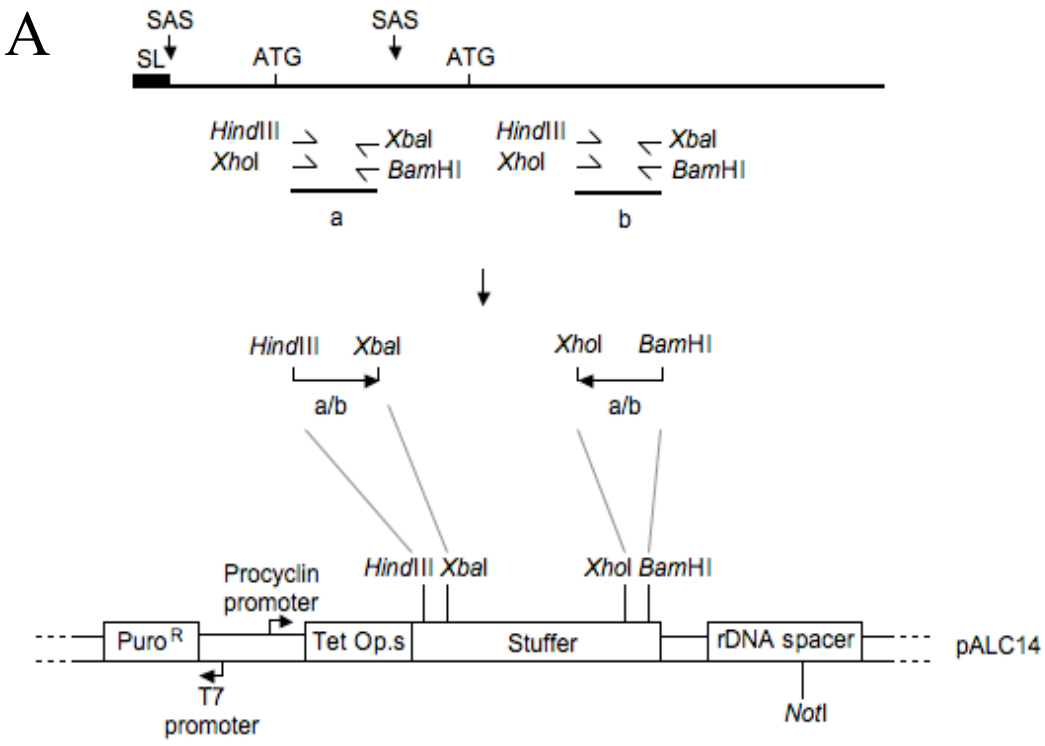
It was identified that there could potentially be differential expression across the lifecycle of the Tb11.47.0019 transcripts produced from trans-splicing at different locations (see Section 4.7.2 and Figure 4.7), generating a long and a short form of the transcript and, consequently, protein. Due to the potentially increased expression of the long form of the transcript in stumpy forms, it would be interesting to investigate if there was a functional consequence of this for differentiation. To study this, constructs for the RNAi-mediated ablation of transcripts containing either the ‘a’ or ‘b’ regions of Tb11.47.0019 (described in Section 4.7) were generated. This would mean that only the long transcript form would be ablated upon RNAi involving the ‘a’ region and both the long and short transcript forms would be ablated upon RNAi involving the ‘b’ region. It is not possible to ablate the short transcript form only,

because all the sequence should be identical to the long transcript form. Following construct generation and the creation of transfectant cell lines, the ability of cells to differentiate to procyclic forms could be tested after induction of RNAi.

4.9.1 Generation of RNAi cell lines

For generation of RNAi cell lines, the plasmid pALC14 (Bochud-Allemann and Schneider, 2002) was used. This plasmid contains a 'stuffer' region possessing *Hind*III, *Xba*I, *Xho*I, and *Bam*HI restriction enzyme site sequences, downstream of a procyclin promoter and tetracycline operator sequences. A schematic diagram of the sub-cloning steps into this vector is shown in Figure 4.14 (A). Insertion of the same gene fragments in opposite orientations can be achieved through amplication of the same DNA sequence with primers containing *Hind*III and *Xba*I (forward and reverse primers, respectively) restriction site sequences, and another amplification with primers containing *Xho*I and *Bam*HI (forward and reverse primers, respectively) site sequences. These DNA sequences were sequentially inserted into the vector by digestion with the appropriate restriction enzymes and ligation reactions. Therefore, inducible expression of double-stranded RNA would be produced with the addition of tetracycline. The pALC14 plasmid also contains the puromycin resistance gene for selection of cells containing the transfected plasmid.

For sub-cloning of the Tb11.47.0019 'a' and 'b' regions the plasmid preparations of these amplified regions in pGEM T-easy vector (which were used for generation of probes to these regions for northern blot analysis, see Section 4.7) were used as template in PCRs. For the 'a' region, PCRs contained either primers 67 and 68 or primers 69 and 70. For the 'b' region, PCRs contained either primers 71 and 72 or primers 73 and 74. The resulting fragments were then inserted into pALC14.



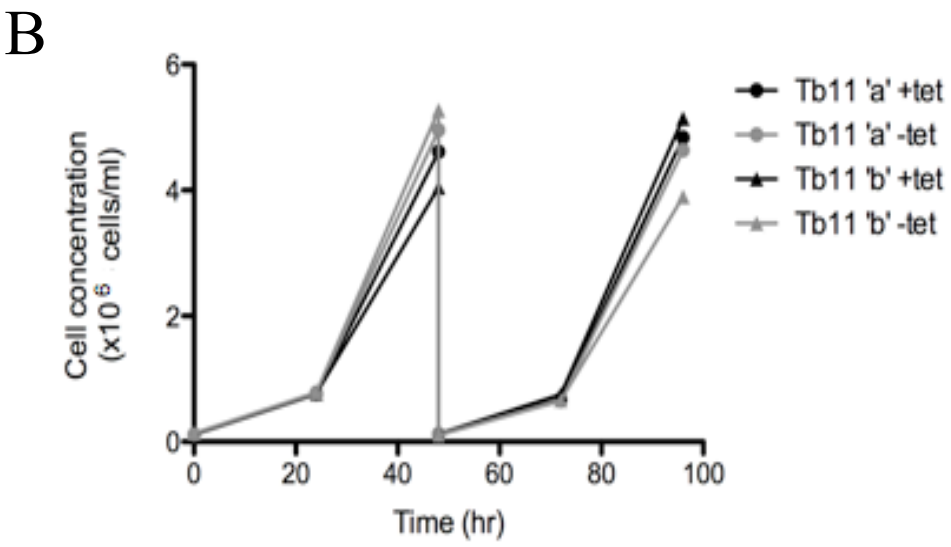


Figure 4.14 continues on the following page

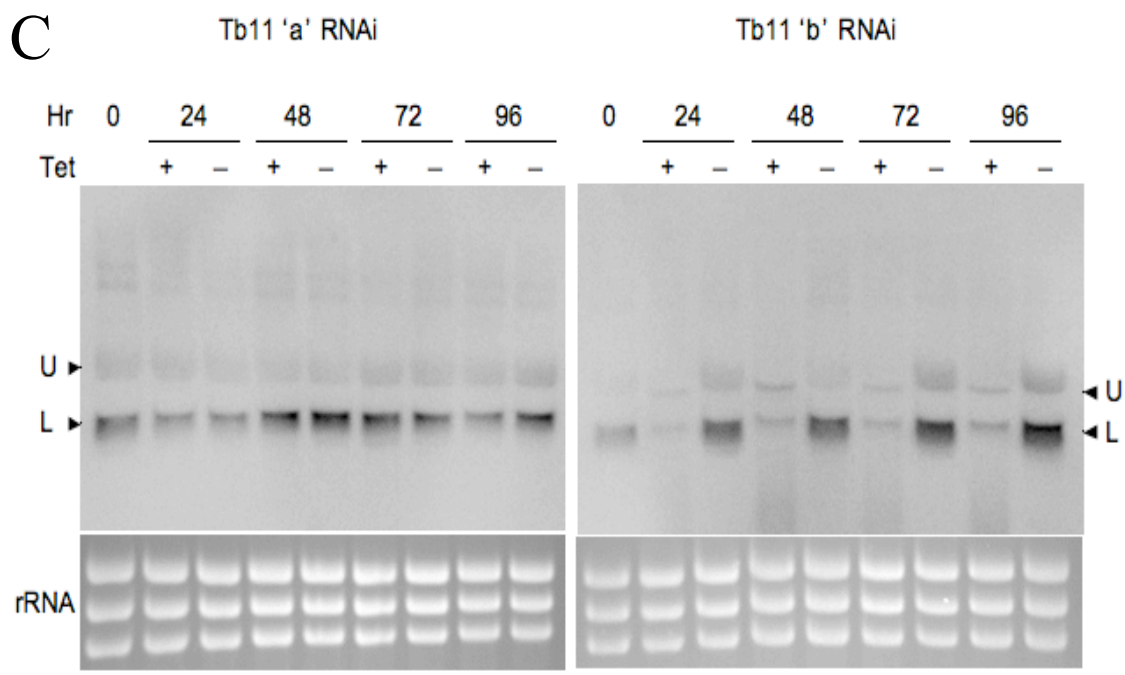


Figure 4.14 Preliminary analysis of RNAi-mediated ablation of Tb11.47.0019 'a' or both Tb11.47.0019 transcripts. A. Schematic diagram of the sub-cloning of the 'a' or 'b' regions of Tb11(Tb11.47.0019) into pALC14. For each of 'a' and 'b', two PCRs were performed with primers (depicted by bent arrows) containing the sequences of the restriction enzyme sites indicated to give two products; one for insertion between the *Hind*III and *Xba*I sites in pALC14, and one for insertion between the *Xho*I and *Bam*HI sites. Shown at the bottom is a representation of the pALC14 vector showing the puromycin resistance gene ($Puro^R$), tetracycline operators (Tet Op.s), stuffer region (stuffer), rDNA spacer, procyclin and T7 promoters, and restriction enzyme site locations. B. Growth analysis following induction of RNAi-ablation of either the 'a' or 'b' regions of Tb11.47.0019. Shown are cell concentrations ($x10^6$ cells/ml) at various time points (in hours) either after induction (+ tet; 1 μg/ml tetracycline) or in non-induced cells (- tet). At 48 hours, the cell concentrations were returned to ~$0.1x10^6$ cells/ml. C. Northern blot analyses of RNA samples taken every 24 hours from cells shown in B. using a digoxygenin-labelled RNA probe against the 'b' region of Tb11.47.0019. Arrowheads indicate the upper (U) and lower (L) bands detected by this probe previously (Figure 4.7). Shown below each blot is the ethidium bromide staining of rRNA present in each sample to indicate loading.

4.9.2 Preliminary assessment of effect of RNAi-mediated ablation of transcripts

As a preliminary analysis of the effect of RNAi-mediated ablation of the Tb11.47.0019 transcripts, cell growth after induction was analysed in one clone of cells obtained from transfection with either the 'a' or 'b' construct described above. For the first analyses, the cells used for transfection were *T. b. brucei* single marker

monomorphic bloodstream form cells, due to initial attempts at transfection of pleomorphic cells with these constructs being unsuccessful. Successful transfectant cells were selected using 0.5 μg/ml puromycin using the clonal selection described in Section 2.2.3.

The growth of cells containing either the Tb11.47.0019 'a' or 'b' RNAi constructs was assessed every 24 hours in cells either following induction of RNAi (+ tet; treated with 1 μg/ml tetracycline) or uninduced (- tet). The resulting cell concentrations measured at each time point are shown in Figure 4.14 (B). As can be seen, similar growth rates were observed for the induced and uninduced cells whether either the long transcript form ('a') or both transcript forms ('b') were targeted for RNAi. There was however a small increase in growth rate for cells at 96 hours after induction of RNAi for both sequences analysed, with respect to uninduced cells. Further analysis would be needed to confirm or disprove this, but the slight decrease in growth rate in these conditions observed at 46 hours makes it appear unlikely that this is due to RNAi induction.

Figure 4.14 (C) shows northern blot analyses of RNA samples taken from cells at the time points shown in Figure 4.14 (B). In these northern blot analyses the probe to the 'b' region of Tb11.47.0019 (discussed in Section 4.7) was used, as it should detect both the long and short transcript forms. From the northern blot analyses, it was observed that there was effective transcript ablation with induction of the Tb11.47.0019 'b' RNAi. This reduced both the upper and lower bands previously detected (see Figure 4.7), indicating that it is likely the upper band does contain the Tb11.47.0019 'b' region sequence and represents the long transcript form. It would be expected that RNAi using the 'b' region would ablate both the long and short transcript forms and this appears to be observed. For the induction of Tb11.47.0019 'a' RNAi, ablation of the long transcript form should result and so it would be expected that the intensity of upper band signal would decrease specifically. However, this did not appear to be the case. This could indicate that the RNAi ablation was not effective in this cell line.

The results shown in Figure 4.14 of this preliminary analysis, indicate that ablation of either of both transcript forms of Tb11.47.0019 generated little affect on the cellular growth rate in monomorphic slender cells. A specific effect from targeting the longer form only, however, could not be determined due to the failure of the Tb11.47.0019 ‘a’ RNAi cell line. Further analyses of any effects attributable to either the long or short form would be best carried out in pleomorphic cells capable of stumpy formation.

4.11 Discussion

4.11.1 uORFs

In other eukaryotes, one mechanism for transcripts to escape global translational repression, such as during stress conditions, is uORFs. Therefore, it was investigated whether uORFs are used as a mechanism for stumpy-enriched transcripts to escape translational repression in stumpy forms.

In *T. cruzi*, a 22-codon uORF present in the tuzin 5'UTR has been reported to substantially reduce reporter gene expression (Teixeira *et al.*, 1999). Additionally, in *T. cruzi*, another four genes were reported to possess uORFs (Jaeger and Brandao, 2011), although no functional analysis was made. Also, a functional uORF has been identified in *Plasmodium falciparum*, another protozoan parasite, which represses translation of *var2csa* (Amulic *et al.*, 2009). Therefore, it appears possible that functional uORFs may also be present in *T. brucei*, although to the best of my knowledge, none have been identified to date.

Using a bioinformatic approach to predict the occurrence of uORFs in the *T. brucei* genome revealed that 25.38% of annotated genes had one or more potential uORFs (Table 4.1), with a median uORF number of 1 and mean of 1.2. This analysis was performed using the definition of an uORF as a methionine codon followed by a stop codon with at least 2 amino acid residues in between and analysed 100nt of upstream sequence for each annotated gene. This length was chosen because the median and mean lengths of known 5'UTRs were previously found to be 68nt and 129nt, respectively (Benz *et al*., 2005). Therefore, using this method, the identified uORF sequences are potential uORFs because it is not known whether each uORF sequence is within the actual 5'UTR sequence for each gene (i.e. there could be many false positive assignments) and additionally, whether the uORF is functional (i.e. whether it affects the expression of the main ORF).

For vertebrate and human 5'UTRs, estimates regarding the frequency of uORFs in the genome vary greatly, being between 11-48% of genes (Meijer and Thomas,

2002). However, for mice and human mRNAs, about 25% were found to have one or more uORFs, with an average of about 1.9 uORFs (Crowe *et al.*, 2006). This is similar to the frequency that was identified here for *T. brucei*. However, the methodology used by Crowe *et al*. (2006) differed substantially. Firstly, the authors examined mature mRNA sequences obtained from each organism, and restricted their analysis to uORFs of 20 to 99 codons in length. At the time the analysis of uORF predictions in *T. brucei* was performed, the actual SAS locations had not been identified on a genome-wide scale. However, with data available from RNA-seq studies, performed by for example, Siegel *et al*. (2010), the identification of uORF sequences in true 5'UTR sequences would be possible. For further work, it would be interesting to repeat the uORF analyses using the data of the 5'UTRs obtained from these studies.

An additional criticism that could be made of this approach could be that many potential uORFs could be missed, these being any uORFs where: i) the start codon precedes the 100nt of upstream sequence, and/or ii) the stop codon is within the main ORF coding sequence, and therefore, the uORF overlaps the main ORF. However, my analysis was performed to obtain a genome-wide prediction of the abundance of potential uORFs and unless each gene was analysed individually, which would be a huge task on a genome-wide scale, these inaccuracies are inevitable.

Analysis of potential uORFs in two datasets of coregulated genes, those upregulated during heat shock and in stumpy forms, suggested that there might be a slight increase in the abundance of potential uORFs in stumpy forms (Table 4.1). Therefore, two genes were chosen for further analysis, these being Tb09.211.2300 and Tb11.47.0019. The ESAG9-EQ 5'UTR was also included in these same analyses as it was found to also possess a potential uORF (Table 4.3).

Using a CAT reporter gene assay, each of the 5'UTRs from these three genes (Tb09.211.2300, Tb11.47.0019 and ESAG9-EQ) was found to repress gene expression in slender forms, at the levels of mRNA and protein (Figure 4.2). This was not due to the replacement of the aldolase 5'UTR present in the reporter gene

plasmid alone because replacement with the TbACO 5'UTR resulted in only a small decrease in CAT expression (Figure 3.3 B). This suggested that the 5'UTR sequences were responsible for the repression of gene expression in slender forms. However, the potential uORF present in each 5'UTR was not responsible for this repression, because when each uORF was deleted through mutation of the uORF start codon this did not alleviate repression (Figure 4.3). Additionally, the 5'UTR sequences were not responsible for release of gene expression repression upon development to stumpy-like forms with 8-pCPT-2'-O-Me-cAMP treatment (Figure 4.4). However, it is possible that alleviation of repression was not observed because gene expression was analysed in stumpy-like forms and not true stumpy forms, where repression may be relieved, but reporter gene expression in true stumpy forms was not examined in this study.

For ESAG9-EQ, the main SAS location was confirmed by Nilsson *et al*. (2010) to be that predicted by Benz *et al*. (2005). Therefore, the EQ 5'UTR sequence analysed in the CAT reporter gene assays should contain a functional SAS. Hence, it is unclear why the *CAT* mRNA produced from the EQ 5'UTR construct was of a much larger molecular weight than predicted (Figures 4.2 B, 4.3 C, and 4.4 B) when the SAS in the EQ 5'UTR sequence would produce a short 5'UTR (41nt). This, combined with the low CAT protein expression levels obtained with this construct in each experiment (Figures 4.2 A, 4.3 A, and 4.4 A), suggests that the EQ 5'UTR was not functional in these experiments. It is possible that additional sequence is required from the endogenous context for correct mRNA processing. But, additionally, it is also possible that stumpy-specific factors may be required for correct mRNA processing, these not being present in slender forms or in the stumpy-like forms analysed. Although, arguing against this second possibility is the observation that the same SAS was identified as the main SAS in slender and stumpy forms in the Nilsson *et al*. (2010) study, suggesting that this SAS is used in both life-cycle stages.

4.11.2 Different trans-splicing sites

As mentioned in Section 4.4, it was discovered in previous studies (Siegel *et al*., 2010; Nilsson *et al*., 2010) that the Tb09.211.2300 and Tb11.47.0019 genes were

members of a group of genes identified to use alternative SAS locations to those predicted by genome annotation (Benz *et al*., 2005), and therefore, were predicted to use an internal start codon.

For Tb09.211.2300, both studies identified only one SAS, located at position +261 relative to the original start codon annotation (Siegel *et al*., 2010; Nilsson *et al*., 2010). Using northern blot analysis or 5'RACE, Tb09.211.2300 transcripts produced from *trans*-splicing reactions using the predicted SAS were not detected (Figures 4.5 and 4.6 B). The SAS that was used in the Tb09.211.2300 transcript detected by 5'RACE was determined to be that at position +261. Therefore, it appears likely that the Tb09.211.2300 5'UTR sequence used in the CAT reporter gene assays did not contain a functional SAS. Consequently, this may explain why this sequence did not support CAT gene expression (Figures 4.2, 4.3, and 4.4).

The Tb09.211.2300 transcript was originally identified as stumpy-enriched by microarray analysis (Kabani *et al*., 2009) performed using an oligonucleotide sequence (marked in bold font in Figure 4.1 A) that was complementary to sequence upstream from the SAS at position +261. However, as mentioned above, the RNA-seq studies, and northern blot analysis and 5'RACE results presented here, only detected usage of the SAS at position +261. This would suggest that Tb09.211.2300 was a false-positive result in the microarray analysis. Confounding this, however, is the observation that the Tb09.211.2300 transcript was slightly enriched in stumpy forms in the northern blot analysis (Figure 4.6 B). Therefore, perhaps an explanation might be that the Tb09.211.2300 transcript detected in the microarray study was a processing intermediate, present prior to transcript *trans*-splicing, which was not detected in the northern blot analysis.

For Tb11.47.0019, two main SAS locations were identified by RNA-seq (Siegel *et al*., 2010; Nilsson *et al*., 2010), one at position +322 relative to the original start codon annotation and the other at position +345. The analysis of Tb11.47.0019 transcripts by 5'RACE identified three SASs (Figure 4.5), however, not all SASs used may be identified by the methodology employed. The SASs identified were

those at position +322/+345 and also the SAS predicted by Benz *et al*. (2005). The identification of the predicted SAS was interesting because in the RNA-seq studies (Siegel *et al*., 2010; Nilsson *et al*., 2010), it was only identified as a minor SAS in procyclic forms, but it appears to be used in slender, stumpy, and procyclic forms by 5'RACE analysis (Figure 4.5). The usage of these SAS locations was also interesting because it could possibly result in the production of different length proteins from the Tb11.47.0019 gene. This was also supported by the identification of two bands in the northern blot analysis using a probe to a region of the transcript present in both alternatively spliced forms (Figure 4.6 C). The intensity of the signal for these two bands suggested that the two alternatively spliced transcript forms might have been present with differential abundance in the life-cycle stages analysed. However, unfortunately, variability in the results obtained from northern blot analyses meant that the life-cycle abundance was unable to be determined.

From the northern blot analyses of Tb11.47.0019 transcripts with the 'b' region probe, (Figure 4.6) it was assumed that the upper band represented the longer transcript form (produced from usage of the predicted SAS), but this was not confirmed. Unfortunately, northern blot analyses using a probe to a region present only in the longer transcript form (probe to region 'a') did not result in the detection of any signal (Figure 4.6). However, RNAi-mediated transcript ablation against the 'b' region, which should ablate the long and short transcript forms, resulted in a decrease in the abundance of both bands (Figure 4.13 C). Therefore, it appears that the upper band contains the 'b' region sequence and so would support the assumption that the upper band detected was the long transcript form.

The use of the predicted SAS and those at position +322/+345 would be expected to produce encoded proteins of two different lengths, a long (487aa) form (from the predicted SAS) and a short (360aa) form (from the SASs at +322/+345). The long protein form sequence was found to include a predicted signal peptide and one or two predicted transmembrane helices, which were absent from the short protein form sequence (Figures 4.8 and 4.10), but both were predicted to contain an UBA/TS-N domain (Figures 4.7 and 4.10). Bioinformatic predictions regarding the cellular

localisation for these two protein sequences resulted in very different results (Table 4.5). Presumably due to the presence of the signal peptide and transmembrane helices, the long protein form was predicted to be localised to the plasma membrane, Golgi, extracellular space, or endoplasmic reticulum. In contrast, the short protein form was predicted to localise to the cytoskeleton, cytoplasm, mitochondrion or nucleus. Additionally, a dileucine motif was found in the long protein sequence but not the short protein sequence, however, the motif was still present in the short form sequence. As mentioned in Section 4.8, a limitation of these bioinformatic predictions is that the tools used are not optimised for use with trypanosome protein sequences.

In *T. brucei*, the production of two different length proteins with different localisations has been demonstrated for aminoacyl-tRNA synthetases, which exhibit alternative transcript splicing (Rettig *et al.*, 2011). The isoleucyl-tRNA synthetase (IleRS) is encoded by a single gene but shows dual localisation to the mitochondria and cytosol, which is achieved through the inclusion/exclusion of a mitochondrial targeting signal in long/short protein forms, respectively, obtained by *trans*-splicing at two different locations (Rettig *et al*., 2011). However, the life-cycle abundance of each transcript isoform was not investigated in that study. Therefore, it is possible that in a similar manner, the generation of different protein isoforms with different cellular localisations, through alternative splicing may also occur for Tb11.47.0019.

In contrast to the predicted localisations for the Tb11.47.0019 long and short protein forms, ectopic overexpression of a Ty epitope tagged copy of either protein form resulted in the same punctate pattern of staining when using immunofluorescence analysis to detect the Ty epitope tag (Figure 4.12). Additionally, western blot analysis of the ectopically expressed tagged proteins produced from either the long or short protein form sequences revealed that both resulted in expression of the short protein form only (Figure 4.11). Therefore, it would appear that even when the upstream ATG is present, the downstream ATG is used preferentially. It was not determined for ectopic expression of the long Tb11.47.0019 protein sequence, whether the SAS present upstream in the pHD451 plasmid or the SAS at position

+322/+345 was used. Therefore, it is possible that the short protein form was produced from the long protein form sequence because the originally annotated ATG is not used or because *trans*-splicing occurred using the SAS at position +322/+345.

The possibility that the use of the predicted SAS does not permit production of the long protein form (beginning at the originally annotated ATG start codon) would be consistent with the observations from the CAT reporter gene assay with this 5'UTR sequence (Figure 4.2 A). Additionally, the homologous proteins of the long protein form identified by protein-protein BLAST analysis (see Table 4.4) showed no homology to the N-terminal extension of the long form. Rather, in most cases the homology began at a downstream methionine residue (most commonly M136), which corresponded to the start of the homologous protein (data not shown). Even the protein that exhibited the most similarity, CBH17021.1 from *T. brucei gambiense*, did not share the N-terminal extension (Figure 4.9). These observations would support the possibility that only the short protein form of Tb11.47.0019 is produced. However, this would need to be determined experimentally, for example, by western blot analysis using an antibody raised against the Tb11.47.0019 protein sequence common to both protein forms.

It is also possible, however, that the long protein form is not produced in slender form cells but is produced in stumpy form cells only. The stumpy form was the life-cycle stage that the long transcript form originally appeared to be enriched, and it might be that additional/different factors, present in stumpy forms and not slender forms are required for expression of the long protein form. The ectopically expressed tagged long/short protein forms were only analysed in slender forms in this study. Therefore, to determine if this is the case, the experiment of western blot analysis using a Tb11.47.0019-specific antibody mentioned above could be used to establish the expression profile of the long protein form.

Chapter 5

Results

Functional analysis of two stumpy-enriched translational proteins in *T. brucei*

In addition to the analysis of gene expression regulation of individual transcripts that are enriched in stumpy forms, proteins involved in the process of translational regulation were analysed. In other organisms, the control of translation can be achieved through regulation of the activity of individual proteins involved in translation, such that global levels of translation are affected. The aim of this part of the project was to determine if any known proteins of translation are important for the global downregulation of translation in stumpy forms or for the escape from translational repression for genes that are expressed in stumpy forms.

5.1 Are any translational proteins upregulated in stumpy forms?

A few proteins known in other organisms to be important in the control of translation were selected to assess whether they could be important in control of translation in stumpy forms. These proteins were eIF4E, eIF2α kinases, eIF6 and a deoxyhypusine synthase. As a starting point to see if any of these proteins could be interesting for further study, their RNA expression across the *T. brucei* lifecycle was investigated by northern blot analysis. In particular, those that would be interesting for further study would be any which showed enrichment in stumpy forms, as this could indicate that they are contributing to translational control in this life-cycle stage.

5.1.1 eIF4 homologues

One of the components of the eukaryotic translational machinery, eIF4E, which binds to the mRNA cap structure and to another translational component, eIF4G, is the limiting initiation factor in most cells and is therefore significant for establishing global translation rates (reviewed in) (Raught and Gingras, 1999). There are four potential homologues of eIF4E identified in *T. brucei* (Clayton and Shapira, 2007). These were designated TbeIF4E1-4 (Tb11.18.0004, Tb927.10.16070, Tb11.01.3630, Tb927.6.1870 respectively). Digoxygenin-labelled probes against 300-500nt of the coding sequence for each of these genes were generated using the primers 75-82 (see Table 2.1). These were then used in northern blot analyses, the results of which are shown in Figure 5.1 (A).

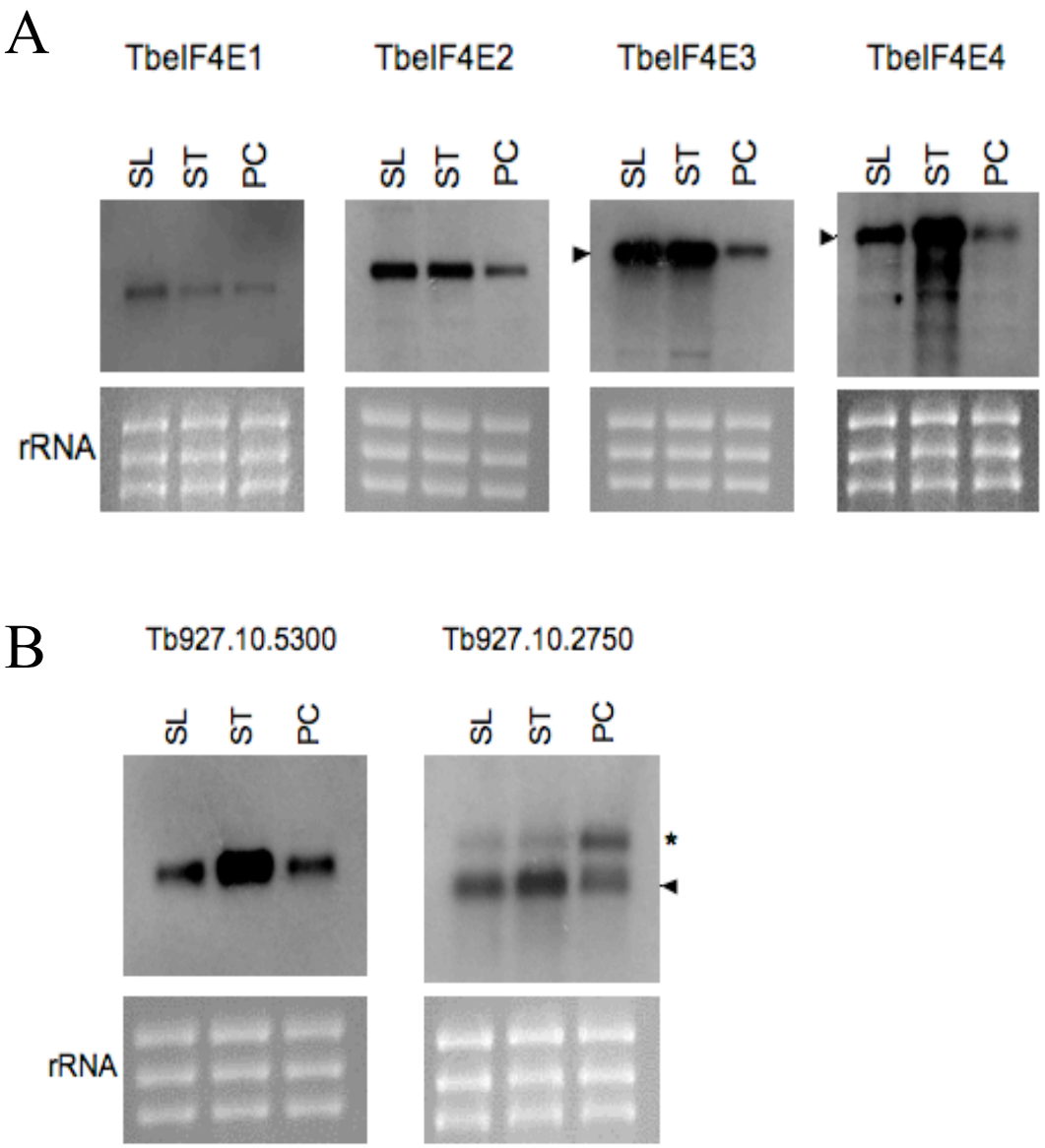


Figure 5.1 Life-cycle abundance of various translational protein homologue transcripts. Transcript abundance was assessed through northern blot analyses using digoxygenin-labelled probes against 300-500nt of the coding sequence for each gene indicated (A: TbeIF4E1-4; B: Tb927.10.5300 [TbeIF6] and Tb927.10.2750 [putative deoxyhypusine synthase]). Arrowheads indicate the main band detected, the asterisk indicates the other band detected with the Tb927.10.2750 probe. SL: slender form; ST: stumpy form; PC: procyclic form. The panel below each blot shows the ethidium bromide staining of the rRNA present to indicate loading of each lane.

The transcripts of the four eIF4E homologues were found to show lifecycle regulation of abundance. The *TbeIF4E1* transcript appeared to be enriched in the slender form, *TbeIF4E2* and *TbeIF4E3* transcripts were downregulated in procyclic forms, and the *TbeIF4E4* transcript was enriched in stumpy forms. The probes

against the *TbeIF4E3* and *TbeIF4E4* transcripts also resulted in the detection of cross-reactive bands that may be the *rRNA* transcripts. Hence, the main signal detected by these probes is marked by an arrowhead in Figure 5.1 (A); this being presumed to be the *TbeIF4E3* or *TbeIF4E4* transcript.

5.1.2 eIF2α kinase homologues

eIF2α kinases mediate phosphorylation of the alpha subunit of eIF2, which causes an inhibition of protein synthesis but activates the translation of specific transcripts, as part of the stress-activated regulatory pathway. There are three eIF2α kinase homologues identified in *T. brucei* (Moraes *et al*., 2007); these were designated TbeIF2K1-3 (Tb11.02.5050, Tb927.4.2500, Tb927.6.2980 respectively). Digoxygenin-labelled probes against 300-500nt of the coding sequence for each of these genes were generated using the primers 83-88 (see Table 2.1), which were then used in northern blot analyses. However, unfortunately the signal achieved with the use of these probes was not great enough to conclusively determine transcript abundance.

5.1.3 eIF6 and deoxyhypusine homologues

Two other translational proteins where also investigated, a putative eIF6 (Tb927.10.5300, the selection of this gene is discussed below in Section 5.3) and a putative homologue of deoxyhypusine synthase (Tb927.10.2750). Mammalian eIF6 is required for efficient translational initiation and may be rate-limiting in translation and growth and acts downstream of extracellular signals (Gandin *et al.*, 2008). Eukaryotic eIF5A promotes translation elongation and possesses a unique modification of lysine to hypusine, which is required for association of eIF5A with ribosomes (Saini *et al.*, 2009; Merrick, 2009). The transcript for a putative deoxyhypusine synthase (Tb927.10.2750) was identified as enriched in stumpy forms by microarray analysis (Kabani *et al.*, 2009). Therefore, both these proteins appeared to be potentially interesting and digoxygenin-labelled probes against 300-500nt of the coding sequence for each of these genes were generated using the primers 89-92 (see Table 2.1).

The results of northern blot analyses with these probes are shown in Figure 5.1 (B). For the putative eIF6, Tb927.10.5300, the mRNA for this gene was found strongly enriched in stumpy form cells compared to slender and procyclic forms. The probe against the putative deoxyhypusine synthase (Tb927.10.2750) resulted in the detection of two bands; the main band is indicated in Figure 5.1 (B) with an arrowhead and a higher molecular weight band is indicated with an asterisk. The transcript of the main band was somewhat enriched in stumpy forms and the other band was enriched in procyclic forms. Interestingly, three different 5'UTR lengths were identified for this gene, 624, 604, and 118 bp (Siegel *et al.*, 2010). Therefore, it is possible that the two bands detected here could represent the transcripts with 624/604 bp length 5'UTRs and the transcript with the 118 bp length 5'UTR. However, the abundance of the transcripts observed in the Siegel *et al*. (2010) study do not match the abundances shown in Figure 5.1 (B).

The enrichment of the *TbeIF4E4* and *TbeIF6* transcripts in stumpy forms highlighted them as interesting for further study, as they may be contributing to the global repression of translation in stumpy forms and/or the escape from translational repression of stumpy-expressed transcripts. Therefore, further characterisation was made of these proteins.

5.2 Preliminary analysis of TbeIF4E1-4

Due to the presence of four homologues of the eIF4E protein, it was possible that these proteins may have different functions. One such possibility could be that one of the homologues functions as a negative regulator, similarly to the 4E homologous protein (4EHP) which binds to the mRNA cap but not to eIF4G, thereby preventing translation initiation. Therefore, the protein sequences of the eIF4E homologues were analysed for conservation of the residues found to be important in cap or eIF4G binding in other organisms.

Firstly, a ClustalW2 multiple sequence alignment was performed with the protein sequences of TbeIF4E1-4, human eIF4E (accession: P06730), *S. cerevisiae* eIF4E

(accession: P07260) and *Drosophila melangaster* 4EHP (accession: Q8T3K5; see Appendix 3 for the alignment). Conservation of important residues was determined by the presence or absence of the same residue in the aligned position. The conservation of residues important for binding to the mRNA cap or eIF4G is summarised in Table 5.1.

All of the four eIF4E homologues show some lack of conservation of residues important for cap binding and binding of eIF4G. The levels of conservation are similar to that found for the 4EHP protein. Some residues important for cap binding are however conserved in all the sequences analysed (W102, W166, E103, G111). This was also true for two residues important for eIF4G binding, G139 and W75. Although in TbeIF4E1 W75 was not conserved and there is a phenylalanine (F) residue at this position, in yeast it was found that eIF4G binding was maintained with a W75F mutation (Ptushkina *et al.*, 1998), so this could represent a conservative mutation.

Of the 14 cap binding residues investigated, 11 were found to be conserved in TbeIF4E1, 10 in TbeIF4E2, 7 in TbeIF4E3, 7 in TbeIF4E4 and 8 in 4EHP. Therefore, this would suggest that TbeIF4E1 and TbeIF4E2 would be likely to bind the mRNA cap, with TbeIF4E3 and TbeIF4E4 less likely. However, 4EHP also lacked a similar number of these residues and yet can still bind to the mRNA cap. Therefore, all homologues may bind the mRNA cap, but perhaps with differing affinities. Of the 8 eIF4G binding residues investigated, 5 were found to be conserved in TbeIF4E1, 3 in TbeIF4E2, 6 in TbeIF4E3, 4 in TbeIF4E4 and 5 in 4EHP. Hence, this would suggest that TbeIF4E3 would be likely to bind to the cap. However, all homologues have a similar number of conserved residues to the 4EHP protein, which does not bind to eIF4G. Overall, therefore, prediction of eIF4G binding for each homologue is difficult.

Table 5.1 Conservation of eIF4E residues important for cap binding or eIF4G binding in TbeIF4E1-4 and 4EHP[a]

Residue[b]	Function	Details	Reference[c]	TbeIF4E1	TbeIF4E2	TbeIF4E3	TbeIF4E4	4EHP
Cap binding:								
W43	Cap binding	W43L: complete loss	1	Y	N (G)	Y	Y	N (Y)
W46	Cap binding	W46L: complete loss	1	Y	N (K)	N (Y)	N (F)	Y
W56	Cap binding	W56L: ~80% loss	1	Y	Y	N (L)	N (S)	N (Y)
W102	Cap binding	W102L: complete loss	1	Y	Y	Y	Y	Y
W166	Cap binding	W166L: complete loss	1	Y	Y	Y	Y	Y
W113	Cap binding	W113L: ~80% loss	1	N (L)	Y	N (I)	N (V)	Y
W130	Cap binding	W130L: complete loss	1	Y	Y	N (F)	Y	Y
E103	Strenghtens cap binding	E103A: >80% loss	1	Y	Y	Y	Y	Y
G111	Cap-binding	G111D: reduces affinity	2	Y	Y	Y	Y	Y
H200	Required for cap binding		1	Y	Y	Y	N (R)	Y
H37	Required for cap binding		1	N (G)	N (T)	N (S)	N (S)	N (G)
K/R112	Interacts with the cap		3	N (K)	Y	N (T)	N (I)	N (Q)
R157	Interacts with the cap		3	Y	N (Q)	Y	Y	N (K)
R/K162	Interacts with the cap		3	Y	Y	N (T)	N (M)	N (S)
eIF4G binding:								
E72 *	Involved in eIF4G binding	Mutation reduced binding	4	Y	N (E)	Y	N (S)	N (Q)
H37 *	Involved in eIF4G binding	Mutation reduced binding	4	N (G)	N (T)	N (S)	N (S)	N (G)
V71 *	Involved in eIF4G binding	Mutation reduced binding	4	Y	N (A)	Y	N (I)	Y
G139 *	Involved in eIF4G binding	Mutation reduced binding	4	Y	Y	Y	Y	Y
W75 *	Involved in eIF4G binding	Mutation reduced binding	4	N (F*)	Y	Y	Y	Y
L131	Involved in eIF4G binding		5	N (E)	N (E)	N (M)	Y	N (E)
L/M135	Involved in eIF4G binding		5	N (I)	N (V)	Y	N (C)	Y
E140	Involved in eIF4G binding		5	Y	Y	Y	Y	Y

[a] Presence (Y) or absence (N) of residue at aligned position; in parentheses are the single letter amino acid codes for non-conserved residues, asterisk here indicates that this residue was found to maintain binding

[b] Asterisk indicates numbering for the yeast eIF4E protein

[c] 1: Morino *et al*. (1996), 2: Altmann and Trachsel (1989), 3: Marcotrigiano *et al*. (1997), 4: Ptushkina *et al*. (1998), 5: Marcotrigiano *et al*. (1999)

In conclusion, the residue comparison predicted no obvious differential binding to eIF4G of the homologues, which may result in the protein functioning as a negative regulator. The protein structure for each homologue was also predicted with homology modelling using SwissModel (http://swissmodel.expasy.org/). However, an interpretation of whether any of the homologues should bind to the cap or eIF4G could not be made from these structures. Moreover, it should be noted that trypanosomes possess a highly modified cap structure, termed cap-4, and four eIF4G homologues. Hence different interactions may be important for the binding to these molecules. Consequently, functional characterisation of the TbeIF4E proteins would be needed to determine cap binding and eIF4G binding ability.

5.2.1 TbeIF4E RNAi-mediated ablation

Due to the enrichment of the *TbeIF4E4* transcript in stumpy forms (Figure 5.1 A), this protein appeared to be a good candidate for further study, despite the inability to predict whether this protein may be a negative regulator of translation through bioinformatic methods. Therefore, to analyse the function of TbeIF4E4, the transcript was ablated through RNAi.

For tetracycline-inducible RNAi, the pQuadra set of plasmids (Inoue *et al.*, 2005) was used. A schematic diagram of construct generation is shown in Appendix 2. The same amplicon used to generate the TbeIF4E4 probe for northern blot analysis (Section 5.1) was used for the insertion into the pQuadra 3 plasmid. Two independently derived stable transfectant cell lines were generated using *T. b. brucei* Lister 427 'single marker' bloodstream form cells; clones B2 and B4.

Unfortunately, it was subsequently discovered that another laboratory was investigating the TbeIF4E proteins. Therefore, work on the TbeIF4E4 protein was ceased. However, a preliminary analysis was made of the RNAi cell lines before this and the results are given briefly below.

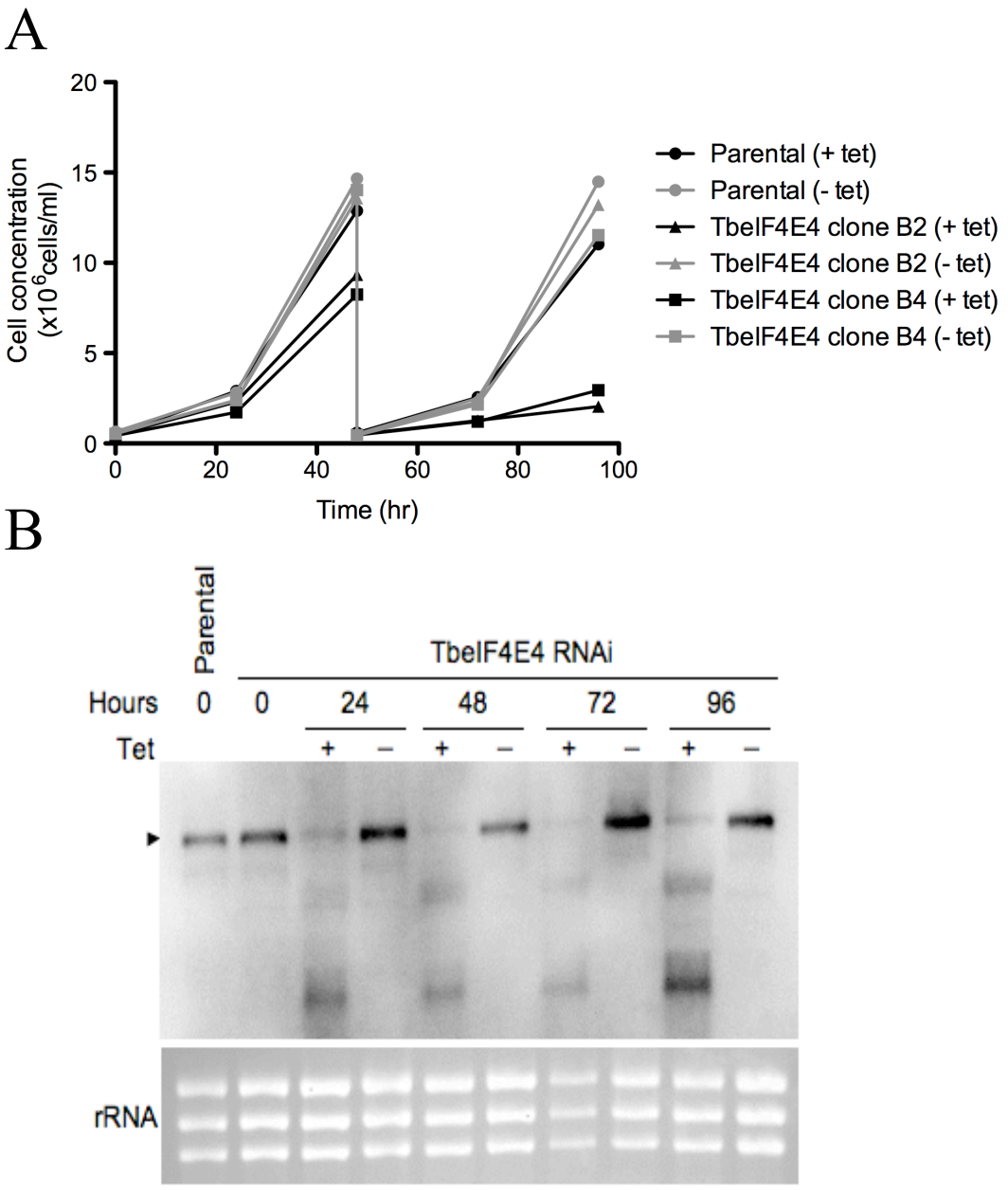


Figure 5.2 TbeIF4E4 RNAi-mediated ablation. A. Growth analysis following induction of TbeIF4E4 RNAi-mediated ablation at 0 hours with 10 μg/ml tetracycline, in two independently derived clones (B2 and B4). The growth of the parental cell line treated with 10 μg/ml tetracycline (Tet), or untreated, is shown as a control. The data shown is from a representative experiment. B. Northern blot analysis of RNA samples from induced (+ Tet) and uninduced (- Tet) cells using a digoxygenin-labelled probe against part of the TbeIF4E4 coding region. Shown below the blot is the ethidium bromide staining of rRNA present to indicate loading of each lane.

Figure 5.2 (A) shows the cell growth after induction of TbeIF4E4 RNAi in each independently derived clone (B2 and B4), compared to the parental cell line also

treated with tetracycline from a representative experiment. Inhibition of cell growth rate was apparent by 48 hours after the induction of RNAi with tetracycline with a further decrease in cells induced for TbeIF4E4 RNAi at 96 hours (22.7% and 27.5% of the cell concentration of uninduced cells for clone B2 and B4 respectively).

Figure 5.2 (B) shows the northern blot analysis of RNA samples taken from one of the TbeIF4E4 cell lines induced or uninduced for RNAi and parental cells, with the same probe used for the northern blot analysis shown in Figure 5.1 (A). The signal detected of the *TbeIF4E4* transcript is clear in the parental and TbeIF4E4 cell line RNA from 0 hours and the uninduced (- Tet) RNA samples, and is marked in Figure 5.2 (B) with an arrowhead. At 24 hours after induction of TbeIF4E4 RNAi with tetracycline (+ Tet lanes), the transcript was significantly reduced in abundance. This was maintained throughout the duration of the experiment, although it should be noted that for the 72 and 96 hour '+ Tet' lanes less RNA was present compared to the '- Tet' control lanes due to the decreased cell concentration of the induced cells.

Therefore, Figure 5.2 shows that induction of TbeIF4E4 RNAi results in an inhibition of bloodstream form cell growth rate starting at about 48 hours but apparent by 72 hours, and confirms knockdown of the *TbeIF4E4* transcript by treatment with tetracycline. A preliminary investigation to assess whether the inhibition of cell growth rate was attributable to a block in the cell cycle was made (as described below in Section 5.5). However, an accumulation of cells in a particular phase of the cell cycle after induction of RNAi was not observed.

Next, it was investigated whether TbeIF4E4 is involved in translation. If so, then this would suggest that it does indeed bind to the mRNA cap and eIF4G protein. Therefore, the polysome profiles were compared from cells induced for TbeIF4E4 RNAi and control, uninduced, cells.

This experiment involved obtaining cell lysate from TbeIF4E4 RNAi cells either uninduced, or induced, for 48 hours with 10 μg/ml tetracycline. Then lysates were layered on to a 15-50% sucrose gradient, centrifuged at 40 000 rpm for 2 hours, and

fractions made of a certain size (see Section 2.9.3). The absorbance was measured at 254 nm throughout fractionation and the resulting traces are shown in Figure 5.3.

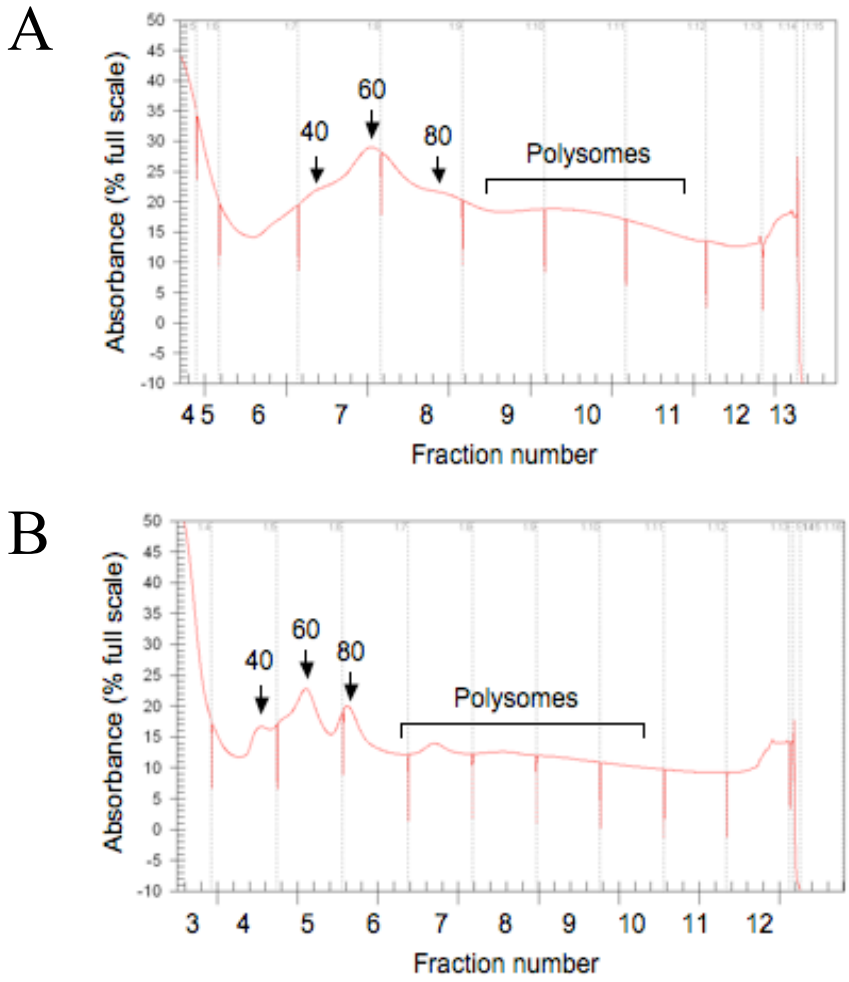


Figure 5.3 Polysome analysis after TbeIF4E4 RNAi-mediated ablation. Cells were harvested for polysome analysis either from control cell populations (uninduced; A) or 48 hours after induction of TbeIF4E4 RNAi (B) after treatment with 100 μg/ml cycloheximide. Cell lysate was layered on to a 15-50% sucrose gradient, centrifuged at 40 000 rpm for 2 hours, and fractions taken at a set number of drops (see section 2.9). Absorbance was measured at 254 nm throughout fractionation. Shown are the absorbance traces with red vertical lines separating the fractions generated. The control was processed in parallel with the polysome analysis of the induced cells. The absorbance peaks corresponding to the assumed 40S, 60S, 80S and polysome locations are indicated.

Unfortunately, the polysome profile of the control cells (Figure 5.3 A) did not appear to be very successful as the peaks of the 40S and 60S monosomal subunits, and 80S monosomes are not very well resolved. In contrast, the polysome analysis of induced cells was more successful, with the monosomal subunits, monosomes and polysomes being well defined on the absorbance trace (Figure 5.3 B). From comparison to absorbance traces from other successful polysome analyses, it appeared that there was a reduction in the abundance of polysomes in the induced sample. This is consistent with the function of eIF4E in translation initiation, because an inhibition of translation initiation would cause a decrease in the number of ribosomes associated with mRNAs. Although further confirmation would be necessary, this supports the function of TbeIF4E4 as an eIF4E protein.

5.3 A putative eIF6 protein in *T. brucei*

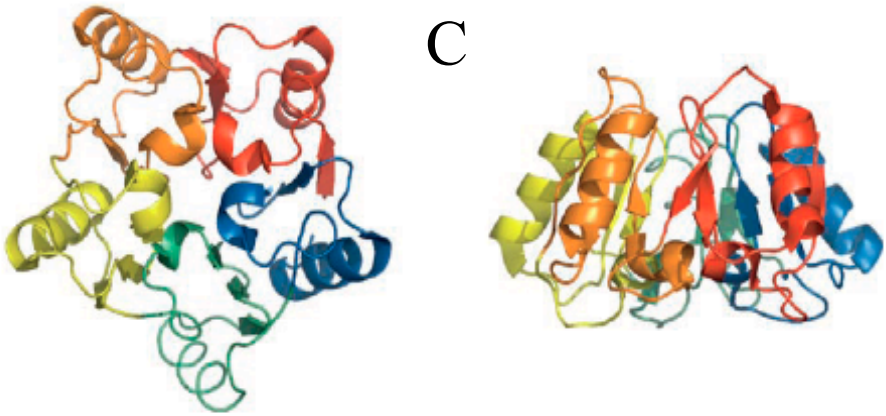
The eukaryotic translation initiation factor, eIF6, appeared to be an interesting candidate protein to investigate because it was upregulated in stumpy forms. Also, it has been found that the expression of this protein is upregulated in archaea during cold shock (Benelli *et al.*, 2009) and cold shock is a trigger involved in differentiation in *T. brucei* (Engstler and Boshart, 2004). Therefore, the presence of a potential eIF6 homologue in *T. brucei* was investigated.

The protein sequence for an archaeal IF6 (*Sulfolobus solfataricus*; accession: Q980G0) was used as query in a protein BLAST search against the *T. brucei* genome entries in the TriTryp database (http://tritrypdb.org/tritrypdb/). This returned one significant hit in the *T. brucei* TREU 927 genome (Tb927.10.5300; e value: 7.4 $x10^{-26}$). The same result was obtained using human eIF6 as query (accession: P56537, e value: 2.0 $x10^{-80}$). In the TriTryp database this gene is annotated as encoding a putative eIF6 protein and a reciprocal BLASTP search using the protein sequence identified an extensive list of sequences annotated as eIF6 proteins.

A

```
human   MAVRASFENNCEIGCFAKLTNTYCLVAIGGSENFYSVFEGELSDTIPVVHASIAGCRIIG 60
TbeIF6  MTLRTRFESSDDVGVFARLTNAYCLVAAGASQNFYSVFEQELASHIPVVYTSIGGSRVVG 60
        *:::: **.. ::* **:***:***** *.*:******* **:. ****::**.*.*::*

human   RMCVGNRHGLLVPNNTTDQELQHIRNSLPDTVQIRRVEERLSALGNVTTCNDYVALVHPD 120
TbeIF6  RLTIGNRHGLVVPSITTDQELQHLRNSLPDSVKVQMVEERLSALGNCVVCNDHVALIHTD 120
        *: :******:**. ********:******:*::: ********** ..***:***:*.*

human   LDRETEEILADVLKVEVFRQTVADQVLVGSYCVFSNQGGLVHPKTSIEDQDELSSLLQVP 180
TbeIF6  LSRETEEVIRDTLQVQTFRTSIAENALVGSYAAVTNKGCMVHPKTPAQDMDEIASLLQVP 180
        *.*****:: *.*:*:.** ::*::.*****...:*:* :*****. :* **::******

human   LVAGTVNRGSEVIAAGMVVNDWCAFCGLDTTSTELSVVESVFKLN---EAQPSTIATSMR 237
TbeIF6  VVAGTINRGNAAIGSGLVVNDWAAFCGLNTTATEITVVERIFQLRRDLGGDESNLLQQLR 240
        :****:***. .*.:*:*****.*****:**:**::*** :*:*.    .: *.:  .:*

human   DSLIDSLT 245
TbeIF6  DTLVDELA 248
        *:*:*.*:
```

B C



D E

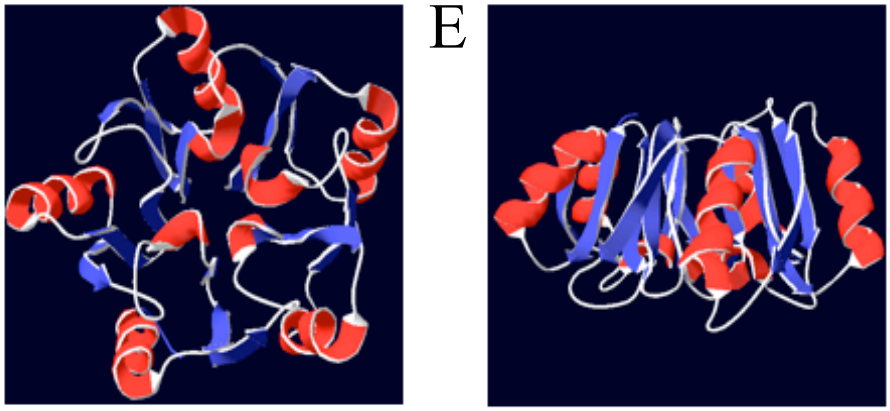


Figure 5.4 Conservation of eIF6 sequence and structure in TbeIF6. A. ClustalW2 alignment of the human eIF6 protein sequence (accession: P56537) and *T. brucei* eIF6 homologue (TbeIF6, Tb927.10.5300) protein sequence. Asterisk (*): single, fully conserved residue; colon (:): conservation between groups of strongly similar properties; period (.): conservation between groups of weakly similar properties. Underlined: region of purified mouse anti-eIF6 antibody immunogen (BD Biosciences). B and C. Structure model of *S. cerevisiae* eIF6 (from Miluzio *et al.* 2009) showing top (B) and side (C) views. D and E. Homology model of TbeIF6

using SwissModel (http://swissmodel.expasy.org/) based on *S. cerevisiae* eIF6 PDB structure (1G62A), shown are the top (D) and side (E) views; blue: predicted beta sheets, red: predicted alpha helices.

A ClustalW2 alignment of the human eIF6 protein sequence and the Tb927.10.5300 protein sequence is given in Figure 5.4 (A). The sequences show a high degree of conservation, with most aligned residues being fully conserved, or possessing similar properties.

The eIF6 proteins of archaebacteria and *S. cerevisiae*, have been shown to form a unique 'pentein' structure, formed by five quasi identical α/β subdomains arrayed around a five-fold axis of pseudosymmetry (Groft *et al.*, 2000). A model of *S. cerevisiae* eIF6 is given in Figure 5.4, showing the top (B) and side (C) views (from (Miluzio *et al.*, 2009). The tertiary structure of the Tb927.10.5300 protein sequence was predicted through homology modelling using SwissModel (http://swissmodel.expasy.org/) and the Protein Database (PDB) structure for *S. cerevisiae* eIF6 (accession: 1G62A). The top and side views of the predicted model are shown in Figure 5.4 (D and E, respectively).

Comparison of Figure 5.4 (B) with (D) and (C) with (E) reveals that the protein sequence of Tb927.10.5300 is predicted to form a highly similar teriary structure to that found for the *S. cerevisiae* eIF6 protein. The 'pentein' structure is predicted to form, but there are some minor differences in the exact folding. However, the predicted structure supports that Tb927.10.5300 may encode an eIF6 protein.

In summary, the bioinformatic analyses demonstrate that there is a gene encoding a putative eIF6 protein in *T. brucei*, which shows a high degree of similarity to known eIF6 proteins in protein sequence and predicted structure. The protein encoded by Tb927.10.5300 will therefore be referred to as 'TbeIF6' hereafter.

5.4 Analysis of TbeIF6 with RNAi-mediated transcript ablation

As a first step to assess the function of TbeIF6, the transcript was ablated through RNAi. For tetracycline-inducible RNAi, the pQuadra set of plasmids (Inoue *et al.*, 2005) was used, as for the TbeIF4E4 RNAi described above. Similarly, the same amplicon used to generate the TbeIF6 (Tb927.10.5300) probe for northern blot analysis (Section 5.1) was used for the insertion into the pQuadra 3 plasmid. Only one independently derived stable transfectant cell line was generated using *T. b. brucei* Lister 427 'single marker' bloodstream form cells; additional transfectant cell lines could not be isolated despite multiple attempts.

Figure 5.5 (A) shows the cell growth after induction of TbeIF6 RNAi compared to the parental cell line also treated with tetracycline, from a representative experiment. The data shown is the cell concentration as a percentage of that of uninduced cells and the mean ± standard error of two independent experiments. Inhibition of cell growth rate was apparent by 48 hours after the induction of RNAi with tetracycline, with complete growth inhibition leading to a further large decrease in the cell concentration of cells induced for TbeIF6 RNAi at 96 hours (9.9% of the cell concentration of uninduced cells).

Figure 5.5 (B) shows a northern blot analysis of RNA samples taken from cells induced or uninduced for TbeIF6 RNAi or parental cells, using the same probe used for the northern blot analysis shown in Figure 5.1 (B). The signal detected of the *TbeIF6* transcript was clear in the parental and TbeIF4E4 cell line RNA from 0 hours and the uninduced (- Tet) RNA samples, and is marked in Figure 5.5 (B) with an arrowhead. As with TbeIF4E4 RNAi, the growth inhibition of induced cells at 72 and 96 hours resulted in a reduction in the amount of RNA collected from these cells. Nonetheless, at 24 hours after induction of TbeIF6 RNAi, the *TbeIF6* transcript was significantly reduced in abundance compared to the RNA from uninduced cells. This was maintained throughout the duration of the experiment, although the apparent further decrease in transcript abundance at 72 and 96 hours is likely to be due to the reduction in the amount of RNA present for these samples.

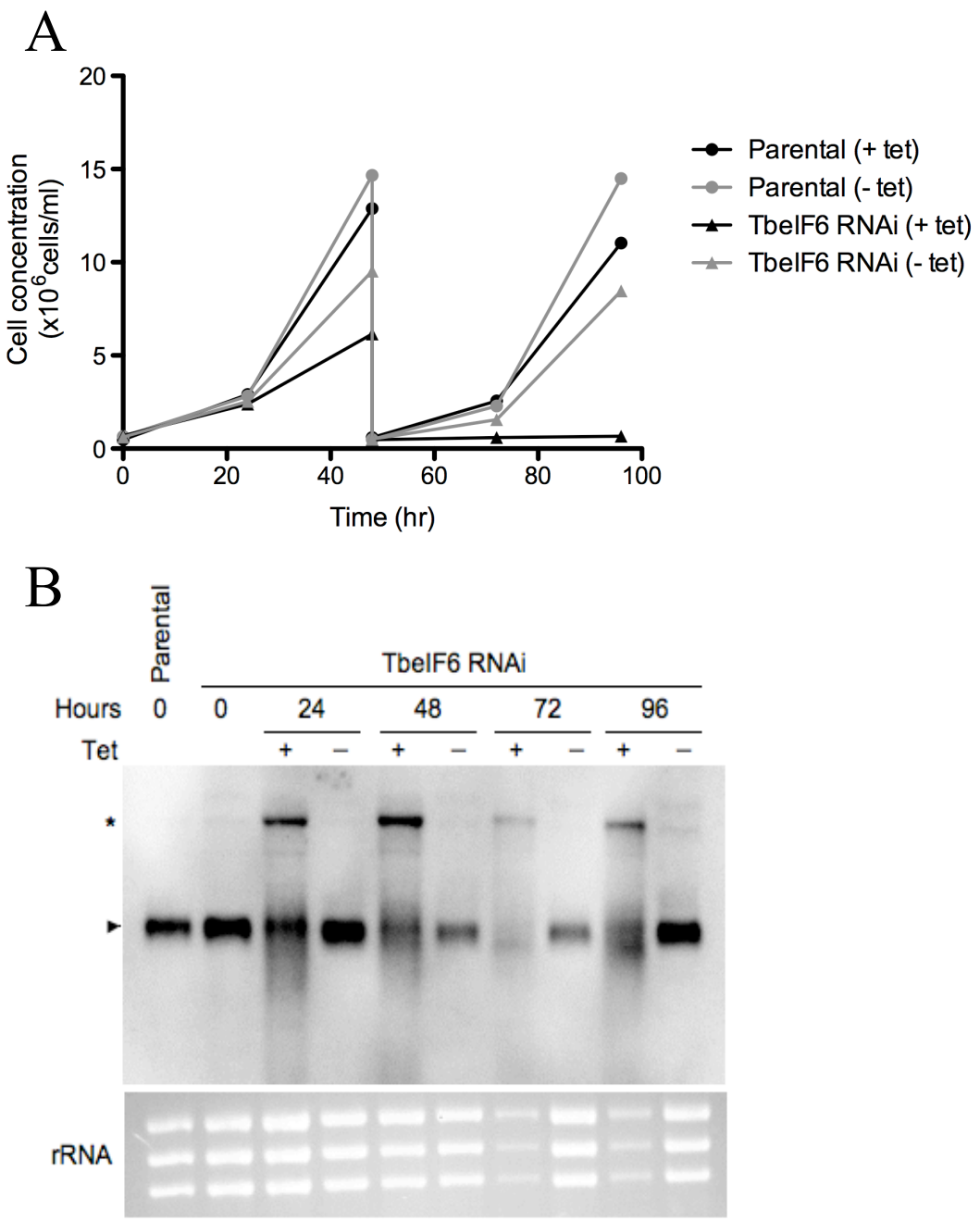


Figure 5.5 TbeIF6 RNAi-mediated ablation. A. Growth analysis following induction of TbeIF6 RNAi-mediated ablation with 10 μg/ml tetracycline (Tet) at 0 hours. The growth of the parental cell line treated with 10 μg/ml tetracycline, or untreated, is shown as a control. The data shown is from a representative experiment. B. Northern blot analysis of RNA samples from induced (+ Tet) and uninduced (- Tet) cells using a digoxygenin-labelled probe against part of the TbeIF6 coding region. The arrowhead indicates the *TbeIF6* mRNA and the asterisk indicates the higher molecular weight band observed with TbeIF6 RNAi induction. Shown below the blot is the ethidium bromide staining of rRNA present to indicate loading of each lane.

5.4.1 TbeIF6 RNAi induction causes the appearance of a higher molecular weight RNA

Also apparent in the RNA samples from the induced cells in Figure 5.5 (B) was an additional band detected by the TbeIF6 probe. This band was of a higher molecular weight and is indicated with an asterisk at the left of the blot in Figure 5.5 (B).

To investigate whether the abnormal processing of RNAs was a general phenomenon for mRNAs with induction of TbeIF6 RNAi, these RNA samples were also probed with the TbeIF4E4 probe using northern blot analysis. This is shown in Figure 5.6 (A) and the appearance of a higher molecular weight *TbeIF4E4* transcript was not detected in the RNA from cells induced for TbeIF6 RNAi. This suggests that irregular RNA processing was not a general consequence of TbeIF6 RNAi.

To determine if the higher molecular weight band was derived from the TbeIF6 RNAi plasmid, a version of the TbeIF6 probe was generated complementary to the antisense strand of *TbeIF6* RNA (TbeIF6 'sense' probe). Again, this 'sense' probe was used in northern blot analysis with the same TbeIF6 RNAi induced and uninduced RNA samples, the result being shown in Figure 5.5 (B). As expected, there was no signal detected using the 'sense' probe with RNA from the parental cell line. or uninduced RNA samples. However, a band (indicated with an arrowhead) and smear was detected using this probe in the RNA from induced cells (Figure 5.6 B). This was not detected in the RNA from uninduced cells, which is consistent with the tight tetracycline induction seen in Figure 5.5 (B). This main band appeared to be a similar size to the higher molecular weight band in Figure 5.5 (B), suggesting that the higher molecular weight band is of TbeIF6 RNAi plasmid origin.

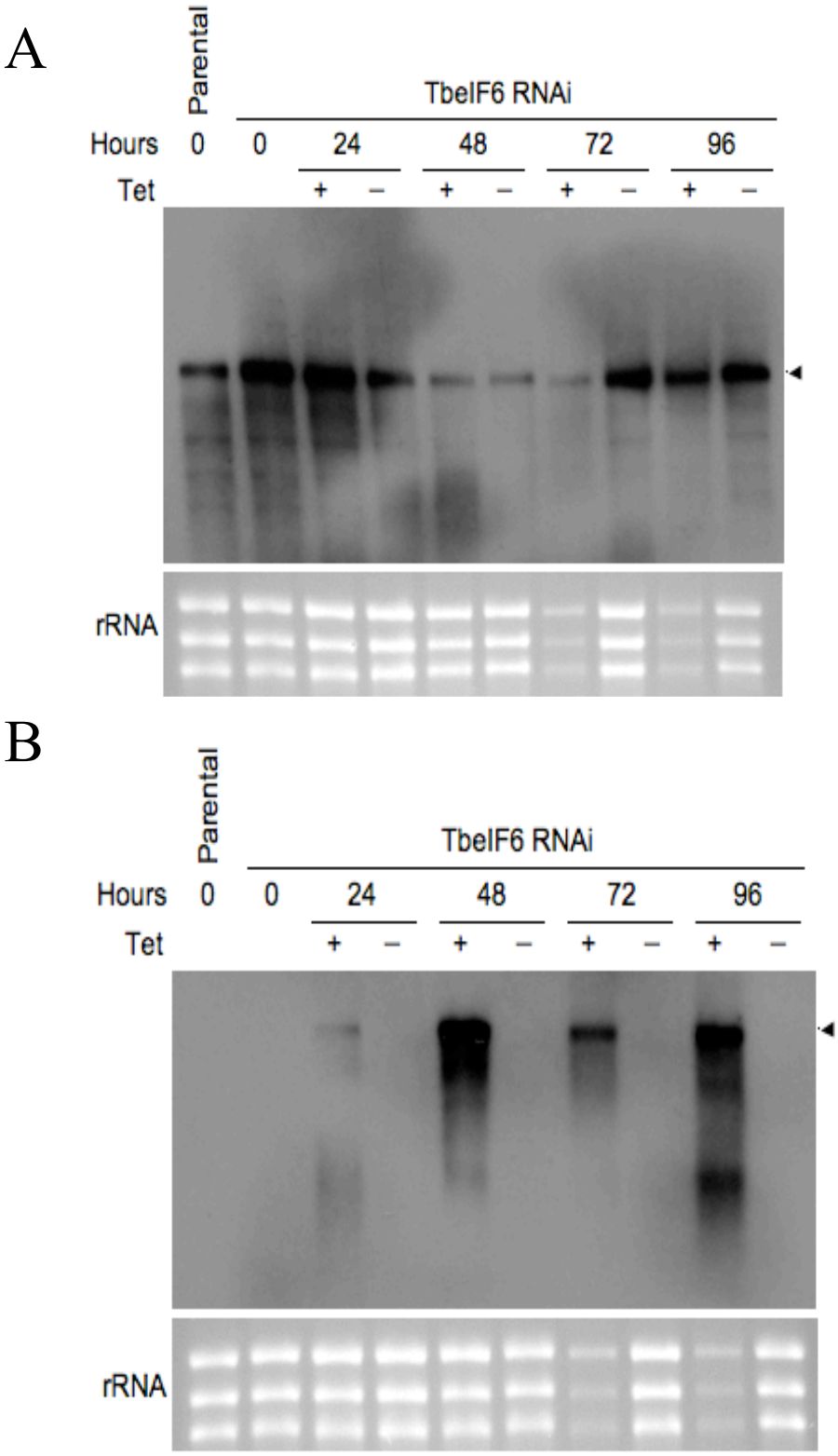


Figure 5.6 continues on the following page

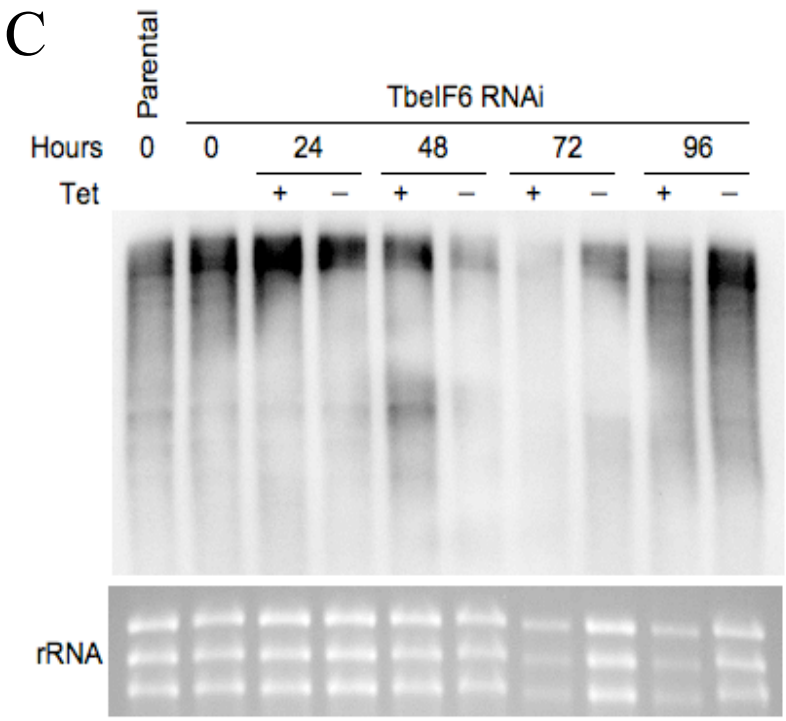


Figure 5.6 Abundance of various transcripts following TbeIF6 RNAi induction. Northern blot analyses with digoxgenin-labelled probes against part of the TbeIF4E4 coding region (A), or to detect the anti-sense strand of *TbeIF6* RNA (B; 'sense' probe), or against part of the INGI sequence (C). RNA samples are from the parental cell line and TbeIF6 RNAi cell line either induced (+ Tet) or uninduced (- Tet) at the times indicated in hours. The arrowheads indicate the main bands detected by the probes. The panel below each blot shows the ethidium bromide staining of the rRNA present to indicate loading of each lane.

As eIF6 has been reported to be involved in miRNA silencing and RNAi in other organisms (see Section 1.5.4), it was investigated whether TbeIF6 plays a role in RNAi in *T. brucei*. Therefore, a transcript known to be subject to RNAi in *T. brucei*, the INGI retroposon, was also investigated using RNA from the TbeIF6 RNAi cell line. It has previously been demonstrated that RNAi deficiency through the generation of *AGO1* knockout (*ago1*$^{-/-}$) cells resulted in the increase in high molecular *INGI* transcripts (Janzen *et al.*, 2006). A digoxygenin-labelled probe against the INGI sequence was generated via amplification of that region by PCR with primers 93 and 94 (see Table 2.1) on genomic DNA. The region of the INGI sequence selected was that used previously for an INGI probe (Janzen *et al.*, 2006) using the ingi-3 sequence (Kimmel *et al.*, 1987).

The northern blot analysis using this INGI probe is shown in Figure 5.6 (C). In the RNA samples from the cells induced for TbeIF6 RNAi there may be an increase in high molecular weight *INGI* transcripts at 24 and 48 hours compared to the uninduced cells. However, the detection over some regions of the blot is obscured and comparison of the 48 hour induced lane to the 24 hour uninduced lane reveals similar detection levels of *INGI* transcripts. Comparison between RNA samples from induced and uninduced cells is also hindered in the 72 and 96 hour samples due to the low amounts of RNA present for the induced cells at these time points.

The data presented in Figure 5.6 suggest that the induction of TbeIF6 RNAi inhibits the processing of the antisense *TbeIF6* transcript produced from the RNAi construct, which does not appear to be a transcript-wide phenomenon. Also, TbeIF6 RNAi induction might affect the RNAi pathway as there may be some accumulation of high molecular weight *INGI* transcripts, however, further experiments would be needed to confirm this.

5.5 Analysis of TbeIF6 through ectopic overexpression

The enrichment of the *TbeIF6* transcript in stumpy forms suggests that the protein expression may be upregulated in this lifecycle stage. Therefore, the effect of TbeIF6 overexpression was investigated in slender form cells, to determine any effects on translation.

The TbeIF6 protein was overexpressed though tetracycline inducible expression of a Ty epitope tagged ectopic copy using the pHD451 plasmid modified to insert a Ty tag at the C terminus of the expressed protein (see Section 4.9.1). The TbeIF6 coding region (excluding the stop codon) was amplified by PCR using primers 95 and 96 on genomic DNA. Thereafter, two independently derived stable transfectant cell lines (clones C5 and C1) were generated using *T. b. brucei* Lister 427 'single marker' bloodstream form cells.

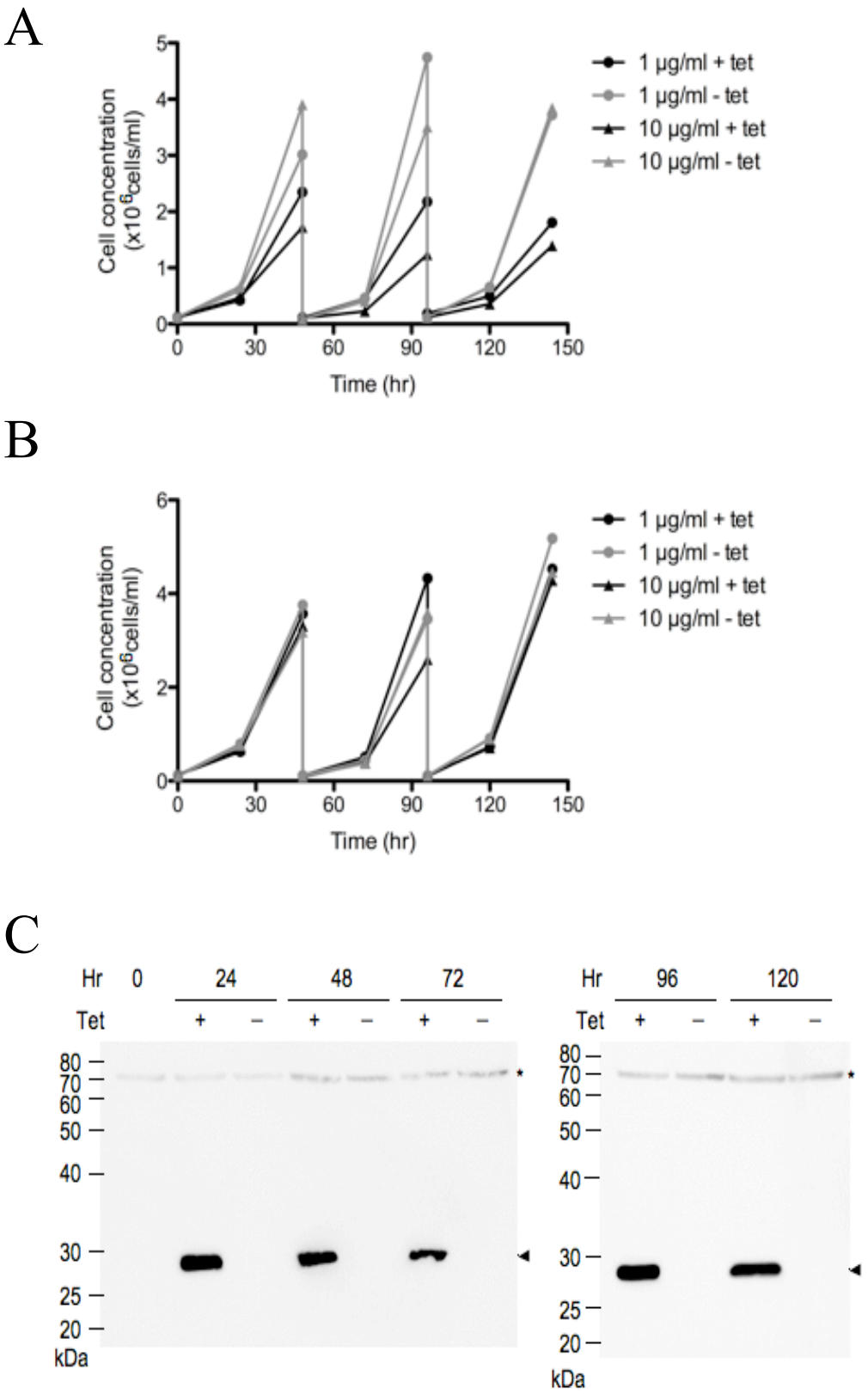


Figure 5.7 continues on the following page

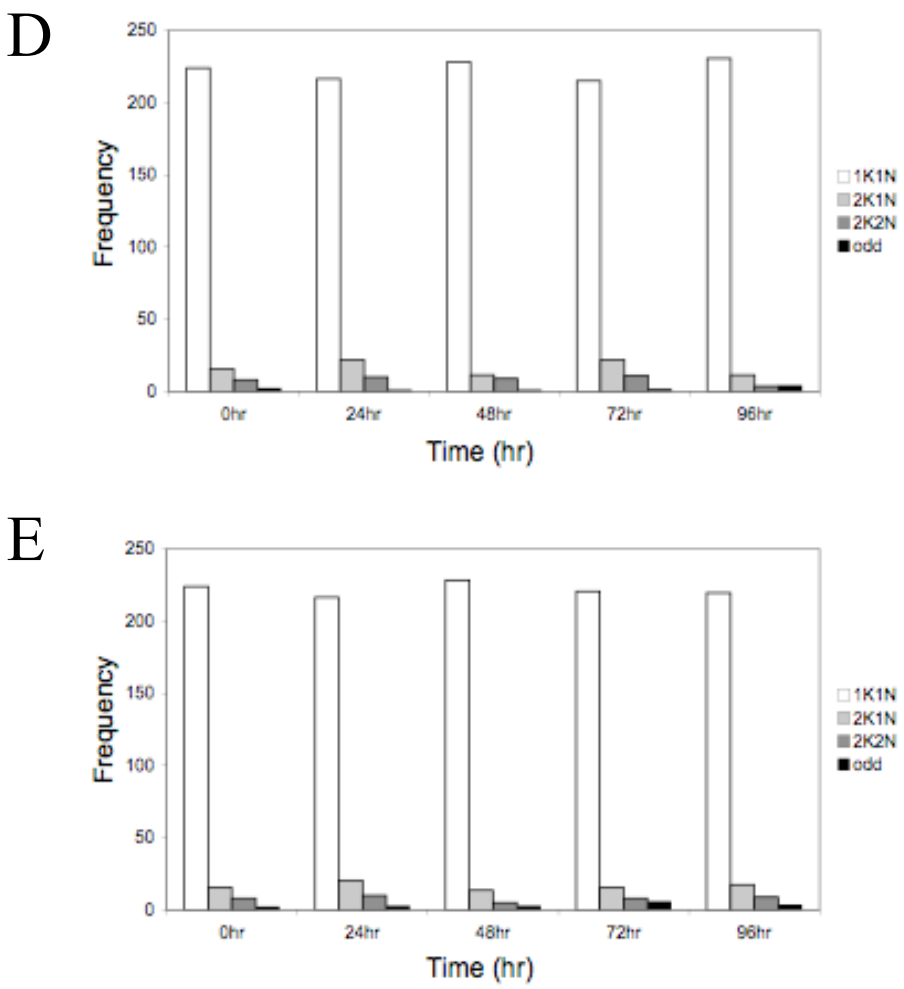


Figure 5.7 Ectopic expression of TbeIF6 protein with a Ty epitope tag. A. Growth analysis following induction of TbeIF6 ectopic expression with 10 μg/ml or 1 μg/ml tetracycline (+ Tet) at 0 hours and control, untreated (- Tet), cells. B. Growth of the parental cell line treated with 10 μg/ml or 1 μg/ml tetracycline and control, untreated, cells. The cell concentration of each culture was returned to ~1 x 10^5 cells/ml every 48 hours. C. Western blot analysis of RNA samples from uninduced (- Tet) and induced (+ Tet) cells of one TbeIF6 overexpression clone, using a mouse monoclonal antibody to the Ty epitope tag (BB2 antibody) and anti-mouse IgG-HRP secondary antibody. Arrowhead indicates the ectopic Ty epitope tagged TbeIF6 protein; asterisk indicates a cross-reactive band. D and E. Cell cycle analysis of 250 uninduced (D) and induced (E) cells at each time point, through staining of DNA with DAPI to determine nuclei and kinetoplast number (1K1N, 2K1N, 2K2N or odd number, where K is kinetoplast number and N is nuclei number).

Figure 5.7 (A) shows the cell growth after induction of TbeIF6 overexpression for one clone with two concentrations of tetracycline (10 μg/ml and 1 μg/ml), due to concerns that the 10 μg/ml concentration used in earlier experiments was too high. The other clone showed a similar effect on cell growth upon induction. It was clear

that induction of TbeIF6 overexpression caused an inhibition of cell growth, present to a smaller extent at 48 hours but increasingly apparent after 96 and 144 hours. The inhibition of cell growth with TbeIF6 protein overexpression was greater in the cells treated with 10 μg/ml tetracycline compared to those treated with 1 μg/ml. However, this is unlikely to be due to an unspecific effect of tetracycline because, except for at the 96 hour time point, the parental cells did not show the same growth difference with the different tetracycline concentration treatments (Figure 5.7 B). Therefore, it seems likely that this was due instead to a greater induction of ectopic TbeIF6 protein expression.

The inducible expression of the TbeIF6 ectopic copy possessing a Ty epitope tag was verified through western blot analysis using an antibody that recognises the Ty epitope tag (mouse monoclonal BB2 antibody) and an anti-mouse-FITC secondary antibody. Representative western blots from one experiment are shown in Figure 5.7 (C).

The predicted molecular weight of the protein encoded by Tb927.10.5300 in the TriTryp database is ~27 kDa. The ectopic Ty epitope tagged protein migrates a little heavier than this, presumably due to the presence of the Ty epitope (which has an estimated molecular weight of 1.167 kDa), and is indicated in Figure 5.7 (C) with an arrowhead. The expression of the ectopic copy was very tightly inducible with tetracycline because expression was not detected in any of the samples from uninduced cells. A cross-reactive band was also detected with these antibodies (see also Figure 4.11) migrating at around 70 kDa. However, its presence served as a useful control for the presence of protein in each lane.

An investigation into whether the inhibition of cell growth rate was attributable to a block in the cell cycle was made through analysis of cells at each time point. Cells were fixed with methanol, stained with the DNA stain, DAPI (4,6-diamidine-2,phenylindole), and the number of kinetoplasts and nuclei present in each cell scored to determine the cell cycle stage. The categories made were 1K1N (G1 and S), 2K1N (G2 and mitotic), 2N2N (post-mitotic; where K is kinetoplast and N is

nuclei), and any other configurations were classified as 'odd'. The frequency of each category in the samples of uninduced cells is shown in Figure 5.7 (D) and for induced cells in Figure 5.7 (E). As can be seen by comparison between the uninduced and induced cells, the accumulation of cells in a particular phase of the cell cycle after induction of ectopic protein overexpression was not observed.

5.6 Is TbeIF6 upregulated during cold shock?

As mentioned above, the investigation of this protein was initiated because the transcript was enriched in stumpy forms (Figure 5.1 B) and in archaeabacteria IF6 was found to be upregulated during cold shock (Benelli *et al.*, 2009). Therefore, an investigation was made of whether TbeIF6 can be likewise upregulated during cold shock.

5.6.1 Detection of TbeIF6 protein with an anti-eIF6 monoclonal antibody

For the analysis of the expression of eIF6 protein, determination of the levels of the endogenous protein would be most informative; therefore, an antibody to detect the TbeIF6 protein was sought. A monoclonal antibody raised in mice against an immunogen peptide of human eIF6 is commercially available (BD Biosciences). The peptide sequence of the immunogen is underlined in the human eIF6 sequence in Figure 5.4 (A) and demonstrated a good level of conservation with respect to the trypanosome sequence. Hence, this antibody was tested with *T. brucei* cell extracts and after some refinement a western blot protocol to successfully detect TbeIF6 was developed (see Section 2.7).

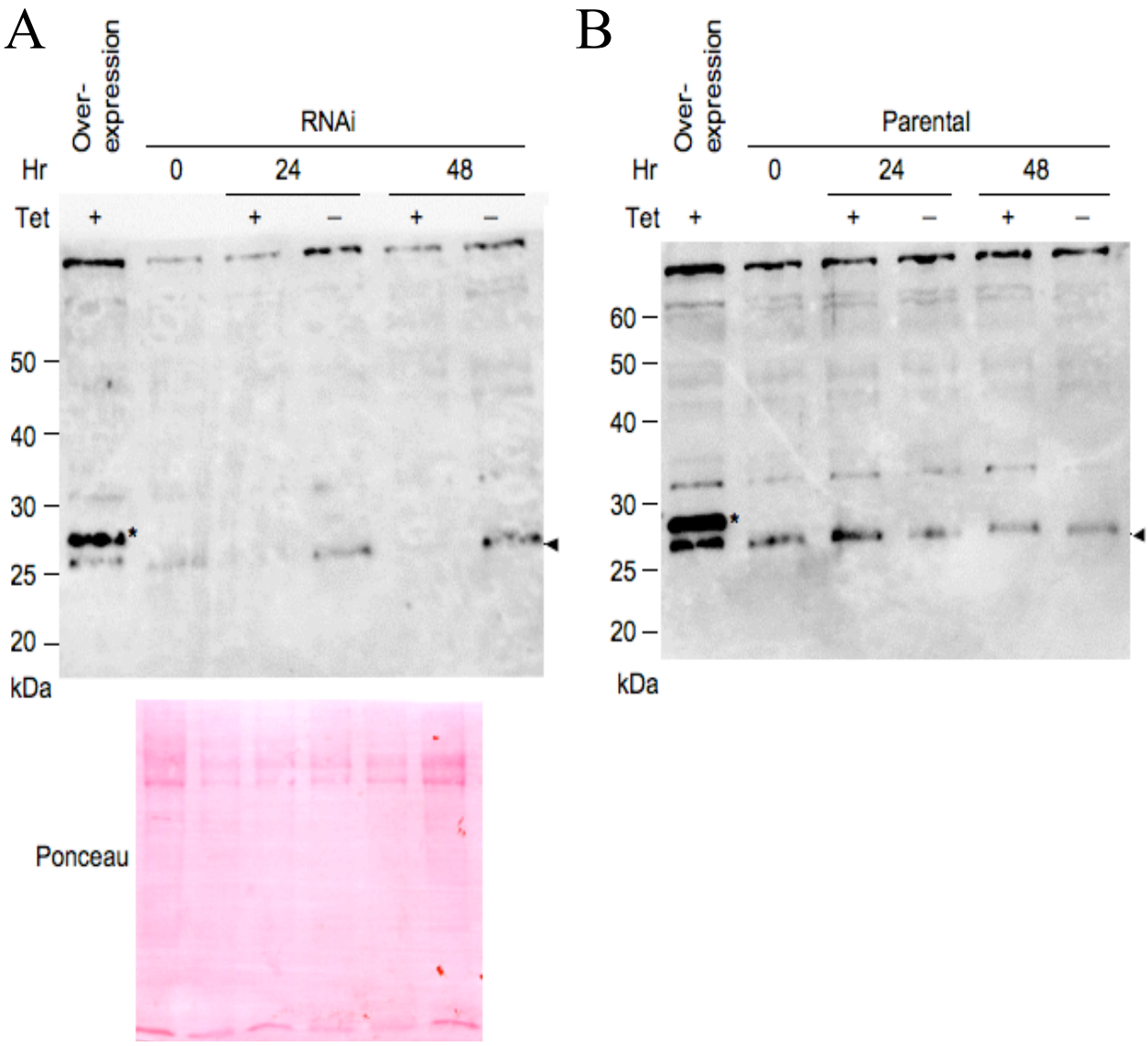


Figure 5.8 continues on the following page

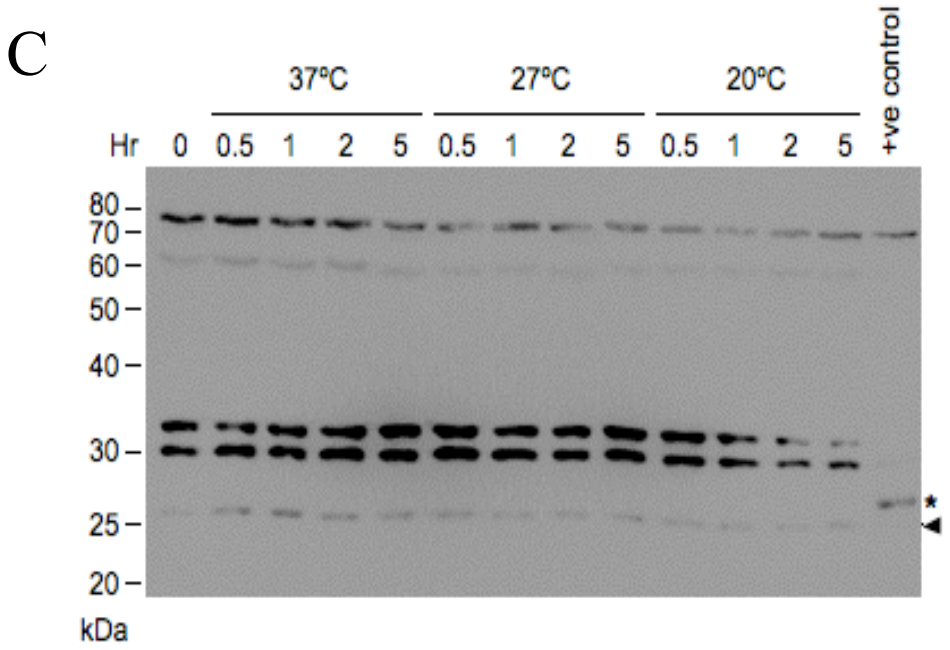


Figure 5.8 Use of a monoclonal mouse anti-eIF6 antibody to detect TbeIF6 protein expression during cold shock. Western blot analyses to detect ectopic Ty epitope tagged TbeIF6 protein and endogeneous TbeIF6 protein using a monoclonal mouse anti-eIF6 antibody raised against a human eIF6 immunogen (BD Biosciences) and anti-mouse IgG-HRP secondary antibody. Cell extracts are from the TbeIF6 RNAi cell line (A) or the parental cell line (B), either treated (+) or untreated (-) with tetracycline (Tet) at the times indicated in hours; or from bloodstream form cells incubated at 37°C, or transferred to 27°C, or 20°C for the time indicated in hours (C). Cell extract previously determined to be ectopically expressing TbeIF6 is included in each blot ('overexpression' or '+ve control' lanes). Ponceau staining of the membrane presented in A is to show protein loading. Asterisks indicate ectopic Ty epitope tagged TbeIF6; arrowhead: endogeneous TbeIF6; protein marker is in kilodaltons (kDa).

Western blot analyses showing the use of this antibody with an anti-mouse IgG-HRP secondary antibody are shown in Figure 5.8. Present on the blots shown in Figure 5.8 (A and B) is cell extract where expression of the ectopic Ty epitope tagged TbeIF6 protein was previously confirmed ('overexpression' lane) and it can be seen that the Ty epitope tagged TbeIF6 was clearly detected, the band migrating at the same size as detected with the BB2 antibody (Figure 5.7 C). This band is marked with an asterisk in Figure 5.8. A number of other bands were detected, but it would appear that endogenous TbeIF6 protein is the band that migrates just below the epitope tagged version (marked with an arrowhead in Figure 5.8). The additional cross-

reactive bands detected by these antibodies provided a useful means to assess protein loading.

Figure 5.8 (A) shows cell extracts from the TbeIF6 RNAi cell line treated or untreated with tetracycline. Although the signal for the endogeneous TbeIF6 was always low, it can be seen that there was a reduction in this signal in the extracts from cells induced for TbeIF6 RNAi. Figure 5.8 (B) shows cell extracts from the parental cell line treated or untreated with tetracycline. Here, it was apparent that there was no reduction in the signal of the band believed to be endogeneous TbeIF6.

Therefore, as this antibody successfully detects TbeIF6 protein, it could be used to assess TbeIF6 protein levels in response to cold shock treatment.

5.6.2 TbeIF6 protein expression during cold shock treatment

To assess the response of TbeIF6 protein expression in cells during cold shock, protein cell extracts were taken from *T. b. brucei* Lister 427 bloodstream form cells after growth at 37°C (0 hours) and at 30 minutes, 1, 2, and 5 hours following transfer of third of the cell population to 27°C or 20°C or continuation at 37°C.

This experiment was performed twice, and a representative western blot analysis of protein cell extracts with the human anti-eIF6 antibody is shown in Figure 5.8 (C). Again a positive control of cell extract where expression of the ectopic Ty epitope tagged TbeIF6 protein was previously confirmed ('+ve control' lane) was included. The ectopic TbeIF6 protein was detected in the western blot and is indicated with an asterisk in Figure 5.8 (C); this provides reassurance that the band detected migrating below this ectopic copy above the 25 kDa marker is endogeneous TbeIF6. Although the detected TbeIF6 signal is faint, it was observed in both experimental replicates that there was no change in TbeIF6 expression in response to cold shock of 27°C or 20°C.

Therefore, unlike the archeaebacteria IF6, TbeIF6 protein does not appear to be significantly upregulated during cold shock, at least not under the conditions investigated.

5.7 Is TbeIF6 involved in translation?

To determine whether TbeIF6 is involved in translation, two types of experiment were performed using the RNAi and overexpression cell lines described above. Firstly, the synthesis of new proteins after induction of TbeIF6 overexpression or RNAi was determined through radiolabelling with [^{35}S]-methionine. Secondly, the polysome profiles of cells were analysed following TbeIF6 overexpression or RNAi induction, as described above in Section 5.2.1.

5.7.1 Polysome profiles of cells following TbeIF6 RNAi or protein overexpression

The effect on translation of induction of TbeIF6 protein overexpression was assessed through analysis of polysome profiles (see Section 2.9). For this, after 48 hours growth either induced for TbeIF6 overexpression or uninduced, cells were treated with 100 μg/ml cycloheximide and harvested.

The uninduced control sample was always processed in parallel to the induced sample. However, unfortunately there were problems of an unknown technical reason meaning that each time only one of the samples produced a successful absorbance trace. Absorbance traces obtained from two uninduced cell samples are shown in Figure 5.9 (A), and from two induced cell samples are shown in Figure 5.9 (B). The absorbance traces at the right of Figure 5.9 (A and B) were obtained from samples processed together but are from cells harvested on different days. Confirmation of the induction, or not in the case of the uninduced cells, of ectopic TbeIF6 expression is shown in Figure 5.9 (E; left blot), with the samples corresponding to the absorbance traces marked with an asterisk at the base of the blot. The ponceau staining of the membrane is also shown to confirm the presence of proteins in the uninduced sample, this being revealed by the presence of the cross-reactive band.

Comparison of absorbance traces shown in Figure 5.9 (A) with (B) suggests a decreased level of 60S subunits, 80S monosomes and possibly polysomes in the induced cells. However, for a firm conclusion to be made this experiment would need to be repeated and absorbance traces obtained from uninduced and induced cells processed in parallel compared.

Additionally, the cells were analysed 48 hours after induction of TbeIF6 overexpression and at this time point there was a difference in the growth rate of uninduced and induced cells, which could affect the polysome profiles obtained. Therefore, cells were instead harvested at 24 hours following induction. Unfortunately, however, there was an unknown technical problem resulting in the absorbance trace from the uninduced cells being completely unsuccessful.

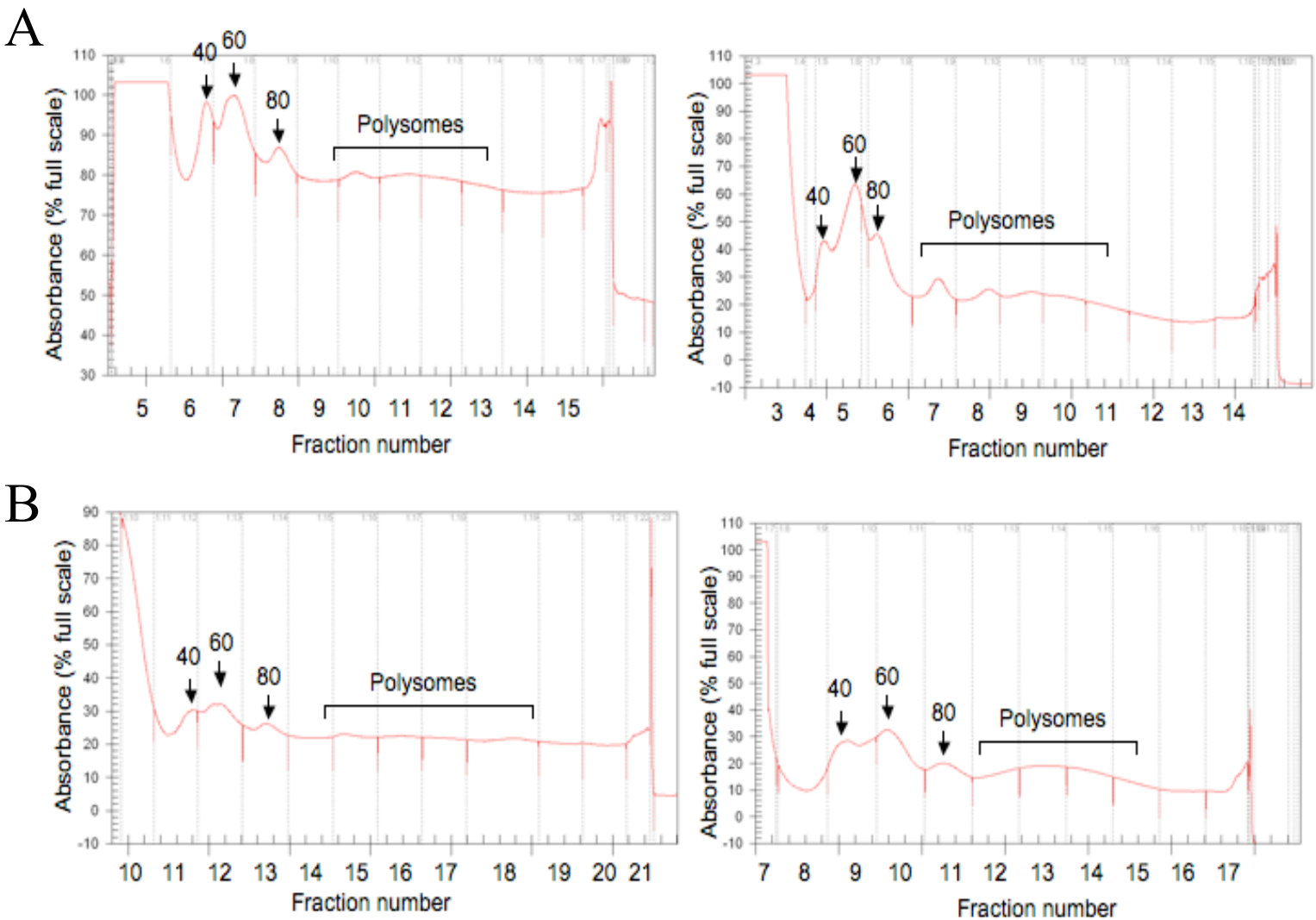
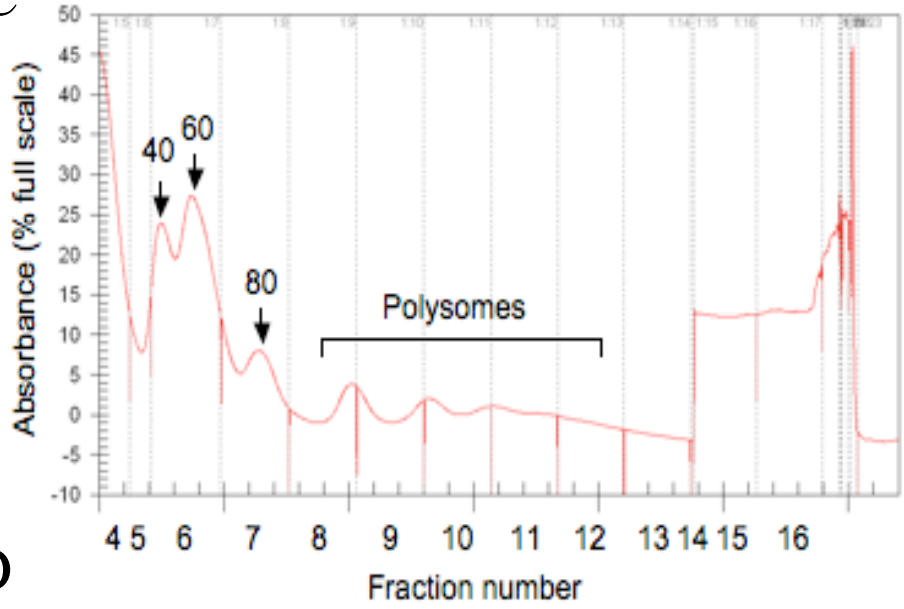


Figure 5.9 continues on the following page

C



D

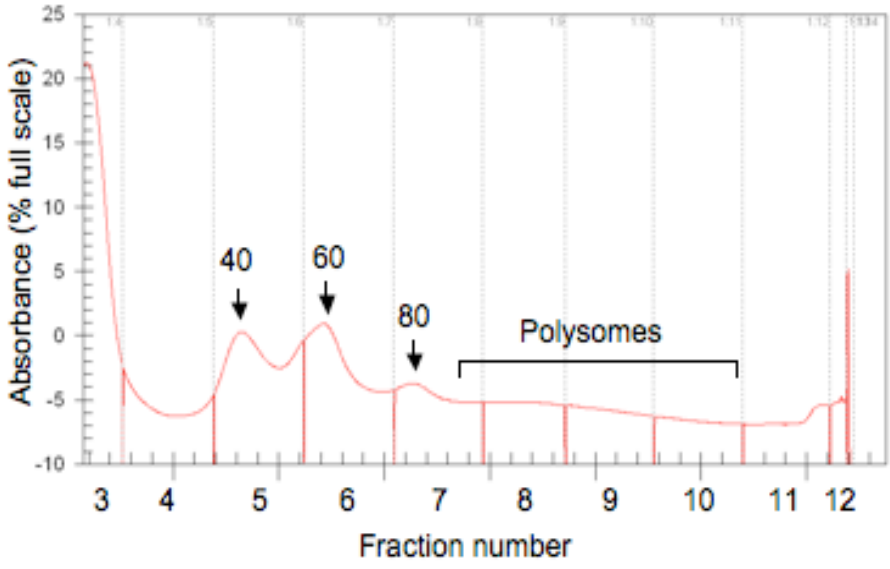


Figure 5.9 continues on the following page

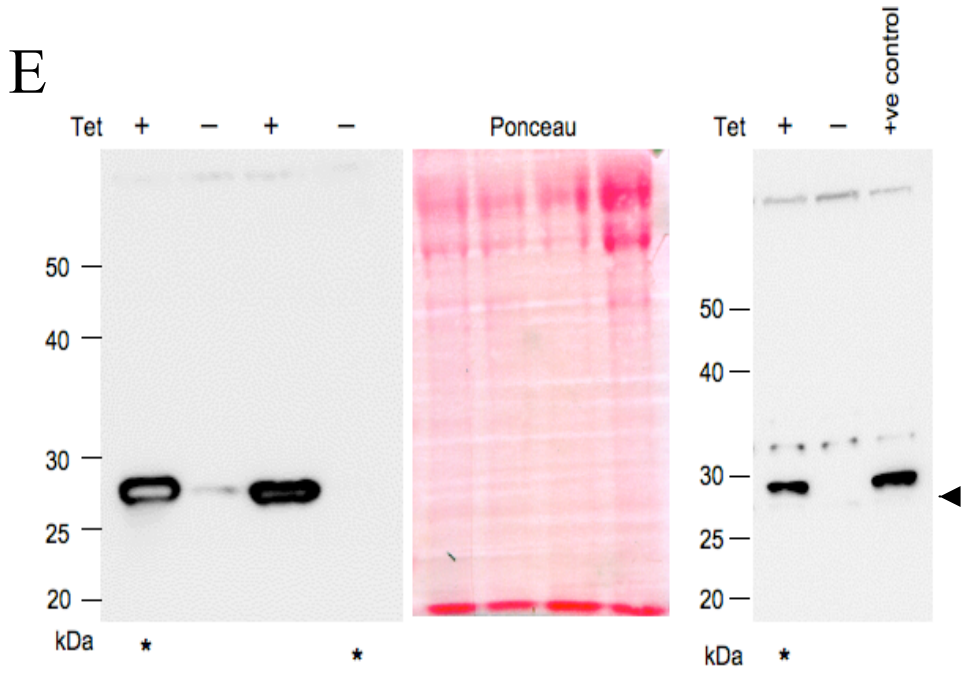


Figure 5.9 Polysome analysis after TbeIF6 ectopic overexpression or RNAi. Cells were harvested for polysome analysis either from control cell populations (uninduced; A), 48 hours after induction of TbeIF6 ectopic expression (B), or 24 hours after induction of TbeIF6 ectopic expression (C), or 48 hours after induction of TbeIF6 RNAi (D), after treatment with 100 μg/ml cycloheximide. Cell lysate was layered on to a 15-50% sucrose gradient, centrifuged at 40 000 rpm for 2 hours, and fractions taken at a set number of drops (see Section 2.9). Absorbance was measured at 254 nm throughout fractionation. Shown are the absorbance traces with red vertical lines separating the fractions generated. The absorbance peaks corresponding to the assumed 40S, 60S, 80S and polysome locations are indicated. E. Western blot analyses of cell extracts taken at the time of cell harvest, stained with BB2 antibody (left panel) or anti-eIF6 antibody (right panel). Ponceau staining of the membrane presented in the left panel is to show protein loading. Asterisks at the base of the left blot indicate samples corresponding to the absorbance traces in the right panels of A and B. Asterisk at the base of the right blot indicates sample corresponding to absorbance trace in C. Arrowhead: ectopic Ty epitope tagged TbeIF6; Tet: tetracycline; '+ve control': cell extract previously determined to express ectopic Ty epitope tagged TbeIF6; protein marker is in kilodaltons (kDa).

Figure 5.9 (C) shows the absorbance trace obtained from the induced cell sample. This may suggest that there is a decrease in the 80S monosomes present in these cells. This would be consistent with the role of eIF6 in other organisms as a molecule that prevents association of 40S and 60S subunits and therefore, the formation of 80S monosomes. The expression of the ectopic TbeIF6 protein was again confirmed by western blot analysis (Figure 5.9 E; right blot).

The effect of TbeIF6 RNAi induction on translation was also investigated through polysome analysis. Again, unfortunately there were problems with the control, uninduced, polysome analysis. Therefore, only the absorbance trace obtained from the induced cells (48 hours after induction) is shown in Figure 5.9 (D). In comparison to absorbance traces from other successful polysome analyses, it would suggest that here there was a reduction in the abundance of 60S monosomal subunits, 80S monosomes, and polysomes. This would be consistent with the proposed function of eIF6 in 60S biogenesis. However, this experiment would need to be repeated before any such conclusions could be made.

5.7.2 Protein synthesis following TbeIF6 RNAi or protein overexpression

The experiment of [^{35}S]-methionine radiolabelling was performed as described in Section 2.10 (performed by Dr. Terry Smith, University of St. Andrews). Each cell line investigated was either treated with 1 μg/ml tetracycline to induced protein overexpression/RNAi or left untreated and, ~24 hours later, following transfer into methionine-free media, incubated with 10 mCi [^{35}S]-methionine for 1 hour. Cell lysate was analysed by SDS-PAGE, stained by fluorography or with Coomassie.

This experiment was performed twice and the results from one representative experiment are shown in Figure 5.10 (A). As can be seen, induction of TbeIF6 RNAi or induction of TbeIF6 protein overexpression caused a decrease in protein synthesis compared to the uninduced controls. This decrease was not observed with the control, parental cell line treated with tetracycline.

In addition to SDS-PAGE analysis, radiolabel incorporation in triplicate samples taken from each cell line +/- tetracycline was quantified through by scintillation counting. The mean counts ± standard error are presented in Figure 5.10 (B) along with the radiolabel incorporation in cells treated with tetracycline as a percentage of that for the corresponding control, uninduced cells. All cell lines in the absence of

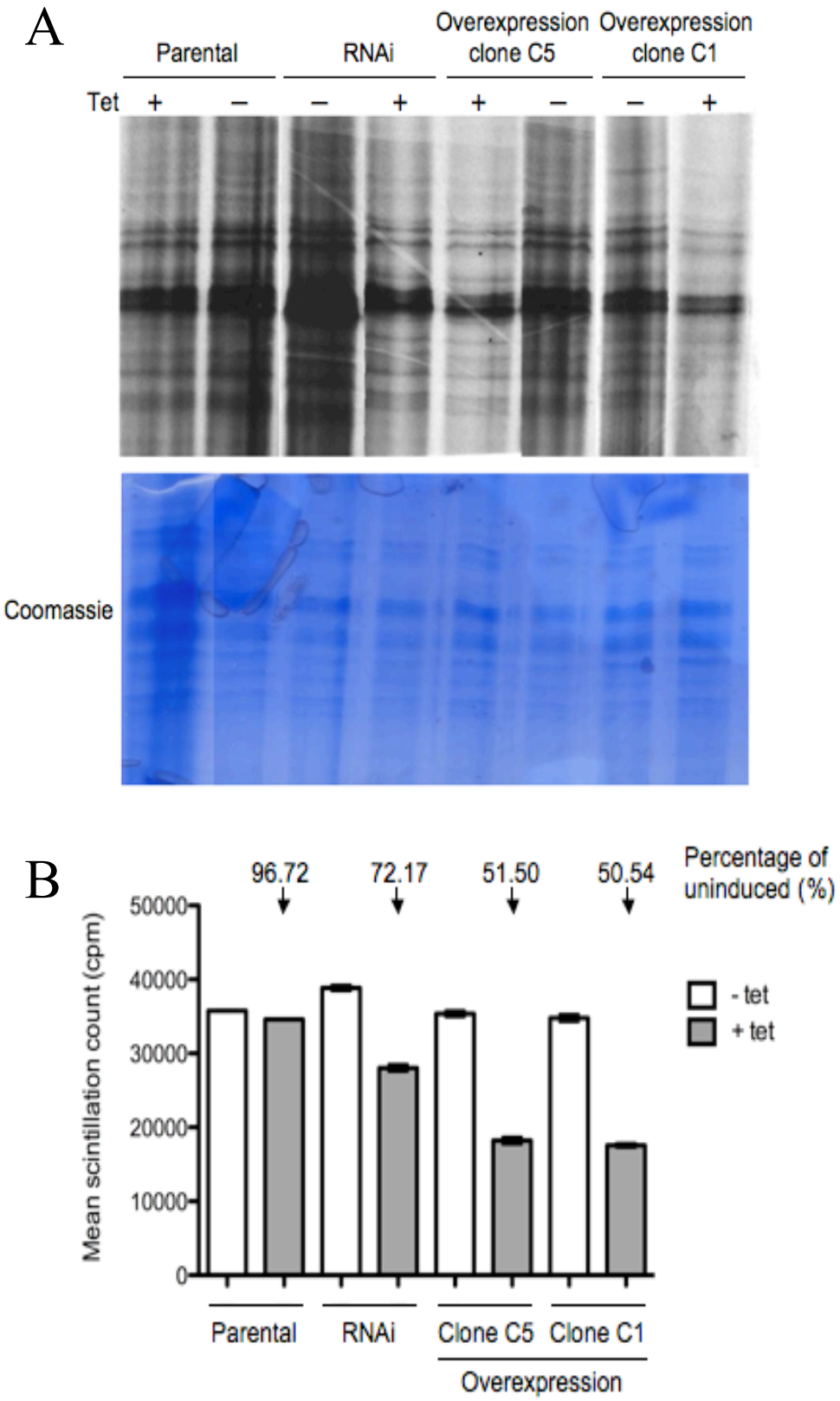


Figure 5.10 continues on the following page

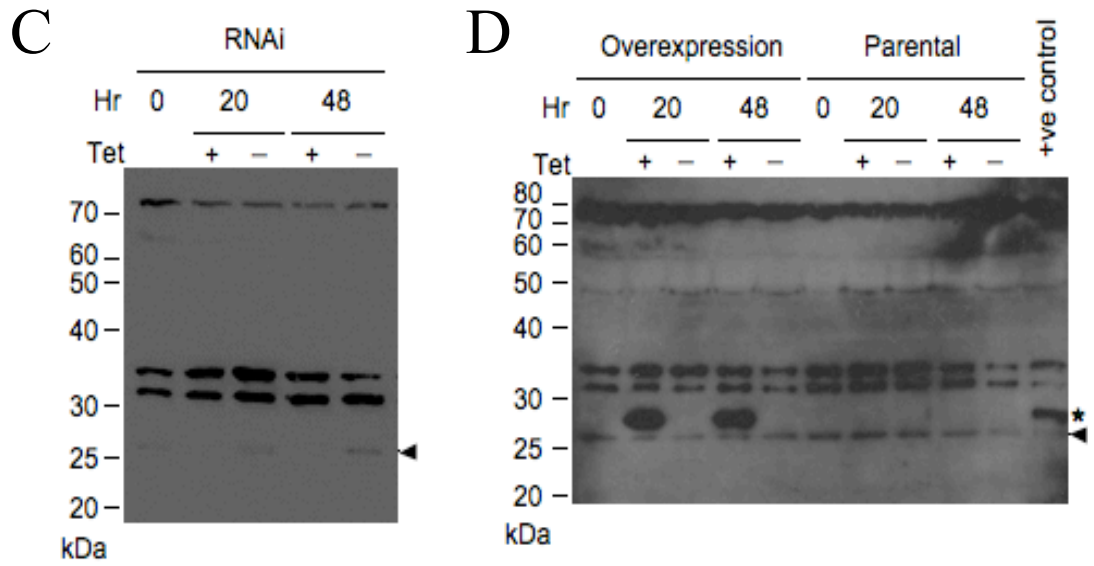


Figure 5.10 Analysis of protein synthesis following TbeIF6 RNAi or ectopic overexpression induction. A. TbeIF6 RNAi or TbeIF6 protein overexpression (clones C5 and C1) cell lines were either treated with 1 μg/ml tetracycline (+ Tet) to induce RNAi/protein overexpression or left untreated (- Tet) and ~24 hours later, following transfer into methionine-free media, incubated with 10 mCi [^{35}S]-methionine for 1 hour. The parental cell line was included as a control. Cell lysate was analysed by SDS-PAGE, stained by fluorography or with Coomassie. B. Mean scintillation counts ± standard error of triplicate samples taken from each cell line +/- tetracycline. Above is shown the radiolabel incorporation in cells treated with tetracycline as a percentage of that for the respective control, uninduced cells. C and D. Western blot analyses with extracts from cells prior to induction (0 hours), prior to radiolabelling (20 hours) and 48 hours following induction for the radiolabelling experiment presented in A and B. A mouse anti-eIF6 antibody and anti-mouse IgG-HRP secondary antibody were used to detect the ectopic Ty epitope tagged TbeIF6 protein (indicated with an asterisk) and endogeneous TbeIF6 protein (indicated with arrowheads). Protein size is shown in kilodaltons (kDa).

tetracycline exhibited a similar incorporation of radiolabel. This was also true for the parental cell line treated with tetracycline (96.72% of the incorporation of untreated cells), demonstrating that treatment with tetracycline does not significantly affect radiolabel incorporation. However, induction of TbeIF6 RNAi caused a reduction of radiolabel incorporation to 72.17% of the uninduced control. Likewise, induction of TbeIF6 overexpression resulted in a reduction of radiolabel incorporation to a greater extent (51.50% and 50.54%, for clones C5 and C1 respectively).

In this experiment, the induction of TbeIF6 RNAi or protein overexpression at 20 hours and 48 hours after induction was confirmed through western blot analysis using a mouse anti-eIF6 antibody raised against the human eIF6 protein (BD Biosciences) and an anti-mouse IgG-HRP antibody. Shown in Figure 5.10 are the resulting western blots for TbeIF6 RNAi (C) and for one of the TbeIF6 overexpression clones (D). In both blots the endogenous TbeIF6 protein is indicated with the arrowhead and the ectopic Ty epitope tagged TbeIF6 is indicated with an asterisk. Although the signal of the endogeneous TbeIF6 protein is weak in the blot of the RNAi samples, there was clear depletion in the tetracycline treated samples. The ectopic expression of TbeIF6 was apparent in the tetracycline treated samples in the blot shown in Figure 5.10 (D) and absent from the untreated samples. Here, a positive control of sample from cells previously determined as ectopically overexpressing the TbeIF6 protein was included. The parental cells showed no change in TbeIF6 protein expression, as expected.

Despite the difficulties with these polysome analyses, from the data presented in Figure 5.10 it is apparent that TbeIF6 is involved in protein synthesis, as induction of RNAi and protein overexpression result in greatly reduced levels of protein synthesis. Analysis of the polysome profiles of induced cells (Figure 5.9) suggests that this may be due to decreased levels of ribosomal subunits/monosomes, leading to a decrease in polysomes. However, at this stage firm conclusions cannot be drawn from these experiments, these would need to be resolved and the experiments repeated to identify the mechanisms by which TbeIF6 is involved in protein synthesis.

5.8 Is TbeIF6 involved in differentiation?

Due to the strong enrichment of the *TbeIF6* (Tb927.10.5300) transcript in stumpy forms compared to other life-cycle stages (Figure 5.1 B), it was postulated that this protein might play a role in differentiation to the procyclic form. To investigate this possibility, the differentiation ability of cells induced for TbeIF6 RNAi or overexpression was tested. If the TbeIF6 protein is involved in differentiation then

the cells induced for RNAi should exhibit a reduced ability to differentiate and those induced for protein overexpression might exhibit an increased ability.

For this experiment, cells were either induced (treated with 1 μg/ml tetracycline) or uninduced and both populations were either treated with 6mM *cis*-aconitate or untreated. After tetracycline/*cis*-aconitate addition, all cells were transferred to 27°C and differentiation was assessed through EP-procyclin expression 72 hours later. As a control, the differentiation of the parental cells under the same conditions was also assessed. EP-procyclin expression was determined by fluorescence-activated cell sorting after staining with an EP-procyclin antibody, a FITC-labelled secondary antibody and a DNA stain, DAPI (see Section 2.8.3). A sample of procyclic cells was also stained to serve as a positive control for EP-procyclin expression.

The results obtained from fluorescence-activated cell sorting were analysed in comparison to the procyclic cell sample. The strategy used for analysis of each sample is given in Figure 5.11. For the procyclic form cell sample two cell populations were gated on the plot of forward and side scatter (Figure 5.11 A). These were of the 'intact' cells (94% of the total population) and 'compromised' cells (3.93%), determined by the cell size and together with the DAPI staining of each population. The 'intact' population was further analysed through selection of 'singlet' cells (87.7% of the 'intact' population) to exclude any aggregated cells (Figure 5.11 B). The DAPI and FITC staining of this population was then analysed for cells positive for FITC staining (98.3% of the 'singlet' population; Figure 5.11 C). Therefore, 98.3% of the intact, singular material in the procyclic form sample was positive for EP-procyclin expression.

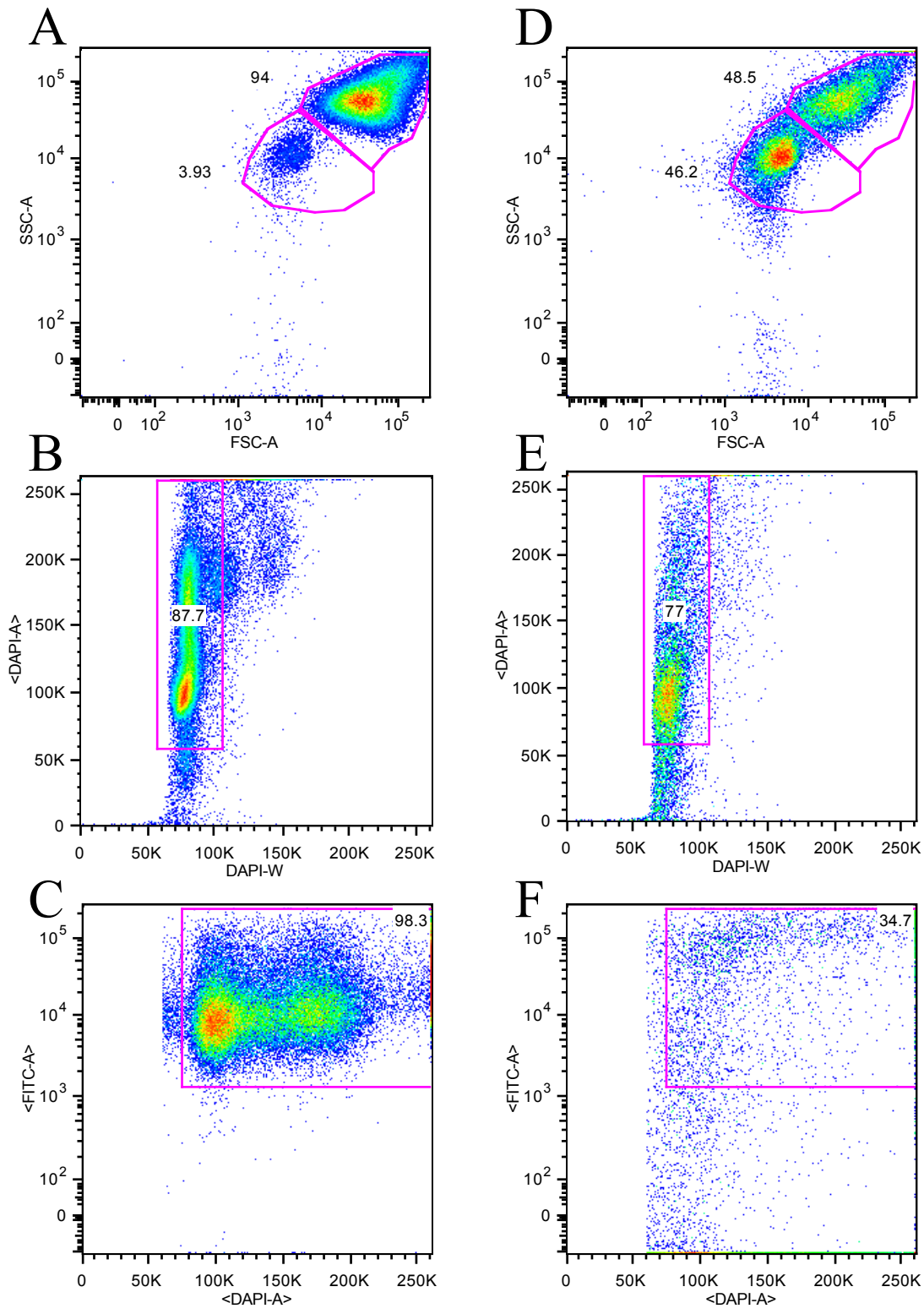


Figure 5.11 Pseudo-colour dot plots obtained from fluorescence-activated cell sorting analysis. Cells were analysed after staining with an EP-procyclin antibody, FITC-labelled anti-mouse IgG antibody and a DNA stain, DAPI (see Section 2.8.3). A-C: procyclic forms cell sample (positive control); D-F: one replicate of parental cells in the absence of tetracycline but treated with 6mM *cis*-aconitate. Pink boxes (gates) represent cell populations selected and percentages of cells within that population are shown. SSC: side scatter; FSC: forward scatter.

These gating parameters were transferred on to the other samples and the suitability of the gates for analysis of the other samples was assessed. As an example, one replicate of the parental cells untreated with tetracycline but treated with *cis*-aconitate is shown with the 'intact' and 'compromised' gates (Figure 5.11 D), 'singlet' gate (Figure 5.11 E), and cells positive for FITC staining (Figure 5.11 F). Each experimental condition of each cell lines was performed in triplicate. The mean for each of the percentage of intact, singular material positive for EP-procyclin expression is shown in Table 5.2.

Table 5.2 Percentage of cells positive for EP-procyclin expression in the intact, singular material, as determined by fluorescence-activated cell sorting[a].

	+ Tet		- Tet	
Cell line	+ CA	- CA	+ CA	- CA
Parental	33.7	0.4	38.4	0.5
RNAi	63.8	3.9	66.1	4.0
Overexpression	15.5	0.2	17.7	0.2

[a] Tet: 1 μg/ml tetracycline; CA: 6mM *cis*-aconitate; percentages are the mean obtained for triplicate cultures

Notably, the TbeIF6 overexpression cell line had a very low level of differentiation competence (17.7% of cells positive for EP-procyclin expression in uninduced cells treated with *cis*-aconitate). The parental cell line had a greater differentiation competence (38.4%) and the RNAi cell line even greater (66.1%). All cell lines exhibited a minimal differentiation competence when untreated with *cis*-aconitate. However, induction of TbeIF6 RNAi or overexpression did not greatly affect the differentiation ability of the respective cell line. In both cases there was a minor reduction in the percentage of EP-procyclin positive cells (66.1% to 63.8% and 17.7% to 15.5% respectively), but this was also observed for the parental cell line treated with tetracycline (38.4% to 33.7%).

Therefore, the result of this experiment suggests that induction of either TbeIF6 RNAi or overexpression does not detectably affect the ability of cells to differentiate to the procyclic form.

5.9 Discussion

Translational activity is often controlled through regulation of components of the translational machinery. Commonly regulated proteins include a member of the cap-binding complex, eIF4E, and eIF2α by phosphorylation. Additionally, other proteins such as eIF6 can regulate the formation of functional 80S ribosomes. As an initial assessment of whether any of these proteins may be involved in translational control in stumpy form *T. brucei*, the expression profiles of some regulatory protein homologues were investigated with northern blot analysis. Because stumpy forms exhibit translational repression (Brecht and Parsons, 1998), any transcripts enriched in this lifecycle stage are likely to be interesting for further study.

5.9.1 TbeIF4Es

It was particularly interesting that one of the four *T. brucei* eIF4E homologues, TbeIF4E4, was identified as enriched in stumpy forms by northern blot analysis (Figure 5.1 A). It has been previously suggested that the multiple trypanosomatid eIF4F isoforms could be associated with the different lifecycle stages (Dhalia *et al*., 2005), however, Freire *et al*. (2011) reported that there was little variation in the mRNA levels of each TbeIF4E between slender and procyclic forms. In contrast, here it was observed that *TbeIF4E1* was enriched in slender forms, *TbeIF4E2* and *TbeIF4E3* downregulated in procyclic forms, and *TbeIF4E4* enriched in stumpy forms (Figure 5.1 A). It is also possible that these homologues may exhibit different functions, such as acting as a negative regulator of translation by binding to the mRNA cap but not eIF4G, similarly to the binding pattern of 4EHP. However, an investigation of potentially conserved cap-binding or eIF4G-binding residues in each homologue (Table 5.1) indicated that each homologue exhibited varying degrees of conservation and, therefore, did not help identify any such function. Additionally, this analysis was hampered by the possibility that different residues may be involved in *T. brucei* eIF4E interactions, due to the use of a highly modified mRNA cap structure, cap 4, and the presence of multiple eIF4G homologues (Yoffe *et al*., 2009).

As TbeIF4E4 was enriched in stumpy forms it was selected for further analysis using RNAi-mediated transcript ablation. A preliminary analysis of the effect of RNAi-mediated transcript ablation revealed that TbeIF4E4 knockdown produced a decrease in cell growth rate that was apparent at 72 hours but also present to a smaller degree at 48 hours after induction (Figure 5.2). This result was very similar to the effect on cell growth of TbeIF4E4 RNAi in slender bloodstream forms reported elsewhere (Freire *et al*., 2011). Freire *et al*. (2011) also reported decreased protein synthesis following induction of TbeIF4E1/TbeIF4E4 double RNAi. Consistent with this, polysome analysis of cells induced for TbeIF4E4 RNAi indicated a decrease in polysome abundance and, therefore, a role in translation (Figure 5.3). However, this polysome analysis was inconclusive due to a problem with the resolution of the profile obtained for the uninduced control (Figure 5.3). Only this preliminary analysis was made, due to the discovery that another group was working on these eIF4E homologues.

5.9.2 TbeIF2Ks and putative deoxyhypusine synthetase

Due to the identification that eIF2α kinases are involved in differentiation in *T. cruzi* and *Leishmania* (see Section 1.7.5), it is unfortunate that the northern blot analyses for the *T. brucei* eIF2α kinases were unsuccessful at determining lifecycle abundance. However, a later Illumina digital tag RNA sequencing experiment (see Chapter 6) indicates that the *TbeIF2K2* (Tb927.4.5200) transcript is enriched in stumpy form polysome-associated RNA compared to slender form polysome-associated RNA (4.01 fold-change, $p = 0$), and so this transcript might exhibit increased translation in stumpy forms. It seems likely that these proteins may be interesting for further study in stumpy form *T. brucei*, due to the recently discovered role for these proteins in differentiation in closely related organisms. Likewise, the putative deoxyhypusine synthetase (Tb927.10.2750) may also be interesting for further study, due to their slight enrichment in stumpy forms (Figure 5.1 B), and the reported drug target potential of deoxyhypusine synthetase proteins in *Plasmodium* and *Leishmania* (Njuguna *et al.*, 2006; Specht *et al.*, 2008; Chawla *et al.*, 2010).

5.9.3 TbeIF6

One eIF6 homologue was identified in the *T. brucei* genome, this being Tb927.10.5300, and was, therefore, named TbeIF6. The sequence of this protein was also predicted to form the highly conserved pentein structure of eIF6 (Groft *et al*., 2000) (Figure 5.4). Interestingly, it was discovered that the mRNA of this gene was strongly enriched in stumpy forms (Figure 5.1 B). Due to the anti-association factor property of eIF6, it provides a connection between 60S ribosome biogenesis and translation, which may be regulated by extracellular signalling (Miluzio et al. 2009). Therefore, this appeared to be an interesting protein to investigate further by RNAi-mediated transcript ablation and ectopic overexpression of a Ty-epitope tagged protein copy.

Induction of either TbeIF6 RNAi or ectopic overexpression resulted in a decrease in the growth rate of slender bloodstream forms (Figures 5.5 A and 5.7 A). Following induction, the decreased growth rates were present to a small degree at 48 hours post induction and growth inhibition steadily increased at later time points measured. The decreased growth rates were likely due to reductions in the level of protein synthesis in induced cells (Figure 5.10), indicating a role for TbeIF6 in translation. Polysome profiles indicated that the reduced protein synthesis was possibly due to perturbations of translation in induced cells (Figure 5.9), however, technical difficulties meant that profiles obtained from uninduced and induced cells could not be directly compared and, thus, firm conclusions could not be made. The indication that there was a decreased level of 60S subunits, 80S monosomes and possibly polysomes in the cells induced for TbeIF6 RNAi (Figure 5.9 D) has also been reported following eIF6 RNAi induction in yeast (Si and Maitra, 1999), supporting this suggestion.

Due to the enrichment of *TbeIF6* in stumpy forms, the effect of RNAi and protein overexpression on differentiation capacity was assessed. However, no effect on differentiation to the procyclic form was observed (Table 5.2). It should be noted, however, that the methodology used in this analysis was not optimal, because the cells were examined for differentiation 72 hours after induction of RNAi or protein

overexpression. At that time there is a significant decrease in the growth rate (Figures 5.5 A and 5.7 A), meaning that there may have been secondary effects on differentiation following induction. Also, there was a high level of cell death in each treatment condition, possibly due to the transfer of cells to 27°C. The mean percentage of each bloodstream form population assumed to represent dead cells was 43.6, in contrast to 3.93% for the procyclic form population. The bloodstream form cells were believed to have a low differentiation capacity, so analysis was made at 72 hours after *cis*-aconitate treatment to maximise the percentage of differentiated cells. However, the low percentages of differentiated cells observed indicate that this approach may not have been successful. To resolve this question, it would be necessary to analyse the effect of RNAi and protein overexpression on differentiation capacity in pleomorphic trypanosomes. However, there was great difficulty in obtaining the transfectant RNAi and protein overexpression cell lines using monomorphic cells, so these constructs would not be ideal for transfection using pleomorphic cells as this procedure is of a lower efficiency.

The reason for the observation of the appearance of a higher molecular weight TbeIF6 RNA species upon induction of TbeIF6 RNAi (Figure 5.5 B) is unclear. Investigation of another mRNA, *TbeIF4E4*, in the cells induced for TbeIF6 RNAi, revealed that the mRNA was the same molecular weight as in uninduced cells (Figure 5.6 A), suggesting that the appearance of higher molecular weight mRNAs was not a general consequence of TbeIF6 RNAi induction. However, it could be possible that this effect only occurs in a subset of transcripts, that includes *TbeIF6* but not *TbeIF4E4*.

A role for TbeIF6 in RNAi was also investigated and induction of TbeIF6 RNAi may affect the RNAi pathway because there was a possible increase in high molecular weight *INGI* transcripts at 48 hours after TbeIF6 RNAi induction (Figure 5.6 C). This would be consistent with some other studies, as eIF6 has been reported to be involved in miRNA silencing and RNAi in other organisms (see Section 1.5.4). However, further experiments are needed to confirm the involvement of TbeIF6 in the RNAi response. Firstly, the northern blot analysis to identify the high molecular

weight *INGI* transcripts needs to be repeated to obtain a more conclusive result. Also, *SLACS* transcripts could be investigated in a similar manner as these are normally suppressed through RNAi (Shi *et al*., 2004). RNAi function could also be tested through the ability of cells induced for TbeIF6 RNAi to produce the FAT phenotype following electroporation of *α-tubulin* dsRNA, similarly to the analysis performed for *AGO1* RNAi (Shi *et al*., 2004).

The successful use of the anti-eIF6 antibody (BD Biosciences) to detect TbeIF6 was particularly valuable. Although other higher molecular weight proteins were detected with the antibody, these appear to be de to cross-reactivity of the antibody, because only the signal believed to be TbeIF6 was lost upon induction of TbeIF6 RNAi (Figure 5.8 A). This antibody would be very informative for other analyses of TbeIF6, such as investigation of the lifecycle abundance of TbeIF6 protein, to determine whether there is enrichment in stumpy forms corresponding to the enrichment of *TbeIF6* mRNA in this lifecycle stage. One analysis that this antibody allowed was an investigation as to whether TbeIF6 is upregulated during cold shock, similarly to aIF6 in Archaea (Benelli *et al*., 2009).

For this experiment, the cold shock temperatures chosen were 27°C and 20°C, the temperature found to increase sensitivity to *cis*-aconitate (Engstler and Boshart, 2004). The time-range of cold shock treatment ranged from 30 minutes to 5 hours, because in Archaea upregulation of aIF6 was observed at 30 minutes of treatment (Benelli *et al*., 2009), and additional time-points were made after that time to allow for a possible increased time for upregulation in *T. brucei*. However, no upregulation of TbeIF6 was observed in two experimental replicates at either cold shock temperature or at any time investigated (Figure 5.8 C).

It appears possible that eIF6 may contribute to the translational repression observed in stumpy forms (Brecht and Parsons, 1998), due to the enrichment of *TbeIF6* in this life-cycle stage (Figure 5.1 B) and the decreased protein synthesis observed with TbeIF6 protein overexpression (Figure 5.10) This might be unusual because increased levels of eIF6 have been associated with proliferating cells in other

organisms (Miluzio et al. 2009), but stumpy form cells are non-proliferative. Therefore, it would have been interesting if TbeIF6 RNAi or overexpression affected the ability of cells to differentiate to the procyclic form. In the experiments performed here this was not observed however (Table 5.2).

Chapter 6

Results

Analyses of global translation in stumpy forms

6.1 Introduction

A comparison of the polysome profiles for different life-cycle stages has demonstrated that stumpy form *T. brucei* shows a dramatic decrease in translational competence compared to the slender form and the procyclic form (Brecht and Parsons, 1998). Through the comparison of 2D-PAGE with radiolabelled methionine-proteins for these life-cycle stages, stumpy forms also showed decreased protein synthesis (Shapiro and Kimmel, 1987). As stumpy forms are non-proliferative but slender and procyclic forms are proliferative this decrease in translation and therefore, protein synthesis, is perhaps expected. However, there are some transcripts that are expressed and so translated in stumpy forms, for example, the ESAG9 family of proteins (Barnwell *et al*., 2010) and PAD1 (Dean *et al*., 2009). These must be part of a small subset of transcripts that escape the normal translational repression in stumpy forms. As stumpy forms are believed to be preadapted for transmission to the insect vector, these transcripts could be genes involved in parasite transmission or competence for differentiation to the procyclic form. Therefore, these genes could be of great interest and importance. The aim of the experiments outlined in this chapter was to identify these genes that escape the translational repression in stumpy forms.

To investigate the global patterns of translation during differentiation and between life-cycle stages, analyses were made between the newly synthesised proteins during the differentiation of stumpy forms to procyclic forms and the polysome-associated transcripts in different life-cycle stages. These global analyses involved two experiments. The first experiment involved comparison of SDS-PAGE with [^{35}S]-methionine-labelled proteins collected at various time-points throughout synchronous differentiation to procyclic forms stimulated by addition of *cis*-aconitate. This sought to identify any increase in the expression of different sets of proteins, or individual proteins, expressed in stumpy forms or throughout differentiation. To quantitate changes, radiolabelled proteins at each time-point were precipitated with TCA and radiolabel incorporation of the samples measured, to

investigate the overall amount of protein synthesis and the dynamics of this throughout differentiation.

A second approach involved comparison of the abundance of mRNA transcripts in four cell samples: total RNA from bloodstream-form monomorphic cells, total RNA from stumpy form cells, polysomal mRNA from bloodstream form monomorphic cells, and polysomal mRNA from stumpy forms. The aim of this experiment was to identify which transcripts in stumpy forms showed polysome-association, and so were actively translated. It was hoped this would assist the identification of genes important for parasite transmission or competence for differentiation to procyclic forms.

6.2 Radioactive methionine labelling throughout stumpy form to procyclic form differentiation

To investigate if there are specific proteins or protein subsets that escape translational repression in stumpy forms and if this profile changes throughout the differentiation to procyclic forms, newly synthesised proteins were analysed by radiolabelling with [^{35}S]-methionine. Trypanosomes of the pleomorphic *T. b. brucei* AnTat1.1 strain, were purified from mice 6 days post-infection (when the cells were >80% stumpy by morphology), resuspended in HMI-9 media and treated with 6mM *cis*-aconitate; except for a small volume which were left untreated as a control. The remainder of the experiment was performed by Dr. Terry Smith (University of St. Andrews), as described in Section 2.10. At 1, 2, 4, 6, 8, 12, and 24 hours after addition of *cis*-aconitate, cell lysate samples were made following incubation with 10 mCi [^{35}S]-methionine for one hour. Proteins in each sample were analysed by SDS-PAGE, stained with Coomassie or analysed by fluorography. This experiment was repeated twice, with similar results observed both times; the results of one representative experiment are discussed below.

In this experiment, cells were pulse labelled with radiolabelled methionine such that only newly synthesised proteins would be labelled. Therefore, by analysing the

profile of radiolabelled proteins by fluorography it was possible to examine the proteins synthesised at different time points throughout differentiation. As expected, there was an overall increase in the amount of radiolabelled protein, and therefore protein synthesis, in the time-course of differentiation from stumpy forms (represented by the 0 hour time-point) to procyclic forms (Figure 6.1 A; shown is only the 8 hour time-point). This was not accounted for by an increase in cell number because during differentiation with *cis*-aconitate the first cell divisions occur after 10 hours, whereas here the increase in radiolabelled protein was observed at 8 hours. However, the resolution of the gels was limited and no differential expression of specific proteins or subsets of proteins was detected during the time course. Analysis of the total loaded protein by Coomassie-staining (Figure 6.1 B) revealed some small variations in the loading of protein between samples, however, this variation could not account for the observed increase in protein synthesis throughout differentiation detected by fluorography.

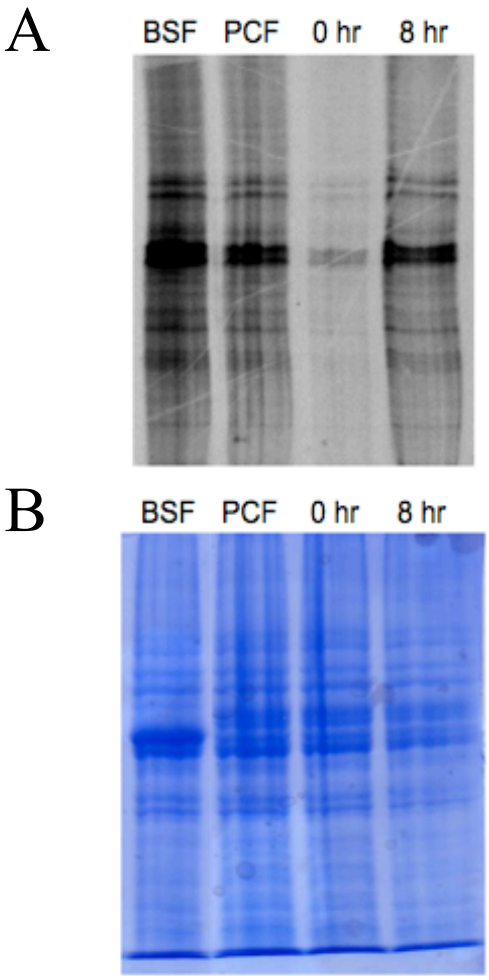


Figure 6.1 Analysis of protein synthesis throughout differentiation by [^{35}S]-methionine labelling. 6mM *cis*-aconitate was added to pleomorphic stumpy form *T. b. brucei* AnTat1.1 cells, followed by labelling with [^{35}S]-methionine for one hour at the time indicated post-addition of cis-aconitate (hr). The 0 hr time-point represents stumpy form cells. Monomorphic slender bloodstream form (BSF) and established procyclic form (PCF) cultures were also analysed. Protein samples were separated by SDS-PAGE and stained either by fluorography (A) to detect [^{35}S]-labelled proteins or with Coomassie (B) to show loading.

The radiolabelling of newly synthesised proteins in this experiment was also analysed by scintillation counting after TCA precipitation of the proteins in each sample. This allowed the quantification of the amount of incorporation of the radiolabel, and so the amount of new protein synthesis for each sample. After cells were incubated for one hour with [^{35}S]-methionine, triplicate samples were taken for TCA-precipitation. The resulting mean scintillation counts are shown in Figure 6.2,

along with the mean count as a percentage of the mean count for established procyclic form cells (100%; right hand axis).

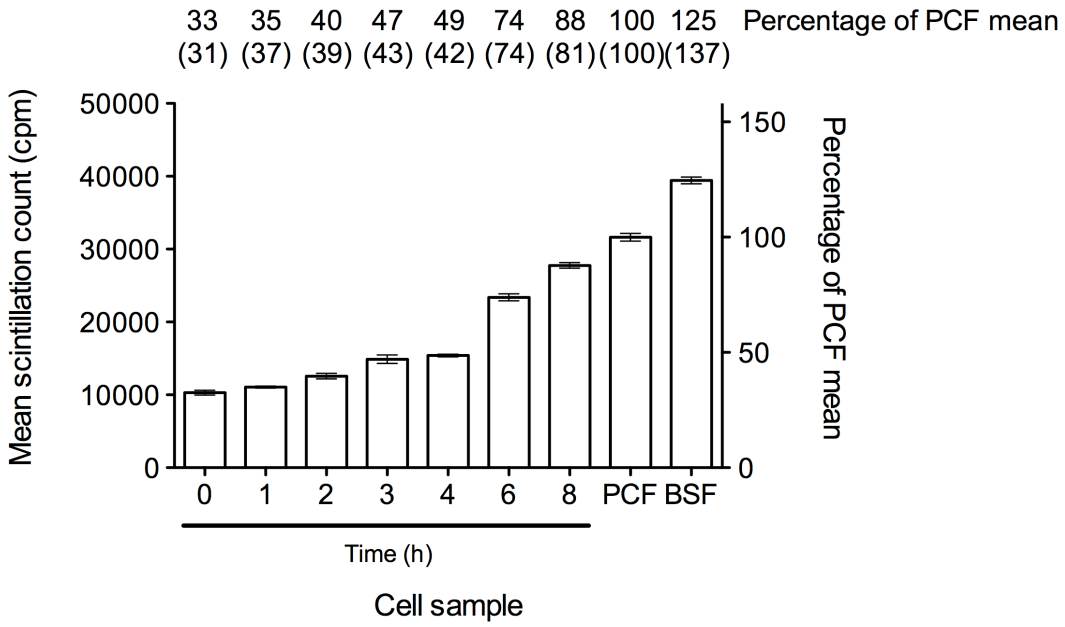


Figure 6.2 Mean scintillation counts of [^{35}S]-methionine incorporation during differentiation from stumpy form to procyclic form. Triplicate samples of cells labelled with [^{35}S]-methionine for one hour were TCA-precipitated and scintillation counts measured to determine protein synthesis. Shown is the mean scintillation count ± standard error. Samples were taken at various time-points, shown in hours (h), post-addition of *cis*-aconitate (6mM); 0 h represents stumpy form cells. Established procyclic form cells (427 449 strain; PCF) and slender bloodstream monomorphic cells (single marker strain; BSF) were also analysed for comparison. Shown on the right *y*-axis is the mean count as a percentage of the mean count for the established procyclic form cells (100%), these values also being shown above each data point. The experiment was performed twice and the numbers in parentheses are those obtained from the second experiment.

Stumpy form cells, represented by the 0 hour time-point, showed a relatively low level of protein synthesis compared to both established procyclic form cells and bloodstream monomorphic cells. However, the amount of protein synthesis increased with time during differentiation of the stumpy forms to procyclic forms upon addition of 6mM *cis*-aconitate. At 8 hours post-addition, the level of protein

synthesis (88%) approached that seen for established procyclic forms. This showed that there was an increase in translation throughout differentiation and, hence, release from the translational repression seen in stumpy forms. Also shown in Figure 6.2 are the derived values of the mean incorporation expressed as a percentage of the mean for established procyclic form cells (100%) for the other of these two experiments. Moreover, for one of the experiments the mean count at 24 hour post-addition of *cis*-aconitate was 117% of the established procyclic form mean, demonstrating that the level of protein synthesis in these cells becomes equivalent to established procyclic form cells (unfortunately, in one experiment contamination with bacteria prevented a 24 hour sample from being assayed). Notably, the largest increase in protein synthesis between the time-points in both experiments was between 4 and 6 hours after the initiation of differentiation (Figure 6.2).

It can be concluded from this experiment that stumpy form cells exhibit a low level of protein synthesis compared to the proliferative bloodstream and procyclic forms, matching earlier studies. However, the level of protein synthesis increases throughout differentiation to the procyclic form, particularly after 4 hours, revealing the release from translational repression. However, the upregulation of expression of individual proteins in either stumpy form cells or throughout differentiation was not sufficiently resolved using the methods employed.

6.3 Investigation of stumpy form polysome-associated transcripts using Illumina digital tag RNA sequencing

In order to resolve proteins that change in expression in stumpy forms or throughout differentiation, an investigation was made at the RNA level. Specifically, to identify which mRNA transcripts escape stumpy form global translational repression, and therefore, are actively translated, the polysome-associated mRNAs were compared with total mRNA from stumpy forms (AnTat1.1 strain) and with total mRNA and polysome-associated mRNA from bloodstream form monomorphic (427 strain) cells. The polysome-associated RNA should represent the actively translated mRNAs, as only mRNAs that are being translated at that time should be ribosomally engaged.

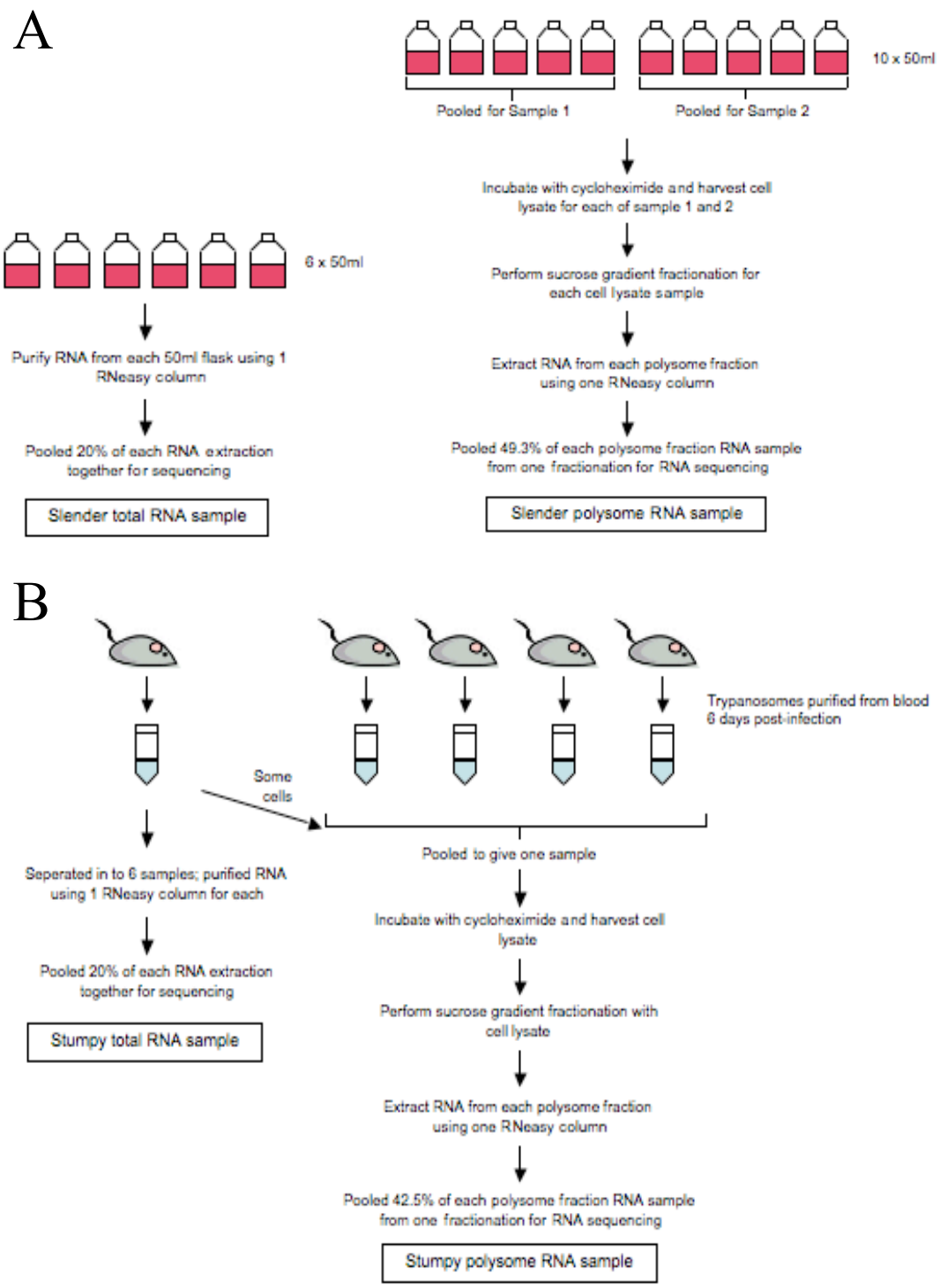


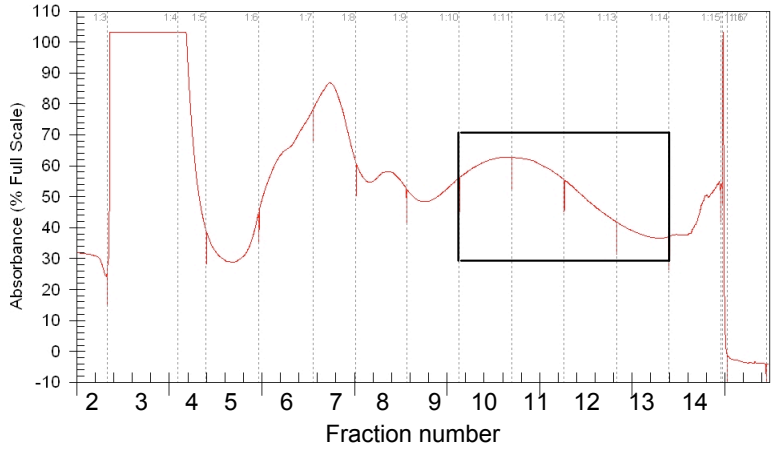
Figure 6.3 Schematic diagram of RNA sample generation for Illumina digital tag RNA sequencing. As outlined in Sections 2.9 and 6.2.1, monomorphic bloodstream form *T. b. brucei* Lister 427 were used for the slender form total RNA and polysome RNA samples (A) and pleomorphic AnTat1.1 were collected at 6 days post-infection in MF1 mice for stumpy form total RNA and polysome RNA samples (B).

6.3.1 Sample generation

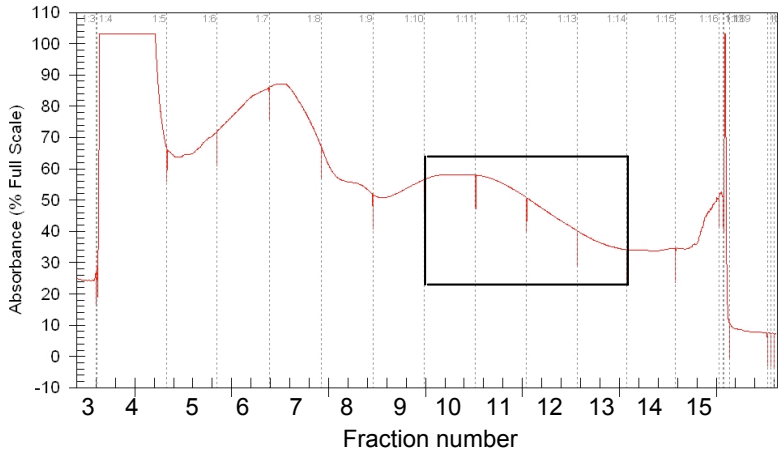
A schematic diagram depicting how each sample was generated for RNA analysis is shown in Figure 6.3. For the bloodstream form monomorphic cell total RNA (hereafter referred to as 'Slender total' RNA sample) six flasks of 50 ml of cells at about 2.5 x10^5 cells/ml were grown to about 2 x10^6 cells/ml. Each 50 ml was washed twice with 20 ml of filter-sterilised PSG to remove any traces of media and then cell lysis and RNA extraction was carried out, as described in Section 2.4.1. For each 50 ml of cells a separate RNeasy column (Qiagen) was used to allow maximum retrieval of RNA from the samples. Some (20%) of each of these six samples was then pooled together to generate the slender total RNA sample for Illumina digital tag sequencing. The remainder was stored to allow transcript validation by northern blotting or quantitative reverse-transcription PCR.

For the bloodstream form monomorphic cell polysome-associated RNA (hereafter referred to as 'Slender polysome' RNA sample) ten flasks of 50 ml of cells were grown to about 1-2 x10^6 cells/ml. Cell lysates were generated for polysome analysis as described in Section 2.9.1. Specifically, cells from five flasks were pooled after the first centrifugation step, so two lysate samples would be generated. Each of these lysate samples were then subjected to sucrose gradient fractionation as described in Section 2.9.3, with the resulting polysome profiles being shown in Figure 6.4. RNA was then extracted from fractions containing only the polysomes (fractions 10-13), with a separate RNeasy column (Qiagen) being used for each fraction to allow maximum retrieval of RNA from the samples. Half of the RNA extracted from each of the four polysome fractions from 'slender cell lysate sample 1' (since the fractionation appeared better for this lysate) was pooled together to generate the slender polysome RNA sample for Illumina digital tag sequencing.

A Slender-form cell lysate sample 1



Slender-form cell lysate sample 2



B Stumpy-form cell lysate

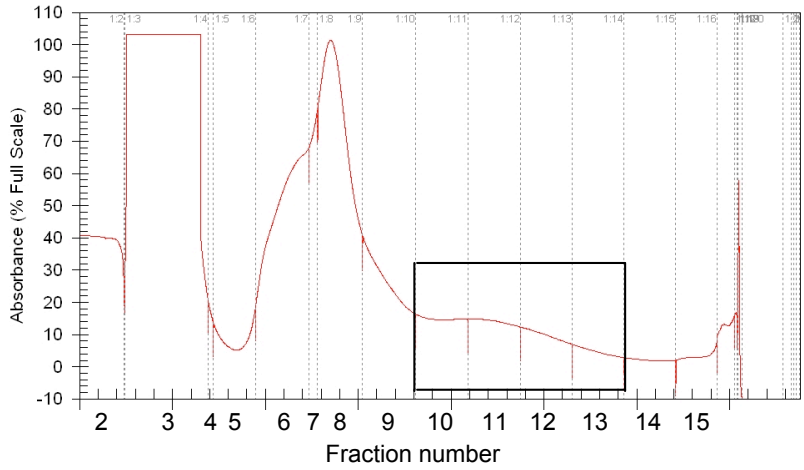


Figure 6.4 Polysome profiles for slender form (A) and stumpy form (B) cells. Lysates from cells treated with cycloheximide were layered on to a 15-50% sucrose gradient, centrifuged at 40 000 rpm for 2 hours, and fractions taken at a set number of drops (each fraction is marked with red vertical lines). Absorbance was measured at 254 nm throughout fractionation. RNA was extracted and purified from polysomal

fractions (fractions 10-13 for each, shown in black boxed regions on profiles) for Illumina digital tag RNA sequencing. See sections 2.9 and 6.2.1 for details.

For the stumpy form total RNA (hereafter referred to as the 'Stumpy total' RNA sample) and for the stumpy form cell polysome-associated RNA (hereafter referred to as the 'Stumpy polysome' RNA sample) five MF1 mice were infected with *T. brucei* AnTat1.1, then trypanosomes harvested 6 days post infection, when about 95% of the trypanosome population were stumpy form by morphology. The parasites were purified over DEAE columns as described in Section 2.2.2, and RNA extracted from the cells purified from one of the mice. For this RNA sample, the resuspended cell pellet was divided into six, and aliquots purified using a separate RNeasy column (Qiagen) to allow maximum retrieval of RNA. Some (20%) of each of these six samples was also pooled together to generate the 'stumpy total' RNA sample for Illumina digital tag sequencing.

Cells purified using two DEAE columns from the four remaining mice, plus some of the cells purified from the mouse used for the 'total RNA' sample, were then used to generate one cell lysate for polysome analysis. Cells from the two purifications were pooled for the wash step. Sucrose gradient fractionation was then performed as described in Section 2.9 with this lysate, the resulting polysome profile being shown in Figure 6.4. RNA was extracted from fractions containing only the polysomes (fractions 10-13) as described in Section 2.9.4. Again, for each fraction, a separate RNeasy column (Qiagen) was used to allow maximum retrieval of RNA from the samples. Some (42.5%) of the RNA extracted from each of the four polysome fractions was pooled together to generate the stumpy polysome RNA sample for Illumina digital tag sequencing. The remainder was stored for future validation experiments.

An analysis using Illumina digital tag sequencing of total mRNA populations from different life-cycle stages had been performed previously (Keith Matthews, unpublished data), so the bloodstream-form monomorphic and stumpy form total RNA samples here also served as replicates for this previous experiment. To further

assist this, total RNA was also obtained from slender form and intermediate form (population ~ 70% intermediate by morphology) pleomorphic cells (AnTat1.1 strain).

The concentration of the RNA samples was verified and controlled for integrity by using nanodrop analysis, and by inspection of each sample (10%) on RNA gels (Figure 6.5), as described in Section 2.4.2. The quality of each RNA sample appeared very good, so the pooled samples were sequenced by Eurofins MWG Operon using Illumina HiSeq 2000 sequencing (see Section 2.11).

The sequencing reads were returned as FASTA sequences, which were mapped to the *T. brucei* TREU 927/4 genome and statistically analysed as follows (with the assistance of Dr. Paul Capewell, University of Edinburgh). Initially, a transcript alignment database was created using all of the identified ORFs from TriTrypDB v3.3 release (http://tritrypdb.org/tritrypdb/), this including mapped UTR data from the Cross Laboratory (Siegel *et al*. 2010). Adapter sequences were trimmed from the reads using the CLC Genomics Workbench and aligned against the transcript database using Bowtie (Langmead *et al.*, 2009) with 1 mismatch (for the four bloodstream monomorphic/stumpy total/polysome RNA samples) or 2 mismatches (for the four total RNA samples used with the results of the previous Illumina digital tag RNA sequencing experiment mentioned above) allowed.

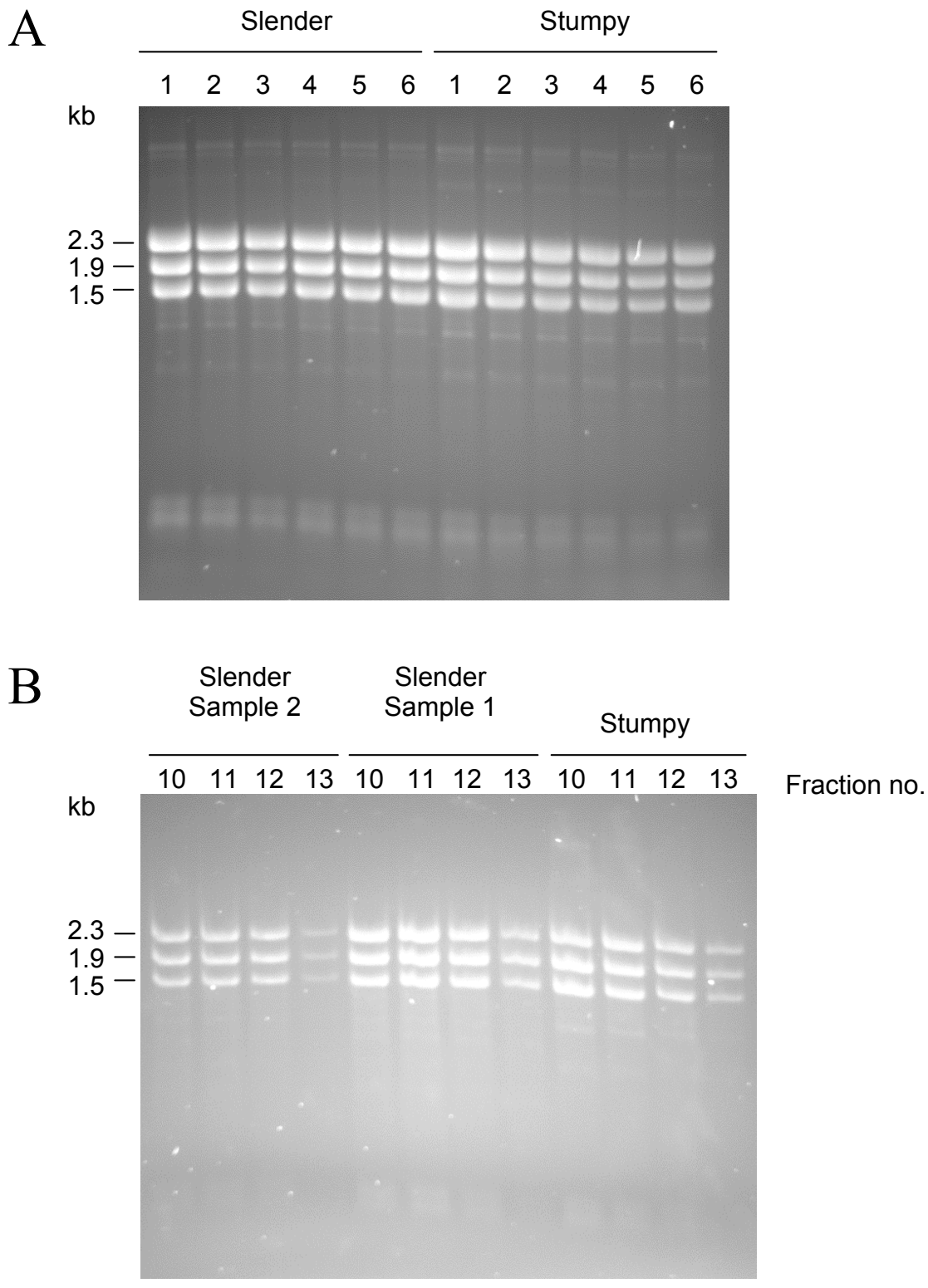


Figure 6.5 Ethidium bromide-stained RNA gels with RNA samples used for Illumina digital tag RNA sequencing. A: Total RNA samples for slender forms (6 samples) and stumpy forms (6 samples). B: Polysome-associated RNA samples for slender forms (samples 1 and 2) and stumpy forms, fraction numbers 10-13 for each. The three major rRNA bands are indicated by their size in kilobases (kb).

Two datasets were prepared for each sample, one that discarded tags that aligned to multiple sites (for identification of unique genes of interest) and another allowing for up to 10 matches per tag (for identification of gene arrays that may also be differentially regulated). The number of tags aligning to a transcript was normalized to the coverage of each transcript (to take into account differing numbers of *Nla*III sites in each gene) and reported as RPKM (reads per kilobase of exon model per million mapped reads). To identify transcripts with differential expression, the RPKM values for each transcript in each dataset were compared using the CLC Genomics Workbench with either a proportional statistical analysis for non-replicated samples (i.e. comparisons with the four bloodstream monomorphic/stumpy total/polysome RNA samples), or Gaussian statistical analysis for samples with replicates (i.e. comparisons with the four total RNA samples used with the results of the previous Illumina digital tag RNA sequencing experiment mentioned above). A stringent *p* value cut-off of 0.01 (corrected for false discovery rate) was used to determine if a transcript was differentially expressed between samples. A *p* value reported as '$p = 0$', means that this *p* value has a value of less than 4.9×10^{-324}.

Using the datasets that discarded tags that aligned to multiple sites, analysis was made of the fold change in slender or stumpy polysome RNA read counts compared with slender or stumpy total RNA read counts, to identify genes whose transcript was enriched by 2-fold or greater (all $p \geq 0.01$). Also returned with each gene identified in each of these comparisons was any associated RNAi phenotype from the study by Alsford *et al*. (2011). Further, a dataset was generated which compared the fold changes for all genes using the datasets allowing for up to 10 matches per tag. The work outlined above, from alignment of reads to the generation of the four datasets and multiple alignment dataset (inclusive), was performed by Dr Paul Capewell (University of Edinburgh).

The number of genes whose transcript was enriched by 2-fold or greater ($p \geq 0.01$) in each comparison and the overlap of these genes between datasets is shown in Figure 6.6. The rest of this chapter will concern analysis of the results of these datasets generated from the Illumina digital tag RNA sequencing experiment.

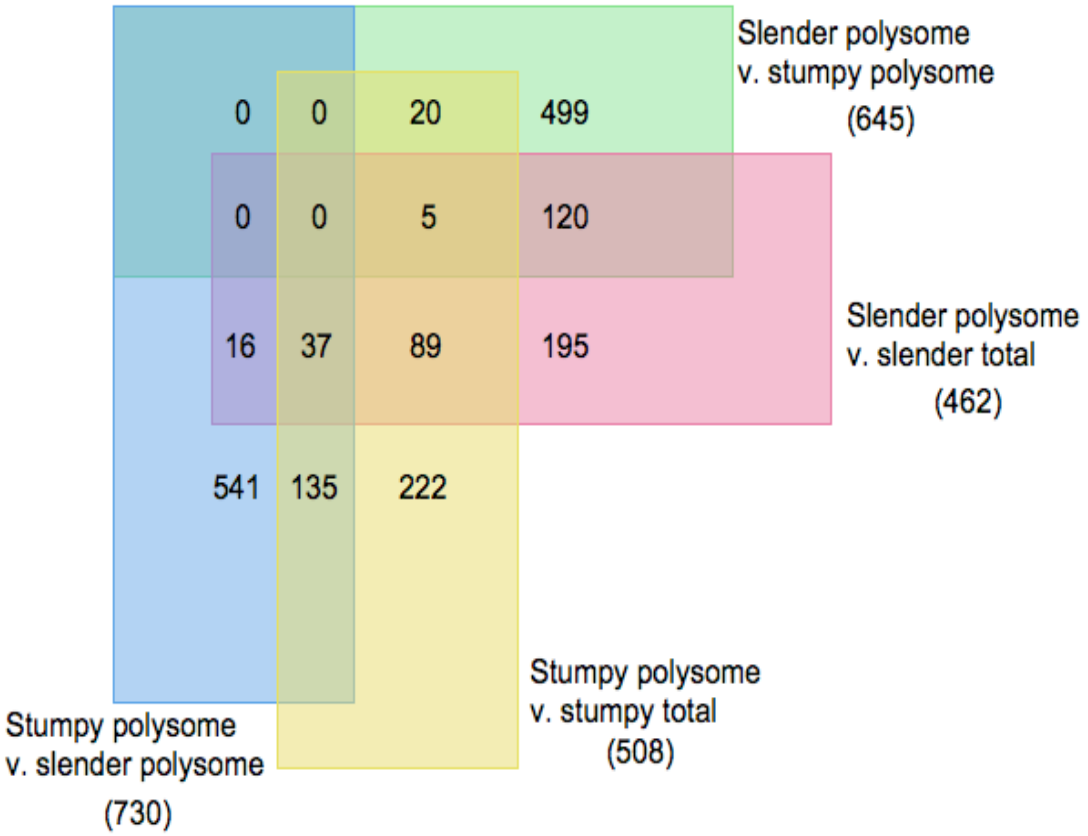


Figure 6.6 Venn diagram of enriched transcripts. Number of genes with transcripts enriched in either slender or stumpy form polysome-associated RNA samples compared to slender or stumpy form total RNA samples. Numbers in parentheses indicate the total number of genes with enriched transcripts in each dataset.

6.4 Which transcripts are actively translated in stumpy cells?

My hypothesis was that genes that are actively translated in stumpy cells would probably include genes important for transmission to the tsetse fly and differentiation competence. Therefore, to detect which transcripts were actively translated in stumpy cells the 508 genes identified as enriched in stumpy polysome RNA by 2-fold or greater (all $p \geq 0.01$) compared to stumpy total RNA were analysed further. The 30 genes showing the greatest fold change are shown in Table 6.1.

Table 6.1 Genes with the greatest fold change upregulation in stumpy form polysome-associated RNA compared to stumpy form total RNA

Gene ID	Fold change	p-value [a]	ST Total RPKM[b]	ST Polysome RPKM	Product Description
Tb11.02.3570	4.68	0	102.15	431.92	hypothetical protein, conserved
Tb927.8.1220	4.58	0	212.01	878.18	hypothetical protein, conserved (pseudogene)
Tb927.7.4060	4.51	0	158.64	647.13	calpain-like cysteine peptidase, putative,cysteine peptidase, Clan CA, family C2, putative
Tb09.160.0680	4.1	5.28E-07	17.58	65.12	sec1 family transport protein, putative (SLY1)
Tb927.10.11820	3.98	1.36E-08	23.13	83.27	hypothetical protein, conserved
Tb927.7.4070	3.97	0	214.96	770.42	calpain-like cysteine peptidase, putative,cysteine peptidase, Clan CA, family C2, putative
Tb927.5.2590	3.93	6.97E-07	18.66	66.26	hypothetical protein, conserved
Tb11.03.0250	3.77	0	60.15	205.18	cyclophilin a,cyclophilin type peptidyl-prolyl cis-trans isomerase (CYPA)
Tb927.6.4130	3.75	6.84E-15	45.71	155.05	hypothetical protein, conserved
Tb09.211.4513	3.73	0	112.04	377.22	kinetoplastid membrane protein KMP-11 (KMP-11)
Tb927.7.2070	3.73	0	262.78	885.25	heat shock protein DNAJ, putative
Tb927.10.10830	3.69	4.87E-03	8.3	27.66	HNRNPA (HNRNPA)
Tb09.211.0700	3.67	0	53.61	177.8	isopentenyl-diphosphate delta-isomerase, putative,isomerase, putative
Tb11.02.2860	3.64	3.87E-11	35.93	118.34	hypothetical protein, conserved
Tb927.8.1750	3.62	0	70.61	230.77	hypothetical protein, conserved
Tb11.02.0020	3.61	0	182.22	593.79	hypothetical protein, conserved
Tb927.10.13710	3.57	0	110.86	357.65	hypothetical protein, conserved
Tb11.02.3230	3.55	8.34E-10	33.14	106.32	hypothetical protein, conserved
Tb927.7.1470	3.54	0	127.78	408.98	ATPase subunit 9, putative
Tb927.7.4290	3.53	8.88E-05	15.34	48.9	hypothetical protein, conserved
Tb11.02.1290	3.5	0	98.77	312.65	hypothetical protein, conserved
Tb927.7.2780	3.5	0	1,139.12	3,605.16	hypothetical protein, conserved
Tb927.7.3590	3.49	0	458.09	1,445.48	hypothetical protein, conserved
Tb11.01.7740	3.48	0	202.05	634.54	hypothetical protein, conserved

Tb927.10.15810	3.47	0	159.88	501.73	hypothetical protein, conserved
Tb11.42.0004	3.45	0	93.36	291.42	hypothetical protein, conserved
Tb927.8.4380	3.45	0	156.26	487.8	hypothetical protein, conserved
Tb11.01.7800	3.44	0	901.04	2,800.43	nucleoside diphosphate kinase (NDPK)
Tb11.01.1180	3.42	0	440.48	1,361.17	hypothetical protein
Tb927.10.6080	3.41	0	154.42	476.18	proteasome beta 5 subunit, putative,proteasome beta 5 subunit (PRCE)

[a] p-value is corrected for false discovery rate (FDR)
[b] ST: stumpy form, RPKM: reads per kilobase of exon model per million mapped reads

6.4.1 Analysis of conserved hypothetical genes

A high proportion (almost two-thirds) of enriched transcripts had an associated gene product annotation of 'hypothetical protein, conserved', and therefore, may represent interesting genes with novel functions or be kinetoplastid-specific genes. To gain more information about possible functions for these genes, initially, the 10 genes with the greatest fold change difference in this dataset were investigated through the information associated with its gene entry page on the *T. brucei* genome database (http://tritrypdb.org/tritrypdb/) (Table 6.2). This yielded more information for the conserved hypothetical proteins, Tb11.02.3570, Tb927.8.1220, Tb927.10.11820, Tb927.5.2590, and Tb927.6.4130.

For Tb11.02.3570, the additional information was that it has an associated GO term indicating that it is a mitochondrial component, is predicted to possess 6 transmembrane domains and an uncharacterised protein family domain (3.70E-06), and was found by mass spectrometry in the *T. brucei* palmitolyated and procyclic form proteomes. Therefore, this information indicates it might be a mitochondrial membrane protein.

For Tb927.8.1220 the predicted protein sequence was found to contain several stop codons, and therefore, this is unlikely to be a functional ORF.

For Tb927.10.11820, this gene was found to be expressed in procyclic forms, has an associated GO term indicating that it is a mitochondrial component, and a tetratricopeptide repeat (TPR)-like protein domain (1.30E-02). Using the protein sequence for this gene as query in a protein BLAST (blastp; http://blast.ncbi.nlm.nih.gov/Blast.cgi) search yielded several genes annotated as pentatricopeptide repeat-containing proteins. In a previous study in *T. brucei*, pentatricopeptide repeat proteins were found to stimulate mRNA adenylation/uridylation to activate mitochondrial translation (Aphasizheva *et al.*, 2011), however, neither of the proteins investigated in this study were Tb927.10.11820. Therefore, it appears this protein is a component of mitochondria which could possibly be involved in mRNA adenylation/uridylation.

Table 6.2 Information from the TriTryp database (http://tritrypdb.org/tritrypdb/) regarding the ten most highly enriched transcripts in stumpy polysome-associated RNA compared to stumpy total RNA.

Gene ID	Notes	GO terms	Protein features	Phenotype	Mass spec.-based expression evidence	Similarities to PDB chains [a]	Additional information
Tb11.02.3570	conserved amongst eukaryotes and plants, detected in *T. brucei* bloodstream plasma membrane fraction	mitochondrial part	6 transmembrane domains, uncharacterised protein family domain	-	*T. brucei* palmitoylated proteome, *T. brucei* procyclic form	-	
Tb927.8.1220	-	-	-	-	-	-	the predicted protein sequence contains several stop codons
Tb927.7.4060	expressed in bloodstream form as detected by proteomic analysis, expressed in procyclic form as detected by proteomic analysis, detected in *T. brucei* bloodstream plasma membrane fraction, detected in *T. brucei* bloodstream cytoskeletal fraction, detected in *T. brucei* flagellum proteome	-	unnamed hypothetical protein, domain of unknown function	abnormal cell cycle during bloodstream stage, abnormal cell morphogenesis during bloodstream stage, decreased growth during bloodstream stage, lethal during bloodstream stage	*T. brucei* proteomics bloodstream form cytosolic phosphoproteome	small myristoylated protein 1 from *L. major* (48), hypothetical protein from *L. major* (35)	
Tb09.160.0680	detected in *T. brucei* bloodstream plasma membrane fraction	vesicle docking involved in exocytosis, protein secretion, vesicle-mediated transport, mitochondrion, protein transporter activity	Sec1/munc 18-like (SM) proteins, Sec1 Sec1 family	-	*T. brucei* palmitoylated proteome, *T. brucei* procyclic form	Sec1 family domain containing protein 1 from *Rattus norvegicus* (45)	
Tb927.10.11820	expressed in procyclic form as detected by proteomic analysis	mitochondrial part	TPR-like	-	*T. brucei* procyclic form	-	

Tb927.7.4070	expressed in procyclic form as detected by proteomic analysis	-	unnamed hypothetical protein, domain of unknown function	abnormal cell cycle during bloodstream stage, abnormal cell morphogenesis during bloodstream stage, decreased growth during bloodstream stage, lethal during bloodstream stage	*T. brucei* procyclic form, *T. brucei* proteomics bloodstream form cytosolic phosphoproteome	small myristoylated protein 1 from *L. major* (49) and hypothetical protein from *L. major* (36)
Tb927.5.2590	-	-	Macro domain profile, Macro domain-like, YBR022w_8, Macro Macro domain	-	-	Macro domain-containing protein 1 from *Homo sapiens* (47), putative polyprotein/phosphatase from *E. coli* (43), Hypothetical protein TTHA0132 from *Thermus thermophilus HB8* (38), Non-structural protein 3 from *Venezuelan equine encephalitis virus* (32)
Tb11.03.0250	expressed in procyclic form as detected by proteomic analysis, expressed in bloodstream form as detected by proteomic analysis	protein folding, cytosol, flagellum, peptidyl-prolyl cis-trans isomerase activity	Cyclophilin-like, Pro_isomerase Cyclophilin type peptidyl-prolyl cis-trans isomerase/CLD, CSA_PPIASE_2 Cylcophilin-type peptidyl-prolyl cis-trans isomerase domain profile, CSAPPISMRASE Cyclophilin peptidyl-prolyl cis-trans isomerase signature (x5)	-	*T. brucei* palmitoylated proteome, *T. brucei* procyclic form	Peptidyl-prolyl cis-trans isomerase A from *Homo sapiens* (73), Cyclophilin A from *Homo sapiens* (74)

Tb927.6.4130	-	mitochondrial part	-	-	*T. brucei* procyclic form	-	One of the orthologs within TriTrypDB is succinate dehydrogenase cytochrome B subunit (SDH3) from *T. cruzi* CL Brener Non-Esmeraldo-like
Tb09.211.4513	expressed in *T. brucei* bloodstream plasma membrane fraction, expressed in *T. brucei* bloodstream cytoskeletal fraction, expressed in bloodstream form as detected by proteomic analysis, expressed in procyclic form as detected by proteomic analysis	defense response, positive regulation of cell proliferation, mitochondrial part	Membrane kinetoplastid protein 11 cytoskeleton microtubule antigen multigene KMP-11 family, KMP11 Kinetoplastid membrane protein 11	-	-	-	

[a] PDB: protein database; numbers in parentheses refer to the percentage identity associated with the PDB chain similarity

Tb927.5.2590 is predicted to possess a Macro domain (1.50E-55), the Pfam annotation (PF01661; http://pfam.sanger.ac.uk) for which states that this is an ADP-ribose binding module found in a range of unrelated proteins, such as the macro-H2A histone protein, with other domains and also alone in a protein family found in bacteria, archaebacteria and eukaryotes. This protein also has some potential structural similarity to PDB chains for human Macro domain-containing protein 1 (p = 7.4 x 10^{-39}), a putative polyprotein/phosphatase from *Escherichia coli* (p = 1.6 x 10^{-27}), Hypothetical protein TTHA0132 from *Thermus thermophilus* HB8 (p = 2.1 x 10^{-18}), and Non-structural protein 3 from *Venezuelan equine encephalitis virus* (p = 9.2 x 10^{-8}; see Table 6.2). The protein with the greatest similarity, human Macro domain-containing protein 1 has been identified as an O-acetyl-ADP-ribose deacetylase (Chen *et al.*, 2011). With the Macro domain present in a wide range of unrelated proteins it is difficult to make any predictions about what the function of this protein may be, however, it is possible that it could function as an O-acetyl-ADP-ribose deacetylase like human Macro domain-containing protein 1.

Tb927.6.4130 also has an associated GO term indicating that it is a mitochondrial component and was found by mass spectrometry in the *T. brucei* procyclic form proteome. Additionally, one of the syntenic orthologs within the TriTryp database listed is annotated as succinate dehydrogenase cytochrome B subunit (SDH3; Tc00.10470553511523.14 from *T. cruzi* CL Brener Non-Esmeraldo-like). The TriTryp database entry page for this gene has a user comment annotating it as the cytochrome b large subunit, and so Tb927.6.4130 may also be a cytochrome b subunit.

6.4.2 Interesting annotated genes in the ten most highly enriched genes

Several genes with more specific annotations also appear interesting, such as two proteins possibly involved in protein secretion or membrane signalling; these are Tb09.160.0680 (putative sec1 family transport protein, SLY1) and Tb09.211.4513 (kinetoplastid membrane protein-11, KMP-11). From the information gained from the gene entry pages for these genes, the putative sec1 family transport protein was detected in the bloodstream plasma membrane, was found in the *T. brucei*

palmitolyated and procyclic form proteomes, and has associated GO terms implicating it in vesicle-mediated transport, exocytosis and protein transporter activity. The kinetoplastid membrane protein-11 was detected by proteomic analysis in the bloodstream plasma membrane and cytoskeletal fractions, and has associated GO terms implicating it in defence response and positive regulation of cell proliferation (Table 6.2). This protein has been determined to be present at much greater levels in procyclic forms than bloodstream forms and increases in abundance during bloodstream form to procyclic form differentiation (Stebeck *et al.*, 1995) and in *T. cruzi* regulation of KMP-11 was suggested to be at the translational level (Thomas *et al.*, 2000), both being consistent with the presence of this transcript in this dataset. Interestingly, KMP-11 from *Leishmania* and African trypanosomes was able to stimulate cellular immune responses, and so KMP-11 may be important for induction of host immune response (Tolson *et al.*, 1994). This additional information confirms these two genes as interesting, as they could be involved in protein secretion or sensing the environment at the plasma membrane.

Another potentially interesting gene is annotated as cyclophilin A (CypA; Tb11.03.0250), as it may be involved in host-parasite interactions. For *T. cruzi* it has been reported that a cyclophilin is secreted by non-dividing trypanomastigotes and is involved in mammalian host cell entry (Moro *et al.*, 1995). A study of the *T. brucei* CypA found evidence for its secretion or release into the *T. brucei* culture medium (Pelle *et al.*, 2002). This study also investigated the transcript and protein expression of CypA throughout the *T. brucei* lifecycle and found that while protein expression was constitutively expressed with little variation in the different developmental stages, the transcript expression was at least 2-fold higher in slender forms than the procyclic form, and low in intermediate and stumpy forms (Pelle *et al.*, 2002). The identification of this gene in this dataset perhaps indicates that the CypA protein is an important protein that is constitutively expressed and so has a high level of association with polysomes, even in stumpy forms where the level of the transcript is lower than other life-cycle stages. This would also suggest that the control of expression for this gene is at the level of RNA decay in slender and procyclic forms

where transcript levels are greater than other life-cycle stages but protein levels are constant.

Additionally, two genes annotated as encoding putative calpain-like cysteine peptidase proteins (Tb927.7.4060 and Tb927.7.4070) had transcripts found to be enriched in stumpy polysome-associated RNA compared to stumpy total RNA. These genes are potentially interesting because cysteine peptidases are involved in various parasite processes, such as virulence and pathogenesis (Caffrey *et al.*, 2011). Tb927.7.4060 and Tb927.7.4070 are members of group 3 calpain-like proteins, termed small kinetoplastid calpain-related proteins (SKCRPs), and are also called TbSKCRP7.1 and TbSKCRP7.2, respectively (Liu *et al.*, 2010). These two proteins have almost identical sequences but only TbSKCRP7.2 was found to be associated with the flagellar membrane (Liu *et al.*, 2010). Moreover, TbSKCRP7.2 was upregulated in procyclic forms 9.8-fold over bloodstream forms as determined by quantitative RT-PCR, whereas TbSKCRP7.1 was not significantly upregulated (Liu *et al.*, 2010). The identification of the TbSKCRP7.2 transcript as enriched in stumpy form polysome RNA is, therefore, consistent with a preparation of these forms for procyclic form differentiation. The detection of TbSKCRP7.1 could be due to high similarity between these gene sequences.

6.4.3 Mitochondrial proteins

A large number of mitochondrial proteins were found to be present in the 508 genes whose transcripts were enriched in stumpy polysome RNA compared to stumpy total RNA. Some putative mitochondrial genes are discussed above (see section 6.4.1), being identified via their mitochondrial GO term. However, also present are a number of genes with a known mitochondrial localisation or association. Due to the change in environment and a resulting switch between aerobic glycolysis and oxidative phosphorylation during the transition between bloodstream forms and procyclic forms, a large number of mitochondrial genes show developmentally regulated expression, being elevated in procyclic forms. A number of these genes are mitochondrially-encoded so will not be present in this analysis, due to alignment of reads with the sequenced TREU 927/4 genome, which does not include the

mitochondrial genome. However, some of these genes are nuclear-encoded and are present in this stumpy-polysome enriched dataset.

For example, the cytochrome oxidase (COX) complex is developmentally regulated in trypanosomes, as it is absent in bloodstream forms but present in procyclic forms (Priest and Hajduk, 1994). Of the components that make the COX complex, at least 10 are nuclear-encoded and for 7 of these the regulation of developmental expression is achieved through differential mRNA stability and translational control (Mayho *et al*. 2006). Two of these components were found in this study to be upregulated in stumpy form polysome RNA compared to stumpy form total RNA; these were cytochrome oxidase subunit VIII (Tb927.4.4620; 2.82 fold change, $p = 2.24 \times 10^{-6}$) and cytochrome oxidase subunit VII (Tb927.3.1410; 2.47 fold change, $p = 0$). The mRNAs for these components show bloodstream stage-specific instability, and 3'UTR analysis of other COX components showed regulation mainly at the translational level but also mRNA stability (Mayho *et al*. 2006). Therefore, from the results shown here it would appear that for COX VIII and COX VII, regulation of expression is also at the translational level. Also detected were cytochrome *c* (see section 6.3.3), a putative cytochrome b5 (Tb11.02.4485; 2.65 fold change, $p = 0$), and a putative cytochrome oxidase copper chaperone (Tb927.3.2650; 2.03 fold change, $p = 0$). It has previously been shown that the levels of cytochrome *b* and COX I and II transcripts, which are undetectable in slender forms, increase in stumpy forms, and mitochondrial rRNAs also increase at this stage (Michelotti and Hajduk, 1987). These authors suggest that stumpy forms accumulate mitochondrial transcripts prior to developing a functional mitochondrion (Michelotti and Hajduk, 1987). Although the transcripts investigated are encoded in mitochondria it would appear from the results presented here that the same is true for, at least some, nuclear-encoded transcripts. This, therefore, provides support for the assumption that stumpy forms are preadapted for transmission to the insect vector and differentiation to procyclic forms.

6.4.4 Other interesting annotated genes

Interestingly, a putative mitogen-activated protein (MAP) kinase was also found in this dataset (2.28 fold change, $p = 1.20 \times 10^{-11}$). This was Tb927.10.16030, which is annotated as putative mitogen-activated protein kinase 4 but has also been named TbMAPK2 (Muller *et al.*, 2002). This protein was shown to be involved in differentiation to the procyclic form, as a null mutant (Δ*mapk2*/Δ*mapk2*) differentiated with delayed kinetics (Muller *et al.*, 2002). The same study also investigated the steady-state levels of the MAPK2 transcript and found that it was 3-fold higher in the bloodstream form than procyclic form. Further, they also report that in the pleomorphic AnTat1.1 strain the transcript expression levels were similar in the slender bloodstream form and the stumpy bloodstream form (Muller *et al.*, 2002). Therefore, the presence of this gene in this dataset is consistent with the previous report, i.e. the transcript is still expressed at high levels in the stumpy form and is also enriched in stumpy form polysomal fractions, due to the requirement for this protein during differentiation to the procyclic-form. This again provides support for the concept that stumpy form cells are preadapted for differentiation to the procyclic form and transmission to the insect vector, as the cells are already starting to express a protein required for differentiation.

It was also noted that there appeared to be several proteins with annotations involved in gene expression, such as in RNA binding, RNA processing, and translation. For example, those involved in RNA binding included putative Y14 (Tb927.7.1170; 2.18 fold change, $p = 5.63 \times 10^{-8}$), putative La (Tb927.10.2370; 3.03 fold change, $p = 3.87 \times 10^{-3}$), as well as RPB25, RBP39, and RBP22 transcripts (Tb927.7.880, Tb11.01.7000, and Tb927.5.1080, respectively). Also, a number of mitochondrial RNA binding proteins were present, these included GBP21, GBP25, and RBP16 (Tb11.55.0009, Tb11.01.4860, and Tb11.02.5770, respectively). Those involved in RNA processing included a putative ribonuclease H (Tb927.10.5070; 3.07 fold change, $p = 0$), ALBA2 (Tb11.02.2030; 2.56 fold change, $p = 0$), ALBA3 (Tb927.4.2040; 2.6 fold change, $p = 2.69 \times 10^{-4}$), and fibrillarin (NOP1, Tb927.10.7500; 2 fold change, $p = 0$). Of those involved in translation these included putative eukaryotic translation initiation factor (eIF) 1A (Tb927.8.5880; 2.9 fold

change, $p = 0$), putative eIF4G (Tb927.8.4820; 2.89 fold change, $p = 1.14 \times 10^{-6}$), putative eIF (Tb11.01.3420; 2.37 fold change, $p = 6.10 \times 10^{-10}$), putative eIF3B (Tb11.01.1370; 2.16 fold change, $p = 3.64 \times 10^{-6}$), putative eIF3D (Tb927.4.1930; 2.29 fold change, $p = 4.36 \times 10^{-3}$), and putative eukaryotic peptide chain release factor subunit 1 (Tb927.7.1710; 2.32 fold change, $p = 1.08 \times 10^{-7}$). The presence of a large number of these types of proteins is expected because the control of gene expression is believed to be through post-transcriptional mechanisms and these protein types would be likely to have an important role in developmental events.

6.5 Which transcripts are more actively translated in stumpy polysomes compared to slender polysomes?

To identify transcripts showing stumpy-enriched translation, the datasets were also analysed for transcripts that were at higher abundance in the stumpy form polysome RNA sample compared to the slender form polysome RNA sample. This identified 732 genes enriched by 2-fold or greater (all $p \geq 0.01$). The top 30 genes showing the greatest fold change of enrichment are shown in Table 6.3.

Table 6.3 Genes with the greatest fold change upregulation in stumpy form polysome-associated RNA compared to slender form polysome-associated RNA

Gene ID	Fold change	*p*-value [a]	SL Polysome RPKM [b]	ST Polysome RPKM [c]	Product Description
Tb927.4.4990	977.52	0	0.51	496.84	ubiquinol-cytochrome c reductase, putative
Tb11.01.3860	943.09	0	2.63	2,494.26	hypothetical protein, conserved
Tb11.01.3840	138.38	0	3.12	433.13	hypothetical protein, conserved
Tb927.7.5400	133.72	1.59E-05	0.17	23.2	hypothetical protein, conserved
Tb927.8.690	120.61	2.07E-07	0.27	32.35	peptidyl-prolyl cis-trans isomerase/rotamase, putative,PPIase, putative (PIN1)
Tb927.3.5760	72.98	2.69E-04	0.24	17.66	hypothetical protein, conserved
Tb09.160.2320	52.44	0	1.76	92.46	hypothetical protein, unlikely (1L12.35)
Tb927.4.100	51.42	2.14E-03	0.26	13.46	retrotransposon hot spot protein (RHS, pseudogene), putative,retrotransposon hot spot protein 1 (RHS1), interrupted
Tb11.03.0920	29.5	6.36E-06	0.92	27.22	hypothetical protein, conserved
Tb11.01.4560	26.71	4.56E-15	2.87	76.96	variant surface glycoprotein (VSG)-related, putative
Tb927.10.15370	25.44	2.09E-10	2.01	51.36	DNA-directed RNA polymerase, alpha subunit, putative,RNA polymerase subunit, putative (RPC40)
Tb927.2.2400	25.13	1.58E-04	0.81	20.32	hypothetical protein, conserved (25N14.65)
Tb927.6.1430	20.68	0	10.06	208.87	hypothetical protein, conserved
Tb927.4.4430	17.3	2.81E-03	0.83	14.45	adenylyl cyclase (GRESAG 4.4)
Tb927.6.5020	14.41	4.56E-15	5.79	83.72	cyclin 7, putative,CYC2-like cyclin, putative (CYC7)
Tb927.6.3070	13.4	0	10.1	135.84	hypothetical protein, conserved
Tb11.01.0720	13.17	6.43E-10	4.13	54.62	cation transporter, putative
Tb927.4.4410	13.01	3.13E-05	2.06	26.93	adenylyl cyclase (GRESAG 4.4)
Tb927.4.3500	12.67	0	9.99	127.06	hypothetical protein, conserved
Tb927.10.9160	12.56	0	15.34	193.4	hypothetical protein, conserved
Tb11.02.5180	11.64	0	50.01	584.5	DNA-directed RNA polymerase II, subunit 9, putative (RPB9)
Tb11.47.0008	10.97	0	26.59	293.01	class I transcription factor A, subunit 3 (CITFA-3)
Tb09.142.0030	10.72	4.40E-11	6.03	64.85	expression site-associated gene (ESAG, pseudogene), putative,expression site-associated gene 3 (ESAG3), pseudogene
Tb927.10.3540	10.66	0	28.61	306.2	hypothetical protein, conserved

Tb927.7.6870	10.56	1.54E-09	5.24	55.54	hypothetical protein, conserved
Tb09.211.1620	10.48	0	18.42	193.83	hypothetical protein, conserved
Tb927.3.2750	10.35	0	28.09	291.98	hypothetical protein, conserved
Tb927.2.200	9.83	0	54.98	542.65	expression site-associated gene (ESAG, pseudogene), putative,expression site-associated gene 3 (ESAG3), degenerate
Tb927.6.3880	9.64	5.50E-04	2.2	21.26	hypothetical protein, conserved
Tb927.6.3260	9.54	0	18.48	176.91	hypothetical protein, conserved

[a] *p*-value is corrected for false discovery rate (FDR)
[b] SL: slender form, RPKM: reads per kilobase of exon model per million mapped reads
[c] ST: stumpy form

6.5.1 Analysis of conserved hypothetical genes

Again, there were many genes encoding proteins annotated as 'hypothetical, conserved' so initially the top 10 of these genes were investigated through the information associated with each gene entry page on the *T. brucei* genome database (http://tritrypdb.org/tritrypdb/) (Table 6.4). This yielded more information for Tb11.01.3860, Tb11.01.3840, Tb927.7.5400, Tb927.3.5760, Tb09.160.2320, Tb11.03.0920, and Tb11.01.4560.

For Tb11.01.3860, the additional information was that it has an associated GO term indicating that it is a mitochondrial component, is predicted to have a signal peptide and was found to be expressed in the procyclic form.

Tb11.01.3840 has a predicted S15/NS1 RNA-binding domain (8.40E-03) implicating it in RNA-binding. Tb927.7.5400 is predicted to possess a signal peptide and 2 transmembrane helices. Using the protein sequence for this gene as query in a protein BLAST (blastp; http://blast.ncbi.nlm.nih.gov/Blast.cgi) search using the 'Reference proteins (refseq_protein)' database yielded no further information. The three significant hits were hypothetical proteins from *T. brucei* (Tb927.7.5420; $p = 0.0$), and *T. cruzi* (Tc00.1047053510305.70 and Tc00.1047053509153.20; $p = 1 \times 10^{-27}$ and 2×10^{-9}, respectively). This, therefore, may be a trypanosomatid-specific gene.

Table 6.4 Information from the TriTryp database (http://tritrypdb.org/tritrypdb/) regarding the ten most highly enriched transcripts in stumpy polysome-associated RNA compared to slender polysome-associated RNA.

Gene ID	Notes	GO terms	Protein features	Phenotype	Mass spec.-based expression evidence	Similarities to PDB chains [a]	Additional information
Tb927.4.4990	expressed in procyclic form as detected by proteomic analysis	-	Non-heme 11 kDa protein of cytochrome bc1 complex (Ubiquinol-cytochrome c reductase) domain	-	-	-	
Tb11.01.3860	-	mitochondrial part	Signal peptide predicted	-	*T. brucei* procyclic form	-	
Tb11.01.3840	-	-	S15/NS1 RNA-binding domain	-	-	-	
Tb927.7.5400	-	-	Signal peptide and 2 transmembrane helices predicted	-	-	-	
Tb927.8.690	expressed in procyclic form as detected by proteomic analysis, expressed in bloodstream form as detected by proteomic analysis	protein folding, peptidyl-prolyl cis-trans isomerase activity	FKBP-like, PPCI_PPIASE_2 PpiC-type peptidyl-prolyl cis-trans isomerase family profile, Rotamase PPIC-type PPIASE domain	-	-	Peptidyl-prolyl cis-trans isomerase from *Arabidopsis thaliana* (49), peptidyl-prolyl cis-trans isomerase from *Candida albicans* (45), peptidyl-prolyl cis-trans isomerase NIMA-inte from *Homo sapiens* (52)	
Tb927.3.5760	-	-	Signal peptide predicted	-	-	-	Only syntenic gene listed (*T. b. gambiense*, Tbg972.3.6390) is annotated as 'hypothetical protein, *T. brucei* spp.-specific protein'; two 927 homologues annotated as putative ESAG9 proteins

Tb09.160.2320	-	-	-	-	-	-	Annotated as 'hypothetical protein, unlikely'; predicted protein sequence of 42 amino acids
Tb927.4.100	-	-	RHSP Retrotransposon hot spot protein, Trypano_RHS Trypano_RHS: trypanosome RHS family	-	-	-	
Tb11.03.0920	-	-	Prefoldin	-	-	-	
Tb11.01.4560	-	-	Signal peptide predicted	-	-	-	

[a] PDB: protein database; numbers in parentheses refer to the percentage identity associated with the PDB chain similarity

Tb927.3.5760 is also predicted to possess a signal peptide. Interestingly, the only syntenic gene listed for this gene is from *T. b. gambiense* and is annotated as a *T. brucei* species-specific protein. Additionally, two of the *T. brucei* 927 homologues are annotated as putative ESAG9 proteins. Therefore, the protein sequence for this gene was used as query in a protein BLAST (blastp; http://blast.ncbi.nlm.nih.gov/Blast.cgi) search using the 'Reference proteins (refseq_protein)' database to see if this could be an ESAG9 gene. This search returned 4 results; the top hit was its own protein sequence ($p = 0.0$), the others were Tb09.160.5400 (a putative expression site-associated gene 9 protein, $p = 1 \times 10^{-13}$), Tb09.142.0380 (a hypothetical protein, $p = 4 \times 10^{-11}$; the entry for this gene on TriTrypDB notes that it has low similarity to ESAG9), and Tb927.5.150 (a hypothetical protein, $p = 6 \times 10^{-11}$; of the orthologs on TriTrypDB two are putative ESAG9 proteins). Therefore, it appears that Tb927.3.5760 shows similarity to ESAG9 proteins and may be an ESAG9 or ESAG9-related gene.

Tb09.160.2320 is annotated as 'hypothetical protein, unlikely' and analysis of the protein sequence on the TriTryp database entry for this gene revealed only a 42 amino acid sequence. Therefore, it is unlikely that this gene is a functional ORF.

Tb11.03.0920 is predicted to have a prefoldin domain (8.10E-04); the Interpro annotation for which states that these proteins are part of a molecular chaperone system which stabilise nascent polypeptide chains and which therefore, promotes their correct folding. Consequently, it is possible that this protein has a similar function.

For Tb11.01.4560, this is predicted to possess a signal peptide, however, this gene is already annotated as a putative VSG-related protein and from the orthologs and paralogs within TriTrypDB listed, it appears that this might be the case, so no further information was gained.

6.5.2 Other interesting proteins

Many mitochondrial proteins were also found to be present in the 732 genes enriched in the stumpy form polysome RNA compared to the slender form polysome RNA. Again, members of the COX complex were present. COX subunits VIII (2.14 fold change, $p = 2.76 \times 10^{-4}$) and VII (3.77 fold change, $p = 0$) were present here, as well as being enriched in the stumpy polysome RNA compared to the stumpy total RNA (see section 6.2.1). Additionally, COX V was enriched 4.03-fold ($p = 0$) as was a putative cytochrome oxidase assembly protein enriched 2.58-fold ($p = 7.25 \times 10^{-8}$). It was previously found from investigations of the COX V 3'UTR that this provided regulation of protein expression almost exclusively at the translational or post-translational level (Mayho *et al*. 2006). This is supported by the enrichment seen here, as it demonstrates that the COX V transcript is more actively translated in stumpy forms than slender forms and so is consistent with regulation being at the level of translation.

As with comparing stumpy polysomal with total RNA, there were several transcripts implicated in translation found enriched in the stumpy polysome material compared to slender polysome sample. Interestingly, one of these was a putative eIF6 (Tb927.10.5300; 3.13 fold change, $p = 0$), which is supportive of the upregulation of this transcript (see Section 5.1.2) and protein in stumpy forms. Also, one of the eIF4E homologues eIF4E3 (Tb11.03.3630; 3.09 fold change, $p = 3.66 \times 10^{-8}$) was found enriched here too. However, this was not the homologue whose transcript was found enriched in stumpy forms by directed northern analysis (eIF4E4; see Section 5.1.1). This may indicate that, although the level of eIF4E3 transcript expression was observed to be consistent in different life-cycle stages (see Section 5.1.1), there may be upregulation of the protein in stumpy forms through an increase in translation. Additionally, a putative deoxyhypusine synthase (Tb927.1.870) was also found enriched in this dataset (2.03 fold change, $p = 3.18 \times 10^{-4}$). However, again, this was not the putative deoxyhypusine synthase found previously found to be upregulated at the mRNA level in stumpy forms by microarray analysis (Kabani *et al*. 2009, see Section 5.1.2) and northern blot analysis (Figure 5.1 B).

It was also noted that many genes annotated as adenylyl cyclase (GRESAG, Genes Related to ESAG, 4/4.4) genes were found in this dataset of transcripts enriched in the stumpy polysome sample compared to the slender polysome sample (Tb927.4.4430, Tb927.4.3750, Tb927.4.4410, Tb927.4.4460; 17.3, 3.09, 13.01, 7.61 fold changes respectively, $p = 3.41 \times 10^{-3}$ - 3.13×10^{-5}). Another study reports that GRESAG 4.1 protein is expressed in both bloodstream and procyclic form cells and at a constant level throughout differentiation, although the absence of a loading control makes assessment of protein levels difficult (Rolin *et al.*, 1993). However, this enrichment of GRESAG genes found here is interesting because transient stimulations in adenylate cyclase activity have been observed during differentiation to procyclic form (Rolin *et al.*, 1993).

Other interesting genes in this dataset include two putative cyclins, 'cyclin 7, putative, CYC2-like cyclin, putative (CYC7, Tb927.6.5020; 14.41 fold charge, $p = 4.56 \times 10^{-15}$) and 'cyclin 3, mitotic cyclin, putative' (CYC3, Tb927.6.1460; 3.52 fold charge, $p = 1.43 \times 10^{-8}$). It could be that these proteins play a role in the cell cycle arrest of stumpy form cells or re-entry into the cell cycle during differentiation, although a previous study reported that neither of these cyclins is likely to be involved in cell cycle control (Li and Wang, 2003). Therefore, it is unlikely that these proteins function in control of the cell cycle in stumpy forms or during differentiation and unclear what the significance of their presence in this dataset is.

Again, MAPK4, also known as TbMAPK2, was found in this dataset as well as being enriched in the stumpy polysome compared to stumpy total RNA (see Section 6.4.4). Another MAP kinase is present in this dataset as well, MAPK5 (Tb927.6.4220; 4.95 fold change, $p = 0$). Like TbMAPK2, this MAP kinase has also been shown to have a role in differentiation (Domenicali Pfister *et al.*, 2006). However, this MAP kinase was shown to be important for differentiation to stumpy forms, where it appeared to be a negative regulator of differentiation, possibly by lowering sensitivity to SIF (Domenicali Pfister *et al.*, 2006). The identification of translation of this transcript in stumpy forms could be consistent with a role of MAPK5 inhibiting differentiation. If MAPK5 lowers responsiveness to SIF, cells

that have already differentiated to the stumpy form would be expected to no longer require responsiveness to SIF. Additionally, as this gene was found enriched in stumpy form polysome transcripts compared to monomorphic slender forms polysome transcripts, this could be because this protein is expressed at low levels in monomorphic slender forms. Again, this is consistent with previous findings that Tb*MAPK5* null mutants in a monomorphic (and so refractory to SIF) bloodstream form strain showed no phenotype in contrast to pleomorphic bloodstream form Tb*MAPK5* null mutants (Domenicali Pfister *et al.*, 2006). However, the authors did not compare endogenous TbMAPK5 protein or mRNA levels in monomorphic cells with pleomorphic cells.

6.5.3 Are there genes actively translated in stumpy forms and not slender forms?

In addition to identifying genes which are upregulated in stumpy form polysomal RNA compared to slender form polysomal RNA, interesting genes might be discovered in a comparison of the genes present in the stumpy polysome RNA reads and absent in the slender polysome RNA gene reads. This is because genes that are highly important for differentiation are not likely to be expressed in the slender form stage. Therefore, the dataset containing all the read counts for each gene was analysed to exclude genes with a value of <1 RPKM in the stumpy polysome sample and then exclude genes with a value of ≥1 RPKM in the slender polysome sample. This yielded a new dataset of 210 genes. Most of these genes have low RPKM values across all sequenced samples and high associated *p* values. However, 6 genes have >20 RPKM value for the stumpy polysome sample; these were Tb927.4.4990 (977.52 fold change, $p = 0$), Tb927.8.690 (120.61 fold change, $p = 2.07 \times 10^{-7}$), Tb11.03.0920 (29.5 fold change, $p = 6.36 \times 10^{-6}$), Tb09.160.1900 (∞ fold change, $p = 6.10 \times 10^{-6}$), Tb927.7.5400 (133.72 fold change, $p = 1.59 \times 10^{-5}$), and Tb927.2.2400 (25.13 fold change, $p = 1.58 \times 10^{-4}$). Each of these genes, however, had already been identified in the stumpy polysome compared to slender polysome analysis and all were present in that dataset within the 15 genes with the greatest enrichment. Therefore, no additional interesting genes were discovered with this method, although this does provide validation of the previous analysis of genes enriched in the stumpy polysome compared to slender polysome samples.

6.6 Is the data for genes with known expression profiles as expected?

To investigate if the data obtained from this global RNA analysis experiment was in agreement with previously published data, the dataset was examined for the expression profile of several genes with known life-cycle expression profiles in *T. brucei*.

6.6.1 Housekeeping gene

Initially, the datasets were interrogated for a housekeeping (i.e. constitutively expressed) gene, β-tubulin. In *T. brucei* the β-tubulin genes are part of an array, being composed of α-tubulin and β-tubulin genes. When the dataset from alignment of reads which discarded tags that aligned to multiple sites was analysed for the read counts of the annotated β-tubulin genes (Tb927.1.2330, Tb927.1.2350, Tb927.1.2370, Tb927.1.2390), the read count for all RNA samples for each of the 4 genes was zero. This result was due to the high degree of similarity between the β-tubulin gene sequences, because when the data from the alignment of reads which allowed for up to 10 matches per tag was analysed, the read counts were very high. The four annotated β-tubulin genes showed very similar read count results to each other within each sample. For slender polysome RNA compared to slender total RNA there was about a 1.54 (range: 1.52 to 1.56, $p = 0$ for each) fold change for the four β-tubulin genes. For stumpy polysome RNA compared to stumpy total RNA there was about a -1.27 (range: -1.24 to -1.32, $p = 0$ for each) fold change. These data reflect that the β-tubulin transcripts are constitutively expressed and translated, with a greater level of translation in slender compared to stumpy forms. For slender polysome RNA compared to stumpy polysome RNA there was about an 8.82 (range: 8.69 to 8.91, $p = 0$ for each) fold change. This appears to reflect the much greater abundance of β-tubulin transcripts present in the slender form compared to the stumpy form. This is as expected because an approximately 5-fold decrease in tubulin transcript levels has been previously reported for stumpy forms relative to slender forms and procyclic forms (Feagin *et al.*, 1986).

6.6.2 Stumpy form upregulated genes

Next, the datasets were interrogated for genes known to be upregulated in stumpy forms. The genes chosen were ESAG9 and the PAD array. None of the ten annotated ESAG9 genes in the genome database (Tb09.160.5400, Tb09.160.5430, Tb09.v1.0330, Tb927.1.5080, Tb927.1.5220, Tb927.3.5790, Tb927.5.100, Tb927.5.120, Tb927.5.4620, and Tb927.7.170) were present in the upregulated in stumpy polysome RNA compared to stumpy total RNA or the upregulated in stumpy polysome RNA compared to slender polysome RNA datasets.

When the data from alignment of reads that allowed for alignment to multiple sequences was analysed for the ten ESAG9 genes, this showed that the read counts for all the ESAG9 genes were low in all four RNA samples. The reason for this is unclear, however, one contributing factor may be that many of the ESAG9 genes are present in telomeric regions. These regions are not sequenced and so are not included in the *T. brucei* TREU 927/4 genome that was used for read alignment, and so this would cause these sequences to be discarded from the analysis. This could explain the low read counts for ESAG9 sequences.

Another possibility is that the ESAG9-like gene found upregulated in the stumpy polysome compared to slender polysome dataset (see Section 6.5.1) is a true ESAG9 gene that was highly expressed at the time of sample generation. Alternatively, it could be a full ESAG9 gene in the AnTat1.1 strain but quite diverged in sequence in the TREU 927 strain.

The PAD gene array was also analysed and the genes were also found to not be present in the upregulated in stumpy polysome RNA compared to stumpy total RNA or the upregulated in stumpy polysome RNA compared to slender polysome RNA datasets. The exception was PAD2, which was found to be upregulated in the stumpy polysome RNA compared to slender polysome RNA dataset (5.7 fold change, p = 5.27 x10^{-10}). PAD2 might have been identified because it has the most distinct 3'UTR of the PAD genes and so reads are more likely to be uniquely assigned to this gene. As with β-tubulin, the absence of most PAD genes in the dataset is probably

due to these genes being part of a family with great similarity between the gene members. Analysis of the alignment of reads that allowed for up to 10 matches per tag for the eight PAD genes revealed few reads in all four RNA samples, although the read counts for PAD1, 2 and 3 were slightly higher.

6.6.3 Other genes

Cytochrome *c* is a nuclear-encoded mitochondrial protein that shows developmental regulation. The protein is not detected in bloodstream forms, which have an inactive mitochondrion and lack cytochrome-mediated respiration, but during differentiation to procyclic forms the protein accumulates, as cytochrome-mediated respiration occurs in the procyclic mitochondrion (reviewed in Priest and Hajduk, 1994). Despite the fact that the cytochrome *c* protein was undetectable in the bloodstream form, the transcripts are present in all life-cycle stages, and therefore, the authors concluded that regulation of cytochrome *c* protein expression occurs at the translational or post-translational level (Torri and Hajduk, 1988). From the data obtained here, cytochrome *c* (Tb927.8.5120) was one of the genes whose transcript was found to be upregulated in stumpy form polysome RNA compared to stumpy form total RNA (2.6 fold change, $p = 0$). This is consistent with the above conclusion that cytochrome *c* protein expression is regulated at the translational or post-translational level. As the authors also found in the above study that the protein was not detectable in stumpy forms, protein turnover might also be occurring.

During differentiation to the procyclic-form a histone H2B transcript was found to be upregulated at 8-12hr after initiation. Two studies have investigated the level of histone H2B transcript in slender and stumpy bloodforms, procyclic forms, and during differentiation by northern blot analysis. Both found that the transcript was enriched (by about 4-fold) in slender forms and procyclic forms compared with stumpy forms, and increased in abundance during differentiation (Matthews and Gull, 1998; Garcia-Salcedo *et al.*, 1999). Interestingly, a putative histone 2B variant (Tb11.02.5250) was found enriched in stumpy polysome RNA compared to stumpy total RNA (3.16 fold change; $p = 3.79 \times 10^{-6}$). This could reflect that the level of the

transcript is low in the stumpy total RNA sample but the protein is still actively translated in stumpy forms.

6.6.4 Analysis of microarray-identified stumpy form upregulated genes

The data from this study was also compared with an analysis of genes found to be upregulated in stumpy forms by microarray analysis. One microarray analysis that identified transcripts upregulated in stumpy forms was the study by Kabani *et al*. (2009), which identified 65 upregulated genes (listed in Kabani *et al*. 2009 Additional file 5). Comparison of these 65 genes in the RNA-sequencing analysis, revealed that a large proportion of the genes had no reads at all in the RNA samples, irrespective of whether multiple tag matches to different genes were considered or not. Most of the genes were mostly annotated as 'hypothetical, unlikely', and therefore, it is likely these are not functional ORFs and not expressed. Moreover, many of the genes had high *p* values associated with the calculated fold changes when comparing RNA samples due to their low read counts. Therefore, a comparison to the RNA-sequencing data was not possible for most genes.

However, for a few genes there was an agreement of transcript upregulation in stumpy forms. These were Tb927.8.1270 (4.4 fold change, $p = 2.29 \times 10^{-8}$) and Tb09.211.2300 (2.79 fold change, $p = 9.40 \times 10^{-6}$) when comparing stumpy polysome to slender polysome RNA. Interestingly, Tb927.8.5670 (2.89 fold change, $p = 2.65 \times 10^{-5}$), Tb927.6.4180 (2.17 fold change, $p = 1.35 \times 10^{-14}$) and Tb927.8.1270 (2.47 fold change, $p = 1.28 \times 10^{-4}$) where found upregulated when comparing stumpy polysome to stumpy total RNA. Therefore, these genes appeared interesting for further investigation.

Table 6.5 Information from theTriTryp database (http://tritrypdb.org/tritrypdb/) for genes found to be upregulated in stumpy form both by microarray analysis (Kabani *et al*. 2009) and by Illumina digital tag RNA sequencing analysis

Gene ID	Notes	GO terms	Protein features	Phenotype	Mass spec.-based expression evidence	Similarities to PDB chains[a]	Additional information
Tb927.8.5670	-	-	-	-	-	-	
Tb927.6.4180	detected in T. brucei bloodstream plasma membrane fraction	-	FUN14 FUN14 family; 2 transmembrane domains	-	-	-	
Tb927.8.1270	expressed in procyclic form as detected by proteomic analysis	-	HSP20-like chaperones	-	*T. brucei* palmitoylated proteome, *T. brucei* procyclic form	-	

[a] PDB: protein database

Tb927.8.5670, Tb927.6.4180 and Tb927.8.1270 are all annotated as 'hypothetical protein, conserved'. To obtain more information that might reveal a possible function for these genes, the TriTryp database was interrogated as before, the results of which are shown in Table 6.5. As the TriTrypDB entry page for Tb927.8.5670 provided no useful information, a basic protein BLAST search was performed using the protein sequence as query. This returned 13 sequences producing significant alignments, all hypothetical proteins from trypanosomatids, except a hypothetical protein from Paenibacillus larvae (e-value = 6.8). Thus, it appears this could be a Trypanosomatid-specific protein. Tb927.6.4180 is predicted to contain a FUN14 domain (e-value = 1.40×10^{-18}). The Pfam (http://pfam.sanger.ac.uk) annotation for this domain (PF04930) states that this family are short proteins that may contain transmembrane helices and are found in eukaryotes and some archaea. This protein has also been detected in the bloodstream form plasma membrane fraction and is predicted to possess two transmembrane helices; hence it may be a transmembrane protein present in the bloodstream plasma membrane. Tb927.8.1270 contains a predicted Hsp20-chaperone domain (e-value = 1.00×10^{-2}). The Interpro (http://www.ebi.ac.uk/interpro) entry for this domain (IPR0008978) states that the Hsp20 family is a small family of mammalian heat-shock proteins most abundant in skeletal muscle and heart. The TriTrypDB entry page states that this protein has been detected in the procyclic form and palmitoylated proteomes (Table 6.5). Therefore, this may be an Hsp20-related protein found in procyclic forms. Tb09.211.2300 is annotated as 'ATP-dependent DEAD/H RNA helicase, putative'. For further information on Tb09.211.2300 see chapter 4.

6.7 Gene Ontology (GO) analysis

In order to obtain an overall view of the classes of genes enriched in each of the polysome RNA samples compared to their corresponding total RNA samples the Gene Ontology (GO) terms associated with the genes in both the comparison datasets were analysed. Gene ontology terms are divided into three domains; 'molecular function', 'biological process' and 'cellular compartment'. Cellular compartment terms refer to the associated cell part or extracellular environment. This can be an

anatomical structure or a gene product group, such as ‘proteasome’. Molecular function terms refer to the molecular activities of the gene product, such as catalytic or binding activities. Biological process terms refer to a series of events completed by one or more molecular functions in a defined order, such as signal transduction or a cellular physiological process. Therefore, analysis of associated GO terms provides an overview of the processes or gene groups associated with a particular condition or state.

6.7.1 Initial GO analysis

The associated GO terms were obtained for the set of transcripts identified as upregulated in stumpy polysome RNA compared to stumpy total RNA and for the set of transcripts identified as upregulated in slender polysome RNA compared to slender total RNA as well (see Section 2.1.3). The GO terms returned from each gene set were then counted and the fold enrichment calculated by comparison of the count number between both datasets to find the GO terms enriched in either stumpy polysomal RNA or slender polysomal RNA.

Tables 6.6 and 6.7 show the GO terms with a fold enrichment of greater-than 1.5, obtained for stumpy form polysomal enriched transcripts and slender form polysomal enriched transcripts, respectively. For the GO terms enriched in stumpy form polysomal enriched transcripts there are terms connected with the proteasome and mitochondrion cellular compartments, ubiquitin and translation initiation processes, and enzyme activities such as endopeptidase, as well as calcium ion and nucleic acid binding functions (Table 6.6).

Table 6.6 Gene Ontology (GO) annotations enriched by greater than 1.5-fold in genes enriched in stumpy form polysomal transcripts compared to genes enriched in slender form polysomal transcripts.

GO ID	GO description	Fold enrichment
Cellular component		
GO:0019774	proteasome core complex, beta-subunit complex	2.45
GO:0005739	mitochondrion	1.74
GO:0005838	proteasome regulatory particle	1.63
GO:0005839	proteasome core complex	1.63
Molecular function		
GO:0009055	electron carrier activity	3.27
GO:0003743	translation initiation factor activity	2.45
GO:0004175	endopeptidase activity	2.45
GO:0005509	calcium ion binding	2.45
GO:0003676	nucleic acid binding	1.63
GO:0003755	peptidyl-prolyl cis-trans isomerase activity	1.63
GO:0004197	cysteine-type endopeptidase activity	1.63
GO:0004221	ubiquitin thiolesterase activity	1.63
GO:0015042	trypanothione-disulfide reductase activity	1.63
Biological process		
GO:0006511	ubiquitin-dependent protein catabolic process	3.00
GO:0006464	protein modification process	2.45
GO:0019538	protein metabolic process	2.45
GO:0000398	nuclear mRNA splicing, via spliceosome	1.63
GO:0006413	translational initiation	1.63

Table 6.7 Gene Ontology (GO) annotations enriched by greater than 1.5-fold in genes enriched in slender form polysomal transcripts compared to genes enriched in stumpy form polysomal transcripts.

GO ID	GO description	Fold enrichment
Cellular component		
GO:0005840	ribosome	3.06
GO:0016021	integral to membrane	1.57
Molecular function		
GO:0003735	structural constituent of ribosome	4.90
GO:0004672	protein kinase activity	4.90
GO:0004674	protein serine/threonine kinase activity	2.86
GO:0016491	oxidoreductase activity	2.45
GO:0005524	ATP binding	2.18
Biological process		
GO:0006470	protein dephosphorylation	3.67
GO:0006412	translation	3.26
GO:0006468	protein phosphorylation	2.75
GO:0051726	regulation of cell cycle	2.45
GO:0008152	metabolic process	1.84

The presence in this dataset of many genes with annotations associated with the mitochondrion has already been noted (see section 6.5.2; also in the dataset of genes upregulated in stumpy polysomal RNA compared to stumpy total RNA, see section 6.4.3), so the presence of the 'mitochondrion' GO term here confirms this observation.

Several of the enriched GO terms were associated with the proteasome. Although not noted elsewhere, the appearance of many genes with annotations associated with proteasome subunits was also observed in both this dataset and the dataset of genes upregulated in stumpy polysomal RNA compared to stumpy total RNA. This suggests that in stumpy forms there is a high level of protein degradation. One of the enriched GO terms is 'proteasome regulatory particle' and several genes with proteasome regulatory subunit annotations were found upregulated in stumpy polysomal RNA compared to stumpy total RNA, for example, Tb927.10.3030 (2.96 fold change, $p = 2.05 \times 10^5$) and Tb927.2.2440 (2.37 fold change, $p = 8.05 \times 10^5$). It could be speculated that the change in abundance of these regulatory proteins provides a change in proteasome composition, and therefore, regulation of proteasome substrates in different life-cycle stages. Connected with this is an enrichment of the GO terms 'ubiquitin-dependent protein catabolic process' and 'ubiquitin thiolesterase activity', which supports the suggestion of an increase in protein turnover in stumpy forms.

A few of the enriched GO terms in this analysis were associated with gene expression; these were 'translation initiation factor activity', 'translational initiation', 'nucleic acid binding', and 'nuclear mRNA splicing, via spliceosome'. As noted in section 6.4.4 many genes with these and other types of annotations associated with gene expression were found in the dataset of genes upregulated in stumpy polysomal RNA compared to stumpy total RNA. This could suggest that in stumpy forms there is a change in the proteins involved in the gene regulation that are elevated to mediate a change in gene expression control. This could mediate gene expression to prepare the parasite for release of translational repression during the differentiation to procyclic form.

For slender form polysomal enriched transcripts the number of enriched GO terms is much less. This is perhaps reflective of an overall increase in many different types of processes, such that there are many low counts for various cellular compartments, processes and functions. In agreement with this, many of the enriched GO terms were general terms such as 'integral to membrane', 'protein dephosphorylation' and 'regulation of cell cycle'. Several terms, however, do appear to be associated with signalling pathways, as there are a number related to protein phosphorylation/dephosphorylation, as well as ATP binding. Again, there are a two related to translation, these being 'ribosome' and 'translation'. Hence, the presence of similar translation-related terms in the results from the genes upregulated in stumpy polysomal RNA may be less specific to that dataset.

Gene ontologies are designated as a hierarchy whereby there are three parental terms at the top of the hierarchy ('molecular function', 'biological process' and 'cellular compartment'). As would be expected, for each of the three parent terms there is no fold enrichment (the fold changes are 0.92, 1.02 and 1.16) between the stumpy form and slender form datasets. Since these terms are equally represented in each dataset, any GO annotations with a large fold change should signify a true enrichment.

6.7.2 GO-Stats analysis

Since many GO term enrichments were based on relatively low underlying reads, there was the potential for false positive calls. For example, the terms with an enrichment of less than 2-fold may not be true enrichments. Therefore, a different analysis of the enriched GO terms was sought. 'GO-Stats' is a GO Toolbox (http://genome.crg.es/GOToolBox/) online tool for identifying statistically relevant over- and under-represented GO terms in gene datasets. This was used to identify GO terms enriched or depleted in either stumpy form polysomal enriched transcripts or slender form polysomal enriched transcripts. However, unlike the earlier analysis, comparison to the genomic frequency of each term was used, rather than between gene sets (see Section 2.1.4). The statistically significant enriched and depleted terms for stumpy form polysomal enriched transcripts and slender form polysomal enriched transcripts, are shown in Tables 6.8 and 6.9, respectively.

Table 6.8 Enriched and depleted Gene Ontology (GO) terms in genes upregulated in stumpy form polysomal RNA compared to stumpy form total RNA, as identified by GO-Stats analysis.[a]

GO ID	GO TERM	No. in Ref[b]	Frequency in Ref	No. in Set[c]	Frequency in Set	*P* value [d]	Fold change
Enrichment (p < 0.01)							
GO:0044257	cellular protein catabolic process	39	0.0137	11	0.0821	0.0005109	5.99
GO:0043632	modification-dependent macromolecule catabolic process	39	0.0137	11	0.0821	0.0005109	5.99
GO:0006511	ubiquitin-dependent protein catabolic process	39	0.0137	11	0.0821	0.0005109	5.99
GO:0019941	modification-dependent protein catabolic process	39	0.0137	11	0.0821	0.0005109	5.99
GO:0051603	proteolysis involved in cellular protein catabolic process	39	0.0137	11	0.0821	0.0005109	5.99
GO:0000502	proteasome complex	32	0.0114	10	0.0735	0.0007826	6.45
GO:0005839	proteasome core complex	14	0.005	7	0.0515	0.0008594	10.30
GO:0043231	intracellular membrane-bounded organelle	433	0.1548	42	0.3088	0.0010563	1.99
GO:0043227	membrane-bounded organelle	433	0.1548	42	0.3088	0.0010563	1.99
GO:0044424	intracellular part	829	0.2963	65	0.4779	0.0013149	1.61
GO:0034962	cellular biopolymer catabolic process	43	0.0151	11	0.0821	0.0014657	5.44
GO:0044237	cellular metabolic process	1137	0.3994	79	0.5896	0.0014853	1.48
GO:0005622	intracellular	839	0.2999	65	0.4779	0.0020475	1.59
GO:0008152	metabolic process	1280	0.4496	85	0.6343	0.0029007	1.41
GO:0044265	cellular macromolecule catabolic process	67	0.0235	13	0.097	0.0046564	4.13
GO:0043170	macromolecule metabolic process	974	0.3421	69	0.5149	0.0064545	1.51
Enrichment (0.01 < p < 0.05)							
GO:0044260	cellular macromolecule metabolic process	908	0.3189	65	0.4851	0.0102599	1.52
GO:0043283	biopolymer metabolic process	948	0.333	67	0.5	0.0106588	1.50
GO:0005739	mitochondrion	163	0.0583	21	0.1544	0.010753	2.65
GO:0034960	cellular biopolymer metabolic process	884	0.3105	63	0.4701	0.0178198	1.51
GO:0009055	electron carrier activity	5	0.0018	4	0.028	0.0191346	15.56
GO:0044238	primary metabolic process	1146	0.4025	76	0.5672	0.0196879	1.41
GO:0009987	cellular process	1515	0.5321	93	0.694	0.0232564	1.30
GO:0044248	cellular catabolic process	90	0.0316	14	0.1045	0.0272371	3.31
GO:0043234	protein complex	248	0.0886	26	0.1912	0.0383897	2.16

*Depletion (*p < *0.01)*							
GO:0031224	intrinsic to membrane	450	0.1608	6	0.0441	0.0077352	0.27
GO:0016021	integral to membrane	450	0.1608	6	0.0441	0.0077352	0.27
Depletion (0.01 < p < *0.05)*							
GO:0016020	membrane	619	0.2212	12	0.0882	0.0107436	0.40

[a] http://genome.crg.es/GOToolBox/ (see Section 2.1.4)
[b] Ref: The reference gene set created by the GO-Stats programme from the *Trypanosoma brucei* genomic distribution of GO terms
[c] Set: The gene set under analysis
[d] *P* values were calculated using the hypergeometric statistical test option and Bonferroni correction

Table 6.9 Enriched and depleted Gene Ontology (GO) terms in genes upregulated in slender form polysomal RNA compared to slender form total RNA, identified by GO-Stats analysis.[a]

GO ID	GO TERM	No. in Ref[b]	Frequency in Ref	No. in Set[c]	Frequency in Set	*P* value[d]	Fold change
*Enrichment (*p < *0.01)*							
GO:0044238	primary metabolic process	1146	0.4025	71	0.6455	4.36E-05	1.60
GO:0044237	cellular metabolic process	1137	0.3994	70	0.6364	8.37E-05	1.59
GO:0008152	metabolic process	1280	0.4496	75	0.6818	0.000146	1.52
GO:0034960	cellular biopolymer metabolic process	884	0.3105	57	0.5182	0.0009043	1.67
GO:0044260	cellular macromolecule metabolic process	908	0.3189	58	0.5273	0.0009313	1.65
GO:0043283	biopolymer metabolic process	948	0.333	59	0.5364	0.0017398	1.61
GO:0043170	macromolecule metabolic process	974	0.3421	60	0.5455	0.0018888	1.59

[a] http://genome.crg.es/GOToolBox/ (see Section 2.1.4)
[b] Ref: The reference gene set created by the GO-Stats programme from the *Trypanosoma brucei* genomic distribution of GO terms
[c] Set: The gene set under analysis
[d] *P* values were calculated using the hypergeometric statistical test option and Bonferroni correction

Although the enriched GO terms resulting from the analysis using GO-Stats were different to those found with the previous analysis, there are similar types of GO terms in both sets (Table 6.8). For GO terms enriched in stumpy form polysome-associated transcripts, 'mitochondrion', 'proteasome core complex', 'electron carrier activity', and 'ubiquitin-dependent protein catabolic process' are enriched in the results of both GO analyses. Many similar terms to these were identified by the GO-Stats analysis too. Therefore, the enrichment of these GO terms appears reliable. Translation-related terms are absent from the results of this GO-Stats analysis and therefore, this is in agreement with the observation from the above GO analysis, that translation-related terms were found enriched in both datasets and so that these are probably not true enrichments in either dataset. Additionally, the GO-Stats analysis identified three depleted terms, all membrane-related. This may indicate a downregulation of membrane proteins in the genes enriched in stumpy form polysome-associated transcripts.

The GO-Stats analysis also identified few enriched terms for the slender polysome associated transcripts (Table 6.9), similarly to the straightforward GO terms analysis. Nonetheless, the GO terms were similar, all being related to metabolic processes. In fact, comparison of Table 6.9 to Table 6.8 reveals that all the GO terms found enriched in slender form polysome-associated transcripts were also found enriched in stumpy form polysome-associated transcripts. This suggests that these metabolic processes could be generally upregulated in bloodstream-form cells rather than specifically upregulated in either slender or stumpy forms.

6.8 RIT-seq analysis

A *T. brucei* genome-wide RNAi-screen was recently been published by Alsford *et al*. (2011). This identified genes that caused the phenotypes of 'defect in bloodstream form growth', 'defect in procyclic form growth' and 'defect in *in vitro* differentiation to procyclic form', as well as an 'overall loss-of-fitness' through RNAi. As stumpy forms are preadapted for transmission to the tsetse fly, the genes they express are likely to be important for differentiation to the procyclic form. It was hypothesised

that there may be an increase in the occurrence of genes associated with a loss-of-fitness in differentiation upon RNAi in the dataset of genes preferentially translated in stumpy forms. Hence, the presence of each RNAi phenotype was investigated in the expression datasets using the data generated by Alsford *et al*. (2011).

The phenotypes observed in the study by Alsford *et al*. (2011) were associated with each gene from the expression analysis using their binary scoring system which produced five groups: significant loss-of-fitness in bloodstream form, in procyclic form, in differentiation, in all experiments, and no significant loss-of-fitness in any experiment. In each expression dataset the occurrence of each phenotype was then compared and shown in Table 6.10. As can be seen, there was no specific enrichment or depletion of any RNAi phenotype in either stumpy form polysome associated transcripts or slender form polysome associated transcripts. As would be expected, the percentage of genes in each dataset with a non-specific loss-of-fitness was almost identical.

Although it was believed that there might be an increase in the occurrence of genes associated with a loss-of-fitness in differentiation upon RNAi among transcripts preferentially translated in stumpy forms, this was not observed. However, this could be due to the methodology used in the RNAi-screen to test for a loss-of-fitness in differentiation. To test for differentiation the authors used bloodstream monomorphic cells stimulated to differentiate with the addition of *cis*-aconitate *in vitro*. In contrast, the dataset of transcripts enriched in stumpy form polysomes used pleomorphic bloodstream cells. Genes associated with the biology of stumpy forms and their role in transmission would, therefore, be missing.

Table 6.10 Percentage of genes upregulated in slender polysome RNA or stumpy polysome RNA with a phenotype from the genome-wide RNAi screen study by Alsford *et al*. (2011)[a].

	Loss-of-fitness in BSF	Loss-of-fitness in PCF	Loss-of-fitness in DIF	Loss-of-fitness all groups	Non-specific loss-of-fitness
Slender	4.53% (21)	2.80% (13)	6.47% (30)	12.28% (57)	36.21% (168)
Stumpy	5.91% (30)	3.35% (17)	5.91% (30)	11.42% (58)	36.61% (186)

[a] Numbers in parentheses are the number of genes in each dataset with the phenotype. BSF: bloodstream form, PCF: procyclic form, DIF: differentiation.

Table 6.11 Genes upregulated in stumpy form polysome RNA which have a 'loss-of-fitness in differentiation' phenotype from Alsford *et al*. (2011).

Feature ID [a]	Product description	Fold change [b]	*p* value [c]
Tb927.7.2070 *	heat shock protein DNAJ, putative	3.73	0
Tb09.211.0700	isopentenyl-diphosphate delta-isomerase, putative,isomerase, putative	3.67	0
Tb927.10.15810	hypothetical protein, conserved	3.47	0
Tb927.6.2820	hypothetical protein, conserved	3.02	0
Tb09.211.3420	riboflavin kinase, putative	2.96	2.98E-07
Tb927.8.580	hypothetical protein, conserved	2.96	0
Tb09.211.3310	replication factor C, subunit 3, putative	2.93	1.64E-07
Tb09.160.2020	thioredoxin (trx)	2.92	0
Tb927.4.4620	cytochrome oxidase subunit VIII (COXVIII)	2.82	2.24E-06
Tb927.5.3810	orotidine-5-phosphate decarboxylase/orotate phosphoribosyltransferase, putative,OMPDCase-OPRTase, putative	2.81	2.95E-10
Tb927.3.4680	RAB GDP dissociation inhibitor alpha, putative	2.8	5.91E-08
Tb927.5.2110	hypothetical protein, conserved	2.8	4.98E-09
Tb09.211.1610 *	phosphatidylserine decarboxylase, putative	2.77	0
Tb11.01.1820	Biotin--acetyl-CoA-carboxylase ligase, putative, biotin--protein ligase, putative	2.68	1.69E-12
Tb09.160.3110	mitochondrial processing peptide beta subunit, putative,metallo-peptidase, Clan ME, Family M16 (1L12.450)	2.64	1.18E-11
Tb927.7.630	hypothetical protein, conserved	2.62	3.85E-05
Tb09.160.0410	hypothetical protein, conserved	2.56	4.71E-07
Tb927.3.5150 *	exonuclease, putative	2.56	1.15E-03
Tb09.211.0310	hypothetical protein, conserved	2.54	2.65E-05
Tb927.10.14860	hypothetical protein, conserved	2.37	8.54E-12
Tb927.7.3100	hypothetical protein, conserved	2.31	4.52E-12
Tb11.02.4900	hypothetical protein, conserved	2.26	1.13E-08
Tb927.7.2740	hypothetical protein, conserved	2.25	0
Tb927.6.4570	hypothetical protein, conserved	2.18	0
Tb11.02.2130	hypothetical protein, conserved	2.14	4.17E-04
Tb927.6.4990	ATP synthase, epsilon chain, putative	2.11	0
Tb927.5.1270	hypothetical protein, conserved	2.1	2.00E-05
Tb927.10.10160	hypothetical protein, conserved	2.08	2.28E-05
Tb927.10.10420	monothiol glutaredoxin, putative	2.02	0
Tb927.8.7010 *	chaperone protein DNAj, putative	2.02	6.77E-12

[a] asterisks indicate genes which are present in both datasets of transcripts upregulated in slender form polysome RNA and transcripts upregulated in stumpy form polysome RNA

[b] Fold change upregulation in stumpy-form polysome RNA compared to stumpy form total RNA

[c] *p* value associated with the fold change; corrected for false discovery rate (FDR)

It was observed that the number of genes with a loss-of-fitness in differentiation phenotype was the same in both datasets (Table 6.10). So to investigate if both datasets contained the same genes identified as important for differentiation, the

genes with the differentiation RNAi phenotype were compared. Table 6.11 shows the genes in the dataset of transcripts enriched in stumpy polysomes that have a loss-of-fitness in differentiation phenotype. With comparison to the transcripts enriched in slender form polysomes, only four genes were present in both datasets (marked with an asterisk in Table 6.11). Hence, genes important for procyclic form differentiation apparently differs between monomorphic and pleomorphic cells.

6.9 Analysis of downstream sequences of enriched genes

Those transcripts enriched in the polysomal fractions of stumpy form cells with regard to the total RNA were analysed for 3'UTR sequence that might control their translation. This involved analysis of the sequences downstream of the gene stop codon for each gene using the online tool, Regulatory Sequences Analysis Tool (RSAT; http://rsat.scmbb.ulb.ac.be/rsat/; see section 2.1.5), which identifies statistically over-represented oligomers in sets of DNA sequences.

The RSAT oligo-analysis tool (van Helden *et al.*, 1998) was used to identify regulatory elements, as this tool is designed to detect over-represented oligonucleotides in co-regulated gene sets, and therefore, to identify regulatory elements. Although used initially to identify DNA binding sites for transcription factors in yeast genes, it can be utilised to identify other regulatory elements (van Helden *et al.*, 1998).

Initially, the Markov model option was used as the background model in the oligo-analysis tool to identify over-represented oligomers in the 300nt downstream-sequences for genes enriched in either stumpy form or slender form polysomes. This length of downstream-sequence was used because it is close to the median length found for all mapped 3'UTRs (400 bp, Siegel *et al*. 2010). However, the results appeared to be fairly unspecific, with many of the oligomers identified present in both analyses. Therefore, to increase specificity the same approach was used as has been used previously, in that oligomer frequencies were compared to those obtained for the 300nt downstream-sequences of genes present on chromosomes 1 and 2

(Mayho *et al*. 2006). Genes present on chromosomes 1 and 2 were chosen as a representative set because it is unlikely that these are co-regulated and so should show no bias in oligomer frequency, except for that which is seen in sequences of 300nt downstream of genes throughout the *T. brucei* genome. For this comparison, the 300nt of downstream-sequence for each of 475 genes present on chromosomes 1 and 2 was retrieved, and used as input in the oligo-analysis tool to generate a frequency file of 8-mers, which was subsequently used as the background model in the oligo-analysis tool with 300nt downstream-sequences of genes in the test dataset. Again, most of the over-represented oligomers (8-mers) identified were present in both datasets, perhaps representing 'bloodstream-form'-specific 8-mer sequences. After these were excluded, 8 unique 8-mers were present for each of the 300nt downstream-sequences of genes enriched in either stumpy form or slender form polysome-associated RNA. These are shown in Table 6.12.

As can be seen in Table 6.12, within each set of 8-mers there appears to be some similarity. For example, in the 'enriched in stumpy form polysomal transcripts' the sequence 'TTATT' occurs in 3 of the 8-mers, and 'TATT' occurs in another. This sequence was not found in any of the unique 8-mers found over-represented in genes enriched in slender form polysomal transcripts. Interestingly, this 'TTATT' 5-mer, along with a 'TTA' 3-mer, 'TATT' 4-mer, and 'TTATTA' 6-mer were found to be over-represented using this same approach in *T. brucei COX* 3'UTRs, as well as in procyclic-enriched transcripts with similar AU-rich oligomers (Matthew Mayho, PhD thesis, University of Edinburgh). Therefore, this appears to confirm the presence of a large number of genes upregulated in procyclic forms in the stumpy form polysome enriched transcript dataset.

Table 6.12 Oligomer (8-mer) sequences, found by Regulatory Sequences Analysis Tool (RSAT; http://rsat.scmbb.ulb.ac.be/) oligo-analysis tool, over-represented in either 300nt of downstream-sequences of stumpy form or slender form polysomal transcripts[a].

Sequence	Expected frequency	Occurance	E-value
Enriched in stumpy-form polysomal transcripts			
attattat	3.0365E-07	39	7.00E-95
tttgtttg	3.0365E-07	39	7.00E-95
aagaaaag	3.0365E-07	38	6.10E-92
tattatta	3.0365E-07	38	6.10E-92
gtttttgt	3.0365E-07	37	5.20E-89
ttattgtt	3.0365E-07	37	5.20E-89
ttgttttg	3.0365E-07	37	5.20E-89
tatatatt	3.0365E-07	36	4.20E-86
Enriched in slender-form polysomal transcripts			
ctcttttt	3.0365E-07	47	1.00E-120
tctctttt	3.0365E-07	47	1.00E-120
tccttttt	3.0365E-07	42	2.00E-105
tgtgtgtt	3.0365E-07	42	2.00E-105
ttcctttt	3.0365E-07	42	2.00E-105
ttttaaa	3.0365E-07	42	2.00E-105
gttttgtt	3.0365E-07	41	2.00E-102
actttttt	3.0365E-07	39	2.00E-96

[a] Using RSAT oligo-analysis, a background model of oligomer (8-mer) sequence occurrence frequency was created from 300nt downstream-sequences for genes on Chromosomes 1 and 2. This was used to identify over-represented 8-mers in the 300nt downstream-sequences of genes enriched in either stumpy form or slender form polysome-associated RNA. The over-represented 8-mers were then compared to find unique 8-mers present in each dataset.

In the unique 8-mers over-represented in slender form polysomal transcripts many contain stretches of many T residues. One of these, 'GTTTTGTT' is very similar to one from the stumpy form polysomal transcripts 8-mers, 'TTGTTTTG'. Therefore, these sequences are unlikely to be specific to either dataset. Except for one, each of the unique 8-mers in this dataset contains 4-5 consecutive T residues. Thus, it would appear that the sequence 'TTATT' is over-represented in the 300nt downstream-sequences of genes enriched in stumpy form polysome-associated RNA and the sequence 'TTTT/TTTTT' is over-represented in the 300nt downstream-sequences of genes enriched in slender form polysome-associated RNA. These sequences could therefore be involved in the control of life-cycle stage-specific translation, either in

promotion of translation in the respective life-cycle stage in which it is over-represented or in repression of translation in the other life-cycle stages.

The ‘feature map’ tool (see section 2.1.5) allowed visualisation of the location of the over-represented unique 8-mers in the 300nt downstream-sequence of the respective dataset. This showed that they are relatively evenly distributed throughout the genes present and with no obvious bias to location along the 300nt for either dataset (see Appendices 4 and 5).

6.10 Discussion

6.10.1 Radioactive methionine labelling throughout stumpy form to procyclic form differentiation

To investigate proteins that escape translational repression in stumpy forms and how this profile changes throughout differentiation to the procyclic form, the synthesis of new proteins was followed using radiolabelled methionine. The amount of protein synthesis was much lower in stumpy forms compared to slender and procyclic forms, and increased during differentiation (Figure 6.1 A). This was consistent with another study (Shapiro and Kimmel, 1987) and of translational repression in stumpy forms (Brecht and Parsons, 1998), and also demonstrated the release of translational repression during differentiation. However, the resolution of this methodology was not sufficient to determine individual proteins that escape repression in stumpy forms or throughout differentiation.

The amount of protein synthesis was also analysed by quantification of radiolabel incorporation. This demonstrated the increasing amount of protein synthesis throughout differentiation with time, and that by 8 hours into differentiation the level of protein synthesis was approaching that observed for established procyclic forms (Figure 6.2). The greatest increase in the amount of protein synthesis throughout differentiation was between 4 and 6 hours after the initiation of differentiation (Figure 6.2). This is coincident with a previously reported dramatic increase in RNA synthesis between 4 and 10 hours into differentiation (Pays *et al.*, 1993).

Interestingly, this result is in contrast to a previous study that reported a decrease in protein synthesis at the onset of *in vitro* differentiation (Bass and Wang, 1992). However, the study by Bass and Wang (1992) differed in that differentiation to the procyclic form was initiated directly in monomorphic slender forms by a temperature decrease alone, whereas in the study presented here, stumpy forms and *cis*-aconitate were used. It is likely that the contrasting results reflect the different methodologies. For example, the decrease in protein synthesis observed in the Bass and Wang (1992) study could have resulted as a cold-shock response to the temperature decrease,

which occurred because the temperature decrease was unaccompanied by the *cis*-aconitate signal. Alternatively, it may have represented a stumpy-like transitional phase needed for differentiation of monomorphic forms to the procyclic form, such that the cells passed through a stage of decreased protein synthesis similar to that observed in stumpy forms.

6.10.2 Analysis of actively translated transcripts using Illumina digital tag RNA sequencing

To identify which mRNA transcripts escape the translational repression in stumpy forms, the polysome-associated mRNAs were compared with total mRNA from stumpy forms (AnTat1.1 strain) and monomorphic slender forms (427 strain) using Illumina digital tag RNA sequencing. There are limitations of this approach, for example, the slender and stumpy populations used were from monomorphic and pleomorphic strains, respectively, and so may exhibit gene expression differences that may bias comparisons. Likewise, only one biological replicate was analysed, therefore, gene expression comparisons between different populations could also result in the identification of false positive transcript enrichments. It would, therefore, be necessary to confirm any transcript enrichment of genes identified for further study through regeneration of the infections and samples, followed by Northern blotting or qRT-PCR.

Also, this approach used an RNA sequencing procedure reliant on the *Nla*III restriction enzyme site sequence, which is not present in every gene sequence. Transcripts that do not possess this sequence will therefore be excluded from the analysis. However, this should not cause a problem because this analysis involved comparison between transcript abundance in different RNA samples rather than comparison between the abundance of different transcripts. Unless, for a given transcript there is differential RNA processing between the two life-cycle stages that results in the loss or gain of an *Nla*III site sequence, and thus gives an apparent change in transcript life-cycle abundance. However, it is likely that this would be a rare occurrence.

The transcripts identified as enriched in the stumpy polysomal RNA appeared to confirm this life-cycle stage as preadapted for transmission and differentiation to the procyclic form. This is because there was a large number of enriched transcripts encoding proteins associated with the mitochondrion or procyclic-enriched transcripts, such as COX VIII, VII and COX V. As noted in Section 6.4.3, it was previously suggested that stumpy forms accumulate mitochondrial transcripts before developing a functional mitochondrion in procyclic forms (Michelotti and Hajduk, 1987) and the data presented here supports this suggestion.

Enrichment of transcripts encoding proteins associated with other types of cellular processes/components was also noted, such as gene expression or the proteasome. This was also confirmed in comparisons of associated GO terms in the datasets of enriched transcripts in slender forms and stumpy forms (Tables 6.6, 6.7, 6.8 and 6.9). The enrichment of transcripts associated with the proteasome in stumpy forms is also consistent with their preadaptation for differentiation, because presumably, each cell will have a high protein turnover when it undergoes the great morphological change that this process involves, in addition to autophagy and glycosomal turnover.

As noted in Section 6.7.1, there was enrichment in proteasome regulatory particle/subunit annotations observed in stumpy form enriched transcripts. The 19S proteasome regulatory particle is important for proteasome substrate binding, unfolding and translocation of the substrate into the proteolytically active core particle, and it therefore, confers substrate specificity (Tomko *et al*., 2011). Thus, it may be speculated that this enrichment indicates that there is a change in regulation of proteasome substrates in different life-cycle stages. For example, in *Arabidopsis thaliana* a proteasome regulatory particle subunit (RPT2) plays a critical role in plant development, with *rpt2a* null mutants exhibiting a range of developmental defects (Lee *et al*., 2011). In addition, yeast cells in a stationary phase exhibit increased proteasome subunit expression, altered proteasome composition and increased levels of individual proteasome proteolytic activites which may play a role in maintaining cell viability during stationary phase (Chen *et al*., 2004). Therefore, suggesting a link

between a non-proliferative cell state and altered proteasome regulation that may support a similar connection in stumpy form trypanosome cells.

Analysis of the associated GO terms, therefore, was useful in confirming the observations of these enriched protein groups. However, one study reported that only 38% of *T. brucei* annotated genes have associated GO terms (Alsford *et al.*, 2011), meaning that a large number of genes are not represented in this analysis. Therefore, it is possible that enrichment of transcripts associated with particular cellular processes/components could have been overlooked in this analysis or that an identified enrichment may be a false positive result.

Additionally, another criticism that could be made of the Illumina digital tag RNA sequencing study presented here is that the polysome fractions used for extraction of polysome-associated RNA may be contaminated with other mRNAs. This is because it is possible that other high-molecular weight complexes that contain mRNAs were present in these fractions. These could be, for example, P bodies or stress granules, and therefore, contain mRNAs for degradation or storage. Thus, some transcripts concluded here to be actively translated in stumpy forms could, in contrast, be degraded or stored in this life-cycle stage.

The transcripts of a number of interesting genes were identified as enriched in the stumpy polysome RNA sample in this study. For example, Tb927.5.2590 which is predicted to possess a Macro domain (see Section 6.4.1); Tb09.160.0680 and Tb09.211.4513, which are annotated as putative sec1 family transport protein (SLY1) and kinetoplastid membrane protein-11 (KMP-11) respectively, Tb11.03.0250 which is annotated as cyclophilin A (see Section 6.4.2); Tb927.7.5400 which is predicted to possess a signal peptide and two transmembrane helices and may be a trypanosomatid-specific gene, Tb927.3.5760 which showed similarity to ESAG9 sequences (see Section 6.5.1); Tb927.8.5670 which may also be a trypanosomatid-specific gene, Tb927.6.4180 which is predicted to contain a FUN14 domain, and Tb09.211.2300 which is annotated as a putative ATP-dependent DEAD/H RNA helicase (see Section 6.6.4). However, analysis of the expression of

these transcripts in slender forms, intermediate forms, stumpy forms, procyclic forms and bloodstream monomorphic forms (from analysis performed by Dr. Paul Capewell of both Illumina digital tag sequencing experiments, see Section 6.3.1) revealed that of these genes only Tb927.7.5400, Tb927.8.5670, and possibly Tb927.3.5760 and Tb09.211.2300 are enriched in stumpy forms.

The next necessary experiment in this work is to validate the Illumina digital tag sequencing results using quantitative PCR or northern blot analysis, to confirm the mRNA expression levels for a number of transcripts. This would be essential to verify the expression profile of any interesting genes selected for further study before that work commenced. For example, the upregulation of Tb09.211.2300 in stumpy forms has already been confirmed by northern blot analysis, as this transcript was investigated in another part of this project (Figure 4.7 B). The further study of interesting genes may involve RNAi-mediated transcript ablation or ectopic overexpression of the encoded protein to assess the protein function.

Chapter 7

Overview

The question that was to be addressed in this project was how are translation and the escape from translational arrest in stumpy form *T. brucei* controlled? This encompassed four specific aims, each of which is addressed below with suggestions of further work.

- What signals control the expression, or escape from translational arrest, of genes expressed in stumpy forms?

For one stumpy-enriched transcript, ESAG9-EQ, a 33nt sequence, termed here the 'element' sequence, in the 3'UTR was identified as contributing strongly to gene silencing in slender forms. This sequence was also determined to be the signal responsible for transcript enrichment upon development to stumpy-like forms in response to 8-pCPT-2'-O-Me-cAMP. Analysis of reporter gene expression in pleomorphic form cells confirmed that this sequence was responsible for the release of translational repression in true stumpy forms. Thus, this sequence represents the first identification of an mRNA sequence responsive to cell-density and stumpy-induction factor signalling. Therefore, it would be interesting to investigate if an RNA-binding protein interacts with the ESAG9-EQ 3'UTR, and specifically the identified element, possibly through the use of an RNA EMSA (electromobility shift assay) before employing one of the techniques discussed in Section 3.10.2, such as tRNA tethering. As mentioned, this could enable dissection of the currently unknown SIF signalling pathway, in the opposite direction to signal transduction.

Additionally, it would be interesting to determine the secondary structure of the element sequence in the full-length 3'UTR experimentally, by RNase enzymatic digestion, for example. If an RNA-binding protein does interact with the element sequence, then mutational analysis of this sequence could determine which nucleotides are important for the interaction and so, together with the experimentally determined element structure, may assist with the discovery of other transcripts bearing this regulatory motif.

- Do mechanisms such as uORFs control the escape from repression, or repression, of genes expressed in stumpy forms?

Although it did appear initially that there was a slight enrichment in the abundance of potential uORFs in stumpy-enriched transcripts compared to heat-shock regulated transcripts or genome abundance (Table 4.1), upon analysis of predicted SAS locations and stumpy-enriched transcripts from an initial Illumina digital tag sequencing experiment, this slight enrichment was not present. This indicates that it is unlikely that uORFs are used as a general mechanism for the control of gene expression in stumpy forms, but this does not exclude the possibility that this mechanism is used to control the expression of some individual genes. However, of the three stumpy-enriched genes possessing potential uORFs that were investigated in this study (ESAG9-EQ, Tb09.211.2300 and Tb11.47.0019) none showed evidence of the identified potential uORFs contributing to the control of their expression.

For Tb11.47.0019, the use of three SAS locations was discovered, potentially giving rise to encoded proteins of different lengths from the annotated ATG or an internal ATG. It has been previously hypothesised that in *T. brucei* the use of differential SAS and PAS could be used to regulate gene expression and also regulate the inclusion/exclusion of various properties in a protein (Siegel *et al*., 2011). Initially, it appeared that the different SAS locations in Tb11.47.0019 may have been used in different ratios during the life-cycle, but variability in the Northern blot analyses meant that these results were inconclusive. The possibility that different protein isoforms of Tb11.47.0019 with different localisations was investigated by ectopic expression of either isoform with a Ty epitope tag. However, it appeared that the short protein form was preferentially expressed, regardless of the presence of the long protein sequence, either due to preferential *trans*-splicing at the downstream location or to preferential recognition of the downstream ATG start codon. Therefore, it appears that differential splicing does not always result in the production of two protein isoforms, at least within slender and stumpy forms. If this project were to be continued, it would be necessary to determine whether the longer protein isoform of Tb11.47.0019 is naturally expressed (for example, in stumpy

forms), which could be achieved by western blot analysis using an antibody raised against the protein sequence common to both isoforms.

- Are specific translational proteins involved in translational control in stumpy forms?

The transcripts of a number of translation initiation factor homologues were investigated for enrichment in stumpy forms, which could be involved in translational control in stumpy forms. As such, the mRNA of one eIF4E homologue, TbeIF4E4, was identified as stumpy-enriched and so may contribute to translational control in this life-cycle stage. Preliminary experiments indicated that RNAi-mediated transcript ablation of TbeIF4E4 resulted in a decreased growth rate and decreased polysome abundance, consistent with a previous study (Freire *et al*., 2011). But this was not analysed any further than the preliminary experiments performed, due to the finding that another group was also researching these eIF4E homologues.

Another translation initiation factor homologue that exhibited transcript enrichment in stumpy forms was the eIF6 homologue, TbeIF6. Either RNAi-mediated transcript ablation or protein overexpression of TbeIF6 resulted in decreased growth and dramatically decreased protein synthesis. This was possibly due to decreased levels of translation following both inductions, although the polysome analysis did not provide conclusive results. TbeIF6 did not appear to be involved in differentiation to the procyclic form, but it remains that the stumpy-enrichment suggests involvement of TbeIF6 in translational repression in stumpy forms. For further work of investigation of this protein, firstly it would be interesting to determine if the TbeIF6 protein is also upregulated in stumpy forms. Additionally, it is necessary to repeat the polysome profiles of cells induced for TbeIF6 RNAi or overexpression alongside uninduced controls, and also to repeat the northern blot analysis to detect *INGI* transcripts using RNA samples from TbeIF6 RNAi-induced cells. Furthermore, it would be interesting to assess the effect of RNAi or overexpression of TbeIF6 in stumpy form cells on differentiation to the procyclic form, and therefore, following

transfection of the relevant constructs into pleomorphic cells. However, this experiment may still suffer from secondary effects due to the induction of either of these conditions resulting in a decreased growth and decreased protein synthesis.

- Which transcripts escape translational repression in stumpy forms?

Protein synthesis in stumpy forms and throughout differentiation to the procyclic form was assessed using radiolabelled methionine. This confirmed that stumpy forms exhibit a decreased level of protein synthesis compared to slender and procyclic forms, and also demonstrated the release from translational repression during differentiation as protein synthesis increased during this process. To identify which transcripts escape translational repression in stumpy forms, Illumina digital tag RNA sequencing was used with total RNA and polysome-associated RNA samples from slender and stumpy forms. Transcripts enriched in the stumpy form polysome-associated RNA sample were determined and some interesting genes were identified for further analysis. These transcripts also appeared to confirm stumpy forms as preadapted for transmission and differentiation to the procyclic form, because there were many enriched transcripts that were associated with the mitochondrion. However, this expression data requires validation by northern blot analyses or quantitative PCR, and this would be the next step in the further work for this project area. Following this, interesting genes could then be investigated for their function and importance in stumpy form biology and differentiation through RNAi-mediated transcript ablation and protein overexpression.

In conclusion, stumpy forms exhibit decreased protein synthesis in comparison to proliferative life-cycle stages and the control of translation in this life-cycle stage may involve translational proteins, such as the eIF6 homologue. For the escape from translational repression in stumpy form *T. brucei*, 3'UTR sequences appear to be important, such as the ESAG9-EQ 3'UTR element sequence. Therefore, stumpy forms are similar to other life-cycle stages studied, where the control of gene expression by *cis*-acting sequences is also important. uORFs do not appear to be a mechanism important for general translational control in stumpy forms, although

they may control the expression of individual genes, but not the three genes analysed in this study. The control of gene expression and production of different protein isoforms may be regulated by different *trans*-splicing locations, however, whilst different SAS locations appear to be used for one gene (Tb11.47.0019) studied, it was observed that this did not result in the production of different protein isoforms, at least in slender forms. Additionally, transcripts enriched in the polysome-associated RNA of stumpy forms were identified in this study, and thus, should represent transcripts that escape the translational repression at this life-cycle stage.

Because stumpy forms are preadapted for transmission to the tsetse fly, a trypanosomiasis treatment developed to target this life-cycle stage could potentially block or reduce parasite transmission. This strategy could, however, result in an increase in virulent slender forms in the host and so may not be a viable treatment strategy unless used in combination with an additional treatment. An alternative method of exploiting stumpy form biology could be the use of compounds to accelerate stumpy form differentiation leading to reduced transmission because of a reduction in parasite numbers through cellular quiescence and host immune system infection control (MacGregor and Matthews, 2010). Consequently, understanding the biology of stumpy forms is an important area for research. At the commencement of this project, nothing was known regarding the control of gene expression in stumpy forms. Therefore, the work presented here makes a significant contribution to understanding the control of translation and escape from translational arrest in this important, transmissible, life-cycle stage of *T. brucei*.

Bibliography

AGABIAN, N. (1990) Trans splicing of nuclear pre-mRNAs. *Cell,* 61**,** 1157-60.
AKSOY, S., WILLIAMS, S., CHANG, S. & RICHARDS, F. F. (1990) SLACS retrotransposon from Trypanosoma brucei gambiense is similar to mammalian LINEs. *Nucleic Acids Res,* 18**,** 785-92.
ALSFORD, S., ECKERT, S., BAKER, N., GLOVER, L., SANCHEZ-FLORES, A., LEUNG, K. F., TURNER, D. J., FIELD, M. C., BERRIMAN, M. & HORN, D. (2012) High-throughput decoding of antitrypanosomal drug efficacy and resistance. *Nature*.
ALSFORD, S., TURNER, D. J., OBADO, S. O., SANCHEZ-FLORES, A., GLOVER, L., BERRIMAN, M., HERTZ-FOWLER, C. & HORN, D. (2011) High-throughput phenotyping using parallel sequencing of RNA interference targets in the African trypanosome. *Genome Res*.
ALTMANN, M. & TRACHSEL, H. (1989) Altered mRNA cap recognition activity of initiation factor 4E in the yeast cell cycle division mutant cdc33. *Nucleic Acids Res,* 17**,** 5923-31.
ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. (1990) Basic local alignment search tool. *J Mol Biol,* 215**,** 403-10.
AMULIC, B., SALANTI, A., LAVSTSEN, T., NIELSEN, M. A. & DEITSCH, K. W. (2009) An Upstream Open Reading Frame Controls Translation of var2csa, a Gene Implicated in Placental Malaria. *Plos Pathogens,* 5.
APHASIZHEVA, I., MASLOV, D., WANG, X., HUANG, L. & APHASIZHEV, R. (2011) Pentatricopeptide repeat proteins stimulate mRNA adenylation/uridylation to activate mitochondrial translation in trypanosomes. *Mol Cell,* 42**,** 106-17.
ARCHER, S. K., LUU, V. D., DE QUEIROZ, R. A., BREMS, S. & CLAYTON, C. (2009) Trypanosoma brucei PUF9 regulates mRNAs for proteins involved in replicative processes over the cell cycle. *PLoS Pathog,* 5**,** e1000565.
ARNOLD, K., BORDOLI, L., KOPP, J. & SCHWEDE, T. (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics,* 22**,** 195-201.
ATAYDE, V. D., TSCHUDI, C. & ULLU, E. (2011) The emerging world of small silencing RNAs in protozoan parasites. *Trends Parasitol,* 27**,** 321-7.
BANGS, J. D. (2011) Replication of the ERES:Golgi junction in bloodstream-form African trypanosomes. *Mol Microbiol,* 82**,** 1433-43.
BANGS, J. D., CRAIN, P. F., HASHIZUME, T., MCCLOSKEY, J. A. & BOOTHROYD, J. C. (1992) Mass spectrometry of mRNA cap 4 from trypanosomatids reveals two novel nucleosides. *J Biol Chem,* 267**,** 9805-15.
BARNWELL, E. M., VAN DEURSEN, F. J., JEACOCK, L., SMITH, K. A., MAIZELS, R. M., ACOSTA-SERRANO, A. & MATTHEWS, K. (2010) Developmental regulation and extracellular release of a VSG expression-site-associated gene product from Trypanosoma brucei bloodstream forms. *J Cell Sci,* 123**,** 3401-11.
BARRETT, M. P. (2010) Potential new drugs for human African trypanosomiasis: some progress at last. *Curr Opin Infect Dis,* 23**,** 603-8.
BASS, K. E. & WANG, C. C. (1992) Transient inhibition of protein synthesis accompanies differentiation of Trypanosoma brucei from bloodstream to procyclic forms. *Mol Biochem Parasitol,* 56**,** 129-40.

BASU, U., SI, K., WARNER, J. R. & MAITRA, U. (2001) The Saccharomyces cerevisiae TIF6 gene encoding translation initiation factor 6 is required for 60S ribosomal subunit biogenesis. *Mol Cell Biol,* 21**,** 1453-62.

BENDTSEN, J. D., JENSEN, L. J., BLOM, N., VON HEIJNE, G. & BRUNAK, S. (2004a) Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng Des Sel,* 17**,** 349-56.

BENDTSEN, J. D., NIELSEN, H., VON HEIJNE, G. & BRUNAK, S. (2004b) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol,* 340**,** 783-95.

BENELLI, D., MARZI, S., MANCONE, C., ALONZI, T., LA TEANA, A. & LONDEI, P. (2009) Function and ribosomal localization of aIF6, a translational regulator shared by archaea and eukarya. *Nucleic Acids Res,* 37**,** 256-67.

BENNE, R., VAN DEN BURG, J., BRAKENHOFF, J. P., SLOOF, P., VAN BOOM, J. H. & TROMP, M. C. (1986) Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell,* 46**,** 819-26.

BENZ, C., NILSSON, D., ANDERSSON, B., CLAYTON, C. & GUILBRIDE, D. L. (2005) Messenger RNA processing sites in Trypanosoma brucei. *Mol Biochem Parasitol,* 143**,** 125-34.

BERBEROF, M., VANHAMME, L., TEBABI, P., PAYS, A., JEFFERIES, D., WELBURN, S. & PAYS, E. (1995) The 3'-terminal region of the mRNAs for VSG and procyclin can confer stage specificity to gene expression in Trypanosoma brucei. *Embo J,* 14**,** 2925-34.

BERRIMAN, M., GHEDIN, E., HERTZ-FOWLER, C., BLANDIN, G., RENAULD, H., BARTHOLOMEU, D. C., LENNARD, N. J., CALER, E., HAMLIN, N. E., HAAS, B., BOHME, U., HANNICK, L., ASLETT, M. A., SHALLOM, J., MARCELLO, L., HOU, L., WICKSTEAD, B., ALSMARK, U. C., ARROWSMITH, C., ATKIN, R. J., BARRON, A. J., BRINGAUD, F., BROOKS, K., CARRINGTON, M., CHEREVACH, I., CHILLINGWORTH, T. J., CHURCHER, C., CLARK, L. N., CORTON, C. H., CRONIN, A., DAVIES, R. M., DOGGETT, J., DJIKENG, A., FELDBLYUM, T., FIELD, M. C., FRASER, A., GOODHEAD, I., HANCE, Z., HARPER, D., HARRIS, B. R., HAUSER, H., HOSTETLER, J., IVENS, A., JAGELS, K., JOHNSON, D., JOHNSON, J., JONES, K., KERHORNOU, A. X., KOO, H., LARKE, N., LANDFEAR, S., LARKIN, C., LEECH, V., LINE, A., LORD, A., MACLEOD, A., MOONEY, P. J., MOULE, S., MARTIN, D. M., MORGAN, G. W., MUNGALL, K., NORBERTCZAK, H., ORMOND, D., PAI, G., PEACOCK, C. S., PETERSON, J., QUAIL, M. A., RABBINOWITSCH, E., RAJANDREAM, M. A., REITTER, C., SALZBERG, S. L., SANDERS, M., SCHOBEL, S., SHARP, S., SIMMONDS, M., SIMPSON, A. J., TALLON, L., TURNER, C. M., TAIT, A., TIVEY, A. R., VAN AKEN, S., WALKER, D., WANLESS, D., WANG, S., WHITE, B., WHITE, O., WHITEHEAD, S., WOODWARD, J., WORTMAN, J., ADAMS, M. D., EMBLEY, T. M., GULL, K., ULLU, E., BARRY, J. D., FAIRLAMB, A. H., OPPERDOES, F., BARRELL, B. G., DONELSON, J. E., HALL, N., FRASER, C. M., et al. (2005) The genome of the African trypanosome Trypanosoma brucei. *Science,* 309**,** 416-22.

BIEBINGER, S., WIRTZ, L. E., LORENZ, P. & CLAYTON, C. (1997) Vectors for inducible expression of toxic gene products in bloodstream and procyclic Trypanosoma brucei. *Mol Biochem Parasitol,* 85**,** 99-112.

BISWAS, A., MUKHERJEE, S., DAS, S., SHIELDS, D., CHOW, C. W. & MAITRA, U. (2011) Opposing action of casein kinase 1 and calcineurin in nucleo-cytoplasmic shuttling of mammalian translation initiation factor eIF6. *J Biol Chem,* 286**,** 3129-38.

BITONTI, A. J., CROSS-DOERSEN, D. E. & MCCANN, P. P. (1988) Effects of alpha-difluoromethylornithine on protein synthesis and synthesis of the variant-specific glycoprotein (VSG) in Trypanosoma brucei brucei. *Biochem J,* 250**,** 295-8.

BLATTNER, J. & CLAYTON, C. E. (1995) The 3'-untranslated regions from the Trypanosoma brucei phosphoglycerate kinase-encoding genes mediate developmental regulation. *Gene,* 162**,** 153-6.

BOCHUD-ALLEMANN, N. & SCHNEIDER, A. (2002) Mitochondrial substrate level phosphorylation is essential for growth of procyclic Trypanosoma brucei. *J Biol Chem,* 277**,** 32849-54.

BOELAERT, M., MEHEUS, F., ROBAYS, J. & LUTUMBA, P. (2010) Socio-economic aspects of neglected diseases: sleeping sickness and visceral leishmaniasis. *Ann Trop Med Parasitol,* 104**,** 535-42.

BORST, P. & SABATINI, R. (2008) Base J: discovery, biosynthesis, and possible functions. *Annu Rev Microbiol,* 62**,** 235-51.

BRECHT, M. & PARSONS, M. (1998) Changes in polysome profiles accompany trypanosome development. *Mol Biochem Parasitol,* 97**,** 189-98.

BREIDBACH, T., NGAZOA, E. & STEVERDING, D. (2002) Trypanosoma brucei: in vitro slender-to-stumpy differentiation of culture-adapted, monomorphic bloodstream forms. *Exp Parasitol,* 101**,** 223-30.

BRUN, R., BLUM, J., CHAPPUIS, F. & BURRI, C. (2009) Human African trypanosomiasis. *Lancet,* 375**,** 148-59.

BURKARD, G., FRAGOSO, C. M. & RODITI, I. (2007) Highly efficient stable transformation of bloodstream forms of Trypanosoma brucei. *Mol Biochem Parasitol,* 153**,** 220-3.

CAFFREY, C. R., LIMA, A. P. & STEVERDING, D. (2011) Cysteine peptidases of kinetoplastid parasites. *Adv Exp Med Biol,* 712**,** 84-99.

CECI, M., GAVIRAGHI, C., GORRINI, C., SALA, L. A., OFFENHAUSER, N., MARCHISIO, P. C. & BIFFO, S. (2003) Release of eIF6 (p27BBP) from the 60S subunit allows 80S ribosome assembly. *Nature,* 426**,** 579-84.

CHAWLA, B., JHINGRAN, A., SINGH, S., TYAGI, N., PARK, M. H., SRINIVASAN, N., ROBERTS, S. C. & MADHUBALA, R. (2010) Identification and characterization of a novel deoxyhypusine synthase in Leishmania donovani. *J Biol Chem,* 285**,** 453-63.

CHEN, D., VOLLMAR, M., ROSSI, M. N., PHILLIPS, C., KRAEHENBUEHL, R., SLADE, D., MEHROTRA, P. V., VON DELFT, F., CROSTHWAITE, S. K., GILEADI, O., DENU, J. M. & AHEL, I. (2011) Identification of macrodomain proteins as novel O-acetyl-ADP-ribose deacetylases. *J Biol Chem,* 286**,** 13261-71.

CHEN, Q., THORPE, J., DING, Q., EL-AMOURI, I. S. & KELLER, J. N. (2004) Proteasome synthesis and assembly are required for survival during stationary phase. *Free Radic Biol Med,* 37, 859-68.
CHENDRIMADA, T. P., FINN, K. J., JI, X., BAILLAT, D., GREGORY, R. I., LIEBHABER, S. A., PASQUINELLI, A. E. & SHIEKHATTAR, R. (2007) MicroRNA silencing through RISC recruitment of eIF6. *Nature,* 447**,** 823-8.
CHOI, S. K., LEE, J. H., ZOLL, W. L., MERRICK, W. C. & DEVER, T. E. (1998) Promotion of met-tRNAiMet binding to ribosomes by yIF2, a bacterial IF2 homolog in yeast. *Science,* 280**,** 1757-60.
CLAYTON, C. & SHAPIRA, M. (2007) Post-transcriptional regulation of gene expression in trypanosomes and leishmanias. *Mol Biochem Parasitol,* 156**,** 93-101.
CLAYTON, C. E. (2002) Life without transcriptional control? From fly to man and back again. *Embo J,* 21**,** 1881-8.
COOPER, H. L., PARK, M. H., FOLK, J. E., SAFER, B. & BRAVERMAN, R. (1983) Identification of the hypusine-containing protein hy+ as translation initiation factor eIF-4D. *Proc Natl Acad Sci U S A,* 80**,** 1854-7.
CROSS, G. A. & MANNING, J. C. (1973) Cultivation of Trypanosoma brucei sspp. in semi-defined and defined media. *Parasitology,* 67**,** 315-31.
CROSS, M., KIEFT, R., SABATINI, R., DIRKS-MULDER, A., CHAVES, I. & BORST, P. (2002) J-binding protein increases the level and retention of the unusual base J in trypanosome DNA. *Mol Microbiol,* 46**,** 37-47.
CROWE, M. L., WANG, X. Q. & ROTHNAGEL, J. A. (2006) Evidence for conservation and selection of upstream open reading frames suggests probable encoding of bioactive peptides. *BMC Genomics,* 7**,** 16.
CZICHOS, J., NONNENGAESSER, C. & OVERATH, P. (1986) Trypanosoma brucei: cis-aconitate and temperature reduction as triggers of synchronous transformation of bloodstream to procyclic trypomastigotes in vitro. *Exp Parasitol,* 62**,** 283-91.
DE MARCO, N., IANNONE, L., CAROTENUTO, R., BIFFO, S., VITALE, A. & CAMPANELLA, C. (2010) p27(BBP)/eIF6 acts as an anti-apoptotic factor upstream of Bcl-2 during Xenopus laevis development. *Cell Death Differ,* 17**,** 360-72.
DEAN, S., MARCHETTI, R., KIRK, K. & MATTHEWS, K. R. (2009) A surface transporter family conveys the trypanosome differentiation signal. *Nature,* 459**,** 213-7.
DENNINGER, V., FIGARELLA, K., SCHONFELD, C., BREMS, S., BUSOLD, C., LANG, F., HOHEISEL, J. & DUSZENKO, M. (2007) Troglitazone induces differentiation in Trypanosoma brucei. *Exp Cell Res,* 313**,** 1805-19.
DHALIA, R., MARINSEK, N., REIS, C. R., KATZ, R., MUNIZ, J. R., STANDART, N., CARRINGTON, M. & DE MELO NETO, O. P. (2006) The two eIF4A helicases in Trypanosoma brucei are functionally distinct. *Nucleic Acids Res,* 34**,** 2495-507.
DHALIA, R., REIS, C. R., FREIRE, E. R., ROCHA, P. O., KATZ, R., MUNIZ, J. R., STANDART, N. & DE MELO NETO, O. P. (2005) Translation initiation in Leishmania major: characterisation of multiple eIF4F subunit homologues. *Mol Biochem Parasitol,* 140**,** 23-41.

DING, Y., CHAN, C. Y. & LAWRENCE, C. E. (2004) Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res,* 32**,** W135-41.

DOMENICALI PFISTER, D., BURKARD, G., MORAND, S., RENGGLI, C. K., RODITI, I. & VASSELLA, E. (2006) A Mitogen-activated protein kinase controls differentiation of bloodstream forms of Trypanosoma brucei. *Eukaryot Cell,* 5**,** 1126-35.

DROZDZ, M. & CLAYTON, C. (1999) Structure of a regulatory 3' untranslated region from Trypanosoma brucei. *RNA,* 5**,** 1632-44.

DUNCAN, R., MILBURN, S. C. & HERSHEY, J. W. (1987) Regulated phosphorylation and low abundance of HeLa cell initiation factor eIF-4F suggest a role in translational control. Heat shock effects on eIF-4F. *J Biol Chem,* 262**,** 380-8.

DURAND-DUBIEF, M. & BASTIN, P. (2003) TbAGO1, an argonaute protein required for RNA interference, is involved in mitosis and chromosome segregation in Trypanosoma brucei. *BMC Biol,* 1**,** 2.

EL-SAYED, N. M., MYLER, P. J., BARTHOLOMEU, D. C., NILSSON, D., AGGARWAL, G., TRAN, A. N., GHEDIN, E., WORTHEY, E. A., DELCHER, A. L., BLANDIN, G., WESTENBERGER, S. J., CALER, E., CERQUEIRA, G. C., BRANCHE, C., HAAS, B., ANUPAMA, A., ARNER, E., ASLUND, L., ATTIPOE, P., BONTEMPI, E., BRINGAUD, F., BURTON, P., CADAG, E., CAMPBELL, D. A., CARRINGTON, M., CRABTREE, J., DARBAN, H., DA SILVEIRA, J. F., DE JONG, P., EDWARDS, K., ENGLUND, P. T., FAZELINA, G., FELDBLYUM, T., FERELLA, M., FRASCH, A. C., GULL, K., HORN, D., HOU, L., HUANG, Y., KINDLUND, E., KLINGBEIL, M., KLUGE, S., KOO, H., LACERDA, D., LEVIN, M. J., LORENZI, H., LOUIE, T., MACHADO, C. R., MCCULLOCH, R., MCKENNA, A., MIZUNO, Y., MOTTRAM, J. C., NELSON, S., OCHAYA, S., OSOEGAWA, K., PAI, G., PARSONS, M., PENTONY, M., PETTERSSON, U., POP, M., RAMIREZ, J. L., RINTA, J., ROBERTSON, L., SALZBERG, S. L., SANCHEZ, D. O., SEYLER, A., SHARMA, R., SHETTY, J., SIMPSON, A. J., SISK, E., TAMMI, M. T., TARLETON, R., TEIXEIRA, S., VAN AKEN, S., VOGT, C., WARD, P. N., WICKSTEAD, B., WORTMAN, J., WHITE, O., FRASER, C. M., STUART, K. D. & ANDERSSON, B. (2005) The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease. *Science,* 309**,** 409-15.

EMANUELSSON, O., NIELSEN, H., BRUNAK, S. & VON HEIJNE, G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol,* 300**,** 1005-16.

ENGSTLER, M. & BOSHART, M. (2004) Cold shock and regulation of surface protein trafficking convey sensitization to inducers of stage differentiation in Trypanosoma brucei. *Genes Dev,* 18**,** 2798-811.

ESTEVEZ, A. M. (2008) The RNA-binding protein TbDRBD3 regulates the stability of a specific subset of mRNAs in trypanosomes. *Nucleic Acids Res,* 36**,** 4573-86.

ESTEVEZ, A. M., KEMPF, T. & CLAYTON, C. (2001) The exosome of Trypanosoma brucei. *EMBO J,* 20**,** 3831-9.

FEAGIN, J. E., JASMER, D. P. & STUART, K. (1986) Differential mitochondrial gene expression between slender and stumpy bloodforms of Trypanosoma brucei. *Mol Biochem Parasitol,* 20**,** 207-14.

FENNELL, C., BABBITT, S., RUSSO, I., WILKES, J., RANFORD-CARTWRIGHT, L., GOLDBERG, D. E. & DOERIG, C. (2009) PfeIK1, a eukaryotic initiation factor 2alpha kinase of the human malaria parasite Plasmodium falciparum, regulates stress-response to amino-acid starvation. *Malar J,* 8**,** 99.

FEVRE, E. M., WISSMANN, B. V., WELBURN, S. C. & LUTUMBA, P. (2008) The burden of human African trypanosomiasis. *PLoS Negl Trop Dis,* 2**,** e333.

FIGUEIREDO, L. M., JANZEN, C. J. & CROSS, G. A. (2008) A histone methyltransferase modulates antigenic variation in African trypanosomes. *PLoS Biol,* 6**,** e161.

FINN, R. D., MISTRY, J., TATE, J., COGGILL, P., HEGER, A., POLLINGTON, J. E., GAVIN, O. L., GUNASEKARAN, P., CERIC, G., FORSLUND, K., HOLM, L., SONNHAMMER, E. L., EDDY, S. R. & BATEMAN, A. (2010) The Pfam protein families database. *Nucleic Acids Res,* 38**,** D211-22.

FLORENT, I. C., RAIBAUD, A. & EISEN, H. (1991) A family of genes related to a new expression site-associated gene in Trypanosoma equiperdum. *Mol Cell Biol,* 11**,** 2180-8.

FREIRE, E. R., DHALIA, R., MOURA, D. M., DA COSTA LIMA, T. D., LIMA, R. P., REIS, C. R., HUGHES, K., FIGUEIREDO, R. C., STANDART, N., CARRINGTON, M. & DE MELO NETO, O. P. (2011) The four trypanosomatid eIF4E homologues fall into two separate groups, with distinct features in primary sequence and biological properties. *Mol Biochem Parasitol,* 176**,** 25-36.

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**,** 4372-80.

GABA, A., JACOBSON, A. & SACHS, M. S. (2005) Ribosome occupancy of the yeast CPA1 upstream open reading frame termination codon modulates nonsense-mediated mRNA decay. *Mol Cell,* 20**,** 449-60.

GANDIN, V., MILUZIO, A., BARBIERI, A. M., BEUGNET, A., KIYOKAWA, H., MARCHISIO, P. C. & BIFFO, S. (2008) Eukaryotic initiation factor 6 is rate-limiting in translation, growth and transformation. *Nature,* 455**,** 684-8.

GARCIA-SALCEDO, J. A., GIJON, P. & PAYS, E. (1999) Regulated transcription of the histone H2B genes of Trypanosoma brucei. *Eur J Biochem,* 264**,** 717-23.

GIFFIN, B. F. & MCCANN, P. P. (1989) Physiological activation of the mitochondrion and the transformation capacity of DFMO induced intermediate and short stumpy bloodstream form trypanosomes. *Am J Trop Med Hyg,* 40**,** 487-93.

GILINGER, G. & BELLOFATTO, V. (2001) Trypanosome spliced leader RNA genes contain the first identified RNA polymerase II gene promoter in these organisms. *Nucleic Acids Res,* 29**,** 1556-64.

GINGRAS, A. C., RAUGHT, B. & SONENBERG, N. (1999) eIF4 initiation factors: effectors of mRNA recruitment to ribosomes and regulators of translation. *Annu Rev Biochem,* 68**,** 913-63.
GRAY, M. W. (1992) The endosymbiont hypothesis revisited. *Int Rev Cytol,* 141**,** 233-357.
GROFT, C. M., BECKMANN, R., SALI, A. & BURLEY, S. K. (2000) Crystal structures of ribosome anti-association factor IF6. *Nat Struct Biol,* 7**,** 1156-64.
GUNZL, A., BRUDERER, T., LAUFER, G., SCHIMANSKI, B., TU, L. C., CHUNG, H. M., LEE, P. T. & LEE, M. G. (2003) RNA polymerase I transcribes procyclin genes and variant surface glycoprotein gene expression sites in Trypanosoma brucei. *Eukaryot Cell,* 2**,** 542-51.
HAGHIGHAT, A., MADER, S., PAUSE, A. & SONENBERG, N. (1995) Repression of cap-dependent translation by 4E-binding protein 1: competition with p220 for binding to eukaryotic initiation factor-4E. *EMBO J,* 14**,** 5701-9.
HARDING, H. P., ZHANG, Y. & RON, D. (1999) Protein translation and folding are coupled by an endoplasmic-reticulum-resident kinase. *Nature,* 397**,** 271-4.
HELM, J. R., WILSON, M. E. & DONELSON, J. E. (2008) Different trans RNA splicing events in bloodstream and procyclic Trypanosoma brucei. *Mol Biochem Parasitol,* 159**,** 134-7.
HENDRIKS, E. F., ROBINSON, D. R., HINKINS, M. & MATTHEWS, K. R. (2001) A novel CCCH protein which modulates differentiation of Trypanosoma brucei to its procyclic form. *EMBO J,* 20**,** 6700-11.
HERSHEY, J. W. (1991) Translational control in mammalian cells. *Annu Rev Biochem,* 60**,** 717-55.
HERTZ-FOWLER, C., FIGUEIREDO, L. M., QUAIL, M. A., BECKER, M., JACKSON, A., BASON, N., BROOKS, K., CHURCHER, C., FAHKRO, S., GOODHEAD, I., HEATH, P., KARTVELISHVILI, M., MUNGALL, K., HARRIS, D., HAUSER, H., SANDERS, M., SAUNDERS, D., SEEGER, K., SHARP, S., TAYLOR, J. E., WALKER, D., WHITE, B., YOUNG, R., CROSS, G. A., RUDENKO, G., BARRY, J. D., LOUIS, E. J. & BERRIMAN, M. (2008) Telomeric expression sites are highly conserved in Trypanosoma brucei. *PLoS One,* 3**,** e3527.
HINNEBUSCH, A. G. (1997) Translational regulation of yeast GCN4. A window on factors that control initiator-trna binding to the ribosome. *J Biol Chem,* 272**,** 21661-4.
HIRUMI, H. & HIRUMI, K. (1989) Continuous cultivation of Trypanosoma brucei blood stream forms in a medium containing a low concentration of serum protein without feeder cell layers. *J Parasitol,* 75**,** 985-9.
HUANG, J. & VAN DER PLOEG, L. H. (1991) Requirement of a polypyrimidine tract for trans-splicing in trypanosomes: discriminating the PARP promoter from the immediately adjacent 3' splice acceptor site. *EMBO J,* 10**,** 3877-85.
IIOKA, H., LOISELLE, D., HAYSTEAD, T. A. & MACARA, I. G. (2011) Efficient detection of RNA-protein interactions using tethered RNAs. *Nucleic Acids Res,* 39**,** e53.
INOUE, M., NAKAMURA, Y., YASUDA, K., YASAKA, N., HARA, T., SCHNAUFER, A., STUART, K. & FUKUMA, T. (2005) The 14-3-3

proteins of Trypanosoma brucei function in motility, cytokinesis, and cell cycle. *J Biol Chem,* 280**,** 14085-96.

IVENS, A. C., PEACOCK, C. S., WORTHEY, E. A., MURPHY, L., AGGARWAL, G., BERRIMAN, M., SISK, E., RAJANDREAM, M. A., ADLEM, E., AERT, R., ANUPAMA, A., APOSTOLOU, Z., ATTIPOE, P., BASON, N., BAUSER, C., BECK, A., BEVERLEY, S. M., BIANCHETTIN, G., BORZYM, K., BOTHE, G., BRUSCHI, C. V., COLLINS, M., CADAG, E., CIARLONI, L., CLAYTON, C., COULSON, R. M., CRONIN, A., CRUZ, A. K., DAVIES, R. M., DE GAUDENZI, J., DOBSON, D. E., DUESTERHOEFT, A., FAZELINA, G., FOSKER, N., FRASCH, A. C., FRASER, A., FUCHS, M., GABEL, C., GOBLE, A., GOFFEAU, A., HARRIS, D., HERTZ-FOWLER, C., HILBERT, H., HORN, D., HUANG, Y., KLAGES, S., KNIGHTS, A., KUBE, M., LARKE, N., LITVIN, L., LORD, A., LOUIE, T., MARRA, M., MASUY, D., MATTHEWS, K., MICHAELI, S., MOTTRAM, J. C., MULLER-AUER, S., MUNDEN, H., NELSON, S., NORBERTCZAK, H., OLIVER, K., O'NEIL, S., PENTONY, M., POHL, T. M., PRICE, C., PURNELLE, B., QUAIL, M. A., RABBINOWITSCH, E., REINHARDT, R., RIEGER, M., RINTA, J., ROBBEN, J., ROBERTSON, L., RUIZ, J. C., RUTTER, S., SAUNDERS, D., SCHAFER, M., SCHEIN, J., SCHWARTZ, D. C., SEEGER, K., SEYLER, A., SHARP, S., SHIN, H., SIVAM, D., SQUARES, R., SQUARES, S., TOSATO, V., VOGT, C., VOLCKAERT, G., WAMBUTT, R., WARREN, T., WEDLER, H., WOODWARD, J., ZHOU, S., ZIMMERMANN, W., SMITH, D. F., BLACKWELL, J. M., STUART, K. D., BARRELL, B., et al. (2005) The genome of the kinetoplastid parasite, Leishmania major. *Science,* 309**,** 436-42.

JAEGER, L. H. & BRANDAO, A. (2011) The composition of upstream open reading frames (uORF) in four genes from Trypanosoma cruzi typical strains. *Parasitol Res,* 109**,** 1205-8.

JANG, S. K., KRAUSSLICH, H. G., NICKLIN, M. J., DUKE, G. M., PALMENBERG, A. C. & WIMMER, E. (1988) A segment of the 5' nontranslated region of encephalomyocarditis virus RNA directs internal entry of ribosomes during in vitro translation. *J Virol,* 62**,** 2636-43.

JANZEN, C. J., VAN DEURSEN, F., SHI, H., CROSS, G. A., MATTHEWS, K. R. & ULLU, E. (2006) Expression site silencing and life-cycle progression appear normal in Argonaute1-deficient Trypanosoma brucei. *Mol Biochem Parasitol,* 149**,** 102-7.

JAO, D. L. & CHEN, K. Y. (2006) Tandem affinity purification revealed the hypusine-dependent binding of eukaryotic initiation factor 5A to the translating 80S ribosomal complex. *J Cell Biochem,* 97**,** 583-98.

JENNI, L., MARTI, S., SCHWEIZER, J., BETSCHART, B., LE PAGE, R. W., WELLS, J. M., TAIT, A., PAINDAVOINE, P., PAYS, E. & STEINERT, M. (1986) Hybrid formation between African trypanosomes during cyclical transmission. *Nature,* 322**,** 173-5.

JENSEN, B. C., SIVAM, D., KIFER, C. T., MYLER, P. J. & PARSONS, M. (2009) Widespread variation in transcript abundance within and across developmental stages of Trypanosoma brucei. *BMC Genomics,* 10**,** 482.

JI, Y., SHAH, S., SOANES, K., ISLAM, M. N., HOXTER, B., BIFFO, S., HESLIP, T. & BYERS, S. (2008) Eukaryotic initiation factor 6 selectively regulates Wnt signaling and beta-catenin protein synthesis. *Oncogene,* 27**,** 755-62.
JINEK, M. & DOUDNA, J. A. (2009) A three-dimensional view of the molecular machinery of RNA interference. *Nature,* 457**,** 405-12.
JOSHI, B., CAMERON, A. & JAGUS, R. (2004) Characterization of mammalian eIF4E-family members. *Eur J Biochem,* 271**,** 2189-203.
KABANI, S., FENN, K., ROSS, A., IVENS, A., SMITH, T. K., GHAZAL, P. & MATTHEWS, K. (2009) Genome-wide expression profiling of in vivo-derived bloodstream parasite stages and dynamic analysis of mRNA alterations during synchronous differentiation in Trypanosoma brucei. *BMC Genomics,* 10**,** 427.
KIEFER, F., ARNOLD, K., KUNZLI, M., BORDOLI, L. & SCHWEDE, T. (2009) The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res,* 37**,** D387-92.
KIMMEL, B. E., OLE-MOIYOI, O. K. & YOUNG, J. R. (1987) Ingi, a 5.2-kb dispersed sequence element from Trypanosoma brucei that carries half of a smaller mobile element at either end and has homology with mammalian LINEs. *Mol Cell Biol,* 7**,** 1465-75.
KLINGE, S., VOIGTS-HOFFMANN, F., LEIBUNDGUT, M., ARPAGAUS, S. & BAN, N. (2011) Crystal structure of the eukaryotic 60S ribosomal subunit in complex with initiation factor 6. *Science,* 334**,** 941-8.
KOLEV, N. G., FRANKLIN, J. B., CARMI, S., SHI, H., MICHAELI, S. & TSCHUDI, C. (2010) The transcriptome of the human pathogen Trypanosoma brucei at single-nucleotide resolution. *PLoS Pathog,* 6**,** e1001090.
KOSTURA, M. & MATHEWS, M. B. (1989) Purification and activation of the double-stranded RNA-dependent eIF-2 kinase DAI. *Mol Cell Biol,* 9**,** 1576-86.
KOZAK, M. (1986) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell,* 44**,** 283-92.
KOZAK, M. (1997) Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6. *EMBO J,* 16**,** 2482-92.
KOZAK, M. (2002) Pushing the limits of the scanning mechanism for initiation of translation. *Gene,* 299**,** 1-34.
KRAMER, S., QUEIROZ, R., ELLIS, L., HOHEISEL, J. D., CLAYTON, C. & CARRINGTON, M. (2010) The RNA helicase DHH1 is central to the correct expression of many developmentally regulated mRNAs in trypanosomes. *J Cell Sci,* 123**,** 699-711.
KRAMER, S., QUEIROZ, R., ELLIS, L., WEBB, H., HOHEISEL, J. D., CLAYTON, C. & CARRINGTON, M. (2008) Heat shock causes a decrease in polysomes and the appearance of stress granules in trypanosomes independently of eIF2(alpha) phosphorylation at Thr169. *J Cell Sci,* 121**,** 3002-14.
KROGH, A., LARSSON, B., VON HEIJNE, G. & SONNHAMMER, E. L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol,* 305**,** 567-80.

LAHAV, T., SIVAM, D., VOLPIN, H., RONEN, M., TSIGANKOV, P., GREEN, A., HOLLAND, N., KUZYK, M., BORCHERS, C., ZILBERSTEIN, D. & MYLER, P. J. (2011) Multiple levels of gene regulation mediate differentiation of the intracellular pathogen Leishmania. *FASEB J,* 25**,** 515-25.

LANGMEAD, B., TRAPNELL, C., POP, M. & SALZBERG, S. L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol,* 10**,** R25.

LANHAM, S. M. & GODFREY, D. G. (1970) Isolation of salivarian trypanosomes from man and other mammals using DEAE-cellulose. *Exp Parasitol,* 28**,** 521-34.

LARKIN, M. A., BLACKSHIELDS, G., BROWN, N. P., CHENNA, R., MCGETTIGAN, P. A., MCWILLIAM, H., VALENTIN, F., WALLACE, I. M., WILM, A., LOPEZ, R., THOMPSON, J. D., GIBSON, T. J. & HIGGINS, D. G. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics,* 23**,** 2947-8.

LAXMAN, S., RIECHERS, A., SADILEK, M., SCHWEDE, F. & BEAVO, J. A. (2006) Hydrolysis products of cAMP analogs cause transformation of Trypanosoma brucei from slender to stumpy-like forms. *Proc Natl Acad Sci U S A,* 103**,** 19194-9.

LEE, J. H., PESTOVA, T. V., SHIN, B. S., CAO, C., CHOI, S. K. & DEVER, T. E. (2002) Initiation factor eIF5B catalyzes second GTP-dependent step in eukaryotic translation initiation. *Proc Natl Acad Sci U S A,* 99**,** 16689-94.

LEE, K. H., MINAMI, A., MARSHALL, R. S., BOOK, A. J., FARMER, L. M., WALKER, J. M. & VIERSTRA, R. D. (2011) The RPT2 subunit of the 26S proteasome directs complex assembly, histone dynamics, and gametophyte and sporophyte development in Arabidopsis. *Plant Cell*, 23, 4298-317

LEE, Y. Y., CEVALLOS, R. C. & JAN, E. (2009) An upstream open reading frame regulates translation of GADD34 during cellular stresses that induce eIF2alpha phosphorylation. *J Biol Chem,* 284**,** 6661-73.

LEFEBVRE, A. K., KORNEEVA, N. L., TRUTSCHL, M., CVEK, U., DUZAN, R. D., BRADLEY, C. A., HERSHEY, J. W. & RHOADS, R. E. (2006) Translation initiation factor eIF4G-1 binds to eIF3 through the eIF3e subunit. *J Biol Chem,* 281**,** 22917-32.

LEUNG, E. K., SUSLOV, N., TUTTLE, N., SENGUPTA, R. & PICCIRILLI, J. A. (2011) The mechanism of peptidyl transfer catalysis by the ribosome. *Annu Rev Biochem,* 80**,** 527-55.

LI, Z. & WANG, C. C. (2003) A PHO80-like cyclin and a B-type cyclin control the cell cycle of the procyclic form of Trypanosoma brucei. *J Biol Chem,* 278**,** 20652-8.

LIU, W., APAGYI, K., MCLEAVY, L. & ERSFELD, K. (2010) Expression and cellular localisation of calpain-like proteins in Trypanosoma brucei. *Mol Biochem Parasitol,* 169**,** 20-6.

MACGREGOR, P. & MATTHEWS, K. R. (2010) New discoveries in the transmission biology of sleeping sickness parasites: applying the basics. *J Mol Med*.

MAIR, G., SHI, H., LI, H., DJIKENG, A., AVILES, H. O., BISHOP, J. R., FALCONE, F. H., GAVRILESCU, C., MONTGOMERY, J. L., SANTORI,

M. I., STERN, L. S., WANG, Z., ULLU, E. & TSCHUDI, C. (2000) A new twist in trypanosome RNA metabolism: cis-splicing of pre-mRNA. *RNA,* 6**,** 163-9.
MANFUL, T., FADDA, A. & CLAYTON, C. (2011) The role of the 5'-3' exoribonuclease XRNA in transcriptome-wide mRNA degradation. *RNA,* 17**,** 2039-47.
MARCOTRIGIANO, J., GINGRAS, A. C., SONENBERG, N. & BURLEY, S. K. (1997) Cocrystal structure of the messenger RNA 5' cap-binding protein (eIF4E) bound to 7-methyl-GDP. *Cell,* 89**,** 951-61.
MARCOTRIGIANO, J., GINGRAS, A. C., SONENBERG, N. & BURLEY, S. K. (1999) Cap-dependent translation initiation in eukaryotes is regulated by a molecular mimic of eIF4G. *Mol Cell,* 3**,** 707-16.
MARSDEN, S., NARDELLI, M., LINDER, P. & MCCARTHY, J. E. (2006) Unwinding single RNA molecules using helicases involved in eukaryotic translation initiation. *J Mol Biol,* 361**,** 327-35.
MARTIN, D., BRUN, C., REMY, E., MOUREN, P., THIEFFRY, D. & JACQ, B. (2004) GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biol,* 5**,** R101.
MATTHEWS, K. R. (2005) The developmental cell biology of Trypanosoma brucei. *J Cell Sci,* 118**,** 283-90.
MATTHEWS, K. R., ELLIS, J. R. & PATEROU, A. (2004) Molecular regulation of the life cycle of African trypanosomes. *Trends Parasitol,* 20**,** 40-7.
MATTHEWS, K. R. & GULL, K. (1998) Identification of stage-regulated and differentiation-enriched transcripts during transformation of the African trypanosome from its bloodstream to procyclic form. *Mol Biochem Parasitol,* 95**,** 81-95.
MATTHEWS, K. R., TSCHUDI, C. & ULLU, E. (1994) A common pyrimidine-rich motif governs trans-splicing and polyadenylation of tubulin polycistronic pre-mRNA in trypanosomes. *Genes Dev,* 8**,** 491-501.
MAYHO, M., FENN, K., CRADDY, P., CROSTHWAITE, S. & MATTHEWS, K. (2006) Post-transcriptional control of nuclear-encoded cytochrome oxidase subunits in Trypanosoma brucei: evidence for genome-wide conservation of life-cycle stage-specific regulatory elements. *Nucleic Acids Res,* 34**,** 5312-24.
MCCULLOCH, R. (2004) Antigenic variation in African trypanosomes: monitoring progress. *Trends Parasitol,* 20**,** 117-21.
MEIJER, H. A. & THOMAS, A. A. (2002) Control of eukaryotic protein synthesis by upstream open reading frames in the 5'-untranslated region of an mRNA. *Biochem J,* 367**,** 1-11.
MENNE, T. F., GOYENECHEA, B., SANCHEZ-PUIG, N., WONG, C. C., TONKIN, L. M., ANCLIFF, P. J., BROST, R. L., COSTANZO, M., BOONE, C. & WARREN, A. J. (2007) The Shwachman-Bodian-Diamond syndrome protein mediates translational activation of ribosomes in yeast. *Nat Genet,* 39**,** 486-95.
MERRICK, W. (2009) Translation: Till termination us do part. *Nature,* 459**,** 44-5.
MICHELOTTI, E. F. & HAJDUK, S. L. (1987) Developmental regulation of trypanosome mitochondrial gene expression. *J Biol Chem,* 262**,** 927-32.

MILUZIO, A., BEUGNET, A., VOLTA, V. & BIFFO, S. (2009) Eukaryotic initiation factor 6 mediates a continuum between 60S ribosome biogenesis and translation. *EMBO Rep,* 10**,** 459-65.

MORAES, M. C., JESUS, T. C., HASHIMOTO, N. N., DEY, M., SCHWARTZ, K. J., ALVES, V. S., AVILA, C. C., BANGS, J. D., DEVER, T. E., SCHENKMAN, S. & CASTILHO, B. A. (2007) Novel membrane-bound eIF2alpha kinase in the flagellar pocket of Trypanosoma brucei. *Eukaryot Cell,* 6**,** 1979-91.

MOREIRA-LEITE, F. F., SHERWIN, T., KOHL, L. & GULL, K. (2001) A trypanosome structure involved in transmitting cytoplasmic information during cell division. *Science,* 294**,** 610-2.

MORINO, S., HAZAMA, H., OZAKI, M., TERAOKA, Y., SHIBATA, S., DOI, M., UEDA, H., ISHIDA, T. & UESUGI, S. (1996) Analysis of the mRNA cap-binding ability of human eukaryotic initiation factor-4E by use of recombinant wild-type and mutant forms. *Eur J Biochem,* 239**,** 597-601.

MORO, A., RUIZ-CABELLO, F., FERNANDEZ-CANO, A., STOCK, R. P. & GONZALEZ, A. (1995) Secretion by Trypanosoma cruzi of a peptidyl-prolyl cis-trans isomerase involved in cell infection. *EMBO J,* 14**,** 2483-90.

MULLER, I. B., DOMENICALI-PFISTER, D., RODITI, I. & VASSELLA, E. (2002) Stage-specific requirement of a mitogen-activated protein kinase by Trypanosoma brucei. *Mol Biol Cell,* 13**,** 3787-99.

MULLER, K., MATUSCHEWSKI, K. & SILVIE, O. (2011) The Puf-family RNA-binding protein Puf2 controls sporozoite conversion to liver stages in the malaria parasite. *PLoS One,* 6**,** e19860.

MURPHEY, R. J. & GERNER, E. W. (1987) Hypusine formation in protein by a two-step process in cell lysates. *J Biol Chem,* 262**,** 15033-6.

NAKAI, K. & HORTON, P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci,* 24**,** 34-6.

NAVARRO, M. & GULL, K. (2001) A pol I transcriptional body associated with VSG mono-allelic expression in Trypanosoma brucei. *Nature,* 414**,** 759-63.

NGO, H., TSCHUDI, C., GULL, K. & ULLU, E. (1998) Double-stranded RNA induces mRNA degradation in Trypanosoma brucei. *Proc Natl Acad Sci U S A,* 95**,** 14687-92.

NILSSON, D., GUNASEKERA, K., MANI, J., OSTERAS, M., FARINELLI, L., BAERLOCHER, L., RODITI, I. & OCHSENREITER, T. (2010) Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of Trypanosoma brucei. *PLoS Pathog,* 6**,** e1001037.

NJUGUNA, J. T., NASSAR, M., HOERAUF, A. & KAISER, A. E. (2006) Cloning, expression and functional activity of deoxyhypusine synthase from Plasmodium vivax. *BMC Microbiol,* 6**,** 91.

OVERATH, P. & ENGSTLER, M. (2004) Endocytosis, membrane recycling and sorting of GPI-anchored proteins: Trypanosoma brucei as a model system. *Mol Microbiol,* 53**,** 735-44.

PANDEY, K. N. (2010) Small peptide recognition sequence for intracellular sorting. *Curr Opin Biotechnol,* 21**,** 611-20.

PARK, J. H., ARAVIND, L., WOLFF, E. C., KAEVEL, J., KIM, Y. S. & PARK, M. H. (2006) Molecular cloning, expression, and structural prediction of

deoxyhypusine hydroxylase: a HEAT-repeat-containing metalloenzyme. *Proc Natl Acad Sci U S A,* 103**,** 51-6.
PARK, M. H., COOPER, H. L. & FOLK, J. E. (1982) The biosynthesis of protein-bound hypusine (N epsilon -(4-amino-2-hydroxybutyl)lysine). Lysine as the amino acid precursor and the intermediate role of deoxyhypusine (N epsilon -(4-aminobutyl)lysine). *J Biol Chem,* 257**,** 7217-22.
PARKER, R. & SONG, H. (2004) The enzymes and control of eukaryotic mRNA turnover. *Nat Struct Mol Biol,* 11**,** 121-7.
PATEROU, A., WALRAD, P., CRADDY, P., FENN, K. & MATTHEWS, K. (2006) Identification and stage-specific association with the translational apparatus of TbZFP3, a CCCH protein that promotes trypanosome life-cycle development. *J Biol Chem,* 281**,** 39002-13.
PATRICK, K. L., LUZ, P. M., RUAN, J. P., SHI, H., ULLU, E. & TSCHUDI, C. (2008) Genomic rearrangements and transcriptional analysis of the spliced leader-associated retrotransposon in RNA interference-deficient Trypanosoma brucei. *Mol Microbiol,* 67**,** 435-47.
PATRICK, K. L., SHI, H., KOLEV, N. G., ERSFELD, K., TSCHUDI, C. & ULLU, E. (2009) Distinct and overlapping roles for two Dicer-like proteins in the RNA interference pathways of the ancient eukaryote Trypanosoma brucei. *Proc Natl Acad Sci U S A*.
PAUSE, A., BELSHAM, G. J., GINGRAS, A. C., DONZE, O., LIN, T. A., LAWRENCE, J. C., JR. & SONENBERG, N. (1994) Insulin-dependent stimulation of protein synthesis by phosphorylation of a regulator of 5'-cap function. *Nature,* 371**,** 762-7.
PAYS, E., HANOCQ-QUERTIER, J., HANOCQ, F., VAN ASSEL, S., NOLAN, D. & ROLIN, S. (1993) Abrupt RNA changes precede the first cell division during the differentiation of Trypanosoma brucei bloodstream forms into procyclic forms in vitro. *Mol Biochem Parasitol,* 61**,** 107-14.
PEACOCK, L., FERRIS, V., SHARMA, R., SUNTER, J., BAILEY, M., CARRINGTON, M. & GIBSON, W. (2011) Identification of the meiotic life cycle stage of Trypanosoma brucei in the tsetse fly. *Proc Natl Acad Sci U S A,* 108**,** 3671-6.
PELLE, R., MCODIMBA, F., CHUMA, F., WASAWO, D., PEARSON, T. W. & MURPHY, N. B. (2002) The African trypanosome cyclophilin A homologue contains unusual conserved central and N-terminal domains and is developmentally regulated. *Gene,* 290**,** 181-91.
PESTOVA, T. V., BORUKHOV, S. I. & HELLEN, C. U. (1998) Eukaryotic ribosomes require initiation factors 1 and 1A to locate initiation codons. *Nature,* 394**,** 854-9.
PESTOVA, T. V., LOMAKIN, I. B., LEE, J. H., CHOI, S. K., DEVER, T. E. & HELLEN, C. U. (2000) The joining of ribosomal subunits in eukaryotes requires eIF5B. *Nature,* 403**,** 332-5.
PRIEST, J. W. & HAJDUK, S. L. (1994) Developmental regulation of mitochondrial biogenesis in Trypanosoma brucei. *J Bioenerg Biomembr,* 26**,** 179-91.
PTUSHKINA, M., VON DER HAAR, T., VASILESCU, S., FRANK, R., BIRKENHAGER, R. & MCCARTHY, J. E. (1998) Cooperative modulation by eIF4G of eIF4E-binding to the mRNA 5' cap in yeast involves a site partially shared by p20. *EMBO J,* 17**,** 4798-808.

PYRONNET, S., IMATAKA, H., GINGRAS, A. C., FUKUNAGA, R., HUNTER, T. & SONENBERG, N. (1999) Human eukaryotic translation initiation factor 4G (eIF4G) recruits mnk1 to phosphorylate eIF4E. *EMBO J,* 18**,** 270-9.
QUEIROZ, R., BENZ, C., FELLENBERG, K., HOHEISEL, J. D. & CLAYTON, C. (2009) Transcriptome analysis of differentiating trypanosomes reveals the existence of multiple post-transcriptional regulons. *BMC Genomics,* 10**,** 495.
RAJKOWITSCH, L., VILELA, C., BERTHELOT, K., RAMIREZ, C. V. & MCCARTHY, J. E. (2004) Reinitiation and recycling are distinct processes occurring downstream of translation termination in yeast. *J Mol Biol,* 335**,** 71-85.
RAUGHT, B. & GINGRAS, A. C. (1999) eIF4E activity is regulated at multiple levels. *Int J Biochem Cell Biol,* 31**,** 43-57.
RETTIG, J., WANG, Y., SCHNEIDER, A. & OCHSENREITER, T. (2011) Dual targeting of isoleucyl-tRNA synthetase in Trypanosoma brucei is mediated through alternative trans-splicing. *Nucleic Acids Res,* 40**,** 1299-306.
ROBLES, A. & CLAYTON, C. (2008) Regulation of an amino acid transporter mRNA in Trypanosoma brucei. *Mol Biochem Parasitol,* 157**,** 102-6.
RODITI, I., SCHWARZ, H., PEARSON, T. W., BEECROFT, R. P., LIU, M. K., RICHARDSON, J. P., BUHRING, H. J., PLEISS, J., BULOW, R., WILLIAMS, R. O. & ET AL. (1989) Procyclin gene expression and loss of the variant surface glycoprotein during differentiation of Trypanosoma brucei. *J Cell Biol,* 108**,** 737-46.
ROLIN, S., PAINDAVOINE, P., HANOCQ-QUERTIER, J., HANOCQ, F., CLAES, Y., LE RAY, D., OVERATH, P. & PAYS, E. (1993) Transient adenylate cyclase activation accompanies differentiation of Trypanosoma brucei from bloodstream to procyclic forms. *Mol Biochem Parasitol,* 61**,** 115-25.
ROM, E., KIM, H. C., GINGRAS, A. C., MARCOTRIGIANO, J., FAVRE, D., OLSEN, H., BURLEY, S. K. & SONENBERG, N. (1998) Cloning and characterization of 4EHP, a novel mammalian eIF4E-related cap-binding protein. *J Biol Chem,* 273**,** 13104-9.
ROWLANDS, A. G., PANNIERS, R. & HENSHAW, E. C. (1988) The catalytic mechanism of guanine nucleotide exchange factor action and competitive inhibition by phosphorylated eukaryotic initiation factor 2. *J Biol Chem,* 263**,** 5526-33.
RUSSELL, D. W. & SPREMULLI, L. L. (1980) Mechanism of action of the wheat germ ribosome dissociation factor: interaction with the 60 S subunit. *Arch Biochem Biophys,* 201**,** 518-26.
RYLEY, J. F. (1962) Studies on the metabolism of the protozoa. 9. Comparative metabolism of blood-stream and culture forms of Trypanosoma rhodesiense. *Biochem J,* 85**,** 211-23.
SAAS, J., ZIEGELBAUER, K., VON HAESELER, A., FAST, B. & BOSHART, M. (2000) A developmentally regulated aconitase related to iron-regulatory protein-1 is localized in the cytoplasm and in the mitochondrion of Trypanosoma brucei. *J Biol Chem,* 275**,** 2745-55.
SAINI, P., EYLER, D. E., GREEN, R. & DEVER, T. E. (2009) Hypusine-containing protein eIF5A promotes translation elongation. *Nature,* 459**,** 118-21.

SCHUMANN BURKARD, G., JUTZI, P. & RODITI, I. (2011) Genome-wide RNAi screens in bloodstream form trypanosomes identify drug transporters. *Mol Biochem Parasitol,* 175**,** 91-4.

SCHWEDE, A., ELLIS, L., LUTHER, J., CARRINGTON, M., STOECKLIN, G. & CLAYTON, C. (2008) A role for Caf1 in mRNA deadenylation and decay in trypanosomes and human cells. *Nucleic Acids Res,* 36**,** 3374-88.

SCORY, S., STIERHOF, Y. D., CAFFREY, C. R. & STEVERDING, D. (2007) The cysteine proteinase inhibitor Z-Phe-Ala-CHN2 alters cell morphology and cell division activity of Trypanosoma brucei bloodstream forms in vivo. *Kinetoplastid Biol Dis,* 6**,** 2.

SENGER, B., LAFONTAINE, D. L., GRAINDORGE, J. S., GADAL, O., CAMASSES, A., SANNI, A., GARNIER, J. M., BREITENBACH, M., HURT, E. & FASIOLO, F. (2001) The nucle(ol)ar Tif6p and Efl1p are required for a late cytoplasmic step of ribosome synthesis. *Mol Cell,* 8**,** 1363-73.

SENGUPTA, D. J., ZHANG, B., KRAEMER, B., POCHART, P., FIELDS, S. & WICKENS, M. (1996) A three-hybrid system to detect RNA-protein interactions in vivo. *Proc Natl Acad Sci U S A,* 93**,** 8496-501.

SHAPIRO, S. Z. & KIMMEL, B. E. (1987) Differential protein synthesis during the life cycle of the protozoan parasite Trypanosoma brucei. *J Protozool,* 34**,** 58-62.

SHAPIRO, S. Z., NAESSENS, J., LIESEGANG, B., MOLOO, S. K. & MAGONDU, J. (1984) Analysis by flow cytometry of DNA synthesis during the life cycle of African trypanosomes. *Acta Trop,* 41**,** 313-23.

SHERWIN, T. & GULL, K. (1989) The cell division cycle of Trypanosoma brucei brucei: timing of event markers and cytoskeletal modulations. *Philos Trans R Soc Lond B Biol Sci,* 323**,** 573-88.

SHI, H., DJIKENG, A., TSCHUDI, C. & ULLU, E. (2004) Argonaute protein in the early divergent eukaryote Trypanosoma brucei: control of small interfering RNA accumulation and retroposon transcript abundance. *Mol Cell Biol,* 24**,** 420-7.

SI, K., CHAUDHURI, J., CHEVESICH, J. & MAITRA, U. (1997) Molecular cloning and functional expression of a human cDNA encoding translation initiation factor 6. *Proc Natl Acad Sci U S A,* 94**,** 14285-90.

SI, K. & MAITRA, U. (1999) The Saccharomyces cerevisiae homologue of mammalian translation initiation factor 6 does not function as a translation initiation factor. *Mol Cell Biol,* 19**,** 1416-26.

SIEGEL, T. M., HEKSTRA, D. R., KEMP, L. E., FIGUEIREDO, L. M., LOWELL, J. E., FENYO, D., WANG, X., DEWELL, S. & CROSS, G. A. M. (2009) Four histone varients mark the boundaries of polycistronic transcription units in *Trypanosoma brucei*. *Genes & Development*.

SIEGEL, T. N., GUNASEKERA, K., CROSS, G. A. & OCHSENREITER, T. (2011) Gene expression in Trypanosoma brucei: lessons from high-throughput RNA sequencing. *Trends Parasitol,* 27**,** 434-41.

SIEGEL, T. N., HEKSTRA, D. R., WANG, X., DEWELL, S. & CROSS, G. A. (2010) Genome-wide analysis of mRNA abundance in two life-cycle stages of Trypanosoma brucei and identification of splicing and polyadenylation sites. *Nucleic Acids Res,* 38**,** 4946-57.

SIMARRO, P. P., JANNIN, J. & CATTAND, P. (2008) Eliminating human African trypanosomiasis: where do we stand and what comes next? *PLoS Med,* 5**,** e55.
SLOBODIN, B. & GERST, J. E. (2010) A novel mRNA affinity purification technique for the identification of interacting proteins and transcripts in ribonucleoprotein complexes. *RNA,* 16**,** 2277-90.
SMITH, T. K., VASILEVA, N., GLUENZ, E., TERRY, S., PORTMAN, N., KRAMER, S., CARRINGTON, M., MICHAELI, S., GULL, K. & RUDENKO, G. (2009) Blocking variant surface glycoprotein synthesis in Trypanosoma brucei triggers a general arrest in translation initiation. *PLoS One,* 4**,** e7532.
SONENBERG, N. & HINNEBUSCH, A. G. (2009) Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell,* 136**,** 731-45.
SPECHT, S., SARITE, S. R., HAUBER, I., HAUBER, J., GORBIG, U. F., MEIER, C., BEVEC, D., HOERAUF, A. & KAISER, A. (2008) The guanylhydrazone CNI-1493: an inhibitor with dual activity against malaria-inhibition of host cell pro-inflammatory cytokine release and parasitic deoxyhypusine synthase. *Parasitol Res,* 102**,** 1177-84.
STEBECK, C. E., BEECROFT, R. P., SINGH, B. N., JARDIM, A., OLAFSON, R. W., TUCKEY, C., PRENEVOST, K. D. & PEARSON, T. W. (1995) Kinetoplastid membrane protein-11 (KMP-11) is differentially expressed during the life cycle of African trypanosomes and is found in a wide variety of kinetoplastid parasites. *Mol Biochem Parasitol,* 71**,** 1-13.
SUBOTA, I., ROTUREAU, B., BLISNICK, T., NGWABYT, S., DURAND-DUBIEF, M., ENGSTLER, M. & BASTIN, P. (2011) ALBA proteins are stage regulated during trypanosome development in the tsetse fly and participate in differentiation. *Mol Biol Cell,* 22**,** 4205-19.
SZOOR, B., RUBERTO, I., BURCHMORE, R. & MATTHEWS, K. R. (2010) A novel phosphatase cascade regulates differentiation in Trypanosoma brucei via a glycosomal signaling pathway. *Genes Dev,* 24**,** 1306-16.
SZOOR, B., WILSON, J., MCELHINNEY, H., TABERNERO, L. & MATTHEWS, K. R. (2006) Protein tyrosine phosphatase TbPTP1: A molecular switch controlling life cycle differentiation in trypanosomes. *J Cell Biol,* 175**,** 293-303.
TEIXEIRA, S. M., KIRCHHOFF, L. V. & DONELSON, J. E. (1999) Trypanosoma cruzi: suppression of tuzin gene expression by its 5'-UTR and spliced leader addition site. *Exp Parasitol,* 93**,** 143-51.
THOMAS, M. C., GARCIA-PEREZ, J. L., ALONSO, C. & LOPEZ, M. C. (2000) Molecular characterization of KMP11 from Trypanosoma cruzi: a cytoskeleton-associated protein regulated at the translational level. *DNA Cell Biol,* 19**,** 47-57.
TOLSON, D. L., JARDIM, A., SCHNUR, L. F., STEBECK, C., TUCKEY, C., BEECROFT, R. P., TEH, H. S., OLAFSON, R. W. & PEARSON, T. W. (1994) The kinetoplastid membrane protein 11 of Leishmania donovani and African trypanosomes is a potent stimulator of T-lymphocyte proliferation. *Infect Immun,* 62**,** 4893-9.

TOMKO, R. J. & HOCHSTRASSER, M. (2011) Order of the proteasomal ATPases and eukaryotic proteasome assembly. *Cell Biochem Biophys*, 60, 13-20
TONELLI, R. R., AUGUSTO LDA, S., CASTILHO, B. A. & SCHENKMAN, S. (2011) Protein synthesis attenuation by phosphorylation of eIF2alpha is required for the differentiation of Trypanosoma cruzi into infective forms. *PLoS One,* 6**,** e27904.
TORRI, A. F. & HAJDUK, S. L. (1988) Posttranscriptional regulation of cytochrome c expression during the developmental cycle of Trypanosoma brucei. *Mol Cell Biol,* 8**,** 4625-33.
UTTER, C. J., GARCIA, S. A., MILONE, J. & BELLOFATTO, V. (2011) PolyA-specific ribonuclease (PARN-1) function in stage-specific mRNA turnover in Trypanosoma brucei. *Eukaryot Cell,* 10**,** 1230-40.
VAN DEN ABBEELE, J., CLAES, Y., VAN BOCKSTAELE, D., LE RAY, D. & COOSEMANS, M. (1999) Trypanosoma brucei spp. development in the tsetse fly: characterization of the post-mesocyclic stages in the foregut and proboscis. *Parasitology,* 118 (Pt 5)**,** 469-78.
VAN HELDEN, J., ANDRE, B. & COLLADO-VIDES, J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol,* 281**,** 827-42.
VASSELLA, E., DEN ABBEELE, J. V., BUTIKOFER, P., RENGGLI, C. K., FURGER, A., BRUN, R. & RODITI, I. (2000) A major surface glycoprotein of trypanosoma brucei is expressed transiently during development and can be regulated post-transcriptionally by glycerol or hypoxia. *Genes Dev,* 14**,** 615-26.
VASSELLA, E., KRAMER, R., TURNER, C. M., WANKELL, M., MODES, C., VAN DEN BOGAARD, M. & BOSHART, M. (2001) Deletion of a novel protein kinase with PX and FYVE-related domains increases the rate of differentiation of Trypanosoma brucei. *Mol Microbiol,* 41**,** 33-46.
VASSELLA, E., REUNER, B., YUTZY, B. & BOSHART, M. (1997) Differentiation of African trypanosomes is controlled by a density sensing mechanism which signals cell cycle arrest via the cAMP pathway. *J Cell Sci,* 110 (Pt 21)**,** 2661-71.
VAUGHAN, S. & GULL, K. (2003) The trypanosome flagellum. *J Cell Sci,* 116**,** 757-9.
VEITCH, N. J., JOHNSON, P. C., TRIVEDI, U., TERRY, S., WILDRIDGE, D. & MACLEOD, A. (2010) Digital gene expression analysis of two life cycle stages of the human-infective parasite, Trypanosoma brucei gambiense reveals differentially expressed clusters of co-regulated genes. *BMC Genomics,* 11**,** 124.
VICKERMAN, K. (1985) Developmental cycles and biology of pathogenic trypanosomes. *Br Med Bull,* 41**,** 105-14.
VICKERMAN, K. (1994) The evolutionary expansion of the trypanosomatid flagellates. *Int J Parasitol,* 24**,** 1317-31.
VILELA, C. & MCCARTHY, J. E. (2003) Regulation of fungal gene expression via short open reading frames in the mRNA 5'untranslated region. *Mol Microbiol,* 49**,** 859-67.
VINCENDEAU, P. & BOUTEILLE, B. (2006) Immunology and immunopathology of African trypanosomiasis. *An Acad Bras Cienc,* 78**,** 645-65.

WALRAD, P., PATEROU, A., ACOSTA-SERRANO, A. & MATTHEWS, K. R. (2009) Differential trypanosome surface coat regulation by a CCCH protein that co-associates with procyclin mRNA cis-elements. *PLoS Pathog,* 5**,** e1000317.

WANG, S., CHEN, A. J., SHI, L. J., ZHAO, X. F. & WANG, J. X. (2012) TRBP and eIF6 Homologue in Marsupenaeus japonicus Play Crucial Roles in Antiviral Response. *PLoS One,* 7**,** e30057.

WANG, S., LIU, N., CHEN, A. J., ZHAO, X. F. & WANG, J. X. (2009) TRBP homolog interacts with eukaryotic initiation factor 6 (eIF6) in Fenneropenaeus chinensis. *J Immunol,* 182**,** 5250-8.

WARNER, J. R., KNOPF, P. M. & RICH, A. (1963) A multiple ribosomal structure in protein synthesis. *Proc Natl Acad Sci U S A,* 49**,** 122-9.

WELLS, S. E., HILLNER, P. E., VALE, R. D. & SACHS, A. B. (1998) Circularization of mRNA by eukaryotic translation initiation factors. *Mol Cell,* 2**,** 135-40.

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.

WOODWARD, R. & GULL, K. (1990) Timing of nuclear and kinetoplast DNA replication and early morphological events in the cell cycle of Trypanosoma brucei. *J Cell Sci,* 95 (Pt 1)**,** 49-57.

YOFFE, Y., LEGER, M., ZINOVIEV, A., ZUBEREK, J., DARZYNKIEWICZ, E., WAGNER, G. & SHAPIRA, M. (2009) Evolutionary changes in the Leishmania eIF4F complex involve variations in the eIF4E-eIF4G interactions. *Nucleic Acids Res,* 37**,** 3243-53.

YOFFE, Y., ZUBEREK, J., LERER, A., LEWDOROWICZ, M., STEPINSKI, J., ALTMANN, M., DARZYNKIEWICZ, E. & SHAPIRA, M. (2006) Binding specificities and potential roles of isoforms of eukaryotic initiation factor 4E in Leishmania. *Eukaryot Cell,* 5**,** 1969-79.

ZAMUDIO, J. R., MITTRA, B., CAMPBELL, D. A. & STURM, N. R. (2009) Hypermethylated cap 4 maximizes Trypanosoma brucei translation. *Mol Microbiol,* 72**,** 1100-10.

ZHANG, M., FENNELL, C., RANFORD-CARTWRIGHT, L., SAKTHIVEL, R., GUEIRARD, P., MEISTER, S., CASPI, A., DOERIG, C., NUSSENZWEIG, R. S., TUTEJA, R., SULLIVAN, W. J., JR., ROOS, D. S., FONTOURA, B. M., MENARD, R., WINZELER, E. A. & NUSSENZWEIG, V. (2010) The Plasmodium eukaryotic initiation factor-2alpha kinase IK2 controls the latency of sporozoites in the mosquito salivary glands. *J Exp Med,* 207**,** 1465-74.

ZHU, N., GU, L., FINDLEY, H. W. & ZHOU, M. (2005) Transcriptional repression of the eukaryotic initiation factor 4E gene by wild type p53. *Biochem Biophys Res Commun,* 335**,** 1272-9.

ZINOVIEV, A., LEGER, M., WAGNER, G. & SHAPIRA, M. (2011) A novel 4E-interacting protein in Leishmania is involved in stage-specific translation pathways. *Nucleic Acids Res,* 39**,** 8404-15.

Appendix 1

SOLUTIONS

Chemicals were purchased from Sigma unless otherwise stated.

PSG (Phosphate-Saline-Glucose) buffer

1 litre:

$NaH_2PO_4.2H_2O$	0.488 g
NaCl	2.55 g
Na_2HPO_4	8.08 g
D-glucose	15 g

pH to 7.8 with HCl

Phosphate buffered saline (PBS)

NaCl
KCl
Na_2HPO_4
KH_2PO_4
pH to 7.4

vPBS (Voorheis's modified PBS)

PBS	250 ml
Sucrose	3.925g
Glucose	0.45g

Mowiol

Mowiol 4-88 reagent (Calbiochem)	10% w/v
Glycerol	25% w/v
Tris (pH8.5)	0.1 M

DTT

DL-Dithiothreitol dissolved in 10 mM sodium acetate (pH5.2), then sterilized by passing through a single-use syringe filter (0.20 μm; Sartorius Stedim Biotech).

NORTHERN BLOT ANALYSIS SOLUTIONS

10X MOPS

MOPS	46.26 g/l
Sodium acetate, pH7.0	50 mM
EDTA, pH8.0	10 mM

Autoclaved and stored at 4°C in the dark.

20X SSC

NaCl	3M
Tri sodium citrate	0.3M

RNA gel loading buffer

Formamide	150 µl
Formaldehyde (37%)	83 µl
Bromophenol blue	0.01%
MOPS	1X
Glycerol	10%
dH_2O	167 µl

Hybridisation buffer

20X SSC	12.5 ml
Formamide	25 ml
10% SDS	100 µl
10% DIG block	10 ml

Maleic acid buffer

Maleic acid	100 mM
NaCl	150 mM

pH7.5 (using NaOH), autoclaved.

DIG block

Blocking reagent (Roche)	10% (w/v)

Maleic acid buffer, pH7.5
Stored at -20°C

DNA MANIPULATIONS

Miniprep solution I

Glucose	50 mM
Tris-HCl, pH8	25 mM
EDTA, pH8	10 mM
Autoclaved	

Miniprep solution II

SDS	1%
NaOH	0.2 M

Miniprep solution III

5M Potassium acetate	60 ml
Glacial acetic acid	11.5 ml
ddH_2O	28.5 ml

1X TAE

Tris-acetate	40 mM
EDTA	1 mM
pH8	

6X DNA gel loading buffer

Bromophenol blue	0.25%
Xylene cyanol	0.25%
Glycerol	30%

TE

Tris-HCl, pH8	10 mM
EDTA	1 mM

WESTERN BLOT ANALYSIS SOLUTIONS

Semi-dry transfer buffer

Tris	2.5 mM
Glycine	15 mM
SDS	0.02%
Methanol	20%

SDS-PAGE gels

		10%	12%
Running gel:	Water	3.12 ml	2.62 ml
	TrisHCl buffer (pH8.8)	1.87 ml	1.87 ml
	Acrylamide	2.5 ml	3.0 ml
	10% APS	50 µl	50 µl
	TEMED	10 µl	10 µl

Stacking gel:	Water	1.5 ml
	TrisHCl buffer (pH6.8)	375 µl
	Acrylamide	625 µl
	10% APS	12.5 µl
	TEMED	4 µl

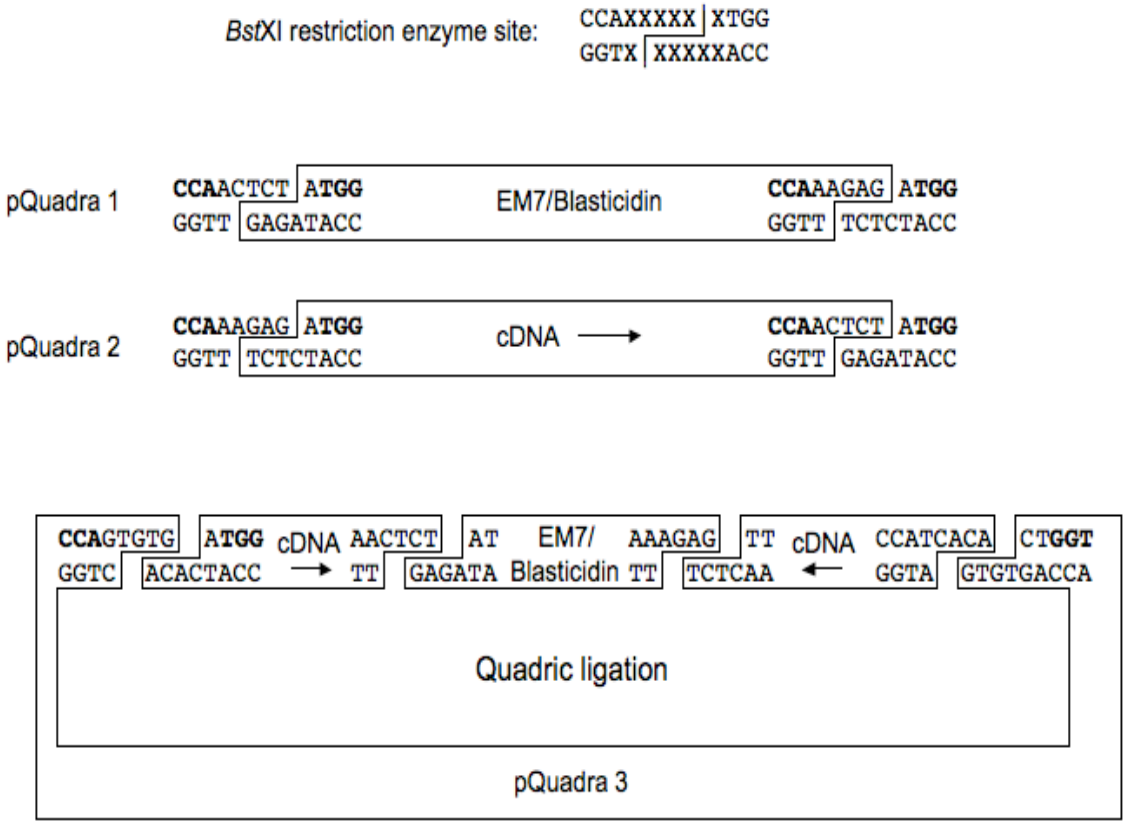
TrisHCl buffer (pH8.8):	TrisHCl	1.5 M
	SDS	0.4%

TrisHCl buffer (pH6.8):	TrisHCl	0.5 M
	SDS	0.4%

Appendix 2

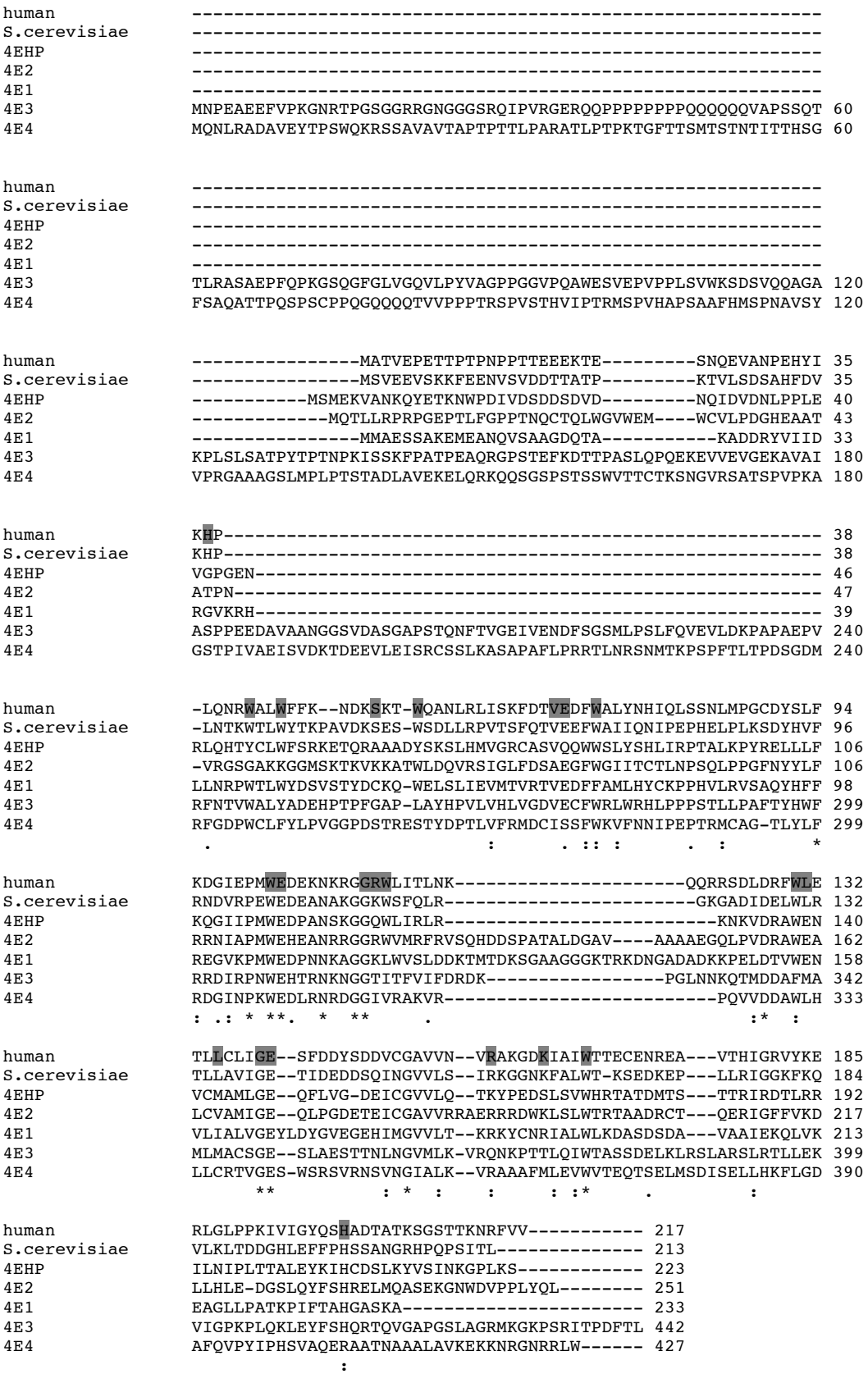
SQL query (written by Christiane Hertz-Fowler, Wellcome Trust Sanger Institute, Hinxton) for retrieving Gene Ontology annotations (Gene IDs entered into the last set of parentheses):

```
SELECT
  gene_product.symbol,
  gene_product.full_name,
  dbxref.xref_dbname,
  dbxref.xref_key,
  species.genus,
  species.species,
  association.is_not,
  evidence.code,
  term.acc,
  term.name
 FROM   gene_product
  INNER JOIN dbxref ON (gene_product.dbxref_id=dbxref.id)
  INNER JOIN species ON (gene_product.species_id=species.id)
  INNER JOIN association ON (gene_product.id=association.gene_product_id)
  INNER JOIN evidence ON (association.id=evidence.association_id)
  INNER JOIN term ON (association.term_id=term.id)
 WHERE
   dbxref.xref_key IN
     ('Tb927.1.700','Tb927.1.710')
 AND
   dbxref.xref_dbname = 'GeneDB_Tbrucei';
```



pQuadra vector quadratic ligation with pQuadra1, pQuadra2, and pQuadra3 vectors, using the *Bst*XI restriction enzyme sequence (depicted at the top of the diagram and shown in bold in other sequences). X: any nucleotide. Arrows indicate 5'-3' directionality of the cDNA sequence.

Appendix 3

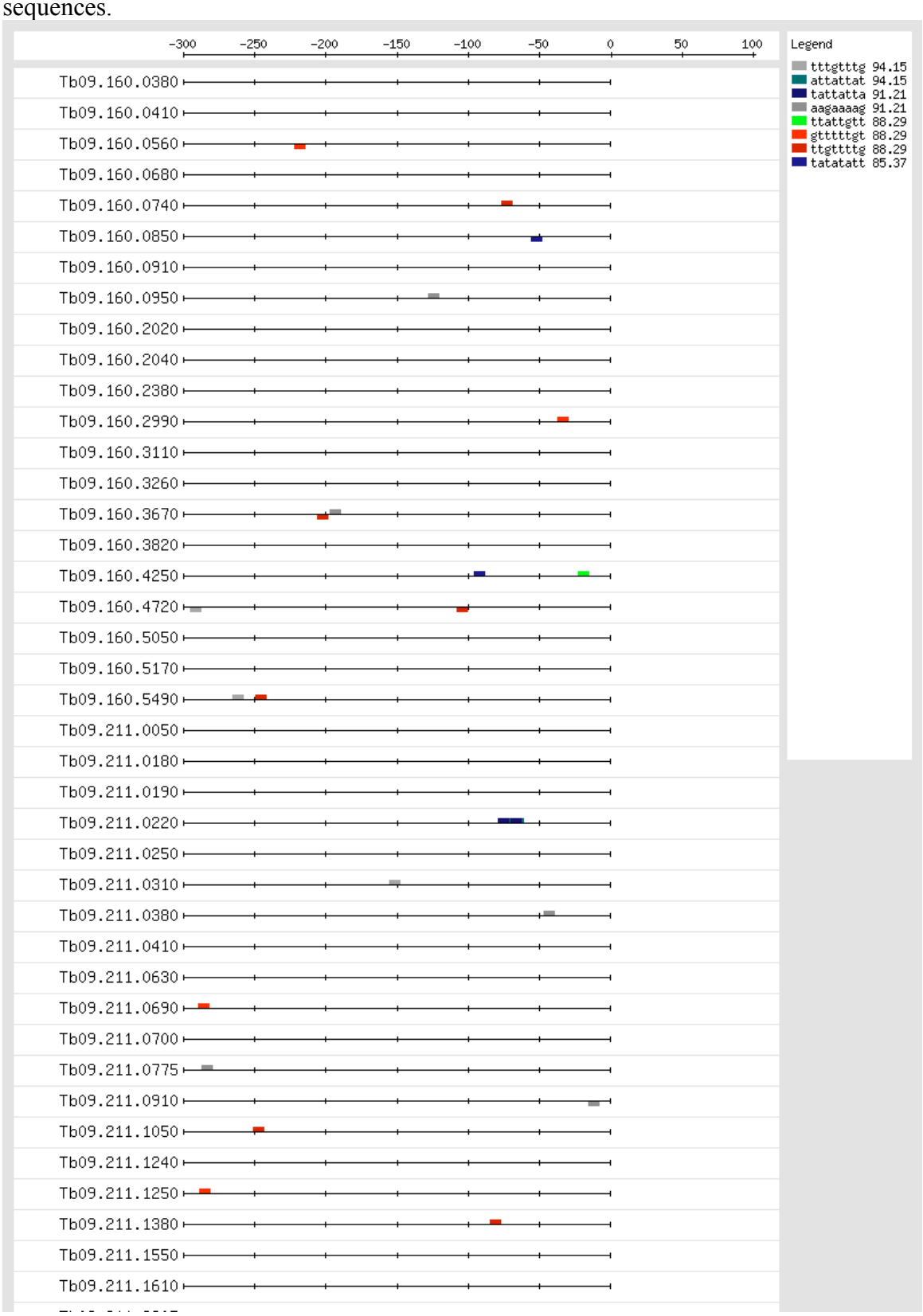


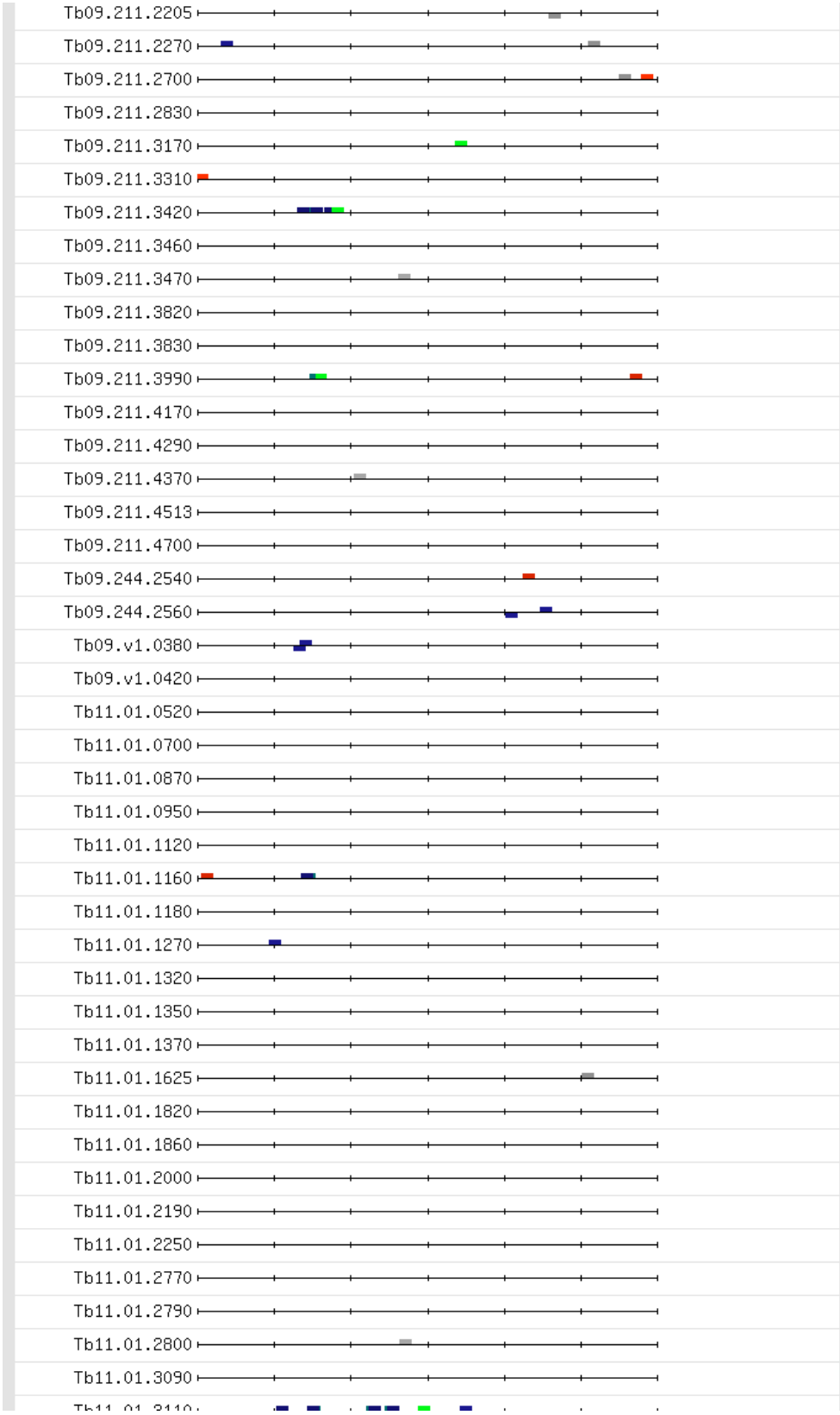
ClustalW2 alignment of the protein sequences of four *Trypanosoma brucei* eIF4E homologues ('4E1-4'), human eIF4E ('human'; accession number: P06730), *S.*

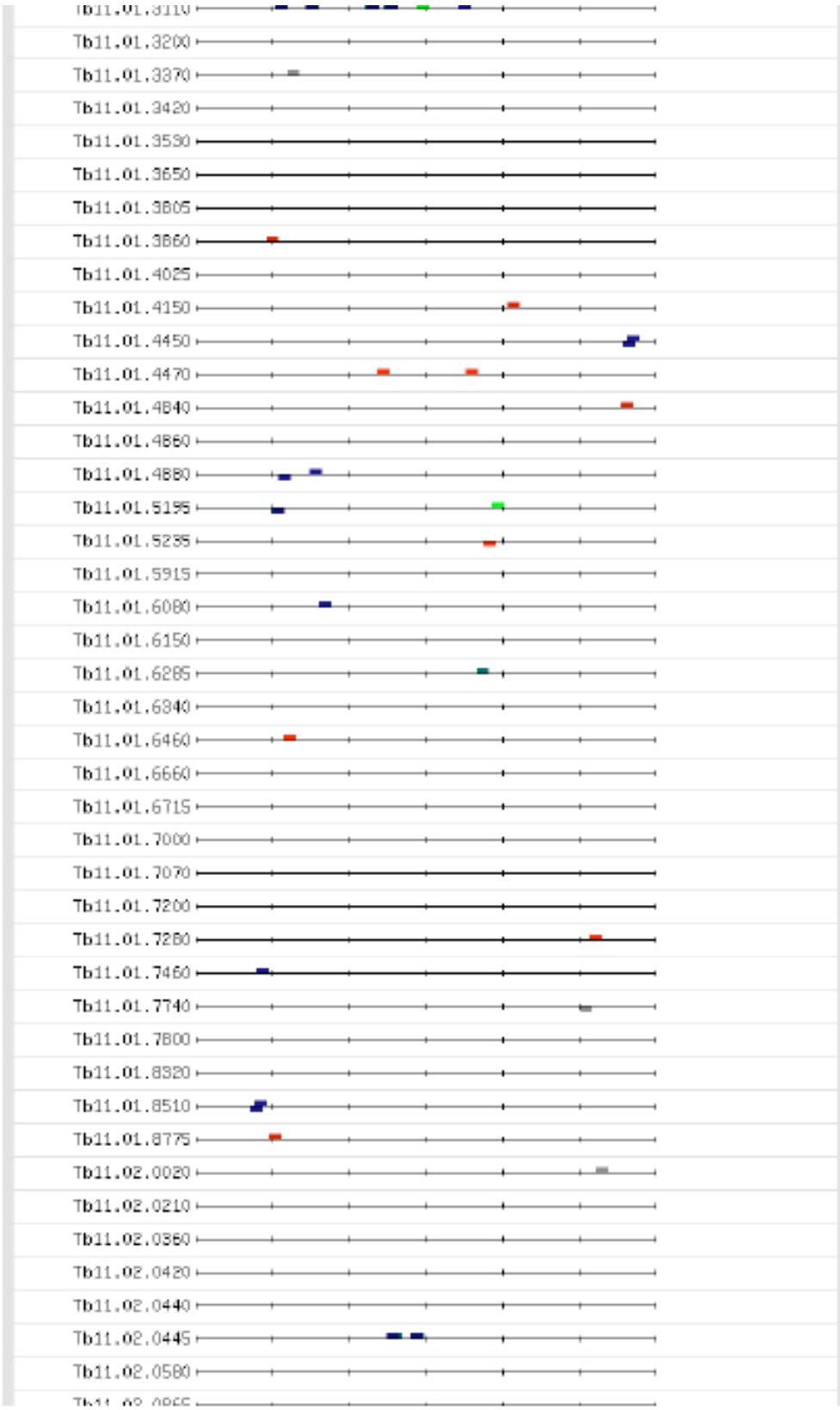
cerevisiae eIF4E ('S.cerevisiae'; accession number: P07260), and 4EHP from *Drosophila melangaster* ('4EHP'; accession number: Q8T3K5). Residues highlighted in grey are those reported to be important for interactions with the mRNA cap or eIF4G (see Section 5.2) and were used for the generation of Table 5.1. Asterisk (*): single, fully conserved residue; colon (:): conservation between groups of strongly similar properties; period (.): conservation between groups of weakly similar properties.

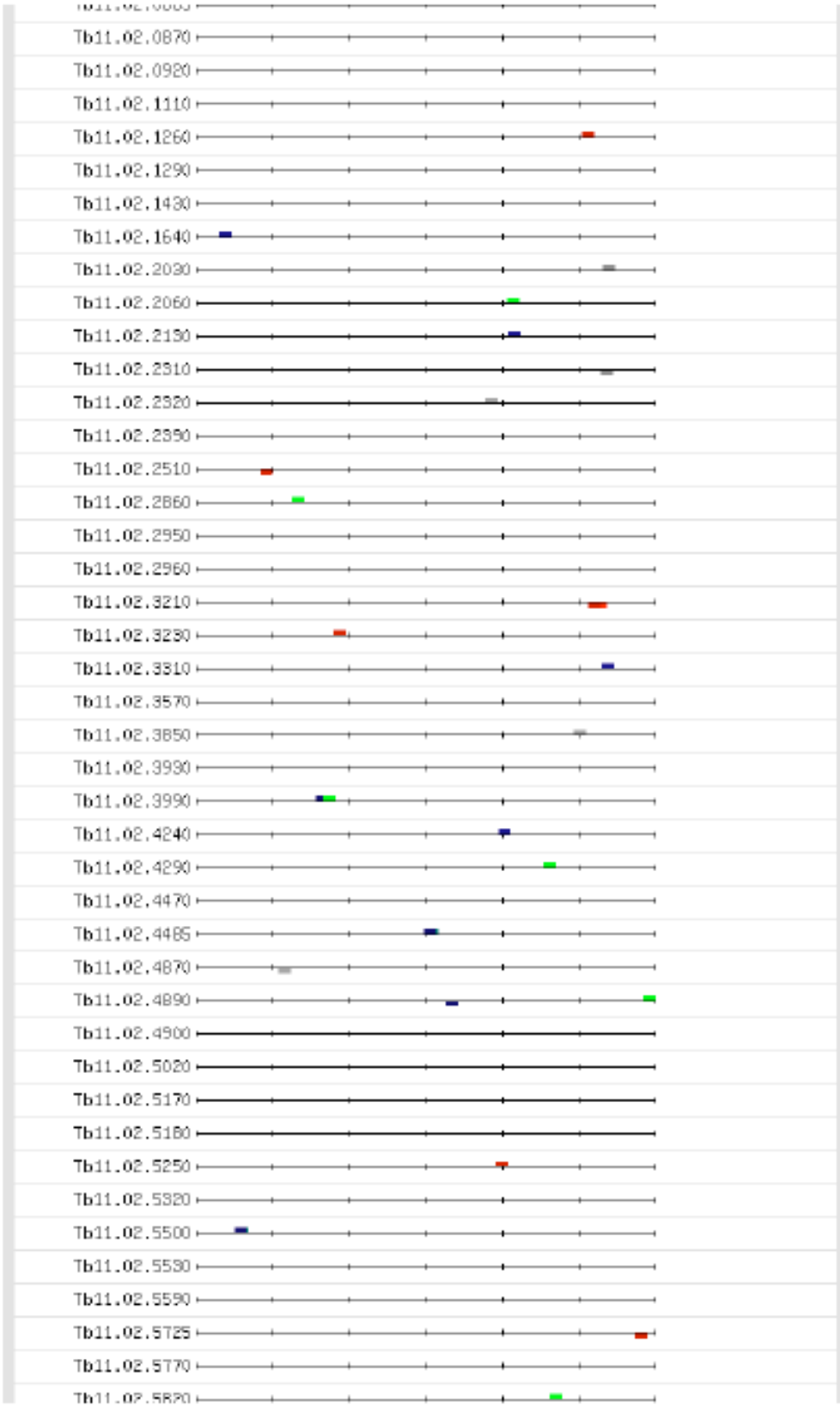
Appendix 4

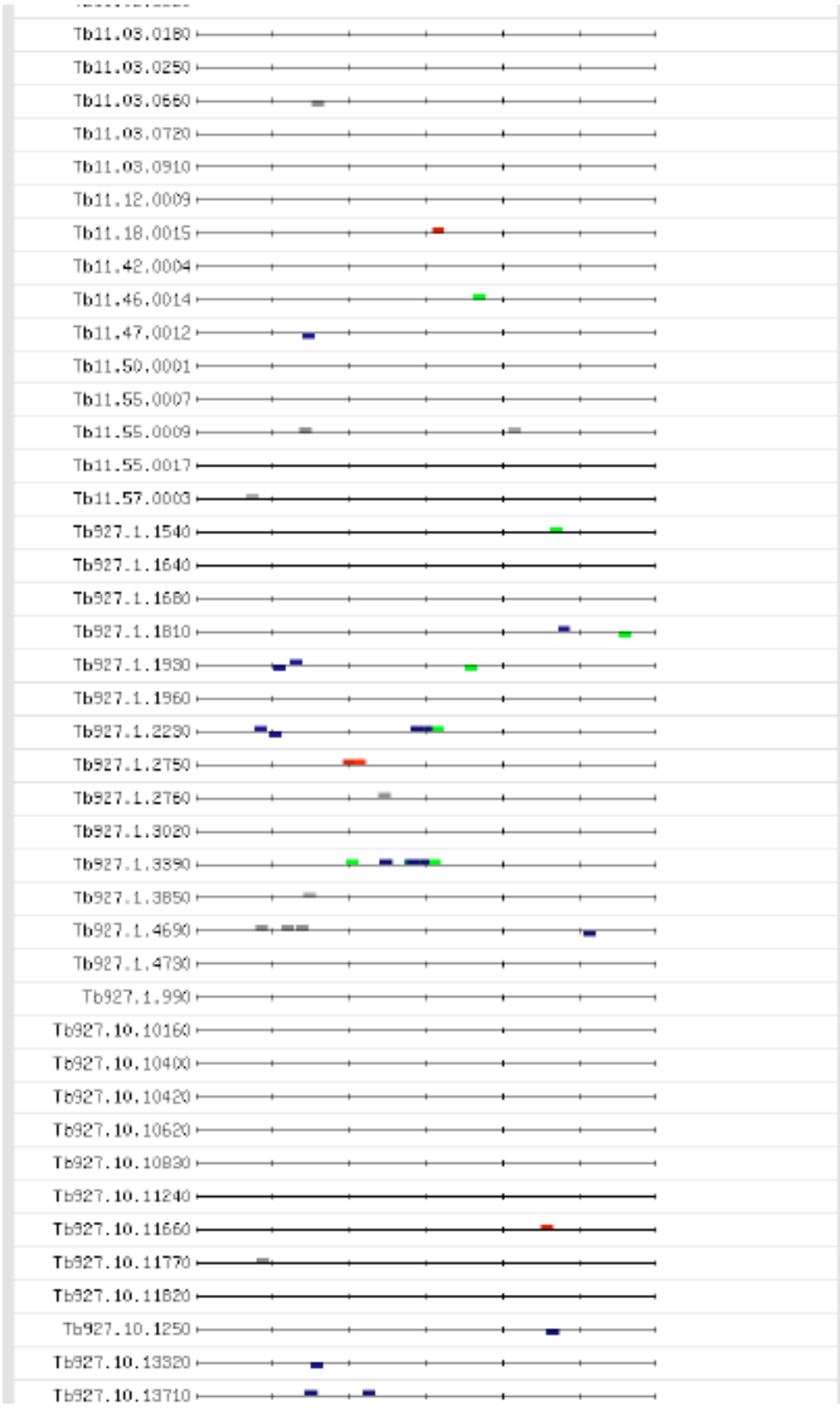
Feature map from RSAT oligo-analysis of genes enriched in stumpy form polysome associated transcripts, showing locations of over-represented unique 8-mer sequences.

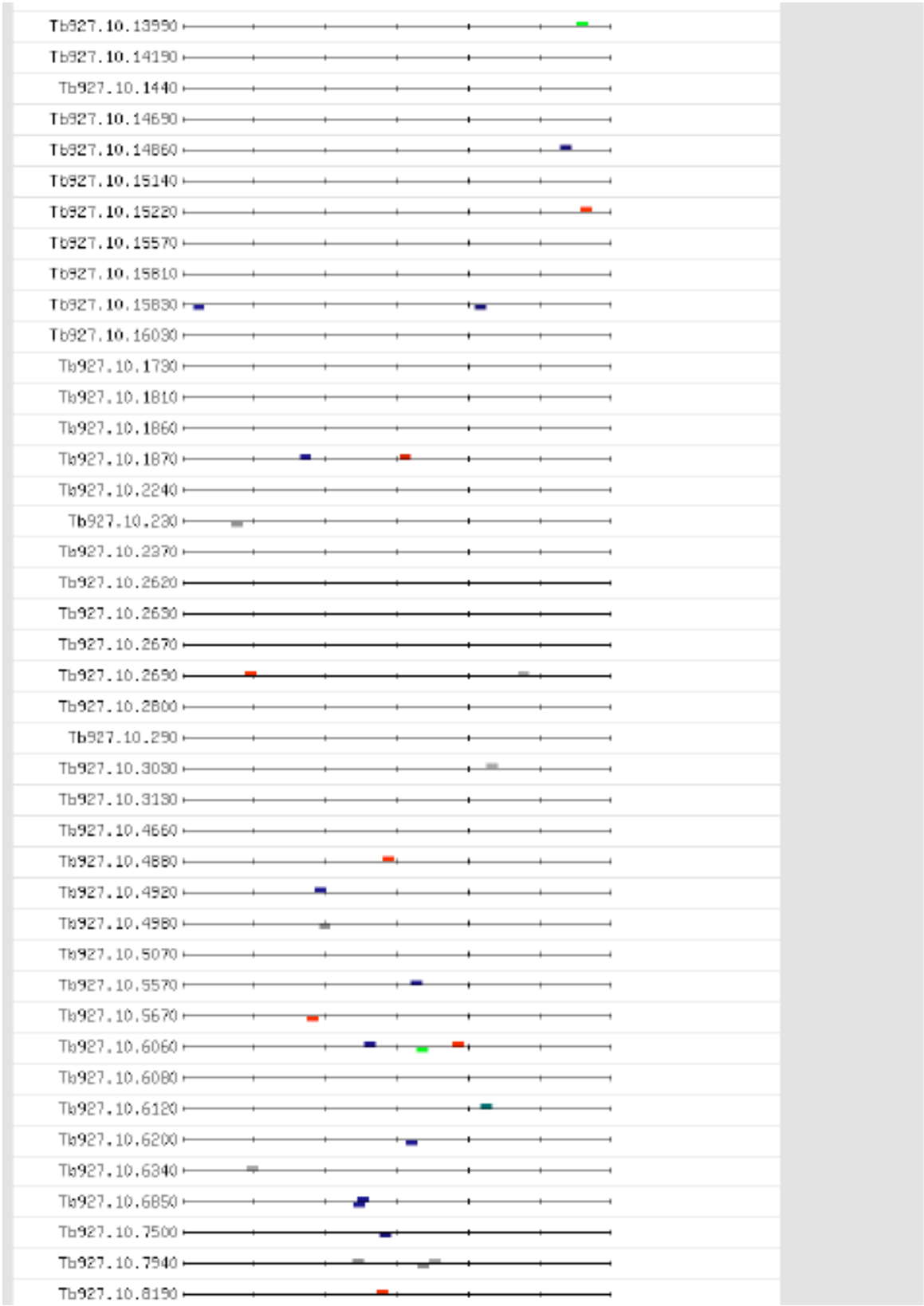


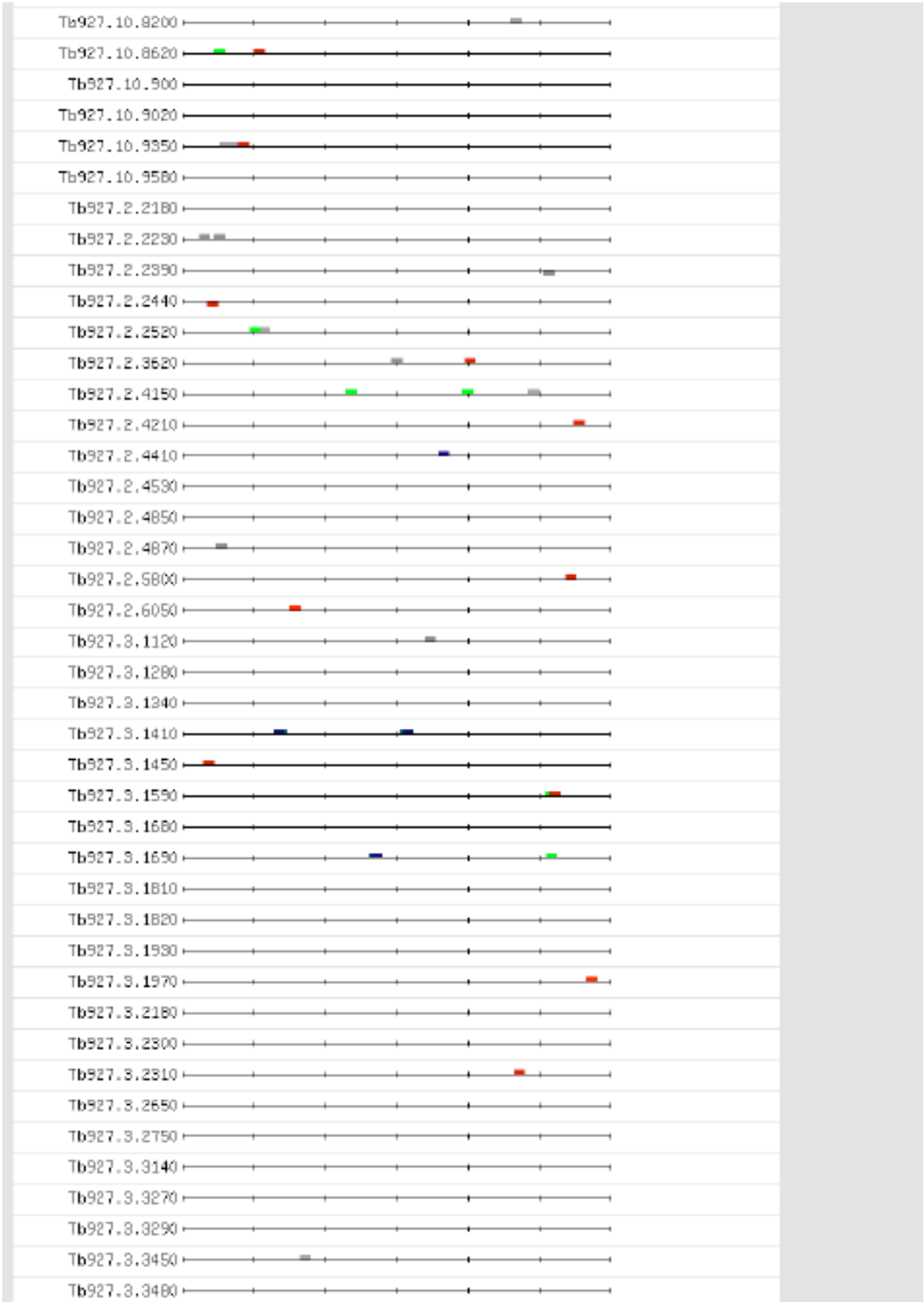

Tb09.211.2205
Tb09.211.2270
Tb09.211.2700
Tb09.211.2830
Tb09.211.3170
Tb09.211.3310
Tb09.211.3420
Tb09.211.3460
Tb09.211.3470
Tb09.211.3820
Tb09.211.3830
Tb09.211.3990
Tb09.211.4170
Tb09.211.4290
Tb09.211.4370
Tb09.211.4513
Tb09.211.4700
Tb09.244.2540
Tb09.244.2560
Tb09.v1.0380
Tb09.v1.0420
Tb11.01.0520
Tb11.01.0700
Tb11.01.0870
Tb11.01.0950
Tb11.01.1120
Tb11.01.1160
Tb11.01.1180
Tb11.01.1270
Tb11.01.1320
Tb11.01.1350
Tb11.01.1370
Tb11.01.1625
Tb11.01.1820
Tb11.01.1860
Tb11.01.2000
Tb11.01.2190
Tb11.01.2250
Tb11.01.2770
Tb11.01.2790
Tb11.01.2800
Tb11.01.3090

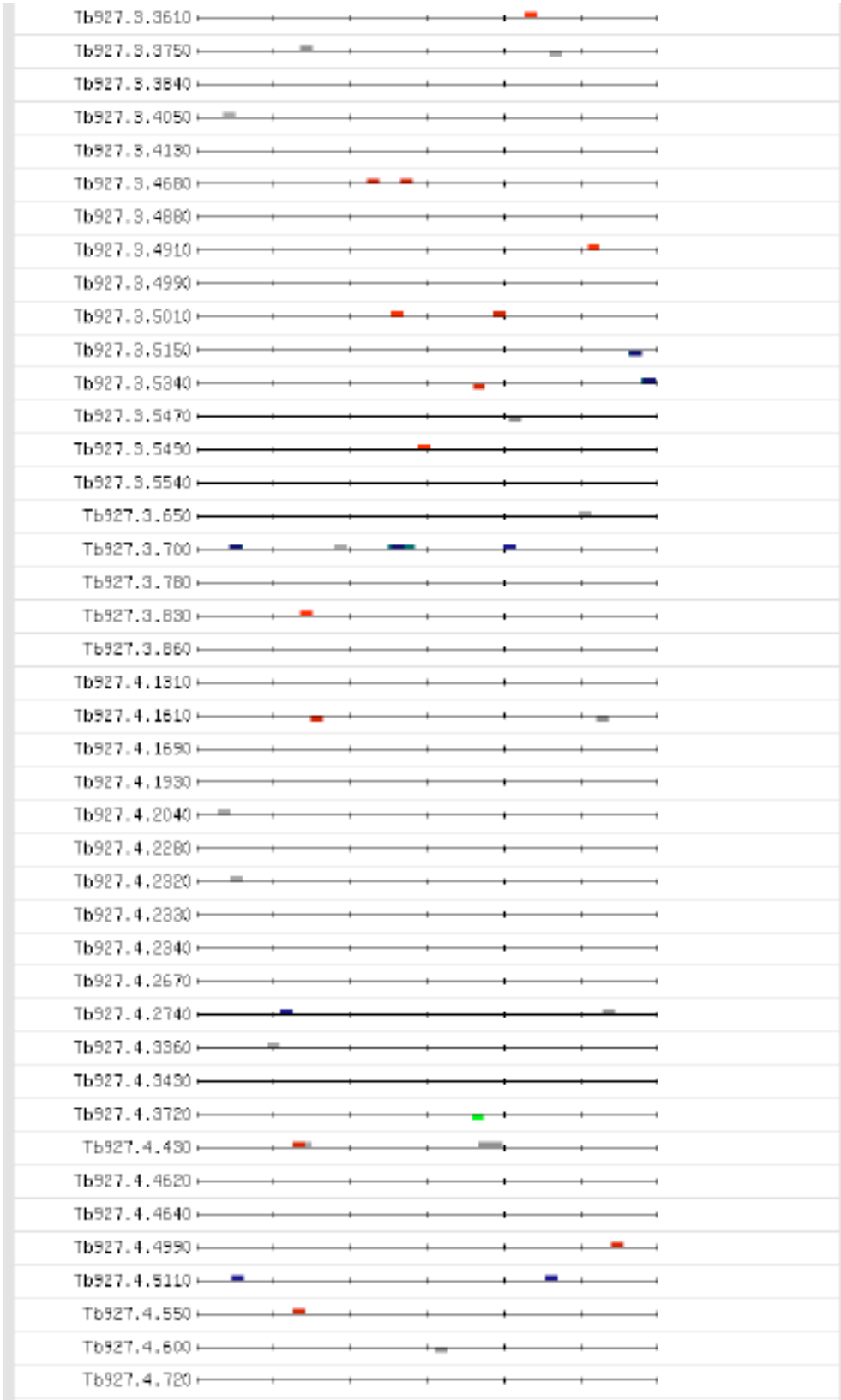

Tb11.01.3110
Tb11.01.3200
Tb11.01.3370
Tb11.01.3420
Tb11.01.3530
Tb11.01.3650
Tb11.01.3805
Tb11.01.3860
Tb11.01.4025
Tb11.01.4150
Tb11.01.4450
Tb11.01.4470
Tb11.01.4840
Tb11.01.4860
Tb11.01.4880
Tb11.01.5195
Tb11.01.5235
Tb11.01.5915
Tb11.01.6080
Tb11.01.6150
Tb11.01.6285
Tb11.01.6340
Tb11.01.6460
Tb11.01.6660
Tb11.01.6715
Tb11.01.7000
Tb11.01.7070
Tb11.01.7200
Tb11.01.7280
Tb11.01.7460
Tb11.01.7740
Tb11.01.7800
Tb11.01.8320
Tb11.01.8510
Tb11.01.8775
Tb11.02.0020
Tb11.02.0210
Tb11.02.0360
Tb11.02.0420
Tb11.02.0440
Tb11.02.0445
Tb11.02.0580

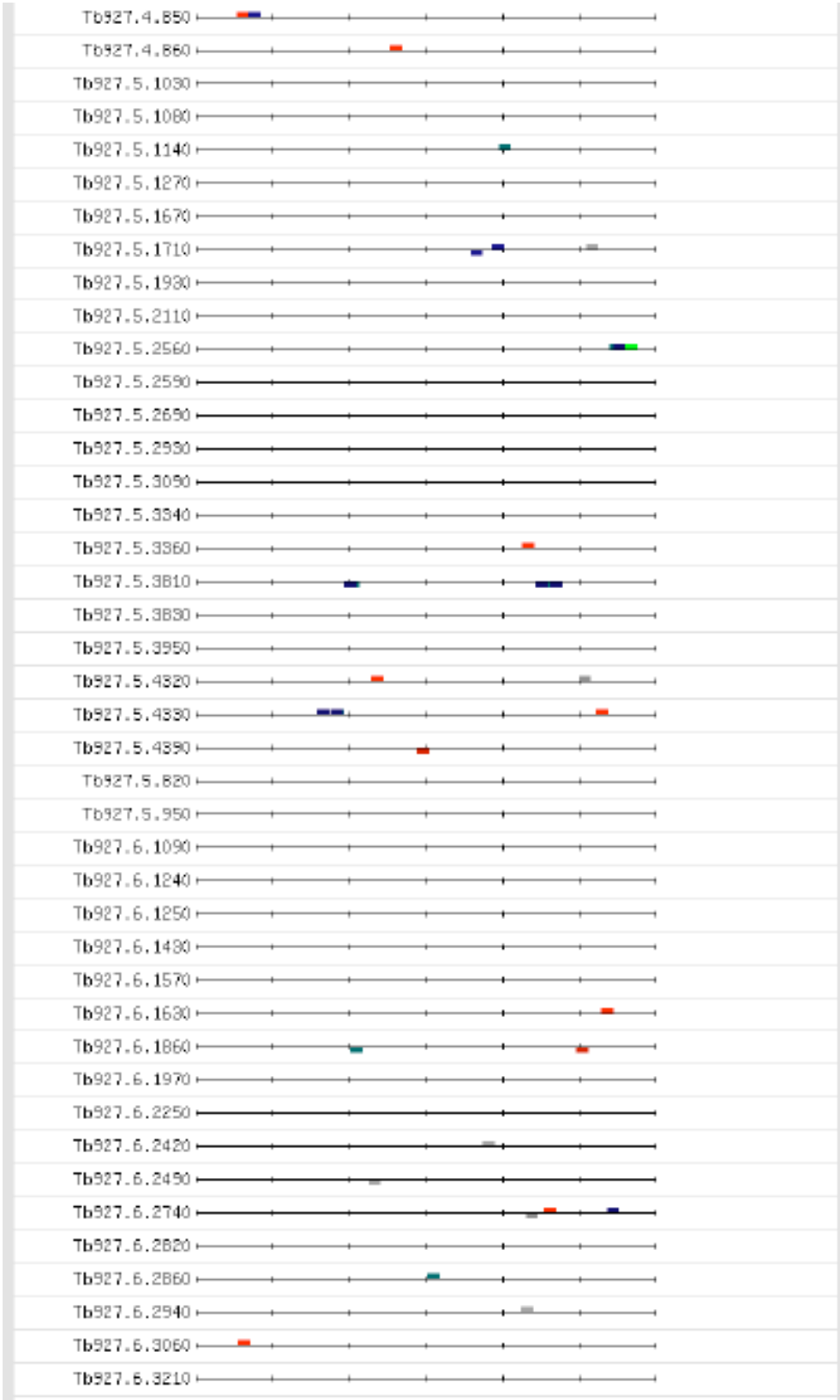

Tb11.02.0870
Tb11.02.0920
Tb11.02.1110
Tb11.02.1260
Tb11.02.1290
Tb11.02.1430
Tb11.02.1640
Tb11.02.2030
Tb11.02.2060
Tb11.02.2130
Tb11.02.2310
Tb11.02.2320
Tb11.02.2350
Tb11.02.2510
Tb11.02.2860
Tb11.02.2950
Tb11.02.2960
Tb11.02.3210
Tb11.02.3230
Tb11.02.3310
Tb11.02.3570
Tb11.02.3850
Tb11.02.3930
Tb11.02.3990
Tb11.02.4240
Tb11.02.4290
Tb11.02.4470
Tb11.02.4485
Tb11.02.4870
Tb11.02.4890
Tb11.02.4900
Tb11.02.5020
Tb11.02.5170
Tb11.02.5180
Tb11.02.5250
Tb11.02.5320
Tb11.02.5500
Tb11.02.5530
Tb11.02.5590
Tb11.02.5725
Tb11.02.5770

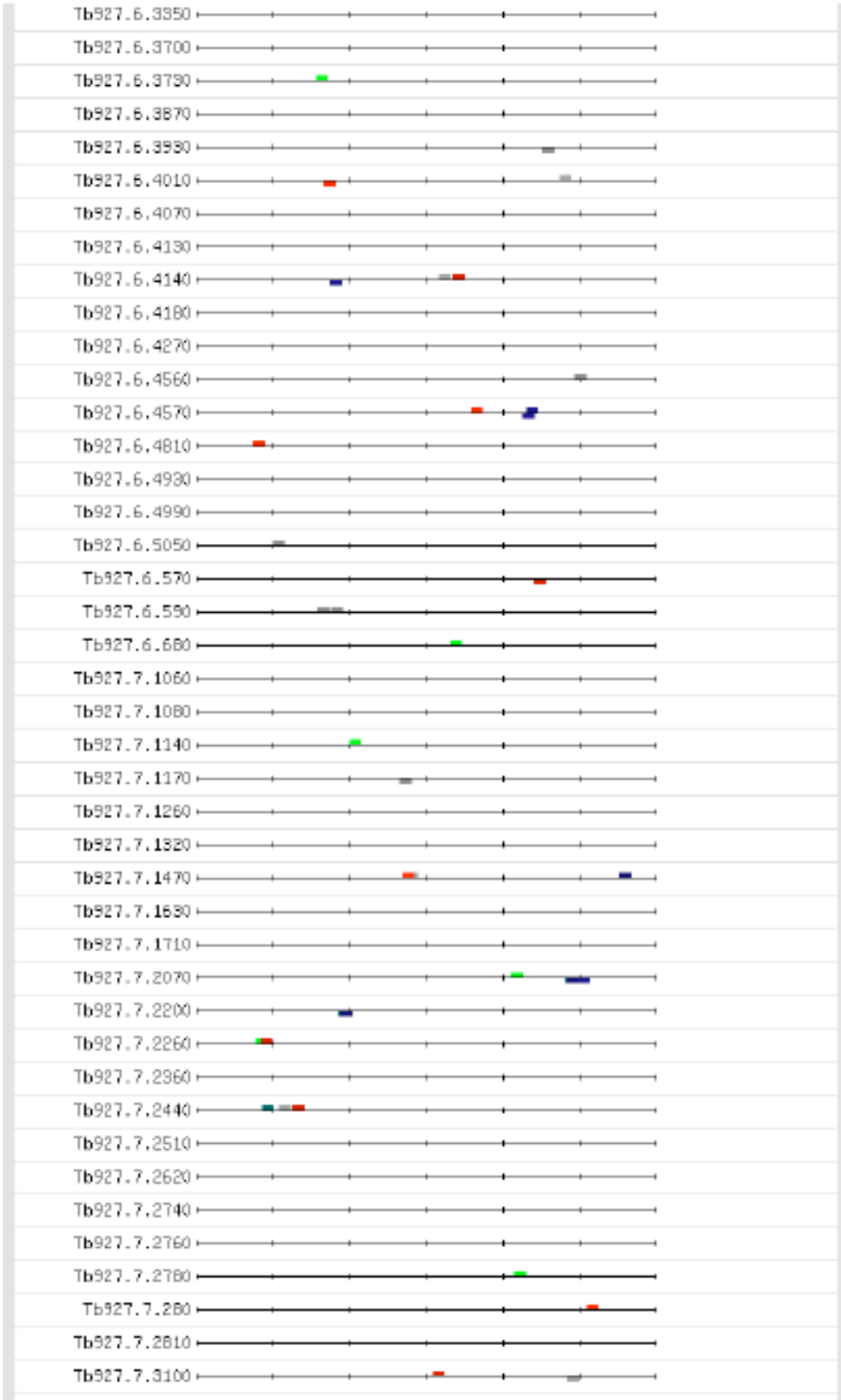

Tb11.03.0180
Tb11.03.0250
Tb11.03.0660
Tb11.03.0720
Tb11.03.0910
Tb11.12.0009
Tb11.18.0015
Tb11.42.0004
Tb11.46.0014
Tb11.47.0012
Tb11.50.0001
Tb11.55.0007
Tb11.55.0009
Tb11.55.0017
Tb11.57.0003
Tb927.1.1540
Tb927.1.1640
Tb927.1.1680
Tb927.1.1810
Tb927.1.1930
Tb927.1.1960
Tb927.1.2290
Tb927.1.2750
Tb927.1.2760
Tb927.1.3020
Tb927.1.3390
Tb927.1.3850
Tb927.1.4690
Tb927.1.4790
Tb927.1.990
Tb927.10.10160
Tb927.10.10400
Tb927.10.10420
Tb927.10.10620
Tb927.10.10830
Tb927.10.11240
Tb927.10.11660
Tb927.10.11770
Tb927.10.11820
Tb927.10.1250
Tb927.10.13320
Tb927.10.13710

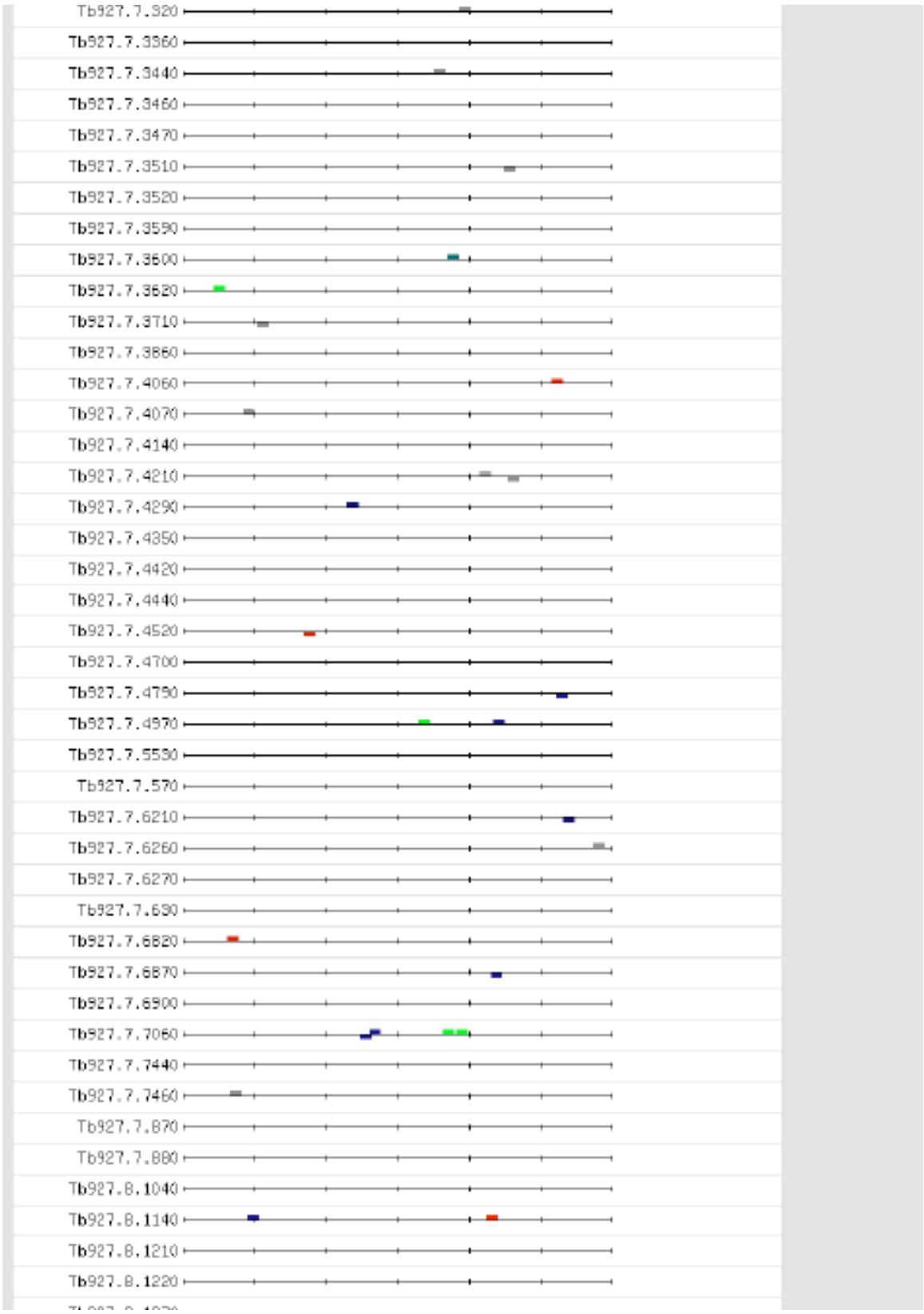

Tb927.10.13950
Tb927.10.14190
Tb927.10.1440
Tb927.10.14690
Tb927.10.14860
Tb927.10.15140
Tb927.10.15220
Tb927.10.15570
Tb927.10.15810
Tb927.10.15830
Tb927.10.16090
Tb927.10.1790
Tb927.10.1810
Tb927.10.1860
Tb927.10.1870
Tb927.10.2240
Tb927.10.230
Tb927.10.2370
Tb927.10.2620
Tb927.10.2630
Tb927.10.2670
Tb927.10.2690
Tb927.10.2800
Tb927.10.290
Tb927.10.3030
Tb927.10.3130
Tb927.10.4660
Tb927.10.4880
Tb927.10.4920
Tb927.10.4980
Tb927.10.5070
Tb927.10.5570
Tb927.10.5670
Tb927.10.6060
Tb927.10.6080
Tb927.10.6120
Tb927.10.6200
Tb927.10.6340
Tb927.10.6850
Tb927.10.7500
Tb927.10.7940
Tb927.10.8150

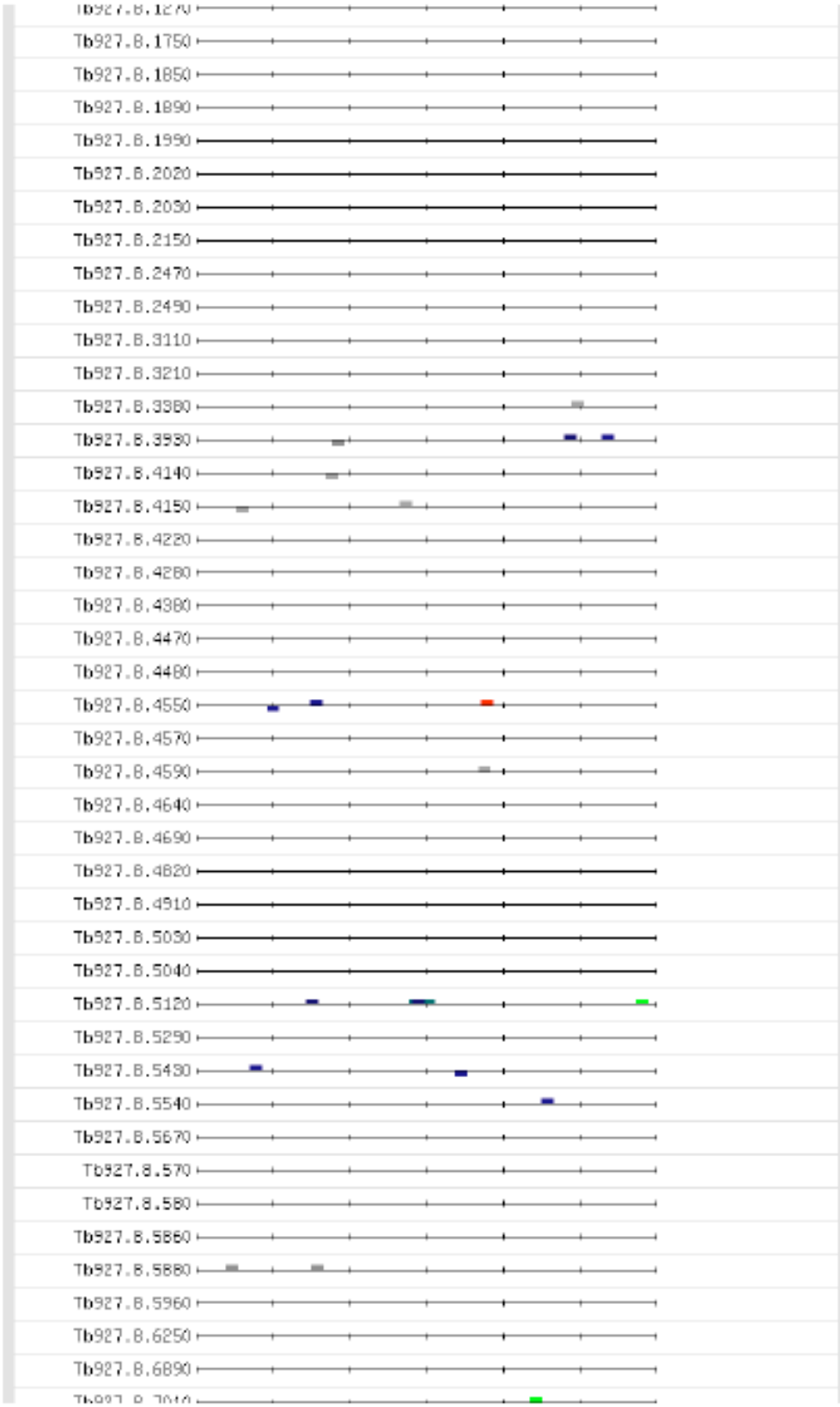

Tb927.10.8200
Tb927.10.8620
Tb927.10.900
Tb927.10.9020
Tb927.10.9350
Tb927.10.9580
Tb927.2.2180
Tb927.2.2230
Tb927.2.2390
Tb927.2.2440
Tb927.2.2520
Tb927.2.3620
Tb927.2.4150
Tb927.2.4210
Tb927.2.4410
Tb927.2.4530
Tb927.2.4850
Tb927.2.4870
Tb927.2.5800
Tb927.2.6050
Tb927.3.1120
Tb927.3.1280
Tb927.3.1340
Tb927.3.1410
Tb927.3.1450
Tb927.3.1550
Tb927.3.1680
Tb927.3.1690
Tb927.3.1810
Tb927.3.1820
Tb927.3.1930
Tb927.3.1970
Tb927.3.2180
Tb927.3.2300
Tb927.3.2310
Tb927.3.2650
Tb927.3.2750
Tb927.3.3140
Tb927.3.3270
Tb927.3.3290
Tb927.3.3450
Tb927.3.3480

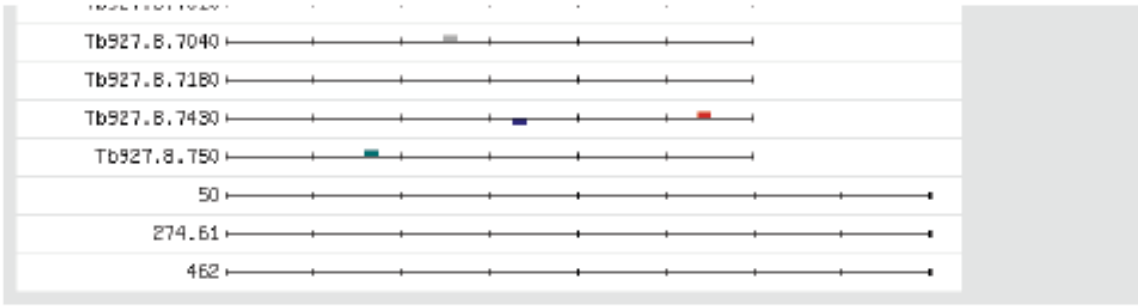

Tb927.3.3610
Tb927.3.3750
Tb927.3.3840
Tb927.3.4050
Tb927.3.4130
Tb927.3.4680
Tb927.3.4880
Tb927.3.4910
Tb927.3.4990
Tb927.3.5010
Tb927.3.5150
Tb927.3.5340
Tb927.3.5470
Tb927.3.5490
Tb927.3.5540
Tb927.3.650
Tb927.3.700
Tb927.3.780
Tb927.3.830
Tb927.3.860
Tb927.4.1310
Tb927.4.1610
Tb927.4.1690
Tb927.4.1930
Tb927.4.2040
Tb927.4.2280
Tb927.4.2320
Tb927.4.2330
Tb927.4.2340
Tb927.4.2670
Tb927.4.2740
Tb927.4.3360
Tb927.4.3430
Tb927.4.3720
Tb927.4.430
Tb927.4.4620
Tb927.4.4640
Tb927.4.4990
Tb927.4.5110
Tb927.4.550
Tb927.4.600
Tb927.4.720


Tb927.4.850
Tb927.4.860
Tb927.5.1030
Tb927.5.1080
Tb927.5.1140
Tb927.5.1270
Tb927.5.1670
Tb927.5.1710
Tb927.5.1930
Tb927.5.2110
Tb927.5.2560
Tb927.5.2590
Tb927.5.2690
Tb927.5.2930
Tb927.5.3050
Tb927.5.3340
Tb927.5.3360
Tb927.5.3810
Tb927.5.3830
Tb927.5.3950
Tb927.5.4320
Tb927.5.4330
Tb927.5.4390
Tb927.5.820
Tb927.5.950
Tb927.6.1090
Tb927.6.1240
Tb927.6.1250
Tb927.6.1490
Tb927.6.1570
Tb927.6.1630
Tb927.6.1860
Tb927.6.1970
Tb927.6.2250
Tb927.6.2420
Tb927.6.2450
Tb927.6.2740
Tb927.6.2820
Tb927.6.2860
Tb927.6.2940
Tb927.6.3060
Tb927.6.3210


Tb927.6.3350
Tb927.6.3700
Tb927.6.3730
Tb927.6.3870
Tb927.6.3930
Tb927.6.4010
Tb927.6.4070
Tb927.6.4130
Tb927.6.4140
Tb927.6.4180
Tb927.6.4270
Tb927.6.4560
Tb927.6.4570
Tb927.6.4810
Tb927.6.4930
Tb927.6.4990
Tb927.6.5050
Tb927.6.570
Tb927.6.590
Tb927.6.680
Tb927.7.1060
Tb927.7.1080
Tb927.7.1140
Tb927.7.1170
Tb927.7.1260
Tb927.7.1320
Tb927.7.1470
Tb927.7.1630
Tb927.7.1710
Tb927.7.2070
Tb927.7.2200
Tb927.7.2260
Tb927.7.2360
Tb927.7.2440
Tb927.7.2510
Tb927.7.2620
Tb927.7.2740
Tb927.7.2760
Tb927.7.2780
Tb927.7.280
Tb927.7.2810
Tb927.7.3100

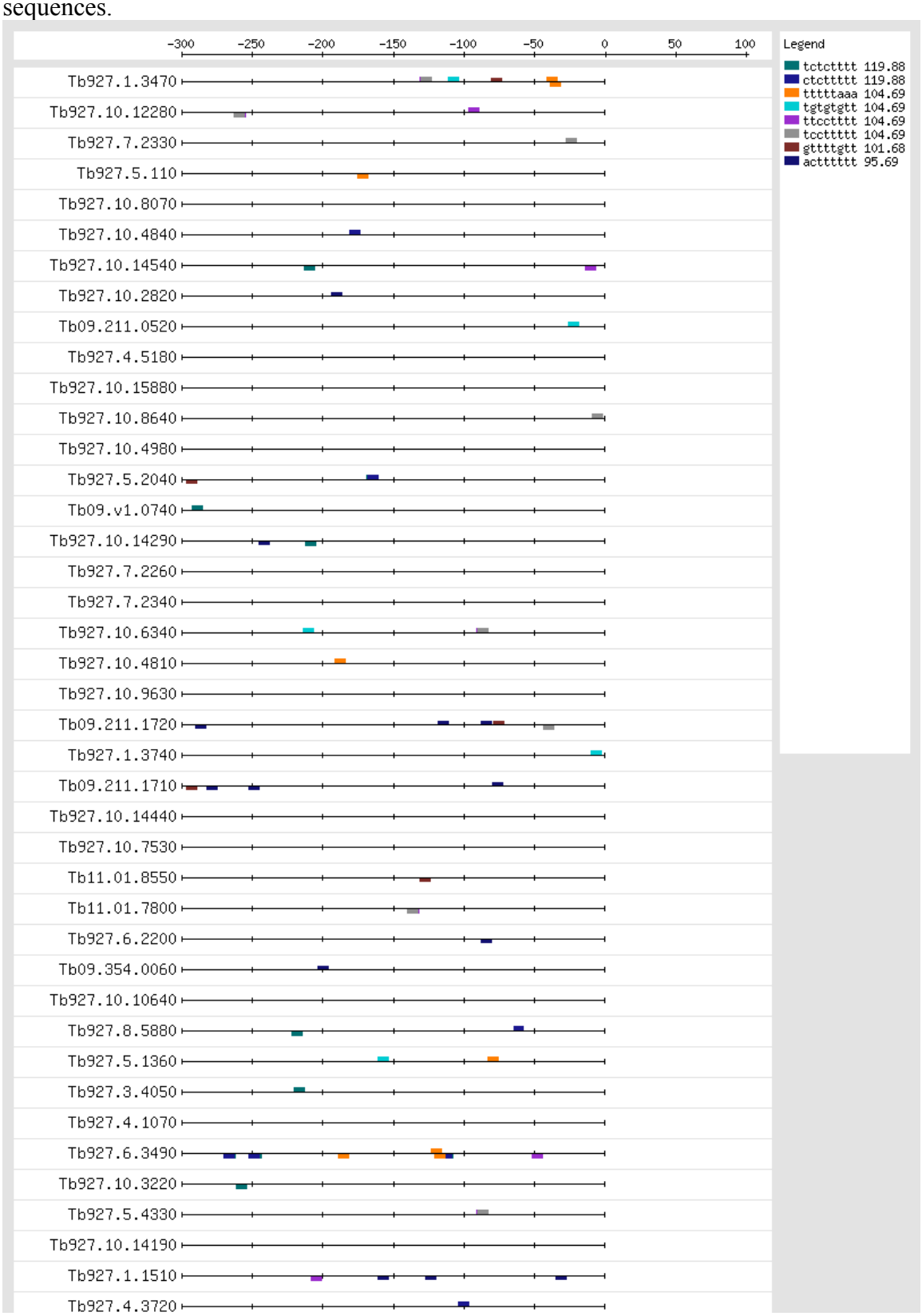

Tb927.7.320
Tb927.7.3360
Tb927.7.3440
Tb927.7.3460
Tb927.7.3470
Tb927.7.3510
Tb927.7.3520
Tb927.7.3590
Tb927.7.3600
Tb927.7.3620
Tb927.7.3710
Tb927.7.3860
Tb927.7.4060
Tb927.7.4070
Tb927.7.4140
Tb927.7.4210
Tb927.7.4290
Tb927.7.4350
Tb927.7.4420
Tb927.7.4440
Tb927.7.4520
Tb927.7.4700
Tb927.7.4790
Tb927.7.4970
Tb927.7.5530
Tb927.7.570
Tb927.7.6210
Tb927.7.6260
Tb927.7.6270
Tb927.7.630
Tb927.7.6820
Tb927.7.6870
Tb927.7.6900
Tb927.7.7060
Tb927.7.7440
Tb927.7.7460
Tb927.7.870
Tb927.7.880
Tb927.8.1040
Tb927.8.1140
Tb927.8.1210
Tb927.8.1220

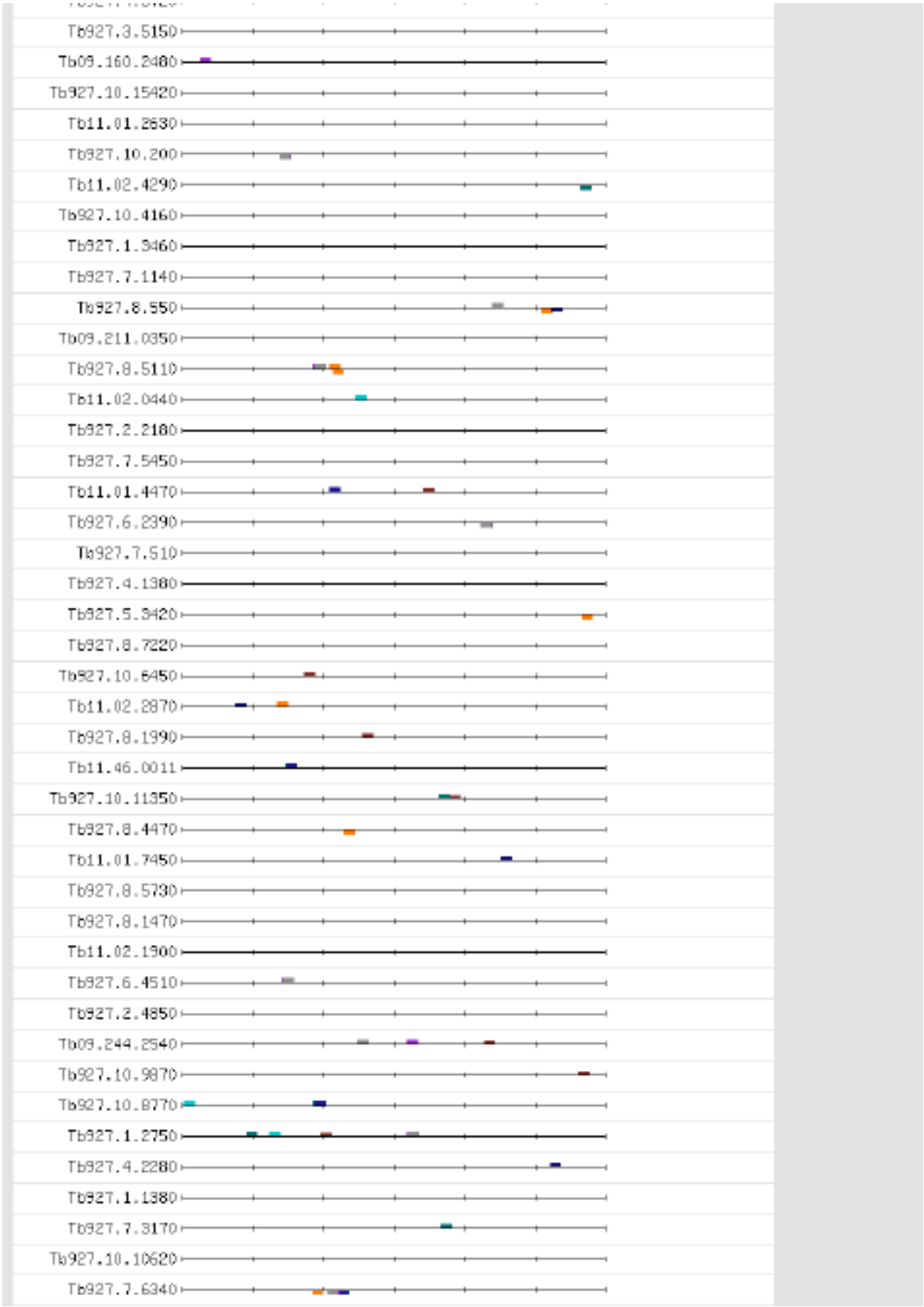

Tb927.8.1750
Tb927.8.1850
Tb927.8.1890
Tb927.8.1990
Tb927.8.2020
Tb927.8.2030
Tb927.8.2150
Tb927.8.2470
Tb927.8.2490
Tb927.8.3110
Tb927.8.3210
Tb927.8.3380
Tb927.8.3930
Tb927.8.4140
Tb927.8.4150
Tb927.8.4220
Tb927.8.4280
Tb927.8.4380
Tb927.8.4470
Tb927.8.4480
Tb927.8.4550
Tb927.8.4570
Tb927.8.4590
Tb927.8.4640
Tb927.8.4690
Tb927.8.4820
Tb927.8.4910
Tb927.8.5030
Tb927.8.5040
Tb927.8.5120
Tb927.8.5290
Tb927.8.5430
Tb927.8.5540
Tb927.8.5670
Tb927.8.570
Tb927.8.580
Tb927.8.5860
Tb927.8.5880
Tb927.8.5960
Tb927.8.6250
Tb927.8.6890

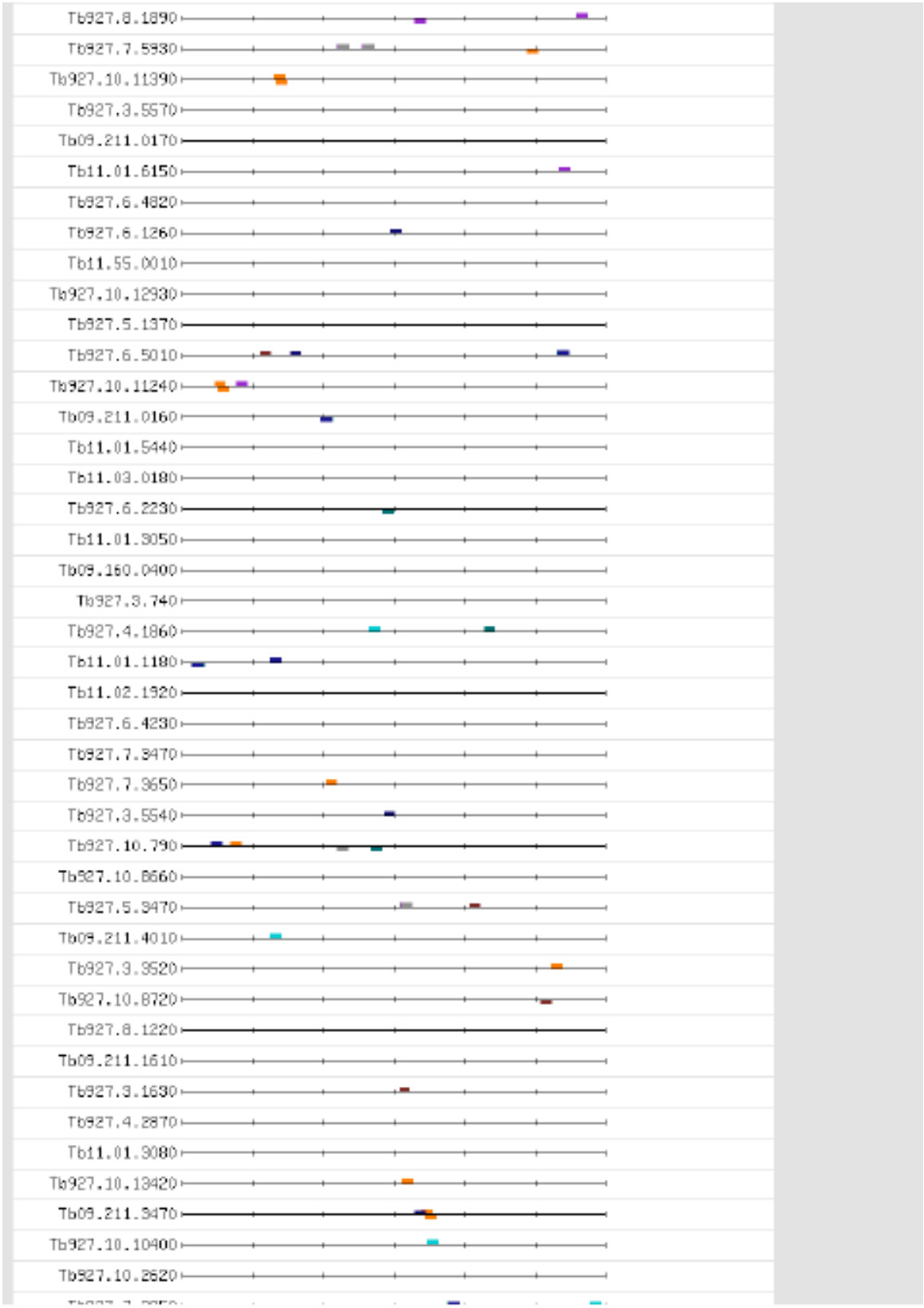

Tb927.8.7040
Tb927.8.7180
Tb927.8.7430
Tb927.8.750
50
274.61
462

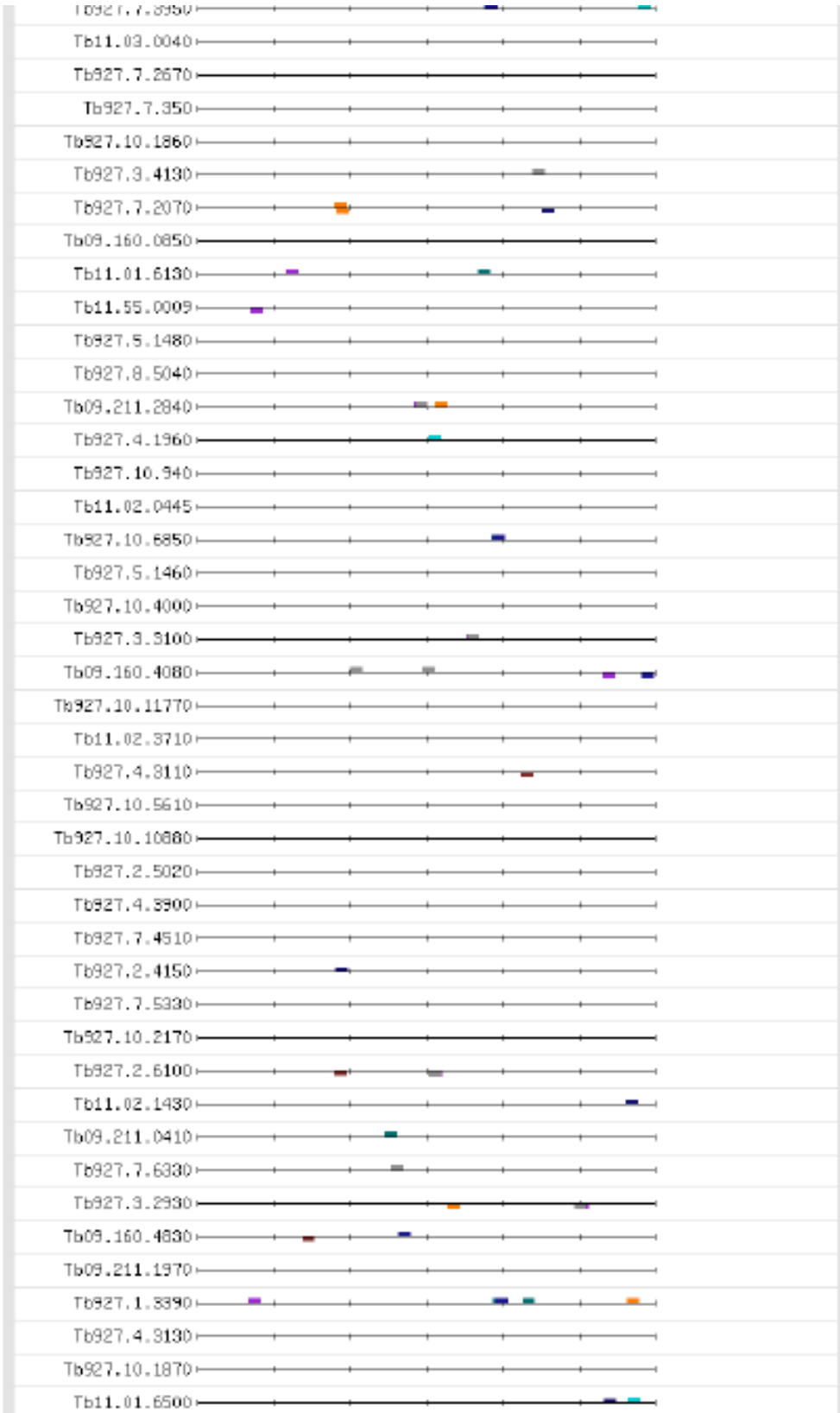
Appendix 5

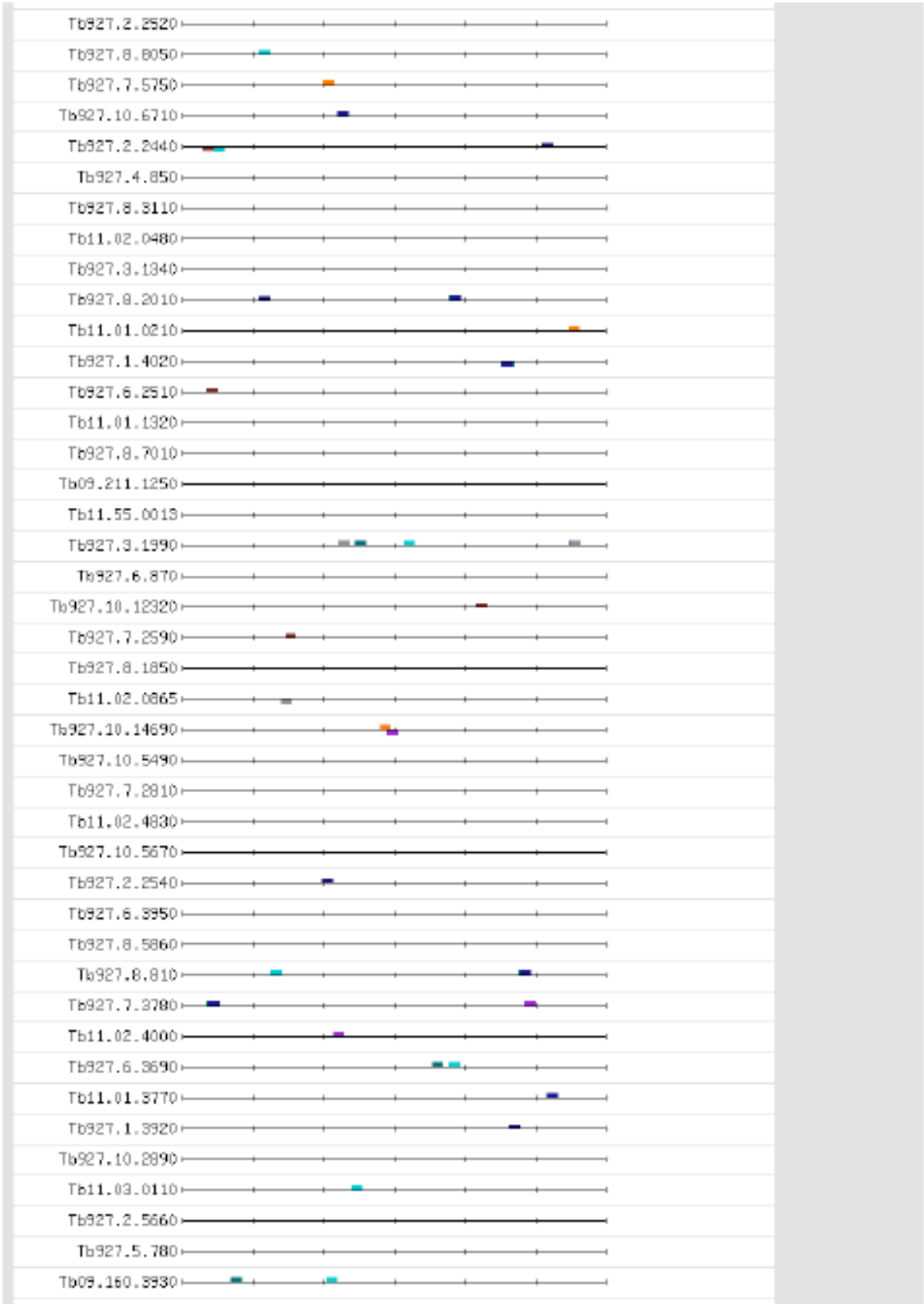
Feature map from RSAT oligo-analysis of genes enriched in slender form polysome associated transcripts, showing locations of over-represented unique 8-mer sequences.

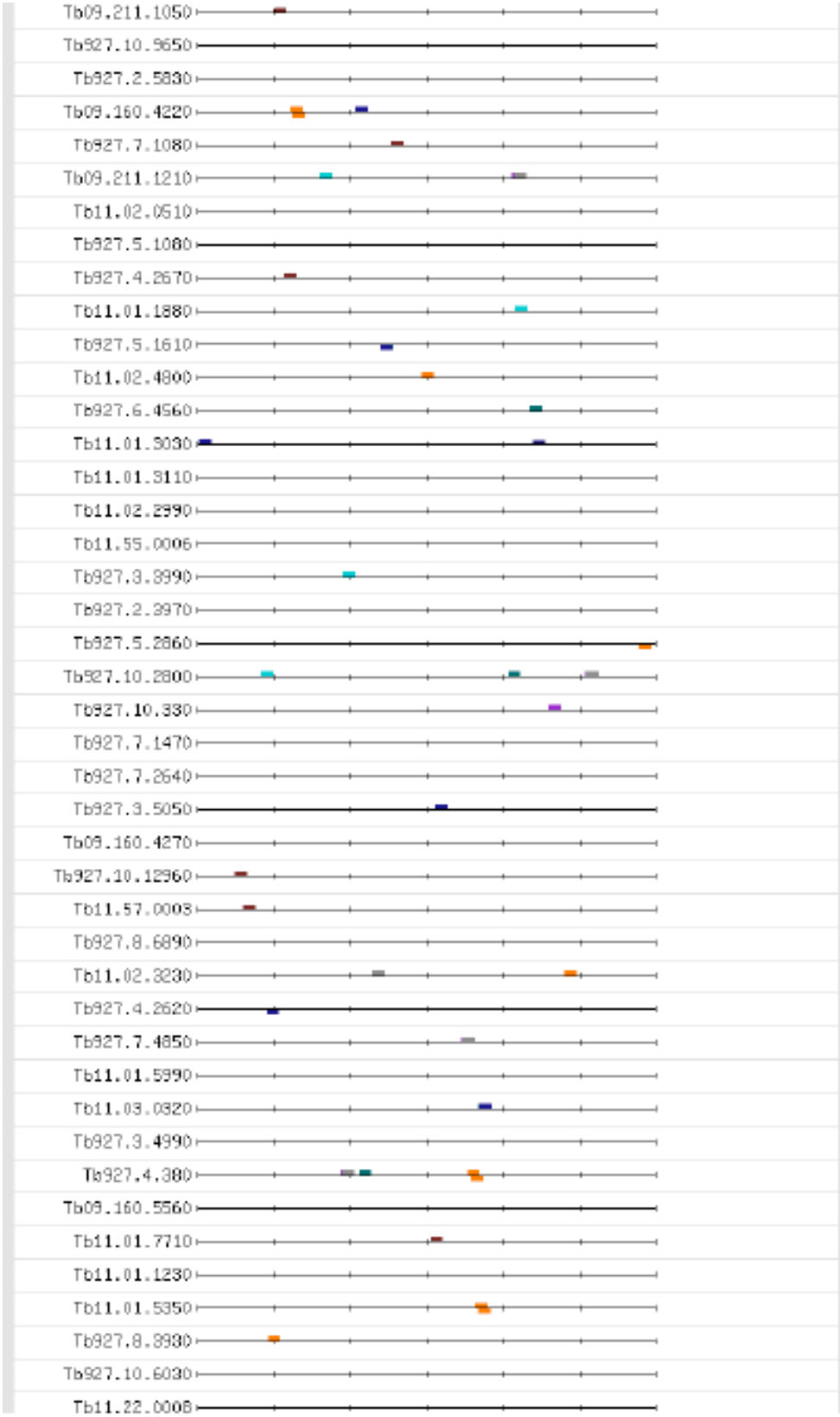


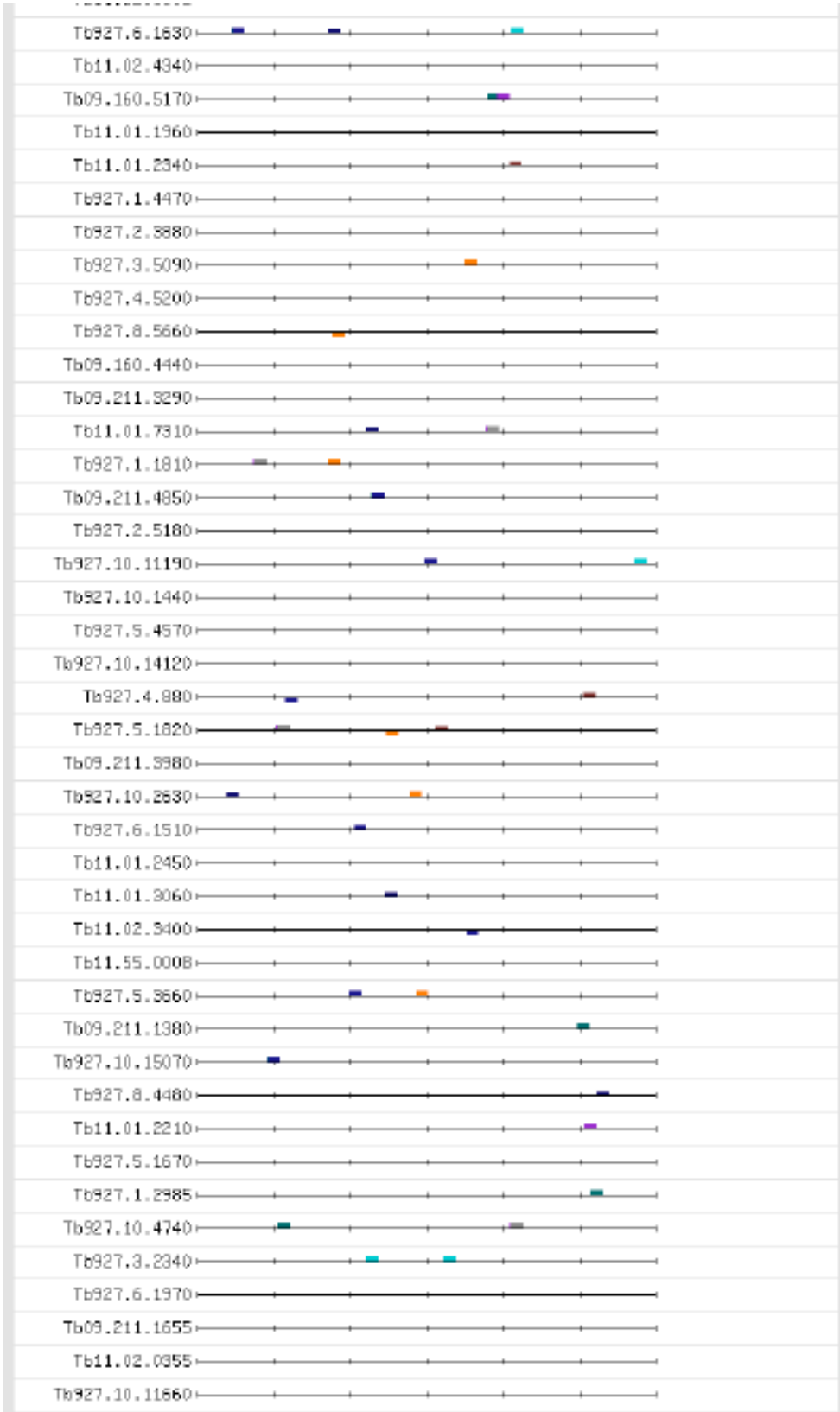

Tb927.3.5150
Tb09.160.2480
Tb927.10.15420
Tb11.01.2630
Tb927.10.200
Tb11.02.4290
Tb927.10.4160
Tb927.1.3460
Tb927.7.1140
Tb927.8.550
Tb09.211.0350
Tb927.8.5110
Tb11.02.0440
Tb927.2.2180
Tb927.7.5450
Tb11.01.4470
Tb927.6.2390
Tb927.7.510
Tb927.4.1380
Tb927.5.3420
Tb927.8.7220
Tb927.10.6450
Tb11.02.2870
Tb927.8.1990
Tb11.46.0011
Tb927.10.11350
Tb927.8.4470
Tb11.01.7450
Tb927.8.5730
Tb927.8.1470
Tb11.02.1900
Tb927.6.4510
Tb927.2.4850
Tb09.244.2540
Tb927.10.9870
Tb927.10.8770
Tb927.1.2750
Tb927.4.2280
Tb927.1.1380
Tb927.7.3170
Tb927.10.10620
Tb927.7.6340

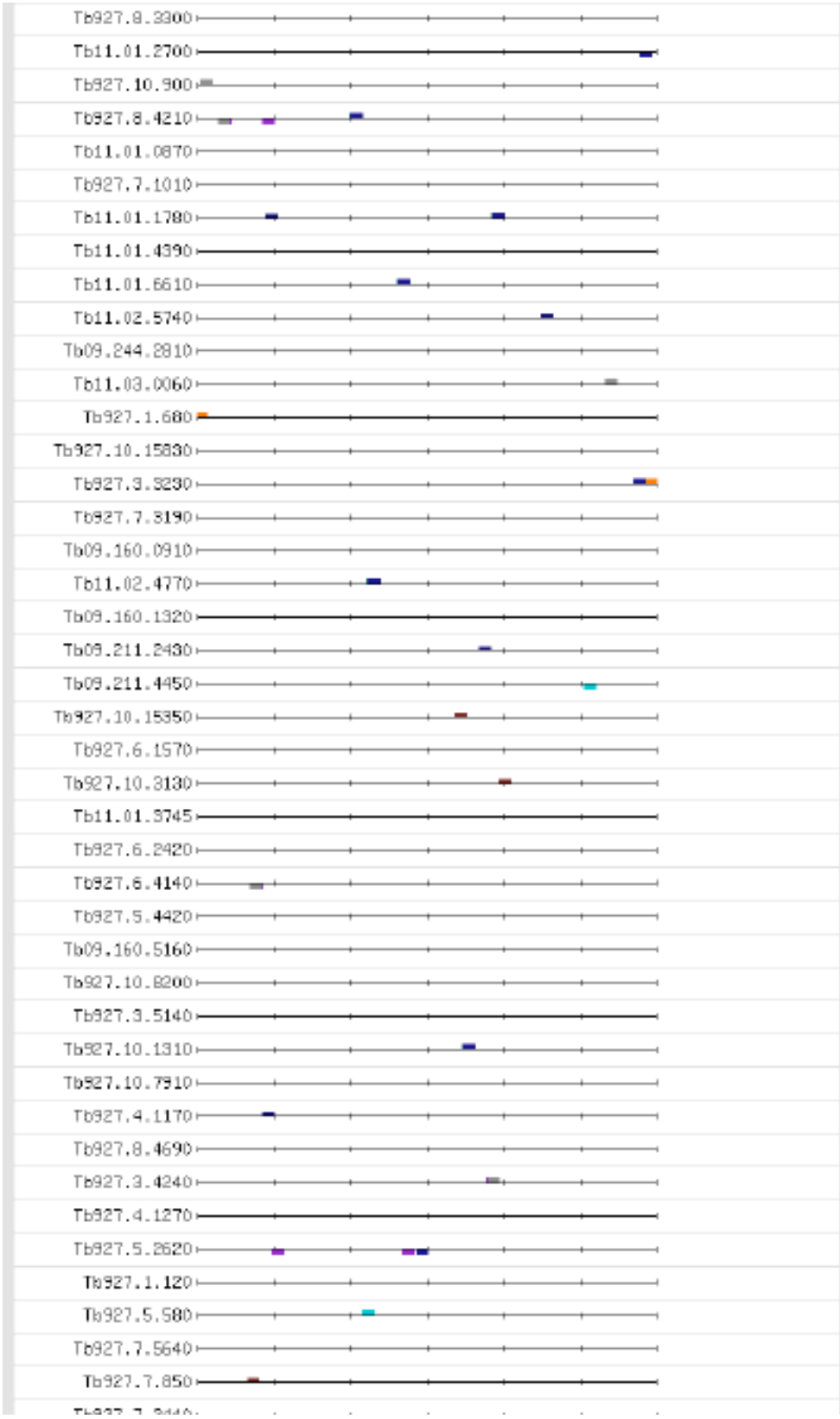

Tb927.8.1890
Tb927.7.5930
Tb927.10.11390
Tb927.3.5570
Tb09.211.0170
Tb11.01.6150
Tb927.6.4820
Tb927.6.1260
Tb11.55.0010
Tb927.10.12930
Tb927.5.1370
Tb927.6.5010
Tb927.10.11240
Tb09.211.0160
Tb11.01.5440
Tb11.03.0180
Tb927.6.2230
Tb11.01.3050
Tb09.160.0400
Tb927.3.740
Tb927.4.1860
Tb11.01.1180
Tb11.02.1920
Tb927.6.4230
Tb927.7.3470
Tb927.7.3650
Tb927.3.5540
Tb927.10.790
Tb927.10.8660
Tb927.5.3470
Tb09.211.4010
Tb927.3.3520
Tb927.10.8720
Tb927.8.1220
Tb09.211.1610
Tb927.3.1630
Tb927.4.2870
Tb11.01.3080
Tb927.10.13420
Tb09.211.3470
Tb927.10.10400
Tb927.10.2620

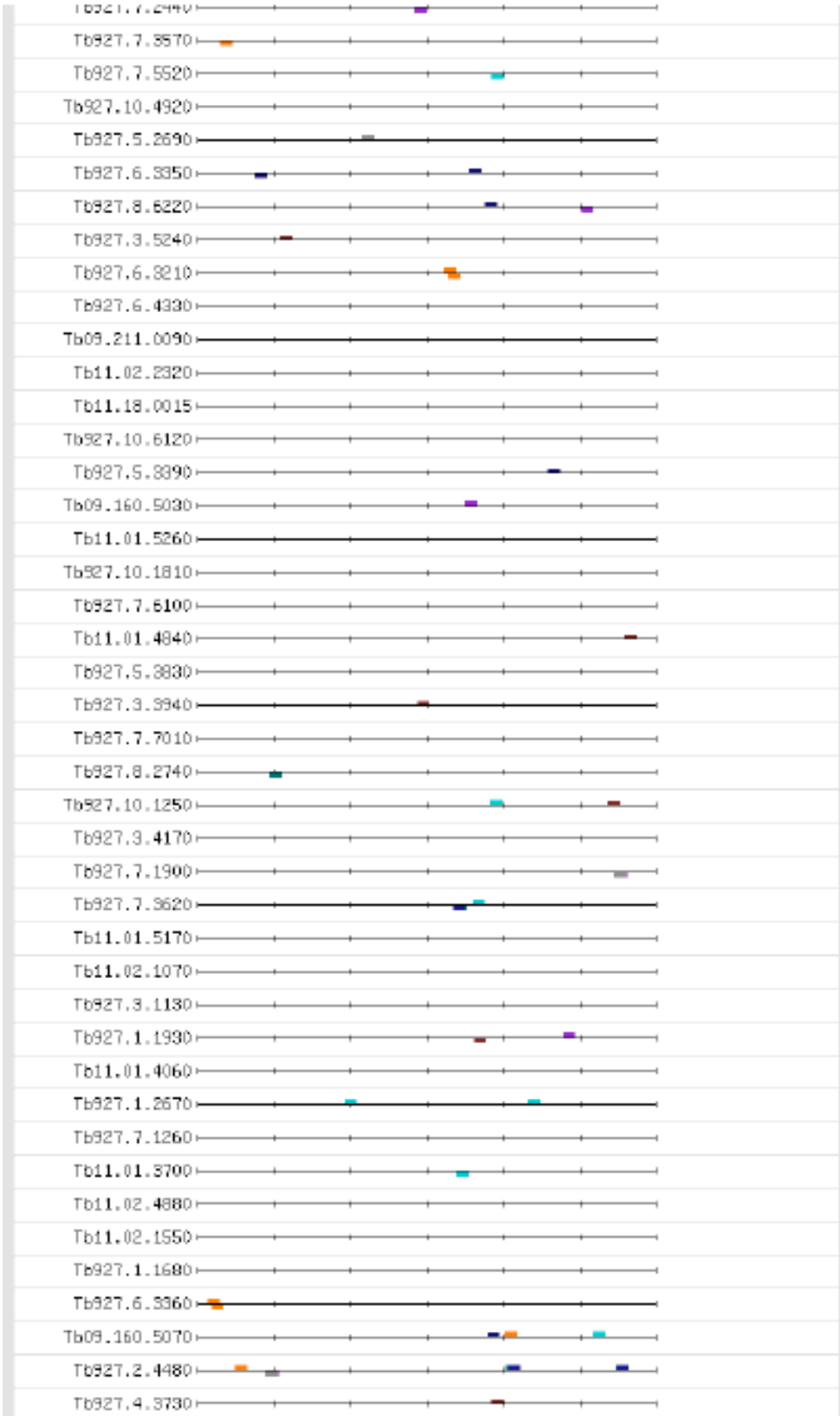

Tb11.03.0040
Tb927.7.2670
Tb927.7.350
Tb927.10.1860
Tb927.3.4130
Tb927.7.2070
Tb09.160.0850
Tb11.01.6130
Tb11.55.0009
Tb927.5.1480
Tb927.8.5040
Tb09.211.2840
Tb927.4.1960
Tb927.10.940
Tb11.02.0445
Tb927.10.6850
Tb927.5.1460
Tb927.10.4000
Tb927.3.3100
Tb09.160.4080
Tb927.10.11770
Tb11.02.3710
Tb927.4.3110
Tb927.10.5610
Tb927.10.10880
Tb927.2.5020
Tb927.4.3900
Tb927.7.4510
Tb927.2.4150
Tb927.7.5330
Tb927.10.2170
Tb927.2.6100
Tb11.02.1430
Tb09.211.0410
Tb927.7.6330
Tb927.3.2930
Tb09.160.4830
Tb09.211.1970
Tb927.1.3390
Tb927.4.3130
Tb927.10.1870
Tb11.01.6500

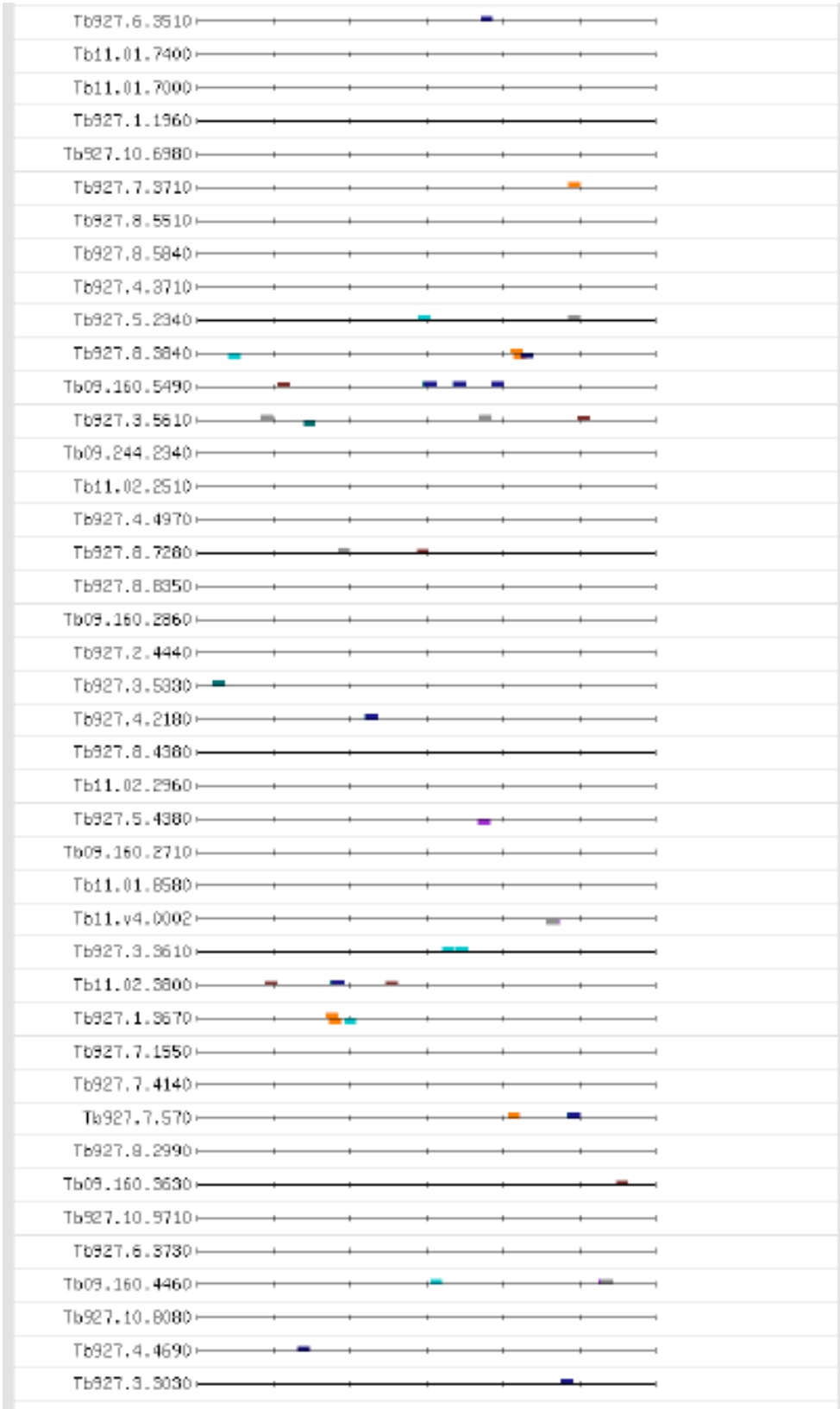

Tb927.2.2520
Tb927.8.8050
Tb927.7.5750
Tb927.10.6710
Tb927.2.2440
Tb927.4.850
Tb927.8.3110
Tb11.02.0480
Tb927.3.1340
Tb927.8.2010
Tb11.01.0210
Tb927.1.4020
Tb927.6.2510
Tb11.01.1320
Tb927.8.7010
Tb09.211.1250
Tb11.55.0013
Tb927.3.1990
Tb927.6.870
Tb927.10.12320
Tb927.7.2590
Tb927.8.1850
Tb11.02.0865
Tb927.10.14690
Tb927.10.5490
Tb927.7.2810
Tb11.02.4830
Tb927.10.5670
Tb927.2.2540
Tb927.6.3950
Tb927.8.5860
Tb927.8.810
Tb927.7.3780
Tb11.02.4000
Tb927.6.3690
Tb11.01.3770
Tb927.1.3920
Tb927.10.2890
Tb11.03.0110
Tb927.2.5660
Tb927.5.780
Tb09.160.3930

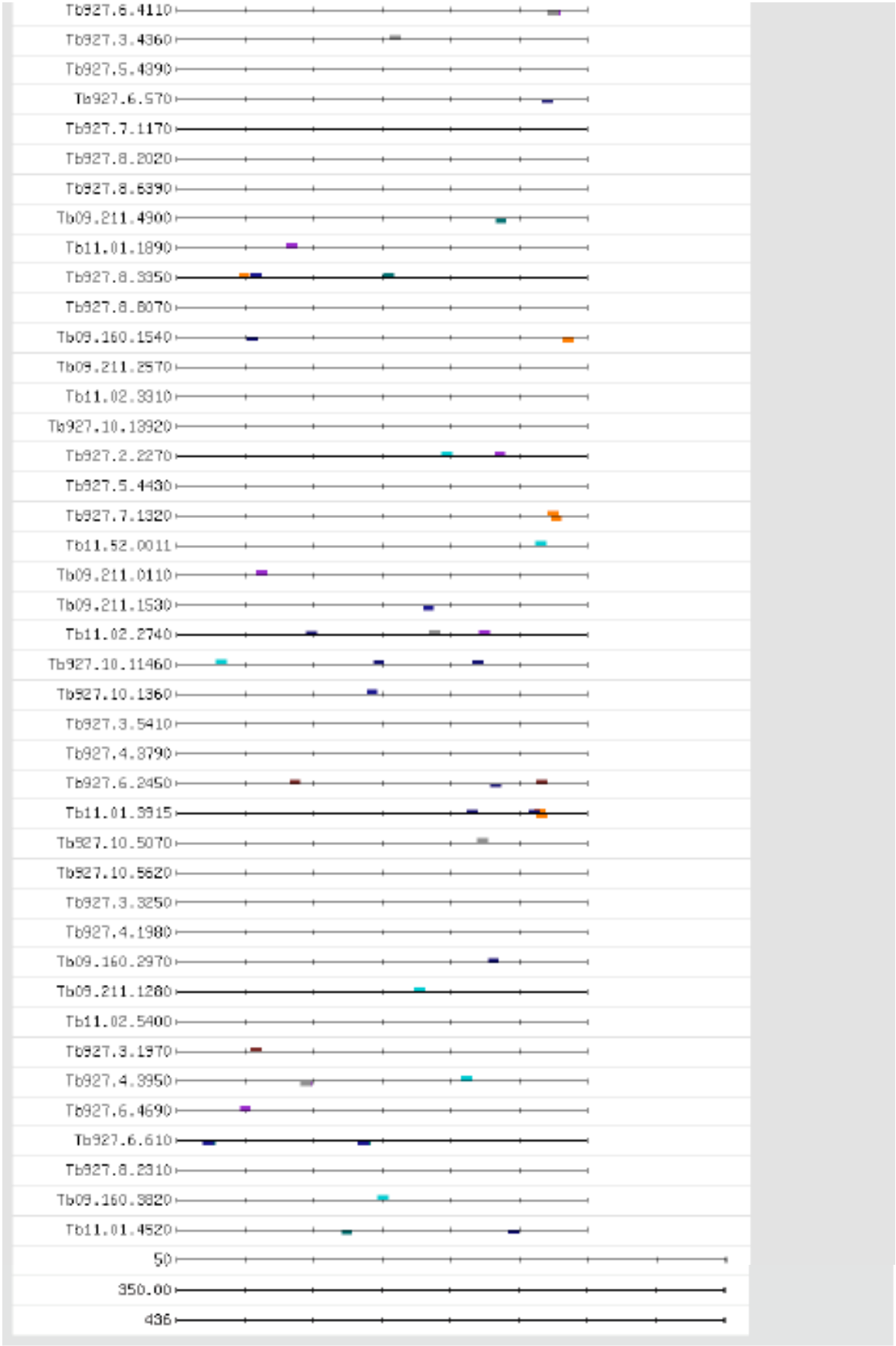

Tb09.211.1050
Tb927.10.9650
Tb927.2.5830
Tb09.160.4220
Tb927.7.1080
Tb09.211.1210
Tb11.02.0510
Tb927.5.1080
Tb927.4.2670
Tb11.01.1880
Tb927.5.1610
Tb11.02.4800
Tb927.6.4560
Tb11.01.3030
Tb11.01.3110
Tb11.02.2990
Tb11.55.0006
Tb927.3.3990
Tb927.2.3970
Tb927.5.2860
Tb927.10.2800
Tb927.10.330
Tb927.7.1470
Tb927.7.2640
Tb927.3.5050
Tb09.160.4270
Tb927.10.12960
Tb11.57.0003
Tb927.8.6890
Tb11.02.3230
Tb927.4.2620
Tb927.7.4850
Tb11.01.5990
Tb11.03.0320
Tb927.3.4990
Tb927.4.380
Tb09.160.5560
Tb11.01.7710
Tb11.01.1230
Tb11.01.5350
Tb927.8.3930
Tb927.10.6030
Tb11.22.0008


Tb927.6.1680
Tb11.02.4340
Tb09.160.5170
Tb11.01.1960
Tb11.01.2340
Tb927.1.4470
Tb927.2.3880
Tb927.3.5090
Tb927.4.5200
Tb927.8.5660
Tb09.160.4440
Tb09.211.3290
Tb11.01.7310
Tb927.1.1810
Tb09.211.4850
Tb927.2.5180
Tb927.10.11190
Tb927.10.1440
Tb927.5.4570
Tb927.10.14120
Tb927.4.980
Tb927.5.1820
Tb09.211.3980
Tb927.10.2630
Tb927.6.1510
Tb11.01.2450
Tb11.01.3060
Tb11.02.3400
Tb11.55.0008
Tb927.5.3660
Tb09.211.1380
Tb927.10.15070
Tb927.8.4480
Tb11.01.2210
Tb927.5.1670
Tb927.1.2985
Tb927.10.4740
Tb927.3.2340
Tb927.6.1970
Tb09.211.1655
Tb11.02.0355
Tb927.10.11660


Tb927.8.3300
Tb11.01.2700
Tb927.10.900
Tb927.8.4210
Tb11.01.0870
Tb927.7.1010
Tb11.01.1780
Tb11.01.4390
Tb11.01.6610
Tb11.02.5740
Tb09.244.2810
Tb11.03.0060
Tb927.1.680
Tb927.10.15830
Tb927.3.3230
Tb927.7.3190
Tb09.160.0910
Tb11.02.4770
Tb09.160.1320
Tb09.211.2430
Tb09.211.4450
Tb927.10.15350
Tb927.6.1570
Tb927.10.3130
Tb11.01.3745
Tb927.6.2420
Tb927.6.4140
Tb927.5.4420
Tb09.160.5160
Tb927.10.8200
Tb927.3.5140
Tb927.10.1310
Tb927.10.7910
Tb927.4.1170
Tb927.8.4690
Tb927.3.4240
Tb927.4.1270
Tb927.5.2620
Tb927.1.120
Tb927.5.580
Tb927.7.5640
Tb927.7.850


Tb927.7.3570
Tb927.7.5520
Tb927.10.4920
Tb927.5.2690
Tb927.6.3350
Tb927.8.6220
Tb927.3.5240
Tb927.6.3210
Tb927.6.4330
Tb09.211.0090
Tb11.02.2320
Tb11.18.0015
Tb927.10.6120
Tb927.5.3390
Tb09.160.5030
Tb11.01.5260
Tb927.10.1810
Tb927.7.6100
Tb11.01.4840
Tb927.5.3830
Tb927.3.3940
Tb927.7.7010
Tb927.8.2740
Tb927.10.1250
Tb927.3.4170
Tb927.7.1900
Tb927.7.3620
Tb11.01.5170
Tb11.02.1070
Tb927.3.1130
Tb927.1.1930
Tb11.01.4060
Tb927.1.2670
Tb927.7.1260
Tb11.01.5700
Tb11.02.4880
Tb11.02.1550
Tb927.1.1680
Tb927.6.3360
Tb09.160.5070
Tb927.2.4480
Tb927.4.3730


Tb927.6.3510
Tb11.01.7400
Tb11.01.7000
Tb927.1.1960
Tb927.10.6980
Tb927.7.3710
Tb927.8.5510
Tb927.8.5840
Tb927.4.3710
Tb927.5.2340
Tb927.8.3840
Tb09.160.5490
Tb927.3.5610
Tb09.244.2340
Tb11.02.2510
Tb927.4.4970
Tb927.8.7280
Tb927.8.8350
Tb09.160.2860
Tb927.2.4440
Tb927.3.5330
Tb927.4.2180
Tb927.8.4380
Tb11.02.2960
Tb927.5.4380
Tb09.160.2710
Tb11.01.8580
Tb11.v4.0002
Tb927.3.3610
Tb11.02.3800
Tb927.1.3670
Tb927.7.1550
Tb927.7.4140
Tb927.7.570
Tb927.8.2990
Tb09.160.3630
Tb927.10.9710
Tb927.6.3730
Tb09.160.4460
Tb927.10.8080
Tb927.4.4690
Tb927.3.3030


Tb927.6.4110
Tb927.3.4360
Tb927.5.4390
Tb927.6.570
Tb927.7.1170
Tb927.8.2020
Tb927.8.6390
Tb09.211.4900
Tb11.01.1890
Tb927.8.3350
Tb927.8.8070
Tb09.160.1540
Tb09.211.2570
Tb11.02.3310
Tb927.10.13920
Tb927.2.2270
Tb927.5.4430
Tb927.7.1320
Tb11.52.0011
Tb09.211.0110
Tb09.211.1530
Tb11.02.2740
Tb927.10.11460
Tb927.10.1360
Tb927.3.5410
Tb927.4.3790
Tb927.6.2450
Tb11.01.3915
Tb927.10.5070
Tb927.10.5620
Tb927.3.3250
Tb927.4.1980
Tb09.160.2970
Tb09.211.1280
Tb11.02.5400
Tb927.3.1970
Tb927.4.3950
Tb927.6.4690
Tb927.6.610
Tb927.8.2310
Tb09.160.3820
Tb11.01.4520
50
350.00
436