# Roles of R-loops in the *Trypanosoma brucei* Genome and Antigenic Variation

**Emma Marie Briggs**

BSc (Hons)

Submitted in fulfilment of the requirements for the Degree of Philosophy

Wellcome Centre for Molecular Parasitology,
Institute of Infection, Immunity and Inflammation,
College of Medical, Veterinary and Life Sciences

**University of Glasgow**

**September 2018**

# Abstract

The genome of the eukaryotic parasite *Trypanosoma brucei* is both dynamic and unconventional in several aspects. In comparison with other eukaryotic genomes, where the majority of protein coding genes are associated with their own transcriptional promoters, *T. brucei* transcribes almost all protein-coding genes polycistronically. Transcription initiates from broad regions that lack defined promoter sequences and RNA Polymerase II then traverses up to hundreds of genes, generating a pre-mRNA that then requires *trans*-splicing and polyadenylation to generate mature mRNAs. Termination of transcription, via virtually unknown processes, occurs where two multigene transcription units converges or, in some cases, adjacent to a downstream transcription initiation site. RNA Polymerase II transcribes the majority of protein-coding genes in this manner, negating any differential gene expression via transcriptional control. A further unusual aspect of the genome is the dedication of as much as a third of the coding capacity to elements of antigenic variation. When infecting the mammalian host, parasites express a dense protein coat of variant surface glycoprotein (VSG). In order to evade host immune elements, *T. brucei* switches expression to antigenically distinct VSGs, employing a repertoire of ~2,000 genes. Both transcriptional and recombination-based strategies enable the parasite to either switch transcription between ~15 expression sites, each housing a distinct VSG, or relocate VSG sequence from silent gene arrays into an active VSG expression site. Although multiple factors have been found to regulate these processes, the events which trigger a VSG switch by either pathway are unclear.

R-loops are three stranded structures containing an RNA-DNA hybrid and displaced single-stranded DNA. Although potentially deleterious to genome integrity, R-loops have been linked to transcription initiation and termination, DNA replication and recombination events. In this study, the potential for R-loop involvement in these fundamental genome functions of *T. brucei* was investigated. Firstly, Ribonuclease (RNase) H enzymes, which resolve the RNA-DNA hybrid portion of R-loops, were characterised, revealing *T. brucei* expresses potentially three distinct catalytic enzymes, two functioning in the nuclear genome and one in the kinetoplast

(mitochondrial) genome. Nuclear RNase H activity was depleted by null mutation or RNAi mediated knockdown of the nuclear RNase H enzymes, showing that while one RNase H, TbRH1, is non-essential, loss of the other, TbRH2, caused several growth and genome integrity defects. As it was hypothesised to increased levels of RNA-DNA hybrids of the genome, RNA-DNA hybrids were mapped in wild type parasites and those lacking RNases H using a specific antiserum, S9.6. This mapping identified the conserved formation of R-loops at centromeres, retrotransposon-associated genes, rRNA and tRNA genes. R-loop enrichment was also uncovered at RNA Polymerase II transcription start sites, as documented in mammalian genomes. DNA damage was specifically increased at these sites after TbRH2 depletion, indicating efficient resolution of these transcription initiation-associated R-loops is critical for genome maintenance. In contrast, R-loops were not associated with DNA replication or transcription termination suggesting RNA-DNA hybrids are not involved in these processes in *T. brucei*. The most abundant sites of R-loop enrichment were found to be at the nucleosome depleted regions located between the coding regions of polycistronically transcribed genes and are associated with polyadenylation and *trans*-splicing, highlighting a novel correlation of R-loops with pre-mRNA processing. Lastly, R-loops were mapped to VSG expression sites where their abundance increased after ablation of RNase H activity, an effect that was associated with both increased DNA damage and VSG switching, uncovering an R-loop-driven mechanism of antigenic variation.

# Table of Contents

# List of Tables

# List of Figures

# Acknowledgements

First, I would like to thank Richard McCulloch for all of his encouragement and enthusiasm during my project. I have truly enjoyed completing my PhD under his guidance and I am hugely grateful for all the opportunities I have had while working in his lab, especially those that involved travelling to Brazil. I would also like to thank my second supervisor Kathryn Crouch and placement supervisor Graham Hamilton for the time they have both taken to teach me bioinformatics (and fix countless error messages). My assessors Tansy Hamilton and Lilach Sheiner have also given me extensive guidance during my PhD, which I sincerely thank them both for.

All past and present members of both the McCulloch lab group and others (Craig, Catarina, Marija, Dan, Vivi, Marcelo, Jeziel, Jenny, Andrea, Samantha, Helena, Mario, Sam, Leandro, Natalia, Carmen, Fernanda, Fernando, Danielle, Jack etc…) have made the GBCR a welcoming and hugely enjoyable place to work. I want to thank you all for teaching me, as well as sharing many enjoyable evenings outside the lab. I want to thank Craig for making me laugh and performing numerous library preps, Leandro for his imaging expertise and Lauren for her help during summer projects.

I want to say thank you to my girls Helena, Vivi and Samantha for being so much fun, and especially Andrea for being such a great friend and enjoying ridiculous amounts of food and cocktails with me.

Lastly, I would like to thank my family and friends outside of the lab for their constant encouragement, especially my mum for her unfailing support over the four years. And finally, I want to thank Scott, for pushing me off the sofa when I didn't want to do work and physically taking my laptop away when I didn't want to stop.

# Author's Declaration

I here declare that this thesis and the results herein presented are the result of my own work, except where otherwise stated and acknowledged. None of the results herein presented have be used previously to obtain a degree at any university.

## Emma Marie Briggs

# List of Abbreviations

| | |
|---|---|
| AGS | Aicardi-Goutières Syndrome |
| AID | activation-induced cytidine deaminase |
| AQR | helicase aquarius |
| ASF/SFS2 | splicing factor 2 |
| ATL | alternative telomere lengthening |
| ATM | ataxia telangiectasia mutated |
| ATR | Ataxia telangiectasia and Rad3-related protein |
| BES | bloodstream from expression site |
| BLAST | basic local alignment search tool |
| BSA | bolvine serum albium |
| BSD | blasticidin |
| BSF | bloodstream from |
| CGI | CpG island |
| ChIP | chromatin-immunnoprecipitation |
| Chk1 | check point kinase 1 |
| Chk2 | check point kinase 2 |
| CTD | C-terminal domain |
| D-loop | displacement-loop |
| DRIP | DNA-RNA hybrid immunoprecipitation |
| DSB | double-stranded DNA break |
| E-value | expected-value |
| EcRHI | *E. coli* RNase HI |
| ESC | embryonic stem cell |
| ESAG | expression site associated gene |
| FA | formaldehyde |
| FACT | factilitates chromatin transcription complex |
| FLC | floral repressor gene |
| G4 | G-quadruplex |
| GO | gene ontology |
| GPI | glycosylphosphatidylinositol |
| gRNA | guide RNA |
| HA | hemaglutinin |
| HAT | human African trypanosomiasis |
| HBD | hybrid binding domain |
| HP1γ | heterochromatin protein 1γ |
| HR | homologous recombination |
| HYG | hygromycin |
| Ig | immunoglobulin |
| IP | immunoprecipitation |
| K | kinetoplast |

| | |
|---|---|
| kDNA | kinetoplast DNA |
| KO | knockout |
| LOH | loss-of-heterozygosity |
| MRE11 | meiotic recombination 11 |
| mtDNA | mitochondrial DNA |
| MTS | mitochondrial targeting sequence |
| mVSG | metacyclic variant surface glycoprotein |
| N | nucleus |
| ncRNA | noncoding RNA |
| NEO | neomycin |
| NER | nucleotide excision repair |
| NUP-1 | nucleoporin-1 |
| ORC | origin recognition complex |
| ORF | open reading frame |
| ORI | origin of replication |
| PAF1 | RNA Polymerase II associated factor 1 |
| PAS | polyadenylation site |
| PBS | phosphate buffered saline |
| PCF | procyclic form |
| PCR | polymerase chain reaction |
| PCR2 | polycomb repressive complex 2 |
| PI | propidium iodide |
| PIC | protease inhibitor cocktail |
| PCNA | proliferating cell nuclear antigen |
| Pol | polymerase |
| poly(Y) | polypyrimidine |
| pre-mRNA | precursor-messenger RNA |
| PTU | polycistronically transcribed unit |
| qPCR | quantitative-polymerase chain reaction |
| RED | ribonucleotide excision defective |
| RER | ribonucleotide excision repair |
| RHS | retrotransposon hotspot |
| RLFS | R-loop forming sequence |
| RMI1 | RecQ mediated genome instability 1 |
| RNAi | RNA interference |
| RNAP | RNA polymerase |
| RNase H | ribonuclease H |
| RPA | replication protein A |
| RpD | (5')-RNA-DNA(3')/DNA |
| SAS | splice acceptor site |
| Sen1 | senataxin |
| seq | next-generation sequencing |
| SETX | senataxin |

| | |
|---|---|
| SIF | stumpy inducible factor |
| siRNA | small-interfering RNA |
| SL | splice leader |
| ssDNA | single-stranded DNA |
| SSR | strand switch region |
| TbRH1 | *T. brucei* ribonuclease H1 |
| TbRH2 | *T. brucei* ribonuclease H2 |
| TbRH3 | *T. brucei* ribonuclease H3 |
| TERRA | telomeric repeat-containing RNA |
| tet | tetracycline |
| TOPO3a | topoisomerase 3a |
| TSS | transcriptional start site |
| TTS | transcription termination site |
| U2AF1 | U2 snRNA auxiliary factor 1 |
| UMS | universal minicircle sequence |
| UMSBP | universal minicircle sequence binding protein |
| VIM | vimentin |
| VSG | Variant Surface Glycoprotein |
| WHO | World Health Organisation |
| WT | wild type |
| XPF | exoderma pigmentosum type F |
| XPG | exoderma pigmentosum type G |
| yH2A | Thr130 phophorylated H2A |

# 1 Introduction

## 1.1 *Trypanosoma brucei*

The eukaryotic parasite *Trypanosoma brucei* is a member of the protozoan *Trypanosomatidae* family within the *Kinetoplastida* order and is further divided into three subspecies: *T. brucei gambiense*, *T. brucei rhodesiense* and *T. brucei brucei*. Commonly referred to as kinetoplastids, members of the *Kinetoplastida* order are distinguished by possession of the kinetoplast, a specialised disk-like network of concatenated circular DNA (kDNA), comprising the mitochondrial genome (Lukes *et al.*, 2002). The *Trypanosomatida* are further identified as having only a single flagellum. In *T. brucei* the flagellum protrudes from the flagella pocket as an extension of the basal body (a cylindrical structure made up of microtubules), to which the kinetoplast is connected (Robinson and Gull, 1991). All *T. brucei* subspecies cause mammalian disease across sub-Saharan Africa where they are found within the limited geographical location of the tsetse fly vector which they also infect (Figure 1.1). Yet, each subspecies' epidemiology, virulence and pathogenicity vary considerably.

Human African trypanosomiasis (HAT) is classed as a neglected tropical disease by the World Health Organisation (WHO) and exists in two documented forms. The slow progressing form found in Western and Central Africa is caused by *T. b. gambiense*, whereas *T. b. rhodesiense* causes a much faster progressing disease in Eastern and Southern Africa (Simarro *et al.*, 2010) (Figure 1.1). In complete contrast, *T. b. brucei* is unable to infect humans, due to sensitivity to the trypanosome lytic factors 1 and 2 found in human serum, but can cause the animal trypanosomiasis disease, Nagana (Vanhollebeke and Pays, 2010). Although HAT is now a rare disease (only 2,804 cases recorded in 2015), cases are still reported in over 20 African countries, where it causes substantial morbidity in some rural communities (Büscher *et al.*, 2017). As only a limited arsenal of treatments and control tools are currently available, infection in these populations still causes suffering and poses the significant threat of epidemic outbreak (Franco *et al.*, 2014). Nagana, however, remains endemic to 40 African countries and causes a still greater economic burden; the high mortality rate caused by infection accounts for massive losses in milk and meat production across the continent (Swallow, 1999).

All *T. brucei* subspecies contribute in a minor capacity to Nagana in cattle along with other *Trypanosomatida*, most predominately *T. brucei's* close relatives *T. congolense* and *T. vivax* (Abebe, Gute and Simon, 2017). Due to the ease with which *T. b. brucei* can be cultured in a laboratory environment and the range of genetic tools available for its manipulation, this subspecies lends itself to molecular research as an ideal model in which to study trypanosome biology.

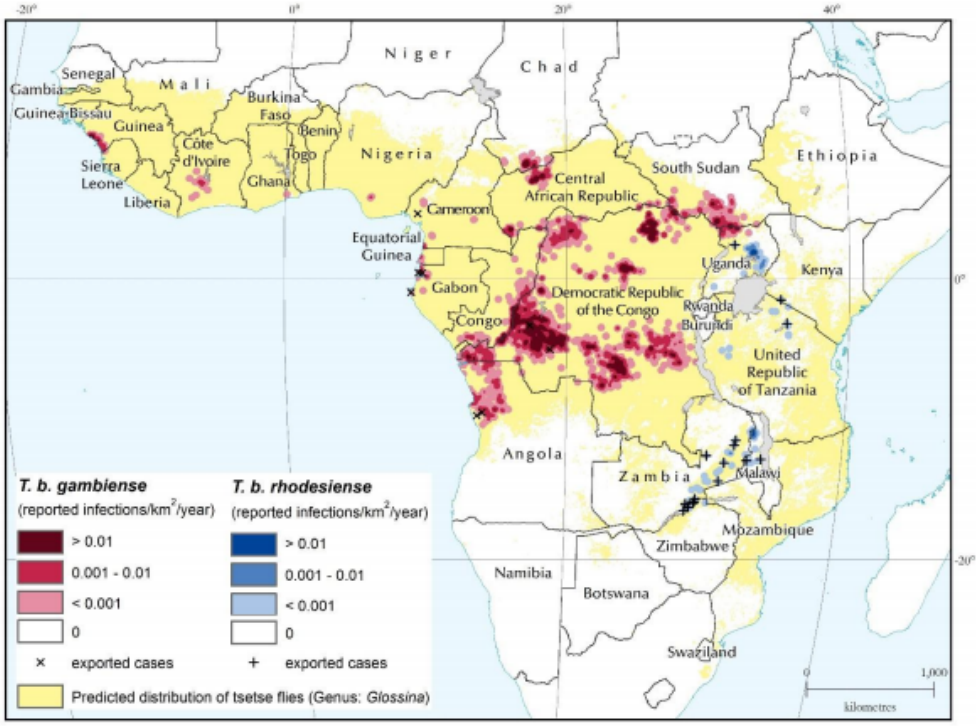



**Figure 1.1 Geographical distribution of human African trypanosomiasis cases reported between 2010 and 2014.**
The number of infections by *T. b. gambiense* and *T. b. rhodesiense* per km$^2$ per year are plotted as density in red and blue respectively, along with the predicted distribution of the tsetse fly in yellow. Plotted data was reported in the WHO atlas of HAT (Simarro *et al.*, 2010). Image source Büscher *et al.* (2017), license number 4410690838783.

### 1.1.1 The *T. brucei* life cycle

All *T. brucei* subspecies are transmitted between mammalian hosts and the blood sucking tsetse fly vector of the *Glossina* genus, although *T. b. gambiense* can additionally be transmitted congenitally (Rocha *et al.*, 2004; De Kyvon *et al.*, 2016). During the life cycle, the parasites undergo metabolic and morphological changes to allow specialisation of each distinct form to the disparate environments of the mammal and tsetse fly, with infective and proliferative forms of the parasite found

in at least three stages in the life cycle (Figure 1.2). Within infected tsetse fly salivary glands reside metacyclic form trypanosomes. These parasites are preadapted for mammalian infection, such as by expressing a Variant Surface Glycoprotein (VSG) coat (see below), and remain arrested in the cell cycle until injected into the mammalian host skin when the fly takes a blood-meal (Matthews and Gull, 1997). In the mammal, metacyclic forms differentiate into 'long-slender' bloodstream form (BSF) cells, which proliferate locally for several days before rapidly spreading via the blood and lymph circulatory systems.

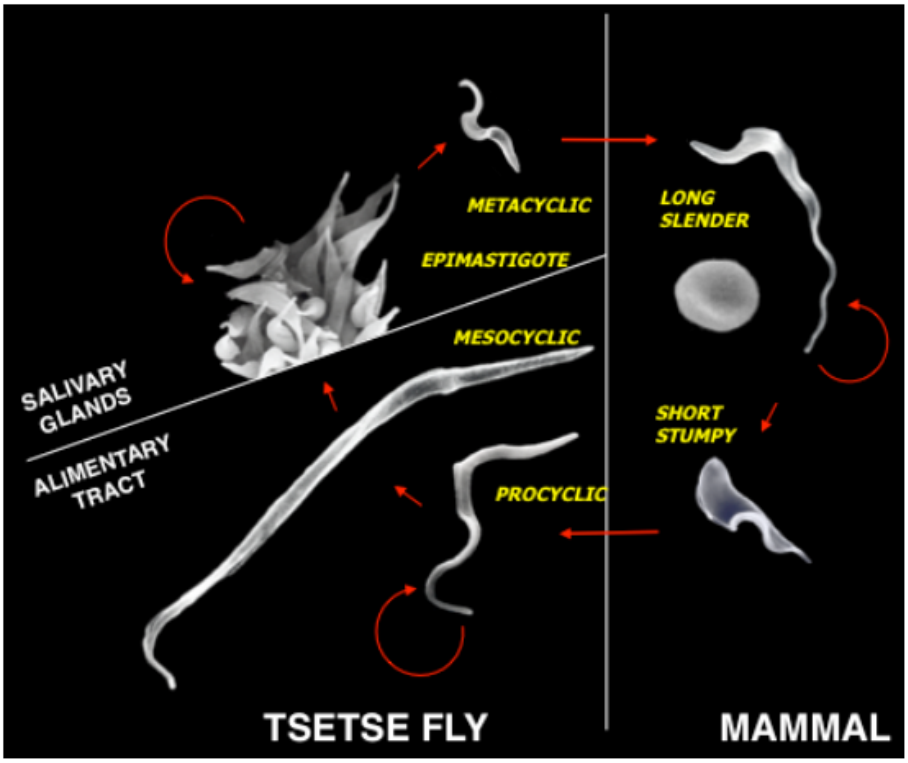



**Figure 1.2 The *T. brucei* parasitic life cycle.**
Scanning electron micrograph images of each life cycle stage are shown to scale. Hosts and environments within the tsetse fly host are labelled. Circular red arrows indicate replicative parasites while straight arrows show direction of cycle progression. Image source Barry and McCulloch (2001), license number 4410691284150.

As parasitaemia increases, quorum sensing of the elusive molecule(s) termed the Stumpy Inducible Factor (SIF) triggers transformation of long-slender BSF cells into 'short-stumpy' BSF parasites (Reuner *et al.*, 1997; Vassella *et al.*, 1997; Rico *et al.*, 2013). These cell cycle-arrested parasites must be ingested from either the blood or skin (Caljon *et al.*, 2016; Capewell *et al.*, 2016) by the insect vector if they are to survive (Turner, Aslam and Dye, 1995). If successfully transmitted, short-stumpy

BSF cells pass into the midgut of the tsetse fly where they differentiate into replicative procyclic form (PCF) trypanosomes (Fenn and Matthews, 2007). A portion of the PCF parasites can transform into proventricular mesocyclic forms, which migrate to the salivary glands of the fly (Van Den Abbeele *et al.*, 1999). Here, transformation into replicative epimastigotes takes place before differentiation into mammalian infective metacyclic cells completes the life cycle (Van Den Abbeele *et al.*, 1999).

As *T. brucei* is strictly extracellular, the parasite has evolved a sophisticated mechanism of antigenic variation in order to evade the adaptive immune defences of the mammal. Metacyclic parasites are coated in metacyclic VSG (mVSG) dimers, which provide initial protection from the mammalian immune system upon transmission (Tetley *et al.*, 1987; Turner *et al.*, 1988). Once differentiated into long-slender BSFs the mVSG coat is replaced by BSF VSG, here referred to simply as VSG. VSG dimers are highly immunological and, although they shield underlying invariant membrane proteins from innate immune attack (Bartossek *et al.*, 2017), the dense coat elicits a specific antibody response (Guirnalda *et al.*, 2007). However, a huge repertoire of antigenically distinct VSGs is available to the BSF parasites, which switch between the proteins to prevent the accumulation of effective immunological memory in the host immune system. The details of this complex process are discussed in detail in section 1.3.

### 1.1.2 *T. brucei* cell structure and cell cycle

*T. brucei* BSF cells are highly polarised, with microtubules running under the membrane from the anterior to the posterior of the cell body to form the cytoskeleton structure (Figure 1.3; reviewed in Matthews 2005). The microtubules open at the posterior end to form the flagella pocket, from which the flagellum extends and attaches laterally along the side of the cell body with only the distal end free (Wheeler et al. 2013). Many of the cell organelles and structures, including the flagellum, flagella pocket, kinetoplast and nucleus, are single copy and specifically positioned within the organised cell body (Figure 1.3).

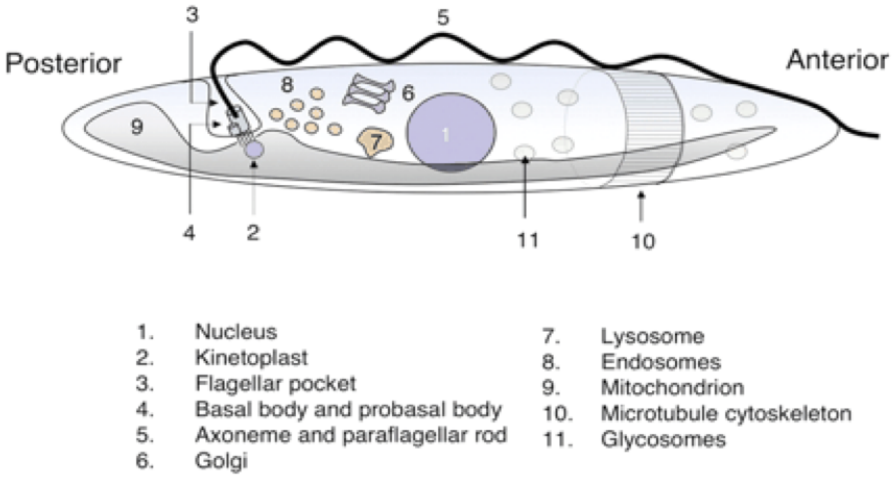



**Figure 1.3 The cellular structure of bloodstream form *T. brucei*.**
A simplified diagram of BSF parasite morphology. The microtubule cytoskeleton (10) that is found beneath the cell membrane is presented in the anterior end. Image source (Matthews, 2005).

Commencing synthesis of a new flagellum by duplication of the basal body, marks the initiation of the cell cycle (Woodward and Gull, 1990; Wheeler *et al.*, 2013). This event is immediately followed by duplication of the Golgi complex (Ho *et al.*, 2006). Two coordinated, temporally overlapping S phase events then take place (Figure 1.4). kDNA replication initiates first, before nuclear DNA replication commences as the kDNA comes to the end of S phase and enters the G2 stage (Figure 1.4; Benz *et al.*, 2017). The newly formed kinetoplast network is segregated from the old as the duplicated basal bodies are separated via microtubule movement, during which nuclear G2 phase takes place (Liu *et al.*, 2005; Jensen and Englund, 2012). At this stage cells possess one nucleus (1N) and two kinetoplasts (2K), before mitosis divides the duplicated nuclear genome, resulting in cells with two nuclei and two kinetoplasts (2N2K) (Ogbadoyi *et al.*, 2000). After nuclear mitosis, cytokinesis commences as a cleavage furrow forms from the anterior to the posterior between the old and new flagella that develops continuously during the previous stages (Wheeler *et al.*, 2013). Finally, abscission and remodelling of the new cells' posteriors completes the cell cycle (Figure 1.4).

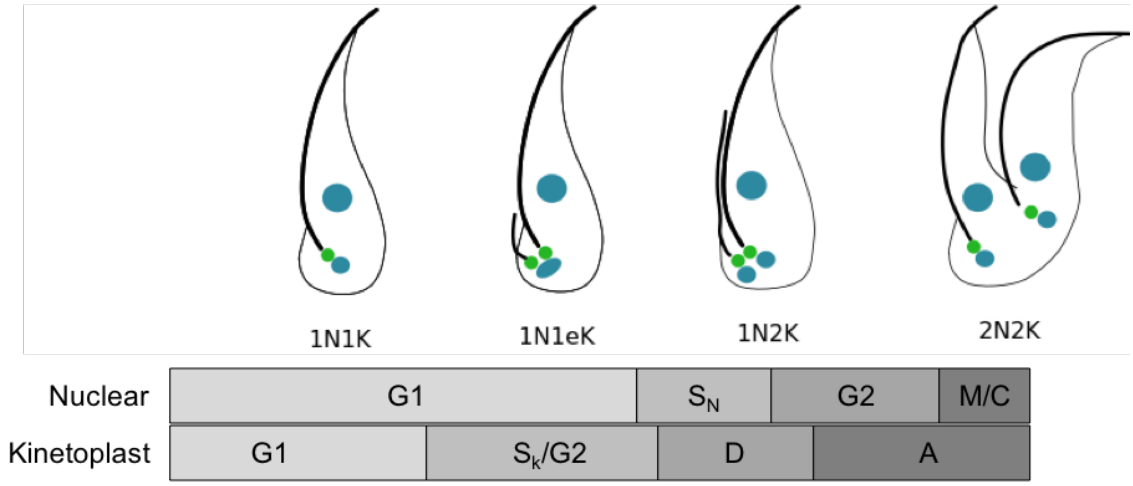



**Figure 1.4 Model of the *T. brucei* cell cycle.**
Events of the *T. brucei* cell cycle are shown as a simplified schematic above the relative timing of coordinated nuclear DNA and kDNA duplication. The nuclei (large blue circles), kinetoplasts (small blue circles), basal bodies (small green circles) and flagella (black) are shown. In the G1 phases of both the nuclear and kinetoplast cycles, the cells possess 1 nucleus and 1 kinetoplast (1N1K), before kDNA synthesis and kinetoplast division generates 1N1eK (elongated kinetoplast) then 1N2K cells. Before the end of $S_k$/G2, synthesis of the nuclear genome ensues and continues while the kinetoplast divides. Nuclear G2 follows before mitosis and cytokinesis complete the cycle. G1 – gap phase 1, $S_k$ – kDNA synthesis, $S_N$ – nuclear DNA synthesis, G2 – gap phase 2, M – mitosis, D – kinetoplast division, C – cytokinesis, A – after kinetoplast division. Recreated from (Benz *et al.*, 2017), under the Creative Commons Attribution 4.0 International License.

### 1.1.3 Human African trypanosomiasis

The clinical features of HAT depend greatly on whether disease is caused by *T. b. gambiense* or *T. b. rhodesiense* infection, host response, disease stage and subspecies strain (Beckers *et al.*, 1981; MacLean *et al.*, 2010). Infection of either subspecies is nearly always fatal if untreated, though instances of healthy carriers or self-cure have been documented in the case of *T. b. gambiense* HAT (Jamonneau *et al.*, 2012). With either subspecies disease progresses in two stages: the first, hemo-lymphatic stage (blood and lymph infection) is followed by a second, meningo-encephalitic stage where parasites cross the blood-brain barrier and invade the central nervous system. Symptoms of the first stage predominantly consist of a long-lasting intermittent fever, headaches, itching and lymphadenopathy (abnormal lymph nodes). Neuopsychiatric symptoms, including the characteristic sleeping disorder that elicited the name *sleeping sickness*, add to these when the second stage is reached. The slowly progressing form of HAT caused by *T. b. gambiense* tends to be chronic with a mean duration of around 3 years (Checchi *et al.*, 2008). However, acute infection by *T. b. rhodesiense*

progresses from the initial stage within a few weeks and leads to death within 6 months (Odiit, Kansiime and Enyaru, 1997).

Treatment options depend on the causative subspecies and stage of disease: pentamidine and suramin are available for first stage treatment, and melarsoprol, eflornithine and nifurtimox for the second (Keating *et al.*, 2015). As first stage drugs are unlikely to effectively treat a second stage infection, and those for the later stage are too toxic to justify use in the first, the stage must first be determined by lumbar puncture to examine parasite load in the cerebral-spinal fluid. Several rapid diagnostic tests have been developed for *T. b. rhodesiense* infection as the characteristic high levels of parasitaemia allow identification by microscopy examination of lymph or concentrated blood samples, but no field-applicable tests are yet available for *T. b. gambiense* HAT (reviewed by the WHO in Büscher et al. 2017).

Pentamidine remains the first-line treatment for *T. b. gambiense* HAT, and an alternative for *T. b. rhodesiense* HAT for which suramin is the first choice. *T. b. gambiense* infection has been effectively treated with pentamidine for decades (95-98% effective) and is well tolerated. Suramin is effective against both subspecies but is more complex to administer and can cause some adverse reactions (Simarro *et al.*, 2012). Second stage treatment induces more adverse reactions (abdominal pain, vomiting and headache) than either first stage drug, but treatment against *T. b. gambiense* HAT with nifurtimox-eflornithine combination therapy is better tolerated and has higher cure rates (95-98%) than monotherapy with melarsoprol or eflornithine (Priotto *et al.*, 2009; Franco *et al.*, 2012). Use of melarsoprol is restricted to treatment of second stage *T. b. rhodesiense* HAT due to the high frequency of severe and life-threatening adverse reactions, the most severe of which is encephalopathic syndrome that occurs in 5-18% of treated patients and is fatal in 10-70% of cases (Lutje, Seixas and Kennedy, 2010; Büscher *et al.*, 2017).

Cases of drug resistance to melarsoprol and pentamidine have been documented (Graf *et al.*, 2013; Munday *et al.*, 2014; Fairlamb and Horn, 2018), but the introduction of nifurtimox-eflornithine combination therapy has decreased the

probability of resistance emergence (Büscher *et al.*, 2017). Additionally, two new drugs are under clinical development: fexinidazole (Torreele *et al.*, 2010) and benzoxaborole (Robert T Jacobs *et al.*, 2011). Importantly, both are easily administered orally, target both *T. b. gambiense* and *T. b. rhodesiense* and are intended against both stages of HAT (Torreele *et al.*, 2010; Robert T. Jacobs *et al.*, 2011).

### 1.1.4 Animal African trypanosomiasis/Nagana

Cases of Nagana have been documented in a vast range of domesticated and wild animals including pigs, sheep, goats, horses, ruminants, lions and leopards, nonhuman primates and some rodents. Cattle infections remain problematic as the high mortality rates due to the associated wasting lead to an estimated loss in productivity of up to 20% across a range of factors, including meat and milk production (Swallow, 1999). Nagana is somewhat controlled via selective bush clearing to remove vegetation relied upon by the vector, employing insecticide use and in some cases breeding of naturally resistant cattle (reviewed in, Meyer et al. 2016). However, meaningful and standardised evalution of these control operations is required in order to inform the implementation of future control programmes (Meyer *et al.*, 2016).

Particularly in areas outside of vector control operations, a handful of trypanocides are widely employed in cattle, with diminazene aceturate and isometamidium used most heavily. Threateningly, resistance to these drugs has been increasing, with between 35 and 70 million doses of trypanocides used across sub-Saharan Africa annually (Holmes, 2013; Giordani *et al.*, 2016). Despite high demand, there has been a lack of interest from pharmaceutical companies due to their anticipation of low profits, and so novel licensed compounds are unlikely to be available for several years (Giordani *et al.*, 2016). However, recent investment from the Bill and Melinda Gates Foundation and the UK Department for International Development (via the Global Alliance for Livestock Veterinary Medicines, GALVmed), hopes to aid the development of new anti-trypanosome therapies, which will be essential if the burden of Nagana is to be tackled.

# 1.2 The unconventional genome of *T. brucei*

## 1.2.1 Genome composition

The ~26 Mb *T. brucei* genome consists of 11 megabase-sized chromosomes, ranging from ~1-5.2 Mb, an unspecific number of mini and intermediate-sized chromosomes (30-700 kb) and a circular kDNA mitochondrial genome (Melville *et al.*, 1998; Wickstead, Ersfeld and Gull, 2004; Berriman, 2005).

The kDNA genome is composed of interlocking circles of two types: a few dozen maxicircles, which encode typical mitochondria genes, and thousands of minicircles, which encode guide RNAs (gRNAs) (Lukes *et al.*, 2002). The gRNAs store information needed to restore a viable coding sequence to the maxicircle transcripts, which require uridine insertion or deletion before being translated (Hajduk and Ochsenreiter, 2010; Aphasizhev and Aphasizheva, 2011). Maxicircles range in size from 20 to 40 kb, whereas minicircles are only 0.5 to 10 kb in length, encoding 50-60 nt long gRNAs. Although all kinetoplastids have minicircles and maxicircles, the dense, inter-locked, disk-like structure that the kDNA forms in *T. brucei* appears unique to trypanosomatids (Marande, Lukes and Burger, 2005).

There are also several unusual characteristics of the *T. brucei* nuclear genome compared to other eukaryotic genomes. The most striking of these distinct features is the organisation of nearly all protein-coding genes into polycistronically transcribed units (PTUs), in which tens to hundreds of genes are transcribed from the same strand in the same direction (see review, Campbell et al. 2003). Maturation of mRNA requires co-transcriptional processing of the pre-mRNA into 5' capped and 3' polyadenylated mRNA molecules through coupled *trans*-splicing and polyadenylation (see below). Although genes within each PTU can be differentially regulated, all are transcribed from shared transcriptional start sites (TSSs), explicitly limiting any means of transcriptional control via the activation or repression of transcription (Clayton, 2014, 2016). Another unusual feature is the presence of a modified DNA base, termed base J (ß-D-glucosyl-

hydroxymethyluracil), which is associated with the ends of PTUs and repetitive DNA regions such as telomeric sequences (Maree and Patterton, 2014).

PTUs make up the actively transcribed core of the Mb chromosomes and are distinct from the highly variable subtelomeric regions that consist of mainly transcriptionally silent gene arrays (Berriman, 2005). Where the cores of the 11 Mb chromosomes are conserved and diploid, the subtelomeres can differ considerably in composition between chromosome homologues and so could be considered aneuploid (Berriman, 2005).

### 1.2.2 Transcription

As stated above, transcription of the core *T. brucei* genome occurs almost exclusively in a polycistronic manner (Johnson, Kooter and Borst, 1987; Berriman, 2005). This transcription is largely carried out by RNA Polymerase (Pol) II, whereas RNA Pol I and RNA Pol III transcribe specific loci, although transcription by these Pols is also often polycistronic, such as in the case of the RNA Pol I transcribed VSG expression sites (see below). Multiple protein-coding genes are transcribed as a PTU and the resulting pre-mRNA transcripts are processed in *trans* by the addition of a splice leader (SL) RNA 5' cap, and polyadenylation of the 3' end to generate mature mRNA (detailed in section 1.2.2.3; Figure 1.5). Adjacent PTUs meet at regions termed strand switch regions (SSRs), where the units either converge (transcription termination sites; TTSs) or diverge (TSSs). At several sites over the Mb chromosomes transcription can terminate and re-initiate on the same strand (Siegel *et al.*, 2009; Wright, Siegel and Cross, 2010). Although no true strand switch takes place, these regions are termed head-to-tail SSRs for continuity. Short PTUs have also been identified in nematodes (Spieth *et al.*, 1993), and a few cases of dicistronic units have been observed in *Drosophila* (Brogna and Ashburner, 1997) and humans (Lee, 1991). However, the near universal use of polycistronic transcription appears to be unique to the kinetoplastids (Jackson *et al.*, 2016) and raises several questions concerning the control of gene expression in *T. brucei* and its relatives.

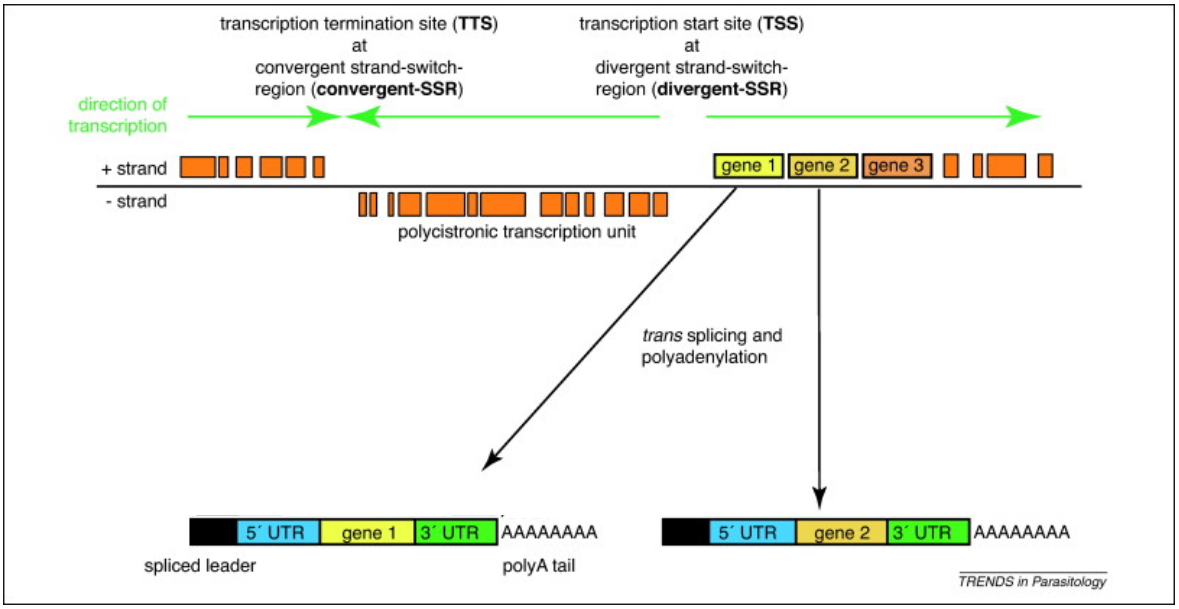



**Figure 1.5 Polycistronic transcription and mRNA maturation in *T. brucei*.**
Transcription by RNA Pol II takes place on either the sense (+) or anti-sense (-) strand, here starting at divergent SSRs. Transcription in opposite directions converges to terminate at convergent SSRs. *Trans*-splicing and polyadenylation dissect pre-mRNA transcripts and generates mature mRNA. Image source Siegel *et al.* (2011), license number 4410700809096.

### 1.2.2.1 RNA Polymerase II transcription initiation

Nuclear run-on assays performed in *Leishmania major*, a tryapnosomatid parasite and close relative of *T. brucei*, first inspired the proposal that RNA Pol II transcription initiates at divergent SSRs and continued until termination at convergent SSRs (Martínez-Calvillo *et al.*, 2003, 2004). Later, genome-wide analysis via Chromatin-Immunoprecipitation coupled to next-generation DNA sequencing (ChIP-seq) revealed the enrichment of multiple histone modifications and histone variants at SSRs (Siegel *et al.*, 2009; Wright, Siegel and Cross, 2010). H4K10ac (histone 4 lysine 10 acetylation) and H3K4me3 (histone 3 lysine 4 trimethylation), along with histone variants H2A.Z and H2B.V, were found to be enriched up to 300-fold at divergent SSRs, indicating their association with transcription initiation in *T. brucei* (Siegel *et al.*, 2009; Wright, Siegel and Cross, 2010). These factors were also found to be enriched within some PTUs, suggesting the presence of head-to-tail SSRs at these sites (Siegel *et al.*, 2009; Wright, Siegel and Cross, 2010). These, along with divergent SSRs, were experimentally confirmed as TSSs using a sequencing technique that enriches for RNA possessing a 5' triphosphate, which only newly transcribed RNA possesses before the addition of a 5' cap (Kolev *et al.*,

2010). Mapping of histone H3 acetylation in *Leishmania major* revealed acetylation of the histone at divergent SSRs as well as towards the chromosome ends and within PTUs, highlighting that the core genome polycistronic structure is conserved between *Trypanosoma* and *Leishmania* parasites (Thomas *et al.*, 2009).

Although these studies have successfully identified broad TSS regions, the mechanisms governing transcription initiation by RNA Pol II are still largely unclear. With the exception of the SL RNA promoter (Günzl *et al.*, 1997), no classical promoter sequences, which in other eukaryotes provide a platform for transcription machinery binding, have been identified in *T. brucei*. However, GT-rich motifs have been found to promote deposition of histone variant H2A.Z to the boundaries of the nucleosome depleted regions found across *T. brucei* divergent (and head-to-tail) SSRs, presumably ensuring a region of relaxed chromatin upstream of the broad TSS (Wedel *et al.*, 2017). This GT-rich motif-directed deposition of H2A.Z is sufficient for transcription initiation in *T. brucei* (Wedel *et al.*, 2017). These RNA Pol II promoters hence appear similar to "dispersed" promoters observed in other eukaryotic genomes that lack both typical motifs, such TATA boxes, and defined TSSs (Juven-Gershon *et al.*, 2008; Lenhard, Sandelin and Carninci, 2012). Instead, they most commonly contain CpG islands, house a broad region where transcription may start, and are associated with constitutively expressed house-keeping genes (Juven-Gershon *et al.*, 2008; Lenhard, Sandelin and Carninci, 2012). In common with divergent SSRs and the initiating portion of head-to-tails SSRs, nucleosomes are well positioned downstream of, and depleted upstream of, the TSSs associated with dispersed promoters in metazoan genomes (Rach *et al.*, 2011; Wedel *et al.*, 2017). Histone variant H2A.Z is also deposited at the flanking sites of nucleosome depleted regions across TSSs of dispersed promoters in metazoans (Rach *et al.*, 2011), as observed at *T. brucei* TSSs (Wedel *et al.*, 2017). These characteristics are in contrast to differentially regulated "focused" promoters, which contain typical motifs, a defined site of initiation and are associated with less organised chromatin (Rach *et al.*, 2011). Hence, RNA Pol II initiation sites in *T. brucei* may be mechanistically similar to dispersed promoter regions (Wedel *et al.*, 2017), in keeping with a lack of transcriptional control over gene expression in the parasite.

#### 1.2.2.2 Transcription termination

Histone variants H3.V and H4.V have been shown to be enriched at TTSs in *T. brucei* (Siegel *et al.*, 2009), as has base J, which flanks PTUs (Cliffe *et al.*, 2010). Depletion of base J leads to transcriptional read through at sites of termination in the *Leishmania* genome (van Luenen *et al.*, 2012; Reynolds *et al.*, 2014). Contrastingly, depletion of base J in *T. brucei* does not lead to complete read through, although specific sites proximal to normal base J deposition do display minor read through (Reynolds *et al.*, 2014). A combination of base J depletion and H3.V deletion does, however, cause antisense transcription of genes downstream of the TTSs in *T. brucei* (Schulz *et al.*, 2016). Interestingly, H3.V has been shown to promote transcriptional termination prior to the end of a PTU to regulate the transcription of specific genes, in particular small-interfering RNA (siRNA) producing loci, suggesting some genes are preferentially located to the ends of PTUs (Reynolds *et al.*, 2016).

#### 1.2.2.3 *Trans*-splicing and polyadenylation

Maturation of the pre-mRNA transcripts is achieved first by *trans*-splicing, which adds a 39 nt SL sequence of heavily modified RNA. The SL RNA substrate is first transcribed independently by RNA Pol II at a locus separate from the primary transcript before 8 methylation events occur across the first 4 nts, generating a "cap 4" structure, which includes the conserved 7-methylguanylate cap ($m^7G$) (Perry, Watkins and Agabian, 1987). nt 28 of the SL RNA is additionally pseudouridylated ($\psi_{28}$) (Liang, Xu and Michaeli, 2002), although this modification is not required for *trans*-splicing (Sturm, Fleischmann and Campbell, 1998; Mandelboim *et al.*, 2002). Only two *T. brucei* genes, encoding a putative poly(A) polymerase (Tb927.3.3160) and an ATP-dependent DEAD/H RNA helicase (Tb927.8.1510), have been found to undergo *cis*-splicing (the process of removing an internal fragment from a transcript that is fundamental to gene regulation in mammals) (Mair *et al.*, 2000; Berriman, 2005; Siegel *et al.*, 2010). Therefore, the vast majority of *T. brucei* protein-coding genes do not have a conventional intron/exon structure. Multiple parallels do, however, exist between *cis*-splicing

and *trans*-splicing: the requirement for a polypyrimidine (poly(Y)) tract (Huang and Van, 1991), an AG dinucleotide 3' splice acceptor site (SAS) where the SL sequence is fused, and formation of intermediate Y structures (Murphy, Watkins and Agabian, 1986).

SASs have been mapped in *T. brucei* by retrieving RNA sequencing reads which span the SAS-coding region junction and removing the SL sequence before aligning the remaining sequence to the genome (Kolev *et al.*, 2010; Nilsson *et al.*, 2010; Siegel *et al.*, 2010). Combined efforts have now mapped >32,000 unique SASs in >8,900 genes in PCF and BSF *T. brucei* (Kolev *et al.*, 2010; Siegel *et al.*, 2010). Most *T. brucei* genes have between 1 and 3 available SASs, and the alternative use of these has been observed in 676 genes when comparing long-slender BSF, short-stumpy BSF and PCF cells (Nilsson *et al.*, 2010).

Polyadenylation is strictly dependent on the downstream *trans*-splicing events (López-Estraño, Tschudi and Ullu, 1998). In fact, the same poly(Y) motif influences the efficiency of polyadenylation of one gene and splicing of the immediately downstream gene (Matthews, Tschudi and Ullu, 1994). An average of 10 polyadenylation sites (PASs) have been mapped per gene in *T. brucei* (Nilsson *et al.*, 2010; Siegel *et al.*, 2010). Binding of poly(A)-binding proteins (PABPs) to poly(A) tails stabilises mRNA transcripts by preventing the deadenylation-dependent mRNA decay pathway, which is initiated by removal of the poly(A) tail and leads to rapid cleavage of the 5' cap (Wilusz, Wormington and Peltz, 2001).

### 1.2.3 DNA replication

Like other eukaryotes, initiation of DNA replication in *T. brucei* depends upon the assembly of the origin recognition complex (ORC) at discrete sites in the genome, termed origins or ORIs. In humans, ORC is made up of five ATPase subunits, ORC1-5, and a sixth unrelated protein, ORC6 (Duncker, Chesnokov and McConkey, 2009). Five subunits of the divergent *T. brucei* ORC have been identified to date; one ORC1-like factor named ORC1B, a highly divergent ORC4-like protein, two even less homologous proteins, Tb7980 and Tb3120, and OCR1/CDC6, a factor which appears

homologous to both ORC1 and CDC6 (Godoy *et al.*, 2009; Dang and Li, 2011; Tiengwe *et al.*, 2012). Indeed, OCR1/CDC6, ORC4 and MCM3 (a DNA replicative helicase) have been found to form part of a high-molecular-weight complex (~530-1011 kDa), implying *T. brucei* does possess a functional ORC (Marques *et al.*, 2016). ORC binding licenses ORIs in the G1 phase through Cdc6-mediated recruitment of MCM, but only some of these primed ORIs are activated upon entering S phase. In *T. brucei*, ORC1B is only found in the nucleus of S phase cells, perhaps indicating a diverged strategy for ORI licensing or activation.(Marques *et al.*, 2016).

#### 1.2.3.1 Origins of replication

In most eukaryotic genomes multiple ORIs initiate replication, although not all ORIs that are licensed for replication are activated when the cells enter S phase. Although the role of ORC appears well conserved among eukaryotes the means by which the complex recognises ORIs is not. For example, ORC binds *S. cerevisiae* ORIs in a sequence-specific manor (Marahrens and Stillman, 1992), and via non-specific AT-rich sequences in both *S. pombe* and *Drosophila* (Kong and DePamphilis, 2001; Vashee *et al.*, 2003). In the human genome the most efficiently activated ORIs are associated with CpG island promoters (Lombraña *et al.*, 2015). The molecular details of this close relationship between replication and transcription initiation are still uncertain, but analysis revealed most human and mouse ORIs contain G-rich repeat elements with the potential to form G-quadruplex (G4) structures (Besnard et al 2012; Cayrou et al 2012). Interestingly, human ORC1 has been shown to efficiently bind G4 structures present in either RNA or single-stranded DNA (ssDNA), both of which are present as transcription takes place, *in vitro* (Hoshina et al 2013).

Replication and transcription processes are also closely link in the *T. brucei* genome, as the ORC subunit ORC1/CDC6 was found to bind to potentially all SSRs, a subset of which were found to be activated as ORIs in the S phase of wild type (WT) cells (Tiengwe *et al.*, 2012). Of these ORIs, 19 were located at divergent SSRs, 3 at convergent SSRs and 18 at head-to-tail SSRs. In all cases replication appears to occur in both directions, with an equal rate of movement in both directions at

convergent and divergent SSRs. At head-to-tail region ORIs, however, replication appears to be faster when occurring in the same direction as transcription. Thus, it can be hypothesised that the *T. brucei* genome has evolved to minimise head-on collisions between the replication and transcription machineries. Further supporting the mechanistic links between replication and transcription, RNAi-mediated depletion of ORC1/CDC6 causes changes in mRNA abundance proximal to SSRs (Tiengwe *et al.*, 2012). In addition, the single actively transcribed VSG expression site (see below) has been shown to be replicated earlier in S phase than all other transcriptionally silent sites, each of which are replicated very late (Devlin et al, 2016).

## 1.3 Antigenic variation in *T. brucei*

The selective forces placed on pathogens by the complex, adaptive mammalian immune system has led to the independent evolution of antigenic variation in many species. All aim to evade the defences of the immunocompetent host and often share common features: monoallelic expression of a single antigen, an archive of silent antigen genes of the same 'family', and antigen expression switching strategies which occur above background levels of recombination or mutation. For example, *Borrelia sp* possess a specialised *vls* locus, which contains an active *vls* surface antigen expression site, *vlsE*, as well as a silent array of *vls* cassettes, and relies on recombination events to move diverse silent *vls* genes into the *vlsE* (Norris, 2014). *Neisseria gonorrhoeae* and *N. meningitidis* also rely on recombination to relocate silent *pilin* genes into the expression loci, *pilE* (Hagblom *et al.*, 1985). *Plasmodium falciparum* similarly employs antigenic variation and is able to recombine ~60 *var* genes during meiosis. However, antigenic variation amongst *var* genes is controlled by epigenetic transcription-driven switching, where the promoter of the active *var* gene is silenced and an alternative *var* gene promoter is activated (Kyes, Kraemer and Smith, 2007). *T. brucei* antigenic variation is impressively complex and employs transcriptional switching as well as various recombination strategies.

BSF *T. brucei* parasites are coated in ~5 x $10^6$ VSG antigen dimers, which comprise 5-20% of each cell's total protein (Cross, 1975). Each protein is approximately 60 kDa in size and is attached to the cellular membrane via a conserved glycosylphosphatidylinositol (GPI) anchor (Ferguson *et al.*, 1988). The resulting dense 'coat' is believed to act as a physical barrier, blocking immune elements from binding other surface proteins (Schwede *et al.*, 2011). However, this understanding has been challenged as some proteins have been shown to protrude above the VSG coat (Higgins *et al.*, 2013). Each BSF parasite expresses a single VSG at a time from one of approximately 15 specialised RNA Pol I-transcribed bloodstream expression site (BES) (Hertz-Fowler *et al.*, 2008). Expression of the VSG antigen is detected by the mammalian host defences and a specific immune attack to be mounted against the parasites (Guirnalda *et al.*, 2007). Antigenic variation ensues as the parasites switch expression to that of a different, antigenically distinct VSG, thus avoiding complete elimination of the infecting population.

### 1.3.1 Genomic elements associated with VSG switching

The ~15 BESs present in the genome are split between the subtelomeres of the 11 Mb chromosomes and the intermediate chromosomes (Figure 1.6). BESs have a generic structure consisting of a RNA Pol I promoter and a string of up to 13 expression site associated genes (ESAGs), separated from the telomere proximal VSG by a series of DNA sequence repeats termed the 70-bp repeats (Figure 1.6) (Hertz-Fowler *et al.*, 2008). Although RNA Pol I typically transcribes *rRNA* genes, BES promoters show limited homology to the *rRNA* promoter (Zomerdijk *et al.*, 1990), as well as the promoter associated with the procyclin genes, which encode the surface proteins of PCF *T. brucei* and are also transcribed by RNA Pol I (Günzl *et al.*, 2003). Each of these RNA Pol I promoters binds slightly divergent RNA Pol I associated transcription factors (Brandenburg *et al.*, 2007), and display different subnuclear localisations: whereas BES transcription occurs in a discrete site known as the expression site body (ESB) (Chaves *et al.*, 1998; Navarro and Gull, 2001), rRNA and procyclin transcription occurs in the nucleolus (Landeira and Navarro, 2007). Similar telomeric mVSG expression sites (ESs) are activated in metacyclic trypanosomes to express mVSGs. Although the promoter sequence here is again

divergent (Ginger *et al.*, 2002), the promoter is recognised by the same general transcription factors employed at BES promoters (Kolev, Günzl and Tschudi, 2017). However, mVSG ESs differ from BESs in lacking ESAGs, and thus are monocistrons that contain only VSGs.

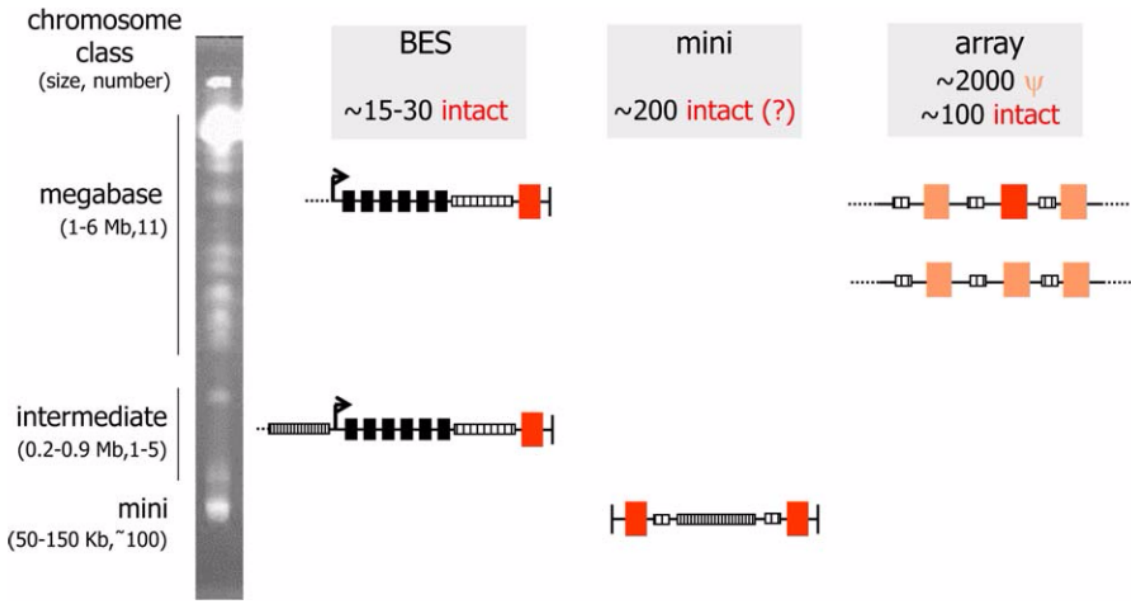



**Figure 1.6 The VSG archive of *T. brucei*.**
Pulsed field gel electrophoresis separated whole *T. brucei* chromosomes are shown to the left. Elements of the VSG archive are shown in line with the chromosomes on which they occur, including the BES found in the telomeric regions of Mb and intermediate chromosomes. Here, the RNA Pol I promoter (flag), ESAG coding regions (black boxes), 177-bp repeats (narrow lined box), 70-bp repeats (wide lined box) and the telomere (vertical line) positions are indicated. Peach boxes indicate pseudo VSGs (ψ) found in the arrays within the subtelomeres of the Mb chromosomes. Mini-chromosomes appear to house only intact VSG that are immediately adjacent to the telomere. Image source (McCulloch, Morrison and Hall, 2015).

Sequencing of different *T. brucei* strains has led to the number of genes in the VSG archive being estimated as between ~1,000 and ~2,500, which accounts for as much as 20% of the coding capacity of the genome (Berriman, 2005; Marcello and Barry, 2007; Cross, Kim and Wickstead, 2014). This bank of VSGs is spread between Mb chromosome subtelomeres, where they are found as arrays (Berriman et al 2005), telomere adjacent regions of the mini-chromosomes (Wickstead, Ersfeld and Gull, 2004) (Figure 1.6), and within telomere proximal BESs and MESs. 90% of VSGs, in all locations, are flanked by 70-bp repeats, although typically the number of repeats is highest within the BESs (Marcello and Barry, 2007). Interestingly, only ~5% of silent VSGs are predicted to be functionally intact, and most of the array VSGs are pseudogenes (Berriman, 2005; Marcello and Barry, 2007). Of these pseudogenes,

~65% appear to have frameshift mutations or premature stop codons in their sequence, 20% of the coding sequences are truncated, and ~10% are classed as atypical in that their predicted folding patterns or post-translation modifications differ to that of characterised VSGs (Marcello and Barry, 2007). The location of the VSGs in the subtelomeres appears to promote diversification of the repertoire, as Mb chromosome sizes differ considerably between isolates and most of this change is due to variation in the size of VSG arrays (Melville, Gerrard and Blackwell, 1999; Callejas *et al.*, 2006). In fact, differences in chromosome length extends to chromosomes homologues within a single cell (Melville, Gerrard and Blackwell, 1999), indicating the arrays are sites of considerable rearrangement.

### 1.3.2 Monoallelic VSG expression and *in situ* switching

Multiple BESs requires a layer of transcriptional control over VSG variant expression: only one BES is actively transcribed at a time, with all others largely silent. However, transcription of the active BES can be silenced, and another activated in an apparently co-ordinated reaction: so-called '*in* situ' VSG switching. As no change in the DNA sequence is needed, *in situ* VSG switching represents a case of transcriptionally controlled gene expression regulation not seen in the polycistronically transcribed core of the *T. brucei* genome. Although transcription of the BESs is dynamic, strict monoallelic expression of a single VSG is evident, since artificially forced selection for WT cells expressing two or more BESs results in inviable parasites, with any such expression being unstable (Chaves *et al.*, 1999; Ulbert, Chaves and Borst, 2002). Both transcription initiation and elongation have been suggested to be supressed in the silent BES (Vanhamme *et al.*, 2000; Nguyen *et al.*, 2014), reducing silent VSG transcript abundance to 10,000–fold below that of the active VSG (Yang *et al.*, 2009).

As well as transcriptional control, the environmental context of the BESs is critically important for monoallelic regulation of VSG expression and, potentially, *in situ* switching. Expression of the active VSG takes place in the ESB, which is geographically separated from the nucleolus (Navarro and Gull, 2001), though both subnuclear structures recruit RNA Pol I for transcription (Landeira and Navarro,

2007). Insertion of the distinct *rRNA* RNA Pol I promoter into a BES results in the divergent promoter adopting the transcriptional status of the BES in question (Horn and Cross, 1995; Rudenko *et al.*, 1995), consistent with features of the ESB exercising control over monoallelic BES expression. The ESB is present in all cell cycle stages of the long-slender BSF trypanosomes, but is lost upon differentiation to the short-stumpy stage, during which the active BES relocates to the nuclear periphery (Landeira and Navarro, 2007). Faithful duplication and segregation of the ESB during mitosis is also required to maintain monoallelic expression of the active BES, as knockdown of cohesin, which is required for accurate chromosome segregation, leads to transcriptional switching (Landeira *et al.*, 2009).

As well as maintenance of active BES expression in the ESB, the inactive BESs must be maintained in a silent state in order to prevent expression of more than one VSG per parasite. The nuclear lamina in mammals is a dense network of lamina filaments that controls nuclear organisation and so influences transcriptional processes (Kind and van Steensel, 2010). In the *T. brucei* nucleus the divergent repetitive protein nucleoporin-1 (NUP-1) performs the role of the lamina, most likely as a nuclear envelope fibre component (Rout and Field, 2001; DuBois *et al.*, 2012). Interestingly, knockdown of NUP-1 leads to derepression of silent BESs, suggesting the inactive BESs are organised into heterochromatin regions regulated by NUP-1 (DuBois *et al.*, 2012).

Studies of chromatin remodelling factors, histone chaperones and the histones themselves, further support the importance of chromatin structure in the transcriptional control of the BESs. Depletion of either the core histone H3 (Alsford and Horn, 2012), or 'linker' histone H1 (Povelones *et al.*, 2012) causes activation of silent BESs. Similar phenotypes are observed when targeting histone chaperones NLP (Narayanan *et al.*, 2011), FACT (Denninger *et al.*, 2010; Denninger and Rudenko, 2014), ASF1A or CAF-1b (Alsford and Horn, 2012). Depletion of the chromatin remodeller ISWI also derepresses the silent BES promoters, though the increase in resulting VSG expression remains significantly lower than that observed from the active VSG (Stanne *et al.*, 2011). Additionally, histone methyltransferase DOT1B, which generates the H379me3 modification, exhibits some influence over

BES repression, as depletion of the protein causes partial expression of previously inactive VSGs as well as slowing *in situ* switching rates (Janzen *et al.*, 2006; Figueiredo, Janzen and Cross, 2008). Lastly, the active BES is specifically depleted of nucleosomes (Figueiredo and Cross, 2010; Stanne and Rudenko, 2010), an effect regulated by the architectural chromatin protein TDP1 (Narayanan and Rudenko, 2013).

Although these findings appear to implicate typical epigenetic factors in the control of chromatin dynamics involved in VSG expression regulation, Glover *et al.* (2016) took steps towards better understanding if trypanosome-specific mechanisms are involved, through the discovery of VSG exclusion 1 (VEX1). VEX1 locates to the active BES where it is required for monoallelic expression, since RNAi leads to increased silent VSG expression (Glover *et al.*, 2016). Indeed, when VEX1 is overexpressed surplus protein binds to other BESs, which also results in expression of more than one VSG. Hence, VEX1 is required for VSG expression and its sequestration to the active BES is critical for monoallelic expression (Glover *et al.*, 2016).

The above discussion demonstrates the clear importance of chromatin structure and subnuclear localisation in the monoallelic expression of VSGs, although no studies to date have demonstrated how *in situ* switching is initiated, or if a clearly defined trigger does in fact exist. The 'winner-takes-all' mechanism proposed by Glover *et al.* (2016) implicates VEX1 as a key protein in the silencing of inactive BESs, but whether the protein actively takes part in an *in situ* switching pathway remains to be investigated.

### 1.3.3 VSG switching via recombination

In order for *T. brucei* to access the full repertoire of VSG genes and pseudogenes available, the parasite employs DNA recombination to translocate sequences into the active BES. Three recombinational VSG switching strategies have been described: VSG gene conversion, segmental gene conversion and, to a lesser extent, reciprocal VSG recombination (Figure 1.7). The latter process involves the exchange

of VSGs and associated chromosome ends between two BESs or between a BES and mini-chromosome, so that crossover occurs without sequence loss (Pays *et al.*, 1985), and the reaction is limited to telomere located VSGs. Gene conversion involves the copying of the whole VSG open reading frame (ORF), as well as some up and downstream sequence, into the BES, displacing the resident VSG. The upstream 70-bp repeats provide homology for recombination (Liu *et al.*, 1983), but the region of copied DNA has been shown to extend beyond the repeats to include some of the ESAGs (Pays *et al.*, 1983; McCulloch, Rudenko and Borst, 1997) and even the promoter (Hertz-Fowler *et al.*, 2008). The region of 3' DNA copied during conversion can terminate in the coding or non-coding sequence of the VSG (Bernards *et al.*, 1981), or as far as the chromosome end in the case of telomeric VSGs (De Lange *et al.*, 1983; Kim and Cross, 2010). Segmental gene conversion is distinct from the other mechanisms, as it instead pieces together segments of VSG sequence to form 'mosaic' VSGs with functional and novel sequences (Roth *et al.*, 1989; Thon, Baltz and Eisen, 1989; Thon *et al.*, 1990). This means the parasite is able to make use of the predicted non-functional VSG pseudogenes, exploiting inter-sequence homology between VSG ORFs, as well as hugely amplifying the number of antigenically distinct VSGs available for expression.

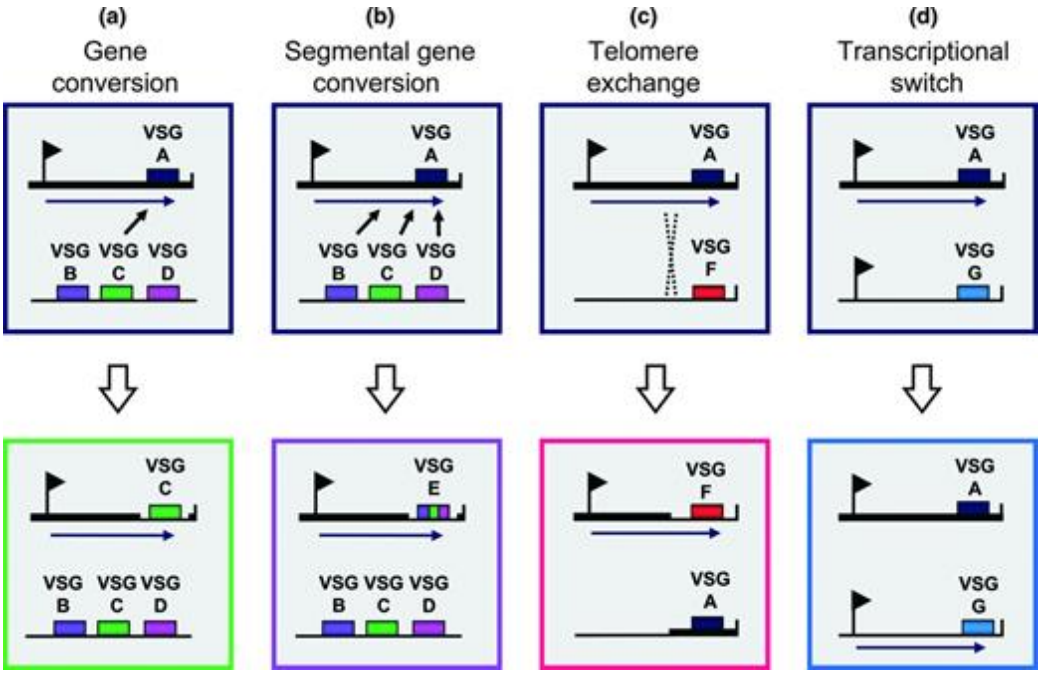



**Figure 1.7 The various mechanism of VSG switching in *T. brucei*.**

a) Gene conversion involves copying the complete sequence of a VSG found elsewhere in the archive into the active BES. b) Segmental gene conversion instead recombines sequences from several VSGs to generate a mosaic VSG gene in the active BES. c) Telomeric exchange occurs when cross-over of chromosome ends, including the telomeres, takes place. d) Transcriptional *in situ* switching does not reply on recombination and instead a switch of active transcription from one BES to another takes place. Black flags indicate promoter positions, blue horizontal arrows indicate active transcription of the site and black arrows show movement of VSG sequence. Dashed lines indicate region of cross-over. Image source Vink, Rudenko and Seifert (2012), license number 4410710021799.

#### 1.3.3.1 The role of homologous recombination in VSG switching

Homologous recombination (HR) facilitates the repair of double-stranded DNA breaks (DSBs) using a related stretch of DNA sequence located elsewhere in the genome, as a template for repair (Figure 1.8). DNA synthesis during the repair process results in the exchange of template DNA sequence from one location to another, and is hypothesised to do so during VSG switching events (McCulloch, Morrison and Hall, 2015). A key HR factor, RAD51, coats 3' ssDNA (generated by 5'-3'resection of DNA strands at the DSB site), forming a nucleoprotein filament that invades the DNA duplex of homologous template DNA, resulting in the formation of a displacement (D)-loop. Once RAD51, facilitated by other proteins including BRCA2 and Rad51 paralogues, has catalysed this strand invasion, break-pair can commence. Break-pair can proceed via several routes, although this occurs most commonly via gene conversion, during which both 3' ssDNA tails can participate in strand invasion (Chen *et al.*, 2007). Firstly, the ssDNA strands of the resulting D-loop are extended via DNA synthesis, having been primed by base hybridisation to the template DNA during strand invasion, and the ssDNA tails pair. Ligation then joins the paired tails to form two Holliday junction structures. Random cleavage of the Holliday junctions by resolvase enzymes results in either crossover, where recombinant chromosomes are formed, or non-crossover events (Chen *et al.*, 2007). During a non-crossover event, retention of the short stretch of DNA synthesised during D-loop extension results in the uni-directional exchange of sequence from the template strand into the chromosome where the DSB occurred (Figure 1.8). Alternatively, dissolution of Holliday junctions by convergent migration can also lead to non-crossover events and gene conversion (Chen *et al.*, 2007).

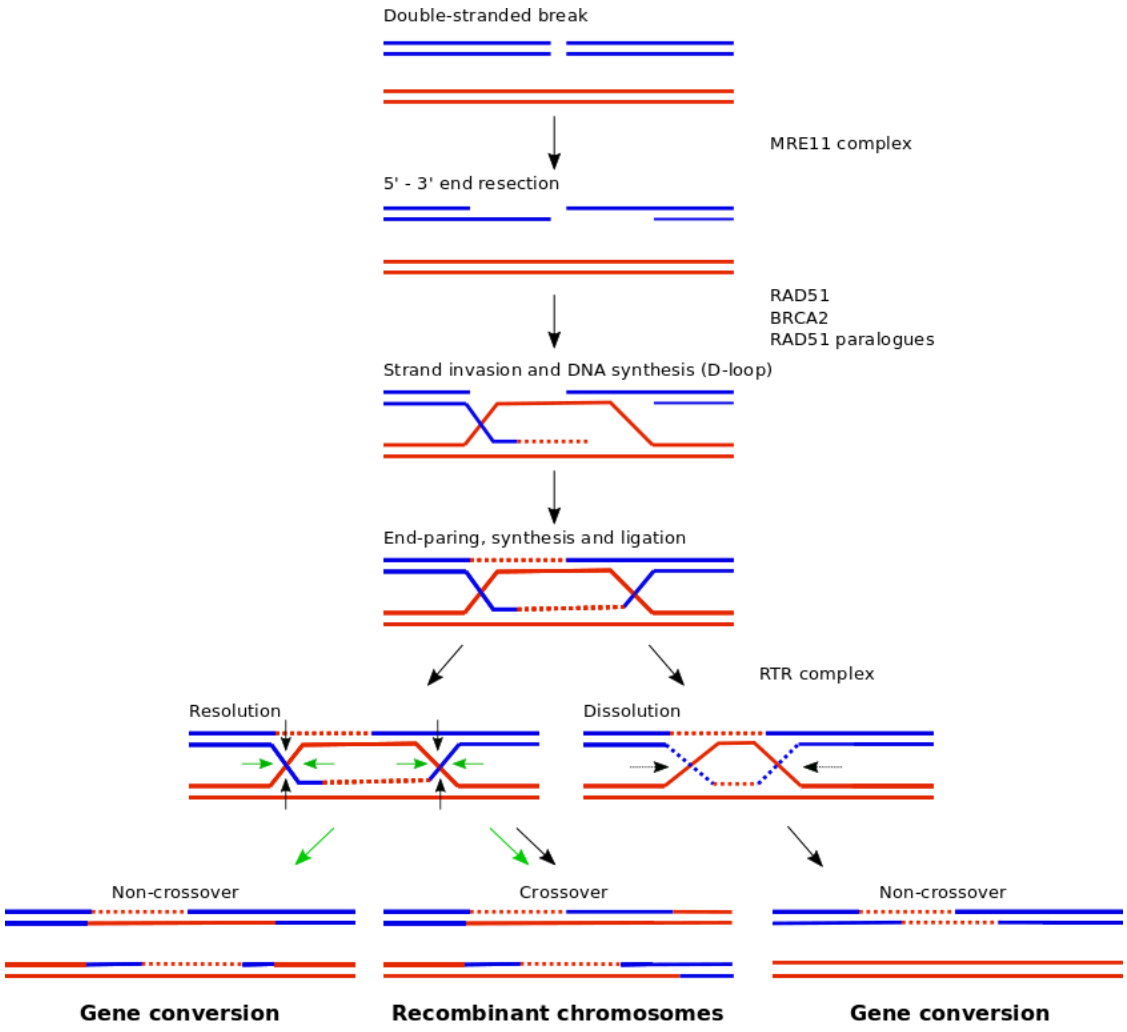



**Figure 1.8 Model of double-stranded break repair pathways.**
The steps of double-stranded break (DSB) repair by Holliday junction resolution and dissolution are depicted along with discussed factors involved in each step. The 5' ends of a DSB are first resected by the action of the MRE11 complex (MRE11, RAD50 and Xrs2/NBS1) to generate 3' ssDNA tails. Strand invasion is then facilitated by RAD51, RAD51 paralogues and BRCA2 to form a displacement (D)-loop, which is extended by DNA synthesis. End-pairing of the ssDNA tails and ligation then form two Holliday junction structures. The RTR complex (RecQ/Sgs1-Top3/TOPO3a-Rmi1/BLAP75/18) enables dissolution of the Holliday junctions by facilitating convergent branch migration (black arrows), resulting in non-crossover events. Alternatively, resolution of the junctions is performed by a resolvase enzyme leading to either crossover (cut with green arrows at both junctions) or non-crossover (cut with green arrows at one junction and black at the other) events. Hence, resolution can lead to either recombinant chromosomes or gene conversion, whereas dissolution by the RTR complex causes gene conversion exclusively. Recreated from (Chen *et al.*, 2007), license number 4413841438071.

Null mutation of RAD51 in *T. brucei* results in a 10-fold reduction in VSG switching rate, including via gene conversion, indicating HR does have a role in VSG switching (McCulloch and Barry, 1999). In a similar manner, mutation of the *T. brucei* homologue of BRCA2 also leads to a reduction in VSG switching, to a similar extent to that seen after loss of RAD51 (Hartley and McCulloch, 2008). Perplexingly, in both cases VSG switching is not abolished and the relative use of recombination and

*in situ* switching is not changed, suggesting two things: some recombination-based switching must occur via a route independent of RAD51 and BRCA2 activity, and transcriptional switching may also be suppressed in the absence of these HR factors. Four Rad51 paralogues exist in *T. brucei* and all function in DNA recombination and repair, but mutation of just one, RAD51-3, has been found to cause significant impairment of VSG switching (Proudfoot and McCulloch, 2005; Dobson *et al.*, 2011).

In contrast to the impaired VSG switching seen in the above mutants, mutation of likely components of the RTR complex (RecQ/Sgs1-Top3/TOPO3a-Rmi1/BLAP75/18) in *T. brucei* results in increased rates of VSG switching (Kim and Cross, 2010, 2011; Devlin *et al.*, 2016). The RTR complex mediates dissolution of Holliday junctions (Figure 1.8) and is known to suppress the occurrence of recombinant chromosomes, as dissolution exclusively results in non-crossover, gene conversion events during DSB break-repair (Wu and Hickson, 2003). Mutation of the RecQ helicase (which pushes the two Holliday junctions towards one another) curiously results in increased levels of VSG switching via gene conversion (Devlin *et al.*, 2016), as does mutation of either Topo3a, the DNA decatenase that resolves the linked DNA strands (Kim and Cross, 2010), or the stabilising protein RMI1 (Kim and Cross, 2011). These findings are somewhat surprising considering that mutation of these factors would be hypothesised to suppress gene conversion events, as their function is to promote non-crossover resolution. It is plausible, however, that mutation of the RTR complex results in less stable Holliday junctions and increased gene conversion via resolution, or an as yet un-investigated pathway.

Mutation of MRE11, another HR factor that acts as part of a complex (MRE11, RAD50 and Xrs2/NBS1) involved in DSB processing, again leads to unexpected findings; *T. brucei* MRE11 mutants display no changes in VSG expression despite high levels of chromosome arrangements due to loss of array VSGs (Robinson *et al.*, 2002; Tan, Leal and Cross, 2002). The MRE11 complex detects DSBs and can act as the initiator for break-repair via HR by performing resection of the DNA strands to generate 3' ssDNA tails, or by recruiting other long-range nucleases to do so (Paull and Deshpande, 2014). Hence, if it is assumed that DSBs act to initiate VSG switching, it

might have been hypothesised that mutation of MRE11 would result in reduced recombination and a lower rate of VSG switching (Robinson *et al.*, 2002; Tan, Leal and Cross, 2002). Though the effects of MRE11 mutation cannot rule out a role for DSBs, execution of VSG switching can clearly occur independently of this complex, or other factors can readily assume its roles. Although a clear link between HR and VSG switching has been demonstrated, mutation of any one HR-associated factor results in only modest changes in switching rates (McCulloch, Morrison and Hall, 2015). How VSG switching may occur independently of RAD51-catalysed HR remains unclear, but one possibility is that at least some of the switching that is observed in HR mutants is through segmental gene conversion, a reaction that has not been dissected genetically. Additionally, microhomology mediated end-joining (MMEJ) has been shown to repair DSBs in *T. brucei* and can occur without the action of RAD51 resulting in deletions at the break site (Conway *et al.*, 2002; Glover, McCulloch and Horn, 2008; Glover, Jun and Horn, 2011). During this process 5' DNA stand resection occurs via the MRE11 complex as in HR, however, as little as 5 – 20 bp of homology is required for the exposed 3' DNA ends to anneal and ligation to take place (Glover, Jun and Horn, 2011). Indeed, MMEJ events were found to contribute to gene conversion events, although, this was only documented to occur at one end of a duplicated sequence and RAD51-dependent HR was required at the other (Glover, Jun and Horn, 2011).

#### 1.3.3.2 70-bp repeats

As noted above, the vast majority of VSGs (~90%) are associated with 70-bp repeats (Marcello and Barry, 2007), although with considerable size variation (Shah *et al.*, 1987). Analysis of the 70-bp repeats located in the BESs revealed the repeat sequence is an average of 76 bp in length and largely AT-rich (78%), although two GC-rich regions were found to be conserved within the repeat sequences (Hovel-Miner *et al.*, 2016). Interestingly, the AT-rich triplet repeat element of the 70-bp repeats has been found to have the potential to become non-H bonded (Ohshima *et al.*, 1996), and may promote recombination, as documented in bacterial plasmids (Pan *et al.*, 2010). Artificial induction of DSBs adjacent to the 70-bp repeats of the active BES can induce VSG switching, an effect not seen after DSB induction when

the 70-bp repeats have been deleted from the BES, suggesting they may act in triggering recombination (Boothroyd *et al.*, 2009). However, VSG switching has been documented in cells lines were the 70-bp repeats have been deleted from the BES, albeit at low levels (McCulloch, Rudenko and Borst, 1997), and the same 70-bp-independent VSG switching was reported in the absence of DSB induction by Boothroyd et al, 2009. More recently, deletion of the 70-bp repeats in the active BES was shown to increase VSG switching before and after artificial DSB induction (Hovel-Miner et al, 2016), despite the previous findings that 70-bp repeat deletion represses switching induced by artificial DSBs (Boothroyd *et al.*, 2009). These data perhaps suggest the role of the 70-bp repeats in VSG switching cannot be over simplified, and most likely extends beyond simply providing homology for VSG recombination (Liu *et al.*, 1983). As discussed above, the boundaries of recombination are frequently seen upstream of the 70-bp repeats (Pays *et al.*, 1983; Proudfoot and McCulloch, 2005) and even encompassing the BES promoter (Hertz-Fowler *et al.*, 2008).

#### 1.3.3.3 Telomeres and VSG switching

Telomere length has also been proposed as the determining factor for VSG switching events (Dreesen, Li and Cross, 2007; Hovel-Miner *et al.*, 2012). *T. brucei* cells with critically short telomeres, revealed by mutation of telomerase, have been shown to have an increased rate of VSG switching compared to those with longer telomeres (Hovel-Miner *et al.*, 2012), supporting this notion. How recombination is promoted in parasites with critically short telomeres remains unclear, however, as ISceI-mediated removal of telomeric repeats failed to increase switching rate (Glover, Alsford and Horn, 2013). Interestingly, after loss of telomere length in telomerase deficient cells, re-expression of the enzyme results in the telomere adjacent to the active BES extending at a higher rate than those associated with silent BES, indicating active transcription influences telomere maintenance, perhaps by enhancing recombination events (Dreesen and Cross, 2006). More recently, depletion of *T. brucei* RAP1, which is known to be critical for telomere stability in mammals (Martinez *et al.*, 2010), has been shown to results in increased levels of telomeric repeat-containing RNA (TERRA), telomeric RNA-DNA hybrids,

DSBs and VSG switching (Nanavaty *et al.*, 2017). Hence, it is likely that telomere instability can cause VSG switching via recombination, but whether this is able to drive antigenic variation in unperturbed WT parasites remains untested.

## 1.4 R-loops

R-loops are three stranded nucleic acid structures formed of a double-stranded RNA-DNA hybrid and displaced ssDNA. The sequence specific, DNA-binding capabilities of RNA were first demonstrated by detection of radiolabelled rRNA hybridised to extrachromosomal single-stranded *rDNA* in *Xenopus* oocytes using autoradiography (Gall and Pardue, 1969). rRNA hybridisation to denatured, double-stranded *rDNA* of both *Drosophila melanogaster* and *S. cerevisiae* was later imaged with electron microscopy, and the triple-stranded structures were termed R-loops (Thomas, White and Davis, 1976; White and Hogness, 1977). A large expansion of research into *in vivo* R-loops has now revealed their frequent occurrence outside of the *rDNA* loci, as well as their central role in several physiological processes, including telomere maintenance, immunoglobulin (Ig) class switching and mitochondrial DNA (mtDNA) replication (Yu *et al.*, 2003; Skourti-Stathaki and Proudfoot, 2014; Santos-Pereira and Aguilera, 2015a). More recently, increasingly diverse roles have been attributed to R-loops, including transcription initiation and termination, DNA replication initiation, and chromatin remodelling (Skourti-Stathaki and Proudfoot, 2014; Lombraña *et al.*, 2015; Santos-Pereira and Aguilera, 2015b). Despite their seemingly diverse functionality, R-loops can also create obstacles for replication forks, have been shown to cause genomic instability and hyper-recombination (Chan, Hieter and Stirling, 2014; Hamperl and Cimprich, 2014; Pan *et al.*, 2014; Sollier and Cimprich, 2015), and have been linked to various neurological diseases and cancers (Groh and Gromak, 2014; Richard and Manley, 2017). Several R-loop capture techniques have now been developed (discussed in more detail in chapter 4) that allow detailed mapping of the transient structures; however, separating functional R-loops from those which are aberrant and deleterious to genome integrity remains challenging.

## 1.4.1 Programmed R-loops and their physiological functions

### 1.4.1.1 R-loops form at gene promoters

Genome-wide mapping of RNA-DNA hybrids across the human genome has revealed strong enrichment for R-loops at the promotor regions of 1,972 unique genes (Ginno *et al.*, 2012). Interestingly, 84% of these were CpG island (CGI) promoters and the majority had areas of particularly strong GC skew downstream of the TSS (65%), whilst the remainder presented more minor GC skew (19%) (Ginno *et al.*, 2012). R-loop formation was also shown to directly protect sequences from transcriptionally repressive *de novo* DNA methylation (Ginno *et al.*, 2012). Hence, R-loop formation is hypothesised to protect CGI promoters from methylation to allow active transcription, an idea supported by the fact that methylation status can be predicted by the level of GC skew, and thus the R-loop forming potential, of CGI promoters (Ginno *et al.*, 2012). Using GC skew levels and the prevalence of CGIs, promoters were classed into four groups (Ginno *et al.*, 2013). The first are CGI promoters with high levels of GC skew, which were largely associated with consistently transcribed “house-keeping” genes. The second contains shorter CGIs, lower levels of GC skew and are less strongly associated with house-keeping functions. These two classes can be considered “dispersed” promoters (section 1.2.2.1) and are associated with R-loop formation. The third class are promoters with low levels of CGI that do not show strong GC skew or R-loop formation. These genes are associated with more tissue-specific functions. Lastly, a fourth class of promoter is defined as those with apparent CGIs but strong negative GC skew found upstream of the TSS. Genes with this final class of promoters were clearly enriched for transcription regulation, cell fate and morphogenesis functions (Ginno *et al.*, 2013). Considering that high levels of GC skew were associated with R-loop formation over GCI promoters, and these are associated with lower levels of DNA methylation (Ginno *et al.*, 2012), it is hence plausible that R-loop formation maintains active transcription of class I, and to a lesser extent class II, promoters (which fall into the dispersed category) to allow continued expression of house-keeping genes (Ginno *et al.*, 2013).

Strand-specific analysis of genome-wide R-loops in human cells also identified R-loops over promoter regions, again most prominently at CGI promoters (Sanz *et al.*, 2016). R-loops over these promoter regions were associated with increased chromatin accessibility, high RNA Pol II occupancy, and several epigenetic markers characteristic of open chromatin (Sanz *et al.*, 2016; Chen *et al.*, 2017) (Figure 1.9). These include, Histone 3 lysine 4 di- and tri-methylation (H3K4me2/3), H3K9 acetylation (H3K9ac), and H3K27ac, which are all markers of active transcription and show marked enrichment in R-loop forming promoters, peaking over the TSS (Sanz *et al.*, 2016; Chen *et al.*, 2017). H3K9me1 is also enriched, but directly over the R-loop peaks that are positioned 1 – 2 kb downstream of the TSS (Sanz *et al.*, 2016; Chen *et al.*, 2017). Histone variant H3.3, which is associated with transcriptionally active chromatin, is also slightly enriched around the TSS of R-loop forming promoters (Sanz *et al.*, 2016). Lastly, the transcription elongation marker H3K36me3 is significantly enriched over the gene body, also downstream of the TSS (Sanz *et al.*, 2016). In addition to the increased levels of open chromatin markers, a decrease in heterochromatin markers H3K9me3 (Sanz *et al.*, 2016) and H3K27me3 (Chen *et al.*, 2017) is evident over the TSS of such promoters, as well as protection from hyper DNA methylation (Sanz *et al.*, 2016). Hence, R-loop forming promoters are heavily protected from epigenetic silencing in the human genome (Sanz *et al.*, 2016; Chen *et al.*, 2017), and mostly likely contribute to continued transcription of house-keeping genes associated with dispersed promoters (Ginno *et al.*, 2013).

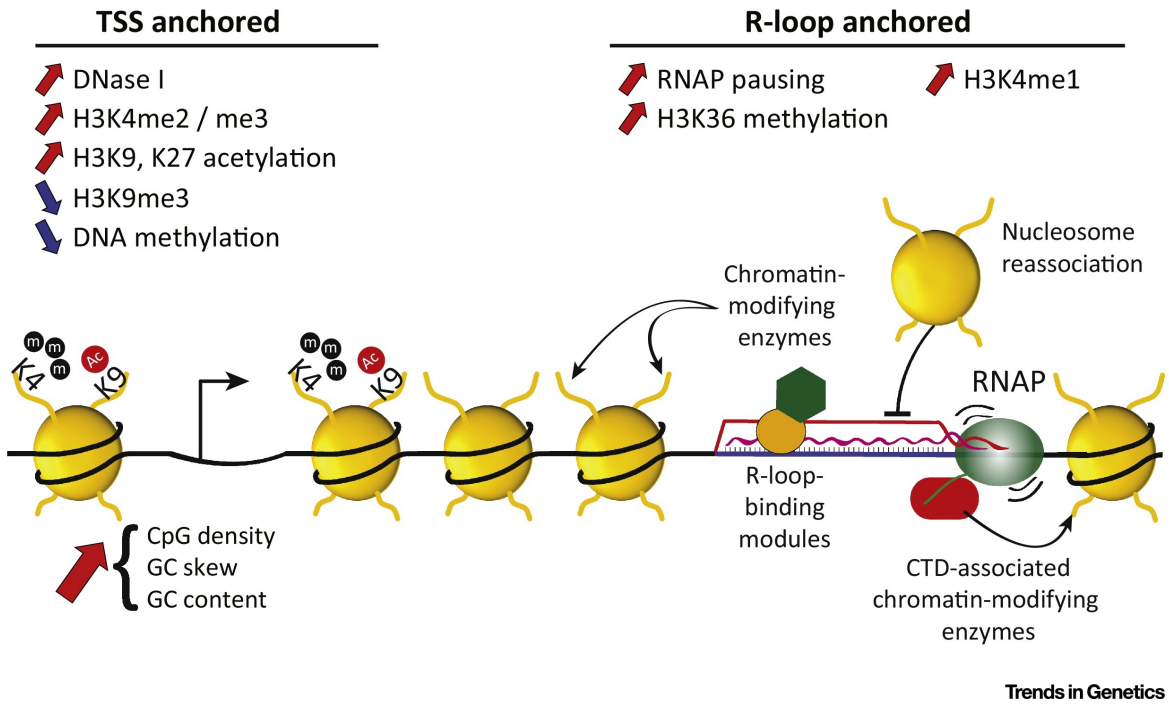



**Figure 1.9 Chromatin features associated with R-loop forming promoter regions of the human genome.**
Several epigenetic markers have been associated with R-loop formation at CpG island promoters, either located directly at the transcriptional start site (TSS anchored) or at the R-loop formation peak, typically found 1 – 2 kb downstream of the TSS (R-loop anchored). Mostly, these promoters display higher GC skew and GC content, as well as increased DNase I sensitivity. Enrichment of several epigenetic markers characteristic of active transcription have been documented at such promoters as well as protection from repressive H3K9 tri-methylation and DNA methylation. R-loop formation has also been correlated with RNA Pol (RNAP) pausing downstream of the TSS, as well as the increased presence of transcription and elongation markers H3K4me1 and H3K36 methylation. Hypothesised to link R-loops and chromatin patterning is the action of chromatin-modifying enzymes, perhaps recruited via interaction with the C-terminal domain (CTD) of paused RNAP, recruitment to the R-loop structure itself, or repression of nucleosome association. Findings and hypothesis depicted are sourced from Sanz *et al*. (2016), Chen *et al*. (2017) and Chédin, (2016). Image source Chédin, (2016), license number 4410710157988.

R-loops which form around human promoter regions have been shown to form co-transcriptionally and to initiate at the TSS, which forms the 5' R-loop boundary (Dumelie and Jaffrey, 2017). The 3' R-loop boundary has been mapped to the first exon/intron junction of intron containing genes, whereas when introns were lacking the 3' boundary varied from gene to gene (Dumelie and Jaffrey, 2017). Despite this fine mapping of R-loops and the clear association with constitutively active promoters, how R-loops may form part of a promoter driven transcription initiation pathway genome-wide is less clear.

Some studies have investigated the function of promoter-associated R-loops at specific loci. In the *Arabidopsis* genome, for example, R-loops have been

associated with both promoter regions and gene bodies, as well as noncoding RNA (ncRNA) (Xu *et al.*, 2017). Detailed investigation of R-loop formation at the floral repressor gene (*FLC*) locus revealed a direct role for R-loops in transcription regulation (Sun *et al.*, 2013). In response to prolonged cold, antisense transcription of the *FLC* locus generates antisense transcripts termed *COOLAIR*, which are able to induce epigenetic silencing of sense *FLC* transcription via various pathways. The *COOLAIR* promoter was found to be GC rich, depleted of nucleosomes, and have the potential to form R-loops (Sun *et al.*, 2013). R-loops originating from the *COOLAIR* promoters can be additionally stabilised via binding of the homeodoman-containing protein, AtNDX, repressing transcription of *COOLAIR* (Sun *et al.*, 2013). Therefore, negative regulation of the *COOLAIR* transcripts via R-loop formation has the potential to positively regulate transcription of *FLC* to regulate flowering (Sun *et al.*, 2013). Sense and anti-sense transcription is also observed at the CGI promoter of a human gene, vimentin (*VIM*) (Boque-Sastre *et al.*, 2015). Anti-sense transcription of *VIM* favours formation of an R-loop structure on the negative strand, reduced nucleosome occupancy and increased transcription factor binding to the displaced ssDNA strand, all allowing active sense transcription of *VIM* (Boque-Sastre *et al.*, 2015). Hence, in both cases R-loop formation on the antisense strand is associated with active sense strand transcription. Whether this model holds true across all R-loop-associated promoters is unlikely, however, as most promoter-associated R-loops form on the sense strand in the human genome (Sanz *et al.*, 2016).

Genome-wide mapping of R-loops in mouse embryonic stem cells again revealed R-loop enrichment at promoter regions, largely in highly expressed genes (P. B. Chen *et al.*, 2015). Interestingly, these R-loops enhance binding of the Tip60-p400 chromatin remodelling complex and, contrastingly, inhibit binding of polycomb repressive complex 2 (PRC2) as well as histone H3 lysine 27 methylation (H3K27me) (P. B. Chen *et al.*, 2015). The Tip60-p400 complex is made up of the Tip60 protein lysine acetyltransferase, which acetylates histone H4 and H3A lysine residues, and the p400 ATPase that incorporates the histone variant H2A.Z. Curiously, the complex is able to simultaneously repress differentiation genes and activate proliferation genes. Since dispersed, CGI promoters display H2A.Z deposition at the

boundaries of the nucleosome depleted region (Rach *et al.*, 2011), it may be possible that associated R-loops enhance binding of the Tip60-p400 complex, facilitating deposition of the histone variant to positively regulate transcription. PCR2 is an H3K27 methyltransferase that is important for gene silencing during development. Disruption of promoter-proximal R-loops increased PRC2 binding and impaired embryonic stem cell (ESC) differentiation (P. B. Chen *et al.*, 2015). Therefore R-loops are able to bind some protein complexes whilst preventing binding of others. How the two activities are distinguished, however, remains unclear.

In light of these experiments, it is clear that promoter proximal R-loops have a role in transcription regulation, which impacts upon important biological processes, such as cell differentiation (P. B. Chen *et al.*, 2015) and flowering (Sun *et al.*, 2013). However, it is probable that R-loops are implicated in many different pathways which influence transcription, potentially both positively and negatively, and that no one model is likely to be conserved uniformly across all species, or even between different genes within the same species.

#### 1.4.1.2 R-loops and transcription termination

R-loop mapping across the human genome also found enrichment to the 3' end of a subset of genes, the majority of which harboured GC skew (Ginno *et al.*, 2013). Interestingly, these highly GC skewed TTSs were somewhat protected from DNA methylation in a similar manner, although less significantly so, to the protection observed over R-loop forming promoter regions (Ginno *et al.*, 2013). Detailed analysis of the human ß-actin gene loci revealed R-loop formation over regions of RNA Pol II pausing at the 3' end of the gene, immediately downstream of the PAS (Skourti-Stathaki, Proudfoot and Gromak, 2011; Skourti-Stathaki, Kamieniarz-Gdula and Proudfoot, 2014). R-loop resolution by senataxin, an RNA-DNA targeting helicase, in these regions is necessary to allow Xrn2 exonuclease access to the 3' cleavage-associated PAS (Skourti-Stathaki, Proudfoot and Gromak, 2011). Hence, R-loop resolution appears to be required before correct mRNA maturation can take place. The terminal RNA Pol II pause regions were also found to be sites of RNAi

activity and epigenetic silencing (Skourti-Stathaki, Kamieniarz-Gdula and Proudfoot, 2014). Antisense transcription and binding of the RNAi factors AGO1 and DICER were detected over the R-loop forming TTS, and were diminished by R-loop degradation (Skourti-Stathaki, Kamieniarz-Gdula and Proudfoot, 2014). R-loop formation also initiated G9a histone lysine methyltransferase recruitment, H3K9me2 enrichment and chromodomain protein HP1γ binding, to promote transcriptional silencing of ß-actin (Skourti-Stathaki, Kamieniarz-Gdula and Proudfoot, 2014). Therefore, R-loop formation is required for correct termination of ß-actin transcription, yet the R-loop must be subsequently resolved to allow maturation of the transcripts.

Genome-wide comparison of human R-loop forming TTSs and non-R-loop forming genes, however, found no overall enrichment of H3K9me2/3 over R-loop positive termination regions (Sanz *et al.*, 2016). This finding indicates R-loop driven recruitment of G9a lysine methyltransferase at the ß-actin locus is not conserved elsewhere. Broadly, global enrichment was observed for binding of the RNAP-interacting complex, PAF1, which has been shown to interact with cleavage and polyadenylation specific factors (Tomson and Arndt, 2013), further linking terminal R-loops to terminal mRNA maturation (Sanz *et al.*, 2016). R-loop positioning around the PAS was also found to correlate with the site of termination; early R-loop peaks corresponded to transcription termination occurring ~7 kb downstream of the PAS, whereas late forming R-loop genes terminated ~17 kb downstream (Sanz *et al.*, 2016). Notably, terminal R-loops were only found in those transcripts undergoing PAS-dependant cleavage and polyadenylation and their formation does not correlate with GC skew in this study, indicating factors which influence R-loop formation are likely to differ in promoter and terminator regions (Sanz *et al.*, 2016).

Evidence of R-loop involvement in termination processes is also found in the *S. cerevisiae* genome, where mapping of RNA-DNA hybrids found >60% form at highly transcribed genes with long 3' homopolymeric dA:dT tracts (Wahba *et al.*, 2016). The likelihood of R-loop formation over the poly(A) tracts was found to be dependent on the length of the tract, where ~92% of poly(A) tracts of ≥ 21 bp were prone to hybrid formation (Wahba *et al.*, 2016). These studies in the yeast and

human genomes reveal a clear link between 3' R-loop formation in a subset of genes, although these may be related to cleavage and polyadenylation rather than termination. Exactly how R-loops participate in any of these processes, or whether R-loops simply form as a consequence of RNA Pol pausing, remains unclear.

#### 1.4.1.3 R-loops and DNA replication

R-loops were first linked to DNA replication when found to initiate replication of the *E. coli* plasmid ColEI; a persistent RNA-DNA hybrid forms within specific origins (Dasgupta, Masukata and Tomizawa, 1987; Masukata and Tomizawa, 1990) and is cleaved by the RNA-DNA hybrid-targeting endonuclease Ribonuclease (RNase) H, allowing it to serve as the primer for DNA synthesis (Itoh and Tomizawa, 1980). RNA-DNA hybrids have also been found to form at the well conserved mtDNA ORI, dependant on the presence of a GC-rich cluster (Chang and Clayton, 1985; Chang, Hauswirth and Clayton, 1985). The RNA strand of the initiation R-loop acts as the primer for elongation by mtDNA polymerase, much like *E. coli* plasmid replication (Xu and Clayton, 1996). Similarly, R-loop formation over a specific DNA unwinding element initiates DNA replication of bacteriophage T4 by providing a primer for leading-strand synthesis, generated by either RNA polymerase or RNase H (Carles-Kinch and Kreuzer, 1997). Alternatively, T4 primase synthesises a primer on the non-template ssDNA strand of the R-loop (Belanger and Kreuzer, 1998).

DNA replication can also be initiated by R-loop formation in the *E. coli* chromosome of RNase H-deficient cells (Asai and Kogoma, 1994) or WT cells under stress conditions (Hong, Cadwell and Kogoma, 1996; Camps and Loeb, 2005; Wimberly *et al.*, 2013). During this alternative route of replication, R-loops form by transcript invasion rather than co-transcriptionally (Asai and Kogoma, 1994), initiating replication at multiple sites that are separate from the single origin and resulting in global changes to replication fork migration patterns and frequent converging forks (Maduike *et al.*, 2014).

R-loops have more recently also been linked to DNA replication in the human genome. As discussed above, R-loops frequently locate to CGI promoters; these are

also the sites of the most efficiently activated and conserved ORIs in mice and humans (Sequeira-Mendes *et al.*, 2009; Cayrou *et al.*, 2011; Besnard *et al.*, 2012; Picard *et al.*, 2014). Hence, like in *T. brucei*, DNA replication is heavily linked to transcription in mammalian genomes and, importantly, R-loops may provide a functional link (Lombraña *et al.*, 2015). How co-transcriptional R-loops may initiate DNA replication in these more complex genomes is, however, unclear. One plausible hypothesis involves the formation of G4 structures (Lombraña *et al.*, 2015), which are present in most mammalian genome ORIs (Cayrou *et al.*, 2011; Besnard *et al.*, 2012). Interestingly, *in vitro* assays have demonstrated that ORC1 is able to bind G4 DNA (Hoshina *et al.*, 2013). Binding affinity of ORC1 to G4 DNA is also dramatically increased by G4 structures from with ssDNA, rather than a DNA duplex (Hoshina *et al.*, 2013), and the ssDNA element of R-loops have the potential to form G4 DNA. In support of this hypothesis, 30% of stringently selected ORIs found at CGI promoters that display R-loop enrichment (Ginno *et al.*, 2012, 2013) also display OCR1 binding upstream of the TSS and G4 formation (Lombraña *et al.*, 2015) (Figure 1.10).

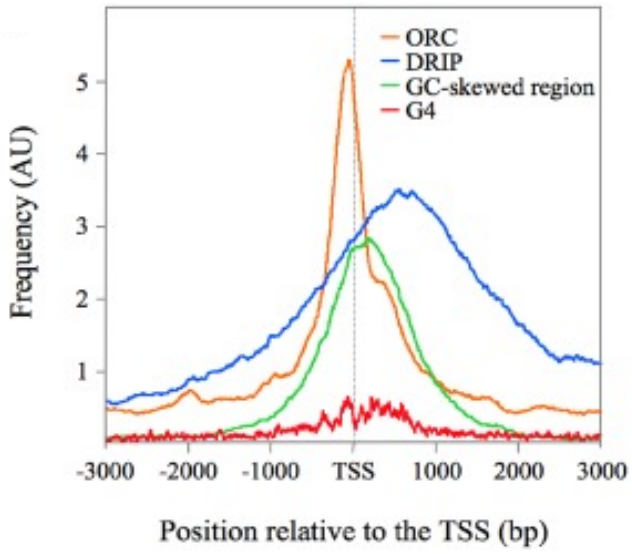



**Figure 1.10 Association of R-loops, G4 motifs and GC skew at sites of ORC binding.**
The distribution of ORC1-binding (orange), DNA-RNA hybrid enrichment (blue), GC skew regions (green) and G4 motifs (red) at CGI promoters that also function as ORIs. In each case hits per bp are mapped over 6 kb centres at the TSSs of 485 CGI promoters. Predicted G4 motifs were identified with algorthym Quadparser (Huppert and Balasubramanian, 2005) and GC-skew regions were sourced from (Ginno *et al.*, 2012). Data, analysis and image source Lombraña *et al.* (2015).

#### 1.4.1.4 R-loops as regulators of chromatin dynamics

As discussed above (sections 1.4.1.1 and 1.4.1.2), R-loop formation has been linked to several epigenetic markers, both of open (Sanz *et al.*, 2016; Chen *et al.*, 2017) and compact (Skourti-Stathaki, Kamieniarz-Gdula and Proudfoot, 2014) chromatin, at promoter and terminator regions, respectively. DNase I hypersensitivity is associated with R-loop formation irrespective of genic position, highlighting the possibility that R-loops influence nucleosome positioning genome-wide (Sanz *et al.*, 2016; Chen *et al.*, 2017). In support of this, R-loops were found to correlate with lower nucleosome occupancy and chromatin opening over the *Snord116* allele (Powell *et al.*, 2013), and conversely, directed R-loop disruption leads to chromatin condensation at the *VIM* loci (Boque-Sastre *et al.*, 2015). Additionally, incorrect transcription-coupled nucleosome positioning may increase R-loop formation, as deletion of the FACT histone chaperone complex in yeast leads to R-loop accumulation (Herrera-Moyano *et al.*, 2014). Interestingly, these observations are consistent with the finding that the presence of RNA within the DNA double helix prevents binding of the core histones (H2A, H2B, H3 and H4) to plasmids *in vitro* (Dunn and Griffith, 1980).

#### 1.4.1.5 Telomeric R-loops

Telomeric repeats undergo transcription by RNA Pol II to generate TERRA, which harbours telomeric repeat sequence as well as elements from the subtelomere (Luke *et al.*, 2008). TERRA is able to form RNA-DNA hybrids at telomeres, which have been proposed to have physiologically relevant functions (Balk *et al.*, 2013; Arora and Azzalin, 2015; Rippe and Luke, 2015). Cancer cell lines which employ alternative telomere lengthening (ATL) strategies exhibit high levels of TERRA and telomeric RNA-DNA hybrids (Arora *et al.*, 2014; Yu, Kao and Lin, 2014). Additionally, RNA-DNA hybrids are elevated in telomerase-null yeast whose telomeres undergo homology directed repair (Balk *et al.*, 2013). Hence, TERRA is thought to form co-transcriptional R-loops in critically short telomeres to promote telomere lengthening and avoid premature cellular senescence (Balk *et al.*, 2013; Rippe and Luke, 2015; Graf *et al.*, 2017). In fact, increased levels of R-loops in critically short

telomeres may promote recombination via clashes with replication machinery, which are known to induce DNA damage and recombination (Aguilera and Gómez-González, 2008; Hamperl and Cimprich, 2016), since telomeric R-loops are cell cycle regulated (Graf *et al.*, 2017) and short telomeres are replicated earlier in S phase (Bianchi and Shore, 2007).

#### 1.4.1.6 R-loops and centromeres

ncRNAs transcribed from centromeric repeats are capable of forming R-loops. In a similar RNAi-dependent mechanism to that observed at the R-loop forming ß-actin terminator region (Skourti-Stathaki, Kamieniarz-Gdula and Proudfoot, 2014), centromeric R-loops mediate heterochromatin formation in yeast (Nakama *et al.*, 2012). Moreover, centromeric R-loops have been found to direct histone H3 serine 10 phosphorylation (H3S10P), a known marker of chromatin condensation, to maintain genome stability (Castellano-Pozo *et al.*, 2013). More recently, ATR (a master check-point kinase) has been shown to locate to centromeres, where the protein is activated via R-loops and the mitotic regulator kinase, Aurora A (Kabeche *et al.*, 2018). Here, ATR is able to promote Aurora B activation, which is critical for correct chromosome alignment during metaphase (Shuda *et al.*, 2009), and faithful chromosome segregation (Kabeche *et al.*, 2018). How R-loops contribute to ATR binding and activation in this scenario remains unclear, although is it possible that ATR binds to replication protein A (RPA)-coated ssDNA, generated by R-loop formation (Kabeche *et al.*, 2018). Indeed, RPA has been shown to locate to R-loops and become phosphorylated by the action of ATR in human cells (Leong *et al.*, 2017). Therefore, R-loops appear to have a role in maintaining heterochromatin at centromeric repeat regions, as well as aiding correct chromosome segregation during mitosis by the activation of Aurora B kinase via ATR (Shuda *et al.*, 2009; Kabeche *et al.*, 2018).

### 1.4.2 R-loops and genomic instability

As well as performing physiological functions in genomes of all types, R-loops pose significant risks to genome integrity. Aberrant R-loop formation, either in terms of

the hybrid's location or timing, leads to DNA damage that undermines the genome and can lead to mutation at the DNA base or chromosome level. In particular, RNA biogenesis mutants are associated with increased levels of R-loops (section 1.4.2.2). How deleterious R-loop formation is converted into DNA damage is debated, as various studies have led to different conclusions being drawn, although generally it is accepted that R-loop formation alone is insufficient, and a secondary event is required. For example, the exposed ssDNA portion of an R-loop is vulnerable to the actions of nucleases and is prone to breakage. R-loops also present potential obstacles to replication forks, which when stalled often collapse, leading to the occurrence of DSBs. For these reasons, all organisms studied thus far express R-loop resolving enzymes (section 1.4.2.3).

### 1.4.2.1 DNA sequence composition influences R-loop formation

Several aspects of DNA sequence composition have been documented to correlate with or determine R-loop formation. Clustering of three or more guanine residues has been shown to be important for the initiation of R-loop formation, and guanine-rich regions were found to allow elongation of the R-loop (Roy and Lieber, 2009). *In silico* analysis of the human genome searching for regions meeting these criteria revealed ~250,000 putative R-loop forming sequences (Wongsurawat *et al.*, 2012; Jenjaroenpun *et al.*, 2015). GC skew, irrespective of clustering, correlates with R-loop formation over CGI promoters and 3' terminator regions of human genes (Ginno *et al.*, 2013). Both positive GC skew and AT skew have also been shown to correlate with R-loop formation in the *Arabidopsis thaliana* genome (Xu *et al.*, 2017). Lastly, poly(A) tracts consisting of >21 bp are sufficient for R-loop formation in *S. cerevisiae* (Wahba *et al.*, 2016).

### 1.4.2.2 RNA biogenesis mutants accumulate R-loops

To date, mutants of >50 proteins have been shown to cause increased levels of R-loop formation, most of them in yeast (Santos-Pereira and Aguilera, 2015a; Al-Hadid and Yang, 2016) (Figure 1.11b). These include proteins involved in transcription elongation (Huertas and Aguilera, 2003) and termination (Wahba *et al.*, 2011), mRNA spicing, polyadenylation, mRNA export and degradation (Santos-Pereira and

Aguilera, 2015a; Al-Hadid and Yang, 2016). Other mutations in proteins that affect mitosis regulation, DNA topology, nucleosome organisation and the DNA damage response also display increased levels of R-loops (Santos-Pereira and Aguilera, 2015b; Al-Hadid and Yang, 2016). In general terms, these studies highlight the importance of efficient RNA maturation and turn over in order to avoid the formation of deleterious RNA-DNA hybrids.

In keeping with the observation that the presence of introns appears to protect genes from R-loop formation in complex genomes (section 1.4.1.1), splicing factor mutants accumulate R-loops (Li and Manley, 2005; Wahba *et al.*, 2011; Stirling *et al.*, 2012; Nguyen *et al.*, 2017). Depletion of the ASF/SF2 splicing factor in chicken B-cells leads to extensive DNA rearrangements due to the accumulation of R-loops (Li and Manley, 2005). Human cells expressing a mutant form of splicing factor U2AF1 (U2 snRNA auxiliary factor 1) similarly display increased R-loop formation as well as DNA damage, as does the pharmacological inhibition of mRNA splicing (Nguyen *et al.*, 2017). Screening of yeast mRNA splicing and polyadenylation factors further linked these processes and the prevention of R-loops, as several mutants were shown to possess higher levels of both DNA damage and R-loops (Wahba *et al.*, 2011; Stirling *et al.*, 2012).

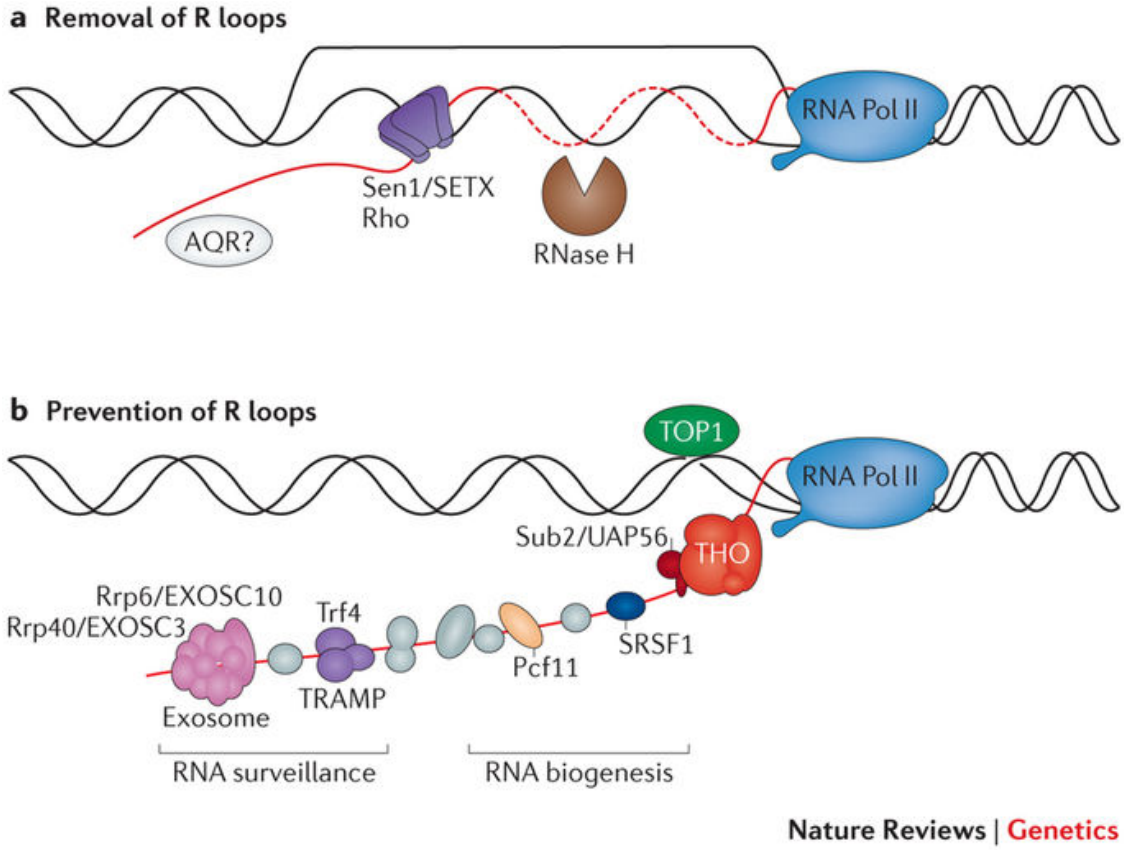



**Figure 1.11 Proteins which remove and prevent R-loops are found across species.**
a) Removal of R-loops can be achieved by different proteins, including endonuclease RNase H enzymes and the helicases Sen1/SETX. Other proteins that may be involved include the helicase aquarius (AQR). b) Other proteins prevent R-loop formation. These include topoiserase I (TOP1), which resolves negative supercoiling, and proteins invloved in RNA biogenesis and RNA surveillance. Image source Santos-Pereira and Aguilera, (2015b), license number 4410710530425.

### 1.4.2.3 R-loop resolving enzymes

As well as actively preventing aberrant R-loop formation, cells also express proteins that resolve R-loops (Figure 1.11a). These proteins include type 1 and type 2 RNase H enzymes that cleave the RNA strand within RNA-DNA hybrids (Cerritelli and Crouch, 2009). Most organisms express one type 1 protein and a type 2 RNase H; in bacteria both classes are is monomeric, but in eukaryotes type 2 RNase H functions as a trimeric complex (Cerritelli and Crouch, 2009). In mammals, both RNase H1 and H2 proteins are essential for development, as RNase H1 is essential for mtDNA replication (Cerritelli *et al.*, 2003) and RNase H2 is critical for maintaining nuclear genome stability (Hiller *et al.*, 2012). However, both types of RNase H are dispensable for cell viability in yeast, although their deletion does lead to detectable levels of mutation and chromosome rearrangement (Wahba *et al.*, 2011;

O'Connell, Jinks-Robertson and Petes, 2015). Type 2 RNase H in eukaryotes has an additional function as the protein is able to initiate the ribonucleotide excision repair (RER) pathway (Sparks *et al.*, 2012; Williams, Gehle and Kunkel, 2017). The enzyme complex cleaves 5' of ribonucleotides that have been mistakenly embedded into newly synthesised DNA by replicative polymerases, allowing them to be removed by the activity of flap endonucleases, polymerase δ and DNA ligase (Sparks *et al.*, 2012). RNase H enzymes are discussed in greater detail in Chapter 3.

RNA-DNA hybrid helicase enzymes of the Sen1 family are also able to resolve R-loops. Mutation in the helicase domain of yeast Sen1 results in the accumulation of DNA damage (Mischo *et al.*, 2011). Interestingly, Sen1 is associated with replication forks across actively transcribed genes, and so may be involved in resolving replication and transcription machinery clashes associated with R-loops (see section 1.4.2.4) (Alzu *et al.*, 2012). The mammalian homolog for Sen1, senataxin (SETX), resolves R-loops particularly arising at G-rich transcription termination pause sites (Skourti-Stathaki, Proudfoot and Gromak, 2011). SETX has a prominent role in responding to R-loop associated DNA damage, significantly during male meiosis (Becherel *et al.*, 2013), and mutations of the gene are associated with neurodegenerative disorders (Chen *et al.*, 2004; Arning *et al.*, 2013).

RPA binds to ssDNA and has been shown to interact with RNase H1 at R-loops to stimulate binding and catalytic activity of the enzyme (Nguyen *et al.*, 2017). R-loops resulting from mutations of splicing machinery in malignant cell lines are also associated with RPA and in this case RPA has been found to be phosphorylated by ATR (Leong *et al.*, 2017). As discussed in section 1.4.1.6, ATR binds RPA-coated centromeric R-loops, where is it activated and stimulates Aurora B allowing correct chromosome segregation (Kabeche *et al.*, 2018) (section 1.4.1.6). These studies highlight signalling pathways initiated by R-loop formation.

#### 1.4.2.4 Transcription-replication clashes

R-loops present significant barriers to the DNA replication machinery, meaning deregulated co-transcriptional R-loops are a significant source of genome instability

as they prevent proper DNA replication and can lead to collapse of the replication fork (Aguilera and García-Muse, 2012; Wahba and Koshland, 2013; Hamperl *et al.*, 2017; Lang *et al.*, 2017). Whether such clashes occur co-directionally or as a result of a head-on clash between the replisome and RNA Pol machinery has consequences for the outcome (Hamperl *et al.*, 2017; Lang *et al.*, 2017) (Figure 1.12). Head-on clashes promote R-loops in sequences prone to their formation and result in the activation of the ATR-dependant DNA damage signally pathway, implying the exposure of ssDNA either within the stalled replication fork or the R-loop itself (Hamperl *et al.*, 2017). The head-on collision further stalls the transcribing RNA Pol, which presumably increases the likelihood of R-loop formation, as well as the replication machinery resulting in incomplete DNA replication (Hamperl *et al.*, 2017). In contrast, co-directional clashes that occur due to R-loop-induced stalling of transcription machinery result in decreased R-loop levels. In this case ATM is activated, implying the presence of DSBs (Hamperl *et al.*, 2017). Hence, the direction of transcription-replication clashes greatly influences R-loop and DNA damage levels, and may explain the bias towards co-directional activity of transcription and replication in the human genome (Hamperl *et al.*, 2017).

Head-on collisions between replication and transcription machinery also increases R-loop formation in bacterial pathogen *Bacillus subtilis* (Lang *et al.*, 2017). Here, R-loops completely blocked replication and resulted in high levels of mutation in genes where head-on collisions occurred, as well as changes in gene expression (Lang *et al.*, 2017).

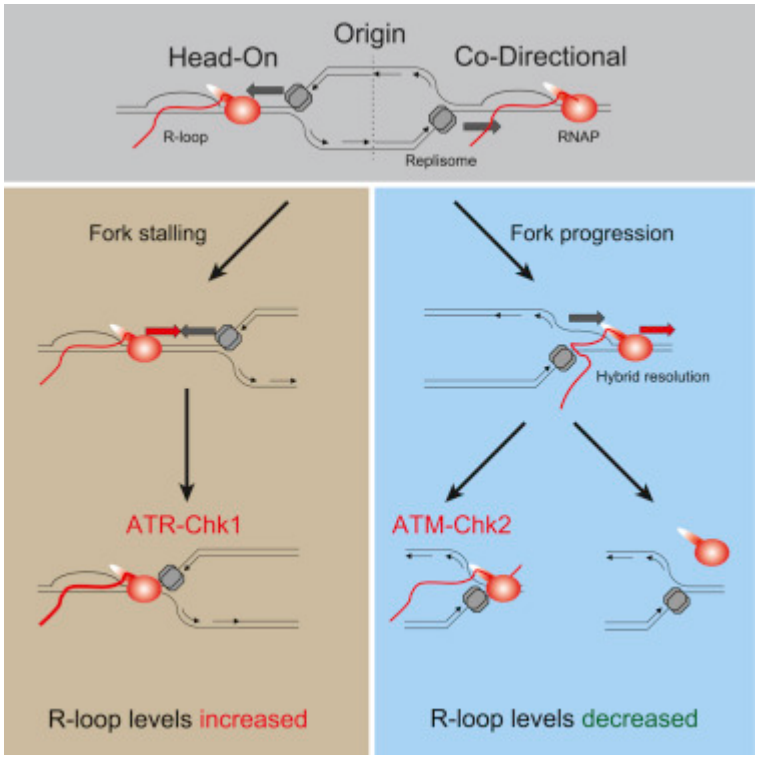



**Figure 1.12 Head-on vs co-directional replicaition-transcription conflicts associated with R-loops.**
Replication forks progress in both directions from ORIs and can collide with transcribing polymerases (RNAP) either head-on or co-directionally. Head-on collision result in stalled replication forks and an increase in R-loop formation behind the blocked transcription bubble. This type of collision is associated with ATR signalling via check point kinase 1 (Chk1). Co-direction clashes allow the replication fork to continue resolving the R-loop behind RNAP. This results in resolution of the R-loop and ATM-Chk2 signalling. Image source Hamperl *et al.* (2017)

## 1.5 Aims and objectives

The gathering wealth of R-loop research is highlighting the diverse roles of these structures, as well as the risks they pose to genome stability. The primary aim of this thesis was therefore to investigate the role of R-loops in the *T. brucei* genome. Several processes that R-loops have been implicated in, namely transcription initiation, termination and DNA replication, are not fully understood in *T. brucei* and so it was hypothesised that the structures may have conserved roles in one or more of these processes, revealing functional overlap with other eukaryotes, or divergence of *T. brucei* in each case. As R-loops are also well-known inducers of DNA damage and recombination events, it was hypothesised that the structures could trigger VSG switching events in BSF *T. brucei*, shedding light on the mechanism(s) of antigenic variation.

To address the above questions, a number of approaches were taken. First, the mode of action of *T. brucei* RNase H enzymes were determined by gene knockout or RNAi-mediated knockdown, and localisation of the predicted RNases H examined. Second, R-loops were mapped across the *T. brucei* genome to determine regions of accumulation. This mapping was undertaken in WT parasites as well as those lacking or depleted of RNase H activity. As RNase H enzymes resolve R-loops, reducing their activity might be predicted to increase the persistence of R-loops in the genome, validating that the precipitation approach adopted captures R-loops. Third, the effects of RNase H loss or depletion on expression of the VSG coat, as well as on R-loop accumulation in the BES and on DNA damage levels were evaluated.

# 2 Materials and Methods

# 2.1 *T. brucei* in vitro culture

## 2.1.1 *T. brucei* strains

The monomorphic Lister 427 strain of long-slender BSF *T. b. brucei* was used as the wild type (WT) cell line (Rudenko *et al.*, 1996; McCulloch, Rudenko and Borst, 1997). Gene deletion mutants and endogenously tagged cell lines were generated using this strain as the parental cell line. Inducible RNAi lines were generated using a genetically modified strain derived from Lister 427, named 2T1 (Alsford *et al.*, 2005).

## 2.1.2 *In vitro* culture of BSF parasites

Lister 427 WT and 2T1 lines, as well as mutant lines originating from each, were maintained in HMI-9 (HMI-9 (Gibco), 20% v/v FCS (Gibco), Pen/Strep (Sigma) (penicillin 20 U/ml, streptomycin 20 µg/ml)), and HMI-11 media (HMI-9 (Gibco), 10% v/v FCS (Gibco), Pen/Strep (Sigma) (penicillin 20 U/ml, streptomycin 20 µg/ml)), respectively. Cells were cultured at 37˚C, 5% $CO_2$, in vented flasks (Hirumi and Hirumi, 1989). For routine culture parasites were maintained in 5 ml at a density below ~2 x $10^6$ cells/ml. 2T1 cells lines were maintained in the presence of 0.2 µg/ml Puromycin (InvitroGen) and 5 µg/ml phleomycin (InvitroGen). Once transfected with a target gene construct, 2T1 parasites were instead cultured with 5 µg/ml phleomycin (Invitrogen) and 5 µg/ml hygromycin B (Calbiochem) to maintain the RNAi construct. RNAi was always induced with 1 µg/ml tet.

| Name | Concentration |
|---|---|
| Hygromycin | 5 µg/ml |
| Neomycin | 2.5 µg/ml |
| Blasticidin | 10 µg/ml |
| Phleomycin | 5 µg/ml |
| Puromycin | 0.2 µg/ml |

**Table 2.1 Selective drug concentrations for *T. brucei* culture.** The final concentration of each selective drug.

For EdU incorporation experiments (section 2.5.4) the RNAi cell line was cultured in a thymidine-free version of HMI-11 media (Iscove's MDM media (Gibco), 10% v/v FCS (Gibco), 1 mM Hypoxanthine, 0.05 mM Bathocuproine disulphic acid, 1 mM sodium pyruvate, 1.5 mM L-cystiene).

Population density was determined by loading 10 µl of resuspended culture into a Neubauer improved haemocytometer (Marienfeld-Superior) and counting at least 2 of the 1 mm square corner regions, taking the average of these. This number was then multiplied by 10,000 to give the number of parasites per ml of culture.

### 2.1.3 Stable transfection of BSF parasites

A 33 ml culture of cells was grown to ~1 x $10^6$ cells/ml, allowing collection of 3 x $10^7$ cells by centrifugation at 432g for 10 min. The majority of the media was poured off before the pelleted parasites were resuspended in the residual media and transferred to a 1.5 ml microcentifuge tube. Cells were then centrifuged at 664 g for 3 min before the media was completely removed. Cells were resuspended in 100 µl of Amaxa Human T Cell Nucleofactor kit solution (Lonza) and transferred to an electroporation cuvette. 10 µg of linearised construct DNA was added to the cuvette and cells were electroporated using Nucleofector™ 2b Device (Lonaz), programme X-001. Once electroporated, cells were immediately transferred into 30 ml of the appropriate media (free of selective drugs) and the tube was inverted several times. Two serial dilutions of the culture were then made by transferring 3 ml of the first into another 27 ml of fresh media, inverting the culture several times and repeating the process once more to generate 1:10 and 1:100 dilutions. This resulted in three cultures at cell densities of approximately 1 x $10^6$ cells/ml, 1 x $10^5$ cells/ml and finally 1 x $10^4$ cells/ml. Each cell dilution was then aliquoted as 1 ml cultures into 24 well plates (Corning). 24 hr later, the appropriate selective drug (Table 2.1) was added to the parasites in a further 1 ml of media (containing double the final drug concentration) per well. Parasites were cultured for 5 - 7 days before 1 ml of each surviving clone was added to 4 ml of fresh drug-containing media in 6 well plates (Corning). The following day, protein (section 2.4.1) or

genomic DNA (section 2.2.1) samples were prepared and either western blot (section 2.4.3) or PCR (section 2.2.2) analysis was performed to confirm the correct generation of modified cell lines.

### 2.1.4 Stabilate preparation and retrieval

Parasites were grown to ~1 x $10^6$ cells/ml before stabilates were prepared. 800 µl of culture was added to 200 µl of 50% glycerol (diluted in the appropriate drug-free media) in 1.8 ml cryovials. Stabilates were stored in a box containing cotton wool and placed at -80°C to allow slow cooling. Parasites were kept at -80 °C for short term storage and transferred to liquid nitrogen for longer storage. To retrieve cells, stabilates were defrosted at room temperature and added to 9 ml of appropriate drug-free media. After 24 hr cells were passaged as a 1:1,000 dilution into media containing any required selective drugs as normal.

### 2.1.5 Growth curves

To begin each growth curve cultures at a density of 5-9 x $10^5$ cells/ml were used to ensure parasites were in log phase. In the case of the RH2A$^{RNAi}$ cell line, cultures were diluted to 1 x $10^5$ cells/ml in 1.2 ml in 12 well plates and tet was added to final concentration of 1 µg/ml to induce RNAi. Every 24 hours, un-induced samples were diluted to a density of 1 x $10^5$ cells/ml to prevent parasites reaching stationary phase and final numbers were corrected according to this dilution to generate cumulative growth curves. For all other cell lines growth curves began at 1 x $10^4$ cells/ml in 1.2 ml and no dilutions were performed.

## 2.2 Basic molecular techniques

### 2.2.1 Genomic DNA extraction

Genomic DNA (gDNA) extracts were made from approximately 4 x $10^6$ BSF cells harvested by centrifugation at 432g for 10 min. Pelleted parasites were resuspended in 1 ml 1X Phosphate Buffered Saline (PBS) to wash away residual media and transferred into a 1.5 ml microcentrifuge tube before centrifuging at 664 g for 3 min to collect cells. The supernatant was removed and parasites were

resuspended in 200 µl fresh 1X PBS. The DNeasy Blood and Tissue DNA extraction kit (Qiagen) was then used following the manufacture's protocol to extract DNA. Samples were eluted in 50 µl of either Buffer AE or $ddH_2O$ and stored at 4 ˚C.

### 2.2.2 Polymerase chain reaction (PCR)

The following standard 50 µl PCR master mix was used to amplify gene fragments for cloning into endogenous tagging and gene deletion constructs: 1X Phusion® HF Buffer (New England Biolabs; NEB), 2.5 µl forward primer (10 pmol/ml stock) and 2.5 µl reverse primer (10 pmol/ml stock), 1 µl 10 mM dNTP mix (Promega) 4 µl gDNA (prepared as in section 2.2.1), 0.5 µl Phusion® High-Fidelity DNA Polymerase (NEB) and $ddH_2O$ to a total volume of 50 µl. PCR was carried out with the following thermocycler sequence: 98˚C for 3 min, 30 cycles of 98˚C for 30 seconds, 54-65˚C for 45 seconds and 72˚C for 1 min per kb (1 min minimum), ending with 10 min at 72˚C. Amplification of gene fragments to be used in Gateway® cloning to generate RNAi constructs was done using the following reaction: 10 µl 5X Phusion® HF Buffer (NEB), 2.5 µl forward primer (10 pmol/ml stock) and 2.5 µl reverse primer (10 pmol/ml stock), 1 µl 10 mM dNTP mix (Promega), 2 µl gDNA, 0.5 µl Phusion® High-Fidelity DNA Polymerase (NEB), 32 µl ddH2O. PCR was then performed with the following thermocycling conditions: 98˚C for 3 min, 30 cycles of 98˚C for 30 seconds, 60˚C for 30 seconds and 72˚C for 30 seconds, ending with 10 min at 72˚C. Where PCR was used to confirm the correct integration of a construct into a mutant cell line or presence/loss of an ORF, Taq Polymerase (NEB) was used as appose to Phusion Polymerase, as was Thermo Buffer (NEB), and the extension temperature was reduced to 68 ˚C. All other reagents and parameters were as described above.

| No. | Target Gene | Sense | Sequence (5'-3') | Cloning sites | Purpose |
|---|---|---|---|---|---|
| 1 | TbRH1 | Fwd | CGACGAAGCTTCTGCGGATGACGGTAATG | HindIII | Myc epitope tagging |
| 2 | | Rev | CGACGTCTAGAGCCAAAGGGCGATTCACA | XbaI | |
| 3 | TbRH1 | Fwd | CCTGTTTTGGTGGGTGTAA | N/A | Correct tag integration |
| 4 | pNAT$^{x12M}$ | Rev | GACAGAAGGACGAAGGAG | N/A | Correct tag integration |
| 5 | TbRH2A | Fwd | CGACGAAGCTTGAACACGCTTAGCCATCAAAC | HindIII | HA epitope tagging |
| 6 | | Rev | CGACGTCTAGAAGGGACTTCCCGCGACAAA | XbaI | |
| 7 | TbRH2A | Fwd | GAGGAAAAGCGTTGTTAGC | N/A | Correct tag integration |
| 8 | TbRH3 | Fwd | CGACGAAGCTTCCACACTTACATTGACACGG | HindIII | Myc epitope tagging |
| 9 | | Rev | CGACGTCTAGATGAGTGGCTATCTTTCATTTTG | XbaI | |
| 10 | TbRH3 | Fwd | CCGATGAGGAGTTTCAGAT | N/A | Correct tag integration |
| 11 | TbRH1 | Fwd | CGACGGGATCCTTGCCTTACCCGTGTTTT | BamHI | 5' UTR amplification |
| 12 | | Rev | CGACGTCTAGACCTTTTCTTTCCCATGGAC | XbaI | |
| 13 | TbRH1 | Rwd | CGACGCCCGGGAGGTGTGTATGGGAATGA | SmaI | 3' UTR amplification |
| 14 | | Rev | CGACGCTCGAGGCACCACCCAGTATAGAAA | XhoI | |
| 15 | NEO | Fwd | GGATTCATCGACTGTGGC | N/A | Correct 3' NEO integration |
| 16 | TbRH1 | Rev | GGGCAAAAAAGTCCACCA | N/A | Correct 3' NEO/BSD integration |
| 17 | BSD | Fwd | TTTTCTCGTCTTGGTGTG | N/A | Correct 3' BSD integration |
| 18 | TbRH1 | Fwd | GTGGAGGAAGCGGTCTTT | N/A | Correct 5' NEO/BSD integration |
| 19 | NEO | Rev | GCGTGCAATCCATCTTGTTC | N/A | Correct 5' NEO integration |
| 20 | BSD | Rev | AAGCGGTCTTTCATATGG | N/A | Correct 5' BSD integration |
| 21 | TbRH1 | Fwd | GAGCCCGTTTTAAGTCTT | N/A | TbRH1 ORF |
| 22 | | Rev | CTGATTGGTCTGGGGTTC | N/A | |
| 25 | TbRH2A | Fwd | CGACGGCGGCCGCGGATCCCTTGAGTGGGCATAGTTA | BamHI + NotI | 5' UTR amplification |
| 26 | | Rev | CGACGAAGCTTTTACTACTGGAAGACCGA | HindIII | |
| 27 | TbRH2A | Fwd | CGACGATCGATTCCCTTAGACACCCCTGA | ClaI | 3' UTR amplification |
| 28 | | Rev | CGACGGCGGCCGCCTCGAGGGACTGAGATGATGCATAA | XhoI + NotI | |
| 29 | TbRH2A | Fwd | GTTTCGTTCTTTCGTTCC | N/A | Correct 3' NEO/BSD integration |
| 30 | TbRH2A | Fwd | CACGCTTAGCCATCAAAC | N/A | TbRH2A ORF |
| 31 | | Rev | CCATCACCAGAACATTTG | N/A | |
| 32 | TbRH2A | Fwd | GGGGACAAGTTTGTACAAAAAAGCAGGCTCCGCAGCTATGACAGGTGTA | AttB | TbRH2A RNAi fragment amplification |
| 33 | | Rev | GGGGACCACTTTGTACAAGAAAGCTGGGTCTCGAAGACGATAGGGATGC | AttB | |

**Table 2.2 Primers used for cloning and mutant cell line confirmation.**
Sequences of primers used for construct generation and confirmation of mutant *T. brucei* cell lines are shown along with their targets. Restriction digest and attB cloning site sequences are underlined. All sequences are shown 5'-3'.

### 2.2.3 Agarose gel electrophoresis

DNA fragments were separated with 1% (w/v) UltraPure™ Agarose (Life Technologies) in gels made with 1X TAE buffer (40 mM Tris base, 19 mM acetic acid, 1mM EDTA) and SYBR® Safe DNA Gel Stain (Life Technologies), added at a 1:20,000 dilution. Loading dye was added to DNA samples before loading alongside 8 µl 1 Kb Plus ladder (Invitrogen™, Life Technologies). Gels were run for 30-60 min in 1X TAE buffer at 100-120 v.

### 2.2.4 Restriction digest

All enzymatic restriction digest reactions were set up as multiples of the following 30 µl reaction according to the amount of DNA which required digestion: 3 µl 10X appropriate buffer, 1-6 µl of eluted DNA, 0.5 µl restriction enzyme(s), $ddH_2O$ up to 30 µl. Reactions were incubated at the appropriate temperature for the enzymes used, usually 37 °C, for 1-12 hr. Products were resolved by agarose gel electrophoresis (section 2.2.3) and the target fragments were purified (section 2.2.5).

### 2.2.5 DNA extraction from agarose gel

DNA bands were visualised on a UV light box, cut from agarose gel with a sterile razor blade and transferred into 1.5 ml microcentrifuge tubes. The DNA was then purified using the QIAquick® Gel Extraction Kit (Qiagen) as per the manufactures protocol. DNA samples were eluted in 30-50 µl of Buffer EB or $ddH_2O$.

### 2.2.6 DNA ligation

Ligation reactions between digested DNA fragments and construct backbones were set up as follows: 1 µl 10X T4 DNA Ligation Buffer (NEB), 0.5 µl T4 DNA Ligase (NEB), 6.5 µl plasmid backbone and 2 µl purified insert fragment. Reactions were incubated at 16 °C overnight. 3 µl of ligation reaction was used to transform 50 µl of *E. coli* MAX Efficiency® DH5-α™ competent cells (Life Technologies) as described in section 2.2.8.

### 2.2.7 Gateway® cloning to generate RNAi constructs

A derivation of Gateway® cloning technology was used to produce tet-inducible RNAi constructs (Jones *et al.*, 2014). PCR to generate gene fragments was performed as described in section 2.2.2 using gene specific primers designed to include attB sites (Table 2.2). Products were resolved on 1% (w/v) agarose gel and extracted, as described in sections 2.2.3 and 2.2.5 respectively. The Gateway® BP Clonase® II Enzyme mix kit (Life Technologies) was used to perform a BP reaction between the PCR product and pGL2084 vector as follows: 5 µl PCR product, 1 µl pGL2084, 3.75 µl TE buffer (10 mM Tris, 1mM EDTA) and 0.25 µl BP Clonase® II (Jones *et al.*, 2014). Reactions were incubated at room temperature before being stopped by the addition of 1 µl Proteinase K solution (2 µg/µl) and incubation at 37 ˚C for 10 minutes. 3 µl of the reaction mix was then used to transform 50 µl of *E. coli* MAX Efficiency® DH5-α™ competent cells (Life Technologies) as described in section 2.2.8.

### 2.2.8 *E. coli* transformation and plasmid purification

*E. coli* MAX Efficiency® DH5-α™ competent cells (Life Technologies) were incubated on ice for 30 minutes with the construct of interest. Cells were then heat shocked at 42 ˚C for 45 seconds and returned to ice for 5 min. 1 ml of S.O.C buffer (2% tryptone, 0.5% yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM MgCl2, 10 mM MgSO4, and 20 mM glucose) was added to the cells which were incubated at 37˚C for 1 hr with agitation. Cells were then pelleted by centrifuging at 700 g for 1 min. Most of the supernatant was removed leaving ~50 µl in which cells were gently resuspended and transferred to a selective agar plate containing 100 µg/ml Ampicillin (Amp). Plates were incubated over night at 37 ˚C. Once established, colonies were picked using a pipette tip and added to 5 ml L-Broth containing 100 µg/ml Amp, then incubated at 37 ˚C overnight in a shaking incubator. Transformed *E. coli* DH5-α cultures were pelleted by centrifuging at 25,000 g for 10 minutes and the supernatant was removed. QIAprep Spin Miniprep Kit (Qiagen) was then used according to the manufactures protocol to purify plasmid DNA from the pelleted bacteria.

### 2.2.9 DNA sequencing

For Sanger sequencing to confirm the generation of correct DNA constructs, 2 µl of sequencing oligos (10 pmol/ml) listed in Table 2.3 were added to 15 µl plasmid DNA (prepared as in section 2.2.8) in Mix2Seq bar-coded tubes. DNA sequencing was carried out by Eurofins Genomics (Ebersber, Germany) and the resulting sequence was aligned to the appropriate reference construct using CLC Genomics Workbench. For next-generation sequencing applications, libraries were prepared by Craig Lapsley (section 2.6.4) and sequenced with the Illumina NextSeq 500 platform by Glasgow Polyomics.

| No. | Target | Sense | Sequence (5'-3') |
|---|---|---|---|
| 78 | pNAT$^{x12M}$ contruct | Fwd | CGTTGGCCGATTCATTAATGC |
| 79 | | Rev | TAATGACGAACGGGAAATGC |
| 80 | KO constructs | Fwd | AGGGTTTTCCCAGTCACGACGTT |
| 81 | | Rev | GAGCGGATAACAATTTCACACAGG |

**Table 2.3 Primer sequences used for Sanger sequencing.**
Primers used for Sanger sequencing (Eurofins Genomics) are shown 5'-3' along with their target constructs.

## 2.3 RNA analysis

### 2.3.1 RNA extraction

1 x $10^7$ cells were collected by centrifugation at 432g for 10 min, media was removed, and the cells were washed in 1 ml 1X PBS. Cells were then pelleted at 664 g for 3 min. Pelleted cells were either stored at -80 ˚C or used immediately for total RNA extraction. RNeasy Mini Kit (Qiagen®) was used following the provided protocol "Purification of Total RNA from Animal Cells Using Spin Technology", starting by adding 600 µl RLT Buffer to each sample and pipetting up and down several times to ensure the cells were homogenized. On-column DNase I (RNase-Free DNase Set, Qiagen®) treatment was performed after step 5 of this protocol where the column was incubated for 15 minutes with DNase I. RNA was eluted in 30 µl RNase-free water and stored at -80 ˚C.

### 2.3.2 First-strand cDNA synthesis

The SuperScript™ First-Strand Synthesis System for RT-PCR (Invitrogen, Life Technologies) was used with 1 µg of total RNA to generate complementary DNA (cDNA). The manufactures protocol was followed using random hexamer primers and SuperScript™ II Reverse Transcriptase. Synthesis was followed with RNase H treatment as per the protocol to remove the resultant cDNA-RNA hybrids. All cDNA samples were stored at -20 ˚C.

### 2.3.3 Quantitative real-time PCR (qPCR)

Primers, designed using Primer Express version 3.0 software (Bio Rad), were first assessed for efficiency by analysing calibration curves and melting profiles. Those which achieved efficiency of approximately 100% (within a 15% interval) were deemed suitable for use (Table 2.4). For every primer pair, each cDNA sample was run in triplicate using Precision™ qPCR MasterMix with SYBR Green. Samples were set up as follows in MicroAmp® Optical 96-well Reaction Plates (Applied Biosystems®, Life Technologies): 12.5 µl 2X Master Mix, 2µl forward primer (3.3 pmol stock), 2 µl reverse primer (3.3 pmol stock), 5 µl of cDNA (prepared as described in section 2.3.2 and diluted 1:10) and 5.5 µl RNase-free water (Qiagen). Plates were sealed using MicroAmp® clear adhesive film (Applied Biosystems®, Life Technologies) and run using a 7500 Real Time PCR system (Applied Biosystems®) as follows: 10 min at 95 ˚C, 40 cycles of 95 ˚C for 15 sec and 60 ˚C for 1 min. Fluorescence intensity data was collected at the extension step (60 ˚C for 1 min). Amplification curves were analysed using the 7500 software version 2.3 (Applied Biosystems®) and results exported as an Excel (Microsoft Office) spreadsheet. For normalisation between samples, primers targeting the Tb927.10.12970 gene (C1) (Kabani et al 2009) were used as the reference gene. The $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001) allowed analysis by normalising sample $C_T$ to C1 $C_T$ values and calculating the fold-change of induced samples relative to uninduced.

| No. | Target Gene | Sense | Sequence |
|---|---|---|---|
| 23 | TbRH1 | Fwd | ACGAACCCCAGACCAATCAG |
| 24 | | Rev | CGCGCAAGACATGGATGA |
| 34 | TbRH2A | Fwd | CCGAGCGGTGCCAAGA |
| 35 | | Rev | TGCATAAATTCGGAAACATCG |
| 36 | rRNA promoter | Fwd | GTCAATACAACACACAATAGG |
| 37 | | Rev | CTTAACTGAGGAAGTGTCATA |
| 38 | rRNA 5.8S | Fwd | GGATGACTTGGCTTCCTAT |
| 39 | | Rev | ATTGATACCACTTATCGCACT |
| 40 | rRNA 28Sβ | Fwd | GTAAGTTCGCAAGAAGCAT |
| 41 | | Rev | ACCAGAAGGAGGTTAGTAGATA |
| 42 | NEK22 | Fwd | TTACTCCTCGCCTCTCATT |
| 43 | | Rev | TGTTCTGGACGTAGGAACTA |
| 44 | HDAC3 | Fwd | CGA CAG CGT GAG GAA TAA G |
| 45 | | Rev | CAC GGG AGA CCT TTC ATT C |
| 46 | ATR | Fwd | TCTAGTATCGCCGGCAGCTT |
| 47 | | Rev | TTGGCTGTATGGCCTGATCTC |
| 48 | GPI-8 | Fwd | TCTGAACGCGCGCACTTC |
| 49 | | Rev | CCACTCACGGACTGCGTTT |
| 50 | ORC1B | Fwd | ACGTCAACTGTGCGGATATG |
| 51 | | Rev | TCCAAGCGAACCTGTGAAC |
| 52 | Actin | Fwd | CGGACGAGGAACAAACTGC |
| 53 | | Rev | TTTCCATGTCATCCCAATTGG |
| 54 | tRNA Glycine | Fwd | CGGACTGCGCAGACTTTTC |
| 55 | | Rev | CGGCGGTAGACTATTATCGTTTG |
| 56 | tRNA Lycine | Fwd | CCCCAGCGGCATGTTC |
| 57 | | Rev | TTTGTGTGGAGAATGTTGGATCTAG |
| 58 | snRNA 1 | Fwd | GGGAGGTGTGAACGCAAGA |
| 59 | | Rev | AAGGAGTACCGGATATAGTATTGCACTAG |
| 60 | snRNA U1 | Fwd | TCCACGGGCCCTGTTG |
| 61 | | Rev | AAAATCAGAGCCCCCAAACAC |
| 62 | ESAG6 | Fwd | TGGGAGGGATGGATGTAATTT |
| 63 | | Rev | CCGACCCCCCTTCCAAT |
| 64 | ESAG8 | Fwd | CGGATGCGTCGTGGAA |
| 65 | | Rev | CCTCCGATACGCCGTTGA |
| 66 | VSG221 | Fwd | AGC AGC CAA GAG GTA ACA GC |
| 67 | | Rev | CAA CTG CAG CTT GCA AGG AA |
| 68 | VSG121 | Fwd | AGGAAGGCAAATACGACCAG |
| 69 | | Rev | TTGGGGTAAAAGTCCTTG |
| 70 | VSG800 | Fwd | ACAGACCGCCGACAGTATC |
| 71 | | Rev | GTATCTTTGTAGGCCGCTGC |
| 72 | VSG T3 | Fwd | CAGCGCAAGTCCAGGACG |
| 73 | | Rev | TGCTTCGTCGTCGCTTAC |
| 74 | VSG13 | Fwd | ATAACGCATGGCCATCTTGCA |
| 75 | | Rev | GTCGTTGCTGTGGATTGCTC |
| 76 | VSG224 | Fwd | GACGCAGCAGAATCAACAC |
| 77 | | Rev | GCTTATTTTGTGTCTGTCGC |

**Table 2.4 Primers used for qPCR analysis.**
All primers used for qPCR analysis are listed with the target genes.

### 2.3.4 RNA-seq library preparation and sequencing

Total RNA samples prepared as in 2.3.1 were used to prepare libraries for Illumina sequencing. All RNA library preparation was performed by Craig Lapsley using the TruSeq Stranded mRNA kit (Illumina). mRNA strands were poly(A) selected and Illumina adapters ligated to give a final fragment size of approximately 300 bp. Sequencing was performed with the NextSeq 500 platform giving paired-end reads of 75 bp in length (Glasgow Polyomics).

## 2.4 Protein analysis

### 2.4.1 Whole cell protein extraction

2.5 x $10^6$ cells per sample were harvested by centrifuged at 432g for 10 min. Cells were washed in 1 ml 1X PBS before being resuspended in 10 µl Loading Buffer (1X NuPAGE® LDS Sample Buffer (Life Technologies), 0.1% ß-mecaptoletanol, 1X PBS) per 2.5 x $10^6$ cells. Samples were then boiled for 10 min before being analysed immediately or stored at - 20 ˚C.

### 2.4.2 Sodium-dodecyl-sulphate-polyacrylamide gel electrophoresis (SDS-PAGE)

If previously frozen, protein extracts were boiled again for 10 min. 4-10 µl of sample (1-2.5 x $10^6$ cells) was loaded per well of NuPAGE® Novex® 10-12% Bis-Tris Protein Gels 1.0 mm (Life Technologies), alongside 10 µl Novex® Sharp Protein Standard ladder. Gels were run in 1X NuPAGE® MOPS SDS Running Buffer at 200 v for 50 min.

### 2.4.3 Western blotting

SDS-PAGE separated protein was transferred to PVDF Transfer Membrane (ThermoFisher) in 1X Transfer Buffer (25 mM Tris, pH 8.3, 192 mM Glycine, 20% (v/v) methanol) at 100 v for 70 min. Ponceau S solution (Sigma Aldrich) was used to confirm the transfer of protein to the PVDF membrane. Membranes were subsequently washed twice in 1X TBST (1X PBS pH 7.2, 0.01% Tween-20) for 5 min

on a rocker. A Blocking Solution of 5% dried milk in 1X TBST was used to incubate the membrane for 1 hr at room temperature or overnight at 4 ˚C. Membranes were then incubated with the primary and then secondary antisera in Blocking Solution (see Table 2.5 for dilution factors) for 1 hr each, separated by three 10 min washing steps in 1X TBST. The membrane was washed again three times in 1X TSBT for 10 min, incubated for 2 min with SuperSignal West Pico Chemi-luminescent Substrate (Thermo Scientific) and finally exposed to X-ray film (Kodak).

| Antibody | Serotype | Concentration | Source |
|---|---|---|---|
| α-Myc | mouse | 1:7,000 | Millipore |
| α-HA | mouse | 1:10,000 | Sigma Aldrich |
| α-EF-1α (elongation factor-1 alpha) | mouse | 1:25,000 | Millipore |
| α-yH2A | rabbit | 1:1,000 | R. McCulloch |
| α-mouse HRP conjugate | goat (Ig H+L) | 1:1,000 | Thermo Scientific |
| α-rabbit HRP conjugate | goat (Ig H+L) | 1:5,000 | Thermo Scientific |

**Table 2.5 Antisera used for western blot analysis.**
All antibodies used for western blot analysis are listed along with dilutions used and manufacturers. HRP – horseradish peroxidase.

## 2.5 Immunofluorescent analysis and imaging

### 2.5.1 DAPI staining and cell cycle analysis

Parasites were maintained under 1 x $10^6$ cells/ml to ensure cultures were in log phase, diluting cultures with appropriate media and selective drugs where necessary between time points. Per sample, 1 - 2 x $10^6$ cells were centrifuged at 432g for 10 min. Centrifuged cells were washed in 500 µl 1X PBS and pelleted by centrifugation 664 g for 3 min. Most of the supernatant was remove, leaving ~30 µl which the pellet was resuspended in and transferred to the well of a 12 well glass slide (Thermo Scientific), spread with Poly-L-Lysine (Sigma Aldrich). Cells were allowed to settle for 5 min before fixing with 25 µl 4% FA for 4 min. Cells were washed twice in 50 µl 1 X PBS for 5 min before 5 µl of Fluoromount G with DAPI (Cambridge Bioscience, Southern Biotech) was added and incubated in the dark for 4 min. A glass coverslip was applied, and the slide was sealed with nail varnish.

## 2.5.2 Standard immunofluorescent analysis of BSF *T. brucei*

For standard immunofluorescence to localise epitope-tagged proteins and phosphorylated histone variant, γH2A, 1 - 2 x $10^6$ cells were collected by centrifugation at 432 g, resuspended in 500 µl 1X PBS and pelleted at 664 g for 3 min. All but ~ 30 µl of supernatant was removed and cells were resuspended in the remaining 1X PBS. Cells were transferred to wells of a 12 well slide (Thermo scientific) spread with Poly-L-Lysine (Sigma Aldrich), before being incubated for 5 min allowing cells to settle and adhere. 1X PBS was removed and replaced by 25 µl 4% FA in 1X PBS for 5 min. 0.01% Triton X-100 (Promega) in 1X PBS was added for 5 min to permeabilise the cells before quenching with 25 µl 100 mM glycine twice for 10 min. Cells were then incubated in 50 µl Blocking Solution (3% Bovine Serum Albium (BSA), Sigma Aldrich, 0.002% Tween-20, Sigma Aldrich, 1X PBS) for 1 hr in a humidity chamber. Blocking Solution was then replaced with primary antibody diluted in Blocking Solution to the concentration stated in Table 2.6 and incubated for a further 1 hr. Wells were then washed twice in 50µl 1X PSB for 5 min, before incubation with secondary Alexa Fluor antibodies for 1 hr. Finally, wells were washed three times in 1X PBS for 5 min and Fluoromount G with DAPI (Cambridge Bioscience, Southern Biotech) was added and incubated for 4 min, before a glass coverslip was applied and sealed with nail varnish.

| Antibody | Serotype | Concentration | Source |
|---|---|---|---|
| α-Myc Alexa Fluor® 488 conjugated | mouse | 1:5,000 | Millipore |
| α-HA | mouse | 1:1,000 | Sigma Aldrich |
| α-yH2A | rabbit | 1:1,000 | R. McCulloch |
| α-VSG 221 | rabbit | 1:10,000 | R. McCulloch |
| α-VSG 221 | rat | 1:10,000 | Gift, L. Glover |
| α-VSG 121 | rabbit | 1:10,000 | Gift, L. Glover |
| α-mouse Alexa Fluor® 594 Goat | goat (Ig H+L) | 1:1,000 | Molecular Probes® |
| α-rabbit Alexa Fluor® 594 Goat | goat (Ig H+L) | 1:1,000 | Molecular Probes® |
| α-mouse Alexa Fluor® 488 Goat | goat (Ig H+L) | 1:1,000 | Molecular Probes® |

**Table 2.6 Antisera used for immunofluorescent analysis.**
All antisera used for immunofluorescent detection are listed with dilutions and source. R. McCulloch antisera were produced by previous member of the McCulloch lab.

### 2.5.3 VSG immunofluorescent analysis

VSG immunofluorescent analysis was performed as previously described (Glover et al. 2016). Cells were centrifuged at 432g for 10 mins, resuspended in 1X PBS to wash cells before centrifuging again at 664 g for 3 min to re-pellet cells. 1% FA was used to fix cells in suspension at room temp for 15 min before adhering to a slide coated in Poly-L-Lysine solution (Sigma Aldrich). Cells were then blocked in 50% FBS serum for 15 min before incubation with primary, α-VSG221 and α-VSG121, and secondary, Alexa Fluor 594 goat α-rabbit (Molecular Probes) antibody and Alexa Fluor 488 goat anti-rat (Molecular Probes), antibodies were carried out at room temp for 45 min in both cases. Cells were then mounted in Fluoromount G with DAPI (Cambridge Bioscience, Southern Biotech).

### 2.5.4 Assay of EdU incorporation

For 5-ethynyl-2'-deoxyuridine (EdU) incorporation assays, parasites were culture in thymidine-free HMI-11 media (section 2.1.2) and were maintained in log phase (1 x $10^5$ - 1 x $10^6$ cells/ml) during time course experiments. 4 hr prior to each time point, 4 ml of culture was transferred to a 6-well plate and incubated with 150 μM EdU (Life Technologies). After 4 hr, cells were collected by centrifugation at 432 g for 10 min. The media was removed, and cells were washed in 1 ml 1X PBS before centrifuging again at 664 g for 3 min. The majority of the supernatant was then removed leaving ~ 30 μl. Parasites were resuspended and transferred to wells of a 12 well glass slide (Thermo Scientific) spread with Poly-L-lysine (Sigma Aldrich), where they were left to settle for 5 min. Supernatant was then removed and parasites were incubated in 20 μl 3.7% FA for 5 min at RT. FA was removed, and cells were washed in 25 μl 3% BSA for 5 min, twice. After removing 3% BSA, 20 μl 0.2% Triton X-100 (Promega) was added and cells were incubated for 10 min at RT to permeabilise. Cells were then washed twice with 3% BSA for 5 min. Click-iT® (Life Technologies) reaction mix was prepared as multiples of the following for each sample: 21.5 μl 1X Reaction Buffer, 1 μl $CuSO_4$, 0.25 μl Alexa Fluor 555 Azide and 2.5 μl 1X Additive Buffer. BSA was removed and parasites were incubated for 1 hr at RT in 25 μl of the Click-iT® reaction mix, in a dark chamber. After incubation, cells

were washed 6 times in 3% BSA. For dual staining with γH2A antiserum, samples were then incubated with primary γH2A antiserum in 3% BSA for 1 hr, washed 3 times in 3% BSA and then incubated in anti-rabbit Alexa Fluor 594 (Table 2.6) for 1 hr. Finally, parasites were washed 3 times in 1X PBS before 5 μl Flouromount G with DAPI (Cambridge Bioscience, Southern Biotech) was added and coverslips were seals to slides with nail varnish.

### 2.5.5 Fluorescent imaging and analysis

For counting purposes, parasites were imaged using an Axioscope 2 fluorescence microscope (Zeiss) and a Zeiss PlanApochromat 63x/1.40 oil objective. DAPI staining was visualised with 440 nm filter, Alexa Fluor 488 staining with a 520 nm filter and Alexa Fluor 594 staining with a 620 nm filter. Background fluorescence was subtracted from images using Fiji software rolling ball background subtraction plugin (http://imagej.net/Rolling_Ball_Background_Subtraction). Counting was performed with Fiji software (Schneider, Rasband and Eliceiri, 2012). For increased resolution, an Olympus IX71 DeltaVision Core System microscope (Applied Precision) was used, running a SoftWoRx Suite v2.0 (Applied Precision) software. Either an Olympus PlanApo 60x/1.42 or an UplanSApo 100x/1.40 oil objectives were used. Z-stacks of 5-6 μm thickness were acquired in 25 sections, and then deconvolved with the 'conservative' method and high noise filtering. Fiji software was then used to generate maximum projection images.

### 2.5.6 Super-resolution structured illumination microscopy (SR-SIM)

SR-SIM was performed with imaging specialist Dr Leandro Lemgruber Soares, who also performed image processing. Imaging was performed with an ELYRA PS.1 Microscope (Zeiss), using a Plan-Apochromat 63x/1.40 Oil DIC objective and 405, 488 and 594 nm lasers. Z-stack sections were ~0.15 μm in thickness and totalled ~10 μm. Image reconstruction was performed with ZEN Black software (Ziess) and 3D rendering with Imaris software (Bitplane) to produce 3D models.

# 2.6 Chromatin-immunoprecipitation (ChIP)

## 2.6.1 Fixing and Shearing Chromatin

The ChIP-IT® Express Enzymatic (Active Motif) kit was used to perform ChIP reactions. Approximately 1 x $10^8$ BSF cells were pelleted by centrifugation at 1,200 g at 4 °C for 10 mins. The media was removed, and pellets were resuspended in 10 ml Fixing Solution (10 % FA, 1X HMI-9, Gibco) and incubated shaking at room temperature for 5 min. 10 ml 2X Glycine-Stop Solution (2 ml 10X Glycine Buffer, 8 ml 1X PBS) was added to fixed cells and briefly mixed by pipetting. Cells were pelleted by spinning at 1200 g at 4°C for 10 min and the supernatant was removed. The cells were then washed in 1X Glycine-Stop Solution (1 ml 10X Glycine Buffer, 9 ml 1X PBS) and pelleted as before. Glycine-Stop Solution was removed, and the pellet was resuspended in 500 µl Lysis Buffer (supplemented with 2.5 µl 100 mM PMSF and 2.5 µl Protease Inhibitor Cocktail (PIC)) and incubated on ice for 30 minutes. The suspension was transferred to an ice-cold dounce homogenizer and repetitively dounced until all nuclei were realised. Cells were transferred to microcentrifuge tubes and the nuclei were pelleted by centrifuging at 2,400 g at 4°C for 10 min. The supernatant was discarded, and the nuclei pellet was resuspended in 350 µl Digestion Buffer (supplemented with 1.75 µl 100 mM PMSF and 1.75 µl PIC). This was incubated at 37°C for 5 min before 17 µl Enzymatic Shearing Cocktail (200 U/ml, 50% glycerol) was added and chromatin was incubated for a further 5 min at 37°C. The digest reaction was stopped with the addition of 7 µl ice-cold 0.5 M EDTA and incubated on ice for 10 minutes. Lastly, sheared chromatin was centrifuged at 18,000 g at 4°C for ten min and the supernatant was collected in fresh microcentrifuge tubes. Samples were stored at -80°C until use.

A 50µl aliquot of each chromatin sample was reverse crosslinked and cleaned according to the DNA Clean Up protocol of the ChIP-IT® Express Enzymatic kit. The concentration of dsDNA in each sample was determined using the Qubit 3.0 Fluorometer (Thermo Fisher Scientific).

## 2.6.2 Chromatin-immunoprecipitation (ChIP)

Sheared chromatin was thawed on ice and used in 200 µl IP reactions according to the ChIP-IT® Express Enzymatic kit (Active Motif) instructions. Each reaction contained 1-2 µg chromatin, 25 µl Protein G magnetic beads, 20 µl ChIP Buffer 1, 2 µl PIC, antibody (4.5 ng S9.6, Kerafast or 10.7 µl α-γH2A, R. McCulloch) and $ddH_2O$ to 200 µl in total. Reactions were incubated at 4°C overnight on an end-to-end rotator. The following morning the microcentrifuge tubes were placed on a magnetic rack to collect the beads. Supernatant was removed, and beads were washed once in 800 µl ChIP Buffer 1 and twice in 800 µl ChIP Buffer 2. At this point ChIP samples using the S9.6 antibody (DRIP) were used to perform an on-bead RNase H control (see section 2.6.3). To elute DNA the beads were resuspended in 50 µl elution Buffer AM2 and incubated at room temperature on an end-to-end rotator for 15 min. 50 µl Reverse Cross-Linking Buffer was added and mixed by pipetting. The beads were pelleted using the magnetic stand and supernatant containing the eluted chromatin was transferred to fresh microcentrifuge tubes. At this point 10 µl of Input sheared chromatin (200 - 300 ng) was thawed on ice. 78 µl of ChIP Buffer 2 and 2 µl 5M NaCl was added to make a total volume of 100 µl. Both eluted ChIP and Input samples were then incubated at 65°C for 3 hr. To degrade protein in the samples, 2 µl Proteinase K was added to each ChIP and input sample and incubated at 37°C for 1 hr. Lastly 2 µl of Proteinase K Stop Solution was added to each sample and DNA was stored at -20 °C.

## 2.6.3 RNase H treatment of S9.6 DRIP samples

In the case of S9.6 ChIP samples (DRIP), each IP reaction was set up in duplicate as described in section 2.6.3. After washing ChIP samples, beads were resuspended in 300 µl RNase H Buffer (50 mM Tris-HCl, 75mM KCl, 3 mM MgCl2, 10 mM DTT, 4% glycerol and 20 µg/ml BSA). 15 µl of E. coli RNase H (30 Units, Invitrogen) or RNase-free $ddH_2O$ was added to samples which were then incubated at 37°C with shaking at 1,000 rpm for 3 hr. Beads were then washed as before; once with 800 µl ChIP Buffer 1 and twice with 800 µl ChIP Buffer 2. DNA was then eluted for each sample as described in section 2.6.2.

### 2.6.4 DNA library preparation

The TrueSeq ChIP Library Preparation Kit (Illumina) was used to prepare all γH2A ChIP and DRIP DNA samples for next-generation sequencing with the exception of WT and *Tbrh1*-/- γH2A ChIP samples where the NEBNext® ChIP-sep Library Preperation kit was used. In all cases 300 bp fragments (including adaptor sequences) were selected for with Agencourt AMPure XP beads (Beckman Coulter). Preparation of all samples was carried out by Craig Lapsley.

### 2.6.5 ChIP-qPCR

qPCR was performed using DRIP/ChIP input and eluted (IP) DNA samples diluted 1:100 and 1:10 respectively. 1 μl of DNA and 400 nM of primers (Table 2.4) were added to SYBR Select Master Mix (Life Technologies) to a total of 21 μl per reaction. All samples were run using a 7500 Real Time PCR system (Applied Biosystems) using the following conditions: 50 °C for 2 min and 95 °c for 2 min, followed by 40 cycles of 95 °C for 15 sec, 59 °C for 15 sec and 72 °C for 1 min. Fluorescence intensity was measured at the end of extensions step, 72 °C for 1 min. $C_T$ values were adjusted according to the dilution factor of each samples (minus log2(dilution)) and the percentage of input DNA from in the IP DNA of each sample was calculated. The calculation methodology is described by Thermo Fisher Scientific (https://tinyurl.com/y7uuzvay).

## 2.7 Flow cytometry

### 2.7.1 Cell cycle analysis by flow cytometry

For each sample, 3 x $10^6$ cells were harvested by centrifugation at 432 g for 10 min. Media was removed, and pellet was resuspended in 800 μl 1X PBS. Cells were centrifuged again at 664 g for 3 minutes and the supernatant was removed. Cells were then suspended in 200 μl 1% FA and incubated at RT for 10 min. Fixed cells were pelleted by centrifuging at 6,000 g for 3 min, supernatant was removed, and parasites were resuspended in 1X PBS for storage at 4 °C. Once collected, all samples were centrifuged again at 6,000 g for 10 min. PBS was removed, and cells

were resuspended in 0.01% Triton X-100 (Promega) in 1X PBS before incubating on ice for 30 min. Cells were pelleted again by centrifugation at 6,000 g for 10 min and supernatant was completely removed. Cells were then resuspended in 1ml of 100 µg/ml RNase A (Sigma Aldrich), 15 µg/ml propidium iodide, 1 X PBS and incubated at 37 °C for 30 min before analysing with the BD FACSCalibur™ (BD Biosciences). Data was analysed on FlowJo_V10™ software (FlowJo, LLC).

## 2.8 Bioinformatics

### 2.8.1 Sequence retrieval

DNA and amino acid sequences of interest were downloaded in FASTA format from either TriTrypDB (Aslett *et al.*, 2010) or GenBank® (http://www.ncbi.nlm.nih.gov/genbank/) databases. For construct assembly and mutagenesis, the latest gene sequences were retrieved from the *T. brucei* Lister427 v4 assembly. For DRIP-seq, ChIP-seq and RNA-seq analysis (targeting the whole transcriptome) a hybrid genome containing the 11 Mb chromosome assemblies of the *T. brucei* TREU927 v5.1 genome, 14 BES contigs and 5 mVSG ES contigs (Hutchinson, Glover and Horn, 2016) was used for alignments. For RNA-seq analysis of VSG transcripts specifically, reads were aligned to a collection of 2470 VSG coding regions of the *T. brucei* Lister427 strain (Cross, Kim and Wickstead, 2014).

### 2.8.2 Primer and construct design

The CLC Genomics Workbench V7 (CLC Bio, Qiagen®) was used to design primers for the amplification of gene fragments to be used in endogenous tagging and gene deletion constructs. CLC was also used to generate construct maps, align sequencing data and identify restriction digest sites. Primers to be used in the generation of RNAi constructs were generated using the TrypanoFAN:RNAit target selection script (http://trypanofan.bioc.cam.ac.uk/software/RNAit.html). This online service uses MIT Primer3 and NCBI Blast to identify suitable primers to amplify a targeting gene fragment. The ORF of targeted genes was uploaded and primers were generated using default parameters. attB sites were then added to the 5' end to give the final primers.

## 2.8.3 Protein sequence alignment

Amino acid alignments were completed using the multiple sequencing alignment (MSA) tool Kalign from EMBL-EBI (http://www.ebi.ac.uk/Tools/msa/kalign/). Kalign is a MSA tool with equal accuracy to other popular methods but is up to ten times faster and concentrates on local regions (Lassmann and Sonnhammer). FASTA sequences were uploaded to the online form and output was given in a ClustalW format using default settings. Annotations were made manually using annotation information obtained from NCBI databases (http://www.ncbi.nlm.nih.gov/protein).

## 2.8.4 DRIP/ChIP-seq analysis

### 2.8.4.1 Read mapping and filtering

Sequencing read quality was first checked using FastGC, before reads were trimmed using TrimGalore (https://github.com/FelixKrueger/TrimGalore) under default settings to remove Illumina adaptor sequences and bases with quality scores < 20. Remaining reads were then mapped to the hybrid reference genome (see above) using Bowtie2 v2.2.9 (Langmead and Salzberg 2012) in "very-sensitive" mode. Reads with a MapQ value <1 were removed using SAMtools (Li et al. 2009) leaving at least 30 million aligned reads per sample (Hutchinson, Glover and Horn, 2016). Quality control read mapping and filtering was carried out by Dr. Kathryn Crouch.

### 2.8.4.2 Calculating DRIP/ChIP-seq signal enrichment

The fold-change between the read depths of input and IP DNA was found for each sample in 50 bp bins using the DeepTools bamCompare tool (library size was normalised by the SES method and fold-change was expressed as a ratio) (Ramirez *et al.*, 2014) and visualised as tracks with IGV (Robinson et al. 2011). In the case of γH2A ChIP-seq mapping in $RH2A^{RNAi}$ parasites, fold-change was calculated in the same manner using DeepTools bamCompare in tet-induced data relative to un-induced.

#### 2.8.4.3 Defining and classifying DRIP enriched regions

To identify regions enriched of RNA-DNA hybrids, the 50 bp bins with a fold-change over 1.2 were extracted as a list of chromosomal start and end coordinates. Custom Python script, DRIP_Enriched_Regions.py (Figure 8.1), was used to find the start and end coordinates of adjacent 50 bp bins, without allowing gaps, to collate a list of RNA-DNA hybrid signal enriched regions.

Bedtools (v2.26.0) Intersect (Quinlan and Hall, 2010) was used to ask if each DRIP enriched region overlapped with genome annotations of SSRs, CIR, VSG arrays, RHS, snRNA, tRNA, rRNA, procyclin, mVSG and VSG expression sites, all inferred from the reference genome annotation and supplied as lists of coordinates in BED files. Results were exported to Excel (Microsoft) and the number of enriched regions overlapping the features within each group was calculated. The remaining DRIP enriched regions were assumed to be located within Pol II transcribed PTUs. These were further classified as being associated with the 5' UTR, 3'UTR or CDS, this time asking only where the middle base of each enriched region located. The remaining enriched regions were assumed to be associated with the regions between UTRs (or within unannotated UTRs) and were classified as intergenic.

#### 2.8.4.4 Plotting DRIP/ChIP-seq signal profiles and heatmaps

Normalised ratio files (described in section 2.8.4.2) were used to generate plots and perform K means clustering analysis using the deepTools tools, computeMatrix, plotProfile and plotHeatmap (Ramirez et al. 2014). The computeMatrix function calculates scores (in this case based on the fold-change ratio) per genome regions, supplied as chromosome coordinates in BED format. Matrices of the scores for all regions listed in the BED files were generated and either exported and plotted using GraphPad Prism or used as an intermediate file for plotProfile and plotHeatmap functions.

#### 2.8.4.5 GC/AT skew analysis

GC and AT skew were calculated using the custom script GC/ATskew.py (Figure 8.2) which used the following formulas: GC skew = (#G - #C)/(#G + #C), AT skew = (#A -

#T)/(#A + #T). The skew was calculated in 20 bp bins across a list of regions given as coordinates in a BED file. The average GC and AT skew value for each bin in all regions listed in BED was exported and plotted using GraphPad Prism.

#### 2.8.4.6 Comparing DRIP-seq data with gene expression

Rate of mRNA decay has previously been analysed by Fadda *et al.*, (2014). Genes analysed were ordered according to the rate of mRNA decay in BSF parasites, excluding those for which no data is available. The 690 genes with the fastest rates of mRNA decay were taken as a group of low expression genes and 690 with the slowest rates of decay were classified as high expression genes. 690 genes taken from the centre of the list were taken as an intermediate expression sample. DRIP-seq signal profiles were generated for each group as described in section 2.8.4.4.

#### 2.8.4.7 Predicting R-loop forming regions

R-loop predications were made using the QmRLFS-finder algorithm (Jenjaroenpun *et al.*, 2015) that is based upon the findings by (Roy and Lieber, 2009), where an R-loop forming sequencing (RLFS) was defined as containing three parts: an initiation zone containing G-clustering, a G-rich elongation zone where the R-loop can extend, and a linking sequence between 0 and 50 bp of any composition. To predict RLFS in the *T. brucei* genome, G-clustering was defined as three clusters containing ≥ 3 G residues, or two clusters of ≥ 4 G residues. In both cases clusters were separated by 1 - 10 bp. Elongation zones were defined as comprising ≥ 40% G residues and 100 - 2000 bp in length.

### 2.8.5 RNA-seq analysis

Read mapping, filtering, count analysis and differential expression analysis was carried out by Kathryn Crouch.

#### 2.8.5.1 Read mapping and filtering

As for DRIP-seq data, sequencing reads were first quality checked with FastQC before trimming with Trim Galore to remove Illumina adaptor sequences and base

with quality scores < 20. Any resulting reads < 20 bp in length were discarded along with the associated paired read. Reads were then mapped with HiSat2 v2.0.5 using default parameters with the exception of not permitting splice alignments (--no-splice-alignment), to either the hybrid genome or VSG collection (section 2.8.1). Mapped reads were then filtered to remove reads with MAPQ scores <1 (Hutchinson, Glover and Horn, 2016).

#### 2.8.5.2 Read count analysis

Mapped reads were counted using HTseq-count software, using default parameters. Here only uniquely mapping reads were considered and those mapping to rRNA genes were masked. For alignments to the hybrid reference genome, strand information was considered when carrying out read count analysis. However, strand information was ignored (--stranded no) when analysing alignment to the VSG collection (Cross, Kim and Wickstead, 2014), as no strand information was available for these coding regions.

#### 2.8.5.3 Differential expression analysis

Normalisation and differential expression between data sets was carried out using DESeq2 v1.18.1 (Love, Huber and Anders, 2014). Data from the 24 hr TbRH2A RNAi samples were considered separately from data generated at 36 hr. In each case HTSeq counts data were used from each duplicate sample, to compare counts for each gene between tet-induced un-induced samples. Those with an adjusted p value < 0.05 were considered significantly differentially expressed.

#### 2.8.5.4 RPKM calculations

In order to compare read alignment to VSG coding regions in WT and *Tbrh1*-/- samples the number of reads per kb of transcript per million mapped reads (RKPM) was calculated manually for each VSG (Cross, Kim and Wickstead, 2014) using counts data generated with HTseq (section 2.8.5.2) and the following formulas:

$$\text{RPK} = \text{read counts}/\text{gene length (kb)}$$

$$RPKM = RPK/(\text{total RPK}/1{,}000{,}000)$$

Fold-change in RPKM for each VSG was then calculated for *Tbrh1*-/- relative to WT.

#### 2.8.5.5 Gene ontology (GO) term analysis

For GO term enrichment analysis of genes which have R-loops associated with their coding regions (CDS), TriTrypDB (Aslett *et al.*, 2010) biological processes GO term analysis was performed in the *T. brucei* TRUE927 background and a significance cutoff of $p < 0.05$. For analysis of genes found to be differentially expressed in RNA-seq experiments, GO term analysis was performed using Cytoscape v3.6.1 (Shannon *et al.*, 2003) plugin BiNGO (Maere, Heymans and Kuiper, 2005). BiNGO analysis was performed with hypergeometric statistical testing of significance, and multiple testing was corrected for with the Benjamini and Hochberg False Discovery Rate (FDR) correction. FDR adjusted P values < 0.01 were deemed significantly enriched terms. Networks were mapped also using Cytoscape v3.6.1.

# 3 Characterisation of *T. brucei* RNase H Enzymes

## 3.1 Introduction

RNase H enzymes are endonucleases characterised by the capacity to hydrolyse the RNA strand of RNA-DNA hybrids, frequently those which form as R-loop structures within the DNA helix. Of those studied, most organisms possess at least one RNase H, though most commonly two distinct catalytic proteins are found. Eukaryotic RNases H are larger and more complex than their prokaryotic counterparts, though similarities in structure and function exist. Since their discovery in calf thymus extracts (Stein and Hausen, 1969) and, later, in *E. coli* (Itaya, 1990), standard nomenclature to describe these enzymes has been adopted for clarity: Arabic numerals were assigned to eukaryotic RNase Hs, RNase H1 and RNase H2; and Roman equivalents to the prokaryotic proteins, RNase HI and RNase HII.

### 3.1.1 RNase H1

Eukaryotic RNase H1 consists of highly conserved N-and C-termini, connected by an internal variable sequence. The N-terminal Hybrid Binding Domain (HBD) binds RNA duplexes with a 25-fold increased affinity for RNA-DNA hybrids over RNA-RNA equivalents, as demonstrated with human RNase H1 (Nowotny *et al.*, 2008). The HBD domain binds two consecutive 2'-OH groups of the RNA strand in a non-sequence specific manner (Nowotny *et al.*, 2008). Hydrolysis is accomplished by the C-terminal RNase H domain, which requires a substrate containing at least four consecutive ribonucleotides (Ohtani *et al.*, 1999) and contains four highly conserved carboxylic acid residues (Nowotny *et al.*, 2008). The RNA strand of an RNA-DNA duplex sits in one shallow groove of the RNase H domain, attracted via the minimal four 2'-OH groups, and the DNA strand in another (Nowotny *et al.*, 2005), meaning both the HBD and RNase H domains of type 1 RNases H are involved in binding RNA-DNA hybrid substrates. The Connecting Domain, bridging the HBD and RNase H domain, is highly variable in both length and sequence, and so has been suggested to allow flexible movement of the two domains around substrates (Cerritelli and Crouch, 2009). The RNase H domain of eukaryotic type 1 RNase H is highly similar to that of the *E. coli* RNase H domain, yet few bacterial RNases HI additionally have HBDs, and in these cases, the RNase H domain and HBD are connected by relatively short a Connecting Domain (Kochiwa, Tomita and Kanai, 2007). Evolution of the

HBD appears to have improved activity of eukaryotic RNase H1, as deletion of the domain in human and mouse RNase H1 results in defective RNase H activity; full-length mammalian RNase H1 proteins resolve long RNA-DNA hybrids processively, meaning long stretches of the hybrid are catalysed per interaction event, whereas enzymes lacking the HBD cleave the hybrid substrate in a disrupted manner, much like *E. coli* RNase H1 (Gaidamakov *et al.*, 2005; Nowotny *et al.*, 2005).

Mammalian RNase H1 enzymes contain two in-frame methionine start codons, causing the translation of RNase H1 protein with and without a mitochondrial targeting sequence (MTS) (Cerritelli and Crouch, 2009). Knockout mutation of mouse *Rnaseh1* is embryonically lethal due to failure to replicate mtDNA (Cerritelli *et al.*, 2003). However, the model organisms *S. cerevisiae* and *C. elegans* have no MTS associated with their respective HBD-containing RNase H1 proteins. *S. cerevisiae* mtDNA is vastly different to that of mammals, being significantly larger and linear in conformation, and so its replication is likely to occur via a different mechanism, independent of essential RNase H1 activity (Cerritelli and Crouch, 2009). *C. elegans* mtDNA is similar to that of mammals, but *C. elegans* expresses a second type 1 RNase H-like protein (RNase H1.1) which contains an MTS domain and two alternative start codons (Arudchandran *et al.*, 2002). Though unessential, deletion of RNase H1 in *S. cerevisiae* leads to detectable levels of genome instability and modest sensitivity to DNA damaging agents (Wahba *et al.*, 2011; O'Connell, Jinks-Robertson and Petes, 2015). No human disease has been linked to mutation of RNase H1, although this may be due to the protein's essential role in mtDNA amplification during embryo development (Cerritelli *et al.*, 2003). Taken together, these findings indicate RNase H1 is essential for typical eukaryotic mtDNA replication and has a non-essential role in nuclear genome maintenance.

### 3.1.2 RNase H2

Type 2 RNase H differs from RNase H1 in several aspects. Most obviously, eukaryotic RNase H2 functions as a complex of three subunits (Jeong *et al.*, 2004), where RNase H2A is the catalytic component and RNase H2B and RNase H2C are believed to be largely structural (Chon *et al.*, 2009). However, RNaseH2B does

contain a PCNA (proliferating cell nuclear antigen)-interacting peptide, which directs an interaction between RNase H2 and PCNA (Chon *et al.*, 2009), and accordingly, the complex's localisation to replication forks (Bubeck *et al.*, 2011), pointing to a role of type 2 RNase H in replication-associated repair processes.

Prokaryotic homologues of the catalytic RNaseH2A subunit, termed RNase HII, exist and function as a monomer (Itaya, 1990). In addition, and in contrast to RNase HI, which requires a stretch of at least four hybridised ribonucleotides, bacterial RNase HII enzymes are specialized to detect (5')-RNA-DNA-(3')/DNA (RpD) junctions (Tadokoro and Kanaya, 2009). This specificity allows RNase HII to remove single ribonucleotides embedded in a DNA duplex, which are often aberrantly incorporated by replicative DNA polymerases (Schroeder *et al.*, 2015). Despite the structural elaboration eukaryotic RNase H2 has undergone during evolution, the enzyme complex is still able to detect RpD junctions, as well as RNA-DNA hybrids formed either as tracks of embedded ribonucleotides or R-loops. Hence in eukaryotes, both type 1 and type 2 RNase H are able to resolve R-loops, yet the latter additionally detects single ribonucleotides embedded in the genome (Cerritelli and Crouch, 2009). RNase H2 detects the reactive 2'-OH group of embedded ribonucleotides and proceeds to cleave 5' of the ribonucleotide, initiating the RER pathway (Sparks *et al.*, 2012; Williams, Gehle and Kunkel, 2017). DNA polymerase δ subsequently performs PCNA-dependent nick translation, causing strand displacement of the ribonucleotide that is then removed from the DNA by the action of flap endonuclease FEN1, before DNA ligase repairs the lesion (Sparks *et al.*, 2012).

In further contrast to type 1 RNase H, human RNase H2 has a critical role in the nuclear genome, since mutation in any of the three subunits causes neurological defects and other symptoms associated with Aicardi-Goutières Syndrome (AGS) (Crow *et al.*, 2006). Additionally, full knockout of either mouse RNase H2B or RNase H2C is embryonically lethal (Hiller *et al.*, 2012; Reijns *et al.*, 2012). Fibroblasts harvested from RNase H2B-null embryos show impaired proliferation and accumulation in the G2/M phase (Hiller *et al.*, 2012). However, like yeast RNase H1, *S. cervesiae* RNase H201 (the yeast RNase H2 catalytic subunit) is dispensable

for growth, though its loss causes low levels of genome instability (O'Connell, Jinks-Robertson and Petes, 2015).

The critical role of RNase H2 in the mammalian genome may be attributed to the secondary function of the complex in RER, as embedded ribonucleotides accumulate, along with DNA damage, in the genome of RNase H2-null mouse cells (Reijns *et al.*, 2012). Various studies have been undertaken to try and elucidate the RNase H2 function (RNA-DNA hybrid resolution or RER) whose loss causes extensive DNA instability, with findings leading to contrasting views (Conover *et al.*, 2015; O'Connell, Jinks-Robertson and Petes, 2015). Investigations have been aided by the use of a yeast mutant line termed *rnh201-RED* (Ribonucleotide Excision Defective), which expresses an RNase H2 complex capable of resolving RNA-DNA hybrids but not of excising singularly embedded ribonucleotides (Chon *et al.*, 2013). *rnh201-RED* yeast display mildly elevated levels of genome instability compared to WT, although this does not match levels seen in the *rnh201-null* yeast (Epshtein, Potenski and Klein, 2016; Zimmer and Koshland, 2016; Cornelio *et al.*, 2017). These studies have led to the general conclusion that both functions contribute to instability observed in *rnh201-null S. cerevisiae*, perhaps to different extents in different regions of the genome (Cornelio *et al.*, 2017). However, it is worth noting RNase H2 is not essential in yeast but is essential in mammalian embryo development, and so these studies do not fully clarify the critical function of mammalian RNase H2.

Despite the complexity of the above findings, the available data from yeast and mammals highlight the overlapping yet distinct functions of eukaryotic RNase H1 and H2, as well as the possibility that each are specialized for different RNA-DNA hybrids in the nuclear genome. Although each has a unique function (mtDNA amplification in the case of RNase H1, and RER initiation by RNase H2), the conservation of the catalytic R-loop resolving function in both enzymes perhaps indicates each has specific RNA-DNA hybrid substrates it is most effectively able to target in the nuclear genome. Indeed, the presence of the PCNA-interacting domain in RNase H2B led Chon et al (2009) to suggest a role for RNase H2 in resolving R-loops that cause DNA damage associated with DNA replication.

### 3.1.3 RNase H enzymes in kinetoplastids

RNase H proteins of either type are relatively unstudied in kinetoplastids, with few investigations described to date. Hesslein and Campbell (1997) successfully cloned the RNase H1 gene from *T. brucei*, *Tbrh1*, and discovered the conservation of three predicted catalytic carboxylic acids and other active site residues present in the RNase H domain (Hesslein and Campbell, 1997). The authors were also able to express and purify recombinant TbRH1 fused to a histidine tag and demonstrate the catalytic breakdown of RNA-DNA hybrids *in vitro* by the fusion protein (Hesslein and Campbell, 1997). The N-terminal 47 amino acids of TbRH1, fused to GST, have also been expressed and assessed for RNA binding capabilities via RNA dot plot assays, demonstrating this region is crucial for WT levels of RNA binding, yet dispensable for enzyme activity (Kobil and Campbell, 2000). The connecting sequence that joins the N- and C-termini was, however, found to be essential for TbRH1 function (Kobil, 2000; Kobil and Campbell, 2000). The N-terminal RNA binding motif was also found to direct nuclear localisation of TbRH1 by fusion of the 47 amino acid sequence to GFP and immunofluorescence imaging (Kobil and Campbell, 2000).

In 2005, Misra et al identified four putative RNase H-encoding genes in *Leishmania*, one highly homologous to *Tbrh1* identified by Campbell and colleagues, and three others assumed to be the subunits of RNase H2 (Misra *et al.*, 2005). LmjF36.0330, assigned as LmRNaseH2C, was found to have a MTS within the N-terminal 40 amino acid residues and was enriched in mitochondrial fractions of the cell (Misra *et al.*, 2005). The authors additionally found that expression of LmjF36.0330 is able to restore growth of *E. coli* lacking RNase HI and HII activity, which is only able to form colonies at 30°C, at 42°C (Misra *et al.*, 2005). Given these findings, the identification of this protein as the RNase H2C subunit is doubtful, since in all characterised eukaryotes this subunit is not catalytic and lacks a MTS. No evidence is available regarding whether or not the putative RNase H2C protein forms a complex or co-localises with the putative LmRNaseH2A and LmRNase2B proteins identified. The authors also noted homologs to three of the four RNase H proteins in *T. brucei* and *T. cruzi*, documenting the apparent absence of RNase H2A subunits in the two *Trypanosoma* species (Misra *et al.*, 2005).

### 3.1.4 Chapter aims

As the RNase H enzymes of *T. brucei* have been relatively unstudied and the most recent characterisation of each protein (Misra *et al.*, 2005) presents many inconsistences with what is known about RNases H in other systems, the first undertaking of this chapter was to identify and characterise the complement of RNase H proteins encoded in the parasite. Each putative catalytic protein was then epitope tagged and localised via indirect immunofluorescence. Mutant cell lines were also generated to facilitate initial characterisation of the *T. brucei* repertoire of RNase H enzymes.

## 3.2 BLASTp identification of RNase H proteins

To confirm the previously identified RNase H proteins of *T. brucei*, as well as to investigate the apparent absence of the RNase H2A subunit (Misra *et al.*, 2005), Basic Local Alignment Search Tool (BLAST) searches were carried out using the amino acid sequences of RNase H proteins well characterised in other species (section 2.8.3). This was executed using TriTryDB's BLASTp function, using sequences of all known RNase H proteins from *Homo sapiens*, *M. musculus*, *S. cerevisiae*, *E. coli* and *L. major* (Table 3.1).

| Query protein | Query species | Query ID | Hit ID | E-value | Identity Score |
|---|---|---|---|---|---|
| **RNase HI** | *E. coli* | YP_006780767.1 | Tb427.07.4930 | 4.00E-10 | 55.5 |
| **RNase H1** | *L. major* | LmjF.06.0290 | Tb427.07.4930 | 1.00E-21 | 90.9 |
| | *S. cerevisiae* | NP_013961.1 | Tb427.07.4930 | 7.00E-09 | 55.1 |
| | *M. musculus* | NP_035405.2 | Tb427.07.4930 | 4.00E-31 | 118 |
| | *H. sapiens* | NP_002927.2 | Tb427.07.4930 | 2.00E-31 | 119 |
| **RNase HII** | *E. coli* | YP_006780795.1 | Tb427.10.4730<br>Tb427.10.5070 | 1.00E-16<br>6.00E-05 | 76.6<br>41.6 |
| **RNase H2A** | *L. major* | LmjF.13.0050 | Tb427.10.4730 | 1.00E-08 | 53.9 |
| | *S. cerevisiae* | NP_014327.1 | Tb427.10.5070<br>Tb427.10.4730 | 2.00E-40<br>9.70E-02 | 145<br>32.7 |
| | *M. musculus* | NP_081463.1 | Tb427.10.5070<br>Tb427.10.4730 | 5.00E-42<br>1.00E-05 | 148<br>45.1 |
| | *H. sapiens* | NP_006388.2 | Tb427.10.5070<br>Tb427.10.4730 | 8.00E-40<br>1.00E-06 | 142<br>47.8 |
| **RNase H2B** | *L. major* | LmjF.36.0640 | Tb427.10.5070<br>Tb427.10.4730 | 8.00E-113<br>1.10E-02 | 332<br>35.8 |
| | *S. cerevisiae* | NP_010565.3 | Tb427.07.3160 | 2.10E-01 | 32.3 |
| | *M. musculus* | NP_080277.1 | Tb427.01.4220<br>Tb427.08.4240 | 3.00E-08<br>3.00E-01 | 53.1<br>31.6 |
| | *H. sapiens* | NP_078846.2 | Tb427.01.4220<br>Tb427.08.4240 | 2.00E-08<br>2.50E-01 | 53.5<br>31.6 |
| **RNase H2C** | *L. major* | LmjF.36.0330 | Tb427.10.4730<br>Tb427.10.5070 | 0.00E+00<br>5.40E-02 | 526<br>34.3 |
| | *S. cerevisiae* | NP_013255.1 | Tb427.02.5040 | 1.00E+00 | 26.9 |
| | *M. musculus* | NP_080892.1 | Tb427.01.4730 | 2.40E-01 | 29.6 |
| | *H. sapiens* | NP_115569.2 | Tb427.01.4730 | 8.20E-02 | 31.2 |

**Table 3.1 BLASTp identification of putative *T. brucei* RNase H proteins.**
Query protein sequences from *E. coli*, *L. major*, *S. cerevisiae*, *M. musculus* and *H. sapiens* were used with NCBI-BLAST via TriTrypDB v2.48. Expected (E)-value describes probability of achieving hit by chance. Identity score describes the extent to which the query and hit protein sequences align where higher score equates to better agreement.

As previously identified, Tb427.07.4930 encodes a predicted protein that showed good sequence identity with both prokaryotic and eukaryotic type 1 (RNase H1, HI) proteins, and from here will be referred to as TbRH1. The identification of the catalytic RNase H2 subunit proved interesting as two proteins, encoded by Tb427.10.5070 and Tb427.10.4730, showed good homology to *E. coli* RNase HII and the eukaryotic RNase H2A subunit. The exception to this finding was that Tb427.10.5070 did not show homology to the *L. major* RNase H2A subunit proposed by Misra *et al*. (2005) and encoded by LmjF.13.0050. In fact, Tb427.10.5070 showed far greater identity scores with the *H. sapiens*, *M. musculus* and *S. cerevisiae* RNase H2A subunits (142, 148 and 145 respectively) compared to Tb427.10.4730 (47.8, 45.1 and 32.7), which showed greater homology to the *E. coli* RNase HII protein (76.6 vs 41.6). *T. brucei* proteins Tb427.01.4220 and Tb427.01.4730 showed limited homology to mammalian RNase H2B (53.1 and 53.5) and H2C (29.6 and 31.2) subunits, respectively, but both contain RNase H subunit-like domains, as revealed by InterPro protein analysis (Figure 3.1). These proteins were therefore assumed to be the most likely candidates for the non-catalytic RNase H2 subunits but were not investigated further.

| Query ID | Hit Species | Hit Annotation | Accession number | E-value | Identity score |
|---|---|---|---|---|---|
| Tb427.10.5070 | *Trypanosoma congolense* | putative RNH | CCC93661.1 | 0.00E+00 | 539 |
| | *Trypanosoma cruzi* | RNH | XP_806655.1 | 1.49E-175 | 502 |
| | *Leishmania major* | putative RNH | XP_001686636.1 | 2.80E-115 | 348 |
| | *Leishmania mexicana* | putative RNH | XP_003874409.1 | 1.72E-110 | 336 |
| | *Angomonas deanei* | RNH2A | EPY38111.1 | 2.33E-106 | 325 |
| | *Ricinus communis* | predicted RH2A isoform X1 | XP_002515756.1 | 2.69E-48 | 175 |
| | *Callorhinchus milii* | predicted RH2A | XP_007905341.1 | 6.54E-48 | 174 |
| | *Populus euphratica* | predicted RH2A | XP_011032867.1 | 2.14E-46 | 170 |
| | *Chinchilla lanigera* | predicted RH2A | XP_005381086.1 | 2.21E-46 | 170 |
| | *Ricinus communis* | predicted RH2A isoform X2 | XP_015572727.1 | 2.37E-46 | 168 |
| Tb427.10.4730 | *Trypanosoma congolense* | putative RHII | CCC93631.1 | 0.00E+00 | 786 |
| | *Trypanosoma vivax* | putative RHII | CCC51426.1 | 0.00E+00 | 677 |
| | *Trypanosoma cruzi* | RHII | XP_805682.1 | 0.00E+00 | 669 |
| | *Trypanosoma cruzi* | RHII | XP_809747.1 | 0.00E+00 | 668 |
| | *Leptomonas pyrrhocoris* | putative RHII | XP_015658298.1 | 0.00E+00 | 557 |
| | *Angomonas deanei* | RHII | EPY42595.1 | 0.00E+00 | 538 |
| | *Angomonas deanei* | RHII | EPY34030.1 | 0.00E+00 | 535 |
| | *Leishmania major* | putative RHII | XP_001686603.1 | 0.00E+00 | 529 |
| | *Leishmania mexicana* | putative RHII | XP_003874375.1 | 0.00E+00 | 528 |
| | *Psychrobacter* | RHII | WP_057760473.1 | 2.36E-20 | 1.01E+02 |
| | *Candidatus Stoquefichus massiliensis* | RHII | WP_044600096.1 | 4.90E-18 | 9.28E+01 |
| | *Roseburia sp. CAG:100* | RHII | CDF47048.1 | 1.55E-17 | 9.24E+01 |
| | *Clostridium sordellii* | RHII | WP_057536894.1 | 3.65E-17 | 9.13E+01 |

**Table 3.2 BLASTp identification of Tb427.10.5070 and Tb427.10.4730 homologues.**
Protein sequences were used as a query for NCBI-BLASTp search of the non-redundant protein sequence database. Expected (E)-value describes probability of achieving hit by chance. Identity score describes the extent to which the query and hit protein sequences align where higher score equates to better agreement.

To clarify whether Tb427.10.5070 or Tb427.10.4730 may be the true RNase H2A catalytic subunit, each was used as a query for BLASTp searches using NIH's Standard Protein BLAST search of the non-redundant protein sequence database. The results in Table 3.2 show Tb427.10.5070 had many high-confidence hits to predicted eukaryotic RNase H2A subunits, whereas the top hits for Tb427.10.4730 consisted of other trypanosomatids and then weaker hits to predicted RNase HII proteins in bacterial species. These data indicate the latter protein may be kinetoplastid-specific and shares more in common with monomeric prokaryotic RNase HII than the eukaryotic RNase H2A catalytic subunit. As a result, Tb427.10.5070 was considered the most likely candidate for the catalytic A subunit, and from here is termed TbRH2A, and Tb427.10.4730 is termed TbRH3 for the purpose of this thesis.

Protein domain predictions of TbRH1, TbRH3 and the three putative subunits of TbRH2 were next made using the InterPro protein analysis tool and are depicted in Figure 3.1, along with amino acid alignments for TbRH1, TbRH3 and TbRH2A with selected RNase H proteins and subunits. In the case of TbRH3, the protein was aligned with its orthologues from *T. cruzi* and *L. major*, as well as RNase HII from *E. coli* and *Thermotoga maritina*, since eukaryotic RNase H proteins showed limited homology. The InterPro predictions indicated that all three *T. brucei* RNase H proteins have the conserved domains expected for RNase H enzymes. TbRH1 is predicted to contain both HBD and RNase H domains and, notably, the predicted domains of the three putative TbRH2 subunits follow domain organisation of *H. sapiens* RH2A, RH2B and RH2C. However, unlike HsRH1, no MTS domain was detected in TbRH1 or TbRH2A, but instead this sequence was found at the N-terminus of TbRH3. Beyond the domain predictions, sequence alignments revealed that all catalytic and active site residues that have been characterised in RNase H enzymes of other species are present in TbRH1, TbRH2A and TbRH3 (Figure 3.1). These data strongly support the possibility that *T. brucei*, and perhaps all trypanosomatids, encode a third catalytic RNase H protein, TbRH3, which is

predicted to be mitochondrial. It should be noted that the putative RNase H-domain of TbRH3 is split and separated by non-conserved sequence, but whether this insertion renders the protein non-functional is unclear (Figure 3.1). Nonetheless, these data may indicate that the mitochondrion of *T. brucei* has necessitated the evolution of a dedicated organellar RNase H.

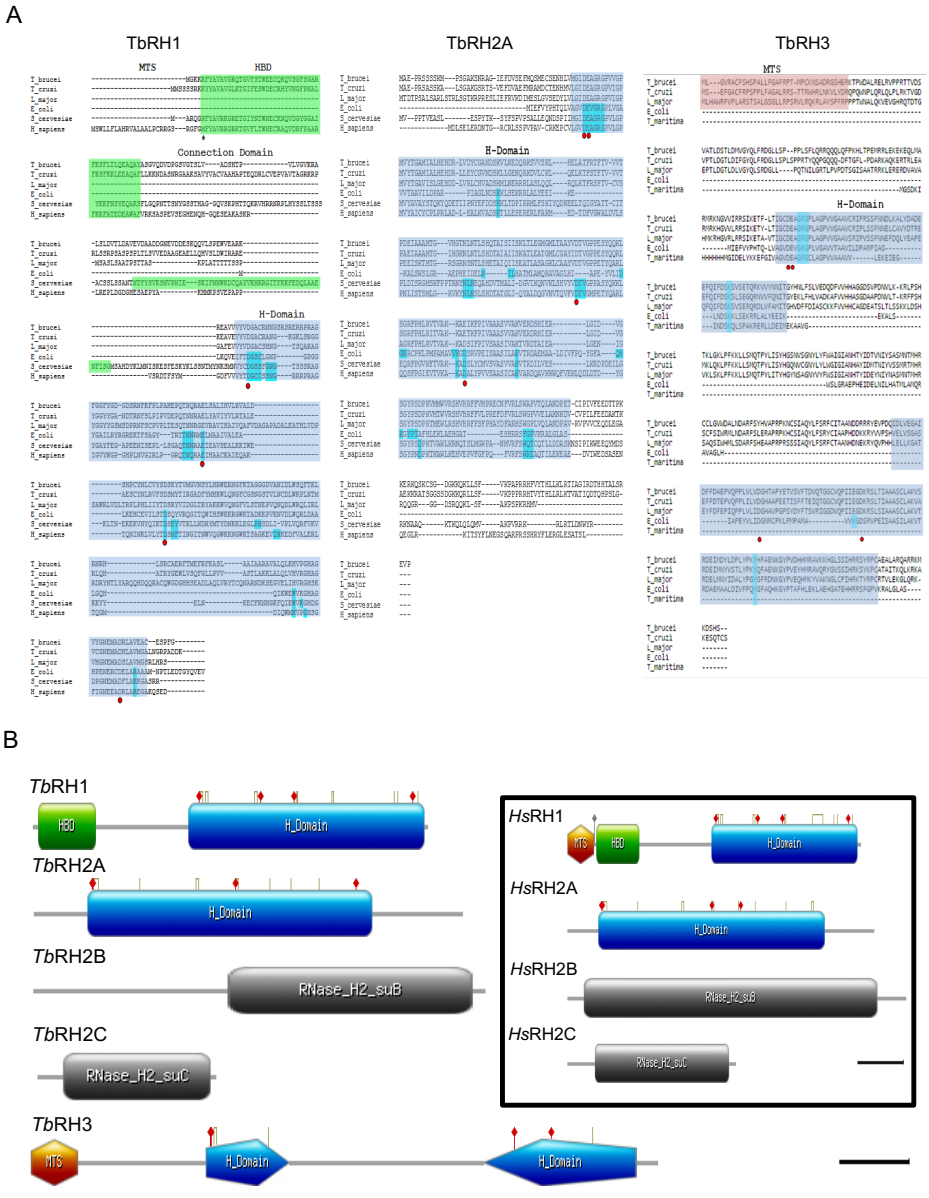



**Figure 3.1 Protein analysis of putative T. brucei RNase H enzymes.**
A) Amino acid sequence alignments of TbRH1, TbRH2A and TbRH3 to appropriate homologs. Hybrid-binding domain (HBD) sequence is highlighted in green, RNase H domain (H-Domain) in blue and mitochondrial targeting sequence (MTS) in red. Conserved active site residues are highlighted in aqua, and catalytic residues are indicated by red dots below alignments. B). Protein domain models of the T. brucei RNase H repertoire as resolved via InterPro analysis (https://www.ebi.ac.uk/interpro/). H. sapiens RNase H proteins are shown for reference. HBD is shown in green, H-domains in blue and MTS in orange. Conserved active site residues are highlighted by white boxes above the models and catalytic residues by red diamonds.

## 3.3 Cellular localisation of *T. brucei* RNase H enzymes

Each putatively catalytic *T. brucei* RNase H protein (TbRH1, TbRH2A and TbRH3) was C-terminally epitope tagged at one endogenous allele in order to detect the proteins' cellular localisation via immunofluorescence. This was achieved with the pNAT$^{x12M}$ vector developed by Alsford and Horn (2008) (Figure 3.2). ~600 bp of each 3' ORF, minus the stop codon, was amplified via PCR for each gene. Restriction sites at the ends of the primers used (Table 2.2) allowed insertion into the tagging plasmids via restriction digest and ligation as detailed in methods sections 2.2.4 and 2.2.6. Sanger sequencing was used to confirm correct generation of constructs (section 2.2.9). The resulting tagging constructs were then linearized using a restriction site within each cloned *T. brucei* ORF sequence (Figure 3.2A), providing two gene fragments as targeting sequences at the ends of the linear plasmids (Figure 3.2B). The linearised plasmids were separately transfected into WT BSF parasites (section 2.1.3), allowing them to undergo single cross-over recombination to integrate at the 3' end of one endogenous allele of the target gene (Figure 3.2B). Resultant drug resistant clones were checked by PCR (using primers indicated in red in Figure 3.2B, Table 2.2, section 2.2.2) and western blot (section 2.4.3), testing for the correct insertion of the plasmid and expression of the tagged RNase H protein, respectively.

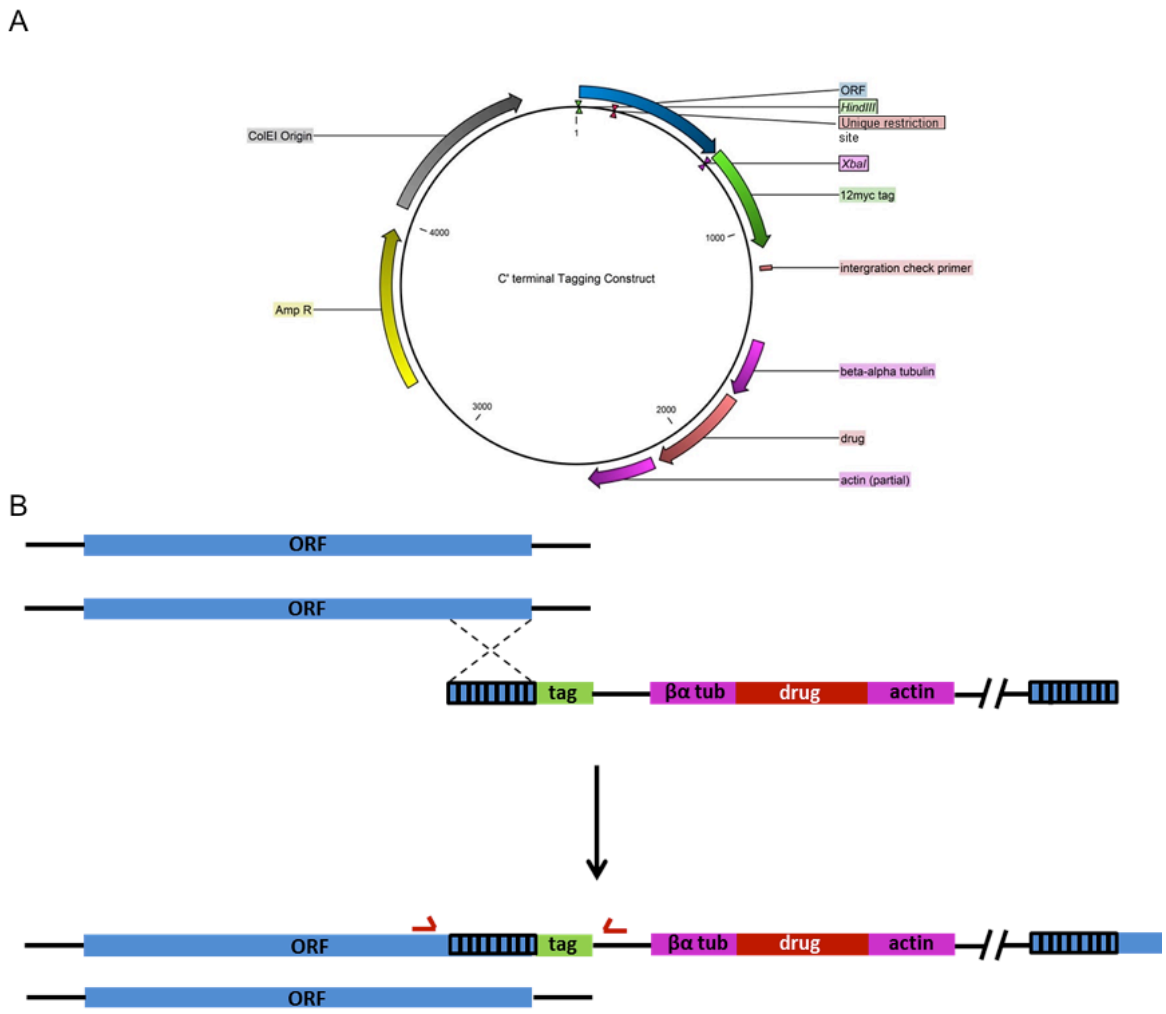



**Figure 3.2 C-terminal epitope tagging strategy.**
A) Plasmid map of pNAT$^{x12M}$ epitope tagging construct (Alsford and Horn, 2008). B) Scheme of single crossover recombination, which results in 3' epitope tagging of the gene of interest. Blue striped areas indicate regions of homology; the full linearised construct is not depicted for simplicity (indicated by dashed gap). Red arrows indicate approximate binding sites of primers used to confirm integration of construct. Diagram not to scale.

## 3.3.1 TbRH1 is a nuclear protein

TbRH1 was endogenously tagged with 12 copies of the Myc epitope, as confirmed by PCR (Figure 3.3B) and western blot (Figure 3.3A) with anti-Myc antibody, verifying expression of a Myc-tagged protein of the expected size (~57 kDa). Growth of three tagged clones, TbRH1$^{12Myc\ cl1-3}$, was also analysed, indicating that addition of the Myc epitope did not result in impaired growth (Figure 3.3C).

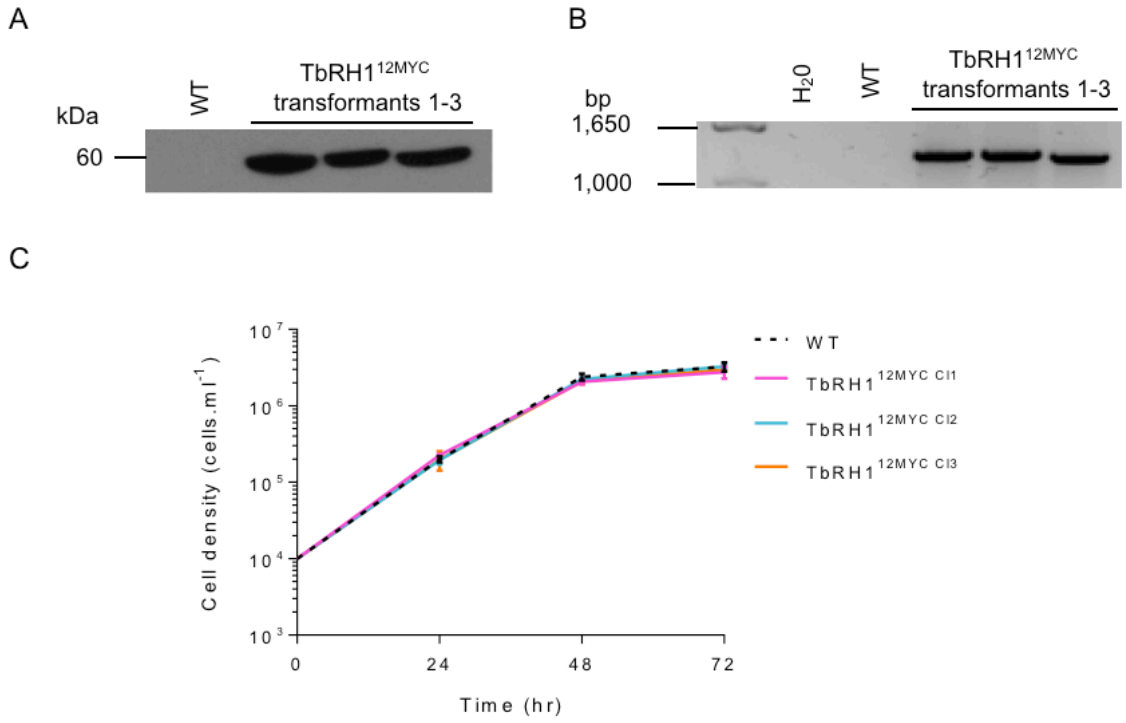



**Figure 3.3 Confirmation of TbRH1 C-terminal epitope tagging.**
A) Western blot of whole cell extracts from three putative TbRH1$^{12Myc}$ expressing transformants and untransformed wild type (WT) cells. B) PCR amplification of *Tbrh1* fragment plus the fused Myc tag sequence from genomic DNA of three putative TbRH1$^{12Myc}$ transformants and WT cells; a control PCR using only water as substrate is also shown. C) Growth curve of three TbRH1$^{12Myc}$ clonal transformants relative to WT. Error bars indicate SD from the mean of three independent experiments.

Indirect immunofluorescence using anti-Myc antibody (section 2.5.2, Table 2.6) revealed signal that co-localised exclusively with DAPI stained nuclear signal, suggesting that TbRH1 is localised to the nucleus of the parasite (Figure 3.4A). Analysis of anti-Myc pixel intensity, as a proxy for TbRH1$^{12Myc}$ protein levels, indicates that TbRH1 is expressed at higher levels in cell cycle stages 1N1eK and 1N2K, which are undergoing nuclear S phase, compared to 1N1K cells (Figure 3.4B). Super-resolution imaging (section 2.5.6) reveals that, although nuclear expression levels change, TbRH1$^{12Myc}$ does not display any discrete localisation within the nuclei at any point during the cell cycle, at least in the BSF life cycle stage (Figure 3.4C).

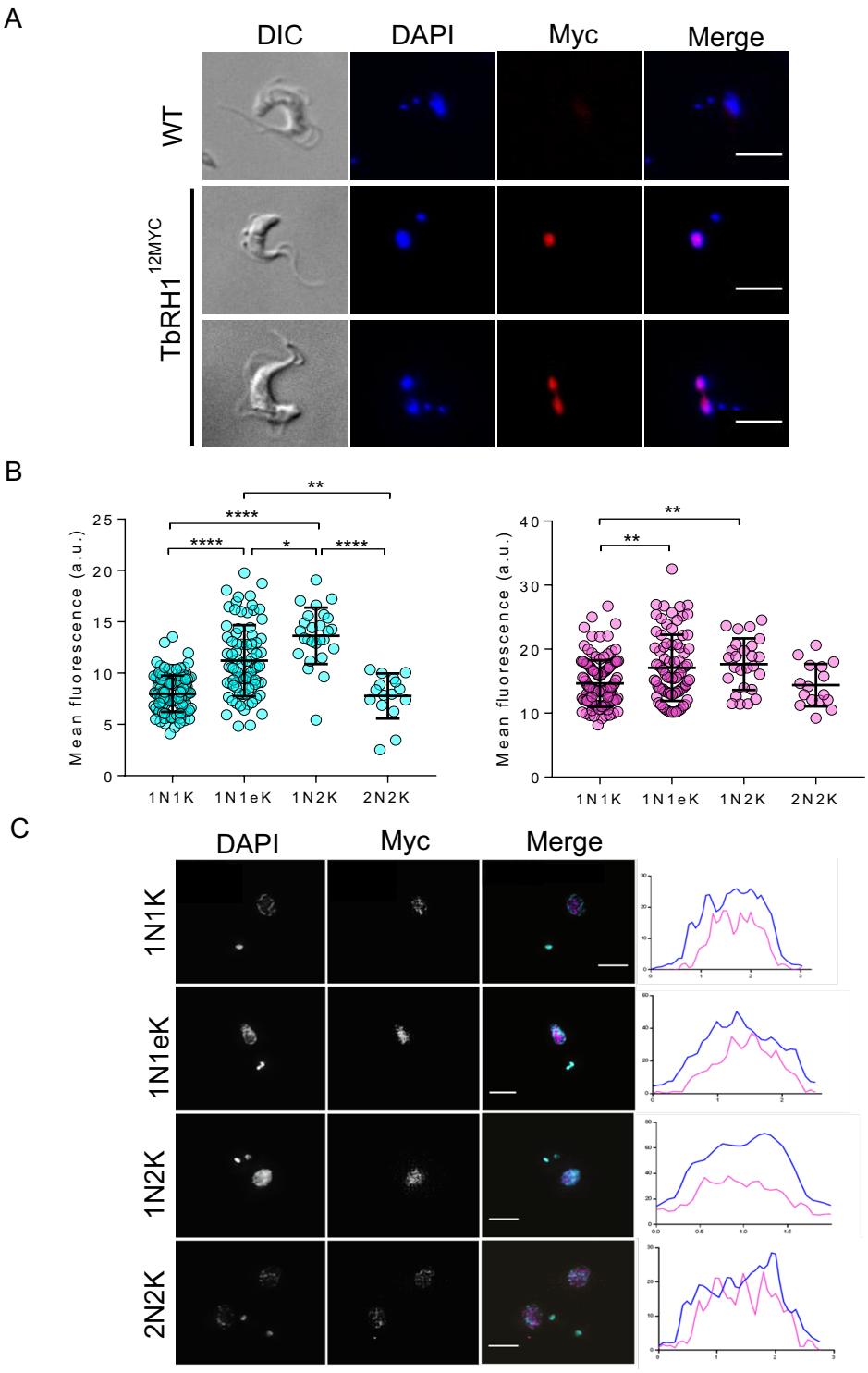



**Figure 3.4 Cellular localisation and expression of TbRH1.**
A) Immunofluorescent imaging of TbRH1$^{12Myc}$ in WT and tagged BSF *T. brucei* using anti-Myc antiserum and DAPI staining. Error bar, 5 μm. B) Mean pixel intensity analysis of immunofluorescent imaging, corresponding to expression of the TbRH1$^{12Myc}$ fusion protein. Mean fluorescence plots for DAPI (cyan) and anti-Myc (magenta) signal in cells displaying different nuclear (N) and kinetoplastid (K) configurations. Significance was determined by Kruskal-Wallis non-parametric test: (*) p-value <0.05; (**) p-value < 0.01; (***) p-value < 0.001; (****) p-value <0.0001. C) Super-resolution imaging of DAPI (cyan) and anti-Myc (magenta) is shown, as well as merged images of the two channels. Graphs shown pixel intensity of each channel (y-axis) across the nucleus (x-axis). Error bar, 5 μm.

### 3.3.2 TbRH2A is a nuclear protein

TbRH2A was tagged using a modified version of same pNAT$^{x12M}$ plasmid, which contains 6 copies of the hemagglutinin (HA) tag sequence rather than 12 copies of Myc. Correct construct integration was confirmed by PCR (Figure 3.5B) using primers No. 7 and No. 4, and expression of the HA-tagged protein (expected size, 43 kDa) was confirmed via western blot using an anti-HA antibody (Figure 3.5A). Growth of three TbRH2A$^{6HA}$ clonal cell lines was no different to that of WT cells (Figure 3.5C).

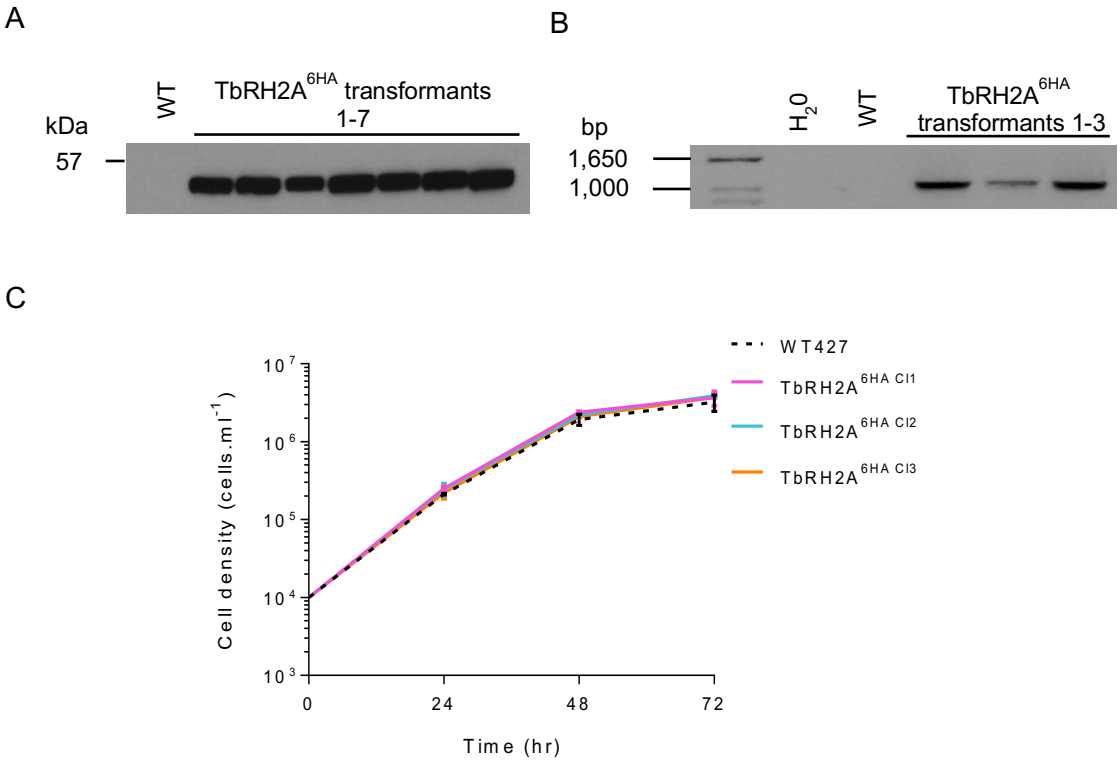



**Figure 3.5 Confirmation of TbRH2A epitope tagging.**
A) Western blot confirmation of TbRH2A$^{6HA}$ expression. B) Amplification of *Tbrh2a* fragment plus HA tag sequence. C) Growth curve of three TbRH2A$^{6HA}$ clonal cell lines. Error bars indicate SD from the mean of three independent experiments.

Fluorescent imaging of the TbRH2A$^{6HA}$ cells using anti-HA antiserum showed that TbRH2A displays a very similar localisation to that of TbRH1; the tagged protein was only detected in the nucleus (Figure 3.6A) and showed no significant sub-nuclear localisation throughout the cell cycle (Figure 3.6C). Also in common with TbRH1, expression of TbRH2A peaks during S phase, as indicated by measuring anti-HA mean fluorescence levels (Figure 3.6B).

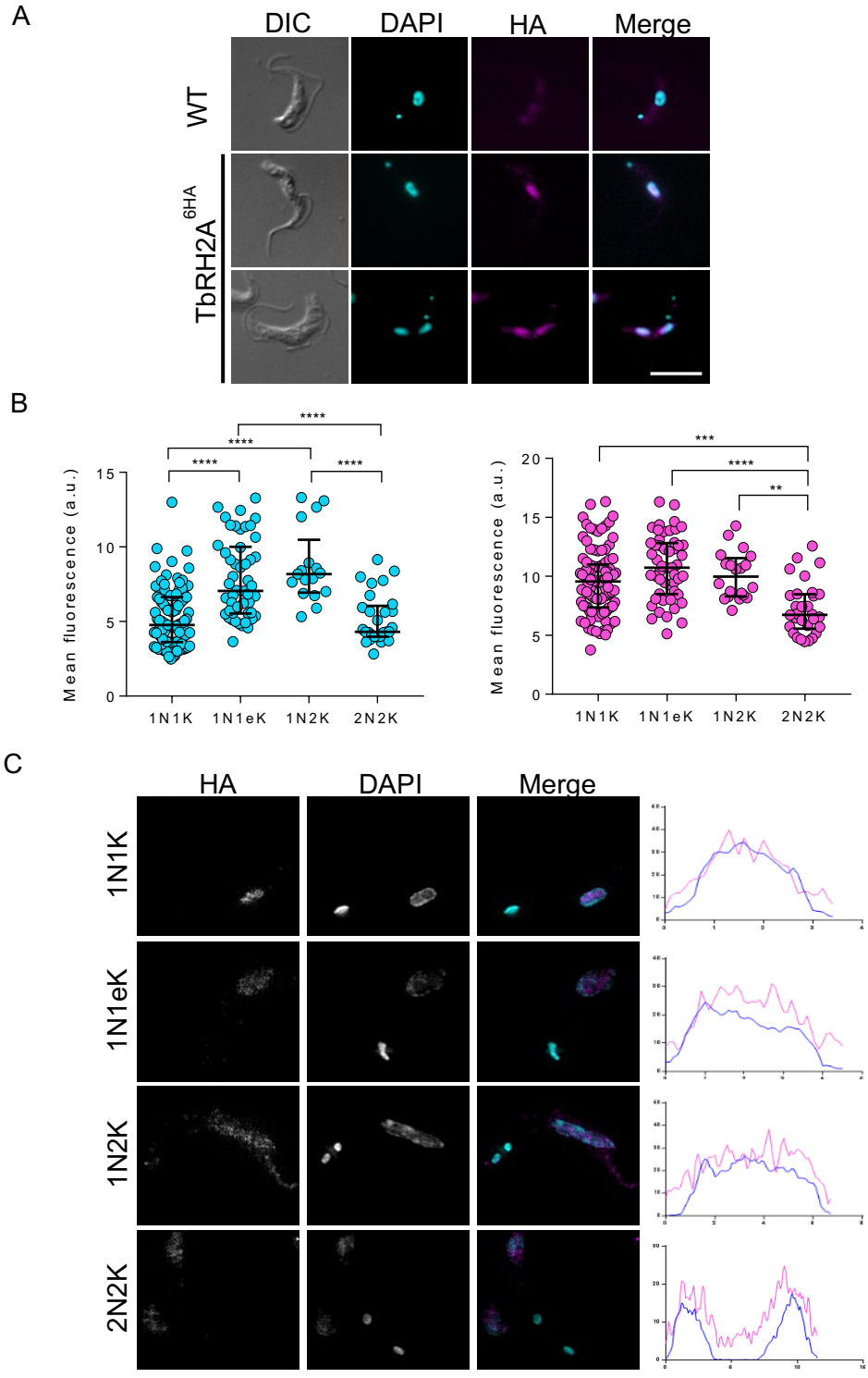



**Figure 3.6 Cellular localisation and expression of TbRH2A$^{6HA}$.**
A) Immunofluorescent imaging of TbRH2A$^{6HA}$ in BSF *T. brucei*. B) Mean pixel intensity analysis of immunofluorescent imaging, corresponding to expression of the TbRH2A$^{6HA}$ fusion protein. Mean fluorescence of each nuclei is shown for DAPI (cyan) and TbRH2A$^{6HA}$ (magenta) signal. Significance was determined by Kruskal-Wallis non-parametric test: (*) p-value <0.05; (**) p-value < 0.01; (***) p-value < 0.001; (****) p-value <0.0001. C) Super-resolution imaging of DAPI (cyan) and TbRH2A (magenta) is shown, as well as merged images of the two channels. Graphs shown pixel intensity of each channel (y-axis) across the nucleus (x-axis).

### 3.3.3 TbRH3 localises to the kDNA

Lastly, TbRH3 was tagged with 12 copies of the Myc epitope. Integration of the plasmid was determined with PCR (Figure 3.7B), and expression of Myc-tagged protein (expected size, 71 kDa) was determined via western blot with an anti-Myc antibody (Figure 3.7A). In contrast with the cells expressing epitope tagged TbRH1 or TbRH2A, growth of three TbRH3$^{12Myc}$ clones appeared somewhat reduced relative to WT cells, perhaps indicating a slight impairment in cell fitness when TbRH3 is C-terminally tagged (Figure 3.7C).

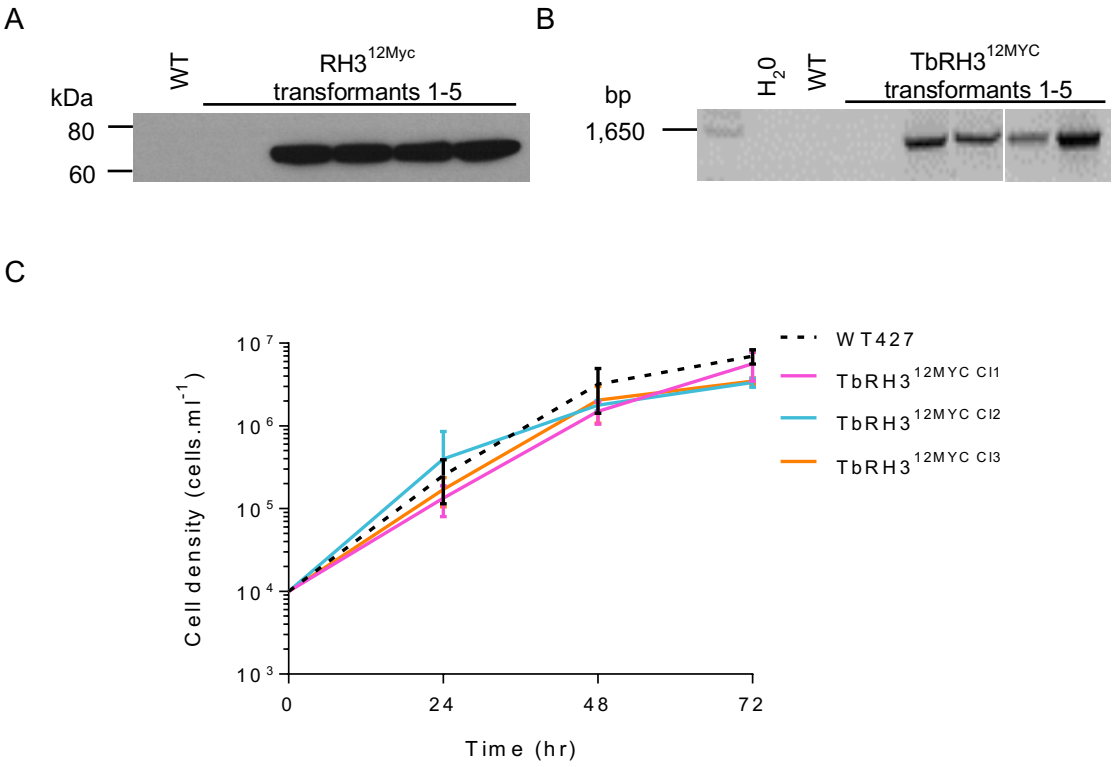



**Figure 3.7 Confirmation of TbRH3 C-terminal epitope tagging.**
A) Western blot confirmation of TbRH3$^{12Myc}$ expression. B) Amplification of *Tbrh3* fragment plus Myc tag sequence. C) Growth curve of three TbRH3$^{12Myc}$ clonal cell lines. Error bars indicate SD from the mean of three independent experiments.

Imaging of TbRH3$^{12Myc}$ cells indicated that TbRH3 shows a greatly different localisation to that of TbRH1 and TbRH2A, as the majority of anti-Myc immunofluorescent signal was detected in co-localisation with kDNA, rather than nuclear DNA (Figure 3.8). Moreover, super-resolution imaging showed the protein localisation changes during the cell cycle; in 1N1K cells anti-Myc signal either co-

localised with the DAPI-stained kDNA, or was seen as a focus beside the kDNA, whereas the signal migrated to form two foci around at opposite poles of the kDNA in 1N2K and 2N2K cells (Figure 3.8B). Some signal could also be observed in the body of the parasite cell (Figure 3.8).

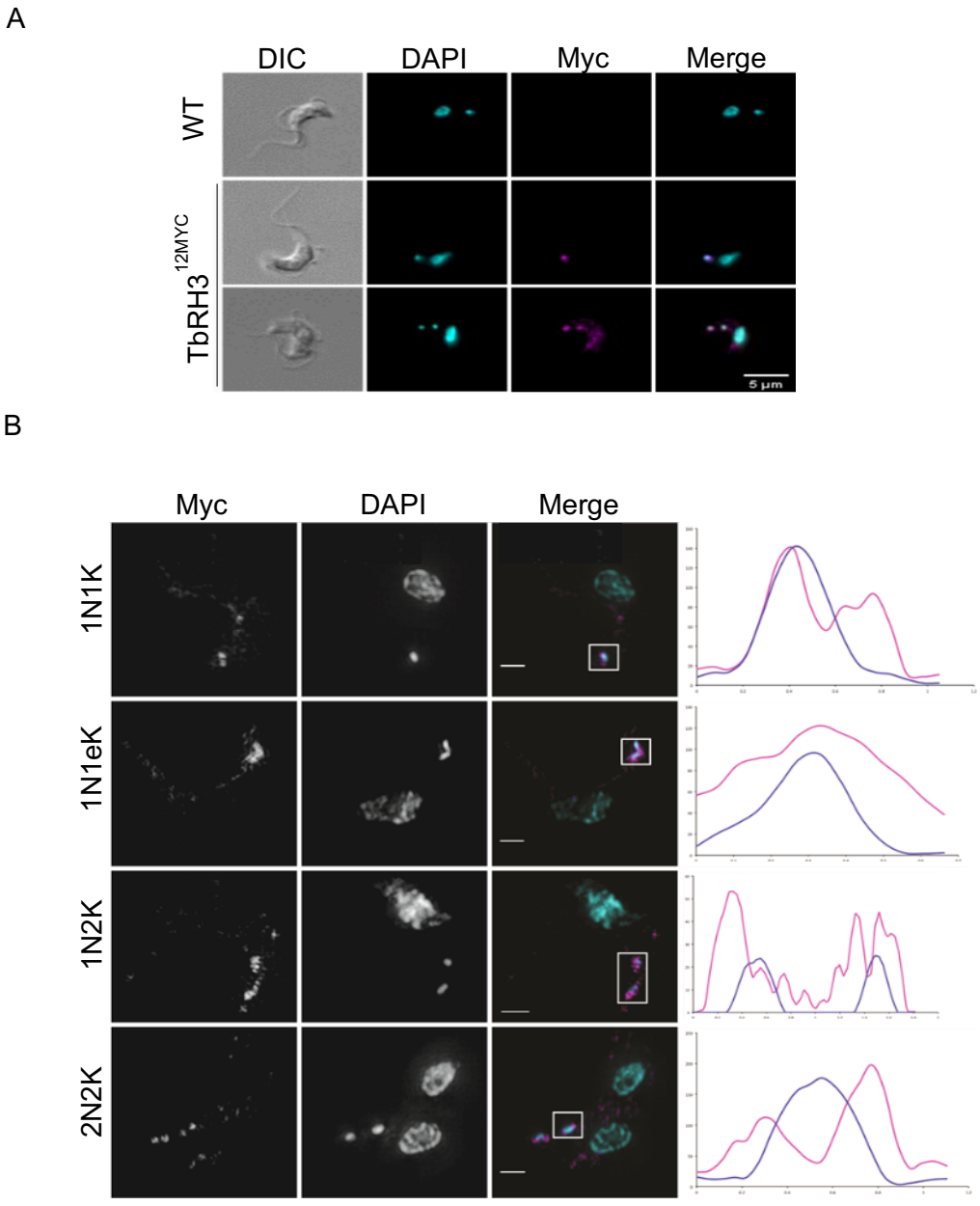



**Figure 3.8 Cellular localisation of TbRH3$^{12MYC}$.**
A) Immunofluorescent imaging of TbRH3$^{12Myc}$ in BSF *T. brucei*. B) Super-resolution imaging of DAPI (cyan) and TbRH2A (magenta) is shown, as well as merged images of the two channels. Graphs shown pixel intensity of each channel (y-axis) across the area indicated by the white box (x-axis).

These localisation experiments confirmed the initial conclusion that Tb427.10.5070 is a much more likely candidate for the TbRH2A subunit than Tb427.10.4730, as the latter protein lacks nuclear localisation.

## 3.4 TbRH1 is non-essential for BSF *T. brucei*

Heterozygous, *Tbrh1*+/-, then homozygous, *Tbrh1*-/-, BSF knockout cells were next generated using the method described by McCulloch and Barry (1999). Two constructs were generated, each containing a drug resistance cassette for neomycin (NEO) or blasticidin (BSD), flanked by *Tbrh1* 5' and 3' UTR sequences (Figure 3.9). 389 bp of the 5' UTR and 435 bp of the 3' UTR were amplified using primers No. 11 and No. 12, No. 13 and No. 14, respectively (Table 2.2). Inclusion of restriction digest sites in these primer sequences, allowed two-step, directional cloning of the UTR sequences into the two plasmids (sections 2.2.4 and 2.2.6). The two resulting constructs were then linearised using BamHI and XhoI restriction digestion, giving linear DNA for each that, when transformed into *T. brucei*, allows homologous recombination on the UTRs to insert the drug resistance cassette into the genome, replacing the *TbRH1* ORF (Figure 3.9A).

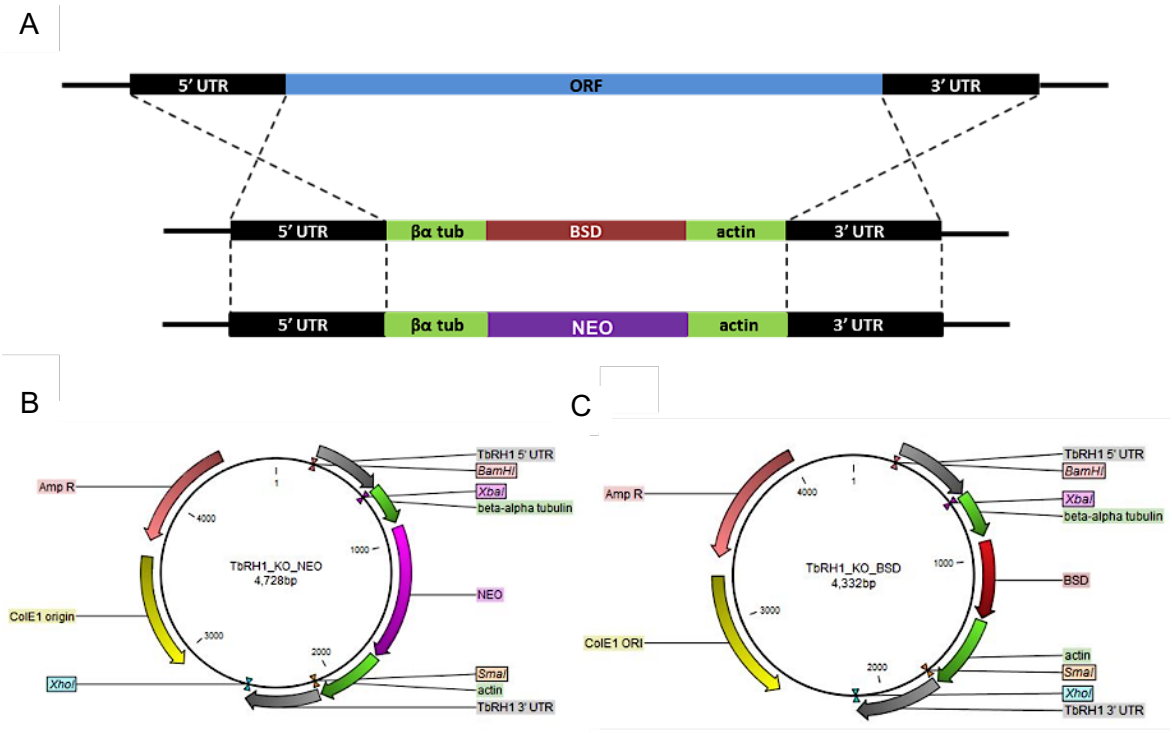



**Figure 3.9 Gene deletion strategy.**
A) Scheme of homologous recombination, which results in deletion of the gene of interest. Dashed lines indicate regions of homology on the UTRs surrounding the ORF. Diagram not to scale. B) Maps of NEO- and (C) BSD-containing knockout constructs for TbRH1 (Devlin *et al.*, 2016).

The NEO-containing linearised *TbRH1* KO construct was first transfected into WT BSF cells, as described in section 2.1.3. Generation of *Tbrh1*+/-$^{NEO}$ clones was tested by PCR, using primers which bound within the NEO resistance gene and within the 3' UTR (Table 2.2), downstream of the flanking sequence used to generate homology regions (Figure 3.10A). The second *TbRH1* KO construct, containing the BSD resistance cassette, was then transfected into one *Tbrh1*+/-$^{NEO}$ cell line and integration was tested in 11 transformants via PCR (Figure 3.10A). Though all 11 clones had integrated the *BSD* construct as expected, loss of the *Tbrh1* ORF, consistent with the generation of a double knockout, was detected in only two of the clones using PCR (Figure 3.10A, B). The nine aberrant, non-null mutants were not examined further, but the two putative successfully generated *Tbrh1*-/- clones were checked for correct integration of the drug resistance cassettes at the 5' end of the *TbRH1* gene using PCR (Figure 3.10B). Finally, loss of detectable *Tbrh1* RNA expression in one of the null mutant clones was determined via RT-qPCR (Figure 3.10C), with the *Tbrh1*+/-$^{NEO}$ cell showing 50% less RNA than WT, consistent with loss of one allele.

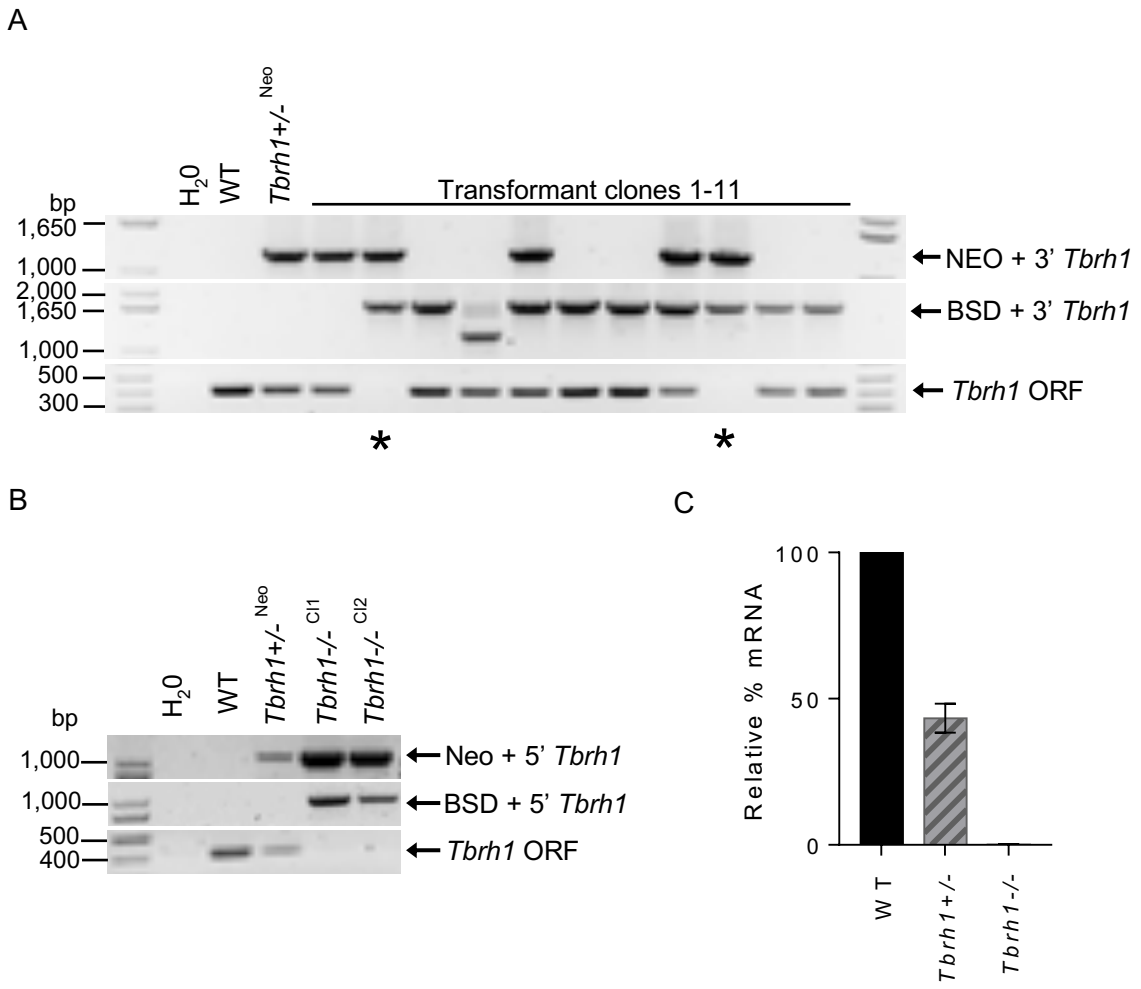



**Figure 3.10 Confirmation of *Tbrh1-/+ and Tbrh1-/-* mutant cell lines.**
A) An agarose gel is shown of PCRs to check 3' integration of NEO (primers No. 15 and No. 16) and BSD (primers No. 16 and No. 17) resistance cassettes, and the presence or absence of the *Tbrh1* ORF (primers No. 21 and No. 21), in WT cells, in the single selected TbRH1+/- cell and in 11 putative *TbRH1* null (*Tbrh1-/-*) clones. B) PCR to check 5' integration of NEO (primers No. 18 and No. 19) and BSD (primers No. 18 and No. 20) resistance cassettes in two *Tbrh1-/-* cells. C) RT-qPCR to determine TbRH1 RNA expression (primers No. 23 and No. 24) after gene deletion in *Tbrh1+/-*$^{Neo}$ and *Tbrh1-/-* $^{cl1}$ relative to WT (set as 100%); error bars show SD from two experiments.

Growth of both the heterozygous and homozygous knockout clone cells was found to be virtually identical to that of the WT progenitor (Figure 3.11A), indicating loss of TbRH1 has no deleterious effect on BSF *T. brucei* in culture. The cell cycle profiles of WT and *Tbrh1-/-* populations, analysed via DAPI staining, also appeared identical, since the proportion of cells in each detectable stage (1N1K, 1N2K and 2N2K, or other) showed no difference (Figure 3.11B). Hence, deletion of *TbRH1* causes no discernible effect on the growth or cell cycle progression in BSF *T. brucei* parasites in culture.

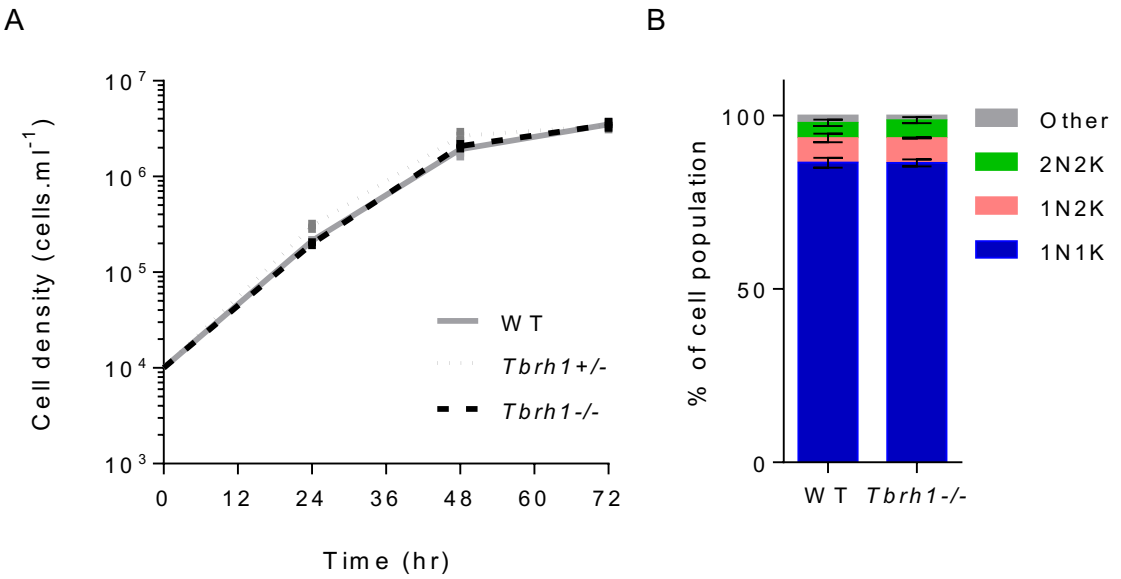



**Figure 3.11 Loss of TbRH1 does not affect growth or cell cycle progression of BSF parasites in culture.**
A) Growth curve of WT. *Tbrh1+/-* and *Tbrh1-/-* cell densities over 72 hr. B) Cell cycle profile of each population presented of the proportion of cells in each cell cycle stage determined by DAPI staining of nuclear (N) and kinetoplast (K) DNA: 1N1K, 1N2K, 2N2K or other (meaning aberrant) cells. Error bars show standard deviation from the mean of three independent repeats counting at least 200 cells.

## 3.5 TbRH2A is an essential protein

Given the success of generating TbRH1 mutants, the same strategy was also attempted to generate knockout mutants lacking TbRH2A in BSF parasites. This was conducted as described in section 3.4, but using primers No. 25 and No. 26, and No. 27 and No. 28 to generate BSD and NEO constructs targeting *TbRH2A* (Table 2.2). Heterozygous cells containing either a NEO- or BSD-cassette integrated at the *TbRH2* locus were successfully generated, as confirmed by PCR (Figure 3.12). However, transfection of the alternate construct in either cell line resulted in no antibiotic resistant clones when the transformant cells were subject to selection with both NEO (2.5 μg/ml) and BSD (10 μg/ml) to select for null mutant cells. When transfections were repeated, but only the antibiotic of the secondarily introduced construct was used for selection, multiple clones could be recovered. However, when the clones were checked by PCR, in all cases these cells had lost the previously integrated construct (presumably via homologous recombination targeting the modified allele) and so still possessed a copy of the *TbRH2A* ORF (Figure 3.12).

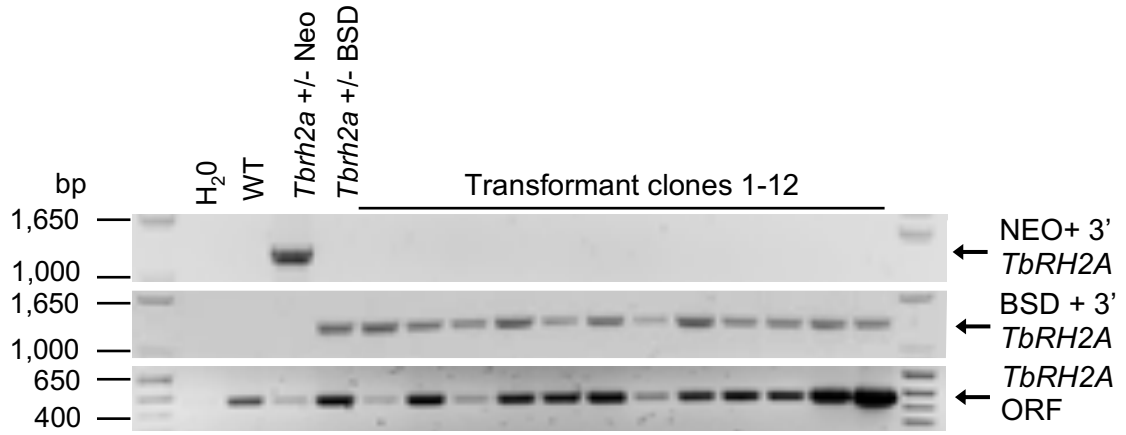



**Figure 3.12 Failure to generate a full TbRH2A knockout cell line.**
PCRs to check 3' integration of Neo- and BDS-containing constructs to delete one allele of *TbRH2A* in heterozygous cell lines *Tbrh2a+/-* $^{Neo}$ and *Tbrh2a+/-*$^{BSD,}$ respectively. Transformant clones 1-12 recovered after transfection of the BSD-containing construct in *Tbrh2a+/-* $^{Neo}$ were checked for integration of each construct (NEO; primers No. 15 and No. 29, BSD; primers No. 17 and No. 29). In all cases the *TbRH2A* ORF was also targeted to check for presence of the gene (primers No. 30 and No. 31).

## 3.5.1 RNAi depletion of TbRH2A

Though not definitive, the above data suggest that TbRH2A, unlike TbRH1, may be essential. To test this and explore the function of TbRH2A further, an RNAi approach was employed to deplete TbRH2A mRNA. The pGL2084 vector produced by Jones et al. (2014), modified from the pRPa$^{ISL}$ vector (Alsford and Horn, 2008), allows integration of two, inverted copies of the target gene sequence into an rRNA spacer locus where they are expressed via a rRNA promoter under the control of a tetracycline (tet) operator, generating long-hairpin RNA. Gateway cloning was used to produce a tet-inducible RNAi construct targeting *TbRH2A* derived from the pGL2084 construct. PCR was used to amplify 432 bp of the *TbRH2A* ORF and add AttB flanking sites (primers No. 32 and No. 33), which facilitated BP cloning of the PCR product as a head-tail duplication into the pGL2084 vector, which contains AttP sites (Figure 3.13A and C). The resulting construct (Figure 3.13B) was then linearised by digestion with AscI and transfected into the BSF 2T1 *T. brucei* cell line generated by Alsford and Horn (2008). An rRNA spacer locus of the 2T1 parasites houses an incomplete hygromycin (HYG) resistance gene, which is completed by successful integration of an RNAi construct that contains the remaining HYG-resistance gene sequence. Hence, the targeted integration of the *Tbrh2a* RNAi

construct into the rRNA spacer was selected for with 5 μg/ml HYG after transfection. RNAi could then be induced by the addition of tet to the resulting clones to trigger RNAi-mediated ablation of *TbRH2A* transcripts.

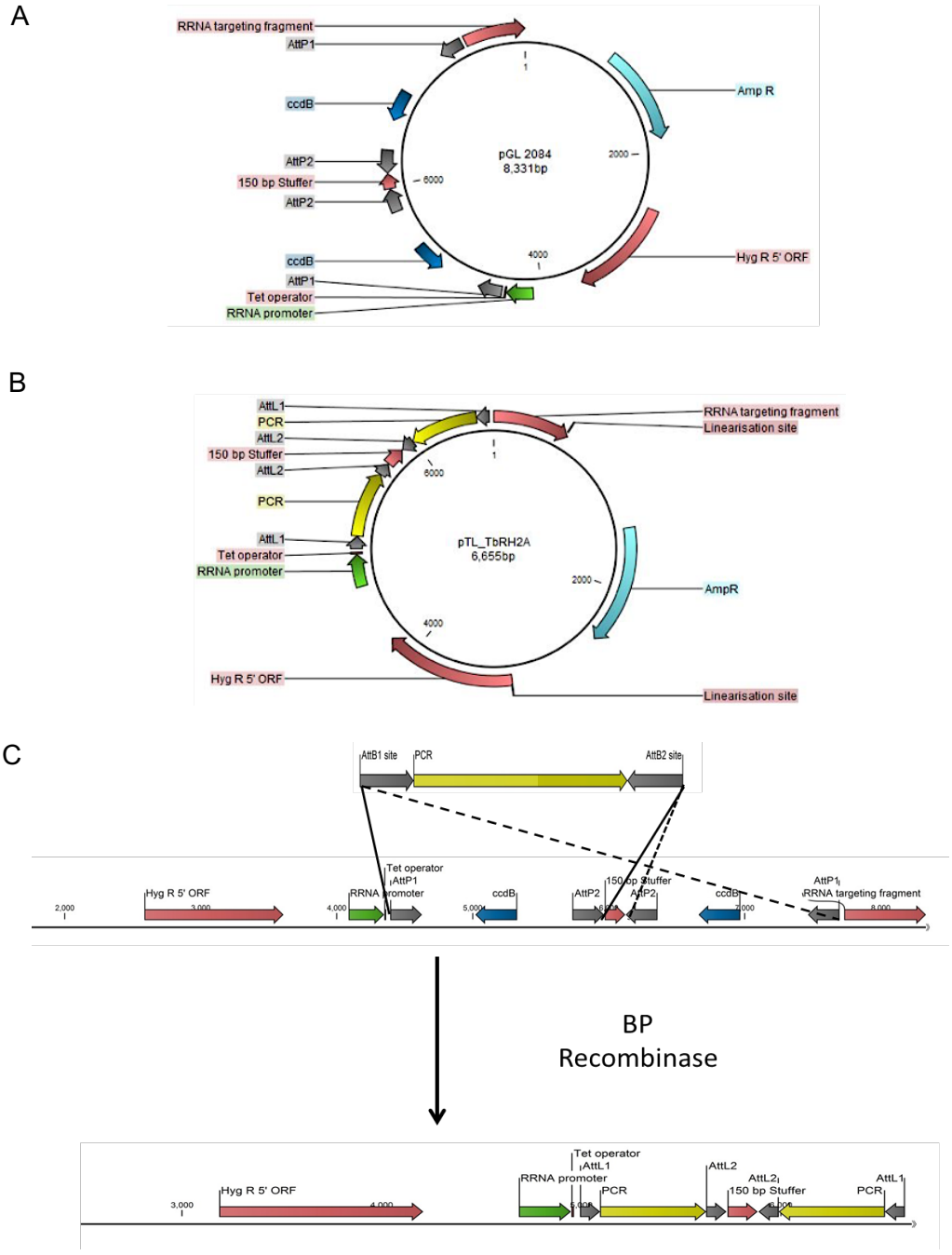



**Figure 3.13 RNAi cloning strategy.**
A) Plasmid map of the pGL2084 construct, which acts as the backbone for integration of RNAi targeting fragments. B) Plasmid map of the Gateway-generated RNAi construct containing two head-tail *TbRH2A* gene fragments. C) Diagram of the BP recombination event that allows cloning of the PCR amplified gene fragment into plasmid pGL2084 via Att recombination sites in the sense and antisense direction.

To evaluate the specificity of TbRH2A targeting in the RNAi clones, levels of *TbRH2A* RNA were measured by RT-qPCR 24 hrs after tet induction and found to be reduced to ~3% of that seen in WT 2T1 *T. brucei* BSF cells (Figure 3.14A). In fact, un-induced TbRH2A$^{RNAi}$ cells were also found to have slightly lower levels (~75 %) of *TbRH2A* RNA compared with the parental 2T1 cell line, indicating the RNAi system is "leaky", causing some knockdown of *TbRH2A* expression in the absence of tet (Figure 3.14A). Growth of the TbRH2A$^{RNAi}$ cells was then analysed with and without the addition of tet (Figure 3.14B). After 24 hr of growth, a small decrease in density of tet-induced cultures was seen relative to cultures free from tet and, from that point on, no further increase in density of the RNAi induced cells was seen, whereas un-induced cells continued to grow unperturbed. These data indicate that TbRH2A RNAi causes a cell growth stall from 24 hrs, a phenotype that could be explained by increased cell death after RNAi or a halt in cell replication.

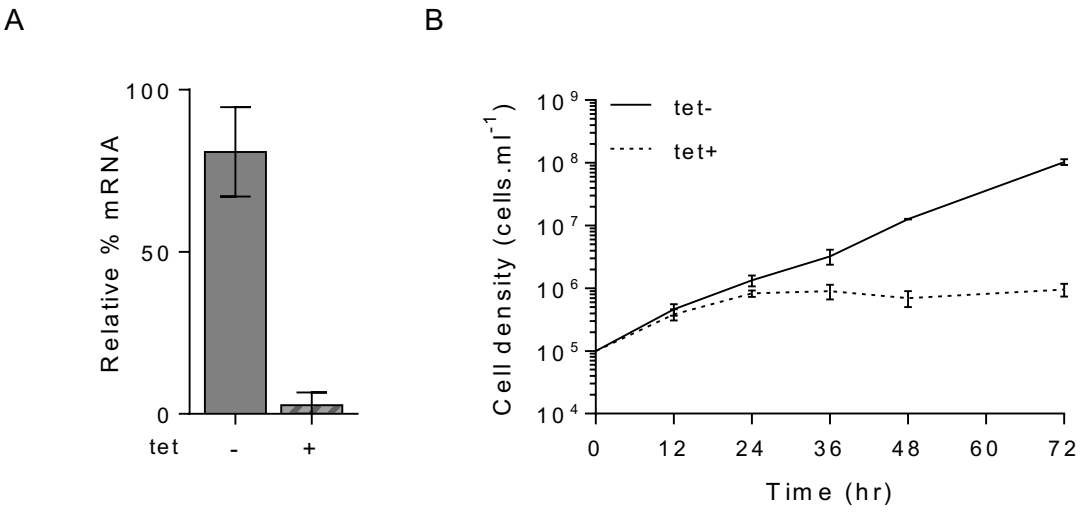



**Figure 3.14 Depletion of TbRH2A leads to stalled growth.**
A) Levels of *TbRH2A* transcripts in tetracycline induced (+) and un-induced (-) cells after 24 hr of culture, relative to 2T1 cells (levels set at 100%), determined by RT-qPCR (primers No. 34 and No. 35). Error bars show standard deviation of two independent repeats. B) Cumulative growth curved of tet + and tet- TbRH2A$^{RNAi}$ cultures, showing cell densities over 72 hr. Error bars show standard deviation from the mean of three independent repeats.

To understand the growth stall after RNAi, the cell cycle profile with and without tet induction was analysed via DAPI staining and counting the N and K configuration of individual cells (Figure 3.15). 24 hr post induction, at the time of the stall, there was a large increase in the amount of 1N2K cells (~40.5 % in tet +, ~9.8 % in tet-),

concomitant with a reduction in 1N1K (~42 % in tet +, ~76.2 % in tet-) and 2N2K (~2.8 % in tet +, ~9.2 % in tet-) cells, indicating a stall at the G2/M phase check point (Figure 3.15A). After 30 hr of RNAi induction, the proportion of 1N1K cells reduced further, an effect that was not due to further increases in 1N2K cells, but mainly due to the appearance of cells (~20%) with 1 nucleus and >2 kDNA foci (1NXK), indicating that kDNA replication and division is unperturbed (Figure 3.15A and B). Coincident with the appearance of these cells, accumulation of cells with abnormal nuclei was seen, including cells with increased nuclear staining or with less clearly defined nuclei, and with >2 kDNA foci (YNXK) (Figure 3.15A and B). The relative percentages of all these cell cycle forms did not clearly change from 30 – 72 hrs, consistent with cell growth having stalled and not with increasing accumulation of aberrant cells.

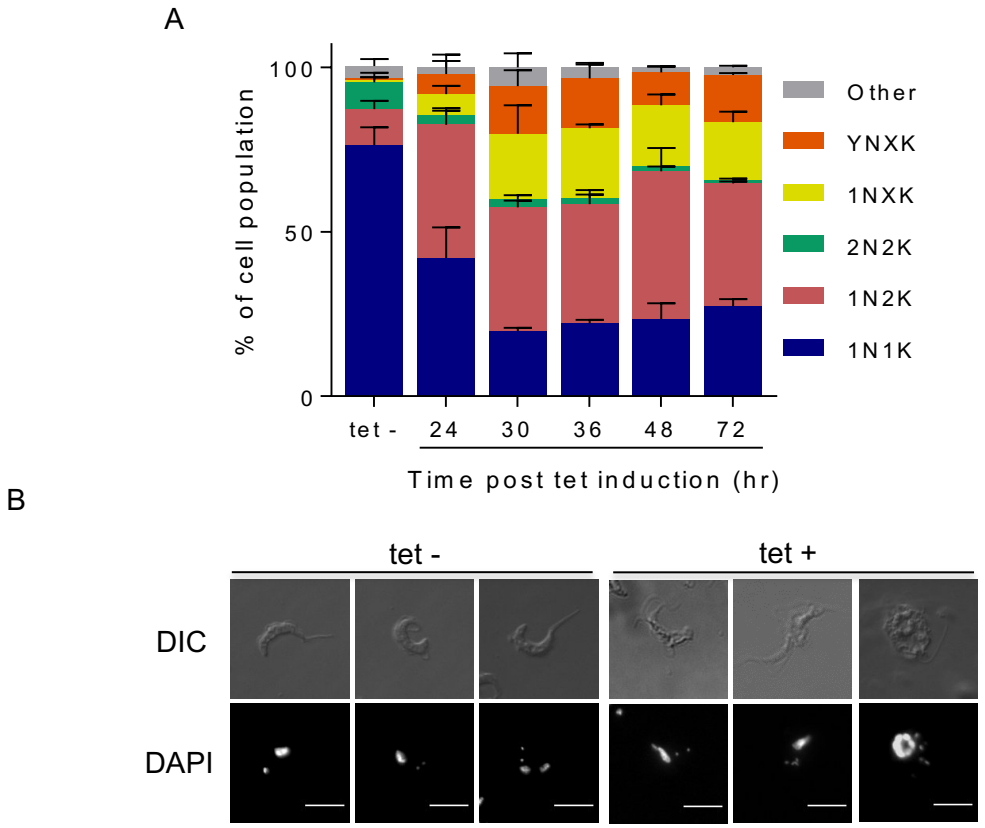



**Figure 3.15 TbRH2A depletion causes a stall in the cell cycle and appearance of abberant cells.** A) Bar graph should the percentage of tet-induced cells at multiple time points in the population that correspond to the following cell cycle stage, as defined by DAPI staning of the nucleus (N) and kinetoplast (K); 1N1K, 1N2K, 2N2K 1NxK (>2 K foci), YNXK (>2 K foci and abberant N number or morphology), and other. tet– shows the average of uninduced samples from all time-points. Error bars show strandard deveation from the mean of three independent replicates. B) Example images of (tet +) induced and un-induced (tet-) cells after 30 hr of growth. Scale bar = 5 μm.

The cell cycle distribution and nuclear content was further investigated using flow cytometry of propidium iodide stained cells (section 2.7). As soon as 12 hr post tet-induction there was a reduction in the proportion of 2N (G1) cells, which appeared to be accompanied by a modest increase in 4N (G2/M) cells (Figure 3.16). However, from 12-36 hr of induction the most notable change was an increase (to 50% at 36 hrs) in cells with DNA content above that of a 2N2K cell (>4N) (Figure 3.16), and with loss of 2N-4N cells. Hence, though DAPI staining suggested RNAi knockdown of *TbRH2A* causes a stall in growth at the G2/M stage, this is partial and continued accumulation of nuclear DNA content is seen from 24 hrs onwards.

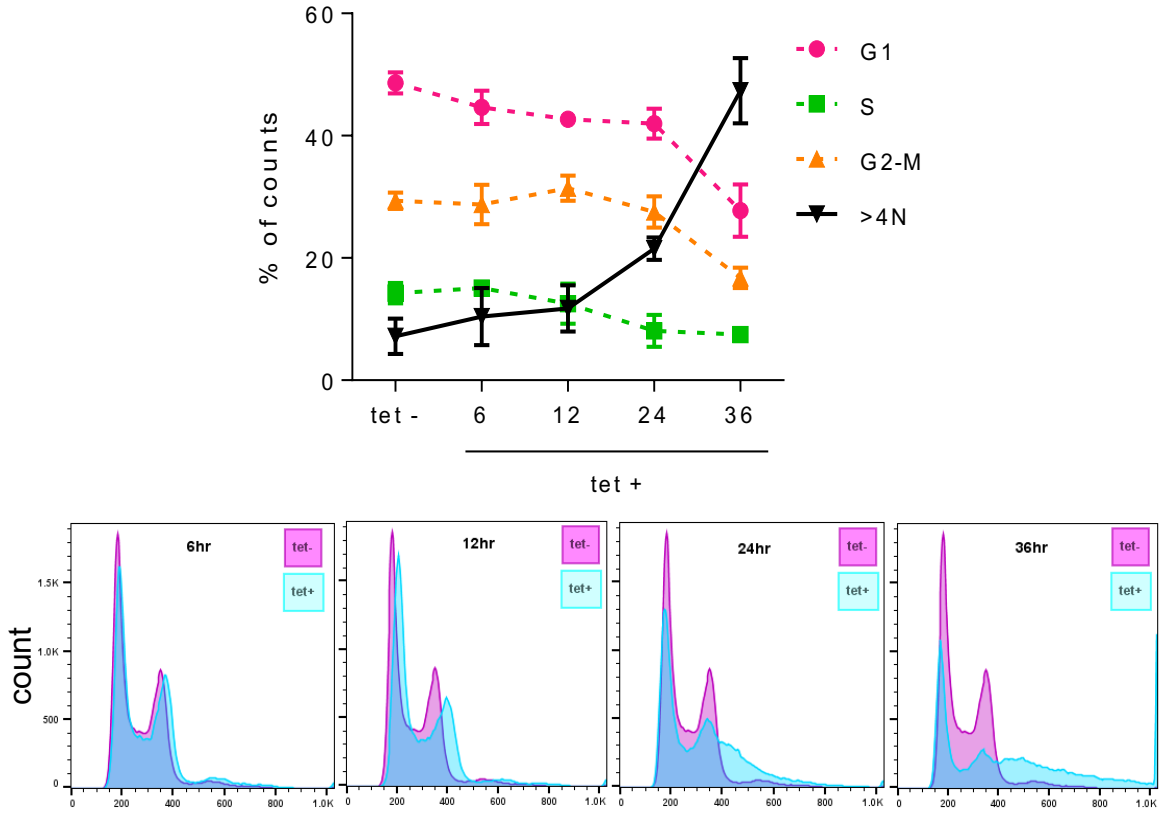



**Figure 3.16 Flow cytometry analysis of TbRH2A depleted cells.**
Profiles of propidium iodide (PI) stained uninduced (tet –, pink) and RNAi induced (tet +, blue) populations after 6, 12, 24 and 36 hr growth are shown. y-axes show cells count and x-axes PI-area fluorescence. The graph above shows the percentage of cells in each cell cycle stage, G1, S and G2-M or cells with genome content >4N, based on measuring proportion of the population with 2N, 2N-4N, 4N and >4N content; tet- shows the average of all tet- time points. Error bars shown the standard deviation from the mean of three independent replicates.

### 3.5.2 Depletion of TbRH2A cause the accumulation of DNA damage

To understand if loss of TbRH2A has wider consequences for nuclear genome integrity, levels of DNA damage were next investigated using the TbRH2A$^{RNAi}$ cell line. Phosphorylation of histone H2A on Thr130 (γH2A), generating the *T. brucei* equivalent of the variant histone γ-H2AX (Kinner *et al.*, 2008), occurs in response to increased levels of lesions, including double stranded DNA breaks (Glover and Horn, 2012). Expression of γH2A was assessed using western blotting and indirect immunofluorescence with anti-γH2A antiserum over 36 hr of culture with and without tet-induction (Figure 3.17). The number of cells in the populations that displayed nuclear γH2A staining increased dramatically with time, rising from ~17% of un-induced cells to ~50% and ~73% of cells after 24 and 36 hr RNAi induction, respectively (Figure 3.17A and B). High-resolution imaging showed that γH2A staining after RNAi induction was often diffuse throughout the nuclei, although in some cells more discrete foci were apparent (Figure 3.17A). Western blot analysis of whole protein extracts using anti-γH2A antiserum confirmed the dramatic increases in modified histone marker 24 and 36hr post tet-induction (Figure 3.17C).

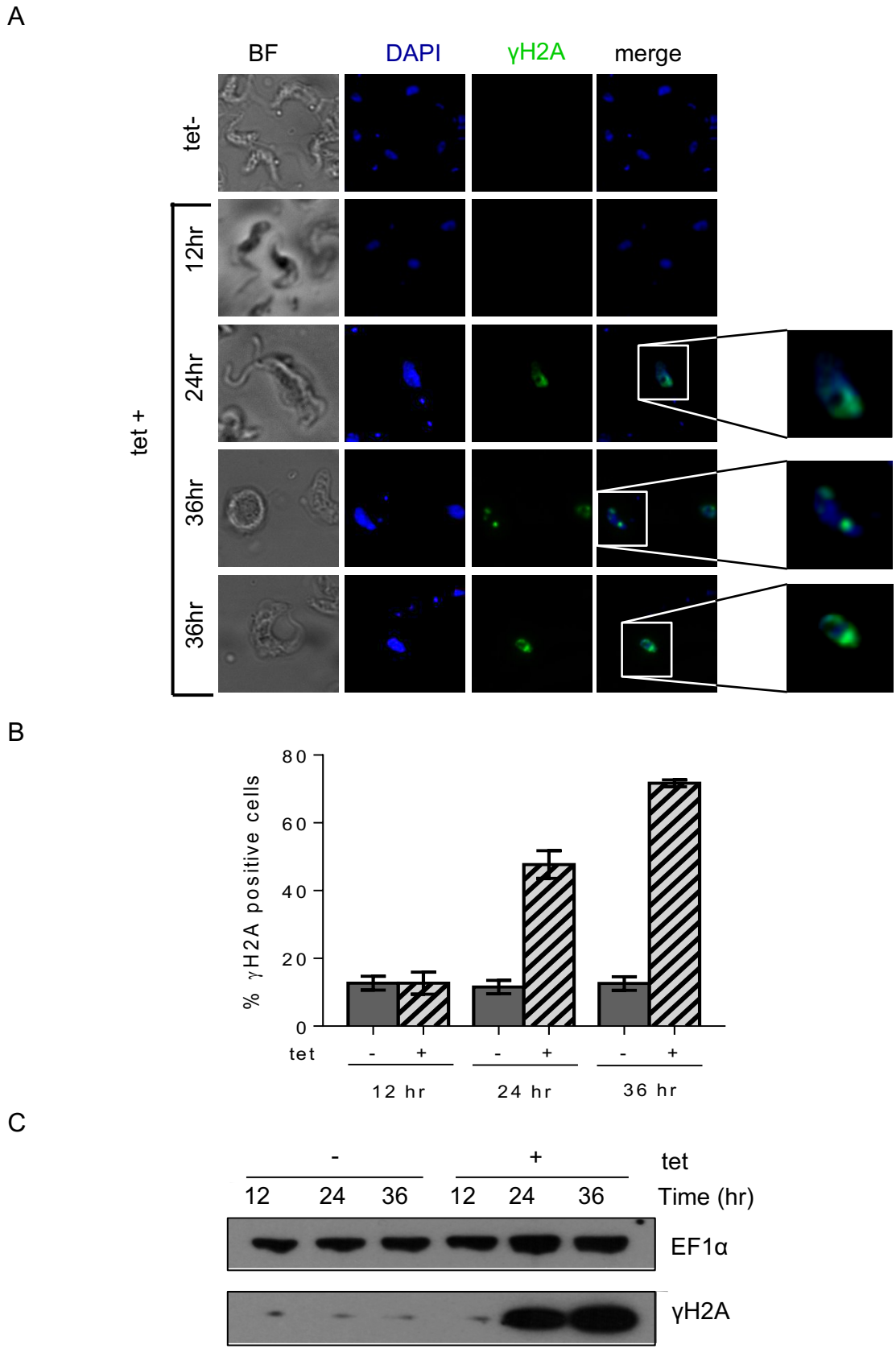



**Figure 3.17 Depletion of TbRH2A causes increased nuclear DNA damage**

A) High-resolution imaging of DAPI (blue) and γH2A (green) immunofluorescence with (tet +) and without (tet-) RNAi induction, at various time points. Images to the right show increased magnification of the boxed nuclear DNA. B) Bar graphs showing the percentage of tet + and tet- populations positively staining for γH2A after 12, 24 and 36 hrs growth; error bars show standard deviation from the mean of three independent repeats. C) Western blot detection of γH2A in whole protein extracts after 12, 24 and 36 hrs growth of tet + and tet- cells. EF1α staining is shown for each sample as a loading control.

### 3.5.3 TbRH2A depleted cells continue to synthesise DNA

To ask if loss of TbRH2 after RNAi was accompanied by impaired DNA replication, incorporation of the thymidine analogue EdU was measured via immunofluorescence using Click-IT® chemistry (section 2.5.4). TbRH2A$^{RNAi}$ cells were cultured with and without tet and, for the last 4 hr of each time point, with the addition of EdU to a final concentration of 150 μM. Cells were then fixed and stained for both EdU and γH2A (using antiserum) to assess whether any change in nuclear DNA replication was associated with changes in DNA damage, given the increase in γH2A (Figure 3.17), and content, as seen by flow cytometry analysis (Figure 3.16). No significant decrease in the number of cells staining for EdU incorporation was seen, as even after 36 hr of tet-induction the number of positive cells reduced to only ~96% of the population (Figure 3.18), with many cells showing both EdU and γH2A signals. Super-resolution imaging of cultures with and without tet after 36 hr of growth revealed co-localisation of EdU and γH2A signals (Figure 3.19). In the absence of RNAi induction most cells showed a single focus of γH2A signal, whereas RNAi induced cells either showed a few foci or signal throughout the nuclei. However, in both cases the signals overlapped with incorporated EdU. These data indicate that TbRH2A-depleted cells contain significantly increased amounts of nuclear DNA damage, which does not halt DNA synthesis.

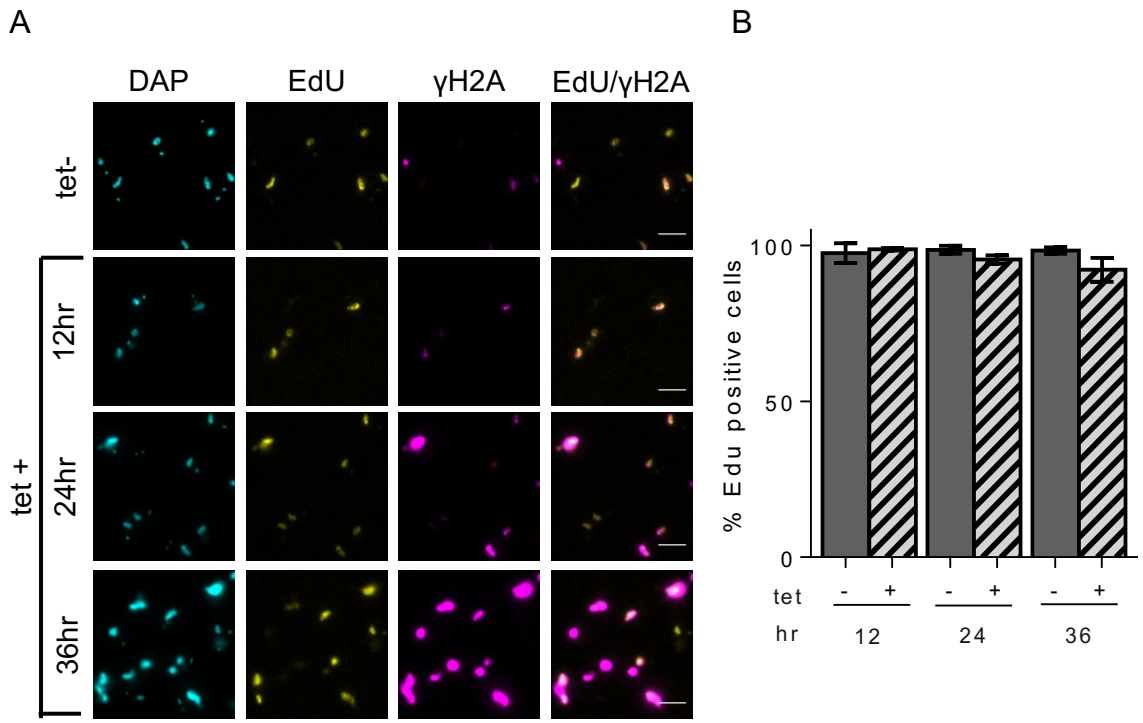



**Figure 3.18 TbRH2A-depleted *T. brucei* cells continue to synthesis DNA.**
A) Immunofluorescence imaging of DAPI (cyan), EdU (yellow) and γH2A (magenta) staining with (tet +) and without (tet-) induction of RNAi. Scale bar = 5 μm. B) Bar graph showing the percentage of tet + and tet- populations positively staining for EdU incorporation after 12, 24 and 36 hrs of growth.

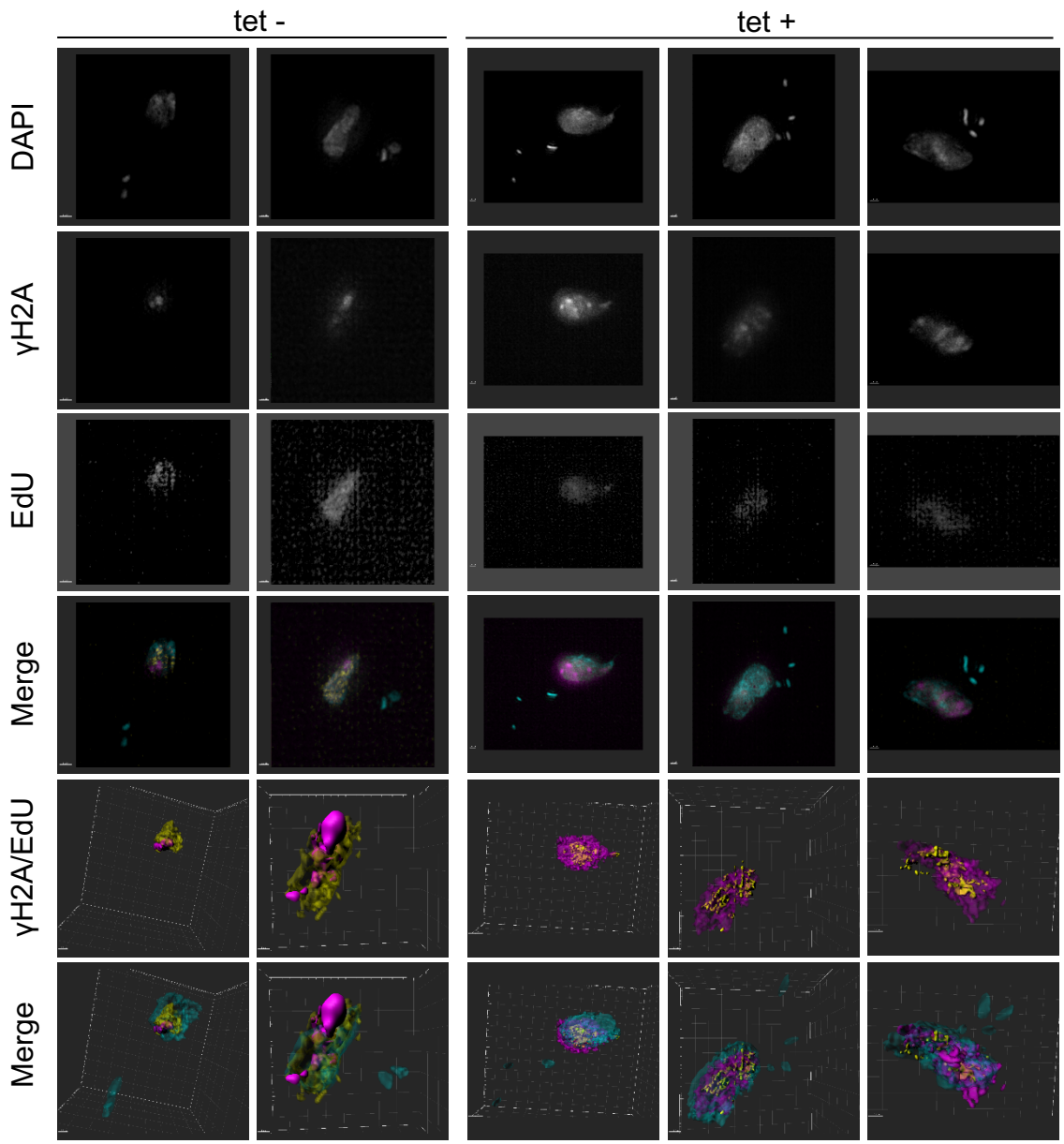



**Figure 3.19 Super-resolution imaging of EdU incorporation and γH2A.**
Super-resolution images of DAPI, γH2A and EdU are shown along with 3D reconstructions of staining demonstrating coloalisation of the signal. The TbRH2A$^{RNAi}$ line was cultured in the presence (tet +) or absence of tet (tet -) for 36 hr before samples were analysed. Scale bar = 1 μm. Imaging and processing was carried out by Dr. Leandro Lemgruber.

## 3.6 Discussion

In this chapter, homology searches were used to clarify the number of putative RNase H enzymes encoded by *T. brucei*, and genetic approaches were used to localise the proteins and examine their importance in BSF cells. BLASTp and protein sequence analysis allowed the identification of three putatively catalytic RNase H proteins, one being the previous identified TbRH1 (Hesslein and Campbell, 1997). The two remaining RNase H domain-containing proteins appear to belong to the Type 2 RNase H family, with TbRH2A (Tb427.10.5070) showing greater homology with eukaryotic RNase H2A species and TbRH3 (Tb427.10.4730) with bacterial RNase HII proteins. In addition to these factors, which all appear to retain catalytic residues, two likely TbRNase H2 structural subunits (Tb427.01.4220 and Tb427.01.4730) were identified. These have a domain organisation consistent with encoding *T. brucei* homologues of eukaryotic RNase H2B and H2C proteins, but this was not tested.

### 3.6.1 TbRH1 is non-essential in cultured BSF *T. brucei*

Immunofluorescence detection of a 12-Myc tagged variant reveals TbRH1 to show nuclear localisation (as previously suggested by Kobil & Campbell 2000), with increased expression during S phase cell cycle stages, possibly indicating a greater requirement for the RNA-DNA hybrid resolution during DNA replication. Additionally, the protein shows no distinct subnuclear localisation and instead remains distributed across the nucleus throughout the BSF cell cycle. Interestingly, TbRH1 lacks a predicted MTS and does not shown additional mtDNA localisation, in contrast to mammalian RNase H1, which is essential for mtDNA amplification (Cerritelli *et al.*, 2003). Hence, TbRH1 appears not to have a role in kDNA replication or maintenance in *T. brucei*.

Analysis of *Tbrh1*-/- parasites shows loss of the protein has no deleterious effect on BSF growth or cell cycle progression, supporting the hypothesis that RNase H1 is not essential for eukaryotic nuclear genome viability under normal conditions, as demonstrated in yeast (Wahba *et al.*, 2011; O'Connell, Jinks-Robertson and Petes, 2015). Yeast lacking RNase H1 show only minor increased levels of chromosome

instability, though these effects are mostly seen when combined with loss of RNase H2, elevating instability to levels above those caused by RNase H2 null mutation alone (Wahba *et al.*, 2011; O'Connell, Jinks-Robertson and Petes, 2015). Interestingly, genomic instability documented in yeast strains lacking various mRNA biogenesis factors that display greater levels of RNA-DNA hybrid formation, is consistently reduced when RNase H1 is over expressed (Wahba *et al.*, 2011). Hence, RNase H1 has a role in resolving nuclear R-loops formed under stress conditions, such as perturbed transcription or RNA processing.

TbRH1 is likely to share a similar nuclear role with other eukaryotic RNase H1 proteins, which are not essential for nuclear genome viability. Indeed, TbRH1 may play a critical role in genome stability during stress conditions where R-loop formation may be more likely. Another possibility is that TbRH1 resolves R-loops that are isolated to specific genomic loci. In either case, RNase H2 is likely to be functionally redundant and so assume the role of R-loop resolution in the absence of TbRH1. Although structural genome changes resulting from a loss of TbRH1 have not been directly investigated, DNA damage and changes to gene expression in *Tbrh1-/-* parasites are explored in Chapter 6.

### 3.6.2 TbRH2A is essential for BSF *T. brucei*

Immunofluorescent detection of HA-tagged TbRH2A showed highly similar nuclear localisation to that of TbRH1, again with increased expression in S-phase cell cycle stages and diffuse localisation across the nucleus throughout the cell cycle. In contrast to TbRH1, however, TbRH2A was found to be essential, a finding that correlates with the essential nature of RNase H2 in mouse embryo development (Hiller *et al.*, 2012; Reijns *et al.*, 2012) and AGS-causing RNase H2 mutations in humans (Crow *et al.*, 2006), but contrasts with the non-essentiality of RNase H2 in yeast (Conover *et al.*, 2015; O'Connell, Jinks-Robertson and Petes, 2015). Mouse fibroblast cells lacking the RNase H2B subunit stall in the G2/M stage of the cell cycle and display increased DNA damage (Hiller *et al.*, 2012), both of which are also observed in *T. brucei* after TbRH2A depletion. However, further investigation by flow cytometry and EdU incorporation shows that, despite the high levels of γH2A

that accumulate after TbRH2 RNAi, the parasites continue to synthesise DNA, resulting in cells with >4N DNA content that fail to undergo faithful nuclear mitosis. RNaseH2B-null mouse fibroblasts also fail to undergo mitosis yet do not continue to synthesise DNA, as flow cytometry analysis showed accumulation in the G2/M phase (Hiller *et al.*, 2012). Hence, in both cases the G2/M check point appears to be activated, yet in *T. brucei* DNA replication continues.

Although tet-induced TbRH2A$^{RNAi}$ cells undergo ~4 rounds of the cell cycle before stalling, TbRH2A mRNA transcripts are likely to persist initially, and so the rapid nature of the phenotype observed after TbRH2A loss in *T. brucei* differs slightly to that of RNaseH2B-null mice that develop into embryos (Hiller *et al.*, 2012). Therefore, while RNase H2 appears essential for both *T. brucei* and mammals, the importance of the complex may be more immediate for the parasite. Interestingly, TbRH2A has conserved all the critical binding sites and catalytic residues predicted to be necessary for both R-loop resolution and initiation of RER. Hence, it is plausible that accumulation of either incorporated ribonucleotides or R-loops (or both) leads to the vast amounts of DNA damage observed in parasites depleted of TbRH2A. Accumulation of DNA damage in RNase H2B-null mouse cells has been attributed to the incorporation of overwhelming amounts of ribonucleotides in the DNA (Hiller *et al.*, 2012; Reijns *et al.*, 2012), although R-loop accumulation was not assessed. This may also be the case in *T. brucei*, yet the rapid onset of stalling implies that rather than a build-up of mutagenic ribonucleotides during each round of the cell cycle leading to accumulation of DNA damage and stalling at the G2/M stage, TbRH2A is completely essential for an aspect of basic genome biology and *T. brucei* is unable to complete the cell cycle without it. For example, TbRH2A may be required for unperturbed replication of the genome, chromosome segregation or transcription of vital genes. Whether TbRH2A acts to remove unresolved R-loops or embedded ribonucleotides that prevent the execution of one or more of these functions is explored further in chapter 5.

### 3.6.3 *T. brucei* encodes a third putatively catalytic RNase H

As in other eukaryotes, with the known exception of *S. cerevisiae*, an RNase H domain-containing protein is required for mtDNA replication (Cerritelli and Crouch, 2009). Given this precedent, is it possible that *T. brucei* kDNA replication also requires such an activity, which is normally assumed by a type 1 RNase H. However, the data provided here indicate that TbRH1 shows no kDNA localisation. In contrast, TbRH3 has a predicted MTS and immonflourescence of Myc-tagged protein indicates localisation to the kDNA. In fact, TbRH3 localises in discrete foci at the poles of duplicated kDNA, which is a pattern shared with other proteins, including a kDNA replication polymerase, Pol ß (Saxowsky *et al.*, 2003), supporting a functional role for TbRH3. However, as TbRH3 has higher sequence identity with bacterial RNase HII proteins than eukaryotic RNase H1, and the RNase H domain is somewhat larger than is found in other type 2 RNase H proteins, its activity may be diverged or specialised when compared to typical RNase H1 functions in RNA-DNA hybrid resolution during mtDNA amplification (Cerritelli *et al.*, 2003). As the protein follows a localisation pattern similar to that of Pol ß, perhaps it functions in the initiation of DNA replication, as reported in *E. coli* plasmid (Itoh and Tomizawa, 1980) and mtDNA replication (Xu and Clayton, 1996). Whatever its role, TbRH3 is likely to be an essential protein in *T. brucei* since efforts to generate even a heterozygous *TbRH3* knockout mutant in BSF parasites were unsuccessful (no data shown), although this was not experimentally confirmed by RNAi. Nonetheless, TbRH3 is conserved in at least *Leishmania* and *T. cruzi*, suggesting kinetoplastids have evolved a third class of RNase H that acts in facets of kDNA function, perhaps reflecting the unique structure of this organellar genome (Jensen and Englund, 2012).

# 4 RNA-DNA Hybrid Mapping in the *T. brucei* Genome

## 4.1 Introduction

A variety of techniques have been used to map R-loops genome-wide; most are variants of immunoprecipitation (IP) using the DNA-RNA hybrid-binding antibody S9.6 (DRIP) followed by quantitative PCR (DRIP-qPCR), next-generation sequencing (DRIP-seq) or hybridisation of isolated DNA to tilling arrays (DRIP-chip) (Halász *et al.*, 2017). Alternatively, catalytically inactive RNase H1 can be used in an affinity pull-down assay to isolate the hybrids the protein binds to, in a protocol termed DRIVE-seq (DNA-RNA *in vitro* enrichment followed by next-generation sequencing of the DNA) (Ginno *et al.*, 2012). Similarly, the "R-ChIP" protocol uses catalytically inactive RNase H1 expressed in human cells to capture R-loops *in vivo* via ChIP-seq by targeting the mutant RNase H, which in theory binds to but does not resolve R-loops, for pull-down (Chen *et al.*, 2017). Both DRIP-seq using native chromatin samples (i.e. not treated with a fixative such as formaldehyde) and DRIVE-seq identified R-loops highly correlated with positively GC-skewed, CGI promoters in the human genome (Ginno *et al.*, 2012, 2013). Though there was a strong overlap in DRIP-seq and DRIVE-seq signal over promoter regions, in total, DRIP-seq identified 20,862 enrichment peaks, whereas DRIVE-seq only revealed 1,224 peaks (Ginno *et al.*, 2012). R-ChIP in human cells identified an intermediate number (12,521) of significantly enriched peaks (Chen *et al.*, 2017). DRIP-chip and DRIP-seq using formaldehyde fixed chromatin (also referred to as S9.6 ChIP-seq) have been effectively used to map R-loops in the *S. cerevisiae* (El Hage *et al.*, 2010; Wahba, Gore and Koshland, 2013; Chan *et al.*, 2014) and *S. pombe* (Ohle *et al.*, 2016) genomes.

More recently, variations on the DRIP-seq protocol have been developed: RNA-DNA immunoprecipitation (RDIP), where several changes were made to the DRIP protocol including switching from enzymatic digestion of the chromatin to sonication and pre-treatment with RNase I to remove single stranded RNA (Nadel *et al.*, 2015); S1-DRIP-seq, where chromatin is pre-treated with S1 nuclease before sonication (Wahba *et al.*, 2016); bis-DRIP-seq, where bisulfite converts cytosine residues to uracil in ssDNA before S9.6 antibody capture of hybrids, allowing identification of ssDNA associated with DNA-RNA hybrids and increasing mapping resolution of R-

loops (Dumelie and Jaffrey, 2017); complementary DRIP (DRIPc-seq) where recovered RNA is reverse transcribed and cDNA is sequenced to reveal strand information (Sanz *et al.*, 2016); and, finally, single-stranded DRIP (ssDRIP-seq) using fixed chromatin to reveal strand information via initial adapter ligation to only the 3' ends of DNA fragments during library preparation (Xu *et al.*, 2017). DRIP-seq using native chromatin has also been referred to as DIP-seq (Skourti-Stathaki, Proudfoot and Gromak, 2011), although there are no notable differences between these protocols. In all cases, pre-treatment of native or fixed chromatin, or "on-bead" treatment of the IP reaction, with *E. coli* RNase H can be used to verify S9.6 binding specificity to R-loops, as a reduction in IP recovered DNA or RNA should be observed in treated samples (El Hage *et al.*, 2014; Halász *et al.*, 2017). Additionally, mapping can be carried out in RNase H mutants as a further control, as DNA-RNA hybrid signal is expected to increase when RNase H activity is reduced.

### 4.1.1 Chapter aims

As R-loops have been linked to various biological processes (see section 1.4.1), as well genomic instabilities (see section 1.4.2), R-loop mapping was undertaken in WT and *Tbrh1*-/- BSF *T. brucei* cells. Several of the predicted R-loop associated processes are not entirely mechanistically understood in *T. brucei*, including DNA replication, transcription termination and transcription initiation, and so identifying regions of R-loop enrichment in the genome was hypothesised to either implicate or rule out the involvement of RNA-DNA hybrids in these fundamental processes. The DRIP-seq approach was taken using the S9.6 antibody and fixed chromatin samples, as performed in *S. cerevisiae* (El Hage *et al.*, 2014) and *S. pombe* (Ohle *et al.*, 2016), since *T. brucei* is also a single celled eukaryote and has a similar sized genome. In this chapter R-loop localisation across the core genome of *T. brucei* is compared in the presence and absence of TbRH1, thus identifying sites of R-loop localisation and inferring likely functions. The mapping revealed conserved sites of R-loop localisation, an association with RNA Pol II transcription initiation, correlation with nucleosome depletion, and a potentially novel role for R-loops in *trans*-splicing and polyadenylation. R-loop localisation in genomic regions associated with VSG switching is discussed in chapter 6.

## 4.2 Identifying RNA-DNA hybrid forming regions in the *T. brucei* genome

DRIP-seq was employed using formaldehyde fixed chromatin samples taken from ~2 x $10^8$ BSF Lister 427 *T. brucei* cells per sample (section 2.6). Briefly, chromatin was extracted and sheared using enzymatic digestion, before incubation with the S9.6 antibody to perform IP of the RNA-DNA hybrids (Figure 4.1). The eluted chromatin was reverse cross-linked, and proteins digested. DNA libraries were prepared using Illumina's TrueSeq ChIP kit and 300 bp fragments including adaptor sequences were selected (carried out by Craig Lapsley) for sequencing with the Illumina Nextseq 500 platform (performed by Glasgow Polyomics). Sequence reads were aligned to a "hybrid" reference genome (Hutchinson, Glover and Horn, 2016) consisting of the *T. brucei* TREU927 v5.1 Mb-sized chromosomes and Lister 427 BES and mVSG ES contig assemblies (carried out by Dr Kathryn Crouch), since the TREU927 genome is, so far, better annotated than the *T. brucei* Lister427 equivalent, allowing for more in-depth downstream analysis (section 2.8.4). DNA isolated from both the DRIP reactions, as well as from samples of 'input' chromatin DNA from each (i.e. a sample of chromatin used for the IP reaction), were sequenced and aligned in this manner. This allowed DRIP-seq signal enrichment to be calculated as ratio of read coverage fold-change for each IP relative to the corresponding input sample (section 2.8.4.3). Hence, enrichment of DRIP-seq signal >1 corresponds to RNA-DNA hybrid formation in the population.

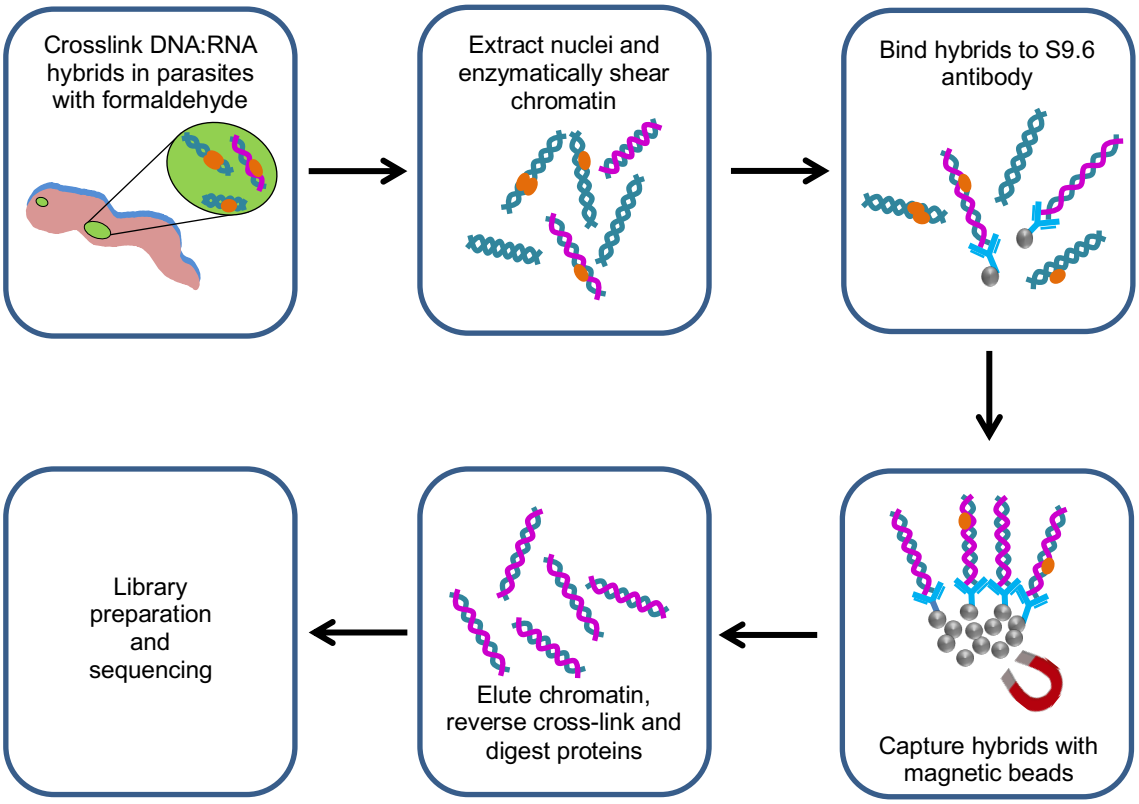



**Figure 4.1 DNA-RNA hybrid immunoprecipitation (DRIP)-seq scheme.**
Parasites were fixed with formaldehyde before the chromatin was extracted and enzymatically digested. A sample of chromatin was incubated with the S9.6 antibody, which binds RNA(pink)-DNA(blue) hybrids. The antibody was adhered to magnetic beads, allowing capture of the hybrids. Chromatin was then eluted and reverse cross-linked, and the proteins digested. The remaining DNA was then prepared as libraries and sequenced. See section 2.6 for detailed methods, and section 2.8.4 for analysis.

Numerous attempts were made to call peaks of DRIP-seq signal enrichment using existing peak-calling software packages SICER (Zang *et al.*, 2009), MACS2 (Zhang *et al.*, 2008) and HOMER (Heinz *et al.*, 2010). However, none were able to call DRIP-seq peaks with satisfactory accuracy, despite several attempts at optimisation (data not shown). Sanz *et al.* (2016) overcame the same problem using a custom hidden Markov Model that involved the manual annotation of DRIP peaks and then the development of a peak-calling pipeline based upon the manually defined model peaks. As this pipeline was developed for use with human DRIP-seq data (where enrichment is several fold higher and in fewer regions than seen in the *T. brucei* DRIP-seq signal, see below), and developing a model requires a larger number of data sets than presented in this study, DRIP-seq peaks were annotated manually. A "peak" was defined as a region where IP signal was enriched ≥ 1.2-fold above the input signal. Enrichment was calculated using non-sliding 50 bp windows, and the

coordinates of each window with sufficient enrichment were extracted (section 2.8.4.3). Coordinates of adjacent enriched 50 bp windows were then combined using a custom script (Figure 8.1) to generate final coordinates of DRIP-seq "peaks", which from here on are referred to as DRIP enriched regions.

### 4.2.1 R-loops are highly abundant in the *T. brucei* genome

Visual analysis of DRIP-seq signal across the 11 Mb chromosomes showed extensive coverage in both the WT and *Tbrh1*-/- samples, with good overlap in signal localisation between them (Figure 4.2). Broadly, DRIP-seq signal enrichment of both WT and *Tbrh1*-/- samples showed some correlation with the locations of tandem array sequences predicted with a tandem repeat finder algorithm that requires no prior knowledge of repeat composition, repeat length or number of repeat copies (lower track; Benson, 1999), as previously reported for many triplet repeats implicated in human genetic disease (reviewed in Freudenreich 2018). Curiously, this enrichment also seemed to be found within the subtelomeres that contain, among other gene types, the VSG tandem arrays, despite these regions being transcriptionally silent. This was particularly evident in the *Tbrh1*-/- sample.


**Figure 4.2 DRIP-seq signal across the 11 Mb-sized chromosomes of *T. brucei*.**
DRIP-seq signal for WT (pink) and *Tbrh1-/-* (green) cells is plotted as fold-change relative to input coverage (scale: 1-4 fold-change). Upper track shows transcripts on sense (black) and antisense (red) strands. Known centromeres are shown as peach circles. Lowest track shows tandem repeat sequences.

A total of 37,527 and 34,207 DRIP enriched regions were identified in WT and *Tbrh1-/-* samples, respectively. These enriched regions amount to ~22.3% (7,461,359 bp) of the genome in the WT sample population, and ~32.2% (10,795,714 bp) of the *Tbrh1-/-* cells (Figure 4.3A). Hence, the increased R-loop coverage in the *Tbrh1-/-* sample indicated that although ~8.6% less R-loop forming regions were identified, R-loops lengthened upon loss of TbRH1 (Figure 4.3B).

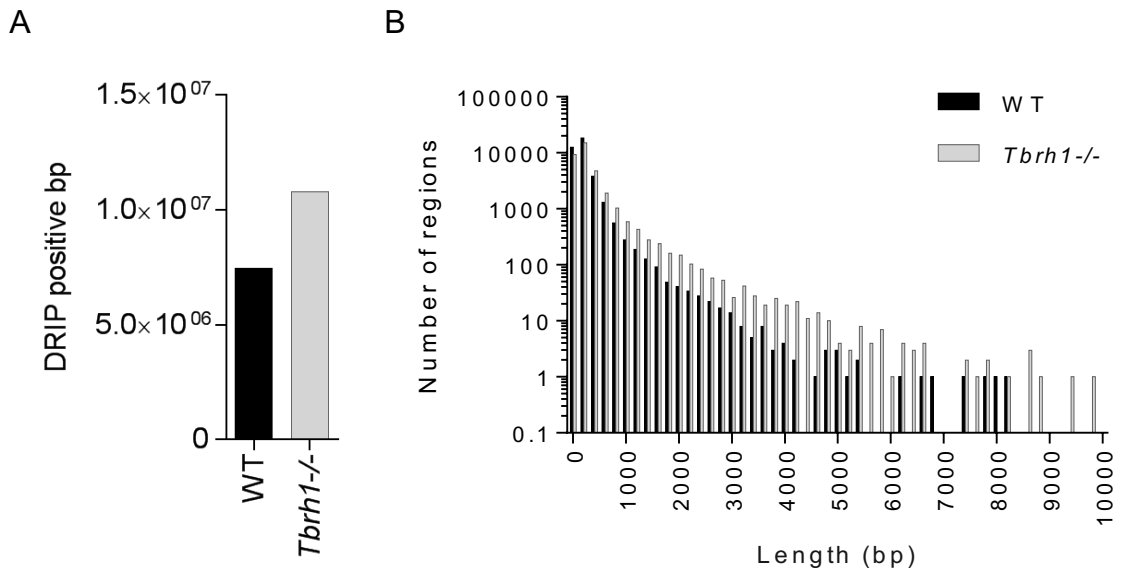



**Figure 4.3 DRIP-seq signal is more widespread after TbRH1 null mutation.**
A) Bar graph of the number of bp occupied by a DRIP enriched region in WT and *Tbrh1-/-* samples. B) Histogram of DRIP enriched region lengths in WT and *Tbrh1-/-* samples.

Localisation analysis revealed the majority of R-loops located to the RNA Pol II transcribed PTUs, with ~94.1% of the WT enriched regions and ~87.2% in the *Tbrh1-/-* population locating here (Figure 4.4). Though the number of PTU-associated DRIP enriched regions did not increase when TbRH1 was ablated, there was an increase in R-loops in other areas of the genome, including the largely non-transcribed VSG arrays, RHS-associated genes and across genes transcribed by RNA Pol I (Figure 4.4). R-loop localisation at each of these regions is explored in more detail below.

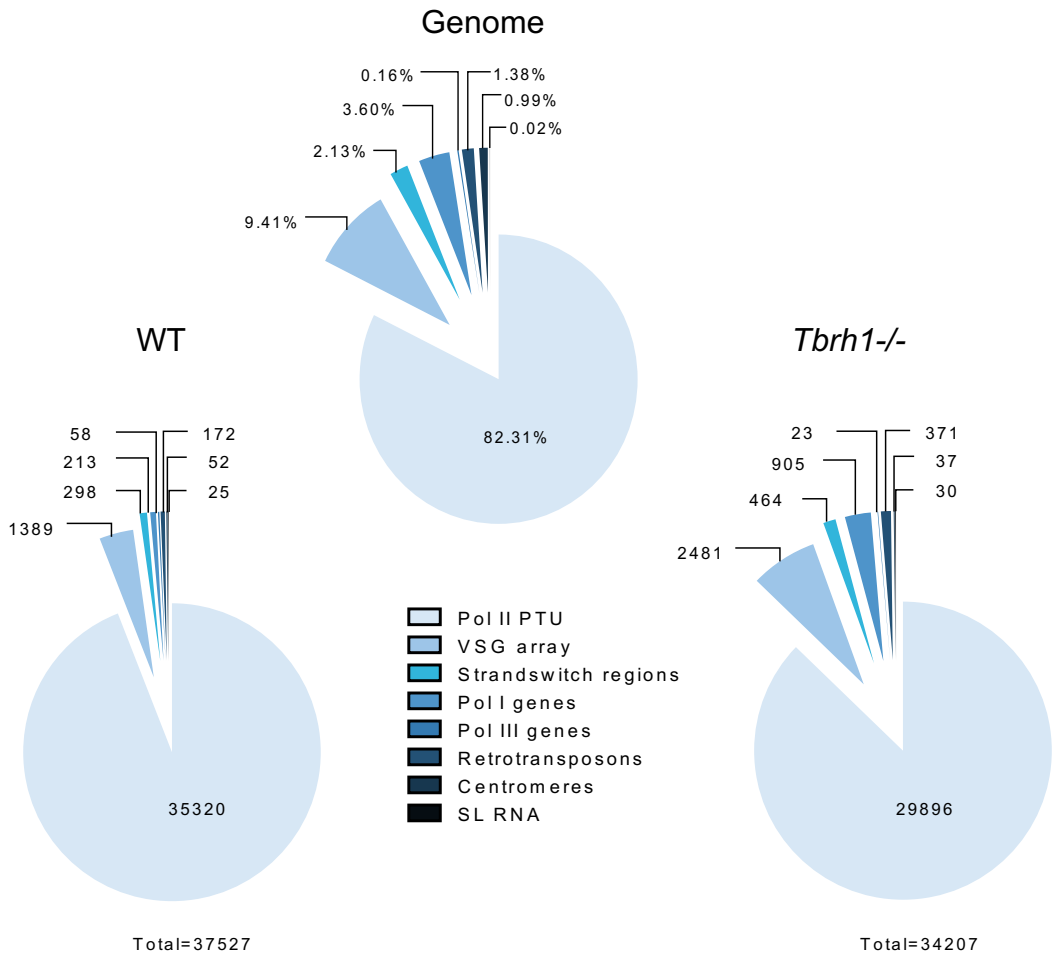



**Figure 4.4 Location analysis of DRIP enriched regions.**
The composition of the *T. brucei* genome (upper), based on the TREU927 plus Lister 427 expression sites hybrid genome assembly (section 2.8.4.3), and the distribution of DRIP enriched regions of WT and *Tbrh1-/-* samples across the genome elements (lower).

## 4.2.2 Many more R-loops form than are predicted

QmRLFS-finder is an algorithm developed with the aim of predicting likely R-loop forming sequences (Jenjaroenpun *et al.*, 2015). This algorithm is based upon findings from (Roy and Lieber, 2009), which indicate G clustering is important for initiating R-loop formation and, downstream of this, G-rich sequences are sufficient for elongation of the RNA-DNA hybrid. When the QmRLFS-finder algorithm was applied to the *T. brucei* TRU927 v5.1 genome sequence (section 2.8.4.7) only 565 R-loop forming sequences were predicted (Figure 4.5). Of these regions, 227 overlapped with DRIP enriched regions in the WT sample and 305 in the *Tbrh1-/-* sample, meaning the vast majority of R-loops found in the *T. brucei* genome, both with and without the presence of TbRH1, were not predicted by the algorithm.

Some predicted regions did, however, show good correspondence to DRIP enriched regions, including those within the locus housing histone H3 coding genes (Figure 4.5B).

A

| Cell line | WT | *Tbrh1-/-* |
|---|---|---|
| **DRIP regions only** | 37300 | 34001 |
| **Predictions only** | 338 | 260 |
| **Correct predictions** | 227 | 305 |

B

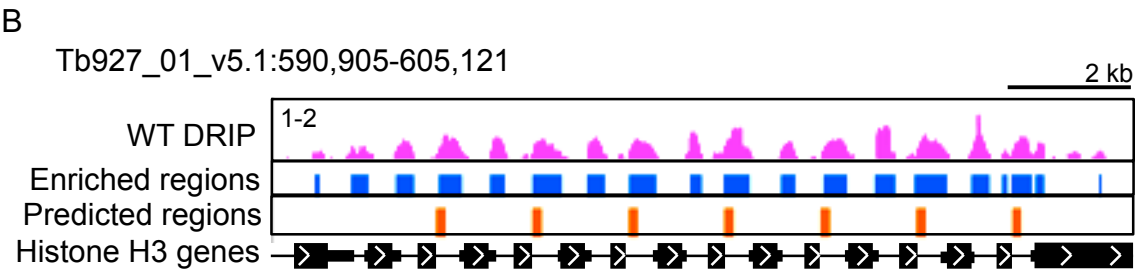



**Figure 4.5 QmRLFS-finder R-loop predictions.**
A) Table showing the number of DRIP enriched regions that were correctly predicted by the QmRLFS-finder algorithm. B) Screenshot of WT DRIP-seq signal (pink) over the locus in chromosome 1 that encodes histone H3 genes. Manually annotated DRIP enriched regions (blue) and QmRLFS-finder prediction (orange) are shown (Jenjaroenpun *et al.*, 2015). The scale in the DRIP track shows fold enrichment in the DRIP sample relative to input.

## 4.3 R-loops form at *T. brucei* centromeres

R-loops have previously been found to form over the centromeric repeats of *S. cerevisiae*, *C. elegans* and human cells, suggesting their conservation across eukaryotes (Castellano-Pozo *et al.*, 2013; Kabeche *et al.*, 2018). Supporting this, weak DRIP-seq signal enrichment was apparent across the annotated centromeric repeats of *T. brucei* chromosomes 1-8 (Echeverry *et al.*, 2012) in the WT sample. Furthermore, the signal was substantially increased in the *Tbrh1-/-* sample, implying TbRH1 targets centromeric R-loops (Figure 4.6).

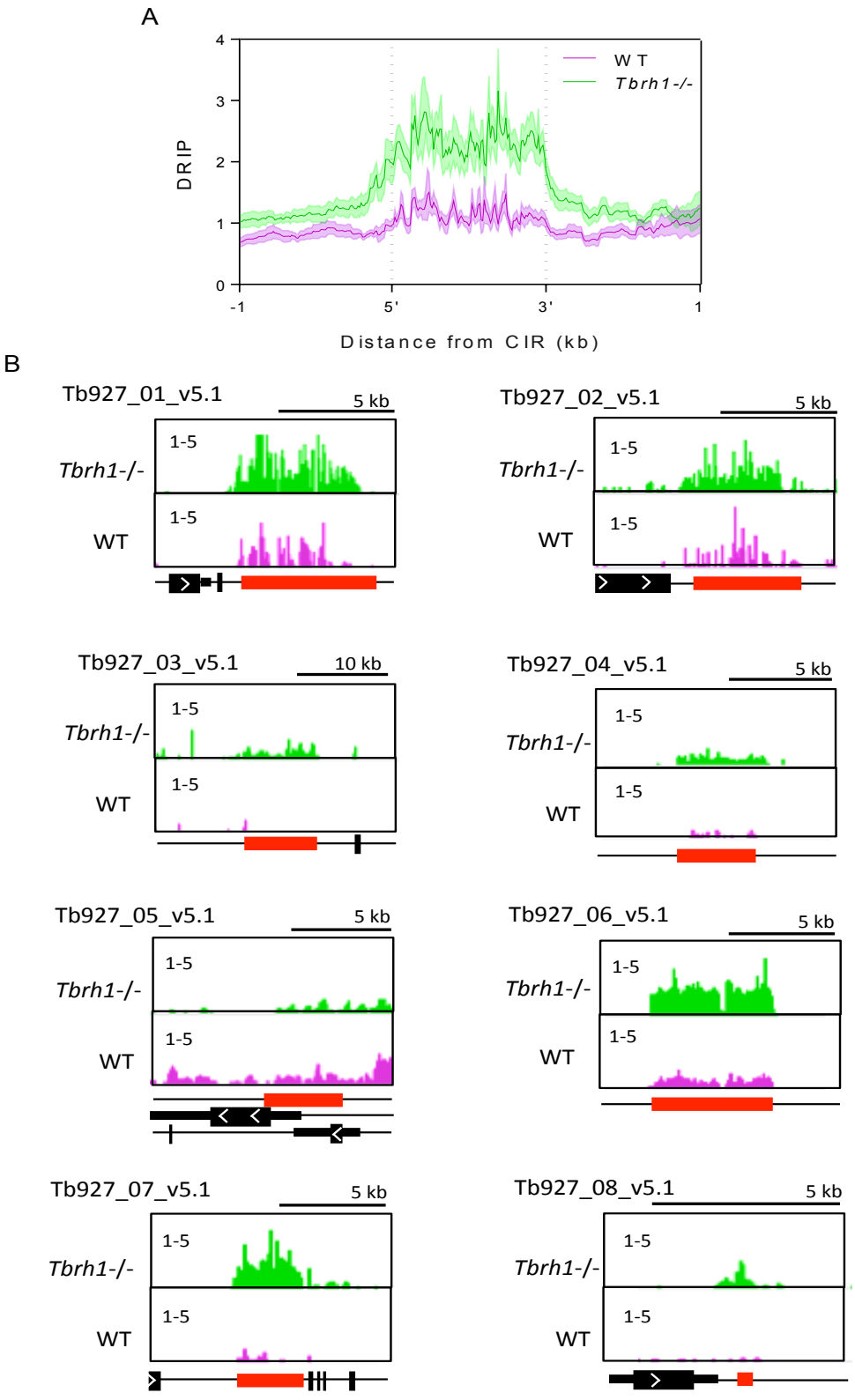



**Figure 4.6 TbRH1-targeted R-loops form at the centromeric repeat of *T. brucei*.**
A) Metaplot of average DRIP-seq signal of WT (pink) and *Tbrh1-/-* (green) samples over scaled centromeric repeats and 1 kb of up and downstream flanking sequence. Standard error is shown as shaded area. B) Screenshots of each annotated centromere DRIP-seq signal for WT (pink) and *Tbrh1-/-* (green) sample. Scale, 1-5 fold-change relative to input. Centromeric repeats are shown in red, CDS in thick black and UTR sequences in thin black lines.

## 4.4 R-loops at RNA pol I transcribed genes

In common with other eukaryotes, *T. brucei* rRNA genes are transcribed by RNA Pol I, as arrays. Such rRNA arrays have previously been shown to form R-loops in yeast (Chan *et al.*, 2014; El Hage *et al.*, 2014). Here, minor DRIP-seq signal enrichment was seen across loci encoding the rRNA subunits in WT parasites. However, this signal increased dramatically when TbRH1 activity is lost (Figure 4.7A, C). DRIP-qPCR was performed with two independently generated DRIP samples targeting three regions of the rRNA loci; the rRNA Pol I promoter, the 5.8S coding region and 28.S coding region (Figure 4.7B). For both WT and *Tbrh1-/-* samples of each replicated, IP reactions were set up in duplicate and one was on-bead treated with *E. coli* RNase HI (EcRHI) (section 2.6.3) to ensure RNA-DNA hybrids were being detected. In each case the percentage of input sequence detected in IP samples was higher in *Tbrh1-/-* samples compared to WT, in agreement with DRIP-seq mapping. EcRHI treatment reduced signal in both WT and *Tbrh1-/-* samples, confirming detection of RNA-DNA hybrids (Figure 4.7B).

Protein-coding genes are also transcribed by RNA pol I in *T. brucei*, including procyclin and associated genes (PAGs) (Günzl *et al.*, 2003), VSGs and associated ESAGs (Günzl *et al.*, 2003), and the monocistronically transcribed mVSGs (Alarcon *et al.*, 1994; Graham and Barry, 1995). The number of DRIP enriched regions increased in the VSG BESs in response to loss of TbRH1 (Figure 4.7), a phenomenon discussed further in Chapter 6. R-loop localisation was also seen in the procyclin loci and in mVSG expression sites, which is somewhat contradictory, since both these loci should be transcriptionally silent in BSF *T. brucei*. In principle, this appears to rule out these R-loops forming *in cis* during transcription, but where they may be generated and then form *in trans* is unclear. Since the mVSG expression sites are immediately downstream of telomeres, which are prominent sites of R-loop formation in human cells (Rippe and Luke, 2015), as well as in *T. brucei* (Nanavaty *et al.*, 2017), the predominant DRIP signal located towards the telomeric repeats may suggest much of the mVSG R-loop detection derives from the telomere. However, some signal can also be observed upstream of the mVSG genes in some cases (Figure 4.7D, Figure 8.3), and its source is less clear.

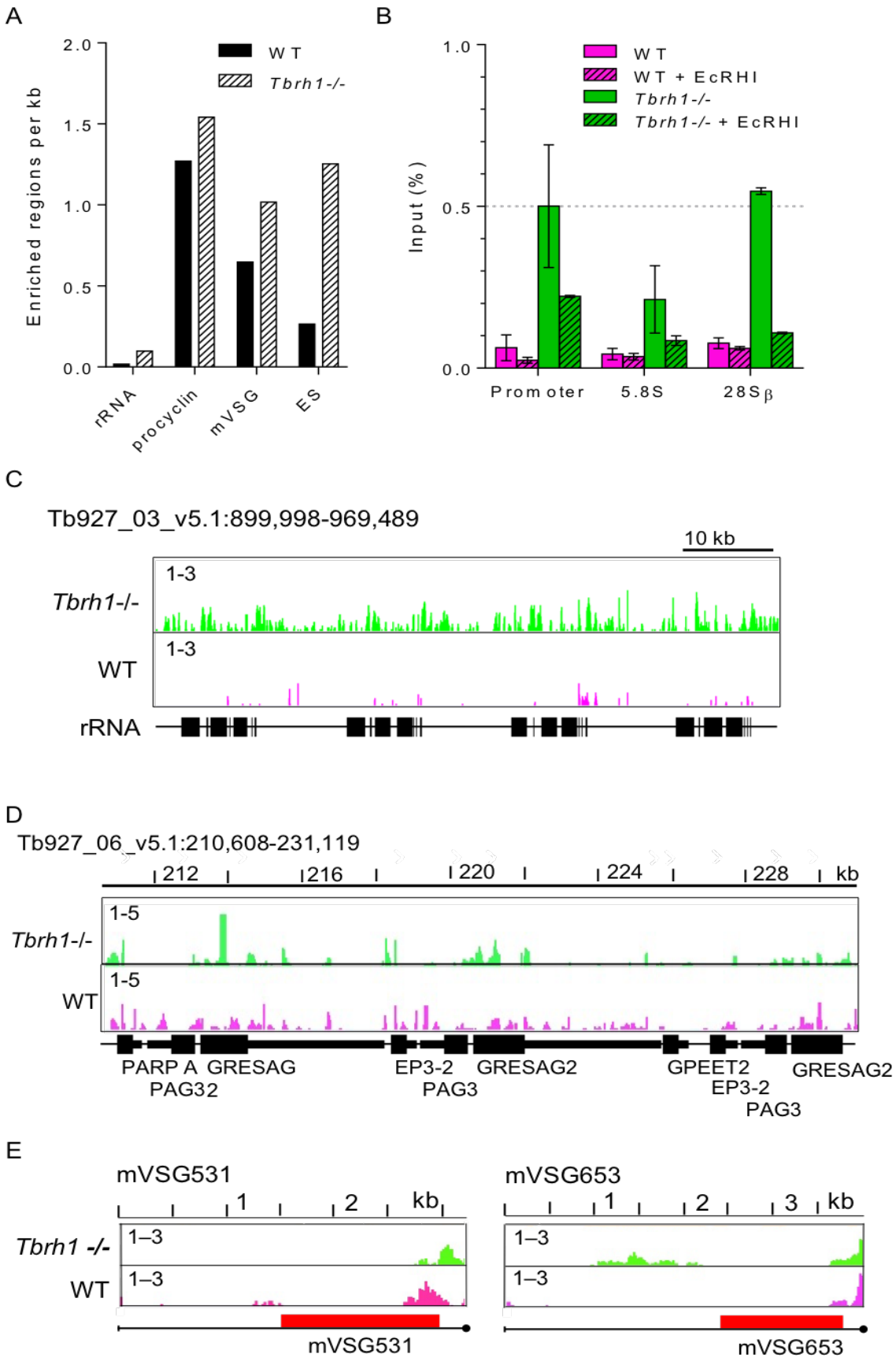

A
Enriched regions per kb
2.0
1.5
1.0
0.5
0.0
WT
Tbrh1-/-
rRNA
procyclin
mVSG
ES
B
Input (%)
1.0
0.5
0.0
WT
WT + EcRHI
Tbrh1-/-
Tbrh1-/- + EcRHI
Promoter
5.8S
28Sβ
C
Tb927_03_v5.1:899,998-969,489
10 kb
Tbrh1-/-
1-3
WT
1-3
rRNA
D
Tb927_06_v5.1:210,608-231,119
212
216
220
224
228
kb
Tbrh1-/-
1-5
WT
1-5
PARP A
PAG32
GRESAG
EP3-2
PAG3
GRESAG2
GPEET2
EP3-2
PAG3
GRESAG2
E
mVSG531
1
2
kb
Tbrh1 -/-
1–3
WT
1–3
mVSG531
mVSG653
1
2
3
kb
1–3
1–3
mVSG653

**Figure 4.7 R-loops form at sites of RNA Pol I transcription in *T. brucei*.**
A) Bar graph showing the number of enriched regions found per kb in each type of RNA pol I transcribed sites in WT and *Tbrh1-/-* cells. B) Graph showing DRIP-qPCR targeting the rRNA promoter, the 5.8S coding region and 28S coding region of the rRNA locus. Percentage of input sequence detected in IP samples is shown for WT (pink) and *Tbrh1-/-* (green) samples. In both cases dash lined bars shown IP reactions treated with EcRHI. Error bars shown SEM of two independent replicates. C-E) Screenshots of DRIP-seq signal for WT (pink) and *Tbrh1-/-* (green) samples over the rRNA array of chromosome 8 (C), procyclin expression site (D) and two representative mVSG expression sites (E).

## 4.5 R-loops form over *T. brucei* RNA Pol III transcribed genes

In yeast, R-loops were found to form over RNA Pol III transcribed tRNA and snRNA genes, the abundance of which increased with double null-mutation of RNase H1 and RNase H2 (El Hage *et al.*, 2014). In this study DRIP-seq signal was enriched over both snRNA and tRNA *T. brucei* genes (Figure 4.8) but, paradoxically, a loss of DRIP-seq signal was seen in the absence of TbRH1. DRIP-qPCR however, reveals increased detection of input sequence in *Tbrh1-/-* IP samples compared to WT over both tRNA and snRNA genes (Figure 4.8B). Signal is reduced by treatment with EcRHI, hence RNA-DNA hybrids have been detected. Although DRIP-seq peaks are still observed in the null mutant line, it is possible that scaling during normalisation of the sequencing data (section 2.8.4) has masked any increase in signal between WT and *Tbrh1-/-* DRIP-seq mapping at these loci.

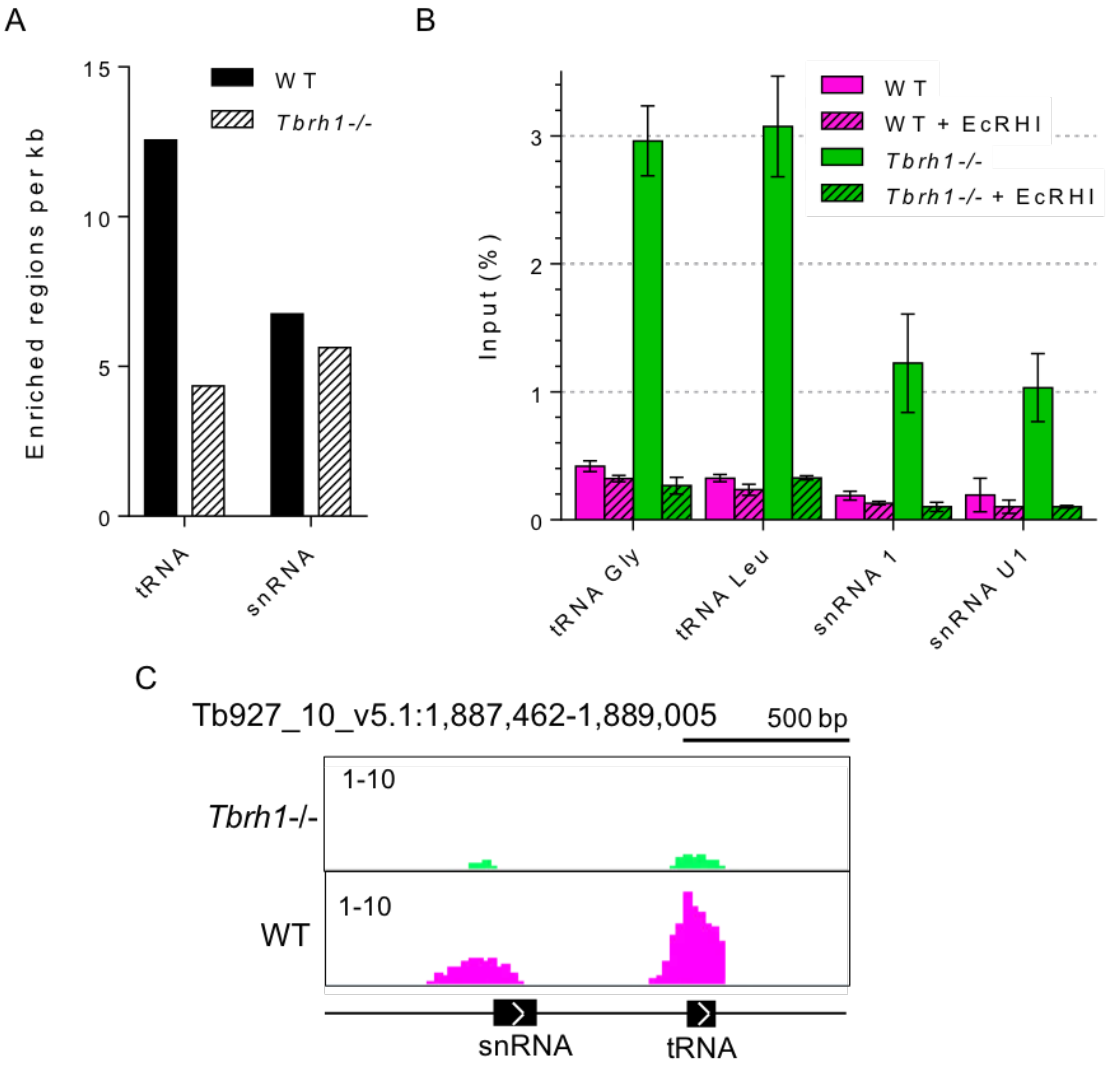



**Figure 4.8 R-loops form over RNA Pol III transcribed snRNA and tRNA gene.**
A) Bar graph showing the number WT and *Tbrh1-/-* DRIP enriched regions per kb found over tRNA and snRNA genes. B) Graph showing DRIP-qPCR targeting two tRNA and two snRNA genes. Percentage of input sequence detected in IP samples is shown for WT (pink) and *Tbrh1-/-* (green) samples. In both cases dash lined bars shown IP reactions treated with EcRHI/ Error bars shown SEM of two independent replicates. C) Screenshot of region on chromosome 10 where both an snRNA and tRNA are located. WT (pink) and *Tbrh1-/-* (green) DRIP-seq signal is shown.

## 4.6 RHS-associated genes are prominent sites of R-loop formation

RHS-associated genes and pseudogenes are a highly abundant and variable family in *T. brucei* that are expressed from subtelomeric arrays (Bringaud *et al.*, 2002). Little enrichment was found over the RHS genes in the DRIP-seq signal of WT cells, but signal was substantially increased with loss of TbRH1 (Figure 4.9).

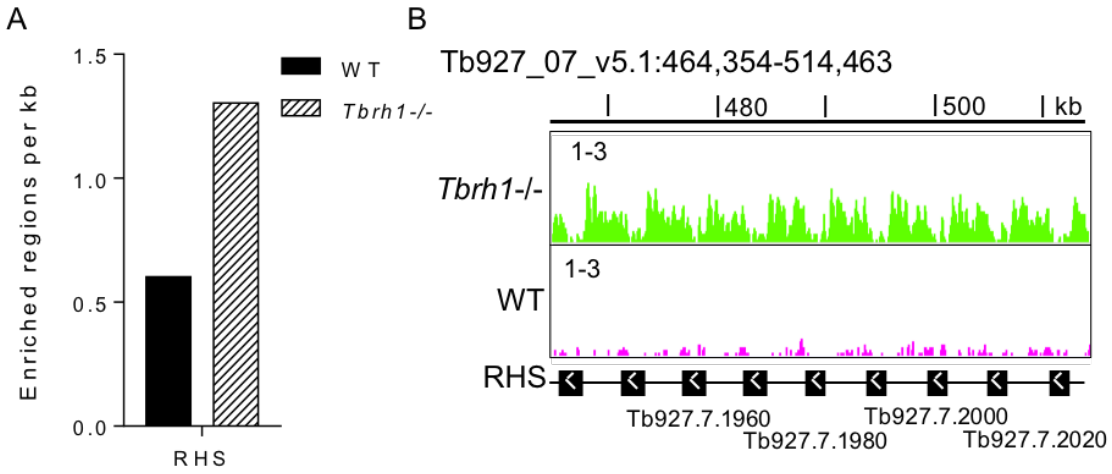



**Figure 4.9 TbRH1-sensitive R-loops form over RHS loci.**
A) Bar graph showing the number of DRIP enriched regions per kb found over the RHS regions in WT (black) and *Tbrh1-/-* (grey, striped) cells. B) Screenshot of DRIP-seq signal enrichment over a section of RHS gene array found in chromosome 7. WT DRIP-seq signal is shown in pink and *Tbrh1-/-* in green.

## 4.7 R-loops at RNA pol II transcribed PTUs

In order to explore the most abundant localisation of R-loops within the RNA Pol II transcribed PTUs, those DRIP enriched regions that were located within the PTUs were further classified according the regions they overlapped most extensively with: CDS, 5' UTR, 3' UTR or intergenic sequences. Most R-loops were found to associate with intergenic or UTR sequences, with only ~34% and ~39% of enriched regions locating to CDS regions in WT and *Tbrh1-/-* cells, respectively, despite this feature constituting 53% of the PTUs (Figure 4.10A). Inspection of DRIP-seq signal across the RNA Pol II PTUs confirmed this, as DRIP-seq signal was found to be most clear and of greater enrichment levels in the regions flanking CDSs, with good correspondence between WT and *Tbrh1-/-* sample profiles (Figure 4.10B). Indeed, the proportion of DRIP enriched regions that located to each PTU feature did not drastically change upon loss of TbRH1 (Figure 4.10).

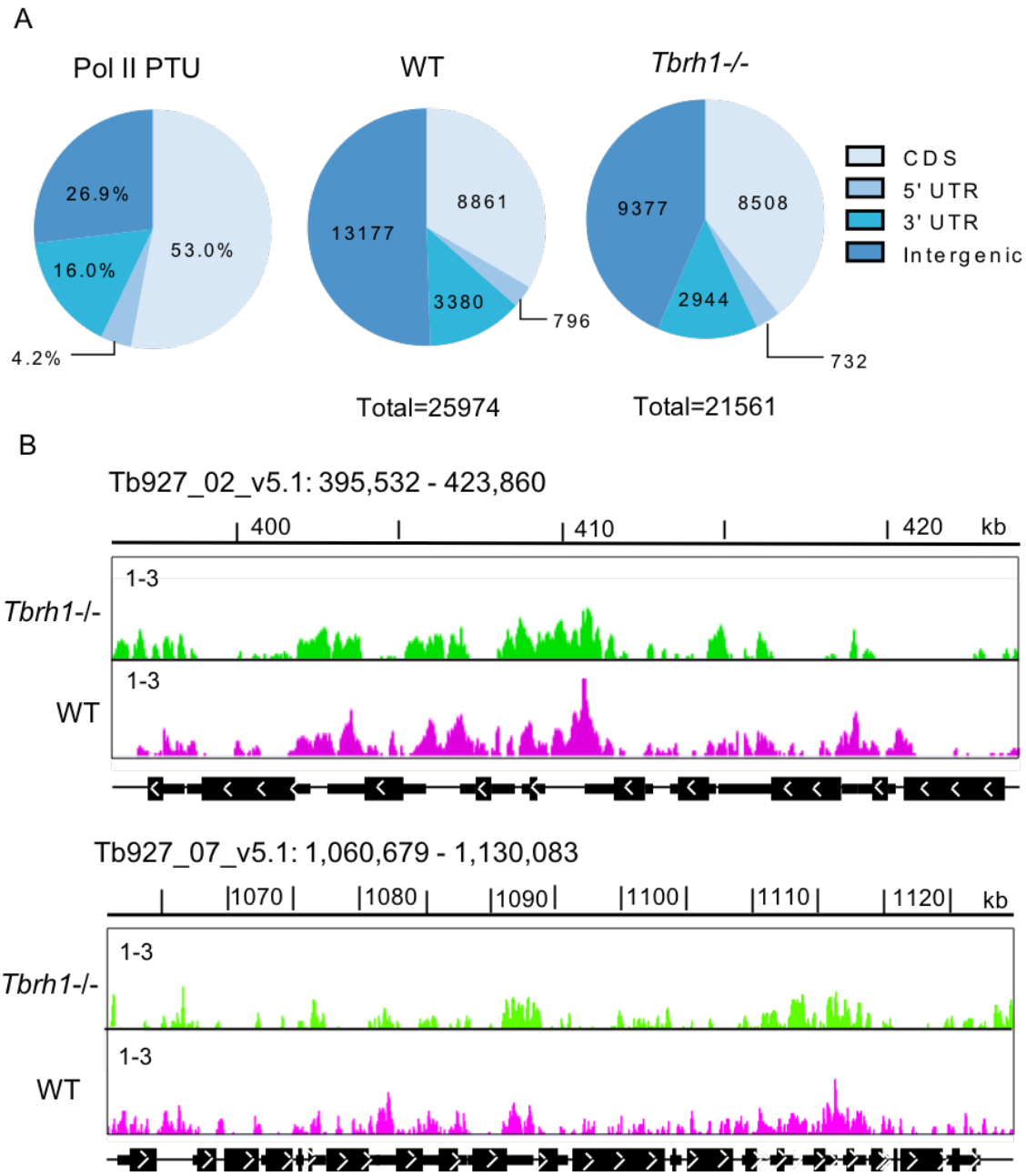



**Figure 4.10 Distribution of DRIP-seq signal and enriched regions over the PTU genomic elements.**
A) The composition of RNA Pol II transcribed PTUs (left) shows the proportion of CDS and UTR sequences. The amount of DRIP enriched regions associated with each element is shown for both WT (centre) and *Tbrh1-/-* (right) samples. B) Screenshots of RNA Pol II transcribed protein-coding genes in two areas of chromosomes 2 and 7. WT DRIP-seq signal is shown in pink and *Tbrh1-/-* in green; scale bars show enrichment in DRIP relative to input, as in previous figures.

DRIP-qPCR, however, revealed increased detection of RNA-DNA hybrids within the coding regions of six RNA Pol II transcribed genes (Figure 4.11). In all cases, greater input sequence was detected in the *Tbrh1-/-* IP samples compared to WT, and these were diminished by EcRHI treatment. Hence, a global increase in RNA-DNA hybrid

levels across the PTUs may occur in *Tbrh1-/-*, which is masked by normalisation of DRIP-seq samples.

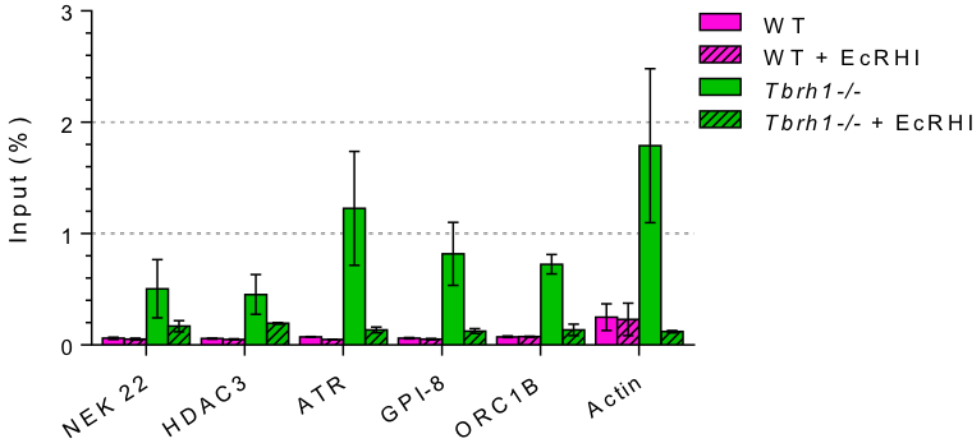



**Figure 4.11 DRIP-qPCR targeting RNA Pol II transcribed coding regions.**
Graph showing DRIP-qPCR targeting six RNA Pol II transcribed coding regions; NEK22 (Tb927.2.2120), HDAC3 (Tb927.2.2190), ATR (Tb927.11.14680), GPI-8 (Tb927.10.13860), ORC1B (Tb927.9.2030) and actin. Percentage of input sequence detected in IP samples is shown for WT (pink) and *Tbrh1-/-* (green) samples. In both cases dashed bars shown IP reactions treated with EcRHI/ Error bars shown SEM of two independent replicates.

Meta-analysis of the DRIP-seq signal over every RNA Pol II transcribed protein-coding gene revealed a striking pattern of signal enrichment over the flanking regions of each CDS, with enrichment peaking in close proximity to both the start and end of the CDS (Figure 4.12). This distribution of signal was both conserved and accentuated in the *Tbrh1-/-* cells.

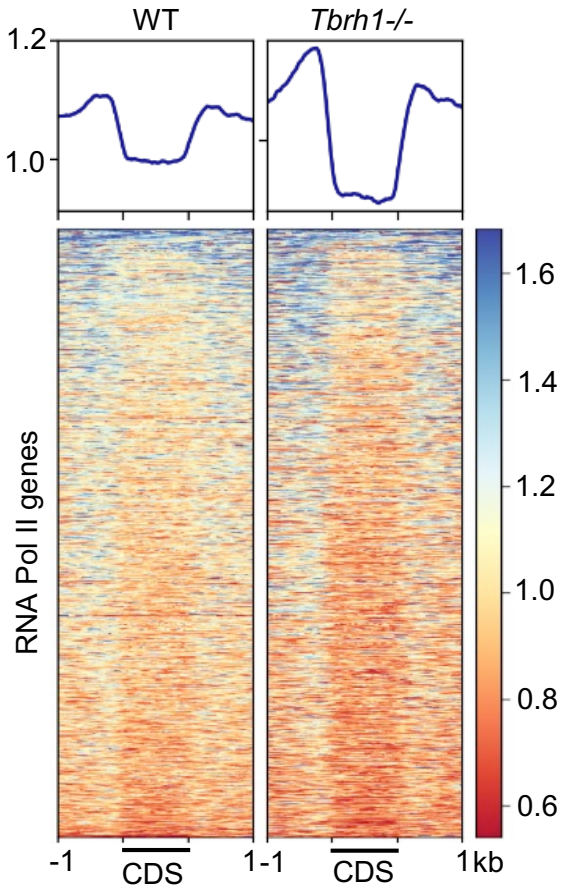



**Figure 4.12 DRIP-seq signal is enriched over the 5' and 3' flanking sequences of each RNA Pol II transcribed CDS.**
Heatmaps of DRIP-seq signal, for WT and *Tbrh1-/-* samples, over each RNA Pol II transcribed protein-coding CDS plus and minus 1 kb of flanking sequence. Metaplot profile summaries of the data are also shown above.

### 4.7.1 R-loops at sites of trans-splicing and polyadenylation

Motif analysis using the MEME algorithm (Bailey and Elkan, 1994) was used to analyse the DNA sequences of regions defined as DRIP-enriched within the PTUs. Three motifs were revealed; two poly(Y) sequences, and one poly(A) tract (Figure 4.13). This finding suggests a relationship between R-loop formation and mRNA maturation, as poly(Y) tracts are critical for both *trans*-splicing and polyadenylation of transcripts (Liang *et al.*, 2003).

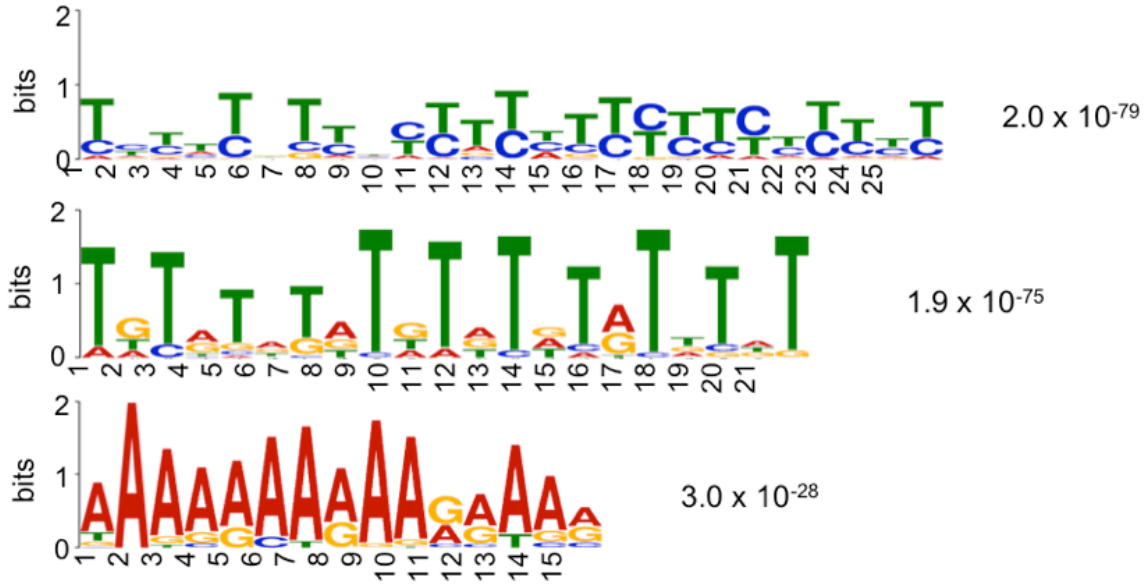



**Figure 4.13 Three motifs are associated with DRIP-seq enriched regions in the PTUs.**
Three motifs found in WT DRIP-seq intra-PTU enriched region sequences are shown. E- values are shown to the right of each motif.

As >50,000 PASs have been mapped for >8,000 *T. brucei* genes (Nilsson *et al.*, 2010; Siegel *et al.*, 2010), and an average of 2.8 SASs have been mapped per gene (Siegel *et al.*, 2011), the correlation between these processing sites and DRIP-seq signal was investigated. Inspection of WT DRIP-seq signal across the PTUs in comparison with SASs (Kolev *et al.*, 2010; Nilsson *et al.*, 2010; Jensen *et al.*, 2014) and PASs (Kolev *et al.*, 2010; Siegel *et al.*, 2010) locations revealed an apparent correspondence between DRIP-seq signal enrichment and PAS location, as higher DRIP-seq peaks were evident over areas more densely populated with PASs (Figure 4.14A). Any correlation with SASs was less clear. To test the significance of this putative relationship, the numbers of PASs and DRIP enriched regions were plotted per gene, per bp, across 5 kb of flanking sequence up and downstream of each RNA Pol II transcribed CDS (Figure 4.14B). This analysis revealed that DRIP-seq enrichment regions correlated strikingly well with PAS location in both WT ($r = 0.985$, $p < 0.0001$) and *Tbrh1*-/- ($r = 0.978$, $p < 0.0001$) data sets (Figure 4.14B).

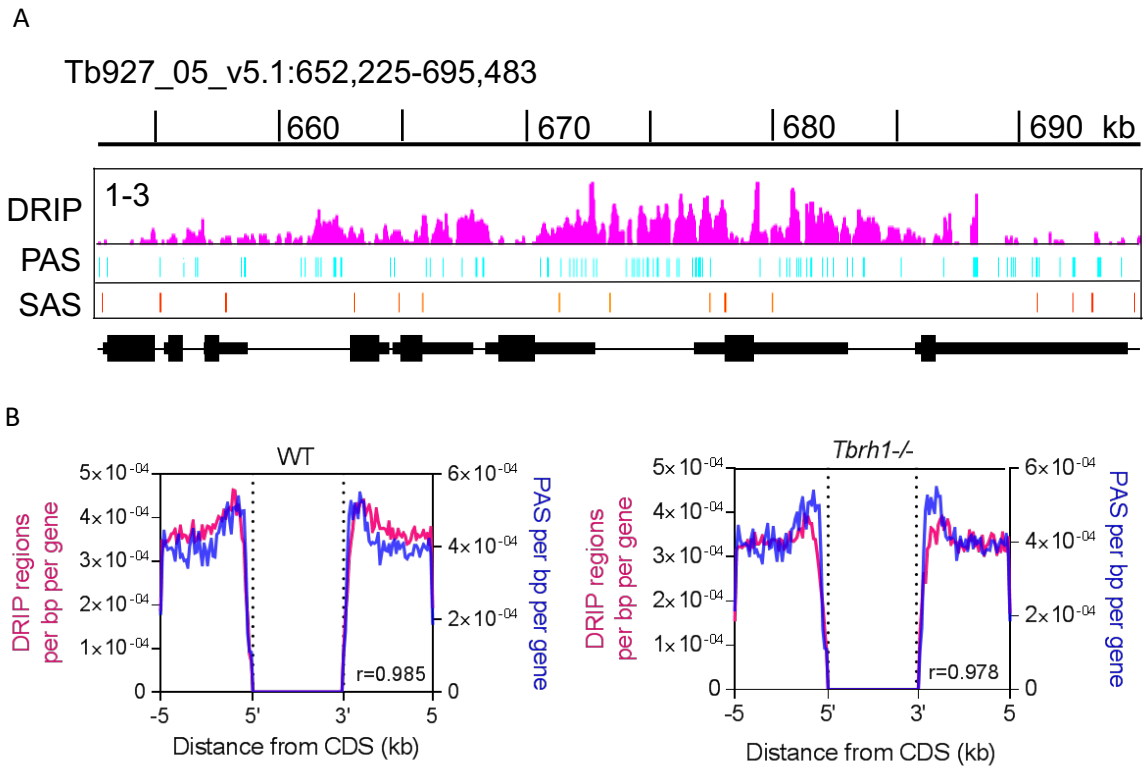



**Figure 4.14 DRIP-seq signal correlates with PASs within the PTUs.**
A) Screenshot of WT DRIP-seq signal (pink) over a region of chromosome 5. PAS (blue) and SAS (orange) locations are shown below. B) Metaplots of the number of identified DRIP enriched regions per bp per gene (pink, left axis) and the number of PAS per bp per gene (blue, right axis) over 5 kb of CDS flanking sequences.

## 4.7.2 Intra-PTU R-loops form at nucleosome depleted regions

Recently, nucleosome occupancy mapping across the RNA Pol II PTUs of *T. brucei*, via histone H3-targetting MNase-ChIP-seq, revealed that nucleosome density is decreased immediately upstream of ATG translational start sites and increases downstream, across the CDSs (Wedel *et al.*, 2017). As R-loops have been implicated in chromatin dynamics and organisation in other genomes (Al-Hadid and Yang, 2016), MNase-ChIP-seq data from Wedel et al (2017) was plotted alongside the WT DRIP-seq data generated here to assess potential correlation. Average signal was plotted from each data set over the 1 kb regions surrounding each RNA Pol II transcribed gene's ATG start site. The first gene of each PTU (n, 110) was separated from all other genes within the PTUs (n, 8,278), allowing comparison of signal found proximal to the sites of transcription initiation, and signal which is found at sites of trans-splicing and polyadenylation. A significant inverse-correlation could be seen between increased DRIP-seq signal and reduced

nucleosome occupancy at the intra-PTU genes (r = -0.8406, p <0.0001) (Figure 4.15A). Though the same pattern was seen at the first predicted genes of the PTUs, it was less clear (r = -0.1779, p = 0.0916). However, when comparing DRIP-seq signal between WT and *Tbrh1-/-* cells over the same sets of genes, a much greater increase was seen in the absence of TbRH1 at the start of the PTUs, indicating RNA-DNA hybrids do form around these genes as well as between genes within the PTU (Figure 4.15B).

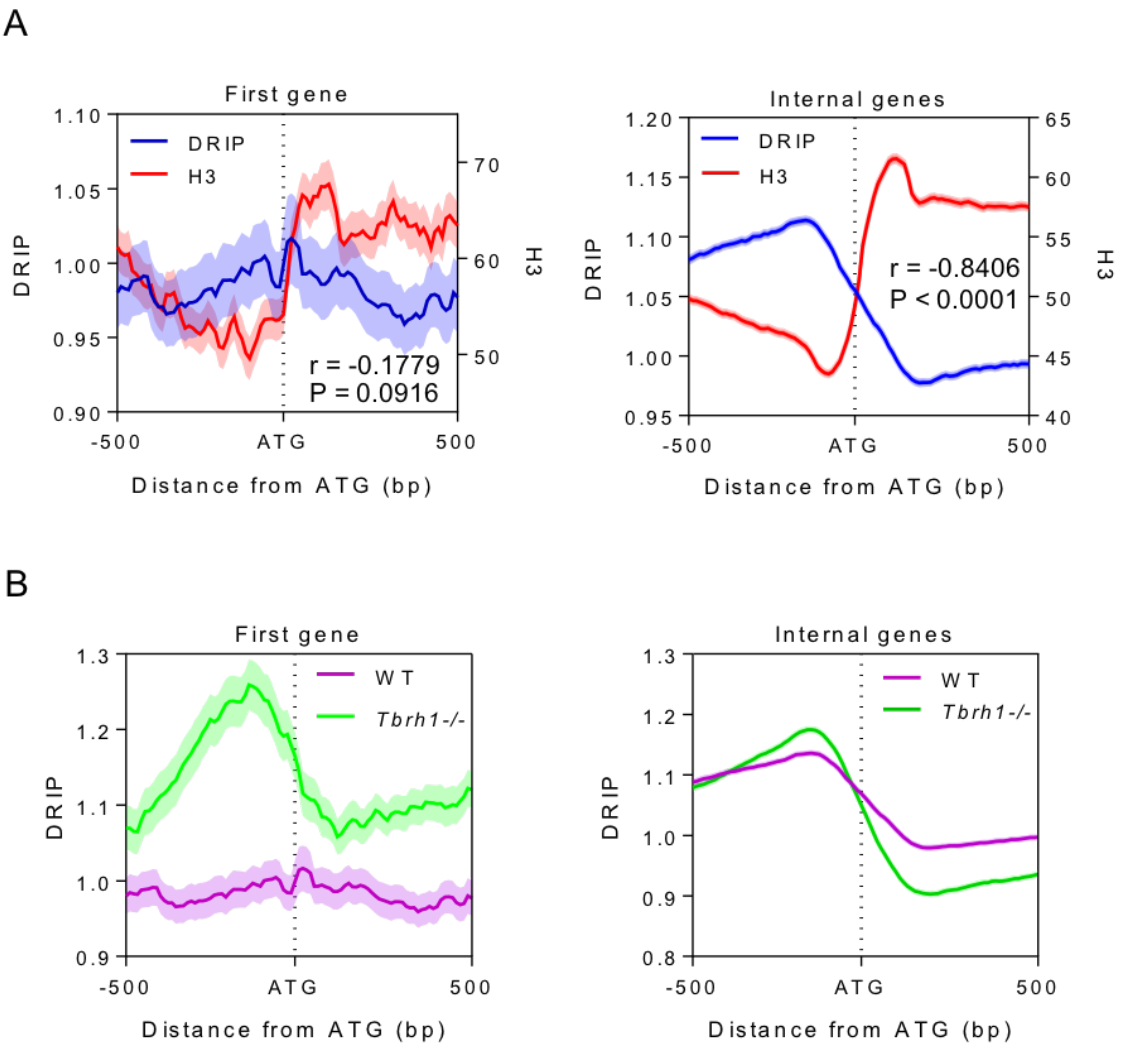



**Figure 4.15 DRIP-seq signal increase over nucleosome depleted regions upstream of ATG start sites.**
A) WT DRIP-seq signal is plotted (blue, left axis) along with histone H3 MNase-ChIP-seq average signal (red, right axis) over the 1 kb regions surrounding the ATG start sites of lead PTU genes (left, n = 110) and PTU-internal genes (right, n = 8,278). H3 MNase-ChIP-seq was previously published by Wedel et al 2017 and is plotted at counts per billion reads. B) WT (purple) and *Tbrh1-/-* (green) average DRIP-seq signal over the same sites. Standard error is shown as shading in each case.

### 4.7.3 DRIP-seq signal correlates with AT and GC negative skew

Positive GC skew has been shown to correlate with R-loop formation over some CGI promoters in the human genome (Ginno *et al.*, 2013), as has both positive GC and AT skew over coding regions of genes in the *A. thaliana* genome (Xu *et al.*, 2017). Unexpectedly, both negative GC and AT skew were found to correlate with intra-PTU DRIP-seq signal surrounding the ATG start sites of both internal (GC: r=-0.2695, p=0.0067; AT: r=-0.472, p<0.0001) and leading (GC: r=-0.5487, p<0.0001; AT: r=-0.8923, p<0.0001) protein-coding genes (Figure 4.16).

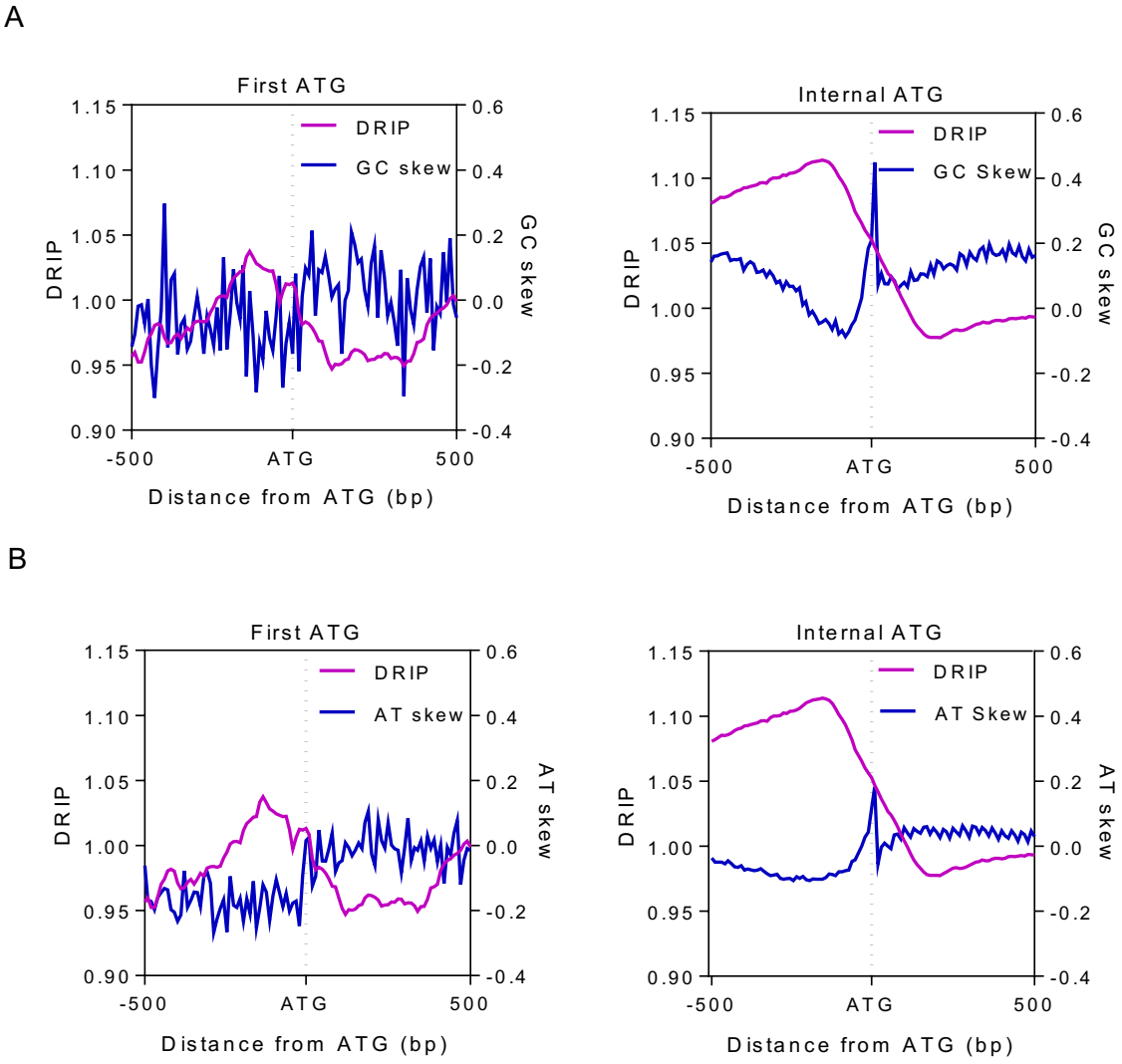



**Figure 4.16 DRIP-seq signal correlates with negative GC and AT skew over ATG start sites.** A) DRIP-seq average signal is plotted over the ATG of the first genes of each PTU (left, n = 110) and internal genes (right, n = 9977) in purple (left axis), alongside GC skew in blue (right axis). B) As in A, however AT skew is shown.

### 4.7.4 Intra-CDS R-loops

As a significant proportion of RNA Pol II PTU-associated DRIP enriched regions showed some overlap with some CDS (Figure 4.10), R-loop formation over these genes was explored further. Intra-CDS DRIP enriched regions were found in 3,224 and 3,179 genes in WT and *Tbrh1-/-* samples, respectively. Of these, 2,254 genes were found in both data sets. Various attributes of intra-CDS R-loop positive genes were compared to those of CDS that do not possess intra-CDS R-loops in any data set (n, 3,071). The WT DRIP-seq signal profile was not significantly different between these gene sets (Figure 4.17A): although DRIP-seq signal was slightly higher over the CDS of R-loop positive genes, as expected, this enrichment was far below the level of enrichment found at the flanking sites in both sets (Figure 4.17A). GC and AT skew were also compared between the two sets, but no difference could be detected between those that possess intra-CDS R-loops and those that do not (Figure 4.17B).

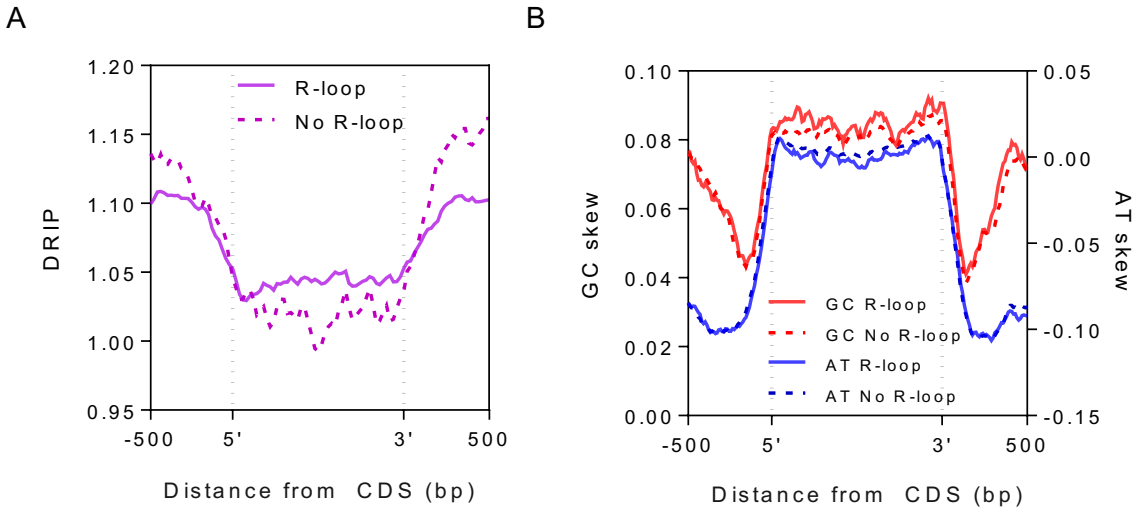



**Figure 4.17 DRIP-seq profile and GC/AT skew do not differ over genes which contain an intra-CDS R-loop and those that do not.**
A) WT DRIP-seq average signal is plotted for genes which contain an intra-CDS DRIP-enriched region (solid purple line) and a random set of genes which do not (dashed purple line). B) GC (red, left axis) and AT (blue, right axis) skew over the CDS +/- 500 bp over genes which contain intra-CDS DRIP enriched regions solid lines) and those which do not (dashed lines).

Gene function was also investigated via gene ontology (GO) term analysis. Unexpectedly, enrichment in 'binding' function terms (including ATP, nucleotide, protein and ion binding) was observed in genes that were found to form intra-CDS

R-loops in the WT sample (Table 4.1). However, the fold-enrichment between the number of genes in the data set relative to the number expected by chance averaged only 1.3 fold, with a standard deviation of 0.15. CDS length, mRNA half-file (Fadda *et al.*, 2014), 5' and 3' UTR length were all compared between genes found to contain an intra-CDS DRIP enriched region in both WT and *Tbrh1*-/- data sets, and genes which did not (Figure 4.18). In all cases no significant difference could be found.

| ID | Name | Result count | Fold enrichment | P-value | Benjamini |
|---|---|---|---|---|---|
| GO:0005488 | binding | 1509 | 1.13 | 7.88E-08 | 3.94E-06 |
| GO:0032559 | adenyl ribonucleotide binding | 397 | 1.32 | 3.16E-06 | 4.84E-05 |
| GO:0030554 | adenyl nucleotide binding | 401 | 1.32 | 3.28E-06 | 4.84E-05 |
| GO:0005524 | ATP binding | 395 | 1.32 | 3.87E-06 | 4.84E-05 |
| GO:0032555 | purine ribonucleotide binding | 453 | 1.27 | 1.42E-05 | 6.55E-05 |
| GO:0032549 | ribonucleoside binding | 452 | 1.26 | 1.55E-05 | 6.55E-05 |
| GO:0001882 | nucleoside binding | 452 | 1.26 | 1.55E-05 | 6.55E-05 |
| GO:0032550 | purine ribonucleoside binding | 452 | 1.26 | 1.55E-05 | 6.55E-05 |
| GO:0001883 | purine nucleoside binding | 452 | 1.26 | 1.55E-05 | 6.55E-05 |
| GO:0017076 | purine nucleotide binding | 457 | 1.26 | 1.59E-05 | 6.55E-05 |
| GO:0097367 | carbohydrate derivative binding | 459 | 1.26 | 1.60E-05 | 6.55E-05 |
| GO:0032553 | ribonucleotide binding | 459 | 1.26 | 1.60E-05 | 6.55E-05 |
| GO:0035639 | purine ribonucleoside triphosphate binding | 451 | 1.26 | 1.70E-05 | 6.55E-05 |
| GO:0003674 | molecular function | 2075 | 1.07 | 3.08E-05 | 0.000109916 |
| GO:0000166 | nucleotide binding | 490 | 1.23 | 6.03E-05 | 0.00018835 |
| GO:1901265 | nucleoside phosphate binding | 490 | 1.23 | 6.03E-05 | 0.00018835 |
| GO:0036094 | small molecule binding | 494 | 1.23 | 6.59E-05 | 0.000193732 |
| GO:0043168 | anion binding | 480 | 1.23 | 7.04E-05 | 0.000195516 |
| GO:0005515 | protein binding | 636 | 1.18 | 0.000103838 | 0.000273259 |
| GO:0097159 | organic cyclic compound binding | 842 | 1.14 | 0.000176392 | 0.000419981 |
| GO:1901363 | heterocyclic compound binding | 842 | 1.14 | 0.000176392 | 0.000419981 |
| GO:0043167 | ion binding | 678 | 1.16 | 0.000262957 | 0.000597629 |
| GO:0016887 | ATPase activity | 112 | 1.44 | 0.001624295 | 0.003531077 |
| GO:0042623 | ATPase activity, coupled | 87 | 1.49 | 0.002549561 | 0.005311585 |
| GO:0016462 | pyrophosphatase activity | 200 | 1.26 | 0.004109398 | 0.007902688 |
| GO:0016818 | hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides | 200 | 1.26 | 0.004109398 | 0.007902688 |
| GO:0017111 | nucleoside-triphosphatase activity | 195 | 1.27 | 0.004382276 | 0.008115326 |

| | | | | | |
|---|---|---|---|---|---|
| GO:0016817 | hydrolase activity, acting on acid anhydrides | 201 | 1.26 | 0.004966523 | 0.008868791 |
| GO:0008026 | ATP-dependent helicase activity | 43 | 1.66 | 0.008623755 | 0.014372925 |
| GO:0070035 | purine NTP-dependent helicase activity | 43 | 1.66 | 0.008623755 | 0.014372925 |
| GO:0016787 | hydrolase activity | 421 | 1.14 | 0.009754508 | 0.015733077 |
| GO:0016773 | phosphotransferase activity, alcohol group as acceptor | 151 | 1.25 | 0.015078862 | 0.023560721 |
| GO:0004386 | helicase activity | 51 | 1.49 | 0.018276414 | 0.027235004 |
| GO:0004672 | protein kinase activity | 122 | 1.27 | 0.018519803 | 0.027235004 |
| GO:0016301 | kinase activity | 161 | 1.23 | 0.020322907 | 0.028621349 |
| GO:0004016 | adenylate cyclase activity | 41 | 1.56 | 0.020607371 | 0.028621349 |
| GO:0003729 | mRNA binding | 110 | 1.28 | 0.022840644 | 0.030393689 |
| GO:0003774 | motor activity | 59 | 1.42 | 0.023099204 | 0.030393689 |
| GO:0004674 | protein serine/threonine kinase activity | 75 | 1.35 | 0.024219967 | 0.030838196 |
| GO:0009975 | cyclase activity | 41 | 1.53 | 0.024670556 | 0.030838196 |
| GO:0003723 | RNA binding | 249 | 1.16 | 0.027699753 | 0.032811201 |
| GO:0044822 | poly(A) RNA binding | 110 | 1.27 | 0.027812196 | 0.032811201 |
| GO:0022804 | active transmembrane transporter activity | 69 | 1.36 | 0.028217632 | 0.032811201 |
| GO:0008094 | DNA-dependent ATPase activity | 21 | 1.82 | 0.031811929 | 0.035760812 |
| GO:0003777 | microtubule motor activity | 55 | 1.4 | 0.032184731 | 0.035760812 |
| GO:0003676 | nucleic acid binding | 436 | 1.11 | 0.033776143 | 0.036562734 |
| GO:0016772 | transferase activity, transferring phosphorus-containing groups | 228 | 1.16 | 0.03436897 | 0.036562734 |
| GO:0005215 | transporter activity | 138 | 1.21 | 0.036986137 | 0.038527226 |
| GO:0008017 | microtubule binding | 44 | 1.42 | 0.044181551 | 0.045083215 |
| GO:0003824 | catalytic activity | 1000 | 1.06 | 0.048957109 | 0.048957109 |

**Table 4.1 Genes with intra-CDS R-loops may be associated with binding activities.**
Table shows results of molecular function GO term analysis. All hit GO terms with a P value > 0.05 are shown with number of genes found (result count), enrichment above the number of genes expected by change (fold-change) P value and Benjamini value.

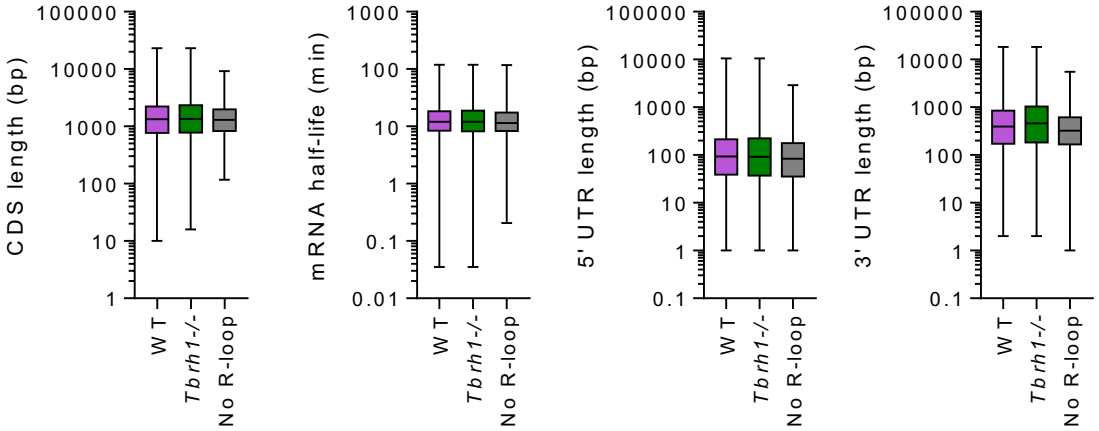



**Figure 4.18 Intra-CDS R-loop formation does not correlate with CDS length, mRNA half-life or UTR length.**
CDS length, mRNA half-life (Fadda *et al.*, 2014), 5' UTR length, and 3' UTR length (from left to right) are compared between genes which were found to have an intra-CDS DRIP enriched region in WT (pink) and *Tbrh1-/-* (green) data sets, and those genes which did not in any data set (grey).

## 4.8 R-loops are not associated with DNA replication origins

As discussed in detail above (section 1.4.1.3), R-loops are known to partake in the initiating steps of DNA replication of bacteria plasmids (Gowrishankar, Leela and Anupama, 2013), the *E. coli* chromosome (Asai and Kogoma, 1994) and mtDNA (Lee and Clayton, 1998). Additionally, R-loop formation is enriched over CGI promoters that also act as ORIs in the human genome (Ginno *et al.*, 2012; Lombraña *et al.*, 2015). Hence, any potential relationship between ORIs and R-loops in the *T. brucei* genome was investigated. Virtually all SSRs have been shown to have OCR1/CDC6 binding potential in *T. brucei*, yet only a proportion of these ORC-bound sites were found to be active ORIs in early S phase (Tiengwe *et al.*, 2012). SSRs were therefore divided according to whether they were found to be active ORIs or not, and average DRIP-seq signal was plotted over each class of SSR (Figure 4.19). Notably, in both cases DRIP-seq signal was reduced across the SSR body but increased in the flanking regions, where the immediately upstream and downstream genes are located. This pattern of R-loop formation was not significantly different between the ORI and non ORI SSRs; in addition, DRIP-seq signal increase after deletion of TbRH1 was seen to similar extents in both cases (Figure 4.19).

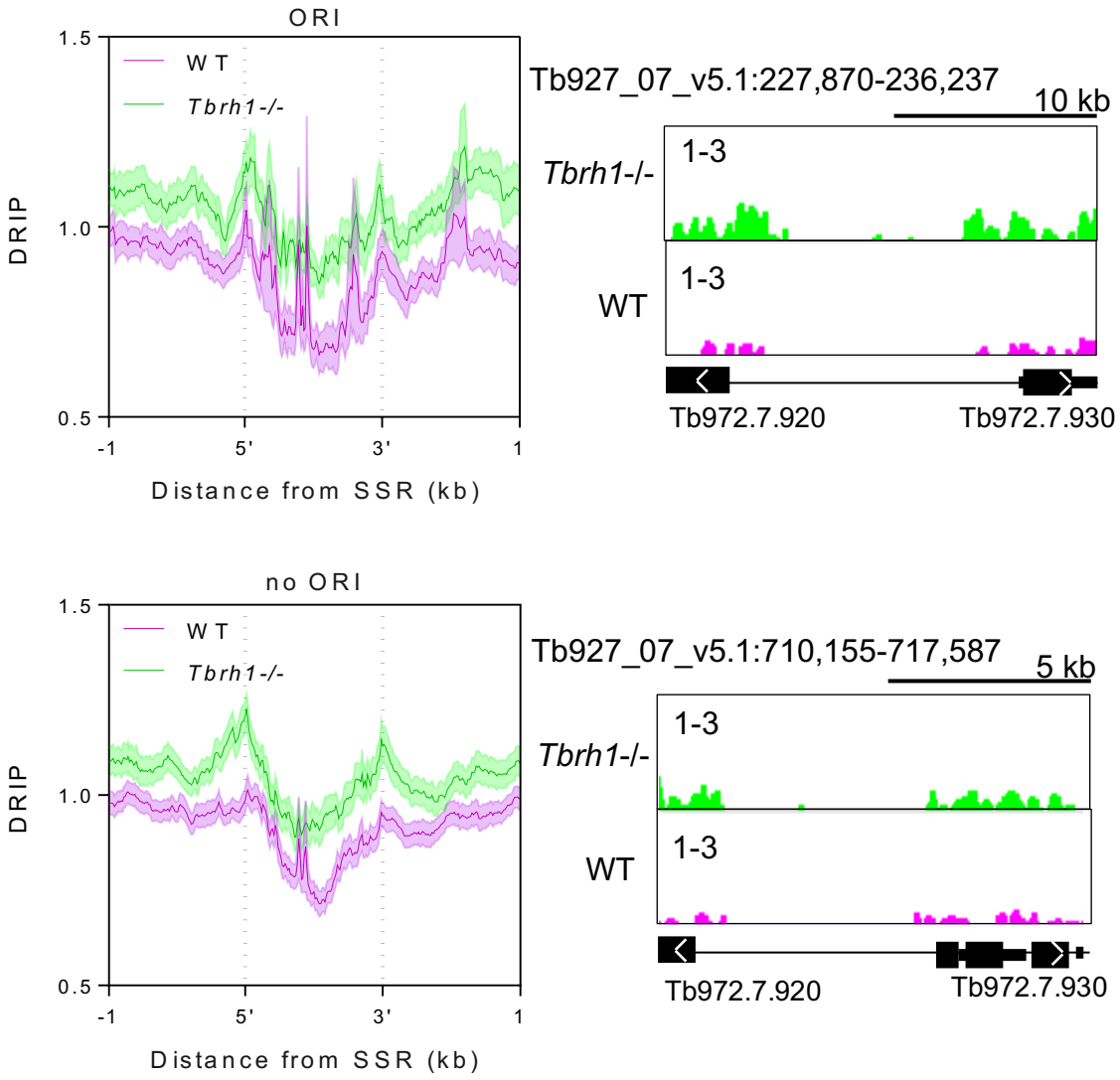



**Figure 4.19 R-loops are not associated with ORIs in *T. brucei*.**
WT (pink) and *Tbrh1-/-* (green) DRIP-seq average signal is plotted across SSRs which are also ORIs (upper left) and those which are not (lower left). Standard error is plotted as shaded regions. Screenshots of DRIP signal across two example SSRs; upper is an ORI and lower is not.

## 4.9 R-loops are associated with RNA Pol II transcription initiation

R-loop formation has also been associated with sites of transcriptional initiation and termination (see sections 1.4.1.1 and 1.4.1.2 ). Hence, DRIP-seq signal was re-plotted over SSRs, in this case by classifying the SSRs into three groups according to their transcriptional, rather than replicative, status: (i) divergent SSRs where transcription initiates in both sense and antisense directions; (ii) convergent SSRs where transcription from both directions terminates; and (iii) head-to-tail SSRs

where transcription terminates then re-initiates on the same strand. A clear association was evident between DRIP-seq signal enrichment and transcription initiation sites, at both divergent SSRs and in the initiation region of head-to-tail SSRs (Figure 4.20). Conversely, DRIP-seq signal was not consistently enriched at sites of transcription termination, either in convergent or head-to-tail SSRs (Figure 4.20). Average DRIP-seq signal did display two peaks of signal within the convergent SSR group, but this was due to the presence of tRNA and snRNA genes, which are known sites of R-loop formation in yeast (El Hage et al. 2014) and where DRIP-seq enrichment was seen in *T. brucei* (see section 4.5). Similar spikes in signal due to RNA-coding genes were also seen in the head-to-tail region average DRIP-seq profile. Besides these signal peaks, DRIP-seq signal remained relatively flat across the convergent SSRs and the termination portion of head-to-tail regions. DRIP-seq signal increased when TbRH1 is lost, but this appeared to be uniform across each class of SSRs, implying R-loops cannot be linked to termination processes in this study.

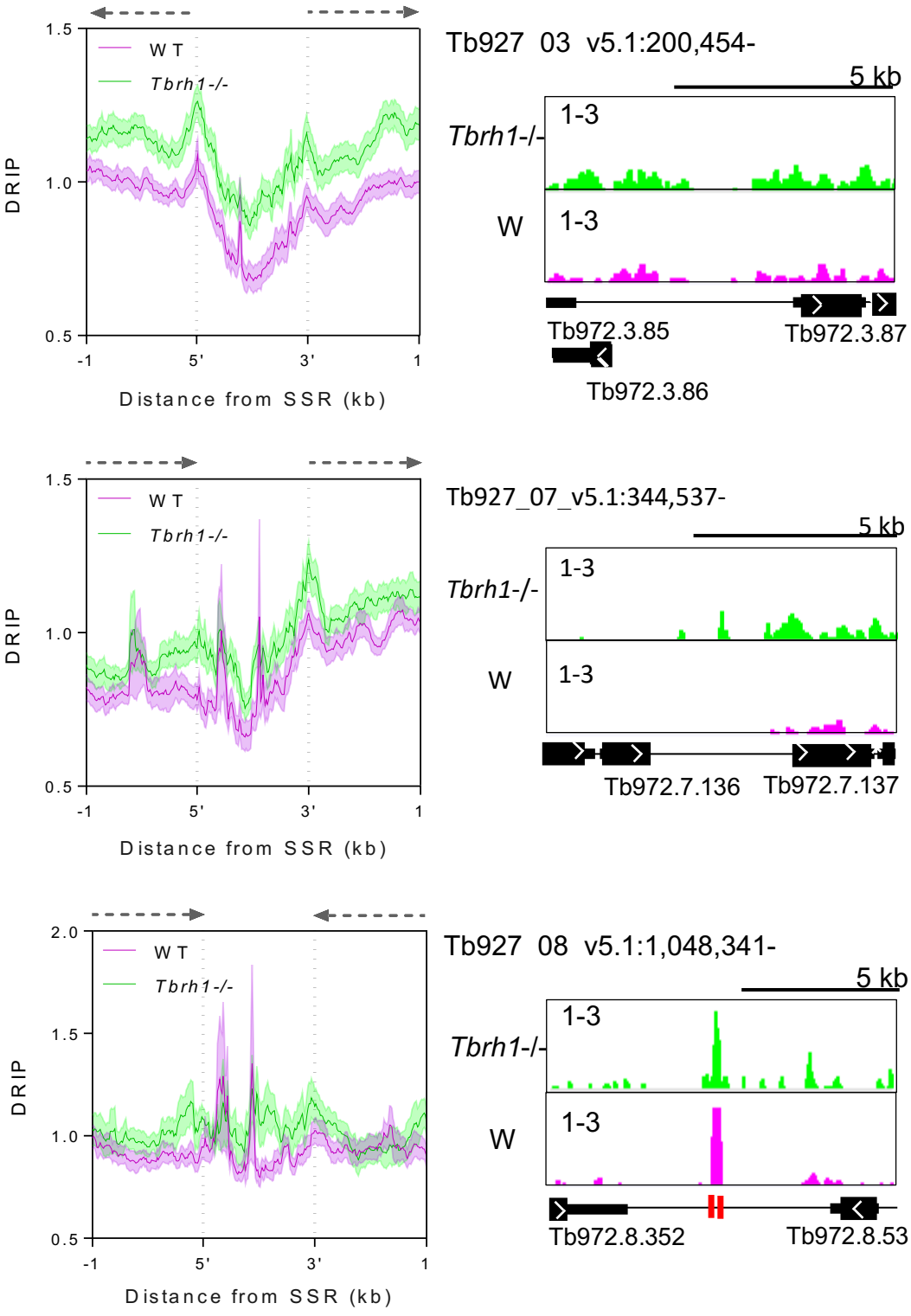



**Figure 4.20 DRIP-seq signal enrichment is associated with sites of transcription initiation.** Average DRIP-seq signal is plotted for WT (pink) and *Tbrh1-/-* (green) data sets over divergent (upper), head-to-tail (centre) and convergent (lower) SSRs. In all cases 5' and 3' denote SSR boundaries defined by flanking transcript coordinates. Transcription direction is shown by dashed arrows. Standard error is shown as shaded regions. A DRIP-seq signal screenshot example of each SSR class is shown to the right of each metaplot; CDS (thick black), UTR (thin black) and snRNA/tRNA genes (red) are shown below the DRIP-seq tracks.

### 4.9.1 DRIP-seq signal correlates with transcription initiation-associated epigenetic markers

To further investigate the association between R-loops and transcription initiation, DRIP-seq signal was again plotted around each class of SSR, this time extending analysis to 8 kb upstream and downstream of the SSR boundaries. WT DRIP-seq signal over these regions was compared to previously published ChIP-seq data sets of four factors that are associated with transcription initiation: histone variants H2B.V and H2A.Z, histone H4 acetylation of lysine 10 (H4K10ac), and bromodomain factor BDF3 (Siegel *et al.*, 2009). Comparison of DRIP-seq signal showed strong correlation with three of the epigenetic markers, H2A.Z, H2B.V and H4K10ac (Figure 4.21). Both DRIP-seq signal and the ChIP-seq signal for all three of these factors peaked slightly downstream of divergent SSR 5' and 3' boundaries, although ChIP-seq signal of each factor diminished further upstream than that of the DRIP-seq signal. This effect is presumably due to inter-PTU R-loop formation across the RNA Pol II transcribed units, maintaining DRIP-seq signal. BDF3 ChIP-seq signal instead peaked directly at the boundaries of the divergent SSRs, and so did not correlate as well with DRIP-seq signal, which peaked slightly downstream of the highest point of BDF3 signal. In all three cases, similarly correlated enrichment was also apparent at the transcription initiation portion of head-to-tail SSRs. Within convergent SSRs, as well as at the termination portion of heat-to-tail regions, ChIP-seq signal of all four epigenetic factors was diminished. Although DRIP-seq signal was also lower in these termination regions compared to initiation sites, it did not reduce completely, again presumably due to the presence of inter-PTU R-loops (Figure 4.21).

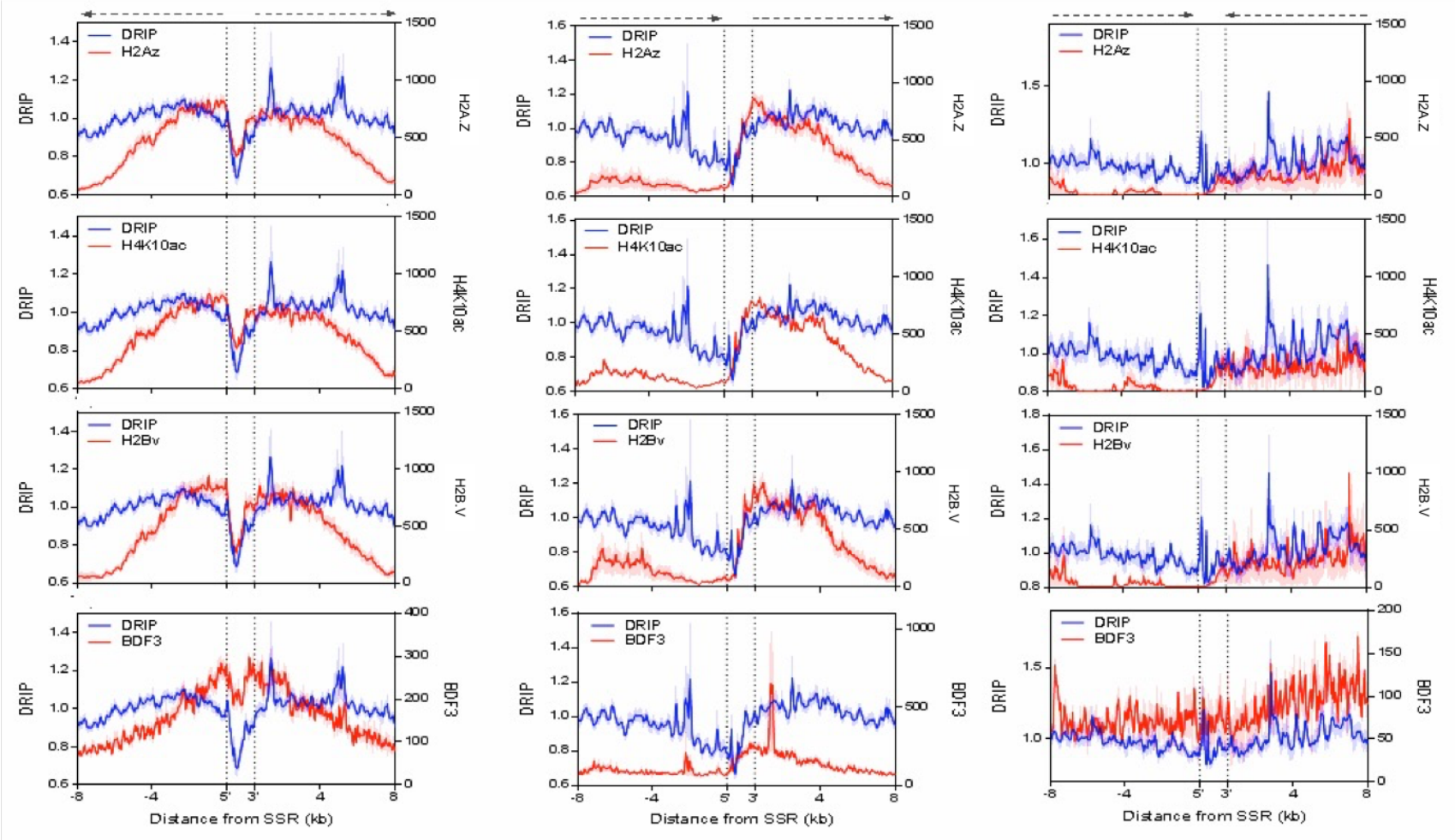



**Figure 4.21 R-loop formation correlates with epigenetic factors associated with transcription initiation.**
WT DRIP-seq signal average (blue) is plotted over each class of SSRs, divergent (left), head-to-tail (centre) and convergent (right). Average ChIP-seq signal for histone H2A.Z, H4K10ac, histone H2B.V and bromodomain protein BDF3 are plotted for comparison in red (Siegel *et al.*, 2009). In all cases standard error is plotted as shaded regions around the mean.

### 4.9.2 DRIP-seq signal does not associate with transcription termination epigenetic markers

Two other histone variants, H3.V and H4.V, have previously been associated with transcription termination (Siegel *et al.*, 2009). Published H3.V and H4.V ChIP-seq signal was therefore plotted alongside WT DRIP-seq signal, as above, to allow comparison. In this case no correlation could be observed between the DRIP-seq signal and ChIP-seq signal of either histone variant (Figure 4.22). ChIP-seq signal of each histone variant was significantly greater than that of DRIP-seq at both convergent SSRs and the terminal portions of head-to-tail-regions, whereas DRIP-seq signal exceeded both at sites of transcription initiation (Figure 4.22).

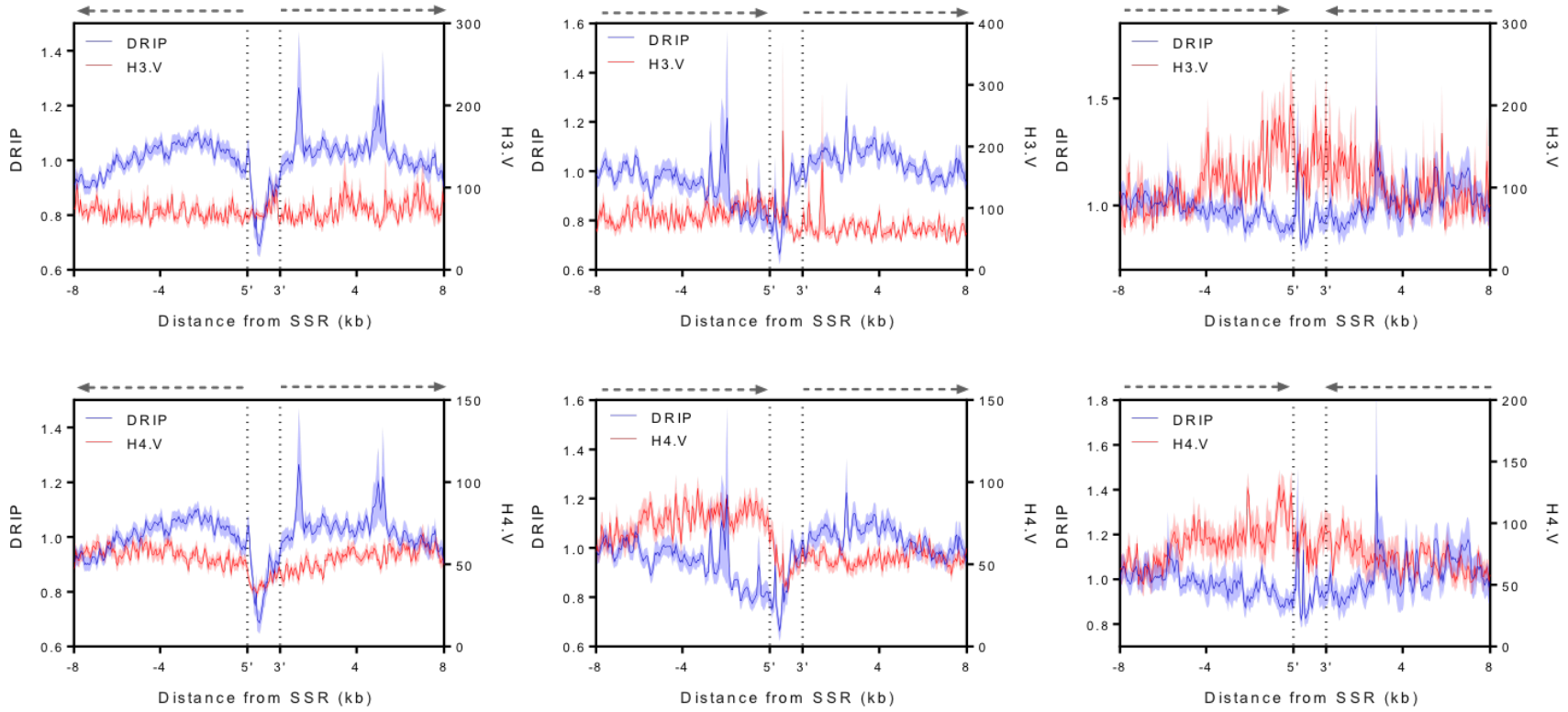



**Figure 4.22 R-loop formation does not correlate with histone variants associated with transcription termination.**
WT DRIP-seq signal average (blue) is plotted over each class of SSRs, divergent (left), head-to-tail (centre) and convergent (right). Average ChIP-seq signal for histone variants H3.V and H4.V are plotted for comparison (red). In all cases standard error is plotted as shaded regions around the mean.

## 4.10 Discussion

R-loop mapping has been carried out in a range of species, revealing widespread roles in DNA replication, transcription initiation and termination, telomere maintenance and recombination events (see section 1.4.1). Although few cases have demonstrated direct mechanistic links (Skourti-Stathaki, Kamieniarz-Gdula and Proudfoot, 2014; Boque-Sastre *et al.*, 2015), several studies have correlated R-loop formation to various epigenetic markers and nucleosome positioning dynamics (Santos-Pereira and Aguilera, 2015b; Al-Hadid and Yang, 2016; Sanz *et al.*, 2016). In this chapter, R-loops have been mapped in the unconventional genome of *T. brucei* to find sites of conserved R-loop formation, and reveal evidence linking R-loops to transcription initiation, *trans*-splicing and polyadenylation, but not to transcription termination or DNA replication.

### 4.10.1 Conserved sites of R-loop formation

#### 4.10.1.1 Centromeres

Repeat sequences are well-established sites of R-loop formation, many of which appear conserved in the *T. brucei* genome. Centromeric repeats are prominent sites of R-loop formation in *T. brucei* WT parasites and even more so in cells lacking TbRH1; hence, TbRH1 acts to resolve these R-loops. In other systems, binding of ncRNAs generates functional RNA-DNA hybrids at centromeres (see section 1.4.1.6, Nakama *et al.*, 2012). In *T. brucei*, siRNAs may act as the RNA component of the R-loops in the three CIR147 bp repeat-containing centromeres at which siRNAs have been detected (Tschudi *et al.*, 2012). These siRNAs are known to function in the RNAi pathway (Patrick *et al.*, 2009) and may therefore act to induce chromatin condensation at these centromeres via a similar ncRNA-dependent pathway to that documented in yeast (Nakama *et al.*, 2012; Castellano-Pozo *et al.*, 2013). However, CIR147 bp repeats are not present in the remaining identified centromeres in *T. brucei*, yet these different repeats all display enriched DRIP-seq signal. It seems plausible, therefore, that R-loop formation in centromeres is a conserved function of centromere-associated repeats across species, perhaps ensuring condensed chromatin (Castellano-Pozo *et al.*, 2013) to direct correct chromosome segregation

(Kabeche *et al.*, 2018). Nonetheless, how this putative conservation in centromere activity correlates with the pronounced divergence in the kinetoplastid kinetochore is unclear (Akiyoshi and Gull, 2014; D'Archivio and Wickstead, 2017). In addition, centromere-associated repeats have not been described at sites of kinetochore factor KKT1 enrichment in *Leishmania* (Garcia-Silva *et al.*, 2017); mapping R-loops may clarify if these loci are true centromeres, since no such activity has been demonstrated to date.

#### 4.10.1.2 rRNA loci

R-loops are also strongly apparent over the *T. brucei* rRNA loci, as has been reported for both eukaryotes (Gall and Pardue, 1969; Thomas, White and Davis, 1976; White and Hogness, 1977; El Hage *et al.*, 2010; Arana *et al.*, 2012) and bacteria (Hraiky, Raymond and Drolet, 2000). In yeast, antisense transcription generates ncRNAs, which form RNA-DNA hybrids at the rRNA RNA Pol I promoter to induce transcriptional silencing of the rRNA via DNA methylation and repressive histone marks (Mayer *et al.*, 2006; Bierhoff *et al.*, 2010; Schmitz *et al.*, 2010). Indeed, small RNAs have been detected from the rRNA loci in *T. brucei* (Zheng *et al.*, 2013) and DRIP-seq signal enrichment is evident over the RNA Pol I transcribed rRNA arrays. These findings highlight the intriguing possibility that R-loops may negatively regulate rRNA transcription in *T. brucei*. However, whether *T. brucei* rRNA-associated R-loops form co-transcriptionally or via antisense small RNAs cannot be extrapolated from this study.

#### 4.10.1.3 tRNA and snRNA

Nascent tRNAs have been shown to engage in R-loop formation in yeast, as well as the snRNAs that are also transcribed by RNA Pol III (Chan *et al.*, 2014; El Hage *et al.*, 2014). DRIP-seq enrichment is observed at tRNA and snRNA genes in WT *T. brucei* parasites, although this signal appears slightly diminished in *Tbrh1-/-* mapping. DRIP-qPCR, however, revealed increased detection of RNA-DNA hybrids after loss of TbRH1, despite this not being reflected in DRIP-seq mapping of the hybrids. As tRNA and snRNA R-loops have previously been demonstrated to be targeted by RNase H enzymes in yeast (El Hage *et al.*, 2014), it is hypothesised that

TbRH1 is also able to target such R-loops, as predicted by DRIP-qPCR. Indeed, it is possible that scaling of input and IP DNA libraries to remove bias introduced by amplification steps of library preparation and sequencing (section 2.8.4), has resulted in the reduction of signal enrichment at tRNA and snRNA genes in the *Tbrh1-/-* sample. In order to confidently compare signal between WT and *Tbrh1-/-* mapping, DRIP-seq would require repeating with the addition of spike-in controls to allow scaling to an exogenous factor (K. Chen *et al.*, 2015).

#### 4.10.1.4 RHS-associated genes

DRIP enrichment is also apparent at RHS-associated genes, and dramatically more so after loss of TbRH1. These genes are highly abundant, variable, are largely expressed in the nucleus and are targets for transposable elements (Bringaud *et al.*, 2002), although their exact function(s) is uncertain. Whether the R-loops detected in the RHS arrays form co-transcriptionally or via binding of siRNAs, which have been detected at RHS-associated sequences (Tschudi *et al.*, 2012), is not clear. RNA-DNA hybrids are associated with retrotransposable elements in the yeast genome, where they most often form with cDNAs produced by reverse transcription, rather than chromosomal genes (El Hage *et al.*, 2014). Transposable elements in *A. thaliana* genes are also sites of R-loop formation, where they have been associated with the heterochromatin markers, H3K27me1 and H3K9me2 (Xu *et al.*, 2017). Although R-loop formation over the *T. brucei* RHS genes may relate to those observed at transposable elements in yeast (El Hage *et al.*, 2014) and plants (Xu *et al.*, 2017), the unique clustering nature of these genes in subtelomeric arrays may indicate a kinetoplastid-specific function of R-loops here.

#### 4.10.1.5 Telomeres

As discussed in section 1.4.1.5, RNA-DNA hybrids that form in telomeres via the binding of TERRA RNA have been implicating in ATL (Arora *et al.*, 2014; Yu, Kao and Lin, 2014). Examination of mVSG ESs, where mVSGs are located directly upstream of the telomeric repeats, shows formation of R-loops in both WT and *Tbrh1-/-* data sets. Moreover, chapter 6 describes R-loops at the telomeres of BESs. It is plausible that telomeric R-loops may play a role in telomere lengthening in *T. brucei* if

replication of these intrinsically difficult to replicate repeats fails (Özer and Hickson, 2018), and it has also been suggested that they are involved in VSG switching (Nanavaty *et al.*, 2017). Indeed, clashes between the replication machinery and telomeric R-loops are hypothesised to trigger homology-directed repair via recombination in ATL (Graf *et al.*, 2017), a proposal which could be extrapolated as a model for inducing recombination-based VSG switching (Nanavaty *et al.*, 2017). This possibility is discussed in the context of DRIP-seq mapping to the BESs in detail in Chapter 6.

### 4.10.2 R-loops are associated with regions of *trans*-splicing and polyadenylation

The most prominent site of R-loop formation in both WT and *Tbrh1*-/- parasites is within the RNA Pol II transcribed PTUs. Here, R-loops were most extensively linked to the intergenic and UTR sequences in virtually all genes. Although several DRIP enriched regions were also found to associate with CDS regions, the enrichment here is minor compared to that of the intra-CDS signal and is not associated with AT or GC skew, UTR length, CDS length, mRNA half-life or gene function. This contrasts with the *A. thaliana* genome, where R-loops are enriched across the coding regions of a subset of genes and are associated with both AT and GC positive skew (Xu *et al.*, 2017). Intra-CDS R-loops have also been documented in the *S. cerevisiae* genome, where they are most strongly associated with highly expressed genes (Wahba *et al.*, 2016). As transcription rate is thought to be largely uniform across the RNA Pol II transcribed genes in *T. brucei*, due to the nature of multigenic transcription, it is unlikely that transcription rates are linked to levels of R-loop formation at different genes.

The formation of R-loops predominantly in the intra-CDS regions does, however, suggest a strong association between these structures and coupled *trans*-splicing and polyadenylation events, reflecting the nature of gene expression organisation in the *T. brucei* genome. Indeed, meta-analysis of DRIP enriched region and PAS occurrence across the CDS flanking sequences, reveals a significant relationship between the two: the density of PAS is positively correlated with R-loop formation in these regions. How R-loops may act in the process of polyadenylation remains

unclear, however. It is possible that increased RNA Pol II pausing may be evident at sites with many PASs, increasing the likelihood of R-loop formation. It is also plausible that the increased number of A bases in 3' UTRs with multiple PASs increases the rate of R-loop formation, or stability of the structure, by influencing the DNA sequence composition.

Motif analysis supports the relationship between R-loops and mRNA processing in *T. brucei*, since R-loop forming sequences are enriched for two poly(Y) tract motifs, which are vital for both splicing and polyadenylation (Matthews, Tschudi and Ullu, 1994). Although, it should be noted that predominant R-loop formation over intra-CDS regions necessitates enrichment of RNA processing-associated poly(Y) tracts and does not necessarily indicate these motifs are directly causal of R-loop formation. However, a third, poly(A) motif is also enriched in R-loop-forming sequences within the PTUs. Poly(A) tracts >21 bp are associated with R-loop formation in the genome of yeast mutants lacking both RNase H1 and H2 activity, along with positive AT skew (Wahba *et al.*, 2016). In this setting the poly(A) tracts have been directly linked to R-loop formation, since deletion of the tracts led to a decrease in R-loops, often to background levels (Wahba *et al.*, 2016). In the *T. brucei* genome poly(A) tracts, but not positive AT skew, are linked to DRIP enrichment, suggesting the possibility that poly(A) tracts themselves are sufficient for R-loop formation, independent of total base skew. However, to date no functional analysis of poly(A) tracts in *T. brucei* has been reported.

DRIP enrichment in the intra-CDS regions is also significantly associated with nucleosome depletion and, thus, with relaxed chromatin. Similar R-loop enrichment in nucleosome depleted regions has been noted in the human genome irrespective of genic location (Sanz *et al.*, 2016; Chen *et al.*, 2017). R-loops are also heavily associated with euchromatin epigenetic markers, particularly at promoters (Sanz *et al.*, 2016; Chen *et al.*, 2017), where they have been shown to directly protect against DNA methylation, a known marker of heterochromatin (Ginno *et al.*, 2012), and chromatin condensation (Boque-Sastre *et al.*, 2015). These findings highlight a potential role of R-loops in regulating nucleosome positioning between coding regions within the *T. brucei* PTUs. However, RNA Pol II is likely to pause at sites of

nucleosome occupancy (Wedel *et al.*, 2017), which may allow more opportunity for RNA to bind the DNA template. Hence, R-loop enrichment may reflect transcription dynamics in *T. brucei* as RNA Pol II pauses at organised nucleosomes positioned over the ATG translational start sites, potentially allowing R-loop formation upstream.

In either case, positioning of nucleosomes 5' of exons has been suggested to promote co-transcriptional *cis*-splicing in eukaryotes by slowing the transcribing polymerase (Schwartz and Ast, 2010). *Cis*- and *trans*-splicing share many mechanistic similarities (Michaeli, 2011) and so it has been suggested that nucleosome positioning also plays a role in *trans*-splicing via a similar mechanism in the *T. brucei* genome, where each CDS is effectively an exon (Wedel *et al.*, 2017). Therefore, R-loops may act to slow the transcribing polymerase, rather than paused polymerases allowing R-loop formation. However, R-loops have so far shown poor correspondence with sites of *cis*-splicing. In yeast, introns have been found to protect genes from R-loop formation (Bonnet *et al.*, 2017), an effect that also seems to be the case in the intron-rich mammalian genomes (Bonnet *et al.*, 2017), where the first exon-intron junction acts as the 3' boundary for promoter-associated R-loops (Dumelie and Jaffrey, 2017). In these genomes, the action of the mRNA cleavage and polyadenylation machinery (Stirling *et al.*, 2012), or spliceosome recruitment alone (Bonnet *et al.*, 2017), appears to protect against R-loop formation. DRIP-seq has shown this is not the case in *T. brucei*, where sites associated with the action of such factors show increased R-loop formation, rather than decreased. Hence, the high levels of intra-CDS R-loops observed in the *T. brucei* genome may indicate a novel role for R-loops in pre-mRNA processing during polycistronic transcription, perhaps via maintaining open chromatin to allow efficient binding of splicing and polyadenylation factors. It is also clear that these R-loops are sensitive to RNase H activity as they showed increased enrichment in *Tbrh1*-/- parasites. Hence, RNase H detection and/or processing of intra-CDS R-loops may also have a role in such processes.

### 4.10.3 R-loops are associated with transcription initiation

As discussed above, R-loops are heavily associated with active transcription at promoters in mammals (Ginno *et al.*, 2012; Boque-Sastre *et al.*, 2015; P. B. Chen *et al.*, 2015; Sanz *et al.*, 2016), *A. thaliana* (Xu *et al.*, 2017), and (less clearly) in yeast (Chan *et al.*, 2014; El Hage *et al.*, 2014). Mapping of RNA-DNA hybrids to the SSRs in *T. brucei* shows R-loops are also associated with transcription initiation in this organism. In metazoan genomes two broad classes of transcriptional promoter have been described: "focused", where transcription initiates from a discreet TSS and is associated with a well-defined promoter sequence (e.g. a TATAA box); and "dispersed" promoters, which initiate transcription from multiple TSSs in a broad region of ~50-100 bp (Lenhard, Sandelin and Carninci, 2012). RNA Pol II transcription initiation in *T. brucei* lacks discreet TSSs and typical promoter sequences, so are likely to be mechanistically more similar to the dispersed promoter class (Kolev *et al.*, 2010). This class lacks defined DNA sequences motifs, which in focused promoters provide sites for the binding of proteins such as transcription factors to allow transcriptional regulation of gene expression (Lenhard, Sandelin and Carninci, 2012). In dispersed promoters open chromatin alone is thought to facilitate transcription initiation of these broadly expressed genes and so negate the need for DNA motifs (Lenhard, Sandelin and Carninci, 2012). As gene expression of RNA Pol II transcribed genes is independent of transcriptional regulation in *T. brucei*, the action of dispersed promoters to simply provide an environment for transcription initiation is in keeping with the nature of the kinetoplastid's genome. Interestingly, dispersed promoters often display high GC content and overlap with CGI promoters, which are prominent sites of R-loop formation (Ginno *et al.*, 2012). At these sites R-loops protect from DNA methylation (Ginno *et al.*, 2012) and are associated with several epigenetic markers and histone variants associated with open chromatin (Sanz *et al.*, 2016; Chen *et al.*, 2017). Therefore, it is hypothesised that R-loops have a similar role in promoting an open chromatin environment favourable for active transcription in *T. brucei*.

In *T. brucei* promoters, a GT-rich motif is capable of conferring RNA Pol II transcription initiation (Wedel *et al.*, 2017). This promoter element contributes in

some capacity to the deposition of histone variant H2A.Z, which is associated with transcription initiation sites (Siegel *et al.*, 2009), to the boundaries of the nucleosome depleted region (Wedel *et al.*, 2017). Supporting the hypothesised role of R-loops and the chromatin environment, DRIP signal is shown to correlate with H2A.Z ChIP-seq signal at the initiation sites, along with other activation markers (histone H2B.V, H4K10ac and, in a more minor sense, bromodomain factor BDF3). This association extends several kb into the PTUs and highlights a role for R-loops in the maintenance of active transcription rather than dynamic gene expression regulation via transcription, which *T. brucei* lacks.

In contrast, no specific R-loop enrichment was associated with sites of transcription termination in either WT or *Tbrh1*-/- parasites, nor with termination associated histone variants H3.V and H4.V. Hence, R-loops were not associated with termination of RNA Pol II in this study, although it remains plausible that another factor, such as a helicase, targets termination associated R-loops in *T. brucei*.

### 4.10.4 Perspectives

Over the last decade, research into the physiological role of R-loops has expanded considerably leading to their implication in many processes, as well as human diseases (Groh and Gromak, 2014; Costantino and Koshland, 2015; Richard and Manley, 2017). Here, R-loop formation in the genome of *T. brucei* has added to understanding of R-loop biology and has linked them to functions in this kineotoplastid genome. DRIP enrichment is evident at rRNA, tRNA and snRNA genes, as well as telomeric and centromeric regions, highlighting the apparent conservation of R-loop formation at these sites. Whether R-loops have functional roles at these sites, and what these may be, remains an interesting avenue of investigation. However, it may be speculated that a functional role this is more probable at some regions (telomeres, centromeres and perhaps rRNA) than others (tRNA and sRNA).

DRIP-seq enrichment analysis in *T. brucei* has additionally been able to tentatively rule out a eukaryotic-wide conserved role for R-loops in typical DNA replication and RNA pol II transcription termination. A lack of specific R-loop enrichment at ORIs in

WT and *Tbrh1*-/- parasites makes it unlikely that R-loop formation has a role in the initiation of DNA replication at ORIs, as suggested in mammals (Lombraña *et al.*, 2015). Although virtually all SSRs have been shown to bind ORC (Tiengwe *et al.*, 2012), a lack of DRIP signal across the bodies of the SSRs and enrichment at the flanks highlights transcription is much more closely linked to R-loop formation than DNA replication in *T. brucei*. Similarly, a lack of specific DRIP signal at sites of transcription termini undermines any potential role for R-loops in these regions.

Analysis of R-loop mapping does, however, provide strong evidence linking R-loops to sites of mRNA processing in the PTUs. R-loop formation between coding regions has not been widely documented in other studies; in fact, introns appear to protect genes from R-loop formation in other eukaryotes. Whether R-loops result from transcriptional pausing or are actively involved in the mRNA maturation process, perhaps by signalling for machinery binding, remains to be investigated. It is also possible that R-loops regulate a chromatin environment permissible for such processes, as they are formed in nucleosome-depleted regions. A similar role in chromatin dynamics may be evident at transcription initiation sites, where DRIP signal correlates with prevalence of epigenetic factors associated with open chromatin. Hence, mapping in *T. brucei* supports a role for R-loops at dispersed promoters, also demonstrated at CGI promoters in mammals (Ginno *et al.*, 2012, 2013; P. B. Chen *et al.*, 2015; Sanz *et al.*, 2016), yet contradicts a role for these structures in gene expression regulation, as *T. brucei* lacks transcriptional control of polycistronically transcribed Pol II genes.

# 5 Why is RNase H2A an Essential Protein for *T. brucei* Parasites?

## 5.1 Introduction

Data discussed in Chapter 3 of this thesis have revealed the requirement for functional TbRH2A, and by extension the RNase H2 complex, by *T. brucei* BSF parasites to maintain an intact genome and propagate in culture. RNAi-induced depletion of the complex's catalytic subunit results in stalled growth of the parasites, which accumulate as G2/M stage and aberrant cells. This stall is accompanied by a substantial accumulation of DNA damage; yet, the parasites continue to synthesise DNA, resulting in excess DNA content per cell. This phenotype differs greatly from the seemingly unperturbed *Tbrh1*-/- parasites but does have some common features with RNase H2 mutants of model eukaryotes (see below). Also in common with other eukaryotes, the cause of DNA damage in *T. brucei* RNase H2 mutants cannot yet be confidently attributed to one, or both, of the complex's two functions: R-loop resolution and activation of the RER pathway.

Like *T. brucei* parasites depleted of TbRH2A, mouse cells that are null mutants of the structural B subunit (*Rnaseh2b*-/-) display impaired cell proliferation, arrest in the G2/M cell cycle stage and accumulate DNA damage (Hiller *et al.*, 2012; Reijns *et al.*, 2012). In contrast to *T. brucei*, however, where DNA synthesis appears to continue after TbRH2A depletion, the number of S phase epiblast cells taken from *Rnaseh2b*-/- mouse embryos is significantly decreased in comparison to WT cells (Reijns *et al.*, 2012). Hence, loss of a functional RNase H2 complex in mice induces genuine cell cycle arrest, whereas in *T. brucei* DNA synthesis continues after the parasites fail to divide. Gene expression profiling of liver cells taken from *Rnaseh2b*-/- mice uncovered p53-induced up-regulation of several genes, including those involved in checkpoint activation, highlighting initiation of DNA damage response pathways (Hiller *et al.*, 2012). In an effort to identify the cause of DNA damage in *Rnaseh2b*-/- cells, Hiller et al (2012) assessed ribonucleotide load within the DNA genome by treating purified DNA with bacterial RNase HII in order to induce detectable DNA nicks at the sites of ribonucleotide inclusion. This technique enabled the authors to detect massive inclusion of ribonucleotides in the DNA of mutant cells, but not in WT equivalents (Hiller *et al.*, 2012). Similarly, alkali-treatment of *Rnaseh2b*-/- mouse DNA also revealed extensive accumulation of

alkali-sensitive ribonucleotides (Reijns *et al.*, 2012). Hence, the role of RNase H2 in the RER pathway may underlie the essential need for RNase H2 in mice, although a role for RNA-DNA hybrids was not investigated in these studies and cannot be ruled out.

Mutations in all of the RNase H2 subunits are known to cause the neuroinflammatory disease Aicardi-Goutières syndrome (AGS) in humans, as are mutations of the 3' - 5' endonuclease *TREX1*, dNTP-depleting enzyme *SAMHD*, adenosine deaminase *ADAR1* and viral-DNA detecting helicase *IFIH1*. In all cases, AGS is believed to result from activation of the innate immune pathway due to accumulated nucleic acids (Crow *et al.*, 2015). AGS-causing mutations of RNase H2 can occur at more than 50 known sites across the three subunits and are normally found as heterozygous mutations (Crow *et al.*, 2015). To better model AGS in mice, Pokatayev et al (2016) generated G37S (glycine to serine) heterozygous and homozygous mouse mutants of the catalytic A subunit. RNA-seq carried out with the perinatal lethal homozygous mutant line, revealed an interferon-driven immune response, similar to that seen in AGS patients (Pokatayev *et al.*, 2016). Importantly, the authors were able to elucidate that the immune response was triggered by the presence of cytosolic DNA, not RNA, in *Rnaseh2a*$^{G37S/G37S}$ mouse cells. Furthermore, this was deemed at least partially responsible for the lethal phenotype of the homozygous mutant mice (Pokatayev *et al.*, 2016). However, as the authors highlight, aberrant processing of either RNA-DNA hybrids or embedded ribonucleotides could result in the free nucleic acids activating the DNA sensing pathway (Pokatayev *et al.*, 2016), and so which role, if not both, is the critical function of RNase H2 in mammals is still unclear.

In complete contrast, yeast is able to tolerate null mutation of RNase H2, even in combination with RNase H1 deletion, although elevated recombination does occur (Conover *et al.*, 2015; O'Connell, Jinks-Robertson and Petes, 2015). To elucidate if increased ribonucleotide load in the genome of RNH201-null yeast (where the catalytic subunit 201 is deleted) causes elevated instability, two studies expressed mutant DNA Polymerases that either incorporated greater or fewer ribonucleotides during replication in the RNH201-null line (Conover *et al.*, 2015; O'Connell, Jinks-

Robertson and Petes, 2015). M644G mutation of DNA Pol ε increases the frequency by which the polymerase incorporates ribonucleotides into the genome, and increases loss-of-heterozygosity (LOH) ~7 fold compared to *S. cerevisiae* expressing the WT Pol ε (Conover *et al.*, 2015). When expressed in RNH201-null yeast LOH raised ~23-fold in comparison to RNH201-null yeast expressing WT Pol ε, indicating the cells inability to initiate RER causes wide spread instability (Conover *et al.*, 2015). However, the same change in LOH could not be observed when either Pol α or δ were manipulated to raise ribonucleotide incorporation in RNH201-null yeast (Conover *et al.*, 2015). Both studies also aimed to see if recombination rates could be lower in RNH201-null cells by expressing a second mutant of Pol ε, M644L, which incorporated fewer ribonucleotides during replication (Conover *et al.*, 2015; O'Connell, Jinks-Robertson and Petes, 2015). Conover et al observed a slight decrease in LOH rate in this line compared to the RNH201-null expressing WT Pol ε, though O'Connell et al detected no change in recombination. Additionally, O'Connell et al highlight that although null mutation of RNase H1 in yeast causes little effect on recombination, dual knock out of RNase H1 and RNH201 increases recombination ~50 % compared to RNH201-null mutation alone. Therefore, RNase H2 represses genome instability in yeast via both R-loop resolution and activation of RER to remove embedded ribonucleotides. Yet, loss of these functions is tolerated in yeast in ways mammalian cells are not able to.

Together the above studies imply both RER and R-loop functions of eukaryotic RNase H2 are key for preventing genomic instability. Why the complex is essential in mammalian cells and not in yeast is most likely due to fundamental difference in the genome biology of each, and how RNase H2 fits into these genomic functions. The phenotype of *T. brucei* parasites depleted of TbRH2A has more in common with mammalian cells than yeast, and so events generating the respective DNA lesions may also be shared. The TbRH2A-tageting RNAi parasite line described in this thesis provides, firstly, a means to try and isolate the cause (or causes) of lethal DNA lesions in a eukaryote with dampened RNase H2 activity, and secondly, an opportunity to further investigate R-loop formation across the *T. brucei* genome.

### 5.1.1 Chapter aims

In this chapter TbRH2A depleted parasites were investigated in more detail. Firstly, R-loops were mapped across the Mb chromosomes to ascertain if sites of RNA-DNA hybrid formation are conserved between samples and look for changes in hybrid location after the subunits depletion. Secondly, γH2A deposition was mapped via ChIP-seq analysis to locate sites of DNA damage after loss of RNase H2 activity. Finally, RNA-seq was carried out to profile gene expression in induced RNAi parasites in comparison to un-induced. Together these studies aim to investigate where DNA damage occurs most frequently in these parasites, if this damage relates to R-loop formation, and how the parasites respond to such damage.

## 5.2 DRIP-seq mapping of R-loops in TbRH2A depleted cells

To evaluate R-loop localisation before and after TbRH2A RNAi, DRIP-seq was performed as described in chapter 4, using the TbRH2A$^{RNAi}$ line described in chapter 3. Parasites were cultured for 24 hr in the presence or absence of tet (1 μg/ml) before fixed chromatin samples were collected and subject to IP with the S9.6 RNA-DNA hybrid targeting antibody (section 2.6). Mapping and analysis were then undertaken using the same methods employed to investigate R-loop formation in WT and *Tbrh1*-/- cells (section 2.8.4).

### 5.2.1 R-loops form across the Mb chromosomes in TbRH2A RNAi parasites

DRIP-seq data was mapped across the *T. brucei* Mb chromosomes and signal was expressed as the ratio of IP read coverage relative to input read coverage, in both induced and un-induced TbRH2A$^{RNAi}$ samples (Figure 5.1). As with WT and *Tbrh1*-/- mapping, R-loop formation was extensive across the Mb chromosomes in both un-induced and induced samples, where the highest peaks again correlated with tandem repeat regions (Figure 5.1). In both samples, extensive DRIP signal was observed in the repetitive subtelomeric regions, housing VSG arrays.

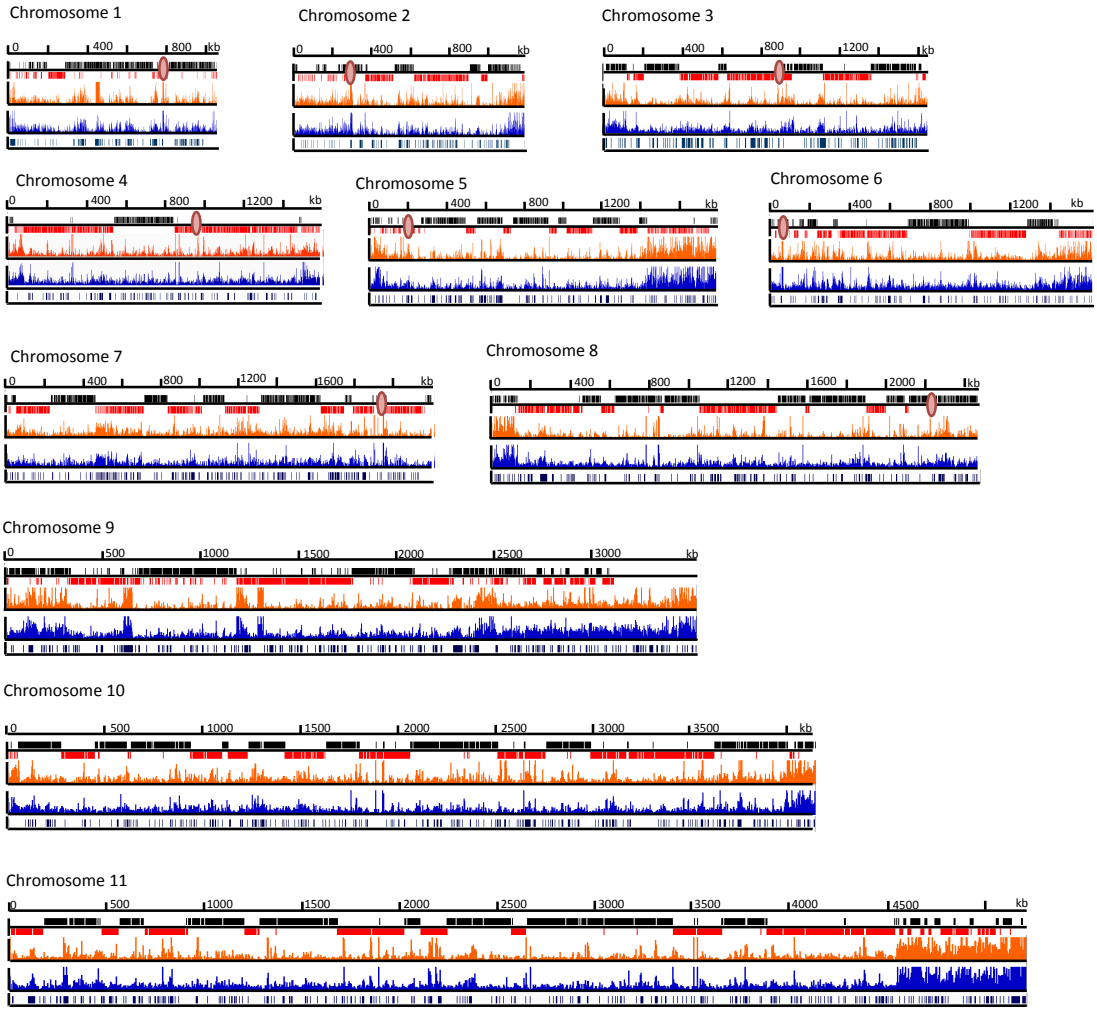



**Figure 5.1 DRIP-seq signal across the 11 Mb-sized chromosomes of TbRH2A RNAi parasites.** DRIP-seq signal for un-induced (blue) and induced (orange) cells is plotted as fold-change relative to input coverage (scale: 1-4 fold-change). Upper track shows transcripts on sense (black) and antisense (red) strands. Known centromeres are shown as peach circles. Lowest track shows tandem repeat sequences.

## 5.2.2 DRIP enriched region classification in TbRH2A depleted parasites

Enriched regions of DRIP-seq signal (>1.2 fold-change) were identified as described for the analysis of WT and *Tbrh1*-/- DRIP data sets (section 4.2). Again the enriched regions, corresponding to sites of RNA-DNA hybrid formation, were classified according to their genomic location and compared to classification of WT DRIP enriched regions to allow consideration of leaky RNAi expression in the tet – parasites (Figure 5.2A). The majority of DRIP enriched regions were located within the Pol II transcribed PTUs, with similar amounts of enriched regions (33,879 tet -, and 35,329 of tet + regions) compared with WT regions (35,320). The total number

of enriched regions between the tet – and tet + samples increased slightly from 39,243 to 40,155, and both increased from the 37,527 WT regions. The location of these enriched regions did not notably differ between tet – and tet + samples. However, compared to classification of WT enriched regions the amount of DRIP enriched regions in the VSG array sites increased from 1,389 WT regions to 3,380 and 2,800 regions in the un-induced and induced samples, respectively. The number of DRIP enriched regions forming in loci housing RNA Pol I transcribed genes also increased from 213 in the WT data to 1,153 tet – enriched regions and 1,102 tet + enriched regions. Lastly, RHS-associated DRIP enriched regions increased from 172 in WT cells to 454 and 393 in tet – and tet + cells, respectively (Figure 5.2A).

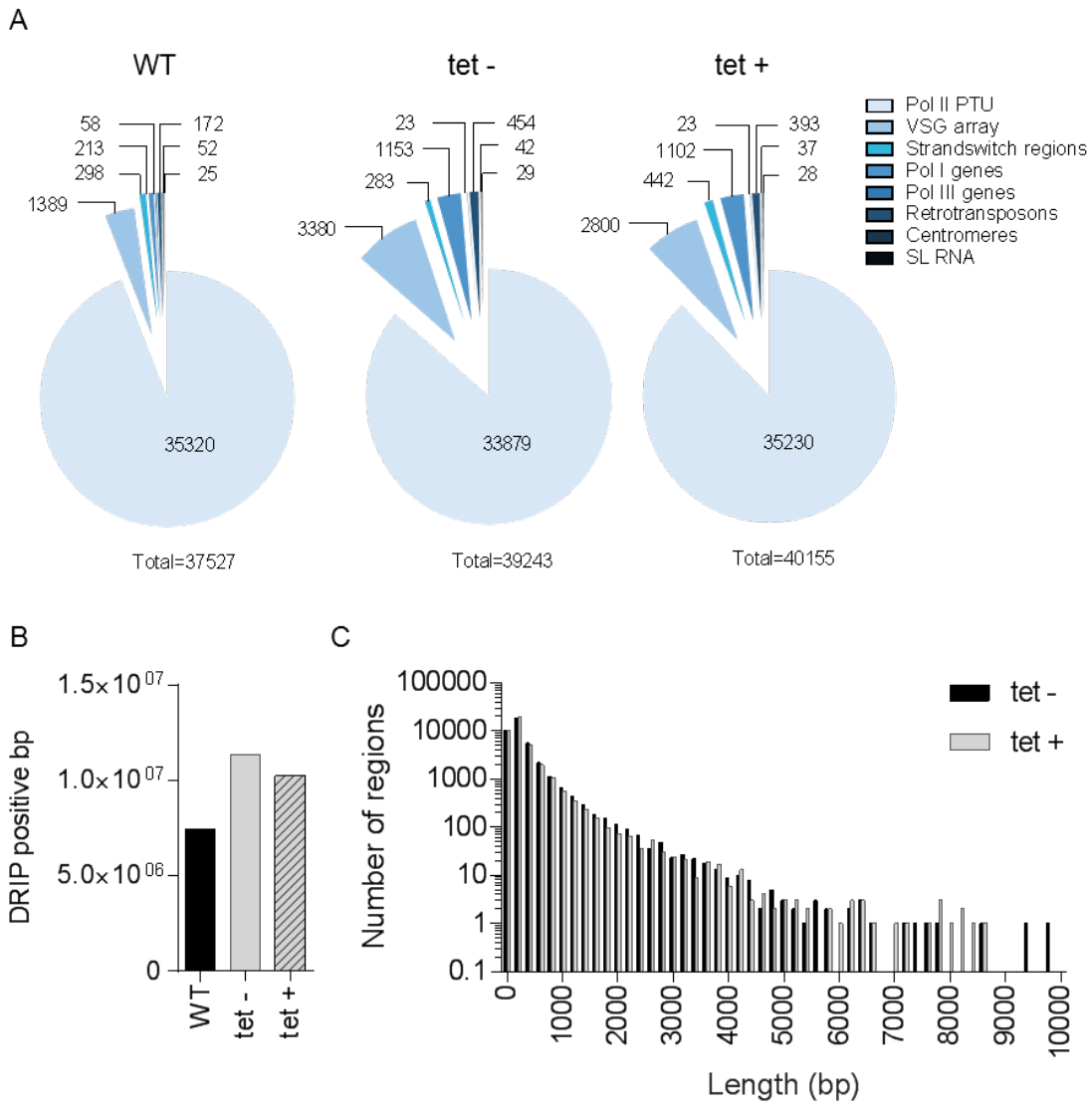



**Figure 5.2 Analysis of DRIP enriched region distribution before and after TbRH2A RNAi knockdown.**
A) The distribution of DRIP enriched regions in WT, TbRH2A RNAi un-induced (tet -) and induced (tet +) samples across the genome elements. B) Bar graph of the number of bp occupied by a DRIP enriched region in WT, TbRH2A RNAi tet- and tet + samples. C) Histogram of DRIP enriched region lengths in tet – and tet + samples.

The number of bp occupied by a DRIP enriched region slightly decreased, from 11,324,282 bp to 10,230,755 bp, after tet induction of the RNAi cells (Figure 5.2B). Again, however, comparison with WT DRIP analysis revealed these values were both increased compared to 7,461,359 bp occupied by an RNA-DNA hybrid in the WT sample. The length of the enriched regions also revealed little difference between tet – and tet + DRIP enriched regions (Figure 5.2B), although this was again increased compared to WT enriched regions.

### 5.2.3 R-loops form in centromeric repeats after TbRH2A depletion

As in *Tbrh1-/-* mutant parasites, in both tet – and tet + TbRH2A$^{RNAi}$ parasites DRIP-seq signal peaks formed over the centromeric repeat regions compared to WT DRIP (Figure 5.3). In both cases DRIP signal was enriched over the repeat regions of all centromeres (Figure 5.3,Figure 8.4), above the level observed in WT DRIP mapping.

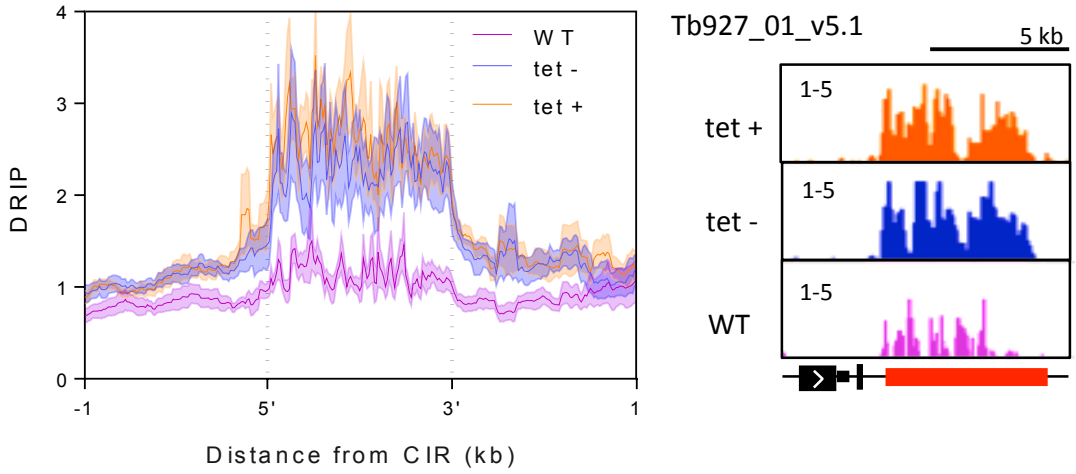



**Figure 5.3 TbRH2A-targeted R-loops form at the centromeric repeats of *T. brucei*.**
Left: Metaplot of average DRIP-seq signal of WT (pink), TbRH2A$^{RNAi}$ un-induced (tet -, blue) and induced (tet +, orange) samples over scaled centromeric repeats plus 1 kb of up and downstream flanking sequence. Standard error is shown as shaded area. Right: Screenshot example of DRIP-seq signal across the annotated centromeric repeats of chromosome 1. Colours are as for metaplot profile. Scale, 1-5 fold-change relative to input. Centromeric repeats are shown in red, CDS in thick black and UTR sequences in thin black lines. For DRIP-seq signal of all other centromeres see Figure 8.4.

### 5.2.4 DRIP signal is enriched at RHS-associated genes in TbRH2A$^{RNAi}$ parasites.

As witnessed at the centromeric loci, in both the tet – and tet + DRIP samples signal was enriched compared to WT DRIP mapping at RHS-associated genes (Figure 5.4). Correspondingly, the number of DRIP enriched regions overlapping RHS-associated gene coding regions increased from 167 WT DRIP enriched regions to 379 and 443 enriched regions in un-induced and induced *Tbrh2a* RNAi samples, respectively (Figure 5.4, left). Notably, DRIP signal was seen to 'dip' over the coding regions of genes associated with RHSs (Figure 5.4, right).

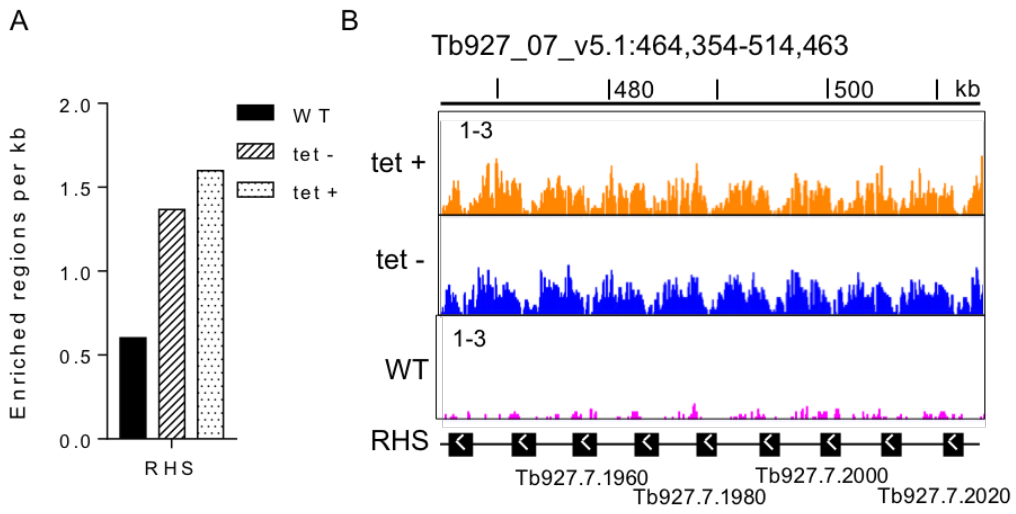



**Figure 5.4 R-loops form over RHS loci of both un-induced and induced TbRH2A RNAi parasites.** A) Bar graph showing the number of DRIP-enriched regions found over the RHS regions in WT (black), TbRH2A RNAi un-induced (tet -, grey striped) and induced (tet +, white dotted) cells. B) Screenshot of DRIP-seq signal enrichment over a section of RHS gene array found in chromosome 7. WT DRIP-seq signal is shown in pink, TbRH2A RNAi un-induced (tet -) in blue and induced (tet +) in orange.

### 5.2.5 DRIP enriched regions increase at Pol I transcribed sites in TbRH2A$^{RNAi}$ cells relative to WT

As seen after *TbRH1* deletion, DRIP signal and DRIP enriched regions were both increased at RNA Pol I transcribed loci in the TbRH2A$^{RNAi}$ parasites, both induced and un-induced, compared to WT parasites (Figure 5.5A). This was most notable within the rRNA loci and BESs (Figure 5.5A and B). Again, the numbers of enriched regions identified in un-induced and induced DRIP samples did not notably differ in any Pol I transcribed loci.

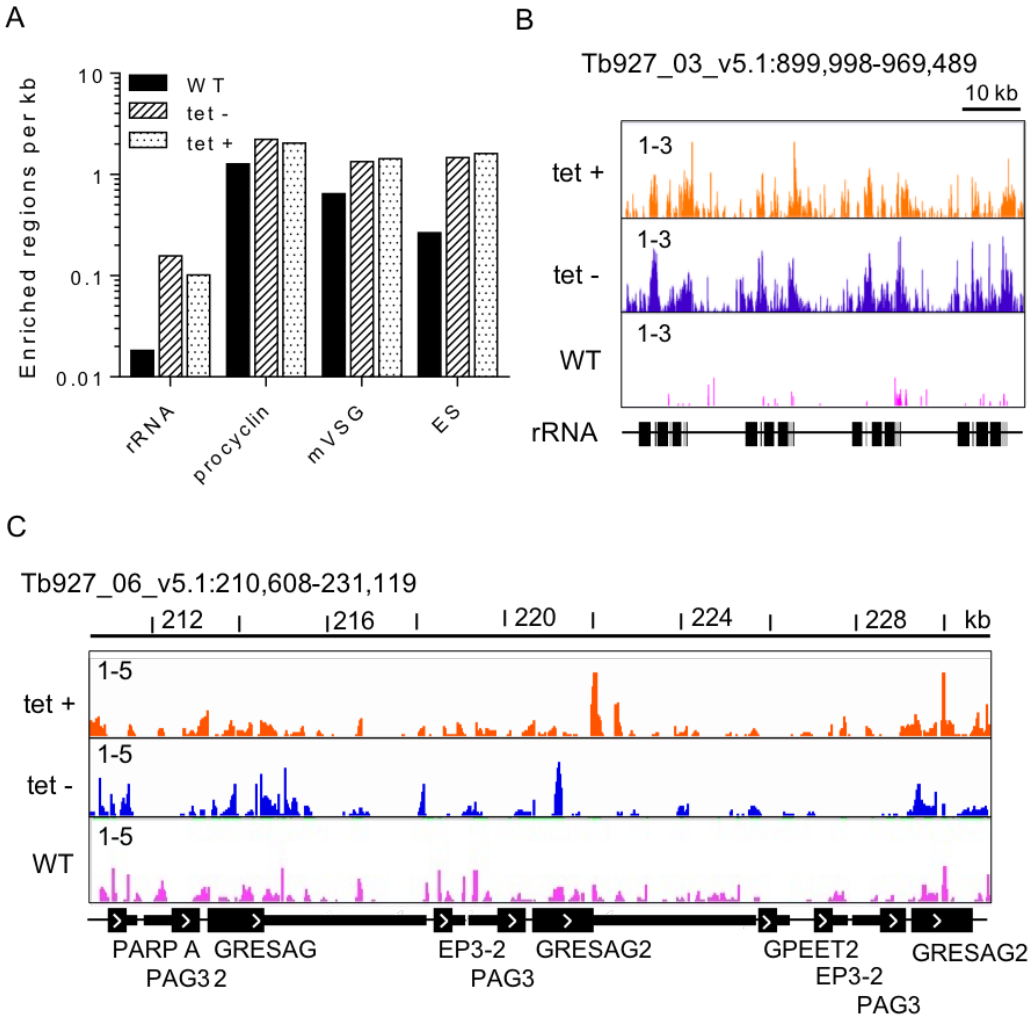



**Figure 5.5 R-loops form at sites of RNA Pol I transcription in TbRH2A RNAi parasites.**
A) Bar graph showing the number of enriched regions found in each type of RNA Pol I transcribed site in WT, TbRH2A$^{RNAi}$ un-induced (tet -, grey striped) and induced (tet +, white dotted) cells. B) Screenshot of DRIP-seq signal for WT (pink), TbRH2A$^{RNAi}$ un-induced (blue) and induced (orange) samples over the rRNA array of chromosome 8. C) Screen shot of WT (pink) TbRH2A RNAi un-induced (blue) and induced (orange) DRIP-seq signal over the procyclin expression locus.

## 5.2.6 DRIP enriched decrease at Pol III transcribed genes in TbRH2A$^{RNAi}$ cells relative to WT

Interestingly, the number of DRIP enriched regions found at RNA Pol III transcribed loci decreased in the TbRH2A$^{RNAi}$ parasites cultured in the presence or absence of tet compared to WT DRIP enriched regions. This was true for both tRNA and snRNA genes, and was comparable with the mapping observed for *Tbrh1-/-* parasites.

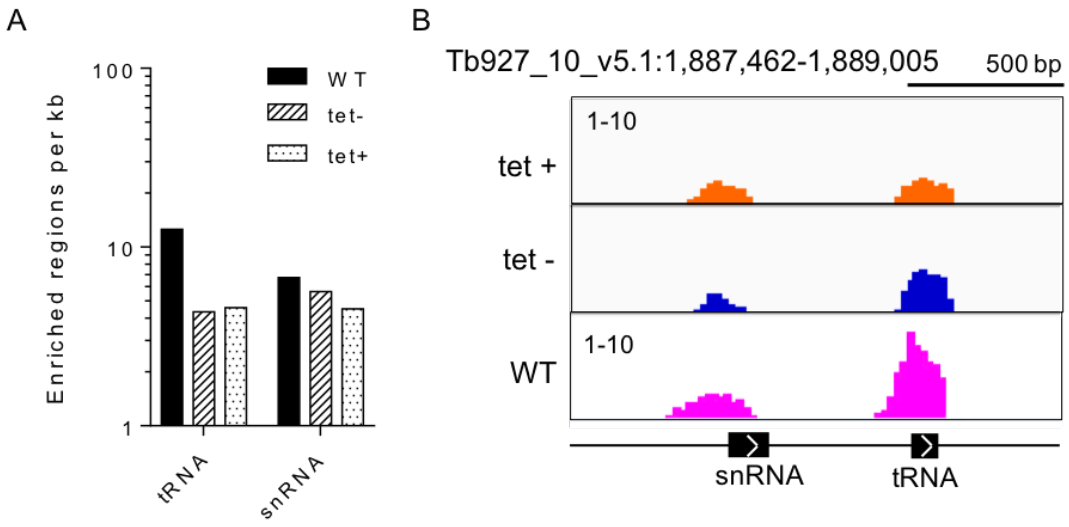



**Figure 5.6 Fewer R-loops form over RNA Pol III transcribed snRNA and tRNA genes in the TbRH2A RNAi line.**
A) Bar graph showing the number WT, TbRH2A RNAi un-induced (tet -, grey striped) and induced (tet +, white dotted) DRIP enriched regions found over tRNA and snRNA genes. B) Screenshot of region on chromosome 10 where both an snRNA and tRNA are located. WT (pink), TbRH2A RNAi un-induced (blue) and induced (orange) DRIP-seq signal is shown.

### 5.2.7 RNA-DNA hybrids form at UTRs and intergenic sequences

DRIP-seq mapping in WT and *Tbrh1-/-* parasites revealed a strong correlation between hybrid formation and non-CDS sequences. Analysis of DRIP enriched regions within the RNA Pol II transcribed PTUs found this correlation is conserved in the TbRH2A$^{RNAi}$ cells, both induced and un-induced: ~58.3% and ~60.1% of DRIP enriched regions overlapped most extensively with intergenic sequences in un-induced and induced TbRH2A$^{RNAi}$ parasites, respectively. This was slightly increased from the 50.7% of intergenic sequence-associated DRIP enriched regions found in WT. The number of 5' UTR, 3' UTR and CDS-associated DRIP enriched regions all decreased in the un-induced sample compared to WT and decreased again in the RNAi-induced sample relative to un-induced. Visualisation of DRIP signal across the PTUs showed the pattern of association of DRIP signal peaks and non-CDS sequences is remarkably conserved between the WT, TbRH2A$^{RNAi}$ un-induced and induced samples (Figure 5.7).

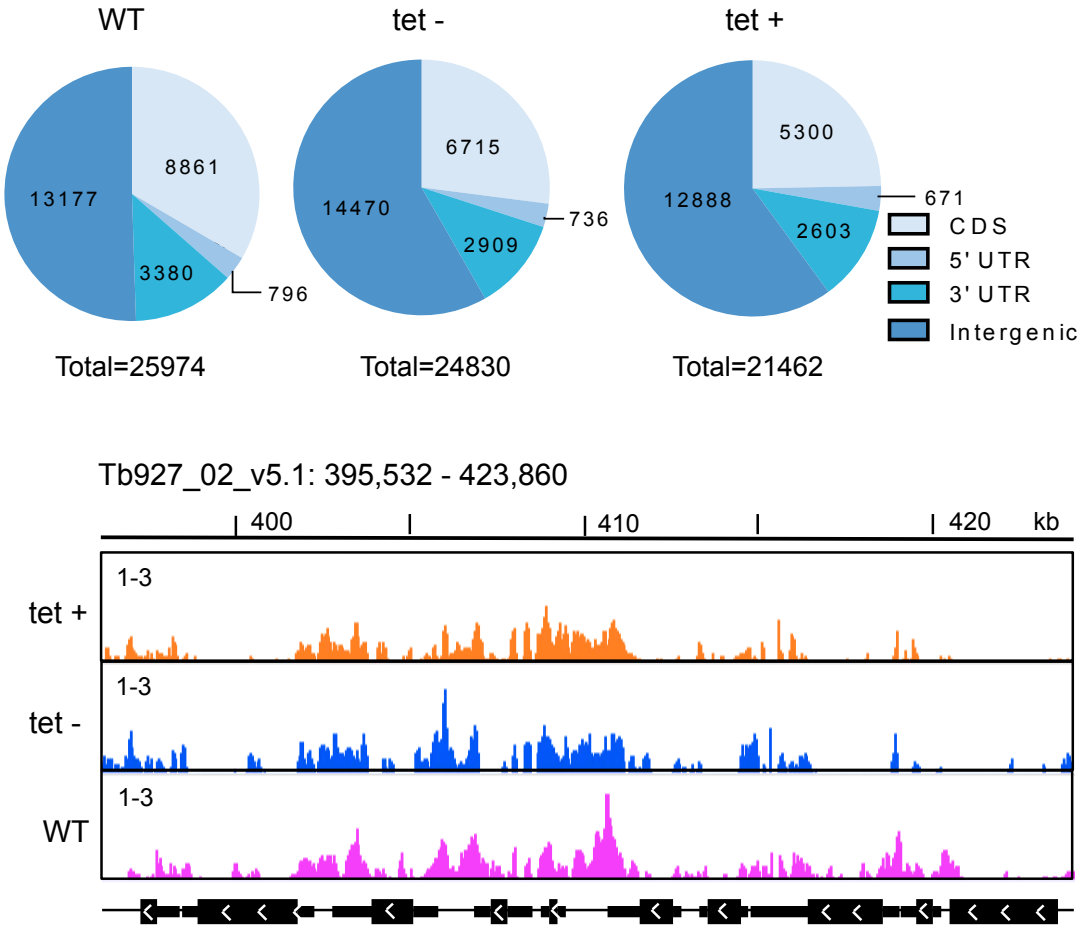



**Figure 5.7 Distribution of DRIP-seq enriched regions over the PTU genomic elements.** Above: The association of WT DRIP enriched regions found within the RNA Pol II transcribed PTUs is shown as the proportion mapping to intergenic, CDS or UTR sequences in WT, TbRH2A RNAi un-induced (tet -) and induced (tet +) samples. Below: Screen shot of RNA Pol II transcribed protein-coding genes in an area of chromosome 2. WT DRIP-seq signal is shown in pink, TbRH2A RNAi un-induced (tet -) in blue and induced (tet +) in orange; scale bars show enrichment in DRIP relative to input, as in previous figures.

DRIP-seq signal was plotted as a heatmap and profile summary of all RNA Pol II transcribed protein-coding genes for both the tet – and tet + DRIP data (Figure 5.8). The WT heatmap and profile are also shown for comparison. In all three cases, distinctive peaks of read enrichment in the 5' and 3' regions flanking the CDS were conserved. In the WT metaplot profile, DRIP signal remained at a value of ~1 across the CDS, meaning hybrids are predominantly lacking across RNA Pol II transcribed CDS (i.e. no enrichment in S9.6 IP vs input). In both the tet- and tet + profiles, the signal dips below 1, indicating that CDS sequences comprise a smaller proportion of the tet – and tet + IP libraries compared to input after normalisation. It is notable that the distance between the signal peaks at CDS flanks and the trough of signal across the CDS is greater in the tet – and tet + profiles compared to WT.

Interestingly, this distance was less in the tet + sample compared to tet -. As observed in section 4.7.1, DRIP enriched regions correlated strongly with PAS in the non-CDS sequences, for both tet – and tet + DRIP data sets (Figure 5.8B and C).

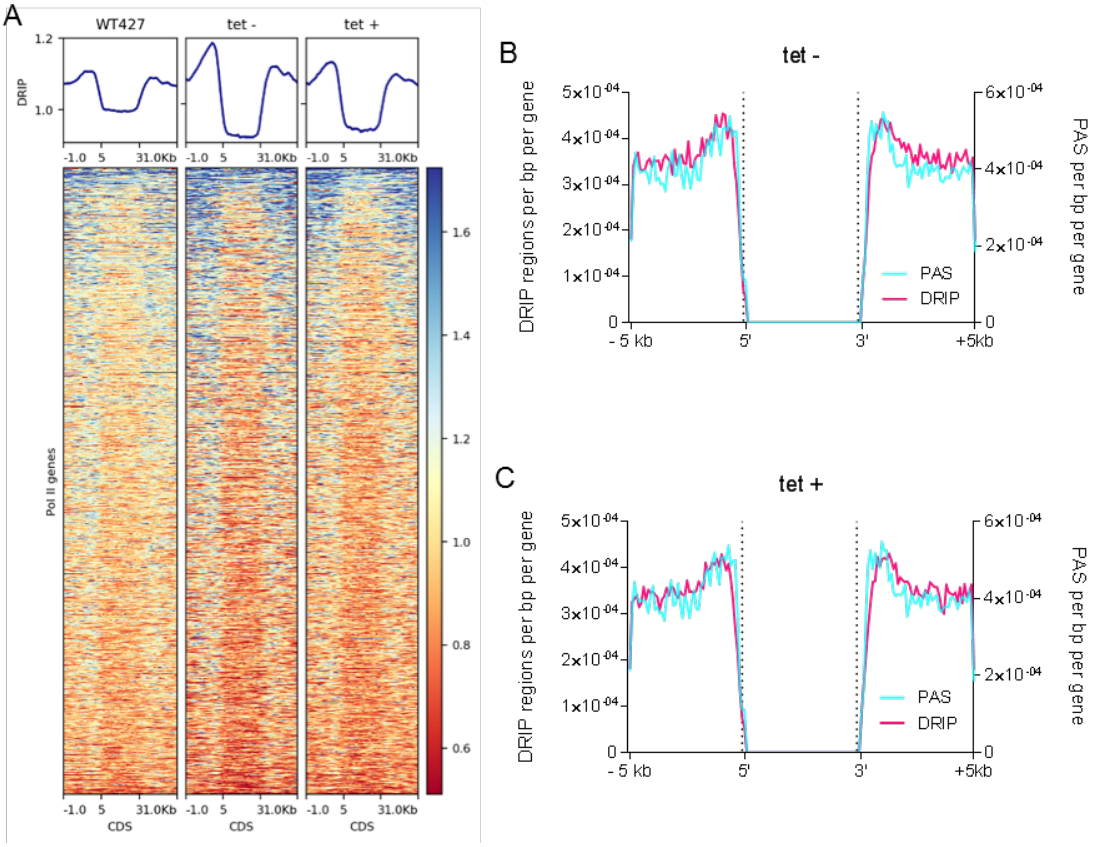



**Figure 5.8 DRIP-seq signal is enriched over the 5' and 3' flanking sequences of each RNA Pol II transcribed CDS.**
A) Heatmaps of DRIP-seq signal, for WT and TbRH2A RNAi induced (tet +) or uninduced (tet -) samples, over each RNA Pol II transcribed protein-coding CDS, plus and minus 1 kb of flanking sequence. Metaplot profile summaries of the data are also shown above. Metaplots of the number of identified DRIP enriched regions per bp per gene (pink, left axis) and the number of PAS per bp per gene (blue, right axis) over scaled CDS regions plus 5 kb of flanking sequences in tet- (B) and tet + (C) samples.

## 5.2.8 DRIP-seq signal deceases after RN2A depletion at sites of transcription initiation

Unlike the data discussed above, investigation of DRIP-seq signal around SSRs revealed differences in signal between the tet-induced and un-induced cells. DRIP-seq signal was plotted over each class of SSR: divergent, representing sites of transcription initiation in both sense and anti-sense directions; convergent sites of transcription termination; and head-to-tail SSRs, where transcription both

terminates and initiates. In each case, WT DRIP-seq signal was plotted as a comparison. In all cases, DRIP-seq signal was highly comparable between the TbRH2A un-induced and WT samples: signal was lacking across most of the SSR but increased to form a distinct peak at the boundary where transcription is predicted to initiate (Figure 5.9). In convergent SSRs, where transcription terminates and does not initiate, no such peaks appeared. Unexpectedly, upon RNAi depletion of TbRH2A, DRIP-seq signal was reduced relative to both the un-induced and WT signal at both divergent SSRs and the regions associated with transcription initiation in head-to-tail SSRs (Figure 5.9). Again, this decrease in the RNAi-induced cells was not apparent in convergent SSRs; instead, signal here was greater after TbRH2A RNAi relative to WT, and comparable with un-induced (as was seen in non-SSR mapping; see above).These patterns were observable in track images of DRIP-seq signal at individual SSRs; several enriched peaks that were apparent in the un-induced sample at the TSSs, were not evident at the equivalent sites of the tet-induced sample. Hence, TSS-associated R-loops were not apparent after TbRH2A knockdown. It should be noted that DRIP-seq signal peaks observed within the divergent and head-to-tail SSRs correspond to RNA Pol III transcribed tRNA and snRNA genes (Figure 5.9, red).

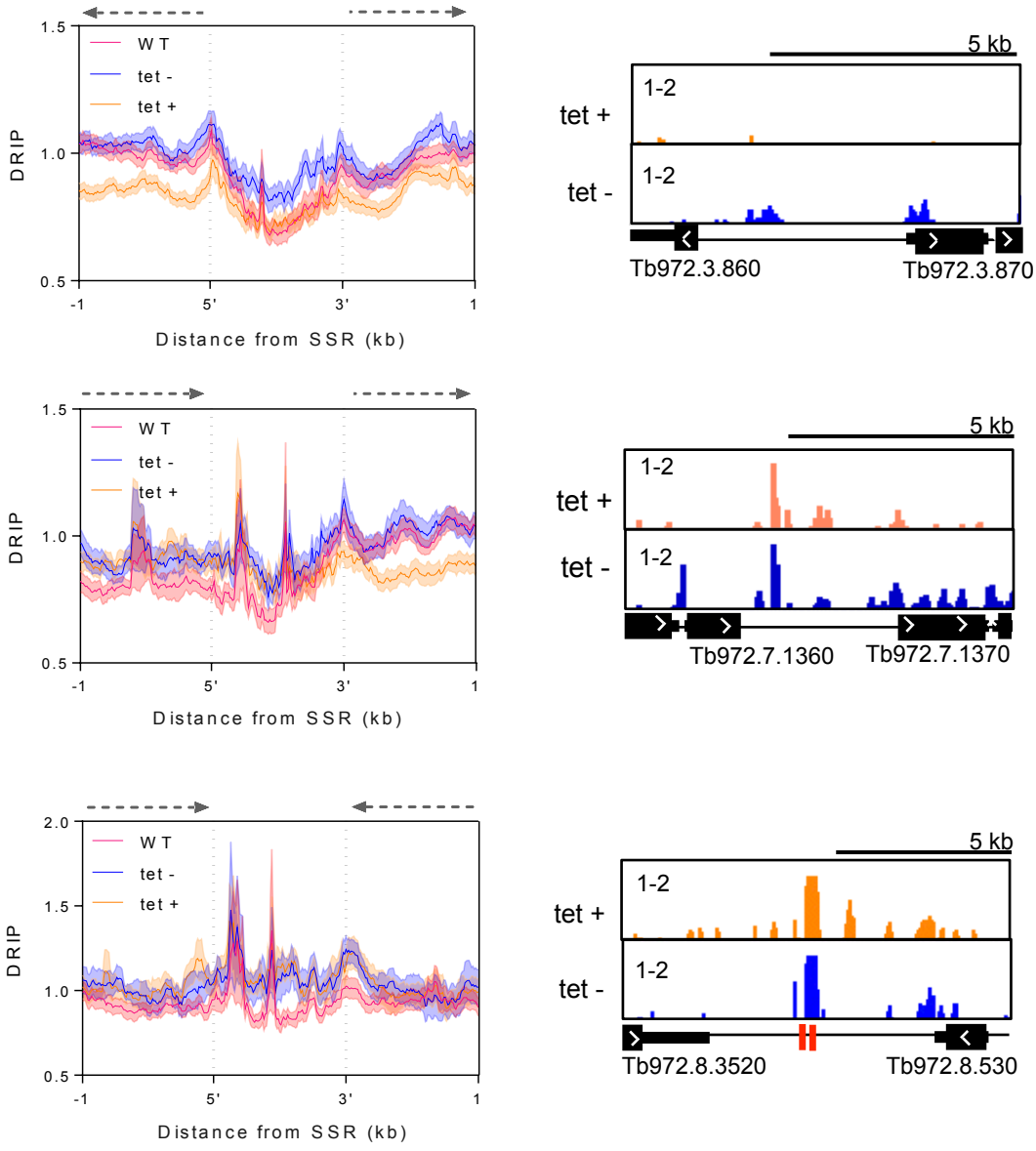



**Figure 5.9 RNA-DNA hybrids are depleted at RNA Pol II transcription initiation sites after knockdown of TbRH2A.**
Average DRIP-seq signal is plotted for WT (pink), TbRH2A RNAi un-induced (tet -, blue) and induced (tet +, orange) data sets over divergent (upper), head-to-tail (centre) and convergent (lower) SSRs. In all cases 5' and 3' denote SSR boundaries defined by flanking transcript coordinates. Transcription direction is shown above the plots by dashed arrows. Standard error is shown as shaded regions. DRIP-seq signal screenshot examples of individual SSRs in each class is also shown to the right of each metaplot; CDS (thick black), UTR (thin black snRNA/tRNA genes (red) are shown below the DRIP-seq tracks.

Next, SSRs were re-grouped according to whether they have additionally been identified as ORIs or not (Tiengwe *et al.*, 2012), and the average DRIP-seq signal was plotted across the two groups. DRIP signal was slightly reduced in the profile of RNAi induced cells of both groups, indicating that it is the transcriptional status,

not replicative status, of the SSRs that dictate this change in DRIP signal (Figure 5.10).

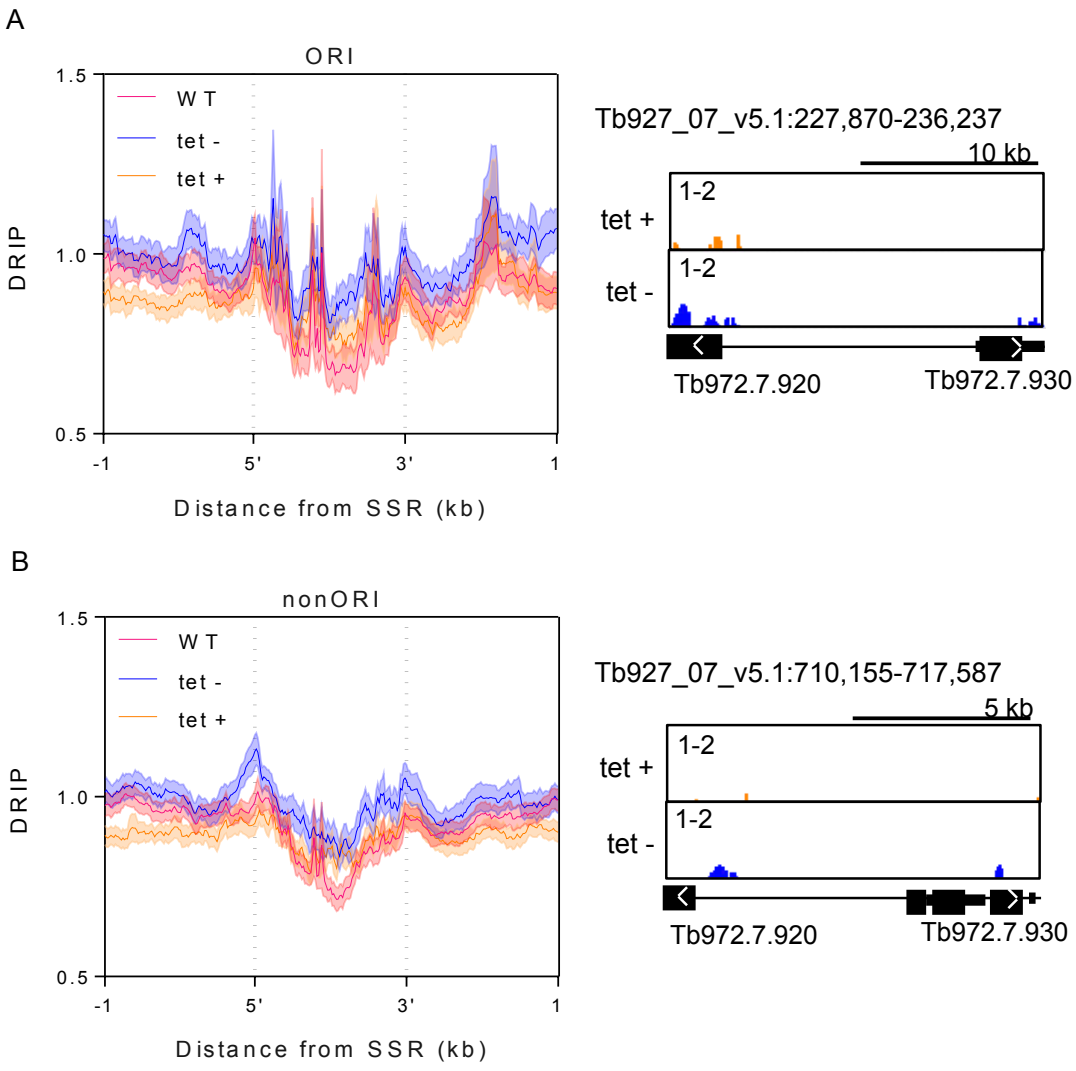



**Figure 5.10 Loss of RNA-DNA hybrids after TbRH2A depletion is not dependent on ORI status of the SSRs.**
WT (pink), TbRH2A RNAi un-induced (tet -, blue) and induced (tet +, orange) DRIP-seq average signal is plotted across SSRs which are also ORIs (upper left) and those which are not (lower left). Standard error is plotted as shaded regions. Screenshots of DRIP signal across two example SSRs for one ORI (upper) and one non-ORI (lower) are also shown.

## 5.3 DNA damage occurs at sites of transcription initiation when TbRH2A is depleted

ChIP-seq targeting histone variant γH2A was next employed to map sites of DNA damage in both tet-induced and un-induced cells. In this case, the analysis was conducted 24 and 36 hr post-induction (Figure 5.11, Figure 5.12 and Figure 8.5),

since IFA (Chapter 3) showed that γH2A signal increased between these time points after TbRH2A depletion. For each condition, ChIP-seq coverage was normalised to the related input sample and plotted across the 11 Mb chromosomes (Figure 8.5). Extensive γH2A signal was observed across the RNA Pol II transcribed PTUs in all cases (Figure 8.5), and little difference was evident between samples. This signal is presumably due to specificity of the anti-γH2A serum, which is added in excess to the IP reaction, resulting in binding not simply to Thr130 phosphorylated H2A but also to unphosphorylated, canonical H2A histone. Hence, in order to focus analysis upon changes after knockdown of *TbRH2A* expression (where increased γH2A staining has been demonstrated via IFA and western blot, Figure 3.17), the fold-change in signal between input-normalised, induced and un-induced ChIP-seq coverage was calculated as a ratio for each time point (Figure 5.11 and Figure 5.12). Therefore, in this analysis, a value of 1 indicates no difference in γH2A ChIP-seq coverage between induced and un-induced samples, whereas values >1 indicate γH2A enrichment after loss of TbRH2A.


**Figure 5.11 γH2A binding occurs at RNA Pol II transcription initiation sites across the 11 mb chromosomes after TbRH2A knockdown.**
γH2A ChIP-seq signal increase in induced relative to un-induced is plotted across the Mb chromosomes, for 24 hr (pink) and 36 hr (green) induction. (Scale 1-3 fold-change). Upper track shows transcripts on sense (black) and antisense (red) strands. Lowest track shows tandem repeat sequences.

Plotting of the ChIP-seq signal enrichment across the Mb chromosomes revealed clear peaks at potentially all known RNA Pol II transcription initiation sites (Figure 5.11 and Figure 5.12). Dense γH2A binding is also clear across the repetitive subtelomic regions that house the VSG arrays (Figure 5.11). To ask how γH2A ChIP-seq signal correlated with the transcription boundaries, signal was analysed across the three classes of SSRs, in order to compare sites of R-loop and γH2A enrichment. 24 hr post-induction, the ratio of induced signal relative to un-induced remained at a value of ~1, indicating there was no γH2A formation (Figure 5.12). However, 36 hr

post-induction, average γH2A ChIP-seq signal raised ~1.2 fold after *TbRH2A* knockdown at both the head-to-tail and divergent SSRs, but not at the convergent SSRs. Notably, the signal increase covered the entirety of the divergent SSRs (and extended into the transcribed flanks) but was only seen around the regions of transcription initiation, not termination, at head-to-tail SSRs. Notably, this pattern shows a clear correlation with pronounced R-loop formation around SSR transcription initiation regions, as well as the SSR regions that showed reduced R-loop levels after TbRH2A RNAi (Figure 5.10).

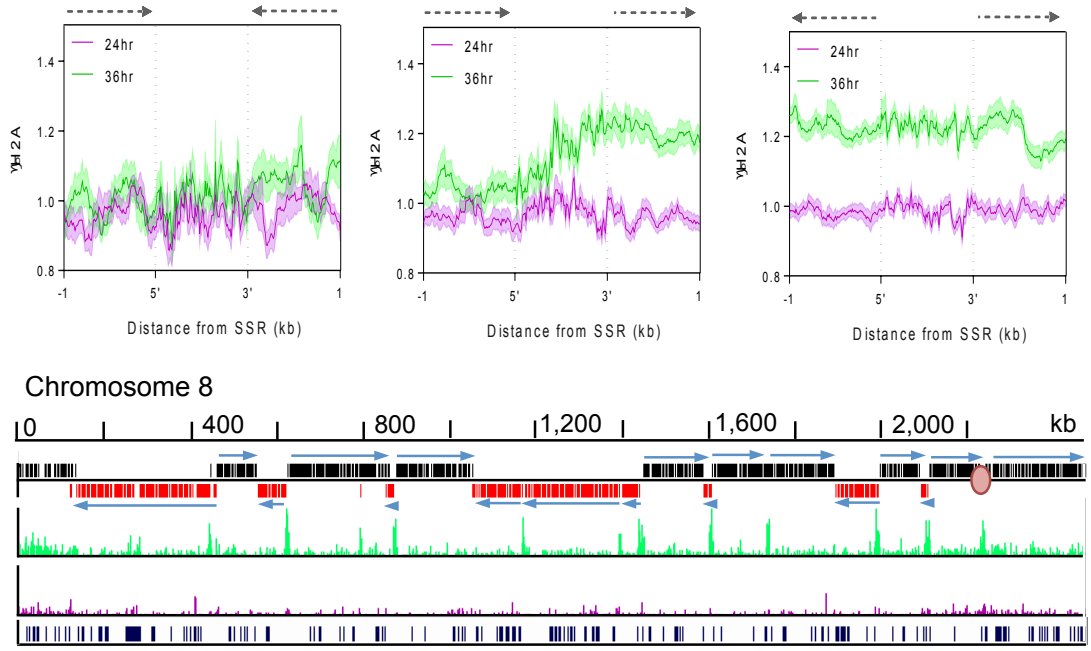



**Figure 5.12 DNA damage is associated with sites of transcription initiation after TbRH2A knockdown.**
Increased γH2A ChIP-seq signal in TbRH2A RNAi induced samples relative to un-induced is plotted for both 24 hr (pink) and 36 hr of RNAi induction. Average signal is plotted across divergent (left), head-to-tail (centre) and convergent (right) SSRs. In all cases 5' and 3' denote SSR boundaries defined by flanking transcript coordinates. Transcription direction is shown above the plots with dashed arrows. Standard error is shown as shaded regions. γH2A ChIP-seq signal increase in induced relative to un-induced cells is also shown plotted across chromosome 8, after 24 hr (pink) and 36 hr (green) induction (scale 1-3 fold-change). Upper track shows transcripts on sense (black) and antisense (red) strands and arrows highlight transcription direction. Lowest track shows tandem repeat sequences.

## 5.4 Depletion of TbRH2A causes up-regulation of VSGs and down-regulation of small molecule biosynthesis pathways

Next, RNA-seq analysis was performed to assess if changes in gene expression in *T. brucei* parasites was detectable after RNAi depletion of TbRH2A. Total RNA

extraction was performed, in duplicate, using parasites that had been cultured for 24 and 36 hr, both with and without tet-induction (section 2.3.1). RNA-seq libraries were generated using poly(A) selection (performed by C. Lapsley, section 2.3.4) and the resulting reads were mapped to the *T. brucei* 927 reference genome (performed by Dr. K. Crouch, section 2.8.4.1). DESeq2 software was then employed to analyse the duplicate data and identify statistically significant, differentially expressed RNAs in the tet + samples relative to tet – at the two time points (performed by K Crouch, sections 2.8.5.2 and 2.8.5.3). Genes that showed differential expression were deemed significant if the assigned adjusted p value was < 0.05 (Figure 5.13).

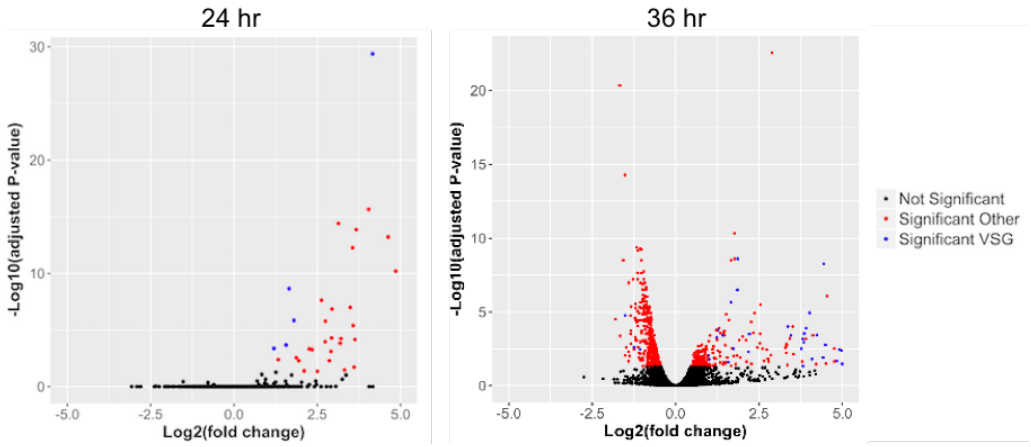



**Figure 5.13 Volcano plots displaying differential expression of genes after TbRH2A knockdown.** Each data point represents a gene; red genes were deemed significantly differentially expressed (adjusted p value < 0.05) relative to uninduced levels. X-axes shows the log2 fold-change between un-induced and induced samples after 24 hr (left) and 36 hr (right) of culture, and Y-axes shows log2 adjusted p value. Data was generated with two independent replicates for each condition. Significantly differentially expressed VSG genes are shown in blue, and all other genes in red. Not significant genes including VSGs are shown in black. (Plotted by Kathryn Crouch).

Given the severe growth phenotype and large accumulation of DNA damage in *TbRH2A* depleted cells, differential expression of many genes might have been anticipated. However, RNA from only 32 genes was significantly differentially expressed after 24 hr of RNAi knockdown and all of these RNAs showed increased reads compared to un-induced samples (Figure 5.14, Table 8.1). VSGs (5) and ESAGs (25) accounted for nearly all of this gene set, and the remaining two genes were procyclin-related (Figure 5.14). These data, indicating altered expression of antigenic variation-associated and RNA Pol I-transcribed genes, are explored in more detail in chapter 6. In fact, after 36 hrs tet-induction these effects were

more marked: from a total of 113 RNAs (Figure 5.14,Table 8.2) that showed significantly increased reads, 18 of the 32 genes seen at 24 hrs remained up regulated; and, in total, 37 VSGs and 14 ESAGs were up-regulated at 36 hrs, along with 10 procyclin and 10 RHS-associated genes (Figure 5.14). Amongst the remaining 42 genes with significantly increased RNA, 14 encode hypothetical proteins, 4 were listed as GRESAGs (genes relating to ESAGs), and a SLAC reverse transcriptase and a centrosomal spinal body protein, CEP44, was detected (Table 8.2).

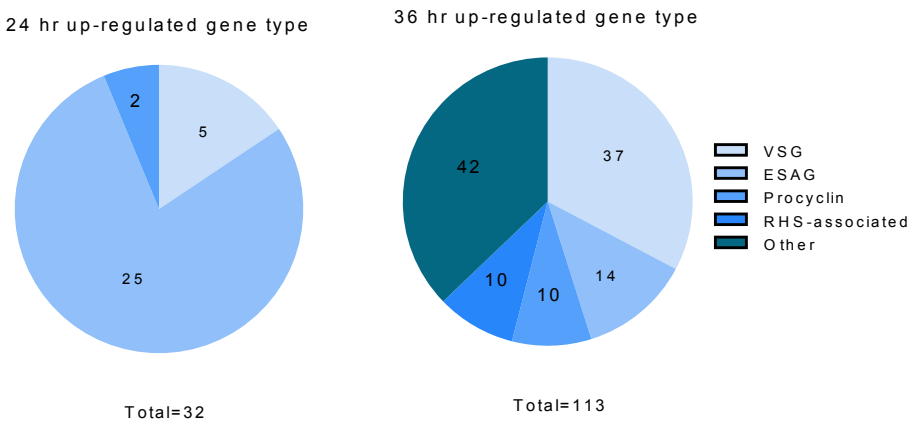



**Figure 5.14 Classes of genes up-regulated after TbRH2A knockdown.**
The number of genes up-regulated in tet-induced TbRH2A RNAi parasites after 24 (left) and 36 hr (right) of growth that are annotated at VSGs, ESAGs, procyclin, RHS-associated and other genes. Total number of genes up-regulated genes are shown below. For full list of genes and associated statistics see Table 8.1and Table 8.2.

In contrast to the exclusive increase in RNA after 24 hrs of TbRH2A RNAi, 396 genes were instead found to be down-regulated after 36hr, of RNAi induction (Table 8.3). Using this set of genes, along with the 113 found to be up-regulated, gene ontology (GO term) analysis was performed to investigate the biological processes which are affected after 36 hrs of TbRH2A knockdown (section 2.8.5.5). GO terms were plotted as a network, where each node represents a GO term and connecting edges show hierarchical relationships. All terms over-represented in the down-regulated (orange) or up-regulated (green) genes were plotted, where colour intensity corresponds with adjusted p value (Figure 5.15). In both cases, GO terms were deemed significantly enriched if an adjusted p value < 0.01 was found. Consistent with the predominance of VSGs and ESAGs, antigenic variation and parent GO terms

were enriched in the up-regulated set of genes. Down-regulated biological processes appeared more diverse, but small molecule biosynthetic processes were the most significantly down-regulated term. Branches of this term include several relating to the synthesis or salvage of both nucleotides and ribonucleotides, particularly purines. Other terms involve the metabolic processing of other small molecules, including cellular carbohydrates, ketones and organic acids. Lastly, cytoskeleton organisation, cellular component movement and molecule transport were also down-regulated upon depletion of TbRH2A (Figure 5.15).

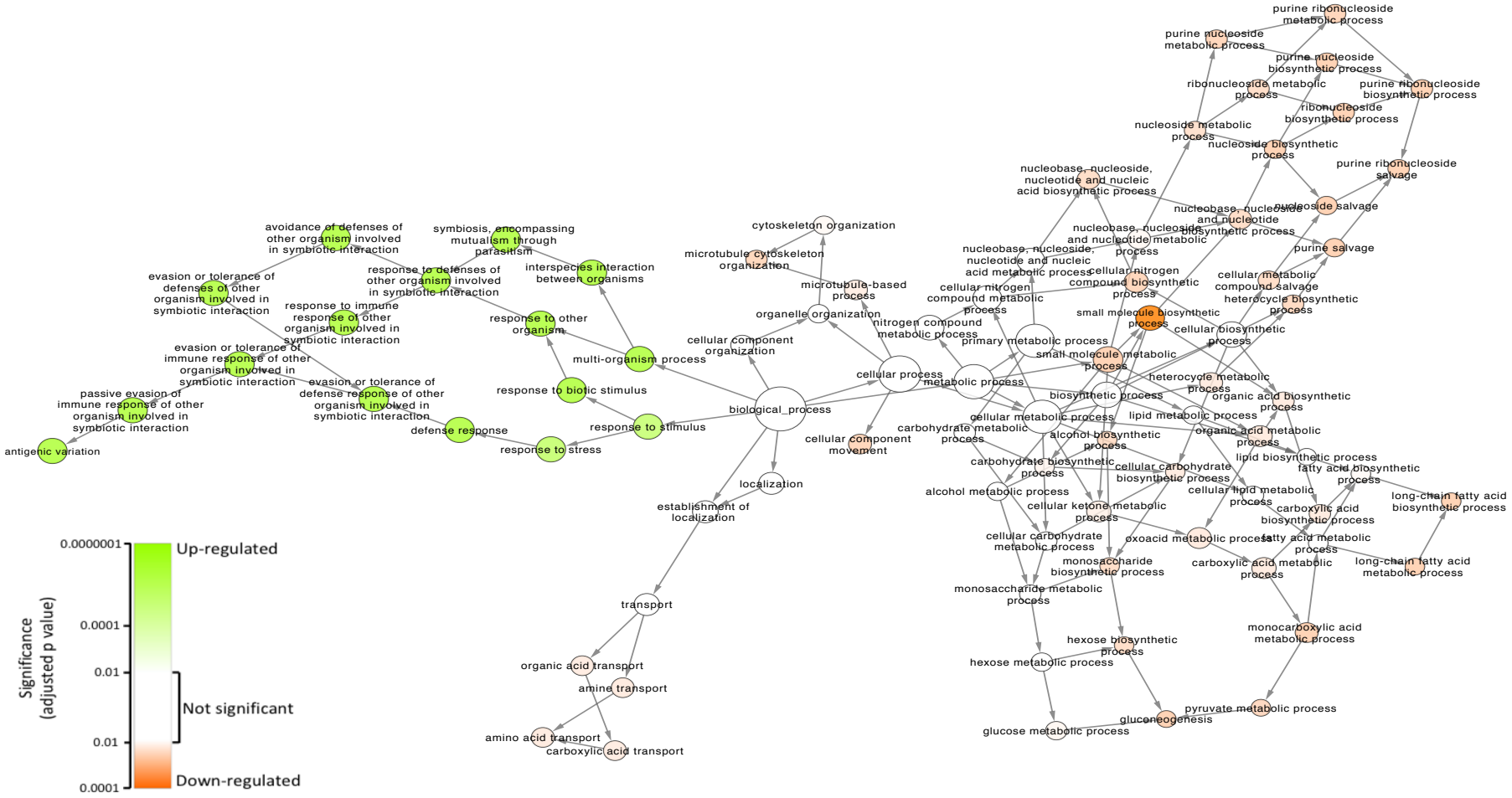



**Figure 5.15 Gene ontology analysis reveals antigenic variation is up-regulated and small molecular biosynthesis down-regulated after TbRH2A knockdown.**

Biological process GO terms linked to genes significantly up- or down-regulated after 36 hr induction of TbRH2A RNAi, as well as parent terms, are plotted as a network. Those enriched in up-regulated and down-regulated genes are coloured green and orange, respectively. Colour intensity indicates significance as adjusted p values. Values > 0.01 are not significant (white).

## 5.5 Discussion

In this chapter the data provide several insights into R-loop and RNase H2 function in *T. brucei*. The first insight is the relatively rigid nature of R-loop localisation across the *T. brucei* genome. DRIP-seq mapping of RNA-DNA hybrids across the Mb chromosomes reveals surprisingly little change in the structures' enrichment after tet-induction of TbRH2A depletion, despite the pronounced growth defects that are seen and the severe damage accumulation. Specifically, similar numbers of DRIP enriched regions located within the RNA Pol II transcribed PTUs are seen before and after RNAi, and inside these regions R-loops are found predominantly around PASs, without pronounced change after loss of TbRH2. Furthermore, the same widespread R-loop localisation in the flanking regions of CDSs is also observed in both WT and *Tbrh1-/-* parasites. This highly reproducible signal pattern highlights that R-loops consistently form across these regions and are not readily displaced, although the mechanism linking PASs and R-loops cannot be elucidated at the point. Although it is likely that RNase H2 is able to target these intergenic R-loops, along with RNase H1, this cannot be fully concluded from the data presented, since the R-loops appear to be detected in similar numbers at these regions in both TbRH2A RNAi induced and un-induced samples, as well as in *Tbrh1-/-* mutants.

However, it should be noted that the necessary scaling of library sizes during normalisation of IP coverage to input coverage, means any global increases or decreases in R-loop formation within the sampled populations may be masked. Indeed, the addition of spike-in chromatin has been used elsewhere to allow normalisation of input and IP library sizes using the sequence depth of alignments to an exogenous genome instead. This additional control enabled the detection of globally reduced histone H3 lysine-79 di-methylation (H3K79me2) and H3K27 tri-methylation (H3K27me3) in human cells treated with inhibitors of methyltransferases DOT1L and EZH2, respectively (Orlando *et al.*, 2014; Egan *et al.*, 2016). Since inter-CDS sequences are the major site of R-loops in the *T. brucei* genome, undetected global changes may similarly be masked here. Hence, a direct quantitative comparison between the WT, TbRH2A RNAi and *Tbrh1-/-* DRIP mapping cannot confidently be made without the additional of spike-in controls that would

allow libraries to be scaled to an unaffected factor. Masking of global changes may in fact explain why both tet – and tet + differ from WT mapping, presumably due to leaky RNAi expression, yet few differences are observed between induced and un-induced samples. Nonetheless, it should be noted that clear and localised changes in deposition of γH2A can be detected by ChIP, and that these changes correlate with potentially altered R-loop abundance at TSSs, perhaps indicating disruption of transcription.

Despite the above concerns about DRIP-seq quantification, regions that do or do not form R-loops in the different cells can be compared between DRIP samples via the presence or absence of signal peaks, irrespective of their enrichment level. In this case, several regions house a greater number of R-loop forming sites (DRIP enriched regions) in both the tet- and tet + parasites compared with WT. These regions include the subtelomeric VSG arrays, RNA Pol I transcribed genes and RHS-associated genes. Interestingly, the same regions show increased DRIP enriched regions after null mutation of TbRH1, and so both classes of *T. brucei* RNase H are likely to target R-loops at these sites. This is most intriguing for the VSG array regions and RNA Pol I transcribed genes, which includes the BESs. This finding is investigated more thoroughly in chapter 6. Also in common with *Tbrh1-/-* DRIP-seq analysis, is the apparent reduction of R-loop formation over RNA Pol III transcribed tRNA and snRNA genes. However, DRIP-qPCR analysis revealed RNA-DNA hybrids are more likely increased after TbRH1 loss and that normalisation of DRIP-seq library sizes masked such an increase. It may be the case that hybrids also increase over tRNA and snRNA genes after depletion of TbRH2A, but this requires investigation with DRIP-qPCR and spike-in controls. Importantly, as all these changes are observed after loss of either RNase H type, R-loop-related defects in these regions are unlikely to explain why TbRH2A is essential and TbRH1 is not.

As discussed in the introduction, the evidence that RNase H2 is essential in *T. brucei* differs from the non-essential nature of the enzyme complex in yeast and is perhaps more comparable to mammals. However, is the basis for the importance of RNase H2 the same in both cases? It seems unlikely that the essentiality of TbRH2 lies in a role in either replicating or segregating the genome. R-loops have been

implicated in initiation of DNA replication (Lombraña *et al.*, 2015), but DRIP-seq could not distinguish patterns of R-loops in SSRs regions that act as origins in *T. brucei* compared with those that do not (Tiengwe *et al.*, 2012). Furthermore, loss of TbRH2A did not appear to impede uptake of EdU, arguing for continued DNA replication in the absence of the RNase H enzyme. R-loops were also found abundantly at centromeres, perhaps arguing that they play are role in chromosome segregation (Kabeche *et al.*, 2018), but potentially increased abundance was seen after TbRH2A RNAi and in TbRH1 null mutants, despite the differences in importance of the RNase H enzymes. The clearest set of data that indicates a potentially specific and critical role for TbRH2 is that which has linked the TbRNase H2 complex to R-loop dynamics specifically at RNA Pol II transcription initiation sites. DRIP-seq mapping revealed a loss of peaks at TSSs within divergent and head-to-tail SSRs. Importantly, this reduction is unique to regions of transcription initiation, not termination, and is seen irrespective of whether or not the SSR is also an ORI. Upon first consideration this finding appears counterintuitive, since loss of the action of an RNase H protein that processes R-loops might be expected to cause an increase in R-loop abundance. Indeed, loss of R-loop formation at the TSSs may be an artefact of normalisation if global changes have occurred after tet-induced RNAi. However, several data argue against this concern. Firstly, the R-loop decrease after TbRH2 RNAi is unique to the Pol II TSSs within the SSRs; at most other locations in the genome DRIP-seq signal enrichment remains comparable between tet – and tet + samples (and in some circumstances shows an increase compared with WT), and the anomalous decreases seen at RNA Pol III genes is common to both tet -and tet + samples relative to WT. Secondly, γH2A mapping via ChIP-seq reveals TSSs as highly discrete, prominent sites of putative DNA lesions specifically after TbRH2A knockdown. Lastly, data discussed in chapter 4 clearly links R-loop formation with RNA Pol II transcription initiation sites, as WT DRIP-seq signal correlates with epigenetic markers of transcription initiation (H2A.Z, H2B.V, H4K10ac and BDF3 binding; Figure 4.21).

Taken together, the DRIP-seq and γH2A ChIP-seq data suggest a highly localised association between R-loops and DNA damage at TSSs. It is hard to imagine such a localised effect is due to increased incorporation of ribonucleotides due to TbRH2

loss, and instead the altered pattern of DRIP-seq signal argues for changed processing of R-loops leading to damage. As DRIP-seq was performed after 24 hr of TbRH2A knockdown, yet γH2A binding at TSSs doesn't appear to occur until 36 hr of knockdown, it may be the case that DNA lesions form after R-loops have been disrupted in some manner at the TSSs. There are several plausible explanations for this, all of which assume RNase H1 is unable to efficiently target initiation-associated R-loops. In the absence of the RNase H2 complex it may be the case that initiation-associated R-loops become fragile sites, vulnerable to breakage, due to action of other R-loop processing enzymes. For example, flap endonucleases, which are able to target the structures at either end of R-loops, might generate either single- or double-stranded DNA breaks (Sollier *et al.*, 2014). A second possibility is that persistent R-loop formation behind an RNA Pol II primed for transcription could lead to transcriptional stalling. This structural change to the local chromatin may alone be sufficient to cause DNA breakage, or the stalled transcription machinery might pose a significant barrier to DNA replication forks. As discussed in section 1.4.2.4, R-loops are known to accumulate at sites of transcription-replication clashes and such clashes have been shown to generate single- or double-stranded breaks (Hamperl *et al.*, 2017; Lang *et al.*, 2017). One reason the SSRs may be a pronounced site of such collisions in *T. brucei* is the known co-localisation of ORIs with SSRs (Tiengwe *et al.*, 2012). This is complicated, however, because although all SSRs appear to bind at least one component of the Origin Recognition Complex, not all SSRs are activated as origins. Despite this, the DRIP-seq data suggests all TSSs accumulate damage, not simply those that are ORI-active. In addition, damage is not seen at convergent SSRs, and ORI activity has been described at a small number of these SSRs. All of these scenarios lead to destruction of the RNA-DNA hybrid and subsequent generation of DNA lesions, perhaps explaining the reduced levels of R-loops, as other activities are recruited to resolve the problems.

γH2A mapping also reveals lower level, sporadic DNA damage across the Pol II PTUs. If TbRNase H2 does indeed have a role in resolving R-loops generated behind transcribing polymerases, it is plausible that similar events generate DNA damage behind RNA Pol II units that have become stalled within the PTU, like at the TSSs. Experiments assessing RNA Pol II movement across the PTUs in parasites depleted of

TbRH2A may help to investigate this further. It is also likely that ribonucleotide load is increased in parasites depleted of TbRH2A, as demonstrated in RNase H2-null mice (Hiller *et al.*, 2012; Reijns *et al.*, 2012). Although this requires experimental investigation, an increased load of ribonucleotides may result in DNA damage across the heavily transcribed genome of *T. brucei*. However, a lack of a functional RER pathway is unlikely to explain such localised damage to the RNA Pol II TSSs.

Gene expression profiling via RNA-seq revealed perhaps surprisingly few changes in the transcriptome of TbRH2A depleted cells. However, two observations can be made. Firstly, changes in VSG and ESAG expression highlight a role of R-loops in VSG switching; a finding explored in more detail in chapter 6 of this thesis. Secondly, nucleoside biosynthesis and salvage pathways are predicted to be down-regulated after TbRH2A depletion, in particular purines pathways. The most likely reason for this is a negative feed-back mechanism trigged by excess nucleotides or ribonucleotides in the cells (Berg *et al.*, 2002). Indeed pyrimidine biosynthesis and salvage pathways are well documented in trypanosomatids (Tiwari and Dubey, 2018), and negative feedback has been shown to operate in *Crithidia fasciculata* where UTP, UDP and CTP inhibit the first enzyme of the *de novo* pyrimidine biosynthesis pathway, carbamoyl phosphate synthetase II (CPSII) (Aoki and Oya, 1987). In contrast, *T. brucei* like other trypanosomatids, lacks *de novo* purine biosynthesis pathways and relies upon purine salvage (Ceron *et al.*, 1979; Berens, Krug and Marr, 1995; Boitz *et al.*, 2012). Interestingly, excess cytosolic DNA is also detected in mouse cells carrying a mutation of the RNase H2A subunit (Pokatayev *et al.*, 2016). In both cases, the question remains, where does the excess nucleic acid originate? In *T. brucei*, GO terms involving both nucleotide and ribonucleotide metabolic pathways are enriched within the genes down-regulated after TbRH2A depletion. Indeed, both may originate from the aberrant breakdown of R-loops in absence of the RNase H2 complex.

Together, the data in this chapter highlight a potential role of RNase H2 in the turnover of R-loops specifically associated with RNA Pol II transcription initiation. Although this possibility would be an exciting avenue of investigation in the future,

the involvement of the RER pathway must first be investigated. It remains possible, if not highly probably, that ribonucleotide load into the DNA genome of *T. brucei* parasites lacking TbRH2A is increased, which would be strongly predicted to cause some level of disruption. Indeed, PCF *T. brucei* cells lacking uracil-DNA glycosylase (UNG), which specifically removes uracil from the DNA genome, display hypermutation rates due mainly to GC to AT transitions and insertions as a consequence of increased uracil load (Víctor M. Castillo-Acosta *et al.*, 2012). Although loss of UNG alone is well tolerated, combined depletion of deoxyuridine 5'-triphosphate pyrophosphatase (dUTPase) that is responsible for removing dUTP from the deoxribonucleotide pool, causes proliferation defects and higher mutation rates in BSF parasites (Víctor M Castillo-Acosta *et al.*, 2012). However, embedded ribonucleotides are unlikely to lead to the localised DNA lesions observed in TbRH2A depleted *T. brucei* and so it remains likely that impaired R-loop processing due to loss of TbRH2A is the primary cause of the localised DNA damage observed here.

# 6 R-loop formation can drive VSG switching

## 6.1 Introduction

As discussed in section 1.3, VSG switching can occur via a number of routes, be they transcription- or recombination-based, and a number of factors have been implicated in the processes (reviewed by McCulloch, Morrison, and Hall 2015). However, the events that trigger such switching are unclear. Several studies support a role for DSBs in initiating VSG switching via recombination, the foremost of these involving the insertion of an endonuclease, ISceI, recognition sequence with the BESs and expression of ISceI to experimentally induce DSBs (Boothroyd *et al.*, 2009). Increased levels of switching through gene conversion were subsequently observed. Interestingly, the elevated levels of switching were dependent upon location of the ISceI recognition site to between the 70-bp repeats and VSG coding regions, and deletion of the 70-bp repeats abolished switching induction, though paradoxically a later study suggested the opposite: the loss of the 70-bp repeats increased VSG switching (Hovel-Miner *et al.*, 2016). Additionally, the use of ligation-mediated PCR in WT parasites lacking ISceI recognition sequences, led to the detection of DSBs within the 70-bp repeats of the actively transcribed BES. DSBs were less detectable within the inactive BESs and ESAGs (Boothroyd *et al.*, 2009). This study therefore implied both the 70-bp repeats and active transcription by RNA Pol I as key factors for VSG switching via gene conversion. A later study by Glover and colleagues added to, and somewhat clouded this picture, showing by ligation-mediated PCR that DSBs are not limited to the 70-bp repeats or the active BES, but confirming that ISceI-mediated DSB formation leads to induction of VSG switching when the lesions were positioned near the 70-bp repeats, and not when at the BES promoter or downstream of the VSG (Glover, Alsford and Horn, 2013).

It is therefore considered possible that, as well as providing homologous sequence, the 70-bp repeats may have an active role in promoting DSBs leading to VSG switching. How such a pathway could operate, however, is less clear. Genetic manipulation of the 70-bp repeat regions in the ISceI expressing *T. brucei* cell line led to the hypothesis that 70-bp repeats even have a role in VSG selection (Hovel-Miner *et al.*, 2016). DSB induction in the active BES containing a shorter repeat

region led to recombination using other BES-housed VSGs as the donor sequences for gene conversion, whereas switching events in cells containing longer 70-bp repeat sequences occurred more frequently with VSGs associated with mini-chromosomes or undetermined sites (Hovel-Miner *et al.*, 2016). It is worth noting that these studies rely on artificially induced VSG switching and, indeed, deletion of the 70-bp repeats was not found to impede low levels of VSG switching observed during *in vitro* culture of unperturbed *T. brucei* (McCulloch, Rudenko and Borst, 1997).

In either case, the proposed model of 70-bp repeat induced VSG switching raises the question of how the repeat region may elicit formation of a DSB. One possibility is that an endonuclease enzyme cleaves this region (Barry, 1997), yet no such native endonuclease has been discovered to date. Alternatively, as active transcription appears to be required, is has been suggested that transcription-associated breakage might trigger VSG switching (McCulloch, Morrison and Hall, 2015). Lastly, it has been proposed that the 70-bp repeats may be able to stall replication, leading to fork collapse and DSBs (McCulloch, Morrison and Hall, 2015). That replication might be a driver of such a form of initiation is consistent with the active BES, unlike all silent BES, being replicated early in S phase (Devlin *et al.*, 2016), and with the effects of RECQ2 (Devlin *et al.*, 2016), TOP3alpha (Kim and Cross, 2010) and RMI1 (Kim and Cross, 2011) mutation.

Another feature of the BESs that may promote recombination-based VSG switching is their proximity to the telomeric repeats (discussed in section 1.3.3.3). Indeed, ligation-mediated PCR suggested that all BESs are fragile, in particular in regions adjacent to the telomeric repeats (Glover, Alsford and Horn, 2013). In addition, critically short telomeres, at least in telomerase-deficient cells, are associated with higher rates of VSG switching (Hovel-Miner *et al.*, 2012). Interestingly, models of VSG switching induced by the 70-bp repeats or the telomeric repeats are not necessarily mutually exclusive (McCulloch, Morrison and Hall, 2015). However, how each may contribute to *in vivo* switching, and how they may impact upon VSG donor sequence selection here, remains to be investigated.

One way in which transcription events can lead to DSBs is via the generation of R-loop structures (reviewed by Sollier and Cimprich 2015). Elevated levels of R-loop formation are documented in many RNA biogenesis mutants, coinciding with increased DNA damage (see section 1.4.2.2). However, R-loops are also likely to lead to DSBs in unperturbed cells. Largely, it is thought that the R-loop structure alone is insufficient to generate breakage and a secondary event is required. Again, several pathways have been suggested to take part in such events. These include the involvement of ssDNA targeting enzymes, processing by flap endonucleases and clashes with replication machinery (Sollier and Cimprich, 2015).

The ssDNA element of R-loops provides an ideal substrate for specific nucleases, including activation-induced cytidine deaminase (AID). AID promotes the conversion of cytosine residues to uracil specifically within ssDNA, leaving the site sensitive to the action of uracil DNA glycosylase, a base excision repair enzyme. Removal of the uracil base by the glycosylase then results in a DNA lesion. This is known to occur at the Ig locus of mature B lymphocytes, where an R-loop is formed. Subsequent processing of both the non-template and template strands of the structure, results in the DSB required to induce Ig class switching recombination (Basu *et al.*, 2011).

R-loop formation also results in nucleic acid flaps at either side of the hybrid that may be processed by structure-specific endonucleases (Sollier *et al.*, 2014). Human flap endonucleases XPF and XPG (xeroderma pigmentosum types F and G) actively process R-loops that form in RNA-processing or topoisomerase I mutant cells, into DSBs (Sollier *et al.*, 2014). As these enzymes normally generate a ssDNA gap during the nucleotide excision repair (NER) pathway, it is possible that they act similarly at R-loop structures and a second event results in DSBs (Sollier and Cimprich, 2015). Alternatively, the enzymes may detect both DNA strands of the R-loops and directly generate DSBs, as supported by *in vitro* studies showing cleavage of both sense and anti-sense strands by purified XPF and XPG (Tian and Alt, 2000).

Lastly, as discussed in section 1.4.2.4, clashes between transcription-associated R-loops and replication machinery lead to DSBs. Replication has been found to be necessary for R-loop-induced DNA damage in several circumstances. This includes in

yeast THO mutants, where transcription during S-phase is required for the associated hyper-recombination phenotype (Wellinger, Prado and Aguilera, 2006), and in ASF or TOP1 null mutant human cells, where DSBs are dependent upon replication (Tuduri *et al.*, 2009; Gan *et al.*, 2011). Importantly, these models are not mutually exclusive and, in fact, could operate in succession to process an R-loop into a DSB. Indeed, R-loop-induced DNA damage in ASF mutant human cells, which is dependent on replication (Gan *et al.*, 2011), is suppressed by knockdown of XPG expression (Sollier *et al.*, 2014), implying the involvement of both DNA replication and flap endonucleases (Sollier and Cimprich, 2015).

### 6.1.1 Chapter aims

As R-loops have been shown to induce DSBs, in some cases leading to recombination events, the aim of this chapter was to ask if R-loops are able to drive VSG switching in *T. brucei*. DRIP-seq mapping to each BES was used to assess R-loop formation in WT cells and in cells lacking either RNase H1 or RNase H2A. VSG switching was also assessed in these mutant cells via RT-qPCR, immunofluorescence assays and RNA-seq. Lastly, γH2A histone binding across the BESs was investigated via ChIP-seq to ascertain whether R-loop formation in absence of RNase H activity could result in DNA damage, capable of triggering VSG recombination events.

## 6.2 TbRH1 and VSG switching

As discussed in Chapter 3, homozygous knockout *Tbrh1*-/- parasites propagate though the cell cycle at the same rate as WT parasites and so loss of the enzyme does not appear deleterious to BSF *T. brucei*. Deletion of the gene is, however, associated with increased levels of R-loops, as demonstrated by DRIP-seq and DRIP-qPCR (Chapter 4), including within the centromere repeats, rRNA loci, and potentially across the RNA Pol II transcribed PTUs. Classification of DRIP enriched regions additionally revealed an increase in R-loop formation within the VSG array regions of the Mb chromosomes in *Tbrh1*-/- parasites. In this section, DRIP-seq mapping in the BESs was analysed in WT and *Tbrh1*-/- cells, as well as levels of VSG switching and BES-associated DNA damage.

### 6.2.1 BES-associated R-loop formation in the absence of TbRH1

DRIP-seq reads for WT and *Tbrh1*-/- cells were mapped across all 14 distinct BESs previously identified and sequenced by Hertz-Fowler et al. (2008). To ensure Illumina reads were only mapped to the correct BES, MapQ filtering was applied (section 2.8.4.1) as described by (Hutchinson, Glover and Horn, 2016) to remove all multi-mapping reads. Little WT DRIP-seq signal was observable across any BES in WT cells, although signal enrichment was clear at the telomere-proximal terminus of each BES contig (Figure 6.1). These peaks of DRIP-seq signal likely represent telomeric R-loops, or TERRA (Cusanelli and Chartrand, 2015; Nanavaty *et al.*, 2017). In contrast, multiple DRIP-seq signal peaks were observable across all BESs in *Tbrh1*-/- cells, most notably across the 70-bp repeat regions (purple boxes, Figure 6.1). In contrast with the predominant localisation of R-loops to inter-CDS regions in the RNA Pol II PTUs (Chapter 5), there was less evidence for such discrete localisation in the BES, whether active or silent.

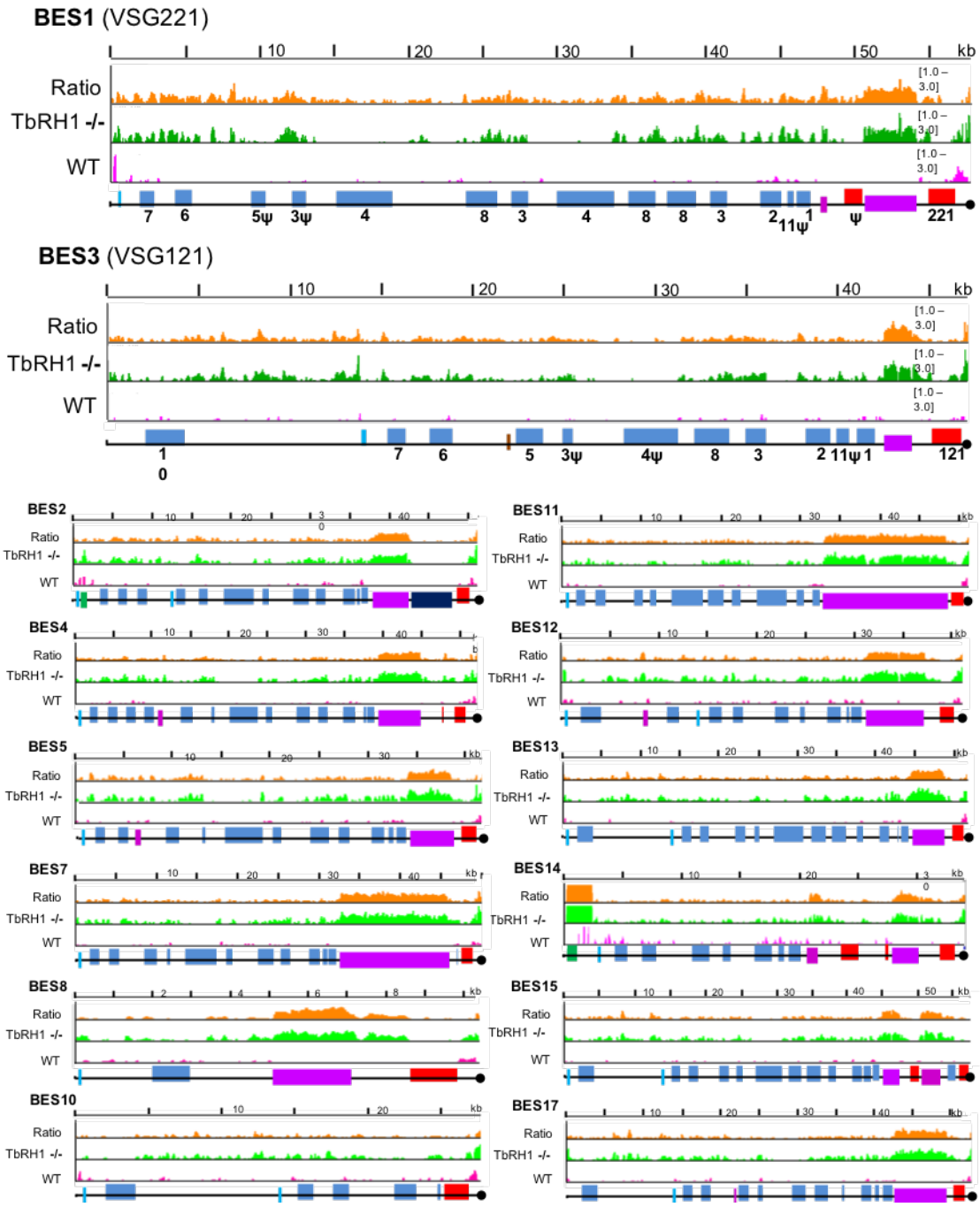



**Figure 6.1 R-loops form across the BESs after TbRH1 loss.**
Localisation of R-loops by DRIP-seq in WT and *Tbrh1-/-* BSF cells. DRIP-seq signal is shown mapped to BES1 (the predominantly active BES of WT cells) and other mainly inactive BESs. Pink and green tracks show normalised ratios of read-depth fold-change (1-3 fold) in IP samples relative to input in WT and *Tbrh1-/-* mutants, respectively, while the orange tracks show the ratio of IP enrichment in *Tbrh1-/-* cells compared with WT. Promoters (aqua), ESAGs (blue, numbered), 70-bp repeats (purple) and VSGs (red) are annotated as boxes; pseudogenes are indicated (ψ), and the end of the available ES sequence is denoted by a black circle.

To validate enrichment of RNA-DNA hybrids within the BESs, DRIP-qPCR was performed on two independent DRIP replicates (Figure 6.2). Primers targeting

ESAG6, ESAG8, VSG221 and VSG121 coding regions were used to amplify sequence in both the input and IP samples. The percentage of each sequence contained within the input that was the recovered during the IP reaction was calculated to indicate levels of DNA-RNA hybrid formation at these regions. In addition, a second IP reaction was set up in parallel, per experiment, and was treated with *E. coli* RNase HI to degrade RNA within the hybrids, and so reduce binding to the S9.6 antibody (section 2.6.3). In all cases, increased input percentage was seen in *Tbrh1-/-* samples compared to WT, and the addition of *E. coli* RNase HI depleted this signal. Thus, the DRIP-qPCR validates the increased DRIP-seq mapping within the BES of the *Tbrh1-/-* mutants (Figure 1.1) and shows this signal is due to R-loops.

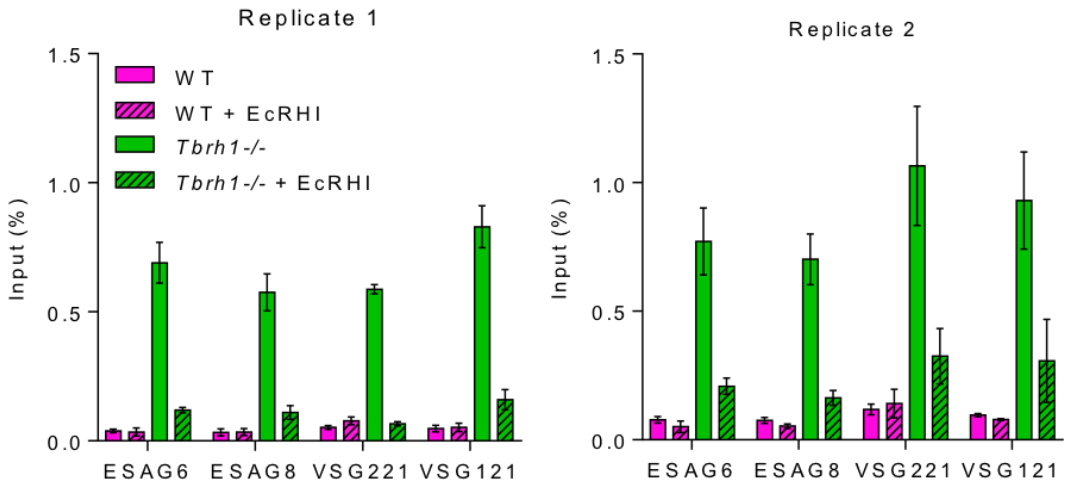



**Figure 6.2 DRIP-qPCR confirms R-loop formation after TbRH1 loss.**
DRIP-qPCR, with or without *E. coli* RNase HI (EcRHI) treatment, showing the percentage of PCR amplification in the IP sample relative to input for WT cells (pink) and *Tbrh1-/-* mutants (green); error bars display SEM for at least three technical replicates and data are shown for two biological replicates (1 and 2).

To analyse patterns of DRIP signal more thoroughly, the BESs were divided into three regions: the ESAG containing region from the 5' end of each BES to the 70-bp repeats; the 70-bp repeats and 500 bp of upstream and downstream flanking sequences; and the VSG, again with 500 bp of flanking sequence (Figure 6.2). In each case, the isolated region of each BES was scaled and DRIP-seq signal plotted per BES as a heatmap for both WT and *Tbrh1-/-* samples (Figure 6.2, upper panels). Additionally, K-means clustering analysis was performed to ascertain whether DRIP-seq signal differed between the normally active BES (BES1) and the other, silent

BESs across each of these three regions. K-means clustering separates samples (in this case each BES) into a specified number of clusters (here two) according to each sample's signal distance from the mean of each cluster. Where BESs were successfully clustered into active (cluster 2, aqua) and silent (cluster 1, blue), the average signal of each cluster was plotted separately as metaplot profiles above the heatmap. Where clustering was unsuccessful, the average signal of all 14 BESs was plotted.

In WT cells, DRIP-seq signal peaks were largely absent across the ESAG-containing portions of each BES (Figure 6.3, left). However, in *Tbrh1-/-* cells there was a broad increase in signal across the region. Interestingly, K-means analysis successfully separated active BES1 (cluster 2) from all other silent BESs (cluster 1), indicating signal differed between these two groups. Separate comparison of the average DRIP-seq signal for WT (purple) and *Tbrh1-/-* (green) samples in the silent and active BESs, indicates a slightly greater increase in RNA-DNA hybrid formation in the active site upon loss of TbRH1 (Figure 6.3 lower panel). However, it is notable that there is only one active BES and 13 silent under consideration, and so significance cannot be confidently assigned. K-means clustering of the 70-bp repeat regions also resulted in separation of the active and silent BESs (Figure 6.3, centre). Here the difference in signal between these groups was more clearly seen, even in the WT DRIP-seq sample. The signal across the silent BESs in the WT sample was notably depleted across the 70-bp repeats, in comparison with the surrounding regions (Figure 6.3, navy). Signal across the actively transcribed 70-bp repeats, however, displayed a slight peak in signal (Figure 6.3, aqua), suggesting low levels of R-loops form here in WT *T. brucei*. In *Tbrh1-/-* cells, hybrid formation clearly increased across the 70-bp repeats of both active and inactive BESs, although the signal profiles were again markedly different. DRIP-seq signal clearly increased across the entire length of the 70-bp repeats after TbRH1 loss in the inactive BESs. However, in the active site DRIP-seq signal increased mainly in the upstream portion of the repeats, before depleting in the downstream portions, and thus following the direction of transcription.

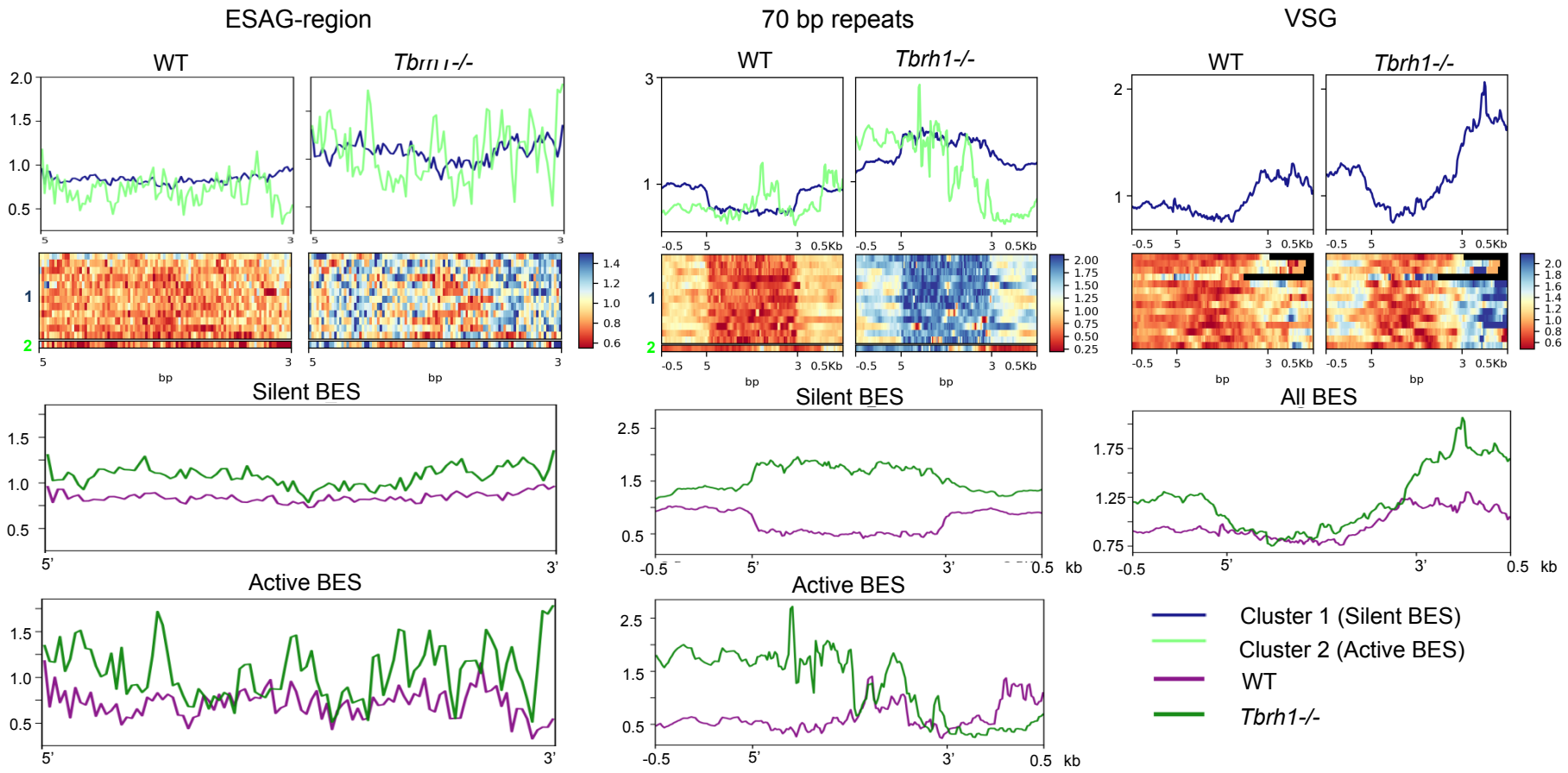



**Figure 6.3 R-loops form most distinctly across the 70-bp repeats.**
DRIP-seq signal fold-change (IP relative to input samples; y-axes) plotted as heatmaps and average signal profiles over BES regions encompassing the ESAGs, 70-bp repeats and VSG; for each region, 5' and 3' (x-axes) denote the upstream and downstream boundaries, and in some cases -/+ 0.5 kb of flanking sequence is shown. Upper two panels: comparison of WT and *Tbrh1-/-* DRIP-seq signal using K-means clustering, which separated the active (light green, cluster 2) and inactive (dark blue, cluster 1) BES when analysing ESAGs and 70-bp repeats, but not VSGs. Lower panels: Overlay of WT (purple) and Tbrh1-/- (green) DRIP-seq signals in the three BES regions, with the active and silent BES displayed separately for the ESAGs and 70-bp repeats.

DRIP-seq signal was not significantly different between the active and inactive BES VSGs in either the WT or *Tbrh1*-/- enrichment profiles, as K-means clustering failed to separate the genes in each analysis (Figure 6.3, right). In all cases, DRIP-seq signal was slightly depleted across the VSG coding regions. In the WT samples, there was signal enrichment downstream of the VSG that, as discussed, is mostly likely due to telomeric R-loops. Upon loss of TbRH1, this telomeric DRIP-seq signal increased, as did signal upstream of the VSG, perhaps due to signal enrichment across the 70-bp repeats (which are positioned 5' to the VSG coding regions).

## 6.2.2 Levels of VSG switching increase after TbRH1 is deleted

To ascertain whether the increased R-loop formation in *Tbrh1*-/- parasites' BESs compared to WT parasites could elicit VSG switching, changes in VSG expression were compared between populations, both at the transcript and protein level.

### 6.2.2.1 Normally silent VSGs are transcribed in *Tbrh1-/-* parasites

RT-qPCR targeting the constitutively active VSG221 and a selection of other, normally silent BES-associated VSGs, was used to ask if changes in VSG transcript levels occur in *Tbrh1*-/- parasites relative to WT. After knockout of *TbRH1*, RNA levels of VSG221 were reduced below that of the WT level (Figure 6.4). In addition, RNA levels of five normally silent VSGs (VSG121, VSG800, VSGT3, VSG13 and VSG224) were increased between ~ 1.25 and 2.25-fold-change relative to WT levels (Figure 6.4), indicating silent VSGs are expressed after TbRH1 loss in a small proportion of the population.

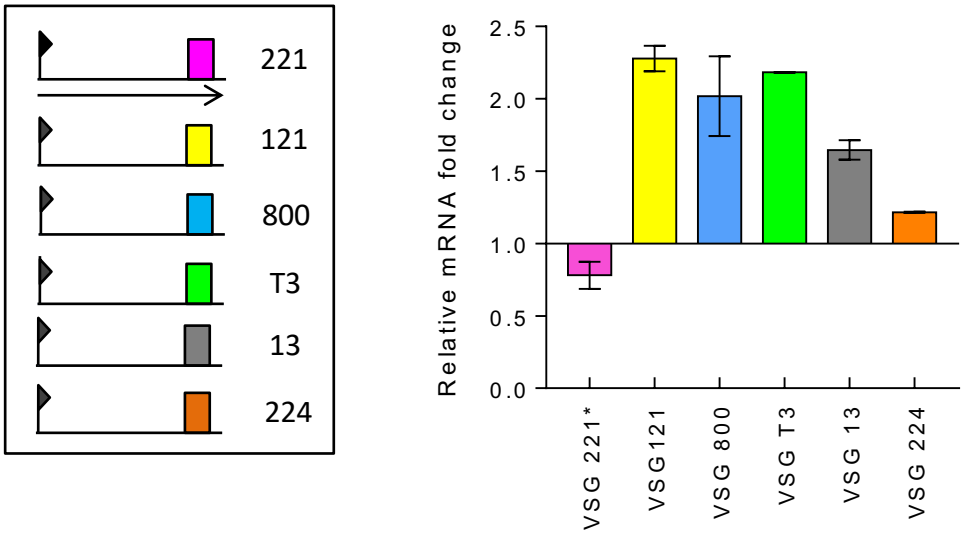
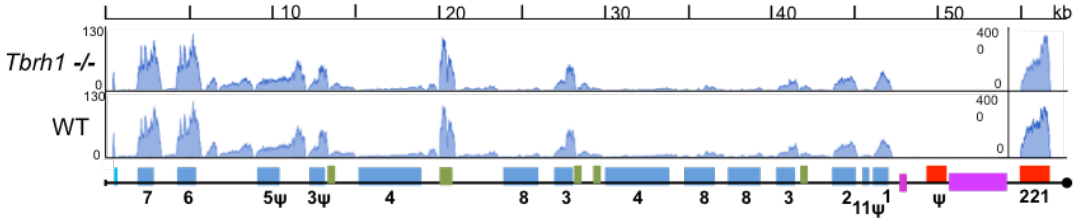


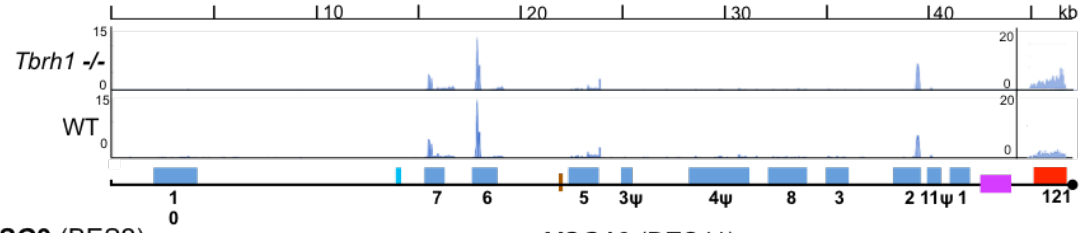

**Figure 6.4 Loss of TbRH1 results in increased transcription of silent BES-associated VSGs.** Left: Simplified diagram of BES used to generate a protective surface coat in the bloodstream *T. brucei* cells; only telomere-proximal VSGs (coloured boxes, numbered) from a selection of BES are shown, and the single BES being actively transcribed (encoding VSG221, pink) is denoted by an arrow extending from the promoter (flag). Right: Graph of VSG RNA levels (corresponding to the BES diagram, and determined by RT-qPCR) in *Tbrh1-/-* cells, plotted as fold-change relative to levels of the cognate VSG RNA in WT cells; * indicates VSG221 is in the active BES of wild type cells, and error bars show SEM for three independent experiments.

To corroborate these findings, RNA-seq was performed using RNA extracted from WT and *Tbrh1-/-* parasites and reads where mapped, again with MapQ filtering, to the 14 BESs (Hertz-Fowler *et al.*, 2008) as well as to a collection of 2470 VSG coding regions documented in *T. brucei* Lister427 (Cross, Kim and Wickstead, 2014). RNA-seq reads are shown mapped to active BES1 and inactive BES3, as well as the VSG coding regions associated with the remaining 12 BESs, in Figure 6.5. Mapping to BES1 showed this BES is the most actively transcribed in the WT and *Tbrh1-/-* parasites, since the greatest numbers of reads, for every gene, were seen here. Mapping to the silent BESs revealed an increase in reads that mapped to some associated VSGs in the *Tbrh1-/-* sample compared to WT cells. This increase was most striking for VSG121, VSG8 and VSG11. Curiously, mapping to VSG224 (in BES7), which increased in expression by RT-qPCR, revealed fewer reads in the *Tbrh1-/-* sample (Figure 6.5), though it should be noted that the populations used for RT-qPCR and RNA-seq are distinct.

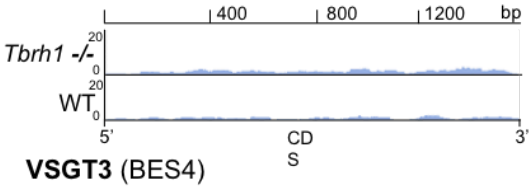
**BES1** (VSG221)



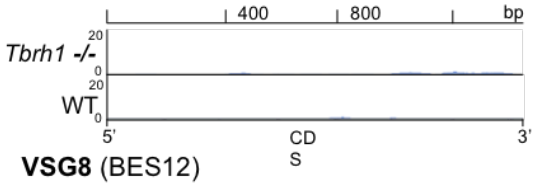

**BES3** (VSG121)



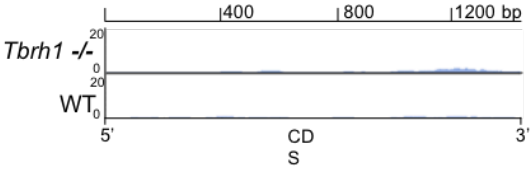

**VSG9** (BES2)



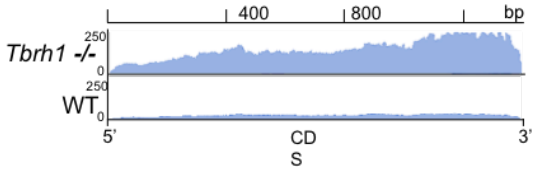

**VSG16** (BES11)



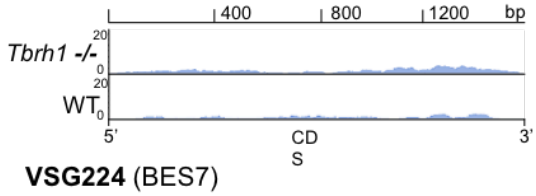

**VSGT3** (BES4)



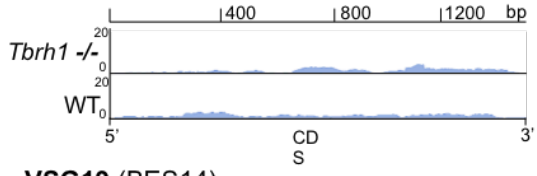

**VSG8** (BES12)



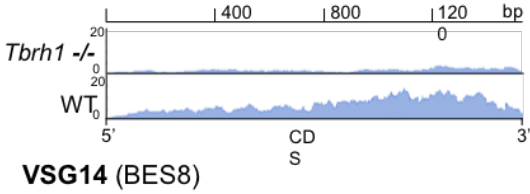

**VSG800** (BES5)



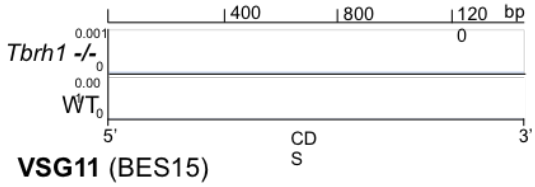

**VSG17** (BES13)



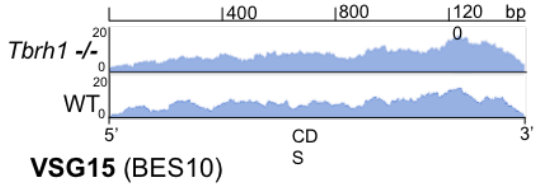

**VSG224** (BES7)



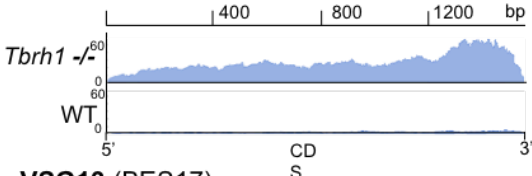

**VSG19** (BES14)



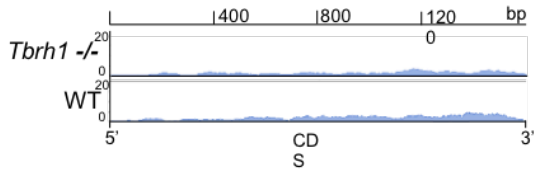

**VSG14** (BES8)




**VSG11** (BES15)




**VSG15** (BES10)




**VSG13** (BES17)

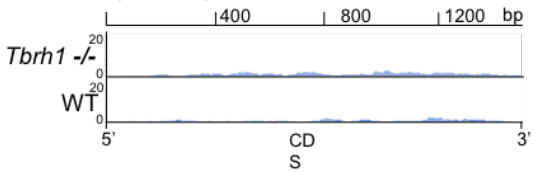


**Figure 6.5 RNA-seq reveals transcription of silent BES-housed VSGs after loss of TbRH1.** Plots of normalised RNA-seq read depth abundance (y-axes) relative to CDS position (x-axes) are shown in WT and *Tbrh1-/-* cells for the complete BES housing VSG221 and VS121, and for just the VSGs housed within the remaining BES identified by (Hertz-Fowler *et al.*, 2008); VSG identity numbers are from (Cross, Kim and Wickstead, 2014).

In some cases, reads were also detected that aligned to silent, non-BES associated VSG coding regions (Figure 6.6). RPKM was calculated for each VSG in the collection and fold-change in RPKM of the *Tbrh1-/-* sample relative to the WT sample was determined in each case (section 2.8.5.4). Of the 2470 VSG coding regions analysed, 65 were found to have an increased RPKM value (> 1.5 FC) in the *Tbrh1-/-* sample relative to WT. Of these, 9 are BES-associated, 2 are housed in MESs, 3 are located in the mini-chromosomes, and 20 within the Mb chromosome arrays. The remaining 31 are predicated to be pseudogenes (Figure 6.6).

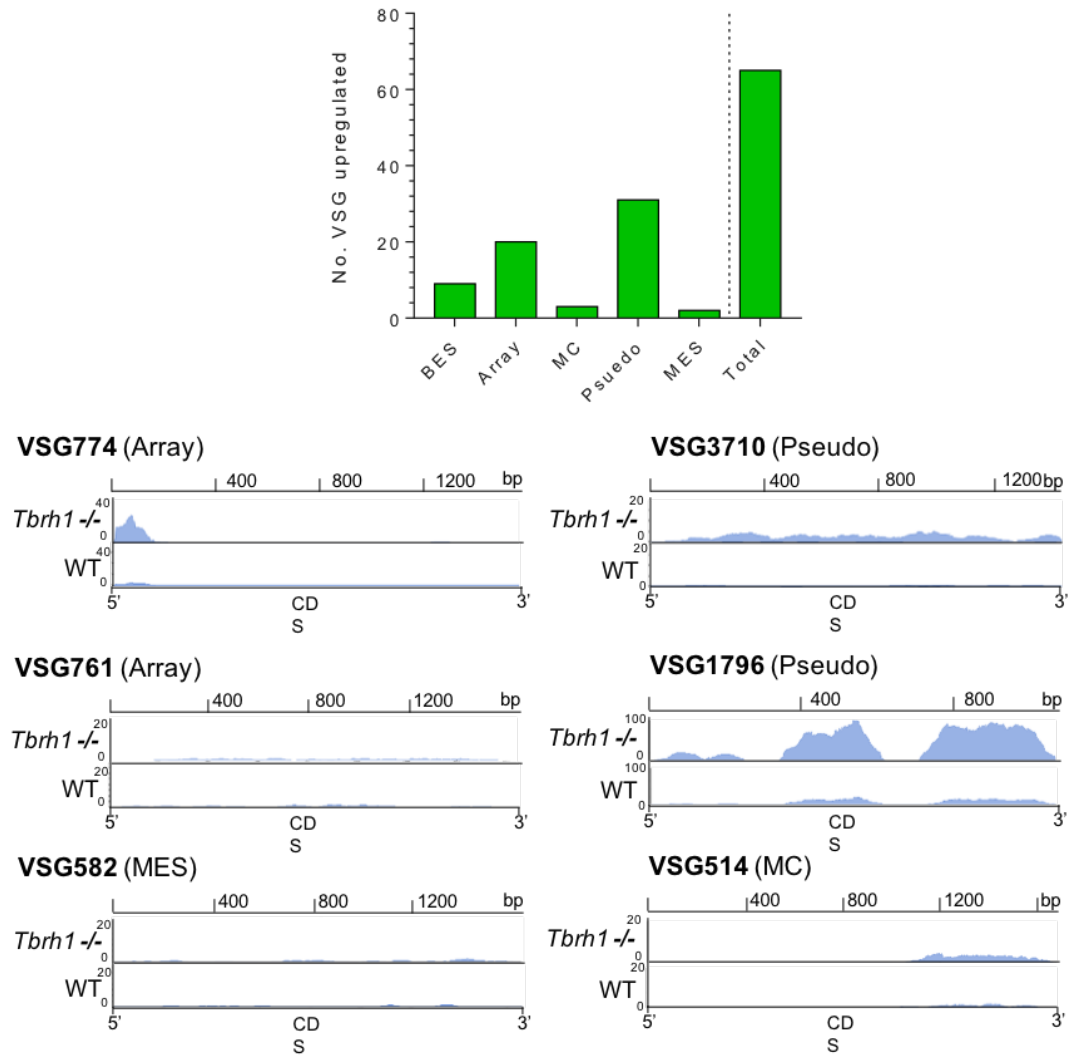



**Figure 6.6 RNA-seq reveals increased transcription of silent VSGs across the repertoire after loss of TbRH1.**
Above: Graph depicting the number of VSG genes that display >1.5-fold increase in RNA abundance (determined by RNA-seq, as RPKM values) in *Tbrh1-/-* cells relative to WT; the total number is sub-categorised depending on whether the VSGs have been localised to the BESs, are intact genes in the subtelomeric arrays (array), are in mini-chromosomes (MC), are pseudogenes (pseudo), or are in a mVSG ES (MES). Below: Plots of normalised RNAseq read depth abundance (y-axes) relative to CDS position (x-axes) for a selection of the above VSGs; VSG identity numbers are from (Cross, Kim and Wickstead, 2014).

### 6.2.2.2 VSG switching frequency is higher in *Tbrh1-/-* parasites

To investigate the stability of (the normally dominant) VSG221 expression in the *Tbrh1-/-* cell line, parasites were cloned to give three clonal lines derived from a single cell (generation 0) of both the WT and *Tbrh1-/-* lines. Immunofluorescent analysis was then performed to assess expression of VSG221 protein coat after ~24 generations (defined as a population doubling) of growth, and at further points over

~45 generations of growth. Cumulative growth was recorded over the time course and showed no difference between the WT and *Tbrh1*-/- parasites (Figure 6.7). >200 parasites for each clone, per time point, were scored for VSG221 expression and VSG221-negative parasites are shown as a proportion of total parasites in Figure 6.7. After ~24 generations *Tbrh1*-/- clonal lines were found to have switched off VSG221 at a frequency of ~0.005. In contrast, a switching frequency of ~0.001 was found for WT parasites. After ~27 generations a higher frequency of switching was observed in *Tbrh1*-/- parasites, ~0.0125, whereas WT frequency remained similar (~0.0005). Switching frequency then remained similar, although switching at generation ~44 of *Tbrh1*-/- was found to be slightly lower, at ~0.007.

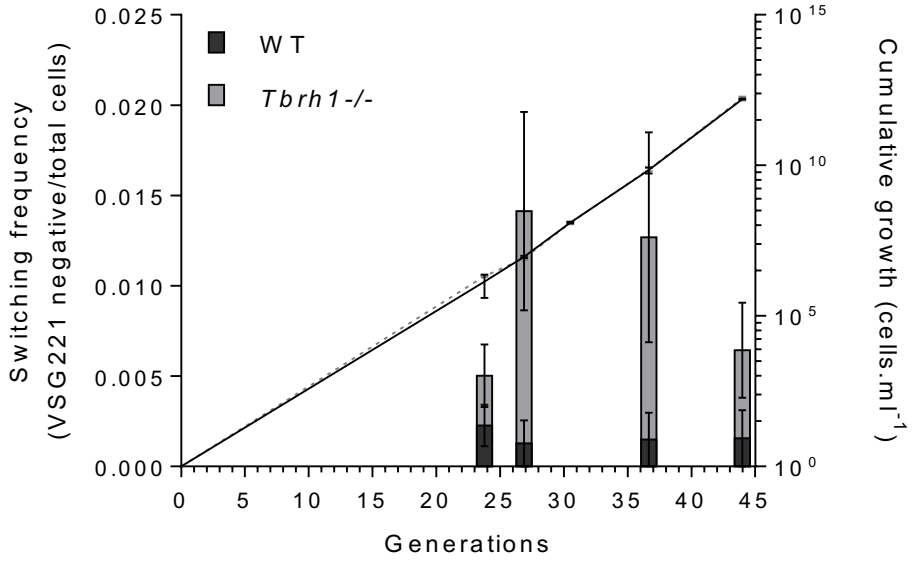



**Figure 6.7 *Tbrh1-/-* parasites switch off expression of VSG221 at a higher frequency than WT parasites.**
Three VSG221-expressing clones of each cell type were generated and serially passaged independently. At multiple time points the number of cells in the populations that do not express VSG221 was assessed by immunofluorescence with anti-VSG221 antiserum; non-VSG221 expressing WT (black) and *Tbrh1*-/- (grey) cells are shown as a proportion of the total population. >200 cells were counted for each clone, at each time point (data shows the average and SD of the three clones). Cumulative density of WT (solid line) and *Tbrh1*-/- cells (dotted line) over the course of the analysis is also shown (average density and SD is shown).

### 6.2.2.3 VSG coat switching is increased in *Tbrh1-/-* cells

Changes in VSG expression were next investigated at the protein level via immunofluorescence. WT 1.2 (the parental line of *Tbrh1*-/- cells, elsewhere referred to as WT in this thesis) and *Tbrh1*-/- parasites were dual stained with

antibodies targeting the WT 1.2 dominant VSG, VSG221, as well as VSG121, which is normally silent in this cell line (Figure 6.8). Another WT *T. brucei* line, WT 1.6, was also dual stained as these parasites have been selected for impaired expression of VSG221 and mainly expresses VSG121 as the dominant VSG (Glover *et al.*, 2007). The percentage of cells expressing one or the other VSG, both VSGs, or neither was calculated in WT 1.2 and *Tbrh1-/-* populations, revealing ~3.12% of the *Tbrh1-/-* parasites did not express either VSG on their surface, whereas parental line WT 1.2 exclusively expressed VSG221 (Figure 6.8B). A small percentage (~0.35%) of *Tbrh1-/-* cells expressed VSG121 on their surface, which was not seen in WT 1.2 cells. Of the *Tbrh1-/-* VSG121-positive parasites, ~25% also expressed VSG221 (Figure 6.8C), indicating either these cells are still in the process of switching VSG protein coats, or monoallelic control of VSG expression has been lost. Of the VSG121 expressing WT 1.6 cells, virtually all parasites expressed VSG121 exclusively.

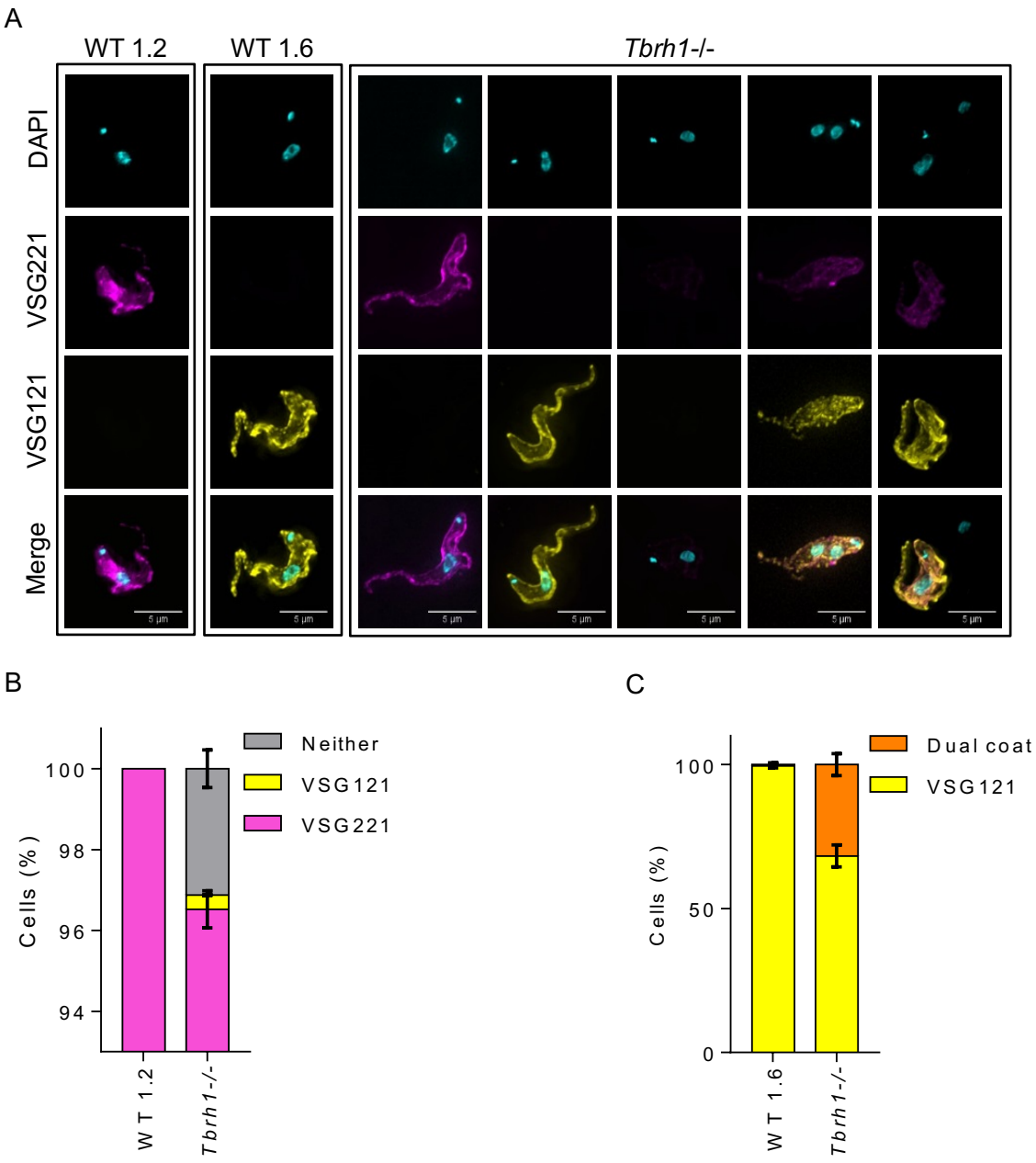



**Figure 6.8 Loss of TbRH1 induces switching of the VSG coat.**
A) Co-immunofluorescence imaging of VSG221 and VSG121 coats. In addition to WT 1.2 cells and *Tbrh1-/-* mutants, an example of a cell is shown from a *T. brucei* strain (WT 1.6) that predominantly expresses VSG121 (Glover *et al.*, 2007) (Scale bar, 5 µm). B) Percentage of WT (WT 1.2) and *Tbrh1-/-* cells expressing VSG221 and/or VSG121 on their surface, as determined by co-immunofluorescence imaging with anti-VSG221 and VSG121 antiserum. The graph depicts the relative proportions of cells in the population in which only VSG221 (pink) or VSG121 (yellow) could be detected, as well as cell with both (orange) or neither (grey) of the two VSG on their surface; >200 cells were analysed for each cell type in each of three replicates (error bars denote SEM). C) Analysis of WT 1.6 and *Tbrh1-/-* mutant cells that express VSG121 on the cell surface, showing the percentages that simultaneously express VSG221 (orange) or only express VSG121 (yellow); >100 cells were analysed in each of three replicate experiments for each cell type.

### 6.2.3 DNA damage levels increase in *Tbrh1-/-* parasites

As DNA damage within BESs is predicted to elicit VSG switching, at least by recombination, levels of the histone variant and DNA damage marker γH2A (Glover and Horn, 2012) were assessed in *Tbrh1-/-* parasites and compared with WT. Deletion of TbRH1 causes no detectible change to growth rate or cell cycle progression (section 3.4), but it is possible that low levels of widespread damage, or highly localised DNA damage may be well tolerated. Western blot analysis of whole protein extracts showed limited differences in the total level of γH2A signal between WT and *Tbrh1-/-* cells (Figure 6.9A). The slight increase in average band intensity of three independent replicates was deemed not significant by a Wilcoxon rank test (Figure 6.9B).

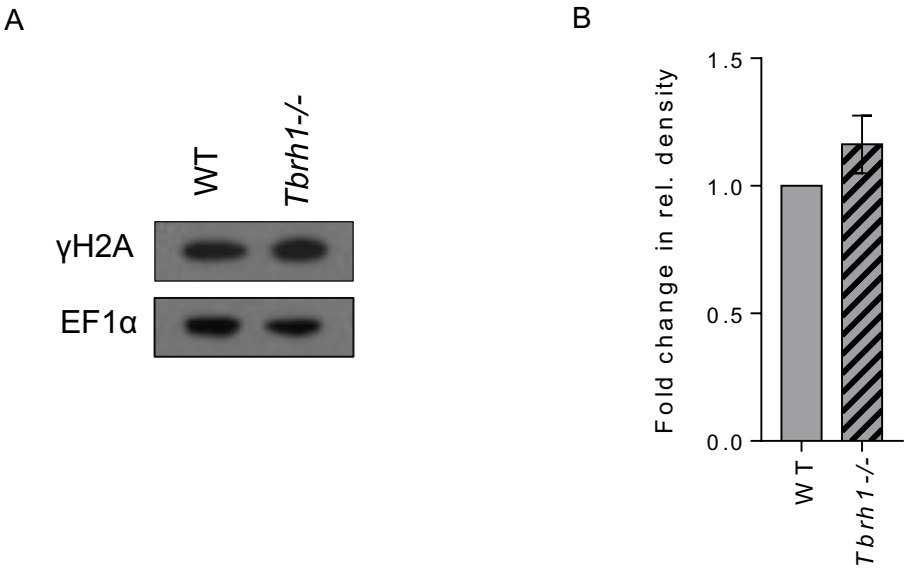



**Figure 6.9 Expression levels of γH2A in *Tbrh1-/-* parasites do not change substantially compared with WT.**
A) Western blot of γ-H2A, detected by specific antiserum, in WT and *Tbrh1-/-* mutants; antiserum detecting EF1-α provides a loading control. B) Relative density of γH2A western blot signal, normalised to EF1-α, is compared in WT (normalised to 1.0) and *Tbrh1-/-* cells.

γH2A was also detected by immunofluorescence using fixed cells. In WT cells ~7% of parasites displayed γH2A staining, in agreement with previous findings by Glover and Horn (2012) (Figure 6.10A and B). In the *Tbrh1-/-* population ~17% of cells displayed γH2A staining, indicating increased nuclear damage. Of these γH2A-positive cells, ~50% contained a single focus of γH2A staining, in both WT and

*Tbrh1-/-* samples (Figure 6.10C). The number of γH2A-positive cells possessing two foci increased from ~20% in the WT population to ~ 33% in the *Tbrh1-/-* parasites. Only very few cells were found to express three foci: ~1% and ~2.7% of WT and *Tbrh1-/-* γH2A–positive parasites, respectively. The remaining population of γH2A-positive cells displayed diffuse staining across the nucleus. This reduced from ~33% of γH2A-positive WT cells to ~18.5% of *Tbrh1-/-* γH2A –positive parasites. Hence, DNA damage most often occurs at defined loci, rather than genome-wide, in *Tbrh1-/-* cells.

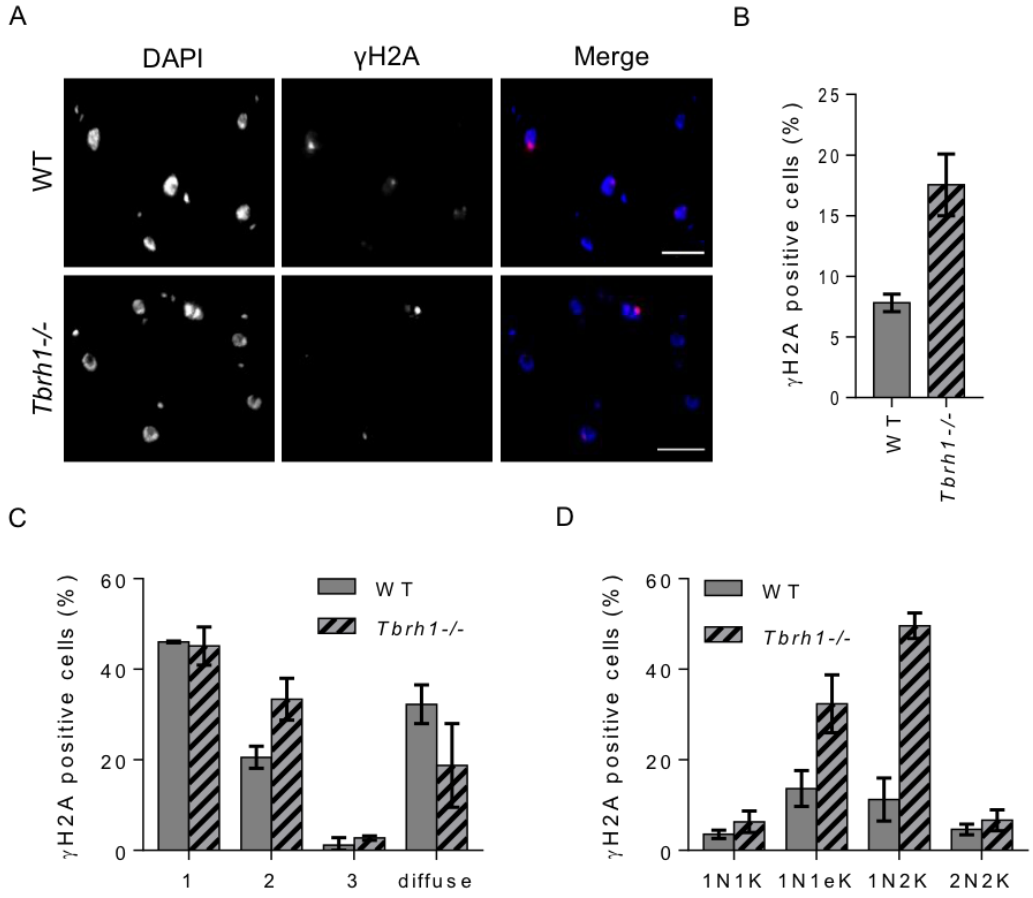



**Figure 6.10 Loss of TbRH1 leads to increased levels of nuclear damage, mainly in replicating cells.**
A) Immunofluorescent imaging of anti-γH2A signal and co-localisation with DAPI staining in WT and *Tbrh1-/-* cells. These are representative images of those analysed in B, C and D. B) Percentage of WT and *Tbrh1-/-* cells (n >200 in each of three replicates) with detectable nuclear anti-γH2A signal. C) Distribution of intra-nuclear γH2A signal in WT and *Tbrh1-/- cells* (n >200 in each of three replicates). Of the γH2A-positive cells those with 1, 2 or 3 foci, or diffuse staining across the nucleus, are shown as a percentage of all positive cells. D) Percentage of γH2A-positive WT and *Tbrh1-/-* cells of each cell cycle stage discernable by DAPI staining (1N1K, 1N1elongated K (1N1eK), 1N2K and 2N2K), expressed as a percentage of total cells counted within each stage. n ≥50 for each cell cycle stage in three replicates. In all cases error bars show SD from the mean.

Counting of γH2A-positive cells was repeated, this time separating parasites according to their cell cycle stage (1N1K, 1N1elongated K (1N1eK), 1N2K and 2N2K). At least 50 cells in each stage were counted and the percentage displaying γH2A staining was calculated for both WT and *Tbrh1-/- T. brucei* lines (Figure 6.10D). A greater percentage of WT cells undergoing nuclear replication (1N1eK) or in in G2-M phase (1N2K) displayed γH2A signal compared to those at the end of M phase (2N2K) or through G1 (1N1K). Analysis of the *Tbrh1-/-* mutants showed that the increased proportion of total cells with γH2A signal (Figure 6.10B) was nearly entirely accounted for by greater numbers of γH2A-positive cells within the 1N1eK (~2.4-fold increase) and 1N2K (~3.6-fold) populations. Hence, γH2A foci form most often in cells which are undergoing, or have recently undergone, DNA replication in the absence of TbRH1. Super-resolution structure-illumination microscopy imaging of anti-γH2A signal and DAPI staining confirmed the discrete nature of γH2A foci in these cells (Figure 6.11).

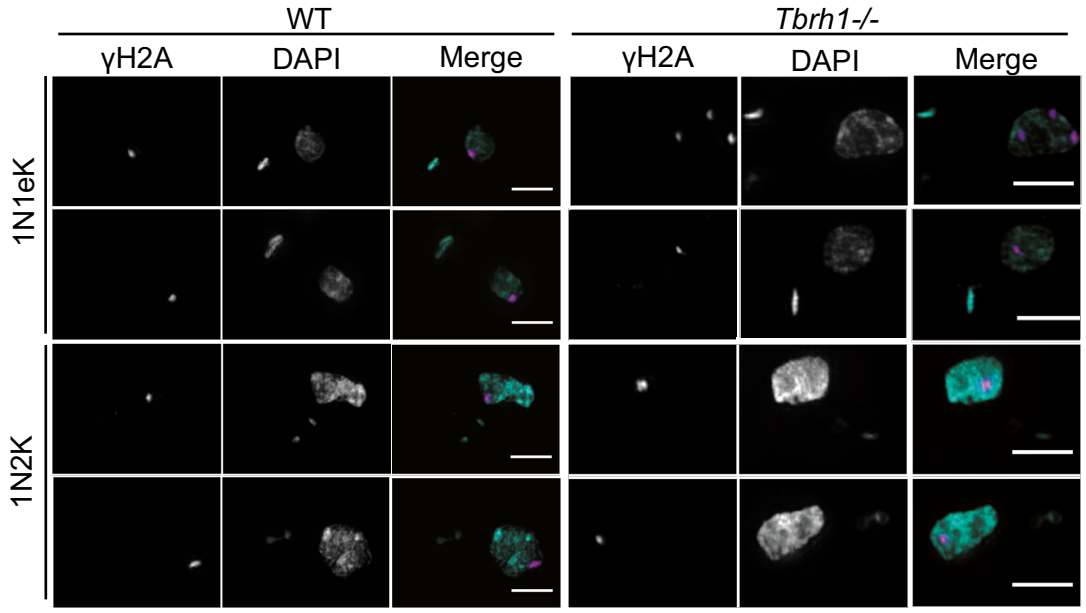



**Figure 6.11 γH2A staining forms discrete foci in WT and *Tbrh1-/-* DNA replication-associated cells.**
Super-resolution structure-illumination microscopy imaging of anti-γH2A signal and DAPI staining of representative replicating (1N1eK and 1N2K) cells. In the merged images of anti-γH2A (pink) and DAPI (aqua), colour us provided. Images were processed by Leandro Lemgruber.

ChIP-seq mapping of γH2A was next employed to ask if the γH2A histone variant localised to the BESs in response to loss of TbRH1. ChIP-seq was performed with WT and *Tbrh1-/-* populations and the fold-change between input and IP sample read

coverage (section 2.8.4.2) was plotted across the 14 BESs (Hertz-Fowler *et al.*, 2008), using MapQ filtering. BES-associated γH2A signal was found almost exclusively in the active BES (BES1) in WT and *Tbrh1*-/- samples (Figure 6.12). The predominantly telomere-proximal signal peaks observed in the WT sample were clearly conserved in the *Tbrh1*-/- ChIP-seq sample, though the peaks were more widespread and also appeared to extend upstream of the 70-bp repeats and VSG. In contrast, only minor increases in γH2A ChIP-seq peaks were evident across the silent BES when comparing the *Tbrh1*-/- cells with WT. Elsewhere in the genome, γH2A ChIP-seq signal was uniformly increased relative to input samples across Pol II transcribed PTUs in both WT and *Tbrh1*-/- samples, most likely due to the antibody binding to non-phosphorylated H2A histone (Figure 8.6).

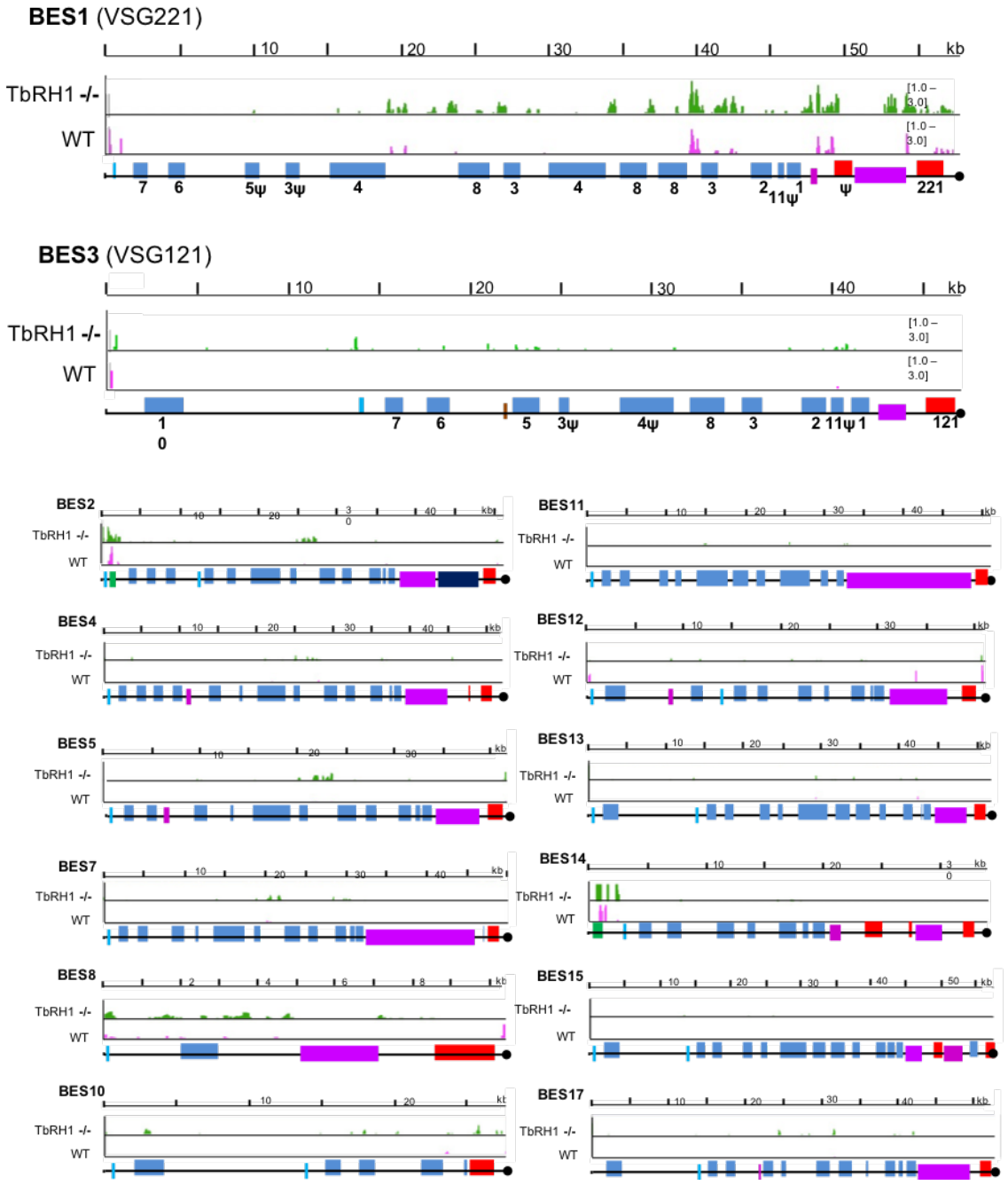



**Figure 6.12 Localisation of γH2A by ChIP-seq in WT and *Tbrh1-/-* cells.**
ChIP-seq signal is shown mapped to the 14 BESs (Hertz-Fowler *et al.*, 2008). Annotations are as described in Figure 6.1. Pink and green tracks show normalised ratios of read-depth fold-change (1-3 fold) in IP samples relative to input in WT and *Tbrh1-/-* mutants, respectively.

## 6.3 RNase H2 and VSG switching

### 6.3.1 *TbRH2A* depletion leads to R-loop formation in the BES

To ask if the VSG effects (described above) seen after loss of TbRH1 are also seen after loss of TbRH2A, DRIP-seq data generated using the TbRH2A$^{RNAi}$ line, after 24 hr

of growth with or without tet-induction, was aligned to the 14 characterised BESs with MapQ filtering and enrichment in coverage in IP samples relative to input was calculated for each. DRIP-seq signal was clearly enriched across the active and inactive BESs relative to WT, both in the tet-induced and un-induced samples (Figure 6.13A), consistent with the RNAi being slightly leaky without tet-induction (Chapter 3). Like DRIP-seq signal derived from *Tbrh1-/-* parasites, signal peaks were clearly seen across the 70-bp repeat regions, as well as across the ESAG containing regions and downstream of the VSGs. Perhaps surprisingly, the peaks observed in the tet-induced sample appeared to be slightly lower than was seen in the uninduced, though it is possible this effect is an artefact of global changes after TbRH2A loss that are masked by data normalisation during mapping (K. Chen *et al.*, 2015).

DRIP-qPCR targeting VSG221 (active BES1), VSG121 (inactive BES3), ESAG6 and ESAG8 coding regions indicated increased levels of RNA-DNA hybrid formation in induced samples compared to un-induced after 36 hr of growth (Figure 6.13B). In all cases, the IP enrichment in the tet + samples was significantly reduced by treatment with EcRHI. However, DRIP-qPCR after 24 hr of growth showed only minor differences in R-loop levels between tet – and tet + samples, for reasons that are unclear, and only in the case of the VSGs did treatment with EcRH1 reduce the enrichment.

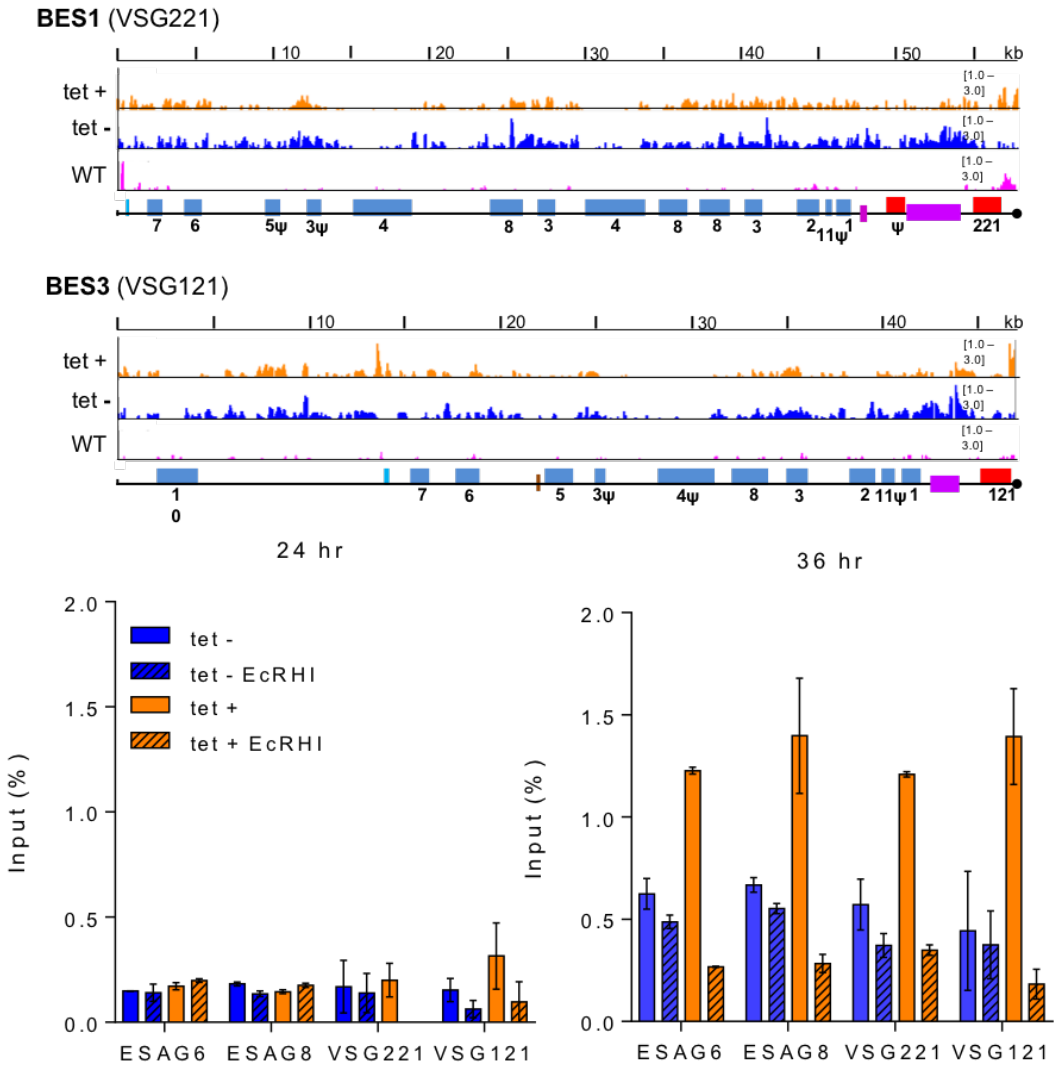



**Figure 6.13 R-loops levels increase across the BESs in cells depleted of *TbRH2A* expression.** A) Localisation of R-loops by DRIP-seq the *TbRH2A* RNAi line grown in the absence (tet -; blue) or presence (tet +; orange) of RNAi induction for 24hr. DRIP-seq mapping in WT parasites is also show (pink). Mapping is shown to BES1 (active site of WT parasites) and BES3 (normally transcriptionally silent site). Annotation of BESs are as in Figure 6.1. B) DRIP-qPCR using primers targeting the sequences of ESAG6, ESAG8, VSG221 (BES1) and VSG121 (BES3), with or without *E.coli* RNase HI (EcRHI) treatment, showing the percentage of amplification in the IP sample relative to input. This is shown for tet induction (tet +) and non-induction (tet -) after 24 (left) and 36 hr (right) of growth. Error bars display SEM for three technical replicates. No VSG221 DNA could be PCR-amplified to detectable levels from 24 hr tet + IP sample after treatment with EcRH1.

## 6.3.2 TbRH2A depletion leads VSG switching in *T. brucei*

### 6.3.2.1 Normally silent VSGs are transcribed after *TbRH2A* depletion

VSG expression changes were investigated via RT-qPCR after 24 and 36 hr tet-induction of RNAi to deplete *TbRH2A* transcripts. Here, no change in VSG221 RNA levels were detectable in induced cells relative to un-induced, both 24 and 36 hr

post-induction (Figure 6.14). In contrast, levels of RNA from all the silent VSGs increased after tet-induction, both at 24 hrs and, even more markedly, after 36 hr of induction, where all silent VSGs increased over 3-fold relative to un-induced samples.

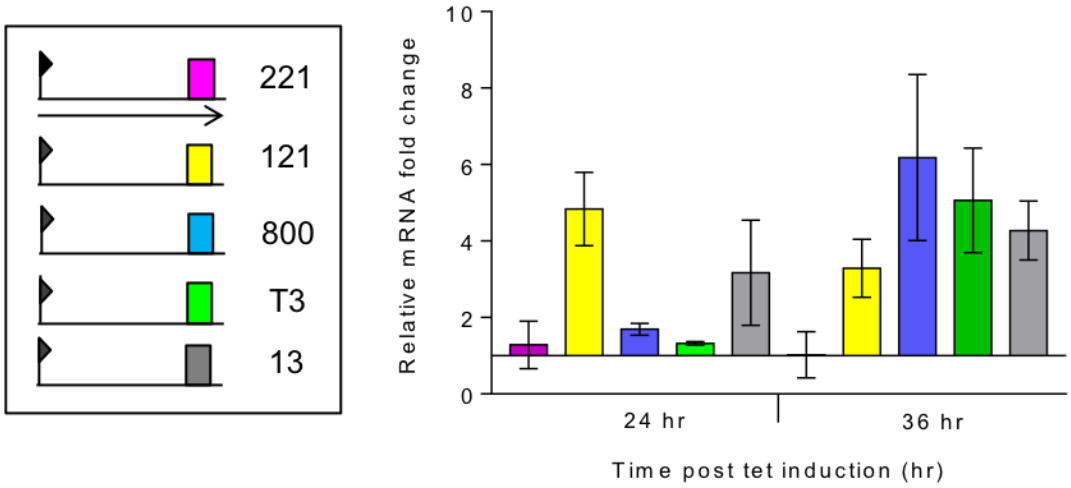



**Figure 6.14 Loss of TbRH2A results in increased transcription of silent BES-associated VSGs.** Left: Simplified diagram of BES used to generate a protective surface coat in the BSF *T. brucei* cells; only telomere-proximal VSGs (coloured boxes, numbered) from a selection of BES are shown, and the single BES being actively transcribed (encoding VSG221, pink) is denoted by an arrow extending from the promoter (flag). Right: Graph of VSG RNA levels (corresponding to the ES diagram and determined by RT-qPCR) of tet-induced TbRH2A$^{RNAi}$ parasites, plotted as fold-change relative to levels of the cognate VSG RNA in un-induced cells after both 24 hr and 36 hr of culture; Error bars show SD for three independent experiments.

RNA-seq data were also aligned, using MapQ filtering, to the 14 BESs (Hertz-Fowler *et al.*, 2008) and VSG coding region collection (Cross, Kim and Wickstead, 2014); in this case total RNA was extracted from the TbRH2A$^{RNAi}$ parasites after 24 and 36 hr of growth with and without tet-induction, with two independent replicates analysed in each case. Mapping to the BES-associated VSGs showed increased read mapping to 11 of the 13 silent VSGs after tet-induction both after 24 hr (Figure 8.7) and 36 hr of induction (Figure 6.15). The predominantly transcribed VSG221, in contrast, showed reduced read coverage at both time points. Curiously, VSG17, which is housed within silent BES13, displayed more mapped reads before induction than after, in both replicates, after 36 hr of growth. In these conditions, unlike in the *Tbrh1-/-* cells, read coverage also increased over the BES coding regions after TbRH2A RNAi, though this effect was predominantly localised to ESAG6 and ESAG7, which are promoter-proximal (Figure 6.15, Figure 8.7).

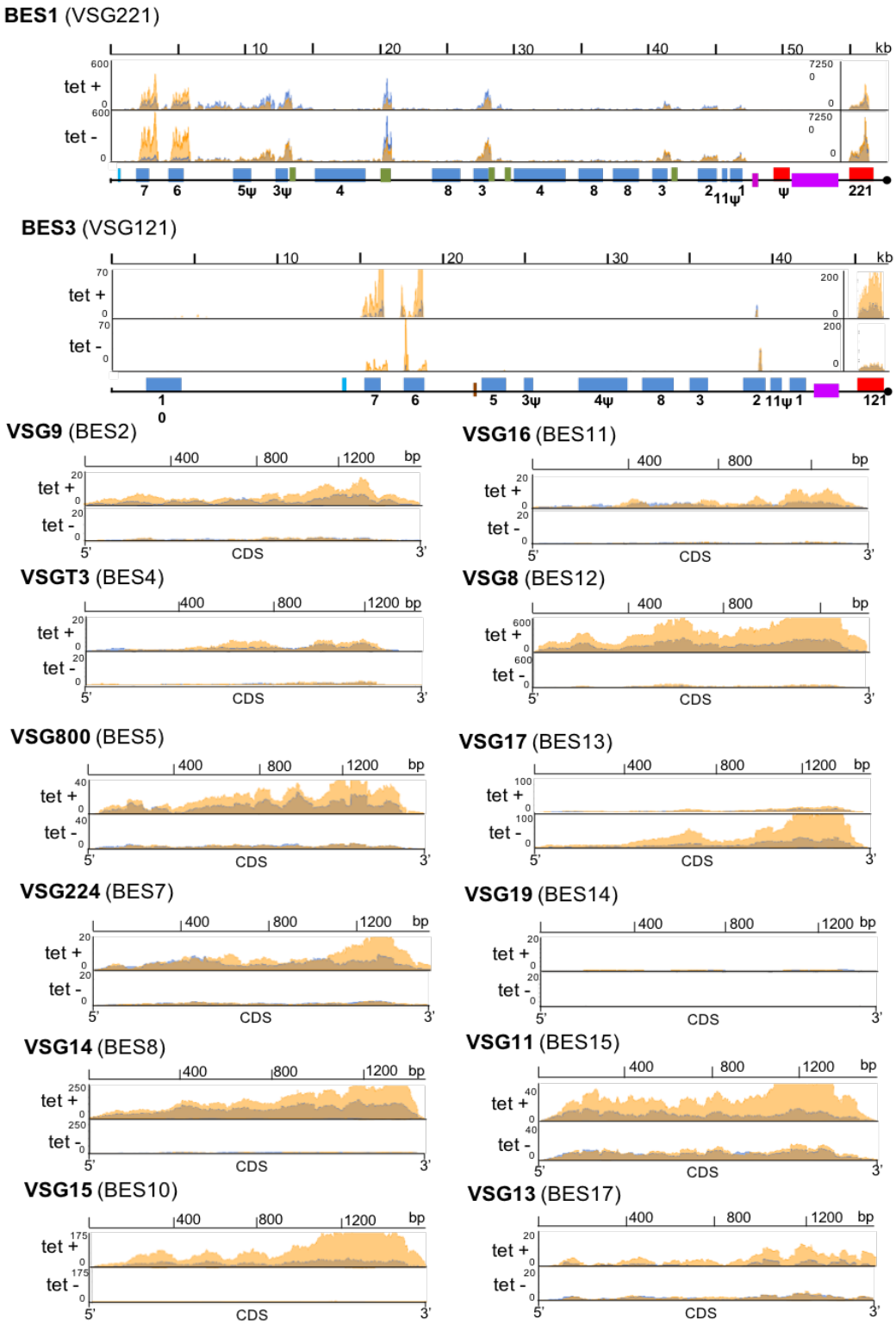



**Figure 6.15 RNA-seq reveals transcription of silent BES-housed VSGs after loss of TbRH2A.** Normalised RNA-seq read depth abundance (y-axes) is plotted for two independent replicates (overlaid orange and blue) of *TbRH2A* RNAi parasites after 36 hr of growth, with (tet +) and without (tet -) induction. Read depth is shown relative to CDS position (x-axes) for BES1 and BES3 as well as a selection of the above VSG housed within BES identified by (Hertz-Fowler *et al.*, 2008); VSG identity numbers are from (Cross, Kim and Wickstead, 2014). For RNA-seq depth abundance after 24 hr of culture see Figure 8.7

As discussed in chapter 5, aligning this RNA-seq data to the *T. brucei* 927 genome and performing differential expression analysis revealed 32 up-regulated genes after 24 hr tet-induction annotated as either VSG, ESAG or procyclin, and after 36 hr of induction 37 VSGs and 14 ESAGs were up-regulated (Figure 5.14). To explore the increase in silent VSG expression further after TbRH2A depletion, differential expression analysis (section 2.8.5.3) was repeated using mapping to the VSG collection of the *T. brucei* Lister 427 stain (Cross, Kim and Wickstead, 2014). 20 VSGs were found to be up-regulated after 24 hr of TbRH2A RNAi induction, and 50 after 36 hr induction (Figure 6.16). Of these, 40% and 60% were classified as pseudogenes (Cross, Kim and Wickstead, 2014) in the 24 hr and 36 hr samples, respectively (Figure 6.16). Of the remaining VSGs with increased expression after 24 hrs, 25% were classified as intact array genes, 20% as intact mini-chromosomal associated, 10% associated with mVSG ESs and the remaining 5% (1 VSG) was BES-associated (Figure 6.16). After 36 hr, a similar proportion of up-regulated VSGs were array-associated (26%), 10% were housed in the BESs and 2% (1 VSG) were mVSG ES-associated (Figure 6.16). No mini-chromosome associated VSGs were significantly up-regulated at this time point. Examples of reads mapping to some of these VSGs are shown for 36 hr post-induction (Figure 6.16).

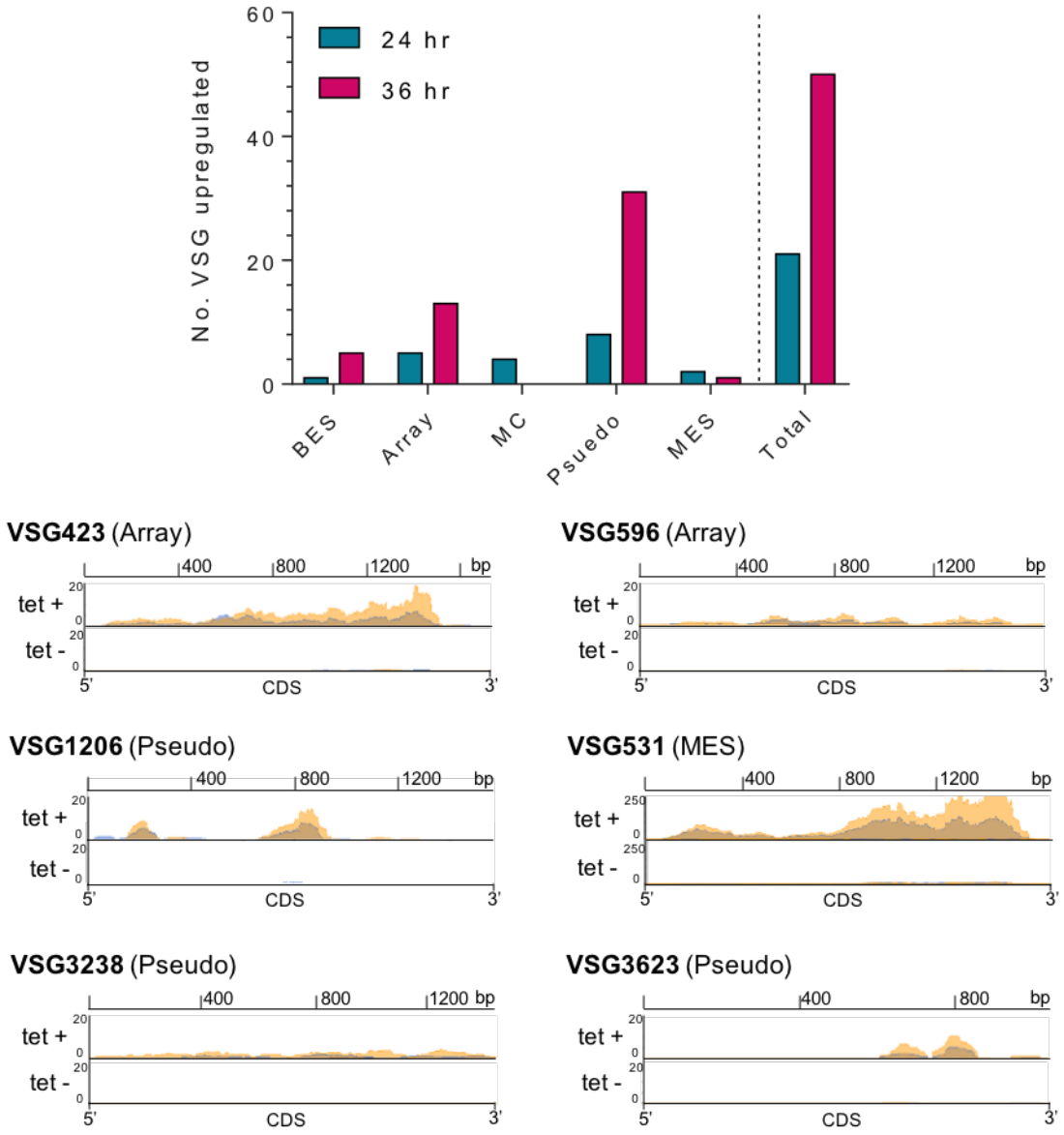
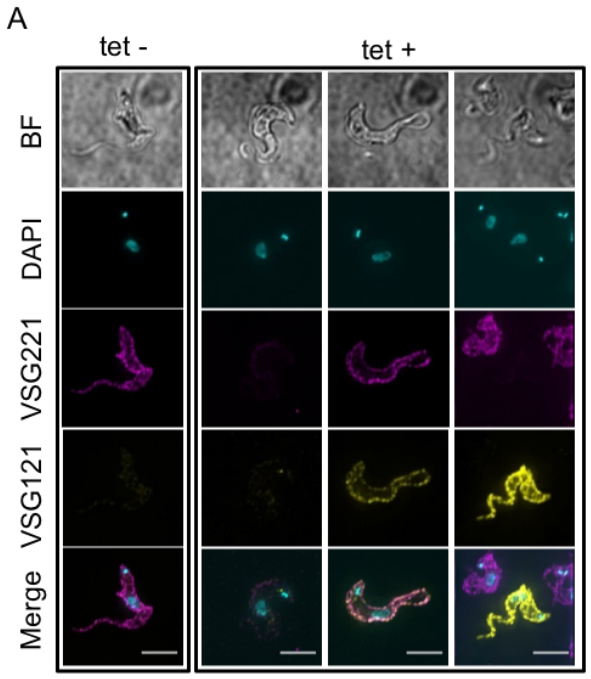


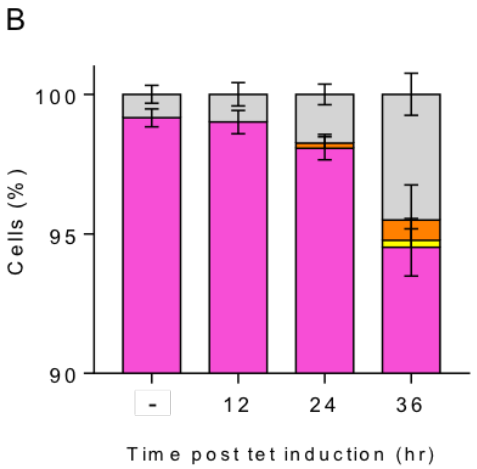

**Figure 6.16 RNA-seq reveals increased expression of silent VSGs across the repertoire after depletion of *TbRH2A*.**
Above: Graph depicting the number of VSG genes that display upregulation in induced samples relative to un-induced at both 24 hr and 36 hr time points. Upregulation was determined by differential expression analysis (section 2.8.5.3) using two independent RNA-seq replicates. The total number is sub-categorised depending on whether the VSGs have been localised to the BES, are intact genes in the subtelomeric arrays (array), are in mini-chromosomes (MC), are pseudogenes (pseudo), or are in a mVSG ESs (MES). Below: Plots of normalised RNA-seq read depth abundance (y-axes) relative to CDS position (x-axes) for a selection of the above VSGs. Two independent repeats after 36 hr of growth are plotted (overlaid in orange and blue). VSG identity numbers are from (Cross, Kim and Wickstead, 2014).

### 6.3.2.2 Changes in the VSG protein coat after TbRH2A depletion

To investigate if changes in the VSG surface occur after depletion of TbRH2A, immunofluorescence analysis of the VSG coat was undertaken. Anti-VSG221 and

anti-VSG121 dual staining of induced and un-induced parasites after 12, 24 and 36 hr of culture was performed. An average of ~0.84% of the un-induced populations from each time point were found to express neither VSG221 nor VSG121 (Figure 6.17B). This appeared to slightly increase to ~1% of cells after 12 hr of tet-induction. After 24 hr, however, ~1.75% of parasites showed no VSG221 or VSG121 signal and a further ~0.18% expressed both VSGs simultaneously. No cells were found to exclusively express VSG121 at this time point. After 36 hr of TbRH2A depletion, ~4.5% of cells expressed neither VSG, ~0.24% expressed VSG121 exclusively and ~0.73% expressed both proteins. Hence, ~1% of cells began to express VSG121 but of these ~75% expressed both VSGs (Figure 6.17B).




| Time | VSG221 | VSG121 | Both | Neither |
|---|---|---|---|---|
| tet - | 99.16 | 0.00 | 0.00 | 0.84 |
| 12 hr | 99.00 | 0.00 | 0.00 | 1.00 |
| 24 hr | 98.07 | 0.00 | 0.18 | 1.75 |
| 36 hr | 94.52 | 0.24 | 0.73 | 4.51 |

**Figure 6.17 Loss of TbRH2A induces switching of the VSG coat.**
A) Co-immunofluorescence imaging of VSG221 (pink) and VSG121 (yellow) surface expression. Example images are shown of cells which have undergone 24 hr of culture with (tet -) or without (tet +) RNAi induction (Scale bar, 5 μm). B) The graph displays the percentage of un-induced (tet -, all time points) and induced cells after 12, 24 and 36 hr of culture with tet, expressing only VSG221 (pink) or VSG121 (yellow) on their surface, as well as cell with both (orange) or neither (grey) of the two VSG on their surface, as determined by co-immunofluorescence imaging with anti-VSG221 and VSG121 antiserum. >200 cells were analysed for each cell type in each of three replicates (error bars denote SEM). The table below shows the average percentage of the three replicates in each case, plotted on the graph above.

### 6.3.3 γH2A is enriched in the active and silent BESs after TbRH2A depletion

γH2A ChIP-seq data generated after 24 and 36 hr of tet-induction of TbRH2A RNAi, explored previously in section 5.3, were next mapped across the 14 characterised BESs (Hertz-Fowler *et al.*, 2008) using MapQ filtering. As described previously (section 5.3), signal enrichment between induced and un-induced samples (each first normalised to the relevant input sequencing) was plotted for both 24 and 36 hr time points (Figure 6.18). Separate plotting of induced and un-induced samples relative to input, for both time points, is displayed in Figure 8.8. For comparison, γH2A ChIP-seq signal enrichment in WT cells is also shown (data as in Figure 1.12). The extent of γH2A signal increase was considerable, since even 24 hr post-induction much greater numbers of peaks were seen than WT cells (Figure 6.18). Notably, signal peaks appeared in all BESs, not predominantly active BES1, and displayed a much wider distribution: peaks were not mainly localised around the 70-bp-repeats but also spread across the ESAG-containing regions. Signal was also evident within, and flanking the VSG coding regions (Figure 6.18). Despite this greater accumulation of BES γH2A, some similarities can be drawn with the BES γH2A ChIP-seq mapping in *Tbrh1-/-* mutants: first, signal enrichment appeared greater in the active BES (BES1, VSG221) than in the silent BESs; and, second, there was a notably lower signal enrichment across the 70-bp repeats. After 36 hrs of RNAi, the γH2A ChIP-seq signal appeared slightly more extensive across active BES1 and also appeared to increase, to a greater extent, across all the silent BESs. Despite this apparent increase, signal enrichment still appeared lower across the 70-bp repeats.

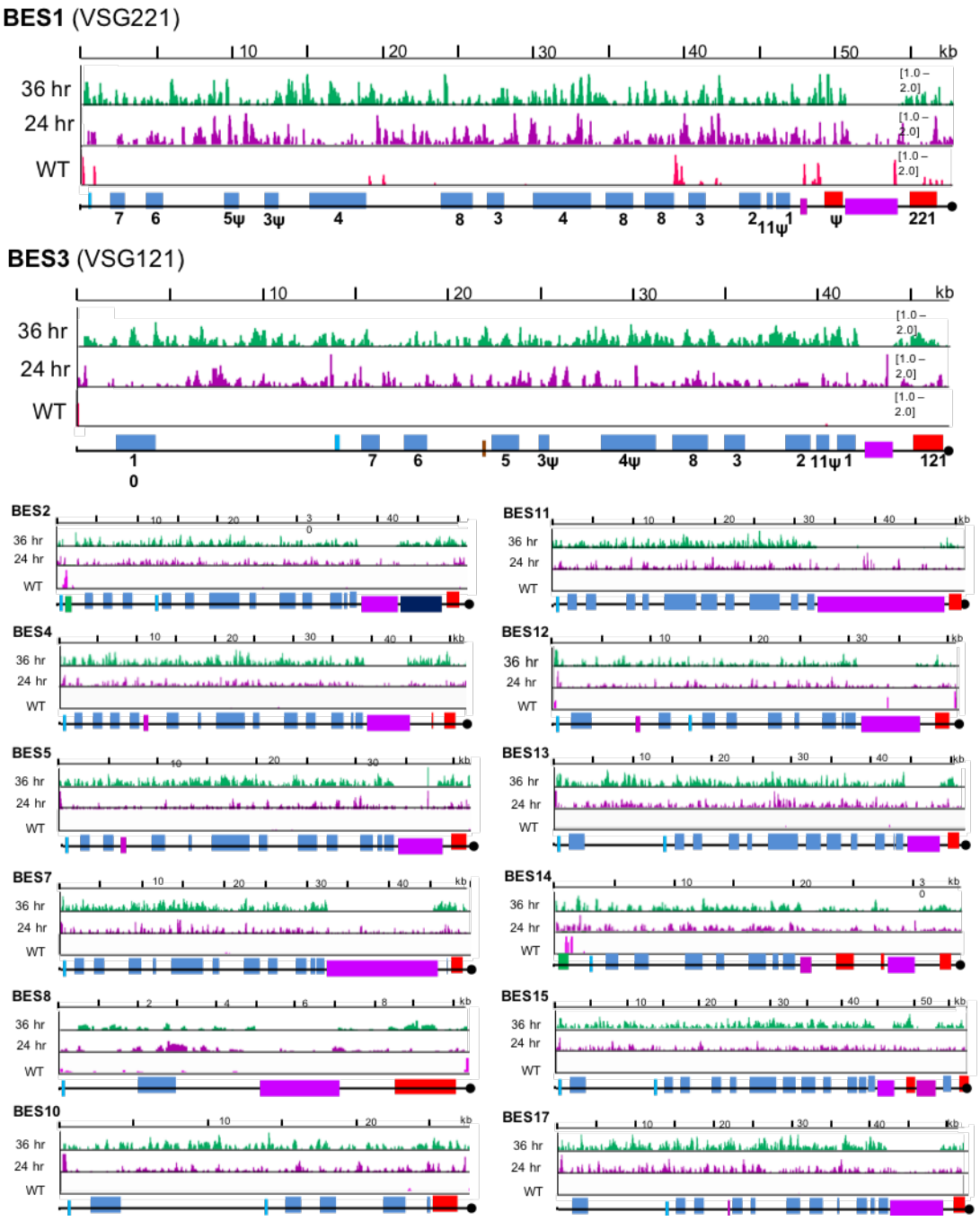



**Figure 6.18 Localisation of γH2A by ChIP-seq mapping to the BESs after depletion of TbRH2A expression with RNAi.**
ChIP-seq signal enrichment is shown mapped to the 14 BESs (Hertz-Fowler *et al.*, 2008) as a ratio of tet-induced samples relative to un-induced (each first normalised to the cognate input sample) for both 24 (purple) and 36 (green) hr time points. WT γH2A ChIP-seq signal (normalised to input) is shown for comparison (pink). Annotations are as described in Figure 6.1.

Increased γH2A signal after 24 and 36 hr induction compared to un-induced samples was corroborated by γH2A ChIP-qPCR, targeting VSG221, VSG121, ESAG6 and ESAG8 coding regions (Figure 6.19). In all four cases, γH2A immunoprecipitated DNA increased significantly in tet + samples compared with tet – counterparts for both

time points. Also, in all four cases, 36 hr tet + IP sample recovery was greater than at 24 hr. Finally, VSG221 showed the highest level of γH2A immunoprecipitation after both 24 and 36 hr of TbRH2A knockdown (Figure 6.19), consistent with greater abundance of modified histone in active BES1 compared with inactive BES3.

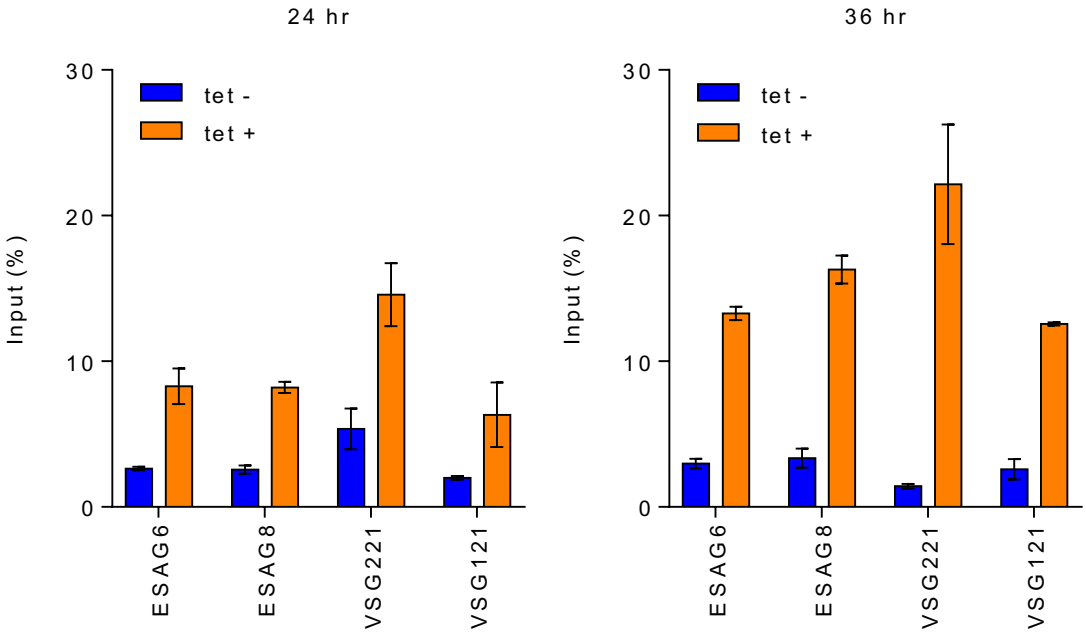



**Figure 6.19 ChIP-qPCR targeting histone variant γH2A shows increased binding of the damage-dependent modified histone after TbRH2A depletion.**
γH2A ChIP-qPCR using primers targeting the sequences of ESAG6, ESAG8, VSG221 (BES1) and VSG121 (BES3), is depicted as the percentage of amplification sequence in the IP sample relative to input. Data is shown for tet induction (tet +) and non-induction (tet -) after 24 (left) and 36 hr (right) of growth. Error bars display SD for two technical replicates.

## 6.4 Discussion

Data presented in this chapter have revealed that increased R-loop formation due to loss of RNase H activity results in increased VSG BES-associated DNA damage and elicits VSG switching in *T. brucei* parasites. Taken together the data suggest BES sequence composition, active transcription and, potentially, DNA replication, appear to interplay and contribute to a model of R-loop driven VSG switching (Figure 6.20).

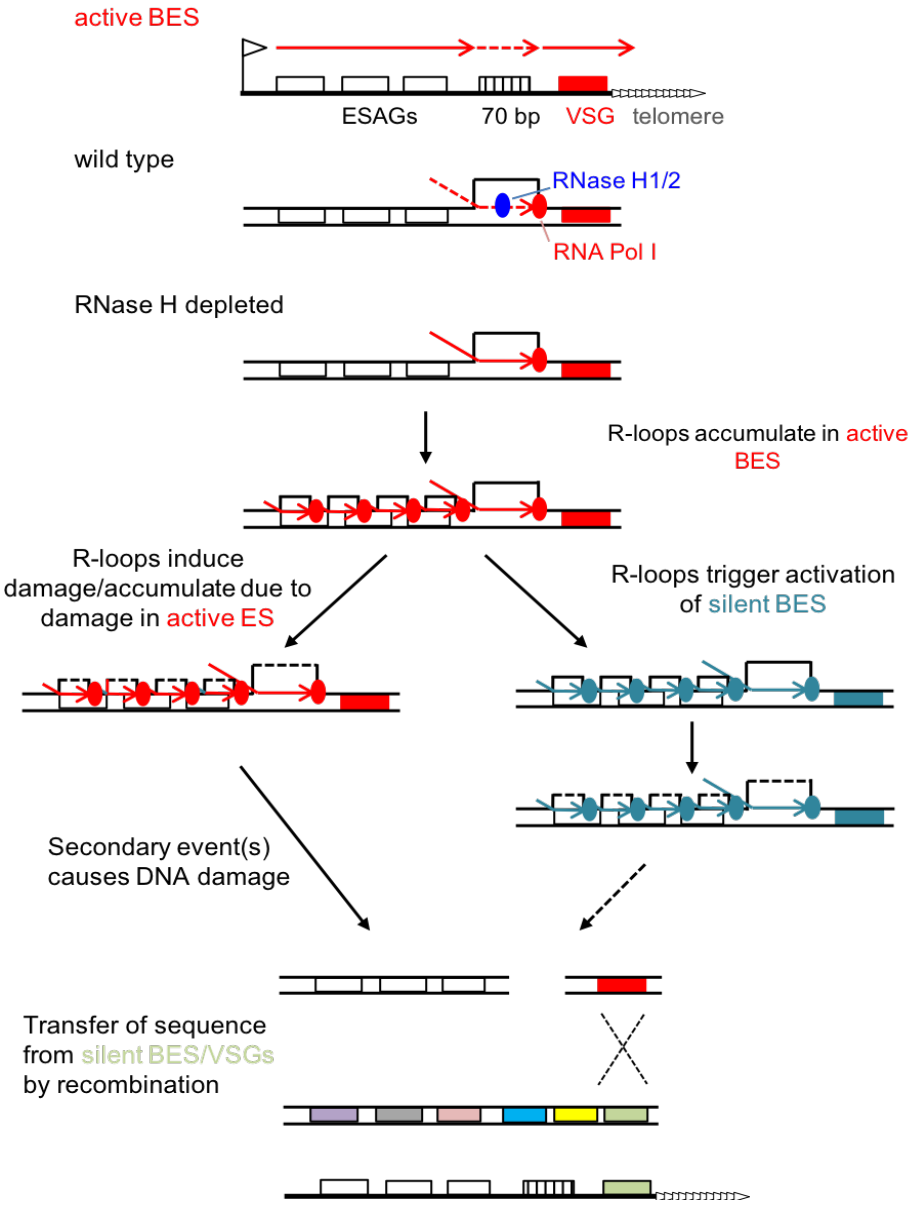



**Figure 6.20 Model of R-loop driven VSG expression changes in RNase H depleted parasites.** The topmost diagram depicts transcription (red arrow) of the active BES, starting at the promoter (flag) to transcribe the ESAGs (white boxes) and VSG (red box) towards the telomeric repeats (arrow heads). Transcription through the 70-bp repeats (lined box) is potentially perturbed in some manner, (indicated by a dashed red arrow) or is R-loop prone as demonstrated by the diagram below. Here, R-loop formation across the 70-bp repeats is depicted in WT parasites, where both RNase H1 and the RNase H2 complex (blue) are able to resolve RNA-DNA hybrids. In either *Tbrh1-/-* or induced TbRH2A RNAi parasites (RNase H depleted), however, R-loops persist and potentially cause the retrograde spreading of R-loops as RNA Pol I (red) continue to transcribe the BES. These may lead to an *in situ* switch to a previously silent BES (teal) due to a lack of VSG expression, or to the accumulation of DNA damage in the active BES. The newly activate BES may also suffer the same fate in the absence of RNase H activity and accumulate R-loops. Recombinational switching is also predicted to occur due to the accumulation of DNA damage. In this case, secondary events such as targeting by flap endonucleases, cytidine deaminases, or potentially clashes with replication machinery, may lead to DSB formation. Here, one DSB is depicted for simplicity, however, DSBs may occur in any location across the BES to allow recombination with silent array VSGs (multiple coloured boxes), VSG found in other BES or mVSG ES. In fact, recombination may occur with elements of pseudogenic VSG sequence to generate mosaic VSGs (not shown).

BES structure, particularly 70-bp repeat sequence composition, appears to encourage R-loop formation, most likely via transcription during expression of the active VSG. DRIP-seq mapping of RNA-DNA hybrids within the active BES of WT cells indicates a small enrichment in signal within the 70-bp repeat region of BES (Figure 6.3), despite a striking paucity of BES R-loops in such cells. Extensive hybrids are then detected across the 70-bp repeats of all BESs after abolition of RNase H activity, since the same effect is seen in both *Tbrh1*-/- null mutants (Figure 6.1, Figure 6.3) and during RNAi knockdown of TbRH2A expression (Figure 6.13). Although DRIP-seq signal peaks are observable across the 70-bp repeats when there is no induction of RNAi (tet – track, Figure 6.13), this is most likely due to leaky RNAi expression (Figure 3.14A) resulting in slight loss of TbRH2A activity. Whether DRIP-seq signal is significantly higher in the RNAi-induced sample vs un-induced cells cannot be confidently ascertained from the DRIP-seq data presented, which is probably not quantitative (K. Chen *et al.*, 2015). To allow clear quantitative comparison between samples, DRIP-seq would require inclusion of spike-in controls (K. Chen *et al.*, 2015). Nonetheless, targeted DRIP-qPCR allied to *E. coli* RNase HI treatment in the *Tbrh1*-/- mutants and after 36 hrs of TbRH2A RNAi argues that R-loops do increase in the BES, though the 70-bp repeats were not directly tested. Characteristics of the 70-bp repeats may make this region a favourable location within the BES for R-loop formation. 70-bp repeats are variable in length and sequence but are in part formed of AT-rich triplicate repeats (Shah *et al.*, 1987; Hovel-Miner *et al.*, 2016). AT-rich repeats are known to become non-H bonded (Ohshima *et al.*, 1996) and even promote recombination (Pan *et al.*, 2010) when transcribed. It is therefore plausible that transcription of these vulnerable repeats leads to R-loop formation and, potentially, initiates VSG switching mechanisms (Figure 6.20). Indeed, it is now widely accepted that the repetitive nature of several fragile sites encourages R-loop formation when transcribed, underlying these regions' vulnerability to breakage (Sollier and Cimprich, 2015; Freudenreich, 2018).

Accordingly, 70-bp repeat associated R-loops could trigger VSG switching by potentially two, non-mutually exclusive, routes: transcriptional *in situ* switching and recombination-based switching. The former may be explained by considering

the fate of the active BES after prominent R-loop formation. DRIP-seq mapping and DRIP-qPCR reveals RNA-DNA hybrids also form in the BES upstream of the 70-bp repeats, within the ESAG coding regions. This effect is seen both after knockout of TbRH1 (Figure 6.1 and Figure 6.3) and after RNAi knockdown of TbRH2A, most markedly after 36 hr of tet-induction, as demonstrated by DRIP-qPCR (Figure 6.13). In all cases, DRIP-seq peaks do not appear to form specifically within a particular sequence type, such as the predominant inter-CDS localisation seen in RNA Pol II PTUs (Chapter 4). Instead, DRIP-seq enrichment appears arbitrary in the BES, and feasibly may accumulate as a consequence of retrograde spreading of transcription blockage, extending from the 70-bp repeats (Figure 6.20). Indeed, transcriptional pausing is known to cause R-loop formation (Chen *et al.*, 2017). Blockage to transcription upstream to the VSG coding region would hence result in a failure to express the active VSG, which is known both to be lethal to *T. brucei* BSF parasites (Sheader *et al.*, 2005) and capable of triggering switching (Aitcheson *et al.*, 2005). Thus, R-loop formation in the active BES due to loss of RNase H activity may result in the selection for parasites that have switched transcription to a different BES.

Data in this thesis support the above model of R-loop driven BES transcriptional switching. RT-qPCR targeting BES-associated VSGs (Figure 6.4 and Figure 6.14), and RNA-seq analysis of RNase H deficient lines reveals increased expression of both BES-associated VSGs and ESAGs (Figure 6.5 and Figure 6.15). *In situ* switching to an alternative BES in RNase H-deficient cells invariably means the newly selected BES is also likely to accumulate R-loops and suffer transcription blockages, resulting in further switching. This would account for the finding that all BES, not just active BES1, display DRIP-seq signal after TbRH1 knockout (Figure 6.1) or TbRH2A knockdown (Figure 6.13). Furthermore, the simultaneous expression of two VSGs (VSG221 and VSG121) after loss of TbRH1 (Figure 6.8) and TbRH2A (Figure 6.17), albeit in small proportions of the populations, indicates either cells in the process of switching, or cells in which monoallelic BES control has been undermined. Although an attractive model, blockage to transcription as a result of R-loop formation requires experimentally validation, by ChIP-seq targeting an RNA Pol I subunit, for example.

Despite the support for the above transcriptional model, the data presented here do not indicate that R-loops induce *in situ* switching alone, but may generate lesions capable of eliciting VSG switching via recombination. (In fact, it might be considered that it is lesions generated by R-loops, and not the RNA-DNA hybrids alone, that impede VSG transcription and cause *in situ* switching.) DNA breakage is widely acknowledged to initiate VSG gene conversion and recombination events (Barry and McCulloch, 2001; Horn, 2004; Boothroyd *et al.*, 2009), but how such lesions are generated in the parasites is unclear. The increased abundance of Thr130 phosphorylated histone H2A, which is a marker of nuclear DNA damage (Kinner *et al.*, 2008; Glover and Horn, 2012), in the BESs after depletion of RNase H activity points to R-loop structures as a potential trigger for initiating DNA lesions. γH2A ChIP-seq reveals predominant binding of the phosphorylated histone to sites close to the 70-bp repeats of the active site (BES1) in WT parasites. In *Tbrh1*-/- parasites this relatively localised accumulation of γH2A in the active BES increases (Figure 6.12), as well as spreading upstream into the ESAG coding region. Little γH2A signal is seen in inactive BESs of *Tbrh1*-/- cells. More extensive γH2A enrichment in the BESs is observable in TbRH2A RNAi induced parasites, relative to un-induced samples, both 24 and 36 hr after induction (Figure 6.18 and Figure 6.19), since increased γH2A ChIP signal is found throughout both the active and silent BES, and is not so clearly limited to regions around the 70-bp repeats. The basis for this difference in damage accumulation in the *Tbrh1*-/- mutants versus TbRH2A$^{RNAi}$ cells probably rests in the severity of the phenotypes, suggesting loss of the former RNase H is tolerated but depletion of the latter is lethal. Furthermore, the difference in the BES γH2A ChIP-seq data is consistent with the more limited, replication-associated accumulation of nuclear γH2A foci in *Tbrh1*-/- mutants relative to the much more severe accumulation of γH2A signal throughout the nucleus after TbRH2A RNAi.

Further data presented in this chapter support the hypothesis that the lesions indicated by γH2A, which increase after loss of RNase H, induce VSG recombination events. Firstly, RNA-seq analyses show increased transcription of silent VSG found outside of the BESs, in both *Tbrh1*-/- (Figure 6.6) and TbRH2A RNAi induced parasites (Figure 6.16). Although analysis of the *Tbrh1*-/- cells is limited by only

having a single replicate, comparison of read depth between WT and *Tbrh1-/-* samples demonstrates increased transcription in *Tbrh1-/-* parasites of intact VSGs found within subtelomeric arrays, mini-chromosomes and mVSG ESs, as well as some pseudogenic VSGs. Differential expression analysis allowed statistically significant differences in VSG expression to be identified after TbRH2A depletion. Again, VSGs from subtelomeric arrays, mVSG ESs and mini-chromosomes, as well pseudo-VSG sequences, were all identified as being more highly transcribed after depletion of TbRH2A, most widely after 36 hr of knockdown. Though it is conceivable these data indicate increased transcription of the genes from their endogenous loci, it is more likely that expression of these non-BES VSGs is due to recombination into the BES. Moreover, RNA-seq mapping provided weak evidence for increased transcription of entire, silent BES and so the proximity of damage, at least in *Tbrh1-/-* mutants, to the 70-bp repeats and VSG may drive the recombination of non-BES VSG genes. One area of uncertainty is that no strong γH2A ChIP-seq signal was detected in the 70-bp repeats, either in the WT cells or the two RNase H mutants, which seems at odds with the suggestion the repeats might be a focus for VSG switch initiation. Nonetheless, nothing is known about nucleosome density in the 70-bp repeats and it is conceivable that lowering of H-bonding limits H2A deposition, and hence limits damage-dependent formation of γH2A. In addition, the nature of the DNA lesion that might form to generate γH2A signal is unknown and may take a different form in the 70-bp repeats and elsewhere in the BESs. It is also conceivable that lack of γH2A ChIP of the 70-bp repeats is merely a technical artefact of ChIP with this antiserum or library preparation on such sequences.

Taken together, the data in this chapter suggest R-loops may in fact initiate both transcription and recombination routes of VSG switching, and mechanistically link the two processes. If found to be the case, this hypothesis is in keeping with findings that both transcriptional switching and VSG gene conversation are impaired in homologous recombination mutants lacking RAD51, BRCA2 and RAD51-3 (McCulloch and Barry, 1999; Hartley and McCulloch, 2008; Dobson *et al.*, 2011). Furthermore, the formation of R-loops and lesions upstream of 70-bp repeats within

the ESAG coding regions is also consistent with frequent observations of intra-BES recombination events (McCulloch, Rudenko and Borst, 1997).

The similarities between the TbRH1 and TbRH2A data sets in this chapter suggest an overlap in functions of the two enzymes during VSG expression and switching: in both cases, increased R-loops, γH2A abundance and VSG switching are observed after RNase H loss. However, differences between the phenotypes do exist and future exploitation of these difference could provide further insight in to switching mechanisms. The differences in BES-associated DNA damage levels between cells lacking TbRH1 and TbRH2A appears to be matched by slightly different levels in VSG switching; immunofluorescent staining with anti-VSG221 antisera reveals a greater proportion of TbRH2A depleted cells that do not express the dominant VSG (VSG221) than *Tbrh1-/-* mutant cells (Figure 6.8 andFigure 6.17), consistent with greater increased expression of silent VSGs after TbRH2A RNAi. The reason for elevated DNA damage and, importantly, damage across all BES and not predominantly active BES1 after TbRH2A loss relative to TbRH1 loss is currently unclear. Although the two enzymes presumably act, at least in part, redundantly at the BESs to resolve R-loops, it may be the case that the TbRH2 complex has a more active role than TbRH1 at the BESs. It also must be considered that the TbRH2 complex is probably able to remove embedded ribonucleotides via initiating the RER pathway, a function that TbRH1 is unlikely to carry out. It is hence possible that higher levels of embedded ribonucleotides in the BES after TbRH2A depletion contributes to more DNA lesions. Indeed, increased incorporation of uracil into DNA, due to loss of uracil-DNA glycosylase, leads to lesions and VSG switching (Víctor M Castillo-Acosta *et al.*, 2012).

How R-loops might cause recombination in the context of VSG switching requires further investigation. However, parallels can be drawn with circumstances in other eukaryotes where R-loops induced DNA breakage and recombination. Firstly, R-loops and transcriptional pausing alone may cause sufficient instability to result in DNA lesions. However, this is somewhat unlikely, since, as discussed previously, it is generally accepted that a secondary event is necessary to process R-loops into DSBs (Sollier and Cimprich, 2015). One possibility is that clashes between transcription-

associated R-loops and DNA replication machinery within the BESs generate DSBs. This hypothesis is supported by the observation that DNA damage in *Tbrh1-/-* parasites is largely associated with DNA replication, as demonstrated by increased γH2A foci in 1N1eK and 1N2K stage parasites (Figure 6.11). An association between DNA damage and DNA replication in TbRH2A depleted cells is, however, less easily understood due to the severe changes to the cell cycle. Nonetheless, EdU uptake experiments (Figure 3.18) and flow cytometry analysis (Figure 3.16) show that TbRH2A depleted cells continue to synthesise DNA. Hence, it is possible that clashes with the DNA replication machinery occur in the actively transcribed BESs of these parasites with greater frequency than in *Tbrh1-/-* mutants. An element of this model may also explain the differential timing of BESs replication; the active BES has been found to replicate early in BSF cells but not PCFs (Devlin *et al.*, 2016).

Another possibility is that R-loops are the target of currently unknown enzymes or factors that elicit VSG switching. AID targets cytosine residues in the exposed ssDNA of R-loops, converting the bases to uracil and leaving the strand vulnerable to breakage during class switching in the IgH locus of B lymphocytes (Basu *et al.*, 2011). Although experimental testing would be required to test such a model, the localised recombination observed in Ig switching has some similarities with the long-range recombination events of VSG switching, though the mode of recombination is quite different. Processing by other enzymes, such as flap endonucleases, is also possible but again would require experimental investigation. Indeed, R-loop induced hyper-recombination of THO mutant yeast is suppressed by overexpression of XPG flap endonuclease (Sollier *et al.*, 2014).

Lastly, it remains possible that R-loops may be a response to DNA lesions rather than the cause. DNA nicks have been found capable of initiating R-loop formation (Roy *et al.*, 2010). Furthermore, RNA-DNA hybrids can form as intermediate structures during HR repair of DSBs, and these require degradation by RNase H (Ohle *et al.*, 2016). Although it is clear that R-loops induce damage and recombination in many settings, distinguishing R-loops which cause lesions from those which result from them remains challenging (Sollier and Cimprich, 2015). All possible routes of R-loop processing into DSBs require investigation before a

tangible model of R-loop driven VSG switching can be established. However, if R-loops do prove to be a key feature in VSG switching, several observations of VSG switching phenotypes in other *T. brucei* mutants may be explained. Firstly, deletion of RAD51 in *T. brucei* impairs VSG switching and so was hypothesised to catalyse switching reactions (McCulloch and Barry, 1999). Mutation of RAD51 in yeast leads to R-loop formation and rearrangements (Wahba, Gore and Koshland, 2013). Conversely, gene conversion events are increased in REQ2 mutant parasites (Devlin *et al.*, 2016). This phenotype is also explicable, since RecQ helicase mutants of both yeast (SGS2) and mammalian cells (BLM) causes R-loop increases and locus-specific instability (Chang *et al.*, 2017). Lastly, RNase H mutant yeast display DNA damage at the RNA Pol I transcribed rRNA loci, which is repaired by break-induced replication facilitated by persistent R-loops (Amon and Koshland, 2016). Interestingly, break-induced replication has previously been proposed to induce VSG switching (Barry and McCulloch, 2001; Kim and Cross, 2010).

To fully understand the role of R-loops in VSG switching mechanisms, many questions need to be addressed. Firstly, do R-loops cause or result from DNA lesions? Exactly what type are these lesions and, if R-loops are causal, what events process R-loops into DNA breaks? How does signalling of the resulting lesion(s) initiate the repair-processes that facilitate VSG switching? Lastly, if R-loops form during HR-mediated repair of DNA lesions, does this result in *trans* formation of R-loops in other BESs and/or silent VSG arrays? Irrespective of the detailed mechanism, data discussed in this chapter strongly argues that BES structure facilitates R-loop formation, particularly in the 70-bp repeat regions whose upstream position to the VSG is advantageous for HR-mediated break repair of lesions and whose sequence directs recombination to the large bank of 70-bp repeat associated silent VSGs.

# 7 Perspectives and future directions

Since the sequence-specific DNA binding capabilities of RNA were first demonstrated (Thomas, White and Davis, 1976; White and Hogness, 1977), investigation into RNA-DNA hybrids and R-loop structures has expanded extensively. Isolation of the hybrids using the S9.6 antibody, as well as the advent of next-generation sequencing technologies, has more recently facilitated global analysis of RNA-DNA hybrid formation across multiple genomes, including that of *S. cerevisiae* (El Hage *et al.*, 2010; Wahba, Gore and Koshland, 2013; Chan *et al.*, 2014), *S. pombe* (Ohle *et al.*, 2016), *Arabidopsis* (Xu *et al.*, 2017), mouse (Sanz *et al.*, 2016) and human (Ginno *et al.*, 2012, 2013; Sanz *et al.*, 2016). In this thesis DRIP-seq has been used to investigate RNA-DNA hybrid localisation across the unusually structured *T. brucei* genome. Mapping of R-loops across the core of the *T. brucei* genome has revealed loci where R-loop formation appears conserved with other genomes, such as the centromeres and RNA Pol III transcribed tRNA and snRNA genes, RNA Pol II transcriptional start sites, and rRNA gene arrays. Additionally, a novel association of R-loops with sites of polyadenylation and, potentially, *trans*-splicing has been uncovered.

Genetic manipulation of putative RNase H enzymes has also allowed the role of these enzymes in *T. brucei* genome maintenance to be explored. R-loop mapping via DRIP-seq in BSF parasites lacking either type 1 or type 2 RNase H activity revealed a potentially specific role of the RNase H2 complex at RNA Pol II transcription initiation regions, a role which RNase H1 appears unable to efficiently perform. Lastly, mapping of both R-loops and the DNA damage marker γH2A to the VSG BESs, as well as analysis of VSG switching in *Tbrh1*-/- and TbRH2A$^{RNAi}$ parasites, demonstrates R-loop formation may drive VSG switching in the parasite, suggesting a localised function with potential parallels to R-loop-induced genome instability in other organisms.

## *7.1* The RNase H repertoire of *T. brucei*

Most organisms express at least one RNase H enzyme, and more commonly possess two types of RNase H enzyme, broadly classified as type 1 and type 2, the latter of which is made up of three subunits in eukaryotes. Three putatively catalytic RNase

H proteins could be identified in the *T. brucei* genome via analysis of the predicted amino acid sequences, as well as two proteins that may act as the RNase H2B and H2C subunits. *T. brucei* is, however, not alone in expressing more than two putatively catalytic RNase H-like proteins. For example, *Candida albicans* possesses three genes which all encode a type 1 RNase H. Two of these show similarity with bacterial RNase HI proteins and appear to be the product of a duplication event (Kochiwa, Tomita and Kanai, 2007), whereas the third is more typical of eukaryote RNase H1 (Cerritelli and Crouch, 2009). *C. elegans* possess four genes encoding RNase H1-like proteins, in addition to a type 2 RNase H complex, yet only two are catalytic (Arudchandran *et al.*, 2002). The varied combinations of RNase H proteins expressed by diverse organisms presumably reflects differing genome structures and, although all catalytic proteins resolve RNA-DNA hybrids in some manner, specific roles the RNase H proteins have evolved to perform in each.

The *T. brucei* RNase H1 protein, TbRH1, examined in this thesis has previously been characterised as a type 1 RNase H (Hesslein and Campbell, 1997). This protein possesses a "typical" type 1 eukaryotic RNase H structure (Cerritelli and Crouch, 2009), containing conserved HBD and RNase H domains and a 42 amino acid Connecting Domain (Kobil and Campbell, 2000), and so is likely to resolve RNA-DNA hybrids containing a minimum of four consecutive ribonucleotide residues (Ohtani *et al.*, 1999). Epitope tagging of the full length TbRH1 protein reveals nuclear localisation of the protein, previously shown to be dependent upon the functional N-terminal HBD (Kobil, 2000). Unlike mammalian type 1 RNase H proteins, TbRH1 lacks an MTS required for targeting to the mitochondria, consistent with a lack of detectable mitochondrial or kDNA localisation of the tagged protein in this study. In mice, and probably humans, RNase H1 is completely essential for mtDNA replication (Cerritelli *et al.*, 2003) but most likely acts redundantly with the RNase H2 complex in the nuclear genome. In accordance with RNase H1 being non-essential for eukaryotic nuclear genome functions, null mutation of the protein is not lethal in yeast, despite resulting in genomic instability (Wahba *et al.*, 2011; O'Connell, Jinks-Robertson and Petes, 2015). It is likely that TbRH1 acts similarly to other eukaryotic RNase H1 proteins in the *T. brucei* nuclear genome, since deletion of the

protein causes no deleterious effects, either to growth rate or cell cycle progression, despite elevated levels of DNA damage.

DRIP-seq mapping of RNA-DNA hybrids and targeted DRIP-qPCR has revealed that TbRH1 potentially targets R-loops genome-wide in *T. brucei*. Although normalisation of sequencing data may have masked expected increases in R-loops after TbRH1 loss in certain areas, increased hybrids were detected across the centromeric repeats, in subtelomeric regions, rRNA gene arrays, RHS-associated genes and in the BESs of *Tbrh1*-/- parasites in comparison with WT. DRIP-qPCR further demonstrated TbRH1 is likely to target RNA-DNA hybrids that form over the RNA Pol III transcribed tRNA and snRNA genes, as demonstrated in yeast (El Hage *et al.*, 2014), and potentially globally across RNA Pol II transcribed PTUs. The data discussed in this thesis therefore do not indicate any site-specific activity for TbRH1, although this would require further experimental investigation, for example via ChIP-seq localisation of the enzyme. Recently, R-loop resolution activity of RNase H1 has been shown to allow progression of replication forks within the human genome (Parajuli *et al.*, 2017). Given the widespread formation of R-loops across the *T. brucei* genome, it is very likely that these structures form barriers to replication machinery in the parasite, and perhaps TbRH1 resolves hybrids to allow unperturbed replication in this setting as well. However, none of the data presented here is able to refute or confirm such a suggestion and further analysis would be necessary, such as by testing for the accumulation of replication machinery at prominent R-loop forming loci in the presence and absence of TbRH1.

In contrast to dual mitochondrial and nuclear localisation of RNase H1 in mammals, a distinct RNase H-like protein (TbRH3) was identified in this thesis that contains an MTS and can be colocalised to the kDNA. This protein shows weak homology with either type 1 or type 2 RNase H enzymes in most eukaryotic species but does have orthologs in other trypanosomatids and shows some homology with bacterial RNase HII monomeric proteins. Although the predicted RNase H domain of this protein is larger than that of bacterial type 2 proteins, all active site and catalytic residues have been conserved, perhaps highlighting specialisation of the domain for a role in kDNA replication or maintenance. These data suggest trypanosomatids have evolved

a dedicated mitochondrial RNase H. One reason may be that the unique structure of the kDNA requires a stand-alone protein. It could be assumed that TbRH3 acts, instead of TbRH1, in kDNA replication, as type 1 RNase H proteins have been shown do in mtDNA replication (Xu and Clayton, 1996) and in *E. coli* plasmid replication (Itoh and Tomizawa, 1980), by excising the RNA strand of R-loops that form at the organellar genome replication origin to generate a ssDNA primer for DNA replication. However, as TbRH3 shows best homology to type 2 bacterial RNase H proteins that are unable to resect tracts of RNA-DNA hybrid in a processive manner like type 1 RNases H, its precise activity is likely to differ somewhat. Hence it is unlikely that TbRH3 catalyses a reaction at kDNA replication origins identical to type 1 RNases H during mtDNA and plasmid replication.

Replication origins of kDNA minicircles contain the universal minicircle sequence (UMS), where leading strand synthesis is initiated by binding of the UMS binding protein (UMSBP) (Abu-Elneel, Kapeller and Shlomai, 1999). Interestingly, the UMS takes on an unusual, but poorly characterised conformation to allow the UMSBP to bind only a single strand of template DNA (Avrahami, Tzfati and Shlomai, 1995). UMSBP additionally binds a distinct hexamer sequence where the first Okazaki fragment is synthesised, beginning lagging strand replication (Abu-Elneel, Kapeller and Shlomai, 1999). In contrast, here UMSBP binds the non-template DNA strand, which logically could be the ssDNA element of an R-loop structure, containing an RNA-DNA hybrid formed by the RNA primer. Interestingly, Okazaki fragment primers remain in replicated minicircles until they migrate to the anti-podal sites for removal (Ryan and Englund, 1989), where TbRH3 appears to migrate during the cell cycle. Here, TbRH3 may remove the RNA primers between Okazaki fragments. Indeed, mammalian RNase H1 is able to degrade most of the RNA primer sequence, although it is unable to remove the 3'-most RNA, which is instead excessed by a flap endonuclease (Turchi *et al.*, 1994). Nonetheless, as noted above, TbRH3 does not contain a 'typical' RNase H type 1 or type 2 domain structure and so its function may be more specialised to the complex kDNA. Indeed, the inability to generate RNAi cells meant it was not possible to directly test for a role in either minicircle or maxicircle kDNA replication.

The RNase H2 complex is essential for mammalian development and has been shown here to be essential for viability of BSF *T. brucei* parasites. Similarities exist between the phenotypes of mouse cell lacking a functional RNase H2 complex and parasite's depleted for TbRH2A, including accumulation at the G2/M cell cycle stage and increased DNA damage (Hiller *et al.*, 2012; Reijns *et al.*, 2012), indicating the protein may have a conserved function. However, the mutant phenotypes differ in the impact upon DNA replication, as mouse cells fail to synthesise nascent DNA to WT levels without RNase H2 (Reijns *et al.*, 2012), whereas *T. brucei* RNAi cells continue to replicate the nuclear genome, causing increased nuclear genome content despite very high levels of damage. This difference may simply reflect differences in the activation of cell cycle check points after RNase H2 loss, or perhaps the greater use of break-induced replication in *T. brucei*. Interestingly, RNase H2 is not essential in the yeast genome, where it is proposed to both resolve RNA-DNA hybrids and remove embedded ribonucleotides to prevent DNA damage and genomic rearrangements (Conover *et al.*, 2015; O'Connell, Jinks-Robertson and Petes, 2015). The lack of at least one, although more likely both, of these functions presumably leads to irreversible damage in the mammalian and *T. brucei* genomes. It may be the case that additional proteins act redundantly with RNase H2 in yeast, or that RNase H2 has assumed another role in mammals and *T. brucei* that other proteins, including RNase H1, cannot perform.

DRIP-seq using *T. brucei* cells depleted of TbRH2A by RNAi revealed little difference in R-loop formation between the RNAi-induced and induced parasites, with the sole clear exception being loss in DRIP-seq signal at the RNA Pol II TSSs. This localised change may indicate a specialisation of RNase H2 to resolve hybrids found at transcription initiation sites. Further evidence for this role lies in the accumulation of DNA damage, as indicated by γH2A mapping, at the same regions after loss of TbRH2. Presumably, the RNA-DNA hybrid is processed into a DNA break in the absence of adequate TbRH2A protein, and so the hybrid can no longer be detected with DRIP-seq. Although it remains possible that TbRH2 removes large amounts of embedded ribonucleotides that cannot be endured by the parasites, why this mis-incorporation should be more pronounced at, or even specific for, transcription initiation sites is unclear. The intriguing possibility that RNase H2, acting on R-

loops, might play an active role in transcription initiation deserves further experimental examination, as this has not yet been described for any type 2 RNase H protein. One route to examine this may be to use ChIP and ask if TbRH2A localises to transcription start sites. Moreover, analysis of TbRH2A mutants that lack ribonucleotide excision activity but retain R-loop resolution functions (Chon *et al.*, 2013), would be very illuminating.

## 7.2 R-loop formation across the *T. brucei* genome

Using the DRIP-seq approach, RNA-DNA hybrids have been mapped across the *T. brucei* Mb chromosomes in WT parasites, as well as those depleted of RNase H2A and in null mutants of TbRH1. Signal peaks were highly consistent across all the cells, with very pronounced enrichment between CDS regions within the RNA Pol II transcribed PTUs (discussed below). Of the other R-loop forming sites described in this work, many have also been demonstrated in other genomes. For example, RNA Pol III transcribed tRNA and snRNA genes are known site of R-loops (El Hage *et al.*, 2014), as demonstrated in *T. brucei* by both DRIP-seq and DRIP-qPCR. The Pol I rRNA array genes are also well-known sites of hybrid formation (Thomas, White and Davis, 1976; El Hage *et al.*, 2010, 2014), again as observed here in *T. brucei*. Tandem repeat sequences are also known to encourage R-loop formation (Lin *et al.*, 2010) and a broad increase in DRIP-seq signal observably correlates with tandem repeats across the Mb chromosomes in *T. brucei*, in particular at centromeric repeats, similar to that observed in *S. cerevisiae* (Castellano-Pozo *et al.*, 2013). Interestingly, centromeric R-loops have recently been directly implicated in a signalling cascade: RPA coats the ssDNA of the R-loops and activates ATR, which in turn stimulates Aurora A, via Chk1, to ensure correct chromosome segregation during mitosis (Kabeche *et al.*, 2018). The finding that *T. brucei* centromeres also form R-loops indicates that this pathway may be conserved, despite the potentially highly variant kinetochore complex that links the centromere with the spindle (Akiyoshi and Gull, 2014; D'Archivio and Wickstead, 2017).

R-loops have been well documented at CGI-associated ORIs in the human genome (Ginno *et al.*, 2013), but the data presented here found no evidence to link R-loops

to DNA replication in *T. brucei*. Although DRIP-seq does reveal hybrids at SSRs, this signal correlates well with transcription initiation (discussed below), rather than being specific to ORI-containing SSRs (Tiengwe *et al.*, 2012). However, it may be the case that ORI-specific R-loops escaped detection in this study if, for example, they are resolved by factors other than RNase H1 and H2. Additionally, as essentially all SSRs display ORC binding, whether or not they are activated as an ORI during S phase (Tiengwe *et al.*, 2012), it remains plausible that R-loops could play a part in ORC recruitment, as suggested in the human genome (Lombraña *et al.*, 2015). Similarly, no evidence was found to implicate R-loops in transcription termination processes in *T. brucei*, as no significant DRIP-seq signal could be observed at the terminal regions of the PTUs. This is in striking contrast to the widespread roles of R-loops in transcription termination in the human (Skourti-Stathaki, Kamieniarz-Gdula and Proudfoot, 2014; Sanz *et al.*, 2016) and *S. cerevisiae* (Wahba *et al.*, 2016) genomes. However, it again remains the case that factors other than RNases H1 and H2 may resolve termination-associated R-loops, such as a homolog of sen1/SETX, which is known to process R-loops at RNA Pol II termination pause sites (Skourti-Stathaki, Proudfoot and Gromak, 2011; Skourti-Stathaki, Kamieniarz-Gdula and Proudfoot, 2014). Nonetheless, it is conceivable that the evolution of multigenic transcription in trypanosomatids has selected for abolishment of R-loop activity in transcription termination, as it would prevent traversal of RNA Pol II to the downstream gene.

### 7.2.1 R-loops and transcription initiation

As discussed in Chapter 4, R-loops are associated with promoters in mammals (Ginno *et al.*, 2012; Boque-Sastre *et al.*, 2015; P. B. Chen *et al.*, 2015; Sanz *et al.*, 2016), *A. thaliana* (Xu *et al.*, 2017) and yeast (Chan *et al.*, 2014; El Hage *et al.*, 2014). In this thesis, R-loops have been detected at RNA Pol II transcription start sites, despite little clear similarity of these regions with conventional RNA Pol II promoters. At CGI promoters in the human genome R-loops provide protection from repressive DNA methylation, promoting active transcription (Ginno *et al.*, 2012). Though a functional DNA methylation pathway has been demonstrated in *T. brucei* (Militello *et al.*, 2008), the modification's role in transcription has yet to be

established. Although DRIP-seq signal correlates well with epigenetic markers of transcription initiation, whether or not R-loops have a role in their deposition is still unknown. Similarly, at human CGI promoters R-loops are correlated with epigenetic markers associated with open chromatin, and increased RNA Pol accessibility (Stork *et al.*, 2016; Chen *et al.*, 2017), yet a pathway directly linking R-loops to regulation of these markers is to be established. Thus, despite the potential similarity to R-loop localisation in the trypanosome genome, further work is needed to determine if there is a functional overlap. R-loops have also been associated with promoter regions in the *A. thaliana* genome (Xu *et al.*, 2017), although a pathway has not been demonstrated to directly link R-loops and chromatin here. In the mouse genome, where R-loops are again found at CGI promoters, binding of the remodelling complex Tip60-p400 is enhanced by R-loop structures, whereas binding of repressive complex PRC2 is suppressed (P. B. Chen *et al.*, 2015). To date such complexes have not been described in *T. brucei*, and so it is unclear if R-loops might suppress the binding of chromatin condensing factors and/or encourage the binding of those that regulate RNA Pol II TSS chromatin structure, such as histone H2A.Z or H2B.V deposition enzymes (Siegel *et al.*, 2009; Wedel *et al.*, 2017).

Although DRIP-seq mapping of RNA-DNA hybrids in *Tbrh1-/-* cells demonstrated TSS-associated R-loops remain present in the absence of TbRH1, mapping in parasites depleted of TbRH2A revealed the hybrids are lost, or at least reduced, at these loci. As localised DNA damage is also found at these sites after TbRH2A depletion, it is likely that R-loops are not correctly resolved in these conditions. These findings reinforce the suggestion that RNase H2 acts specifically in the regulation of TSS-associated R-loops, which may indicate a parasite-specific adaption of their function. Perhaps the RNase H2 complex has binding partners that recruit it to R-loop-prone transcription initiation regions. If so, identification of such factors would be revealing. It is perhaps valuable to note that in humans cells RNase H2-specific recruitment has been demonstrated to replicating DNA and localised sites of DNA damage, both via PCNA binding (Kind *et al.*, 2014). If a similar PCNA-association exists for TbRH2, this would reinforce a link between transcription and DNA replication in *T. brucei* (Tiengwe *et al.*, 2012; Devlin *et al.*, 2016).

### 7.2.2 R-loop formation at sites of *trans*-splicing and polyadenylation

In other genomes, R-loops are poorly associated with splicing sites and are, in fact, protected from RNA-DNA hybrid formation by the presence of introns, an observation made in both yeast (Bonnet *et al.*, 2017) and humans (Bonnet *et al.*, 2017; Dumelie and Jaffrey, 2017). Thus, the data in *T. brucei*, where effectively every CDS within a PTUs is an exon and the intra-CDS sequences are introns, is diametrically opposed. Unlike in other eukaryotes, the DRIP-seq reveals a strong enrichment for R-loops across intergenic and UTR sequences, which is well correlated with both PASs and nucleosome depletion. RNA Pol II is likely to pause at the nucleosome depleted sites between CDS regions (Wedel *et al.*, 2017), and so it is possible that R-loop formation promotes such pausing to facilitate the action of the *trans* splicing and polyadenylation machineries. However, R-loops may equally form as a consequence of such RNA Pol II pausing. As a result, the two possibilities require experimental investigation by, for example, assessing RNA Pol II pausing via ChIP-seq (Wedel *et al.*, 2017) in RNase H deficient parasites where increased R-loops are to be expected. Similarly, R-loop formation could be analysed using DRIP methods in parasites where transcriptional pausing has been induced artificially such as with the use of $\alpha$-amanitin (Das *et al.*, 2006).

## 7.3 R-loops are able to drive VSG switching

Another, equally specific location of R-loop formation in the *T. brucei* genome is within the BESs, particularly across the 70-bp repeats. Multigenic, protein-coding RNA Pol I transcription units of this sort have no counterpart in non-kinetoplastid eukaryotes, indicating R-loop localisation in this environment is truly unique. The data presented, demonstrating the accumulation of BES-associated DNA damage and increased levels of VSG switching in *Tbrh1-/-* mutants and TbRH2A depleted parasites strongly supports a model of R-loop-driven VSG switching, potentially by both transcription and recombination routes. However, several questions remain before a pathway of switching can be clearly drawn. Firstly, what events ensure R-loop formation predominantly, or exclusively in the actively transcribed BES? Does transcription stall as a result of R-loops, causing a decrease in VSG transcript levels capable of causing a transcriptional switch to an alternative BES? Is the R-loop

directly processed to form a DNA break, and if so by what factors? Alternatively, is a secondary event such as collision with the DNA replication machinery required to generate a DNA break? Do R-loops in fact form *in trans* due to the highly homologous sequences of the BESs and within the VSG archive? Lastly, in unperturbed parasite populations that undergo switching during an infection, what events could precede R-loop formation? Data in this thesis suggest RNase H enzymes target these R-loops for resolution; is this also the case during infection? How are sufficient R-loops generated in order to induce optimum levels of switching?

Although the high divergence of antigenic variation strategies used by different infective organisms means R-loop driven antigen variation is likely to be specific to *T. brucei*, this work does contribute to a growing understanding of how R-loops may induce recombination (Skourti-Stathaki and Proudfoot, 2014; Sollier and Cimprich, 2015). Indeed, to date only during mammalian Ig class switching has deliberate R-loop formation been linked to planned recombination (Basu *et al.*, 2011). Thus, dissecting how R-loops effect VSG switching could have wide-reaching implications for the understanding of genome stability in many settings.

## 7.4 Future directions

Data presented in this thesis provide the first genome-wide account of R-loops in a kinetoplastid genome, as well as the first characterisation of all three RNase H proteins in *T. brucei*. To date, R-loops have only been examined in *T. brucei* in the context of telomere biology, where their abundance was shown to be reduced after over-expression of TbRH1 (Nanavaty *et al.*, 2017). By examining R-loop distribution genome-wide a global perspective on how R-loops may contribute to genome stability and function is revealed. These findings offer the opportunity for several exciting lines of research to be explored in the future. Firstly, what exactly is the role of the *T. brucei* RNase H2 complex, and what makes it an essential enzyme? Is it in fact crucial that RNase H2 removes embedded ribonucleotides through the RER pathway? To establish this ribonucleotide load could be investigated in TbRH2A depleted parasites, and phenotype rescue could be attempted with an TbRH2A mutant protein capable of resolve RNA-DNA hybrids alone, as performed in *S.*

*cerevisiae* (Chon *et al.*, 2013). ChIP-seq could also be performed with both TbRH1 and TbRH2A to establish if the two enzymes target the same or different genome loci. Immunoprecipitation and mass spectrometry could also be enlisted to establish binding partners of the RNase H2 complex.

Another intriguing avenue would be to establish if there is a functional link between R-loops and kinetoplastid gene expression. This would actually involve examination of two questions. First, do R-loops act in RNA Pol II transcription initiation, a correlation that is frequently observed in other genomes yet has not be functionally characterised. Second, do R-loops contribute to multigenic pre-mRNA processing, which would be very novel if correct. Mapping of RNA Pol II binding with ChIP-seq in parasites lacking RNase H activity, perhaps by targeting TbRH2A for depletion with RNAi in the *Tbrh1*-/- background, may uncover changes at the TSS. Additionally, changes in RNA Pol II movement across the PTUs would help investigation into the role, if any, of R-loops in *trans*-splicing and polyadenylation. Finally, changes in transcription initiation- and processing-associated epigenetic markers could be investigated in the RNase H deficient cells, via ChIP-qPCR or ChIP-seq targeting H2A.Z, H2B.V and H4K10ac (Siegel *et al.*, 2009).

Detailed investigation of R-loop-driven VSG switching could also be pursued. Here there are many possible approaches: for example, RNA Pol I movement across the BES could be tracked via ChIP-seq in RNase H deficient cells to establish if pausing occurs due to R-loop formation. The 70-bp repeat sequence could also be genetically manipulated to establish if characteristics of the repeats facilitate R-loop formation. DNA replication could also be tracked with DNA combing technology (Calderano *et al.*, 2015) to establish if transcription-replication clashes are encouraged by R-loop formation and, if so, are these co-transcriptional or head-on? Lastly, in order to establish the factors that interact with RNA-DNA hybrids, which could potentially be involved in VSG recombination reactions, the S9.6 antibody might be utilise in a chromatin pull-down of associated proteins that could then be identified via mass spectrometry (Cristini *et al.*, 2018). In fact, this approach may also uncover proteins that might link R-loops to transcription initiation, *trans*-splicing or polyadenylation.

# 8 Appendices

```
## Generate enriched region coodinates from enriched bin coordinate file

# from tabulate import tabulate
import numpy
file = open("RH2A_unind_coverage_gt1.2.bedgraph", "r") ##Reads in the file of bin coordinates
lines = file.read().splitlines() ## Split this file by each line and read lines
length = (len(lines) - 1) ## calculate how many lines in the file
index = 0 ## set up an index for each line
Chrom = []
Start = []
End = []
## Generate a for loop to go through the file line by line
for line in lines:
    if index == 0:              # To print first chro and start coord
        line = lines[index]
        splitLine = line.split("\t")
        Chrom.append(splitLine[0])
        Start.append(splitLine[1])
    if index < length:                  ## only if the index is less then the total length
        line = lines[index]
        next_line = lines[index+1]
        splitLine = line.split("\t")
        splitNextLine = next_line.split("\t")
        if splitLine[2] != splitNextLine[1]:        ## is the end coord of line and start coord if next line different?
                Chrom.append(splitNextLine[0])
                Start.append(splitNextLine[1])
                End.append(splitLine[2])
    if index == length:         # print the last end coord
        line = lines[index]
        splitLine = line.split("\t")
        End.append(splitLine[2])
    else:
        exit
    index = index + 1 #This is nessecary to add the index to the next line
# print chrom
#` print Start
# print End

results = numpy.column_stack((Chrom, Start, End)) ##Create an array for chrom, start and end
# print results

file = open("RH2A_unind_enriched_region_coords.bed", "w") ##wirte out the array to a file
numpy.savetxt(file, results, delimiter = "  ", fmt = "%s")
file.close
```

**Figure 8.1 Python script to generate a list of DRIP enriched regions.**
Once DRIP-seq enrichment was found in 50 bp windows, those with enrichment of IP compared to input samples > 1.2 fold were combined and considered as DRIP enriched regions. 50 bp window coordinates were used as an input for the above script in bed format and were scanned to find those that are adjacent and display > 1.2 fold-change. The start coordinate of the first window to meet the criteria and end coordinate of last downstream window to do so were printed and exported in BED format of enriched regions.

```
1.  ## This script will find the GC skew over binned regions of fasta sequence
2.  ## Input it of fasta format of as many sequences as need.
3.  ## Window can be changes to alter the size of window in bp (non-sliding window)
4.  ## Input and ouput file names can be changed as required
5.
6.  from __future__ import print_function
7.
8.
9.  file = open("WT_Rloop_pos_gene_seq.fasta", "r") # Reads in the file of region coord
    inates
10. lines = file.read().splitlines() # Split this file by each line and read lines
11. index = 0 # Set up an index for each line
12. file = open("WT_Rloop_pos_gene_GCskew.bed", "w") # File to save GC skew data too
13.
14. values = []
15. window = 10
16. for seq in lines:
17.     if not seq.startswith(">"):                  # Skip fasta sequence names
18.         for i in range(0, len(seq), window):
19.             s = seq[i: i + window]
20.             g = s.count('G') + s.count('g')
21.             c = s.count('C') + s.count('c')
22.             try:
23.                 skew = (g - c) / float(g + c)
24.             except ZeroDivisionError:
25.                 skew = 0.0
26.             bins = str(skew) + "\t"
27.             print(bins, sep = ' ', end = '', file = file) # Print skew value of eac
    h bin on the same line, seperated by a tab
28.         print("\r", end = '', file = file) # Start a new line in file of skew data
    for next fasta seq
29.         index = index +1
30. file.close()
```

**Figure 8.2 Python script to find GC and AT skew of DNA sequences**
The above script was used to calculate GC or AT skew for sequences inputted in fasta format. The script for GC skew is shown as an example, AT skew script was identical where G was exchanged for A and C for T. Skew was calculated in 10 bp windows.

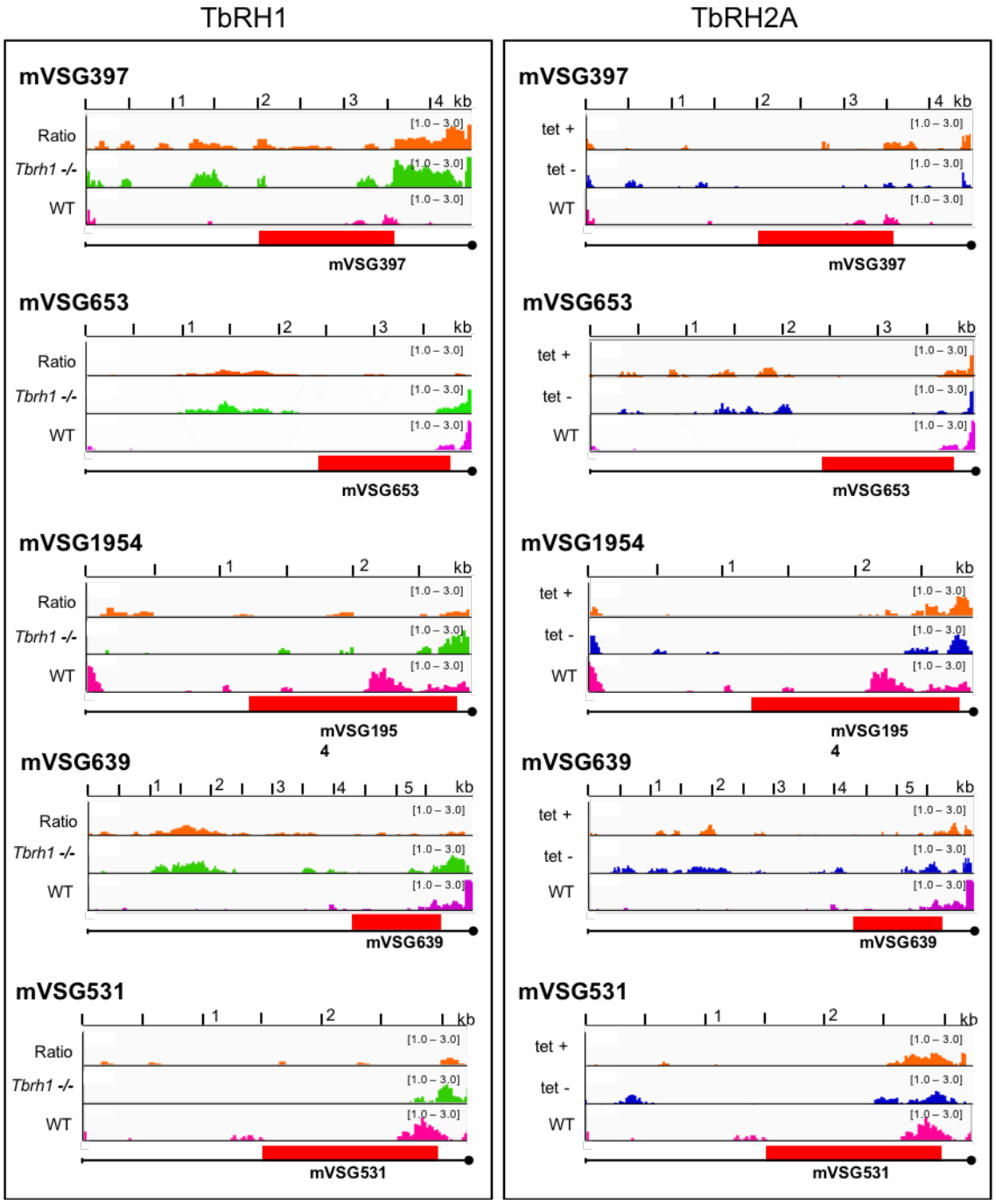



**Figure 8.3 DRIP-seq mapping for RNA-DNA hybrids to the mVSG expression sites.**
DRIP-seq signal enrichment is plotted across 5 available mVSG expression sites, for data generated in WT (pink), *Tbrh1-/-* (green), TbRH2A un-induced (blue) and induced (orange) (24 hr) cell lines. In all cases red box depicts mVSG coding regions, and circles indicate end of avaible contig sequence.

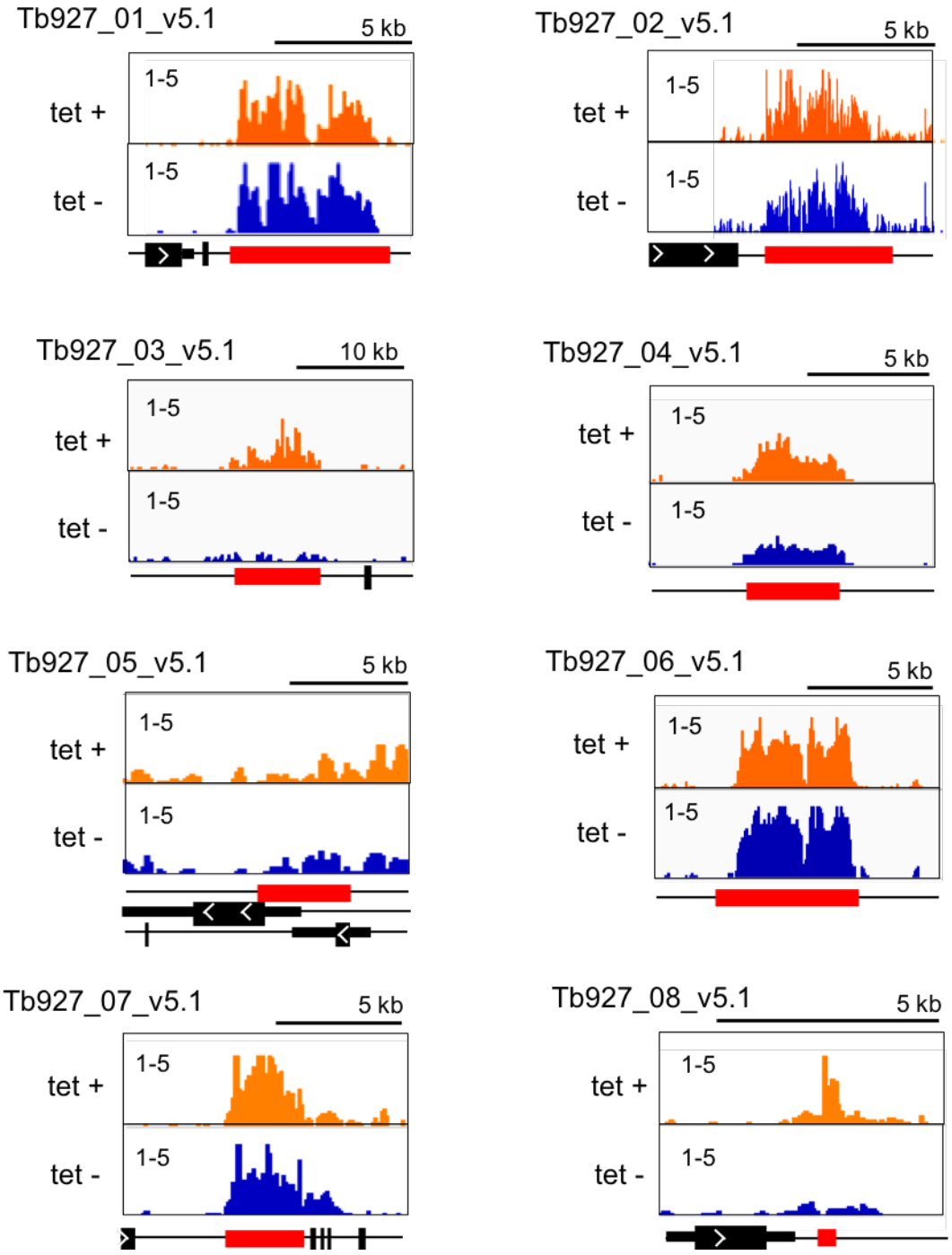



**Figure 8.4 DRIP-seq mapping for RNA-DNA hybrid at centromeres in TbRH2A RNAi parasites.** DRIP-seq enrichment of IP sample relative input is plotted across the 8 known centromeric regions for both induced (orange) and un-induced (blue) sample. In all cases red boxes depict centromeric repeats, thick black boxes are coding regions with transcription direction indicated with white arrows, and UTRs are shown as think black boxes.

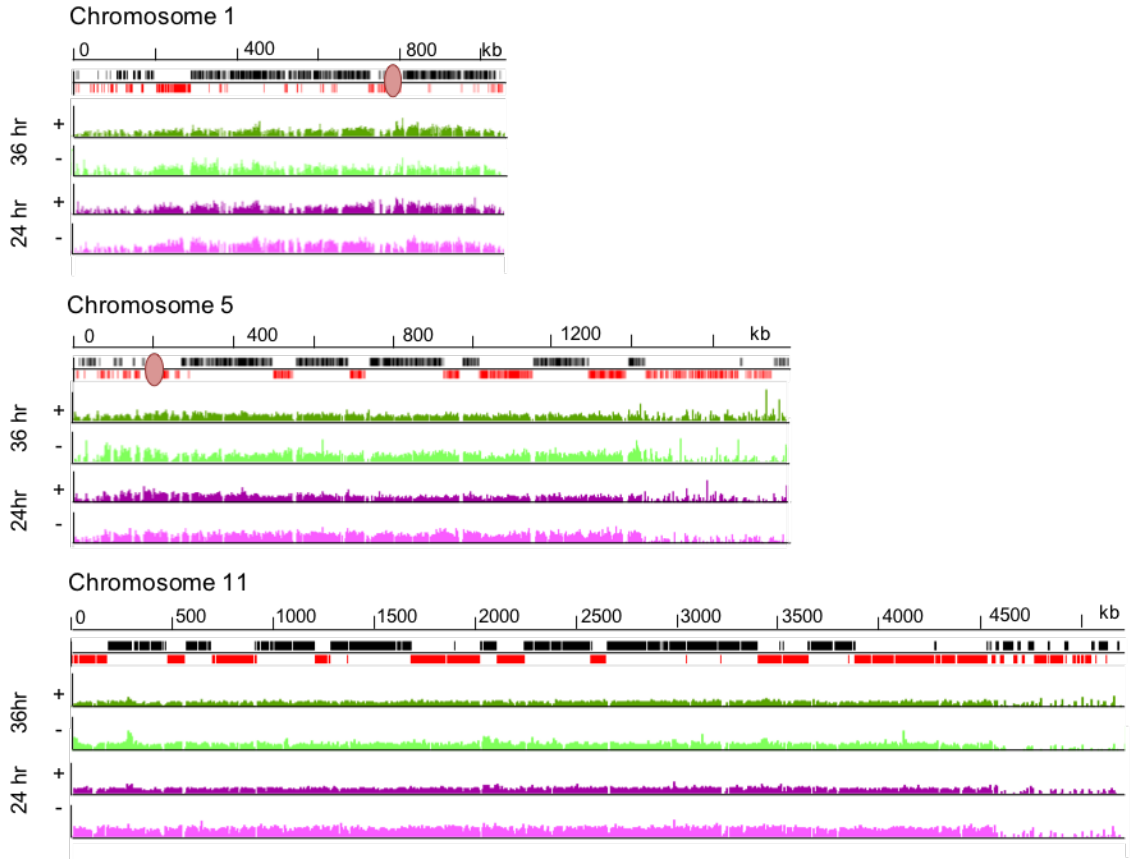



**Figure 8.5 γH2A ChIP-seq mapping to the Mb chromosomes in TbRH2A RNAi parasites.**
γH2A ChIP-seq mapping to three example Mb chromosomes (1, 5 and 11) is shown. Enrichment of IP mapping relative to input is shown for un-induced and induced samples taken after 24hr (pink) and 36 hr (green) of culture. Scales, 1-5 fold-change. Coding regions are shown in the top track with black depicting the sense strand, and red anti-sense.

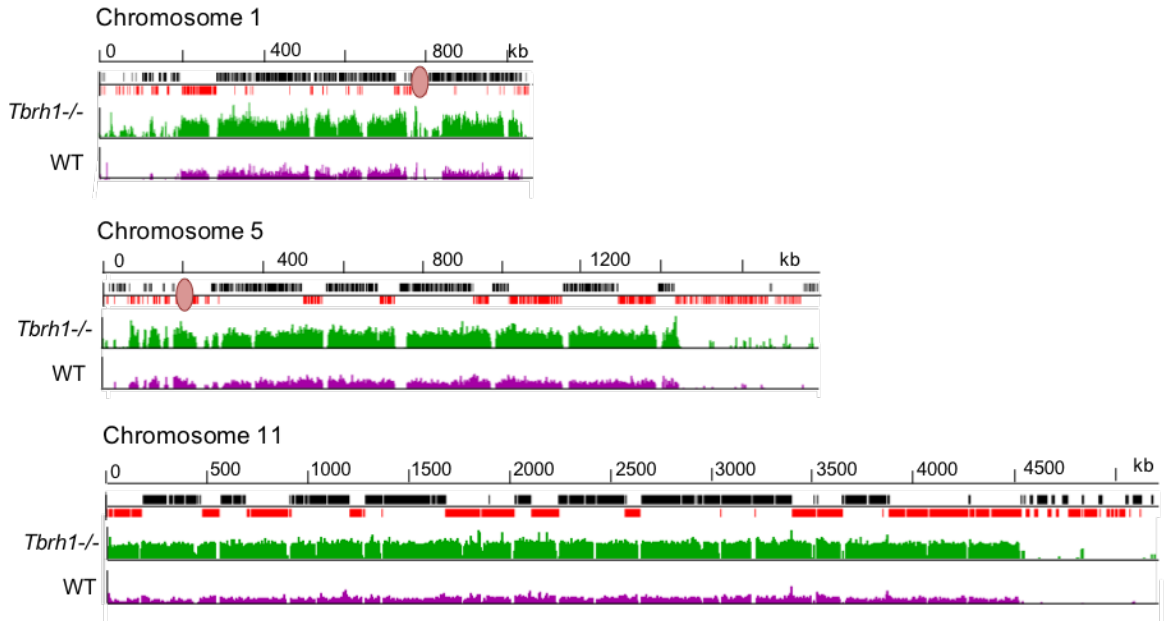



**Figure 8.6 γH2A ChIP-seq mapping to the Mb chromosomes in WT and *Tbrh1-/-* parasites.** γH2A ChIP-seq mapping to three example Mb chromosomes (1, 5 and 11) is shown. Enrichment of IP mapping relative to input is shown for WT (pink) and *Tbrh1-/-* (green) of culture. Scales, 1-5 fold-change. Coding regions are shown in the top track with black depicting the sense strand, and red anti-sense.

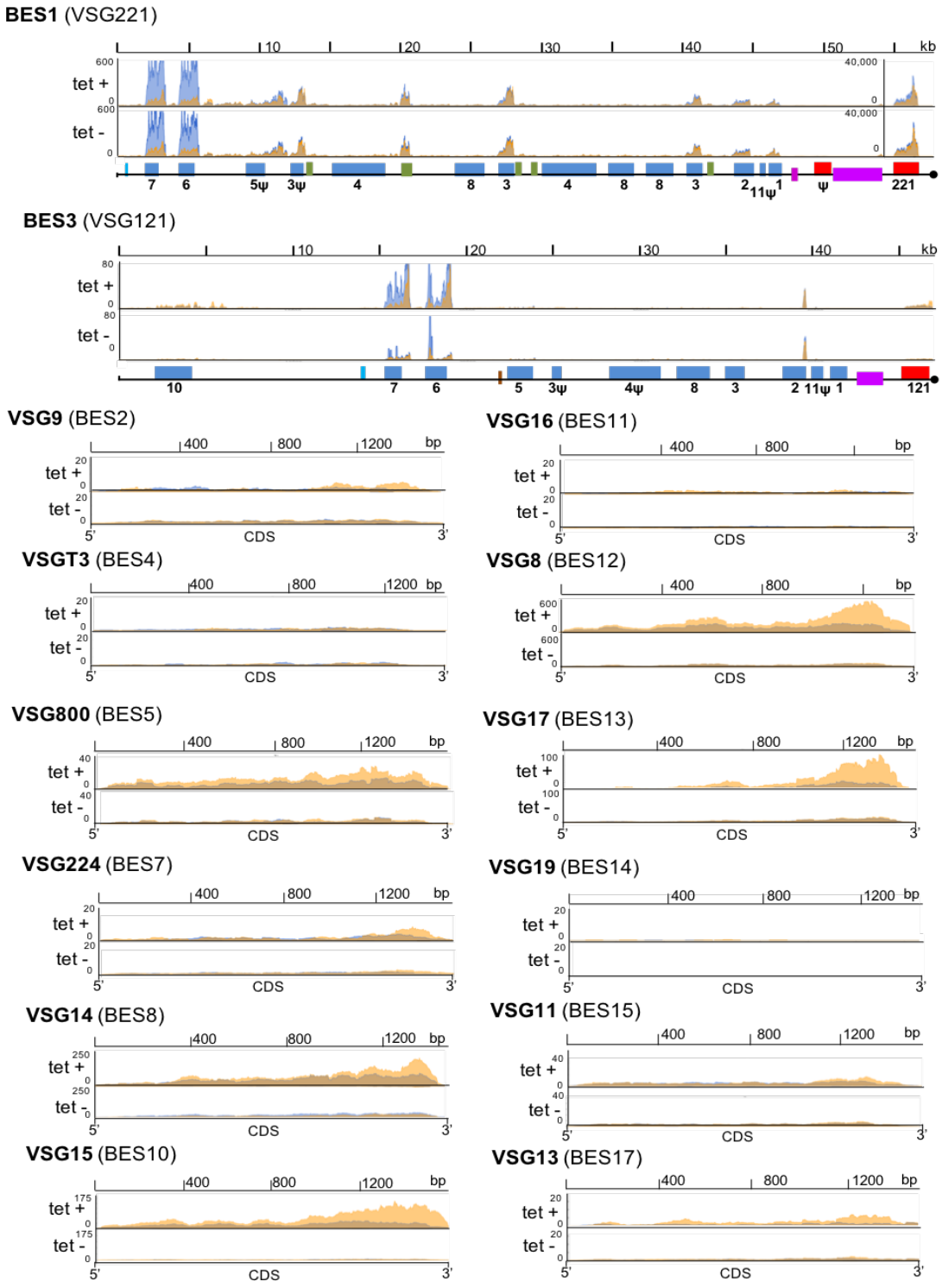



**Figure 8.7 RNA-seq mapping to silent BES-housed VSGs after 24 hr of TbRH2A depletion.** Normalised RNA-seq read depth abundance (y-axes) is plotted for two independent replicates (overlaid orange and blue) of *TbRH2A* RNAi parasites after 24 hr of growth, with (tet +) and without (tet -) induction. Read depth is shown relative to CDS position (x-axes) for BES1 and BES3 as well as a selection of the above VSG housed within BES identified by (Hertz-Fowler *et al.*, 2008); VSG identity numbers are from (Cross, Kim and Wickstead, 2014).

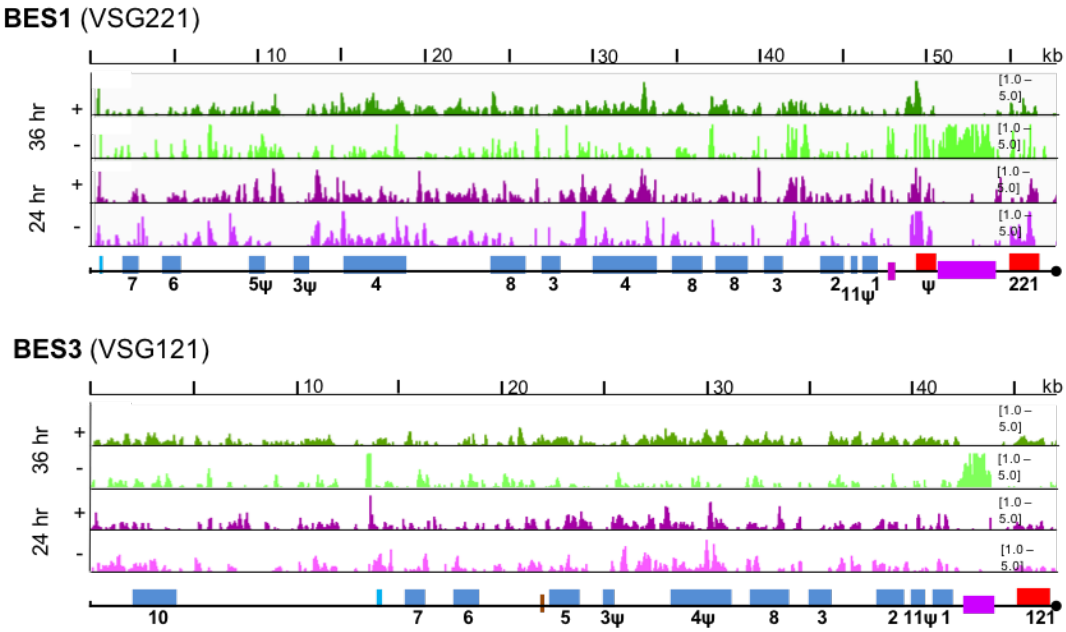



**Figure 8.8 γH2A ChIP-seq mapping to the BESs after 24hr of TbRH2A depletion.**
ChIP-seq signal enrichment is shown mapped to the active BES, BES1, and an inactive BES, BES3, (Hertz-Fowler *et al.*, 2008) as a ratio of IP to input signal for un-induced (lighter) and induced (darker) samples taken at both 24 (purple) and 36 (green) hr time points. Annotations are as described in Figure 6.1.

| Gene ID | Base Mean | Log2 FC | p value | Gene description |
|---|---|---|---|---|
| Tb427.BES98.4 | 80.91446371 | 4.854500301 | 4.76E-14 | expression site-associated gene 3 (ESAG3, pseudogene), putative |
| Tb427.BES134.3 | 196.4132048 | 4.632872805 | 3.40E-17 | transferrin-binding protein, putative |
| Tb427.BES134.6 | 1317.875299 | 4.16131121 | 4.38E-34 | variant surface glycoprotein (VSG) |
| Tb427.BES134.2 | 915.4331785 | 4.040173088 | 4.90E-20 | transferrin-binding protein, putative |
| Tb427.BES98.2 | 1719.674636 | 3.675707873 | 6.12E-18 | transferrin-binding protein, putative |
| Tb427.BES129.6 | 41.67769308 | 3.636744051 | 1.33E-07 | expression site-associated gene 3 (ESAG3, pseudogene), putative |
| Tb427.BES134.5 | 39.19467042 | 3.610256624 | 6.18E-05 | expression site-associated gene 3 (ESAG3, pseudogene), putative |
| Tb427.BES129.3 | 142.5747758 | 3.58134443 | 0.000000006 | transferrin-binding protein, putative |
| Tb427.BES98.1 | 803.919418 | 3.564188717 | 3.56E-16 | transferrin-binding protein, putative |
| Tb427.BES153.1 | 64.8268814 | 3.491335833 | 1.12E-10 | folate transporter, putative |
| Tb427.BES122.1 | 1848.159365 | 3.316942889 | 0.000116136 | transferrin-binding protein, putative |
| Tb427.BES15.1 | 63.08762692 | 3.209696259 | 1.04E-07 | folate transporter, putative |
| Tb427.BES28.1 | 147.0893493 | 3.187391215 | 3.15E-07 | transferrin-binding protein, putative |
| Tb427.BES98.3 | 167.5221562 | 3.134199178 | 1.28E-18 | expression site-associated gene 5 (ESAG5) protein, putative |
| Tb427.BES56.3 | 1015.486523 | 2.947401705 | 1.72E-10 | transferrin-binding protein, putative |
| Tb427.BES29.8 | 74.77286568 | 2.935857152 | 8.58E-08 | expression site-associated gene 1 (ESAG1) protein, putative |
| Tb427.BES56.1 | 58.9491296 | 2.917904869 | 1.99E-06 | folate transporter, putative |
| Tb427.BES98.4 | 38.26632636 | 2.869435937 | 1.59E-05 | expression site-associated gene 3 (ESAG3, pseudogene), putative |
| Tb427.BES29.5 | 104.4553261 | 2.74888531 | 2.35E-09 | expression site-associated gene 3 (ESAG3, pseudogene), putative |
| Tb427.BES29.6 | 144.1569569 | 2.741079309 | 2.24E-07 | expression site-associated gene 2 (ESAG2) protein, putative |
| Tb427.BES56.5 | 113.9970646 | 2.628604888 | 2.27E-11 | expression site-associated gene 3 (ESAG3, pseudogene), putative |
| Tb427.BES29.1 | 37.23482713 | 2.501653448 | 0.000159181 | folate transporter, putative |
| Tb927.10.10210 | 626.8045138 | 2.344035412 | 1.38E-06 | procyclin-associated gene 4 (PAG4) protein |
| Tb427.BES126.5 | 468.6077199 | 2.261115845 | 1.12E-06 | expression site-associated gene 3 (ESAG3, pseudogene), putative |

| | | | | |
|---|---|---|---|---|
| Tb427.BES126.9 | 51.38611308 | 2.113184858 | 0.000139011 | expression site-associated gene 3 (ESAG3) protein, putative |
| Tb427.BES10.2 | 673.9775699 | 1.949775237 | 1.53E-05 | transferrin-binding protein, putative |
| Tb427.BES29.3 | 1423.189122 | 1.874848581 | 7.59E-06 | transferrin-binding protein, putative |
| Tb927.11.17440 | 948.8572373 | 1.801535235 | 1.85E-09 | variant surface glycoprotein (VSG), degenerate |
| Tb427.BES29.9 | 7767.543212 | 1.652226649 | 2.02E-12 | variant surface glycoprotein (VSG) |
| Tb427.BES56.13 | 857.6695129 | 1.562144419 | 4.78E-07 | variant surfave glycoprotein (VSG) |
| Tb927.6.510 | 742.2088554 | 1.333807798 | 1.19E-05 | GPEET procyclin |
| Tb427.BES98.12 | 624.7495503 | 1.204289916 | 9.61E-07 | variant surface glycoprotein (VSG) |

**Table 8.1 Antigenic variation associated genes are upregulated after 24hr of TbRH2A RNAi**
List of genes found to be statistically up regulated after 24 hr TbRH2A knockdown with RNAi. The log2 fold-change (FC) of normalisaed read counts in tet-induced samples relative to un-induced is shown for each gene, as well as p value of significance and gene description.

| Gene ID | Base Mean | Log2 FC | p value | Gene description |
|---|---|---|---|---|
| Tb927.7.6540 | 52.11121318 | 6.925801127 | 3.96E-07 | variant surface glycoprotein (VSG, atypical), putative |
| Tb927.11.19410 | 9.82459566 | 6.732055091 | 0.000654909 | variant surface glycoprotein (VSG), putative |
| Tb927.9.16130 | 9.650666209 | 6.705819231 | 0.000776044 | variant surface glycoprotein (VSG, pseudogene), putative |
| Tb927.9.17970 | 9.611356636 | 6.702508309 | 0.00055714 | variant surface glycoprotein (VSG, pseudogene), putative |
| Tb927.9.16130 | 8.528470356 | 6.529239771 | 0.001068167 | variant surface glycoprotein (VSG, pseudogene), putative |
| Tb927.11.18870 | 14.71649152 | 5.797281008 | 0.000552866 | variant surface protein (VSG), putative (pseudogene) |
| Tb927.9.16100 | 10.15205259 | 5.223211467 | 0.003040099 | variant surface glycoprotein (VSG, pseudogene), putative |
| Tb427.BES153.1 | 50.46346806 | 5.064435689 | 6.70E-09 | folate transporter, putative |
| Tb927.9.400 | 13.4299292 | 4.987509649 | 0.00198716 | variant surface glycoprotein (VSG, pseudogene), putative |
| Tb927.7.6530 | 181.2744897 | 4.958744655 | 0.00012341 | variant surface glycoprotein, degenerate |
| Tb427.BES134.6 | 1557.938318 | 4.908508218 | 0.000103825 | variant surface glycoprotein (VSG) |
| Tb927.9.17430 | 14.35338573 | 4.824059228 | 0.001091433 | variant surface glycoprotein (VSG, pseudogene), putative |
| Tb427.BES134.3 | 137.3508479 | 4.754244349 | 0.00123756 | transferrin-binding protein, putative |
| Tb927.7.1980 | 19.3064374 | 4.7504538 | 8.44E-05 | retrotransposon hot spot protein 7 (RHS7), putative |
| Tb927.7.1960 | 52.957613 | 4.541740671 | 2.87E-09 | retrotransposon hot spot protein 7 (RHS7), putative |
| Tb427.BES134.2 | 622.0362066 | 4.50528691 | 0.001880094 | transferrin-binding protein, putative |
| Tb927.9.16350 | 28.16973622 | 4.495286261 | 4.03E-05 | variant surface glycoprotein (VSG, pseudogene), putative |
| Tb927.7.6520 | 174.0741508 | 4.439723679 | 8.34E-12 | variant surface glycoprotein, frameshift |
| Tb927.9.400 | 17.79488162 | 4.424338667 | 0.000530527 | variant surface glycoprotein (VSG, pseudogene), putative |
| Tb927.9.16130 | 40.31895526 | 4.235957117 | 6.18E-06 | variant surface glycoprotein (VSG, pseudogene), putative |
| Tb927.9.16060 | 13.60311146 | 4.208796447 | 0.002128565 | hypothetical protein |
| Tb927.3.2550 | 53.73419138 | 4.108403155 | 0.000006747 | hypothetical protein, conserved |
| Tb927.9.18130 | 17.97321897 | 4.080665434 | 0.000739989 | variant surface glycoprotein (VSG, pseudogene), putative |
| Tb927.9.16170 | 98.41075253 | 4.011319195 | 8.12E-08 | variant surface glycoprotein (VSG, pseudogene), putative |
| Tb927.11.17440 | 1689.413382 | 3.915725633 | 1.66E-06 | variant surface glycoprotein (VSG), degenerate |
| Tb427.BES56.1 | 33.9857857 | 3.897738906 | 0.000012517 | folate transporter, putative |
| Tb427.BES64.2 | 2360.214434 | 3.857551673 | 4.79E-06 | variant surface glycoprotein (VSG) |
| Tb927.9.16370 | 39.47922286 | 3.848001048 | 2.32E-05 | variant surface glycoprotein (VSG, pseudogene), putative |
| Tb927.9.16160 | 13.56820979 | 3.823669368 | 0.003125917 | variant surface glycoprotein (VSG, pseudogene), putative |

| | | | | |
|---|---|---|---|---|
| Tb927.9.350 | 32.42208327 | 3.806051867 | 1.26E-05 | variant surface glycoprotein (VSG, pseudogene), putative |
| Tb427.BES134.5 | 16.24052381 | 3.79930773 | 0.001260451 | expression site-associated gene 3 (ESAG3, pseudogene), putative |
| Tb927.3.2540 | 27.88594666 | 3.769763367 | 8.76E-05 | variant surface glycoprotein (VSG)-related, putative |
| Tb927.3.2500 | 39.6974236 | 3.536648275 | 0.000265094 | Trypanosomal VSG domain containing protein, putative |
| Tb427.BES15.1 | 43.21821073 | 3.510988205 | 1.24E-06 | folate transporter, putative |
| Tb927.9.16640 | 46.71814216 | 3.453973341 | 7.25E-06 | variant surface glycoprotein (VSG), putative |
| Tb927.3.2580 | 48.73238532 | 3.417205899 | 1.18E-05 | hypothetical protein, conserved |
| Tb927.7.6510 | 117.2153427 | 3.362454632 | 1.25E-06 | variant surface glycoprotein, degenerate |
| Tb427.BES129.6 | 22.19878306 | 3.344558174 | 0.002564161 | expression site-associated gene 3 (ESAG3, pseudogene), putative |
| Tb427.BES29.1 | 33.72605882 | 3.326101763 | 3.50E-05 | folate transporter, putative |
| Tb427.BES129.3 | 92.45314016 | 3.325156116 | 4.51E-05 | transferrin-binding protein, putative |
| Tb927.9.16070 | 31.49144846 | 3.313301532 | 0.000207691 | hypothetical protein |
| Tb427.BES98.4 | 38.95353032 | 3.287313173 | 7.01E-05 | expression site-associated gene 3 (ESAG3, pseudogene), putative |
| Tb927.6.510 | 2743.247315 | 2.888428876 | 3.13E-27 | GPEET procyclin |
| Tb927.9.13160 | 34.81573771 | 2.861135188 | 0.000300634 | hypothetical protein |
| Tb927.9.760 | 34.64970861 | 2.788927427 | 0.001229052 | hypothetical protein |
| Tb927.3.2520 | 25.86796758 | 2.768736243 | 0.000677067 | expression site-associated gene 1 (ESAG1) protein, putative |
| Tb927.3.2530 | 44.31005596 | 2.732369665 | 0.000106956 | expression site-associated gene 11 (ESAG11), degenerate |
| Tb427.BES98.4 | 26.62190256 | 2.654005325 | 0.001033015 | expression site-associated gene 3 (ESAG3, pseudogene), putative |
| Tb11.v5.0353 | 42.78081849 | 2.599420192 | 8.65E-05 | retrotransposon hot spot (RHS) protein, putative |
| Tb927.9.14660 | 224.0843812 | 2.539511815 | 0.000000015 | SLACS reverse transcriptase, putative (pseudogene) |
| Tb927.6.810 | 125.5648416 | 2.527923271 | 4.94E-06 | hypothetical protein |
| Tb427.BES56.5 | 74.55858437 | 2.520131727 | 2.85E-05 | expression site-associated gene 3 (ESAG3, pseudogene), putative |
| Tb927.7.1990 | 54.39904771 | 2.446660193 | 6.77E-05 | retrotransposon hot spot protein 7 (RHS7), point mutation |
| Tb427.BES98.3 | 109.3919206 | 2.365870381 | 8.64E-08 | expression site-associated gene 5 (ESAG5) protein, putative |
| Tb11.v5.0797 | 114.1711689 | 2.327838541 | 4.21E-06 | flagellar calcium-binding protein |
| Tb927.10.13060 | 48.39581998 | 2.274925451 | 0.00197326 | hypothetical protein |
| Tb927.7.7560 | 443.7862501 | 2.268750814 | 0.000000532 | hypothetical protein |
| Tb927.7.6500 | 371.4551197 | 2.209533977 | 5.38E-06 | variant surface glycoprotein (VSG), putative |
| Tb927.9.210 | 46.92075342 | 2.193052773 | 0.000171584 | variant surface glycoprotein (VSG, pseudogene), putative |

| | | | | |
|---|---|---|---|---|
| Tb927.9.4840 | 114.122153 | 2.1298012 | 0.001279235 | glucose transporter (pseudogene), putative |
| Tb927.2.680 | 345.4546485 | 2.111870366 | 1.62E-05 | Protein of unknown function (DUF1181), putative |
| Tb927.6.470 | 485.0195869 | 2.017178561 | 8.74E-06 | gene related to expression site-associated gene 2 (GRESAG2), putative |
| Tb11.v5.0702 | 114.8166629 | 1.938027666 | 0.000548371 | calpain-like cysteine peptidase, putative |
| Tb927.6.110 | 192.7403593 | 1.904900892 | 1.61E-05 | hypothetical protein |
| Tb927.8.7300 | 302.5703561 | 1.858756115 | 2.62E-12 | variant surface glycoprotein (VSG)-related, putative |
| Tb927.11.4080 | 426.878734 | 1.852426671 | 0.000000001 | variant surface glycoprotein (VSG)-related, putative |
| Tb427.BES126.5 | 218.1031774 | 1.843028523 | 0.000354384 | expression site-associated gene 3 (ESAG3, pseudogene), putative |
| Tb927.2.6310 | 60.93971409 | 1.800364433 | 0.002121981 | iron/ascorbate oxidoreductase family protein, putative |
| Tb927.10.10210 | 489.4000215 | 1.761289861 | 2.73E-12 | procyclin-associated gene 4 (PAG4) protein |
| Tb927.10.10230 | 820.7554285 | 1.758441501 | 2.08E-14 | procyclin-associated gene 5 (PAG5) protein |
| Tb927.6.480 | 67.64540321 | 1.739061382 | 0.000383447 | surface protein EP3-2 procyclin precursor |
| Tb927.8.270 | 101.5036729 | 1.731211535 | 8.77E-05 | variant surface glycoprotein, degenerate |
| Tb927.1.3360 | 446.0502814 | 1.682070187 | 8.09E-05 | hypothetical protein |
| Tb927.1.280 | 51.20720474 | 1.663429373 | 0.002921437 | Protein of unknown function (DUF1181), putative |
| Tb927.10.10240 | 770.0499371 | 1.656041053 | 4.02E-12 | procyclin-associated gene 1 (PAG1) protein |
| Tb927.8.7320 | 221.5496491 | 1.652145682 | 9.62E-09 | variant surface glycoprotein (VSG)-related, putative |
| Tb927.4.130 | 416.2726939 | 1.595604064 | 3.23E-07 | receptor-type adenylate cyclase GRESAG 4, pseudogene, putative |
| Tb927.7.150 | 207.3910624 | 1.568650168 | 0.001413761 | variant surface glycoprotein, degenerate |
| Tb927.4.190 | 278.3005905 | 1.541229888 | 0.000516936 | retrotransposon hot spot protein (RHS, pseudogene), putative |
| Tb927.10.10220 | 962.6179284 | 1.509791474 | 1.17E-07 | procyclin-associated gene 2 (PAG2) protein |
| Tb927.10.10270 | 174.210351 | 1.499458754 | 0.001479878 | hypothetical protein |
| Tb927.5.280 | 63.04040369 | 1.49209763 | 0.001821499 | retrotransposon hot spot protein 3 (RHS3), interrupted, degenerate |
| Tb927.2.720 | 345.8331458 | 1.490494453 | 4.40E-05 | retrotransposon hot spot protein 5 (RHS5), degenerate |
| Tb927.6.520 | 1082.406497 | 1.464066503 | 6.88E-06 | EP3-2 procyclin |
| Tb927.2.910 | 436.5109616 | 1.464036617 | 2.97E-06 | expression site-associated gene 4 (ESAG4), degenerate |
| Tb927.4.240 | 300.5563346 | 1.452479593 | 0.00140124 | retrotransposon hot spot protein 4 (RHS4), degenerate |
| Tb927.10.10260 | 2008.912194 | 1.442501305 | 3.94E-06 | EP1 procyclin |
| Tb927.6.100 | 105.8013248 | 1.426997378 | 0.002129678 | receptor-type adenylate cyclase GRESAG 4, pseudogene, putative |
| Tb927.9.18140 | 11383.24334 | 1.418870436 | 7.13E-06 | variant surface glycoprotein (VSG, pseudogene), putative |

| | | | | |
|---|---|---|---|---|
| Tb927.10.10250 | 621.1927578 | 1.349448564 | 4.45E-06 | EP2 procyclin |
| Tb927.6.210 | 367.075393 | 1.337717246 | 0.001514039 | leucine-rich repeat protein (LRRP, pseudogene), putative |
| Tb927.5.4020 | 414.3249901 | 1.334263638 | 0.000379904 | hypothetical protein |
| Tb927.1.275 | 195.0247213 | 1.327932369 | 0.00017662 | expression site-associated gene 4 (ESAG4), pseudogene |
| Tb927.3.2590 | 1759.247304 | 1.309063016 | 1.62E-06 | hypothetical protein |
| Tb927.3.1500 | 1585.561051 | 1.272159875 | 2.37E-06 | variant surface glycoprotein (VSG)-related, putative |
| Tb927.1.70 | 1268.318947 | 1.265998034 | 0.000134422 | retrotransposon hot spot protein 4 (RHS4), putative |
| Tb927.7.6490 | 4323.283585 | 1.250545624 | 1.48E-05 | hypothetical protein, conserved |
| Tb927.3.1510 | 387.2309373 | 1.238675074 | 9.90E-06 | variant surface glycoprotein (VSG)-related, putative |
| Tb927.5.5530 | 1227.091308 | 1.222801364 | 2.39E-06 | Variant Surface Glycoprotein, putative |
| Tb927.6.530 | 156.8191787 | 1.214100699 | 0.000697794 | procyclin associated gene 3 (PAG3) protein |
| Tb927.2.3330 | 151.7381566 | 1.192337209 | 0.002112579 | hypothetical protein |
| Tb927.6.340 | 319.062522 | 1.16057376 | 0.000561548 | receptor-type adenylate cyclase GRESAG 4, pseudogene, putative |
| Tb927.2.660 | 1369.101818 | 1.155089376 | 0.000597791 | expression site-associated gene 4 (ESAG4), degenerate |
| Tb07.30D13.60 | 320.1801847 | 1.145649936 | 5.40E-05 | hypothetical protein, conserved (pseudogene) |
| Tb927.7.6570 | 1149.455028 | 1.128136503 | 5.46E-05 | Enriched in surface-labeled proteome protein 17 |
| Tb927.2.1260 | 660.570418 | 1.116307126 | 0.000201577 | expression site-associated gene 4 (ESAG4), degenerate |
| Tb927.7.6840 | 751.799654 | 1.113600086 | 0.002743568 | Centrosomal spindle body, CEP44, putative |
| Tb927.6.140 | 2162.058647 | 1.10856309 | 9.84E-05 | retrotransposon hot spot protein 5 (RHS5), putative |
| Tb927.6.690 | 490.3490536 | 1.106868329 | 9.41E-05 | hypothetical protein, conserved |
| Tb927.3.5690 | 1919.948201 | 1.034801434 | 7.06E-06 | hypothetical protein, conserved |
| Tb11.v5.0226 | 155.7227032 | 1.029097031 | 0.003270251 | invariant surface glycoprotein, putative |
| Tb927.5.2260 | 167.3798552 | 1.00679401 | 0.002141654 | conserved protein |
| Tb927.11.12710 | 1242.658229 | 1.005622583 | 0.000785991 | variant surface glycoprotein (VSG)-related, putative |

**Table 8.2 Mainly antigenic variation associated genes are upregulated after 36 hr of TbRH2A RNAi**
List of genes found to be statistically up regulated after 36 hr TbRH2A knockdown with RNAi. The log2 fold-change (FC) of normalisaed read counts in tet-induced samples relative to un-induced is shown for each gene, as well as p value of significance and gene description.

| Gene ID | Base Mean | Log2 FC | p value | Gene description |
|---|---|---|---|---|
| Tb927.9.12550 | 419.475039 | -1.803595591 | 2.80E-07 | glycerol kinase, glycosomal |
| Tb927.10.14140 | 9414.531185 | -1.674005926 | 1.03E-24 | pyruvate kinase 1 |
| Tb927.11.11980 | 621.2182858 | -1.665807523 | 7.62E-06 | cytoskeleton-associated protein 15 |
| Tb427.BES40.20 | 8821.839598 | -1.567112409 | 4.38E-12 | expression site-associated gene 1 (ESAG1) protein, putative |
| Tb927.10.8230 | 5464.790321 | -1.519769327 | 1.81E-18 | protein disulfide isomerase 2 |
| Tb927.6.4300 | 434.1677737 | -1.512632505 | 0.000341599 | glyceraldehyde 3-phosphate dehydrogenase, glycosomal |
| Tb427.BES40.22 | 3610039.441 | -1.512092644 | 0.000000137 | variant surface glycoprotein (VSG) |
| Tb927.9.12610 | 760.8971767 | -1.480026204 | 6.75E-05 | glycerol kinase, glycosomal |
| Tb11.v5.0531 | 290.2426473 | -1.414328275 | 0.000164914 | fructose-bisphosphate aldolase, glycosomal, putative |
| Tb927.10.2890 | 92.23681948 | -1.410642832 | 0.001201291 | enolase |
| Tb11.v5.0668 | 534.7367365 | -1.402773873 | 2.69E-10 | protein disulfide isomerase |
| Tb927.9.12630 | 798.5237588 | -1.381058966 | 0.0003749 | glycerol kinase, glycosomal |
| Tb927.1.2390 | 2256.684615 | -1.350381971 | 1.48E-07 | beta tubulin |
| Tb927.11.715 | 370.6775856 | -1.348965155 | 0.000005261 | dynein light chain LC8, putative |
| Tb427.BES40.7 | 2240.75733 | -1.341675389 | 3.11E-05 | receptor-type adenylate cyclase, putative |
| Tb927.1.700 | 5749.735276 | -1.339352437 | 2.86E-06 | phosphoglycerate kinase |
| Tb927.4.4860 | 350.8721809 | -1.330997659 | 2.33E-08 | amino acid transporter 8, putative |
| Tb427.BES40.16 | 15582.84102 | -1.271857462 | 1.40E-10 | expression site-associated gene 3 (ESAG3) protein, putative |
| Tb927.10.6880 | 209.7609699 | -1.24521761 | 9.72E-05 | glyceraldehyde 3-phosphate dehydrogenase, cytosolic |
| Tb927.3.1470 | 186.061793 | -1.242631355 | 6.22E-05 | variant surface glycoprotein (VSG)-related, putative |
| Tb927.11.880 | 31003.87333 | -1.210523996 | 2.22E-07 | cyclophilin a |
| Tb927.11.1090 | 307.1177856 | -1.203734963 | 9.75E-06 | calpain-like protein, putative |
| Tb927.11.18680 | 1094.614237 | -1.197770312 | 1.39E-08 | dynein light chain LC8, putative |
| Tb11.v5.0540 | 691.345731 | -1.192342859 | 3.32E-08 | amino acid transporter 8, putative |
| Tb927.7.1780 | 1760.175306 | -1.185127344 | 3.78E-07 | Adenine phosphoribosyltransferase, putative |
| Tb927.4.4840 | 302.4570039 | -1.184088841 | 1.68E-05 | amino acid transporter 7, putative |
| Tb927.2.4590 | 2454.867917 | -1.162772126 | 3.80E-07 | branched-chain amino acid aminotransferase, putative |
| Tb927.11.7380 | 6439.120443 | -1.161239686 | 2.35E-13 | glycerol-3-phosphate dehydrogenase (FAD-dependent), mitochondrial |

| | | | | |
|---|---|---|---|---|
| Tb427.BES40.10 | 17893.51703 | -1.153481308 | 5.62E-07 | expression site-associated gene 3 (ESAG3) protein, putative |
| Tb927.3.4070 | 10489.68528 | -1.152675423 | 1.69E-05 | Pyruvate transporter, putative |
| Tb927.10.11300 | 682.3798256 | -1.139914534 | 6.19E-08 | paraflagellar rod component, putative |
| Tb927.8.5120 | 3302.43861 | -1.137615658 | 4.57E-07 | cytochrome c |
| Tb927.3.2960 | 10857.00718 | -1.136727203 | 4.03E-11 | inosine-adenosine-guanosine-nucleosidehydrolase |
| Tb11.v5.0516 | 425.7969728 | -1.135311815 | 1.13E-06 | glyceraldehyde 3-phosphate dehydrogenase, cytosolic, putative |
| Tb927.10.4560 | 2090.744141 | -1.132742929 | 1.39E-10 | elongation factor 2 |
| Tb927.9.12110 | 3011.7123 | -1.125345273 | 6.00E-13 | 6-phosphogluconate dehydrogenase, decarboxylating |
| Tb927.3.3270 | 40034.84321 | -1.113142024 | 6.10E-05 | ATP-dependent 6-phosphofructokinase, glycosomal |
| Tb927.5.4530 | 1068.250128 | -1.093810961 | 3.59E-07 | Long chain fatty acyl elongase |
| Tb927.11.16350 | 738.3691773 | -1.093564168 | 0.000552014 | hypothetical protein, conserved |
| Tb927.2.6150 | 278.5767622 | -1.085138987 | 0.000215295 | adenosine transporter 2 |
| Tb927.9.10310 | 2413.384703 | -1.08115305 | 0.000000091 | mitochondrial carrier protein 11 |
| Tb927.11.16760 | 5407.474251 | -1.071429426 | 2.58E-08 | T-complex protein 1, alpha subunit, putative |
| Tb927.1.2360 | 218.0001538 | -1.067127513 | 0.000660849 | alpha tubulin |
| Tb927.3.1520 | 229.1075979 | -1.065007102 | 9.52E-05 | variant surface glycoprotein (VSG)-related, putative |
| Tb927.11.12120 | 1057.604948 | -1.059029585 | 0.00023701 | RNA-binding protein, putative |
| Tb927.10.11700 | 1206.434069 | -1.055943595 | 2.29E-05 | guanylate kinase, putative |
| Tb927.9.6210 | 4910.126175 | -1.05526352 | 3.28E-13 | arginine kinase |
| Tb427.BES40.17 | 628.3313574 | -1.04754019 | 0.000939984 | hypothetical protein |
| Tb927.11.14090 | 4710.993038 | -1.046230144 | 1.02E-08 | Double RNA binding domain protein 18 |
| Tb927.11.5520 | 6890.768777 | -1.042013426 | 2.81E-10 | triosephosphate isomerase |
| Tb927.8.8330 | 7398.56821 | -1.038676453 | 3.18E-08 | cysteine peptidase, putative |
| Tb927.11.13500 | 10235.89331 | -1.029364307 | 4.84E-13 | par1 |
| Tb927.4.4870 | 251.9304803 | -1.027673114 | 0.000911145 | amino acid transporter, putative |
| Tb927.11.9590 | 32030.66425 | -1.025530683 | 3.78E-07 | S-adenosylhomocysteine hydrolase, putative |
| Tb927.6.3840 | 13071.99154 | -1.018677696 | 4.54E-12 | reticulon domain protein |
| Tb927.9.10230 | 9873.811841 | -1.017246088 | 4.42E-11 | Co-chaperone protein P23 |
| Tb927.9.10400 | 2536.600137 | -1.016564649 | 0.000004435 | hypothetical protein, conserved |

| | | | | |
|---|---|---|---|---|
| Tb927.11.7460 | 1799.190188 | -1.007372253 | 0.000002871 | luminal binding protein 1 (BiP), putative |
| Tb927.7.2240 | 4649.670772 | -0.99714677 | 8.67E-09 | hypothetical protein, conserved |
| Tb927.10.4570 | 15523.37547 | -0.992707051 | 1.10E-07 | elongation factor 2 |
| Tb927.10.10920 | 463.921429 | -0.991885845 | 1.89E-05 | heat shock protein, putative |
| Tb927.10.14170 | 1935.180161 | -0.990737218 | 8.98E-10 | aquaglyceroporin 2 |
| Tb927.10.15760 | 3414.968275 | -0.989259681 | 5.43E-09 | hypothetical protein, conserved |
| Tb927.5.3810 | 1667.711073 | -0.979245139 | 0.000586268 | orotidine-5-phosphate decarboxylase/orotate phosphoribosyltransferase, putative |
| Tb927.4.2530 | 7285.637991 | -0.977468582 | 3.48E-10 | hypothetical protein, conserved |
| Tb927.10.3640 | 183.9941052 | -0.976193235 | 0.000932917 | nuclear transmembrane protein, putative |
| Tb927.10.1390 | 2877.29184 | -0.970794792 | 0.003158761 | hypoxanthine-guanine phosphoribosyltransferase, putative |
| Tb927.11.16400 | 3362.795615 | -0.969582914 | 2.33E-09 | kinetoplast-associated protein 3, putative |
| Tb927.10.12820 | 3987.230204 | -0.969090137 | 3.54E-08 | hypothetical protein, conserved |
| Tb927.1.2400 | 402.1086263 | -0.960923446 | 0.000211812 | alpha tubulin |
| Tb927.4.1300 | 5497.119056 | -0.957169605 | 1.01E-07 | amidinotransferase, putative |
| Tb927.9.4680 | 8166.964166 | -0.955930435 | 2.99E-11 | Eukaryotic initiation factor 4A-1 |
| Tb927.9.8950 | 7372.783409 | -0.953330821 | 1.32E-10 | metallo- peptidase, Clan M- Family M48 |
| Tb927.11.1900 | 8820.546413 | -0.951284005 | 0.000001095 | T-complex protein 1, beta subunit, putative |
| Tb927.11.3240 | 7565.151696 | -0.948018576 | 1.86E-08 | T-complex protein 1, zeta subunit, putative |
| Tb927.11.2400 | 1358.051051 | -0.944933316 | 4.35E-07 | Flabarin-like protein |
| Tb927.8.7410 | 8859.500739 | -0.944741761 | 3.05E-07 | calreticulin, putative |
| Tb927.8.5460 | 818.7350987 | -0.941870284 | 0.000272996 | Flagellar calcium-binding 44 kDa protein |
| Tb927.5.3400 | 5619.667734 | -0.941606433 | 3.15E-08 | calcium-translocating P-type ATPase |
| Tb927.10.2190 | 2292.796451 | -0.939270382 | 5.99E-08 | Protein of unknown function (DUF667), putative |
| Tb927.7.5790 | 4694.063269 | -0.938964022 | 4.45E-09 | protein disulfide isomerase, putative |
| Tb927.11.15910 | 3706.934352 | -0.935981401 | 7.55E-09 | iron superoxide dismutase |
| Tb927.10.830 | 2570.191047 | -0.932895213 | 3.53E-07 | adenylate kinase, putative |
| Tb927.11.16130 | 10195.10265 | -0.926740296 | 2.53E-10 | nucleoside diphosphate kinase |
| Tb11.v5.0838 | 2825.059887 | -0.92241439 | 1.10E-07 | leucine-rich repeat protein (LRRP), putative |

| | | | | |
|---|---|---|---|---|
| Tb927.7.4160 | 3030.046525 | -0.920726322 | 0.000000003 | Fatty acid elongase |
| Tb11.v5.0538 | 1221.266373 | -0.918603105 | 0.000025941 | 69 kDa paraflagellar rod protein, putative |
| Tb927.6.3750 | 3590.288892 | -0.91525592 | 2.38E-08 | heat shock 70 kDa protein, mitochondrial precursor, putative |
| Tb927.8.6390 | 4442.273589 | -0.911574639 | 0.000126627 | lysophospholipase, putative |
| Tb927.9.12700 | 2929.855727 | -0.90958146 | 1.74E-05 | phospholipase A1, putative |
| Tb927.10.10140 | 4575.533482 | -0.908684554 | 8.68E-10 | paraflagellar rod component, putative |
| Tb927.10.1060 | 5062.358733 | -0.908296115 | 2.17E-06 | T-complex protein 1, delta subunit, putative |
| Tb927.11.4180 | 2571.802431 | -0.896759972 | 0.000000083 | PSP1 C-terminal conserved region, putative |
| Tb927.4.2310 | 6087.385189 | -0.896683584 | 3.52E-07 | asparaginyl-tRNA synthetase, putative |
| Tb927.10.14150 | 6099.71452 | -0.895496101 | 2.75E-05 | nuclear segregation protein, putative |
| Tb927.10.5840 | 10560.03362 | -0.894073627 | 3.40E-07 | translation elongation factor 1-beta, putative |
| Tb927.11.3250 | 15223.20427 | -0.893883352 | 1.09E-07 | dynein heavy chain, putative |
| Tb927.10.8900 | 714.8880949 | -0.890196988 | 5.19E-06 | choline/ethanolamine phosphotransferase (CEPT) |
| Tb927.11.15100 | 4917.393522 | -0.882412403 | 2.41E-08 | Tb5.20 |
| Tb927.11.2650 | 7151.344863 | -0.881838792 | 4.91E-07 | heat shock protein 84, putative |
| Tb927.2.4710 | 7427.054425 | -0.881574334 | 5.70E-10 | RNA-binding protein, putative |
| Tb927.8.730 | 5186.090853 | -0.876169021 | 3.09E-05 | nucleolar RNA-binding protein, putative |
| Tb927.7.1300 | 12530.48916 | -0.876076512 | 1.24E-06 | protein disulfide isomerase, putative |
| Tb927.10.11710 | 1088.015232 | -0.875296849 | 2.33E-05 | hypothetical protein |
| Tb927.9.2470 | 12702.74154 | -0.874079236 | 0.000000017 | nucleolar protein |
| Tb927.7.2650 | 12067.57219 | -0.873311284 | 7.95E-10 | hypothetical protein, conserved |
| Tb927.10.8190 | 6560.146812 | -0.872024334 | 0.000000009 | T-complex protein 1, theta subunit, putative |
| Tb927.11.18700 | 179.4465319 | -0.871901873 | 0.003342113 | hypothetical protein, conserved |
| Tb927.10.9570 | 5334.357615 | -0.865080828 | 3.98E-08 | paraflagellar rod component, putative |
| Tb927.11.5370 | 2173.403409 | -0.864688689 | 2.68E-07 | hypothetical protein, conserved |
| Tb927.11.6230 | 5593.38036 | -0.862654833 | 4.12E-07 | protein transport protein SEC61 subunit alpha, putative |
| Tb927.11.1430 | 5911.948489 | -0.86105537 | 3.27E-09 | Component of motile flagella 2 |
| Tb927.8.2460 | 2471.84321 | -0.860473491 | 3.56E-07 | hypothetical protein, conserved |
| Tb11.v5.0414 | 834.3044404 | -0.856703292 | 4.07E-06 | ATP-dependent DEAD/H RNA helicase, putative |

| | | | | |
|---|---|---|---|---|
| Tb927.11.13050 | 268.931369 | -0.856082864 | 0.00211502 | calmodulin |
| Tb927.6.2360 | 406.456844 | -0.85579691 | 0.00045849 | adenosine kinase, putative |
| Tb927.3.930 | 16862.19739 | -0.850230763 | 1.96E-07 | dynein heavy chain, putative |
| Tb927.3.1840 | 4820.130574 | -0.84864077 | 4.24E-08 | 3-oxo-5-alpha-steroid 4-dehydrogenase, putative |
| Tb927.9.11270 | 9864.619752 | -0.847988091 | 1.33E-06 | t- complex protein 1 (eta subunit), putative |
| Tb927.11.16360 | 676.0004779 | -0.847280011 | 0.002146365 | hypothetical protein, conserved |
| Tb927.9.6290 | 6348.935551 | -0.844375256 | 0.002886301 | arginine kinase |
| Tb11.v5.0171 | 353.275894 | -0.842532031 | 0.001252249 | 5'-3' exonuclease XRNA, putative |
| Tb927.4.5010 | 12101.22174 | -0.841583991 | 3.92E-06 | calreticulin, putative |
| Tb927.1.3830 | 12671.48689 | -0.839553499 | 0.000771345 | glucose-6-phosphate isomerase, glycosomal |
| Tb927.11.7510 | 12407.25617 | -0.839420266 | 0.000319258 | luminal binding protein 1 (BiP), putative |
| Tb927.11.11520 | 4748.720447 | -0.838783151 | 5.18E-08 | glycosomal membrane protein |
| Tb927.9.13785 | 1416.537587 | -0.837740553 | 0.000934346 | kinetoplastid membrane protein 11-5 |
| Tb927.7.2400 | 547.5891448 | -0.83758287 | 0.000213052 | tyrosyl-tRNA synthetase, putative |
| Tb11.v5.1046 | 12850.99314 | -0.832219859 | 3.47E-08 | elongation factor 1-alpha, putative |
| Tb927.11.1070 | 3738.042377 | -0.826335724 | 5.34E-08 | glycosomal transporter (GAT3), putative |
| Tb927.10.6400 | 5350.929418 | -0.824964243 | 0.00000004 | chaperonin HSP60, mitochondrial precursor |
| Tb927.8.2070 | 627.341065 | -0.821710164 | 6.16E-05 | Present in the outer mitochondrial membrane proteome 39-1 |
| Tb10.NT.148 | 604.3611845 | -0.820436552 | 7.79E-05 | Noncoding RNA, putative |
| Tb927.7.1110 | 3586.745938 | -0.817221501 | 0.001994057 | asparagine synthetase a, putative |
| Tb927.11.13020 | 2053.279768 | -0.817183338 | 0.000459508 | calmodulin |
| Tb927.11.7560 | 4709.471953 | -0.816560157 | 4.03E-07 | Component of motile flagella 15 |
| Tb927.6.3940 | 2649.190638 | -0.815478592 | 2.99E-05 | Autophagy-related protein 27, putative |
| Tb927.11.3030 | 2450.860262 | -0.814057215 | 0.000165459 | phosphoribosylpyrophosphate synthetase, putative |
| Tb927.10.8940 | 26390.36718 | -0.813726115 | 7.79E-08 | flagellum targeting protein kharon1, putative |
| Tb927.9.5860 | 2334.883567 | -0.807933895 | 4.99E-07 | tryparedoxin peroxidase |
| Tb927.7.4470 | 1716.835184 | -0.80587591 | 7.67E-07 | hypothetical protein, conserved |
| Tb927.11.5440 | 3348.834161 | -0.805166612 | 1.32E-05 | NADP-dependent malic enzyme, cytosolic |
| Tb927.11.10940 | 837.503866 | -0.80098237 | 0.000109131 | hypothetical protein, conserved |

| | | | | |
|---|---|---|---|---|
| Tb927.11.13190 | 1982.843303 | -0.800003482 | 6.53E-07 | elongation factor 1 gamma, putative |
| Tb927.7.710 | 5940.138912 | -0.799555675 | 0.000004652 | heat shock 70 kDa protein, putative |
| Tb927.8.4430 | 10160.82582 | -0.794747262 | 0.003300383 | uridine phosphorylase |
| Tb927.11.14880 | 7643.144736 | -0.794742456 | 2.41E-08 | Paraflagellar Rod Proteome Component 9, putative |
| Tb927.10.12330 | 778.5875352 | -0.791999756 | 0.000275023 | zinc finger protein family member, putative |
| Tb927.10.540 | 6619.630867 | -0.790642274 | 0.000000392 | ATP-dependent RNA helicase SUB2, putative |
| Tb927.10.14890 | 9861.283852 | -0.79047051 | 5.69E-07 | C-terminal motor kinesin, putative |
| Tb927.11.16630 | 1960.453787 | -0.790006023 | 0.000121889 | major facilitator superfamily, putative |
| Tb927.9.4210 | 1917.679487 | -0.789390712 | 6.39E-07 | fatty acyl CoA synthetase 3 |
| Tb927.5.2570 | 1625.232207 | -0.788241343 | 1.17E-05 | eukaryotic translation initiation factor 3 subunit b |
| Tb927.10.14500 | 2590.212382 | -0.787971867 | 2.54E-06 | hypothetical protein, conserved |
| Tb927.8.4330 | 2887.143111 | -0.787756132 | 0.00023079 | small GTP-binding protein Rab11 |
| Tb927.11.760 | 11200.07093 | -0.787226985 | 2.41E-06 | protein phosphatase 2C, putative |
| Tb927.11.13090 | 1144.602005 | -0.786986755 | 0.000074046 | elongation factor 1 gamma, putative |
| Tb927.9.4750 | 5200.99029 | -0.786901342 | 7.04E-07 | hypothetical protein |
| Tb927.11.17000 | 6182.683058 | -0.785467003 | 6.79E-06 | AIR9-like protein |
| Tb927.9.4200 | 3195.577923 | -0.783519636 | 5.21E-06 | fatty acyl CoA synthetase 2 |
| Tb927.8.1550 | 3728.80798 | -0.777944341 | 1.74E-06 | paraflagellar rod component, putative |
| Tb927.11.720 | 4891.01396 | -0.776803883 | 2.30E-07 | hypothetical protein, conserved |
| Tb927.2.6000 | 4022.680949 | -0.775445124 | 0.000636964 | glycosylphosphatidylinositol-specific phospholipase C |
| Tb927.5.1210 | 12493.64589 | -0.773157455 | 2.87E-08 | short-chain dehydrogenase, putative |
| Tb927.3.4100 | 5914.761264 | -0.772636032 | 0.000136155 | Pyruvate transporter, putative |
| Tb927.10.6510 | 922.3024971 | -0.771940143 | 3.67E-05 | chaperonin HSP60, mitochondrial precursor |
| Tb927.5.3220 | 3546.687722 | -0.770505076 | 1.61E-06 | signal peptidase type I, putative |
| Tb927.11.1990 | 2742.415818 | -0.768961753 | 0.000330096 | arginyl-tRNA synthetase, putative |
| Tb927.5.3800 | 3417.696804 | -0.768940679 | 0.000491452 | glutamine hydrolysing (not ammonia-dependent) carbomoyl phosphate synthase, putative |
| Tb927.1.4490 | 1567.233241 | -0.765882679 | 0.000101802 | acetyltransferase, putative |
| Tb927.11.6870 | 1772.347747 | -0.764561238 | 1.49E-05 | 14-3-3 protein 2 |

| | | | | |
|---|---|---|---|---|
| Tb11.v5.0671 | 506.4260477 | -0.758123731 | 0.00130034 | protein kinase A catalytic subunit, putative |
| Tb927.6.4000 | 5281.098596 | -0.757764492 | 0.000235423 | small glutamine-rich tetratricopeptide repeat protein, putative |
| Tb927.11.1980 | 4575.549232 | -0.754010598 | 3.77E-06 | zinc finger protein family member, putative |
| Tb927.10.15320 | 923.5530731 | -0.752367069 | 3.26E-05 | hypothetical protein, conserved |
| Tb927.7.470 | 1766.883651 | -0.752351823 | 2.46E-05 | Enriched in surface-labeled proteome protein 14 |
| Tb927.10.12890 | 5349.309811 | -0.751588497 | 3.15E-06 | bifunctional aminoacyl-tRNA synthetase, putative |
| Tb927.10.7680 | 3475.570192 | -0.750609595 | 7.27E-06 | GTPase activating protein, putative |
| Tb927.10.16120 | 5495.020397 | -0.750471427 | 7.86E-05 | inosine-5'-monophosphate dehydrogenase |
| Tb927.2.4210 | 4461.099174 | -0.750081302 | 5.79E-07 | Phosphoenolpyruvate carboxykinase [ATP], glycosomal |
| Tb927.11.14620 | 5145.975341 | -0.749904086 | 0.000403727 | expression site-associated gene 2 (ESAG2) protein, putative |
| Tb927.10.2530 | 1413.598342 | -0.749738348 | 0.000141402 | adenylate kinase, putative |
| Tb927.9.10770 | 31861.98876 | -0.749223513 | 9.85E-07 | polyadenylate-binding protein 2 |
| Tb927.11.15420 | 467.9747692 | -0.745810374 | 0.000543131 | eukaryotic translation initiation factor 3 subunit k |
| Tb927.11.6860 | 1769.523263 | -0.745537237 | 5.67E-05 | hypothetical protein |
| Tb927.4.4040 | 3706.486064 | -0.744831211 | 6.61E-06 | hypothetical protein, conserved |
| Tb927.8.6660 | 7424.204041 | -0.744315234 | 1.76E-07 | paraflagellar rod component, putative |
| Tb927.9.11100 | 1127.979006 | -0.744206736 | 6.12E-05 | cAMP-dependent protein kinase catalytic subunit 1 |
| Tb927.5.4500 | 2065.183116 | -0.742638816 | 3.77E-05 | ras-like small GTPase, putative |
| Tb927.7.3550 | 20003.02941 | -0.741243994 | 8.26E-08 | cytoskeleton associated protein, putative |
| Tb927.9.5770 | 3539.85029 | -0.739708 | 1.77E-06 | tryparedoxin peroxidase |
| Tb927.5.4420 | 3378.378891 | -0.736336992 | 2.89E-06 | nucleolar RNA helicase II, putative |
| Tb927.11.2300 | 1780.441926 | -0.735959989 | 0.002121054 | eukaryotic peptide chain release factor subunit 1, putative |
| Tb927.9.9940 | 2761.148498 | -0.735704383 | 7.68E-06 | PACRGB |
| Tb927.7.3760 | 3468.107835 | -0.734279909 | 1.22E-05 | phosphatidylserine synthase, putative |
| Tb927.10.7930 | 8928.061917 | -0.730427015 | 2.31E-07 | 2,3-bisphosphoglycerate-independent phosphoglycerate mutase |
| Tb927.10.1040 | 2404.534284 | -0.729253654 | 0.00003754 | serine peptidase, Clan SC, Family S10 |
| Tb927.6.4500 | 9729.203503 | -0.726585563 | 3.22E-07 | conserved repeat domain containing protein, putative |
| Tb927.10.200 | 5182.214995 | -0.725986093 | 3.65E-06 | vacuolar ATP synthase, putative |
| Tb927.11.3730 | 3080.33511 | -0.720466791 | 0.000415832 | leucyl-tRNA synthetase, putative |

| | | | | |
|---|---|---|---|---|
| Tb927.10.14790 | 7442.066215 | -0.717009904 | 0.002779163 | metallo-peptidase, Clan MG, Family M24 |
| Tb927.8.650 | 6087.590423 | -0.714684521 | 9.63E-06 | cation-transporting ATPase, putative |
| Tb927.10.13780 | 3641.950217 | -0.712674232 | 0.000145753 | glycogen synthase kinase 3 |
| Tb927.11.14250 | 6874.962822 | -0.712519947 | 2.70E-06 | T-complex protein 1, epsilon subunit, putative |
| Tb927.11.11010 | 11163.12045 | -0.710564573 | 1.64E-05 | hypothetical protein, conserved |
| Tb927.11.1110 | 5934.934901 | -0.708940305 | 1.33E-05 | calpain, putative |
| Tb927.3.5370 | 4151.160645 | -0.708936 | 1.83E-05 | hypothetical protein, conserved |
| Tb927.9.5320 | 4095.503061 | -0.700844443 | 0.001206321 | nucleolar RNA binding protein, putative |
| Tb927.10.13630 | 9715.081017 | -0.700162457 | 4.38E-05 | glucosidase, putative |
| Tb927.10.14700 | 4976.171304 | -0.699286654 | 1.28E-05 | hypothetical protein, conserved |
| Tb927.9.12650 | 13135.2986 | -0.698934509 | 4.17E-06 | ubiquitin-activating enzyme E1, putative |
| Tb927.6.5070 | 7672.412136 | -0.698487454 | 0.000041344 | hypothetical protein, conserved |
| Tb927.11.8970 | 6892.274619 | -0.697200975 | 0.000520844 | ribose 5-phosphate isomerase, putative |
| Tb927.4.2740 | 4085.120321 | -0.696981065 | 0.001236883 | p25-alpha, putative |
| Tb927.11.14950 | 7460.211993 | -0.696766526 | 2.38E-06 | zinc finger protein 2 |
| Tb927.9.14130 | 1740.285574 | -0.696748382 | 0.003413531 | hypothetical protein, conserved |
| Tb927.10.8830 | 3531.744794 | -0.695837941 | 0.000117181 | Flagellum attachment zone protein 5 |
| Tb927.8.6110 | 7393.848792 | -0.695392274 | 0.000879283 | 3-hydroxy-3-methylglutaryl-CoA synthase, putative |
| Tb927.8.760 | 2813.80376 | -0.693792773 | 0.001836736 | nucleolar RNA-binding protein |
| Tb927.10.4760 | 2748.89117 | -0.69211295 | 0.000008787 | Integral membrane protein DUF106, putative |
| Tb927.9.9290 | 5920.605776 | -0.691124039 | 0.001878827 | polyadenylate-binding protein 1 |
| Tb927.10.5810 | 12275.754 | -0.690543672 | 0.000230966 | The ARF-like 2 binding protein BART, putative |
| Tb927.10.9430 | 3235.589583 | -0.690296172 | 0.000533671 | phosphoribosylpyrophosphate synthetase, putative |
| Tb927.11.2610 | 23809.25442 | -0.690260343 | 2.68E-06 | hypothetical protein, conserved |
| Tb927.7.1130 | 1852.951366 | -0.689516116 | 0.000110072 | trypanothione/tryparedoxin dependent peroxidase 2 |
| Tb927.10.13290 | 2427.547374 | -0.688813782 | 0.001158452 | ethanolamine phosphotransferase (EPT) |
| Tb927.7.190 | 4011.789559 | -0.68705547 | 0.001124911 | thimet oligopeptidase, putative |
| Tb927.9.4190 | 3263.844066 | -0.686995059 | 0.000978855 | fatty acyl CoA syntetase 1 |
| Tb927.5.1360 | 3268.335541 | -0.683130341 | 3.14E-05 | nucleoside 2-deoxyribosyltransferase |

| | | | | |
|---|---|---|---|---|
| Tb927.11.15200 | 667.4741549 | -0.67977659 | 0.001815135 | hypothetical protein, conserved |
| Tb927.9.9550 | 5341.743968 | -0.677202996 | 2.56E-05 | conserved protein, unknown function |
| Tb927.7.3330 | 2354.785873 | -0.676531471 | 1.99E-05 | Flagellar attachment zone protein 10 |
| Tb927.11.5820 | 2453.675234 | -0.676079378 | 4.12E-05 | Inhibitor of apoptosis-promoting Bax1, putative |
| Tb927.9.9810 | 3777.762197 | -0.675329414 | 0.000106974 | hypothetical protein, conserved |
| Tb927.10.7090 | 4987.959551 | -0.672022429 | 1.82E-05 | Alternative oxidase, mitochondrial |
| Tb927.7.1470 | 2406.849151 | -0.671712456 | 0.001037493 | Mitochondrial ATP synthase subunit c-3 |
| Tb927.8.3150 | 6052.221477 | -0.670613071 | 0.000265362 | T-complex protein 1, gamma subunit, putative |
| Tb927.10.14780 | 4022.420363 | -0.670016153 | 0.000505851 | Canertinib-binding protein kinase 1 |
| Tb927.9.8880 | 2873.066485 | -0.66954361 | 3.44E-05 | actin B |
| Tb927.11.11250 | 4937.56218 | -0.664988169 | 2.03E-05 | cytosolic malate dehydrogenase |
| Tb927.10.12010 | 1737.166152 | -0.664814339 | 0.000961398 | hypothetical protein, conserved |
| Tb927.8.1990 | 7222.198204 | -0.663549298 | 0.000544874 | peroxidoxin |
| Tb927.10.2370 | 2370.563624 | -0.662950372 | 0.002432723 | Lupus La protein homolog, putative |
| Tb927.10.1630 | 1746.341423 | -0.66189842 | 0.000371762 | atp-binding cassette sub-family e member 1 |
| Tb927.8.6240 | 4042.817898 | -0.660343511 | 5.19E-05 | STOP axonemal protein |
| Tb927.6.4480 | 4737.533062 | -0.659113818 | 1.33E-05 | valyl-tRNA synthetase, putative |
| Tb927.7.4260 | 3690.014914 | -0.656691169 | 2.98E-05 | Enriched in surface-labeled proteome protein 13 |
| Tb927.9.6250 | 655.4390898 | -0.656015448 | 0.000976164 | arginine kinase |
| Tb927.11.6440 | 8469.652534 | -0.655100438 | 2.28E-05 | hypothetical protein, conserved |
| Tb927.3.5340 | 4894.881557 | -0.652816307 | 2.88E-05 | Hsc70-interacting protein (Hip), putative |
| Tb927.11.7170 | 3577.21945 | -0.64962454 | 7.29E-05 | seryl-tRNA synthetase |
| Tb927.10.2240 | 5095.367187 | -0.646103631 | 1.82E-05 | RNA binding protein |
| Tb927.11.14370 | 2260.413925 | -0.645466282 | 4.19E-05 | Microtubule-associated protein 103???kDa |
| Tb927.11.9530 | 16345.36347 | -0.645348987 | 6.57E-06 | 14-3-3 protein 1 |
| Tb927.7.940 | 3883.875344 | -0.643152046 | 0.000193892 | protein kinase C substrate protein, heavy chain, putative |
| Tb927.10.5770 | 1912.828205 | -0.643127017 | 7.96E-05 | Valosin-containing protein |
| Tb927.6.4320 | 3427.036138 | -0.641993951 | 0.000181458 | Protein of unknown function (DUF2817), putative |
| Tb927.4.2040 | 9879.95908 | -0.641990327 | 0.000908813 | DNA/RNA-binding protein Alba 3 |

| | | | | |
|---|---|---|---|---|
| Tb927.11.1150 | 1777.924837 | -0.641603721 | 0.00015664 | radial spoke protein 3 |
| Tb927.10.12670 | 2035.93561 | -0.641346983 | 4.87E-05 | hypothetical protein, conserved |
| Tb927.9.8990 | 1490.821453 | -0.639868972 | 0.000101932 | Ankyrin repeats (3 copies), putative |
| Tb927.7.4180 | 6038.614291 | -0.639052837 | 0.000046176 | Fatty acid elongase |
| Tb927.8.6760 | 6643.116096 | -0.637615717 | 0.000317122 | translationally-controlled tumor protein homolog, putative |
| Tb927.5.4390 | 1602.879124 | -0.6374048 | 0.000190575 | hypothetical protein, conserved |
| Tb927.7.4290 | 1954.585128 | -0.637230915 | 8.99E-05 | Nuclear distribution protein C homolog |
| Tb927.5.3830 | 2296.119081 | -0.636001247 | 0.00022806 | dihydroorotate dehydrogenase (fumarate) |
| Tb927.11.740 | 9080.836753 | -0.635457293 | 3.04E-05 | eukaryotic translation initiation factor 5A |
| Tb927.9.10560 | 1377.697868 | -0.631312943 | 0.000735028 | Archaic Translocase of outer membrane 19 kDa subunit |
| Tb927.10.13580 | 1183.578725 | -0.629968851 | 0.000936651 | major facilitator superfamily protein (MFS), putative |
| Tb927.2.4440 | 751.0335192 | -0.629707486 | 0.001010453 | hypothetical protein |
| Tb927.10.15180 | 6765.25061 | -0.627773169 | 9.47E-05 | nucleosome assembly protein, putative |
| Tb927.9.12510 | 2767.022841 | -0.62728473 | 8.21E-05 | ATP-dependent DEAD/H RNA helicase, putative |
| Tb927.11.4910 | 2202.702 | -0.626253737 | 0.0002334 | predicted ankyrin repeat family protein |
| Tb927.10.13670 | 2017.058615 | -0.625540266 | 0.003096723 | serine/threonine protein phosphatase 5 |
| Tb927.11.2700 | 1344.203977 | -0.622445486 | 0.001997121 | Basal body protein |
| Tb927.5.1260 | 1167.984231 | -0.622075499 | 0.000419251 | Sulfate transporter N-terminal domain with GLY motif/Sulfate transporter family, putative |
| Tb927.9.8000 | 6479.502252 | -0.6220709 | 0.000329646 | ubiquitin-protein ligase, putative |
| Tb927.10.3990 | 7296.227197 | -0.618963922 | 6.53E-05 | DHH1 |
| Tb927.11.1330 | 969.1144962 | -0.614117928 | 0.000517378 | Stress responsive A/B Barrel Domain, putative |
| Tb927.2.5060 | 3695.528124 | -0.613984687 | 0.000676444 | GTP binding protein, putative |
| Tb927.10.2900 | 4004.380459 | -0.613254263 | 3.37E-05 | importin beta-1 subunit, putative |
| Tb927.10.12710 | 15908.62432 | -0.613113248 | 0.002849952 | heat shock protein 110, putative |
| Tb927.11.550 | 5905.182303 | -0.612017602 | 0.001207535 | hypothetical protein SCD6.10 |
| Tb927.11.13230 | 2865.760491 | -0.610763151 | 0.000658763 | VAMP-associated protein, putative |
| Tb927.9.1750 | 5274.501925 | -0.610325976 | 4.60E-05 | Fibronectin type III domain containing protein, putative |
| Tb927.6.4370 | 3196.606925 | -0.607981964 | 0.00280883 | eukaryotic translation initiation factor 3 subunit d |

| | | | | |
|---|---|---|---|---|
| Tb927.11.800 | 3579.648431 | -0.607439367 | 0.000336474 | prefoldin subunit, putative |
| Tb927.6.4440 | 8550.830584 | -0.605392289 | 0.001154272 | RNA-binding protein 42 (RNA-binding motif protein 42) |
| Tb927.11.950 | 3559.365348 | -0.603720977 | 5.06E-05 | RNA-binding protein, putative |
| Tb927.3.1900 | 893.4207317 | -0.603107239 | 0.001634962 | conserved protein, unknown function |
| Tb927.5.4180 | 7069.544345 | -0.602189874 | 0.002746364 | histone H4, putative |
| Tb927.11.9640 | 4290.050426 | -0.599138516 | 0.000304971 | glycyl-tRNA synthetase, putative |
| Tb927.9.4420 | 2970.672373 | -0.598329 | 0.000129565 | hypothetical protein, conserved |
| Tb927.11.3980 | 2337.837761 | -0.598301398 | 0.001589118 | metallo-peptidase, Clan ME, Family M16 |
| Tb927.11.7350 | 3516.663712 | -0.598121171 | 0.000475914 | Histone H2B variant V |
| Tb927.11.1420 | 830.9338361 | -0.597607896 | 0.002124447 | hypothetical protein, conserved |
| Tb927.3.960 | 1380.021671 | -0.594844201 | 0.000507813 | protein transport protein Sec61 gamma subunit, putative |
| Tb927.8.7970 | 2458.204136 | -0.591767242 | 0.000415885 | hypothetical protein |
| Tb927.10.10280 | 1293.330848 | -0.591112887 | 0.00063382 | microtubule-associated protein, putative |
| Tb927.10.8980 | 1252.450132 | -0.590321516 | 0.001139261 | hypothetical protein, conserved |
| Tb927.7.2510 | 3001.979555 | -0.590098038 | 0.001808278 | endoplasmatic reticulum retrieval protein, putative |
| Tb927.3.2050 | 2567.54636 | -0.589798365 | 0.000373327 | Minicircle replication factor 172 |
| Tb927.8.1460 | 1189.852446 | -0.58592677 | 0.000909099 | EamA-like transporter family, putative |
| Tb927.11.7780 | 1878.364315 | -0.584736068 | 0.002772 | Mitochondrial import receptor subunit ATOM46 |
| Tb11.v5.0463 | 820.4250395 | -0.583900872 | 0.002453244 | expression site-associated gene (ESAG) protein, putative |
| Tb927.7.5840 | 1812.166084 | -0.582455744 | 0.001880707 | hypothetical protein, conserved |
| Tb927.6.4140 | 3482.508744 | -0.579784064 | 0.00075765 | paraflagellar rod component, putative |
| Tb927.11.10800 | 1631.791288 | -0.578633049 | 0.00076608 | Domain of unknown function (DUF4496), putative |
| Tb927.6.4180 | 1553.860847 | -0.577214252 | 0.000426252 | FUN14 family, putative |
| Tb927.9.9100 | 2451.654976 | -0.575407644 | 0.000449894 | hypothetical protein, conserved |
| Tb927.5.3170 | 2033.221852 | -0.573464529 | 0.001086622 | ribose-phosphate pyrophosphokinase, putative |
| Tb927.10.14980 | 4095.441603 | -0.573009537 | 0.000564264 | intraflagellar transport protein 52 |
| Tb927.4.590 | 5603.370138 | -0.572944693 | 8.14E-05 | PQQ-like domain/Protein of unknown function (DUF1620), putative |
| Tb927.8.2320 | 1236.732262 | -0.572292605 | 0.000792531 | hypothetical protein |
| Tb927.10.170 | 3634.440867 | -0.571463555 | 0.001318979 | pseudouridine synthase, Cbf5p |

| | | | | |
|---|---|---|---|---|
| Tb927.11.11590 | 5038.184144 | -0.570237236 | 0.003207507 | Eukaryotic translation initiation factor 3 subunit E |
| Tb927.11.1020 | 5539.095699 | -0.570054081 | 0.001990888 | ribokinase, putative |
| Tb927.11.260 | 3843.773896 | -0.569847538 | 0.00018555 | Putative intraflagellar transport protein F6 |
| Tb927.10.13280 | 4122.133928 | -0.5684295 | 0.000116312 | hypothetical protein, conserved |
| Tb927.5.1160 | 5800.990307 | -0.566973939 | 0.001815165 | Degradation arginine-rich protein for mis-folding, putative |
| Tb927.3.1590 | 1419.827744 | -0.566632459 | 0.001054365 | mitochondrial RNA binding complex 1 subunit |
| Tb927.11.15370 | 5772.559169 | -0.566547545 | 0.000145872 | Importin subunit beta-4, putative |
| Tb927.8.2540 | 7231.409674 | -0.566478033 | 0.000345741 | 3-ketoacyl-CoA thiolase, putative |
| Tb927.9.5190 | 1906.749257 | -0.566256141 | 0.000574193 | proliferative cell nuclear antigen (PCNA), putative |
| Tb927.9.1440 | 6405.778807 | -0.565014492 | 0.000755334 | Paraquat-inducible protein A, putative |
| Tb927.11.14300 | 5378.023054 | -0.56426515 | 0.000105345 | Dynein intermediate chain, axonemal |
| Tb927.6.4630 | 3304.136116 | -0.563429267 | 0.000269605 | kinetoplastid-specific phospho-protein phosphatase, putative |
| Tb927.6.3650 | 4833.943644 | -0.5630773 | 0.00011479 | ADP-ribosylation factor, putative |
| Tb927.10.10050 | 3593.081501 | -0.560788343 | 0.00072761 | hypothetical protein, conserved |
| Tb927.11.15130 | 1270.805999 | -0.56037142 | 0.001008246 | DnaJ homolog, putative |
| Tb927.10.13720 | 5550.334358 | -0.55866066 | 0.000169486 | RNA-binding protein 29, putative |
| Tb927.11.2430 | 8958.134103 | -0.558449955 | 0.00038053 | Cytoplasmic dynein 2 heavy chain (DYNC2H2), putative |
| Tb927.5.4480 | 7822.444037 | -0.557114353 | 0.000220809 | paraflagellar rod component par4, putative |
| Tb927.10.14760 | 8948.870998 | -0.557091313 | 0.001159511 | clathrin light chain, putative |
| Tb927.10.7140 | 5414.22425 | -0.556258979 | 0.00014021 | membrane-bound acid phosphatase 2 |
| Tb927.11.2540 | 2596.761637 | -0.556083656 | 0.000401816 | hypothetical protein, conserved |
| Tb927.6.5050 | 1861.418321 | -0.555068951 | 0.000787403 | V-type ATPase, C subunit, putative |
| Tb927.9.11480 | 5106.744673 | -0.553691346 | 0.000249083 | Enriched in surface-labeled proteome protein 9 |
| Tb927.7.2190 | 4980.92797 | -0.553249191 | 0.002273385 | translocon-associated protein (TRAP), alpha subunit, putative |
| Tb927.8.1870 | 15649.21438 | -0.551905249 | 0.001934585 | Golgi/lysosome glycoprotein 1 |
| Tb927.9.6560 | 2980.467002 | -0.549792153 | 0.002486407 | NAK family pseudokinase, putative |
| Tb11.v5.0863 | 1335.720298 | -0.549343641 | 0.001238422 | hypothetical protein, conserved |
| Tb927.11.5840 | 2278.069619 | -0.549228657 | 0.001777706 | Protein translation factor SUI1 homolog, putative |
| Tb927.9.2590 | 2621.252802 | -0.549051287 | 0.000529711 | hypothetical protein, conserved |

| | | | | |
|---|---|---|---|---|
| Tb927.7.1120 | 1486.459929 | -0.547876549 | 0.00339555 | trypanothione/tryparedoxin dependent peroxidase 1, cytosolic |
| Tb927.7.2170 | 3891.991111 | -0.547816876 | 0.001601383 | hypothetical protein, conserved |
| Tb927.10.1230 | 2910.09764 | -0.547140356 | 0.0003076 | Enriched in surface-labeled proteome protein 23 |
| Tb927.8.3790 | 1416.484585 | -0.542014729 | 0.001040868 | paraflagellar rod component, putative |
| Tb927.7.1330 | 866.57155 | -0.540141461 | 0.00330241 | hypothetical protein, conserved |
| Tb927.8.3250 | 7907.303357 | -0.539746588 | 0.000294192 | dynein heavy chain, putative |
| Tb11.v5.0364 | 1255.00935 | -0.539434613 | 0.001543269 | hypothetical protein, conserved |
| Tb927.10.840 | 2923.919308 | -0.538946865 | 0.000733241 | Flagellum attachment zone protein 6 |
| Tb927.11.4450 | 3045.17466 | -0.535162938 | 0.002738864 | ALBA-Domain Protein |
| Tb927.10.290 | 2586.452852 | -0.53473888 | 0.0006195 | proteasome alpha 2 subunit, putative |
| Tb927.11.15140 | 3641.055743 | -0.534313258 | 0.000577158 | bilobe region protein, putative |
| Tb927.6.4460 | 2977.267203 | -0.53098429 | 0.001877408 | S-adenosylmethionine decarboxylase |
| Tb927.9.13070 | 15923.28306 | -0.530061325 | 0.003245095 | Heat shock factor binding 1 domain-containing protein |
| Tb927.11.9860 | 10628.91411 | -0.530025708 | 0.001517248 | EF-hand domain pair, putative |
| Tb927.11.8140 | 1411.373758 | -0.529873645 | 0.002781086 | ER lumen retaining receptor protein, putative |
| Tb927.2.3780 | 3656.171074 | -0.529483275 | 0.001852787 | translation initiation factor IF-2, putative |
| Tb927.11.4290 | 5578.025859 | -0.528866155 | 0.000363895 | 40S ribosomal protein S12 |
| Tb927.11.10990 | 1469.961875 | -0.528708449 | 0.001951443 | methylthioribose-1-phosphate isomerase, putative |
| Tb927.10.3970 | 4296.657657 | -0.527461633 | 0.0006745 | hypothetical protein, conserved |
| Tb927.10.14030 | 6637.264614 | -0.526798765 | 0.002378443 | hypothetical protein, conserved |
| Tb927.9.15890 | 1906.336624 | -0.526279569 | 0.000990718 | expression site-associated gene 5 (ESAG5) protein, putative |
| Tb927.3.5060 | 3183.584813 | -0.52550554 | 0.001403243 | Ankyrin repeats (many copies)/Ankyrin repeats (3 copies), putative |
| Tb927.9.9060 | 3023.240129 | -0.525423942 | 0.000878961 | Lsm12 protein, putative |
| Tb927.11.6460 | 3177.628858 | -0.524873393 | 0.000713521 | hypothetical protein, conserved |
| Tb927.10.3790 | 3547.985992 | -0.52475147 | 0.001018297 | hypothetical protein, conserved |
| Tb927.3.1680 | 3226.641322 | -0.523335324 | 0.002378908 | Eukaryotic translation initiation factor 3 subunit F |
| Tb927.10.8930 | 3318.635644 | -0.521943736 | 0.001400273 | paraflagellar rod component, putative |
| Tb927.11.500 | 13947.00393 | -0.52179048 | 0.000614144 | RNA-binding protein, UBP1 |
| Tb927.4.1270 | 3224.436274 | -0.519573364 | 0.00247714 | ruvB-like DNA helicase, putative |

| | | | | |
|---|---|---|---|---|
| Tb927.11.5850 | 6317.992429 | -0.516168228 | 0.000632155 | RNA-binding protein 38, putative |
| Tb927.11.3380 | 2713.066194 | -0.515285579 | 0.001789019 | Ran-binding protein 1, putative |
| Tb927.11.1080 | 2280.462007 | -0.514951526 | 0.00097302 | hypothetical protein, conserved |
| Tb927.11.3490 | 4638.232328 | -0.514930099 | 0.000456469 | hypothetical protein, conserved |
| Tb927.7.1920 | 1895.539547 | -0.514747294 | 0.001304447 | paraflagellar rod component, putative |
| Tb927.10.14330 | 2761.213642 | -0.513861006 | 0.00317349 | Utp14 protein, putative |
| Tb927.6.4770 | 1098.783899 | -0.511272904 | 0.003283857 | protein mkt1, putative |
| Tb927.11.16340 | 3076.40169 | -0.503942526 | 0.00165277 | Importin-beta N-terminal domain/HEAT-like repeat, putative |
| Tb927.9.12980 | 3329.47373 | -0.499534534 | 0.002515118 | Alpha/beta hydrolase family, putative |
| Tb927.9.7770 | 2751.168475 | -0.498754399 | 0.001252842 | spermidine synthase |
| Tb927.2.4410 | 3204.994786 | -0.496531809 | 0.001753379 | Eukaryotic protein of unknown function (DUF846), putative |
| Tb927.6.4970 | 3645.348818 | -0.494963175 | 0.003170023 | serine/arginine-rich protein specific kinase SRPK, putative |
| Tb927.11.6790 | 1962.82959 | -0.491821989 | 0.002779326 | predicted WD40 repeat protein |
| Tb927.11.14840 | 5410.191681 | -0.490727637 | 0.002014638 | chromosomal passenger complex 2 |
| Tb927.10.13710 | 3789.744984 | -0.488486945 | 0.001699873 | Activator of Hsp90 ATPase, N-terminal/Activator of Hsp90 ATPase homolog 1-like protein, putative |
| Tb927.10.6800 | 2525.334565 | -0.483097639 | 0.001771955 | developmentally regulated GTP-binding protein, putative |
| Tb927.3.720 | 4988.861346 | -0.47973977 | 0.001863358 | zinc finger protein 3 |
| Tb927.11.630 | 3244.023653 | -0.474902713 | 0.003032184 | RNA polymerase I second largest subunit |
| Tb927.11.2360 | 4067.066921 | -0.461982987 | 0.001887451 | phenylalanyl-tRNA synthetase (beta subunit), putative |
| Tb927.8.5450 | 4885.204973 | -0.459366078 | 0.001701591 | amino acid permease 24 |
| Tb927.8.1940 | 2869.968511 | -0.453133135 | 0.003310826 | endosomal integral membrane protein, putative |
| Tb927.5.1000 | 18735.89493 | -0.45016784 | 0.001429049 | ubiquitin-conjugating enzyme E2, putative |
| Tb927.10.14470 | 5196.753709 | -0.445541559 | 0.003074318 | intraflagellar transport protein 140 |

**Table 8.3 List of downregulated genes after 36hr of TbRH2A RNAi**

List of genes found to be statistically down regulated after 36 hr TbRH2A knockdown with RNAi. The log2 fold-change (FC) of normalisaed read counts in tet-induced samples relative to un-induced is shown for each gene, as well as p value of significance and gene description

| Gene ID | Base Mean | Log2 FC | p value | Gene description |
|---|---|---|---|---|
| Tb427VSG-621 | 147.73983 | 4.103694 | 0.0002079 | array |
| Tb427VSG-15 | 1235.56385 | 3.892097 | 0 | BES |
| Tb427VSG-2096 | 11.76122 | 3.382387 | 0.0024056 | Pseudogene |
| Tb427VSG-1299 | 15.86982 | 3.290564 | 0.0007291 | Pseudogene |
| Tb427VSG-631 | 326.97103 | 2.867043 | 0 | array |
| Tb427VSG-775 | 235.98108 | 2.796116 | 0 | array |
| Tb427VSG-582 | 93.64212 | 2.584657 | 0.0000016 | MES |
| Tb427VSG-3591 | 140.76899 | 2.328949 | 0.0000006 | Pseudogene |
| Tb427VSG-5 | 103.85178 | 2.246091 | 0.0000009 | BES |
| Tb427VSG-514 | 53.30968 | 2.083303 | 0.0017956 | MC |
| Tb427VSG-322 | 1720.99756 | 1.716163 | 0.0000028 | array |
| Tb427VSG-636 | 114.55927 | 1.686899 | 0.0004664 | array |
| Tb427VSG-3709 | 192.04112 | 1.682074 | 0.0000412 | Pseudogene |
| Tb427VSG-6 | 546.80796 | 1.630817 | 0.0000417 | BES |
| Tb427VSG-3064 | 448.01582 | 1.555355 | 0.0000221 | Pseudogene |
| Tb427VSG-1954 | 539.42207 | 1.534786 | 0.000236 | MES |
| Tb427VSG-397 | 408.14157 | 1.521798 | 0.0004783 | Pseudogene |
| Tb427VSG-8 | 5879.58614 | 1.412936 | 0.0001566 | BES |
| Tb427VSG-1316 | 2272.83319 | 1.150647 | 0.0012714 | Pseudogene |
| Tb427VSG-1559 | 1098.06947 | 1.093385 | 0.0031419 | Pseudogene |

**Table 8.4 List of expressed VSGs after 24 hr of TbRH2A RNAi**
VSG genes found to be significantly ($p < 0.005$) upregulated after 24 hr of TbRH2A RNAi in induced samples relative to un-induced, expressed as Log2 fold-change (FC) in read counts. Whether each VSG is predicted to be found intact within silent arrays (array), a BES or an mVSG ES, or be pseudogenic is also shown.

| Gene ID | Base Mean | Log2 FC | p value | Gene description |
|---|---|---|---|---|
| Tb427VSG-1206 | 1.63E+01 | 4.548474 | 0.0005312 | pseudogene |
| Tb427VSG-423 | 3.77E+01 | 4.483429 | 0.0000051 | array |
| Tb427VSG-3238 | 9.43E+00 | 3.752305 | 0.009258 | pseudogene |
| Tb427VSG-3623 | 9.27E+00 | 3.716557 | 0.0080107 | pseudogene |
| Tb427VSG-3768 | 2.95E+01 | 3.681719 | 0.0000915 | pseudogene |
| Tb427VSG-596 | 1.01E+01 | 3.654775 | 0.0071546 | array |
| Tb427VSG-1614 | 6.35E+01 | 3.635489 | 0.0000008 | pseudogene |
| Tb427VSG-1641 | 1.08E+01 | 3.577864 | 0.0067922 | pseudogene |
| Tb427VSG-3709 | 3.75E+02 | 3.438951 | 0 | pseudogene |
| Tb427VSG-3064 | 8.58E+02 | 3.258154 | 0 | pseudogene |
| Tb427VSG-5 | 1.72E+02 | 3.144822 | 0.0000001 | BES |
| Tb427VSG-3037 | 1.20E+02 | 3.124865 | 0.0000037 | pseudogene |
| Tb427VSG-15 | 7.42E+02 | 3.088941 | 0.0000022 | BES |
| Tb427VSG-2096 | 2.04E+01 | 2.968342 | 0.0013374 | pseudogene |
| Tb427VSG-1387 | 1.58E+01 | 2.930711 | 0.0081921 | pseudogene |
| Tb427VSG-312 | 5.34E+01 | 2.846268 | 0.000176 | array |
| Tb427VSG-570 | 3.44E+01 | 2.804362 | 0.0013071 | array |
| Tb427VSG-1579 | 2.76E+01 | 2.63234 | 0.0013355 | pseudogene |
| Tb427VSG-1022 | 2.34E+01 | 2.575163 | 0.0025979 | pseudogene |
| Tb427VSG-1157 | 1.45E+02 | 2.569905 | 0.0000327 | pseudogene |
| Tb427VSG-1316 | 2.27E+03 | 2.561984 | 0.0000298 | pseudogene |
| Tb427VSG-3139 | 9.45E+01 | 2.560946 | 0.0001056 | pseudogene |
| Tb427VSG-2272 | 3.90E+01 | 2.501443 | 0.0009362 | array |
| Tb427VSG-3386 | 2.73E+01 | 2.427168 | 0.002898 | pseudogene |
| Tb427VSG-322 | 1.73E+03 | 2.411786 | 0.0000117 | array |
| Tb427VSG-621 | 6.59E+01 | 2.328027 | 0.0035334 | array |
| Tb427VSG-1559 | 1.00E+03 | 2.301547 | 0.0000547 | pseudogene |
| Tb427VSG-3449 | 3.71E+02 | 2.178837 | 0.0002387 | pseudogene |
| Tb427VSG-13 | 1.90E+02 | 2.159051 | 0.0003525 | BES |
| Tb427VSG-14 | 1.07E+03 | 2.147594 | 0.0000755 | BES |
| Tb427VSG-636 | 1.62E+02 | 2.137291 | 0.0008439 | array |
| Tb427VSG-775 | 1.42E+02 | 2.121092 | 0.0015607 | array |
| Tb427VSG-11 | 8.33E+02 | 1.913511 | 0.0018668 | BES |
| Tb427VSG-3577 | 6.52E+01 | 1.899652 | 0.0039648 | pseudogene |
| Tb427VSG-3719 | 2.73E+02 | 1.896749 | 0.0035249 | pseudogene |
| Tb427VSG-1394 | 1.89E+03 | 1.860581 | 0.0012079 | pseudogene |
| Tb427VSG-3221 | 3.63E+02 | 1.844018 | 0.0016085 | pseudogene |
| Tb427VSG-1042 | 9.57E+01 | 1.816259 | 0.0022194 | pseudogene |
| Tb427VSG-1241 | 8.87E+01 | 1.752897 | 0.005665 | pseudogene |
| Tb427VSG-761 | 1.14E+02 | 1.734486 | 0.0037029 | array |
| Tb427VSG-502 | 4.97E+01 | 1.733352 | 0.0105722 | array |
| Tb427VSG-676 | 1.89E+02 | 1.726471 | 0.0079513 | array |
| Tb427VSG-1073 | 6.03E+01 | 1.721688 | 0.0075532 | pseudogene |

| | | | | |
|---|---|---|---|---|
| Tb427VSG-531 | 1.18E+03 | 1.69994 | 0.0016698 | mVSG ES |
| Tb427VSG-3303 | 4.86E+03 | 1.614583 | 0.0055186 | pseudogene |
| Tb427VSG-3104 | 4.56E+03 | 1.599405 | 0.005474 | pseudogene |
| Tb427VSG-1347 | 2.47E+02 | 1.572806 | 0.0044442 | pseudogene |
| Tb427VSG-1521 | 2.09E+02 | 1.487125 | 0.0106162 | pseudogene |
| Tb427VSG-350 | 1.11E+02 | 1.481647 | 0.0101408 | array |
| Tb427VSG-1267 | 7.01E+02 | 1.416988 | 0.0100395 | pseudogene |
| Tb427VSG-2 | 4.64E+06 | -2.990698 | 0 | BES active |

**Table 8.5 List of expressed VSGs after 36 hr of TbRH2A RNAi**

VSG genes found to be significantly ($p < 0.005$) differential expressed after 36 hr of TbRH2A RNAi in induced samples relative to un-induced expressed as Log2 fold-change (FC) in read counts. The final VSG, Tb427VSG-2 (also named VSG221 in this thesis), is housed in the dominant BES and is down-regulated in comparison to all others found to be upregulated. Whether each VSG is predicted to be found intact within silent arrays (array), a BES or an mVSG ES, or be pseudogenic is also show.

# List of References

Van Den Abbeele, J. *et al.* (1999) 'Trypanosoma brucei spp. development in the tsetse fly: characterization of the post-mesocyclic stages in the foregut and proboscis.', *Parasitology*, 118 ( Pt 5), pp. 469–78.

Abebe, R., Gute, S. and Simon, I. (2017) 'Bovine trypanosomosis and vector density in Omo-Ghibe tsetse belt, South Ethiopia.', *Acta tropica*, 167, pp. 79–85.

Abu-Elneel, K., Kapeller, I. and Shlomai, J. (1999) 'Universal minicircle sequence-binding protein, a sequence-specific DNA-binding protein that recognizes the two replication origins of the kinetoplast DNA minicircle.', *The Journal of biological chemistry*, 274(19), pp. 13419–26.

Aguilera, A. and García-Muse, T. (2012) 'R Loops: From Transcription Byproducts to Threats to Genome Stability', *Molecular Cell*. Cell Press, 46(2), pp. 115–124.

Aguilera, A. and Gómez-González, B. (2008) 'Genome instability: a mechanistic view of its causes and consequences', *Nature Reviews Genetics*. Nature Publishing Group, 9(3), pp. 204–217.

Aitcheson, N. *et al.* (2005) 'VSG switching in Trypanosoma brucei: antigenic variation analysed using RNAi in the absence of immune selection.', *Molecular microbiology*, 57(6), pp. 1608–22.

Akiyoshi, B. and Gull, K. (2014) 'Discovery of Unconventional Kinetochores in Kinetoplastids', *Cell*. Cell Press, 156(6), pp. 1247–1258.

Al-Hadid, Q. and Yang, Y. (2016) 'R-loop: an emerging regulator of chromatin dynamics', *Acta Biochimica et Biophysica Sinica*. Oxford University Press, 48(7), pp. 623–631.

Alarcon, C. M. *et al.* (1994) 'A monocistronic transcript for a trypanosome variant surface glycoprotein.', *Molecular and cellular biology*. American Society for Microbiology (ASM), 14(8), pp. 5579–91.

Alsford, S. *et al.* (2005) 'Tagging a T. brucei RRNA locus improves stable transfection efficiency and circumvents inducible expression position effects.', *Molecular and biochemical parasitology*. Europe PMC Funders, 144(2), pp. 142–8.

Alsford, S. and Horn, D. (2008) 'Single-locus targeting constructs for reliable regulated RNAi and transgene expression in Trypanosoma brucei.', *Molecular and biochemical parasitology*, 161(1), pp. 76–9.

Alsford, S. and Horn, D. (2012) 'Cell-cycle-regulated control of VSG expression site silencing by histones and histone chaperones ASF1A and CAF-1b in Trypanosoma brucei', *Nucleic Acids Research*. Oxford University Press, 40(20), p. 10150.

Alzu, A. *et al.* (2012) 'Senataxin Associates with Replication Forks to Protect Fork Integrity across RNA-Polymerase-II-Transcribed Genes', *Cell*. Elsevier, 151(4), pp. 835–846.

Amon, J. D. and Koshland, D. (2016) 'RNase H enables efficient repair of R-loop induced DNA damage.', *eLife*. eLife Sciences Publications, Ltd, 5.

Aoki, T. and Oya, H. (1987) 'Regulatory properties of carbamoyl-phosphate

synthetase II from the parasitic protozoan Crithidia fasciculata.', *Comparative biochemistry and physiology. B, Comparative biochemistry*, 87(4), pp. 655–8.

Aphasizhev, R. and Aphasizheva, I. (2011) 'Uridine insertion/deletion editing in trypanosomes: a playground for RNA-guided information transfer', *Wiley Interdisciplinary Reviews: RNA*, 2(5), pp. 669–685.

Arana, M. E. *et al.* (2012) 'Transcriptional responses to loss of RNase H2 in Saccharomyces cerevisiae', *DNA Repair*, 11(12), pp. 933–941.

Arning, L. *et al.* (2013) 'The SETX missense variation spectrum as evaluated in patients with ALS4-like motor neuron diseases', *neurogenetics*. Springer-Verlag, 14(1), pp. 53–61.

Arora, R. *et al.* (2014) 'RNaseH1 regulates TERRA-telomeric DNA hybrids and telomere maintenance in ALT tumour cells.', *Nature communications*, 5, p. 5220.

Arora, R. and Azzalin, C. M. (2015) 'Telomere elongation chooses TERRA ALTernatives', *RNA Biology*. Taylor & Francis, 12(9), pp. 938–941.

Arudchandran, A. *et al.* (2002) 'Multiple ribonuclease H-encoding genes in the Caenorhabditis elegans genome contrasts with the two typical ribonuclease H-encoding genes in the human genome.', *Molecular biology and evolution*, 19(11), pp. 1910–9.

Asai, T. and Kogoma, T. (1994) 'D-loops and R-loops: alternative mechanisms for the initiation of chromosome replication in Escherichia coli.', *Journal of bacteriology*, 176(7), pp. 1807–12.

Aslett, M. *et al.* (2010) 'TriTrypDB: a functional genomic resource for the Trypanosomatidae', *Nucleic Acids Research*. Oxford University Press, 38(suppl_1), pp. D457–D462.

Avrahami, D., Tzfati, Y. and Shlomai, J. (1995) 'A single-stranded DNA binding protein binds the origin of replication of the duplex kinetoplast DNA.', *Proceedings of the National Academy of Sciences of the United States of America*, 92(23), pp. 10511–5.

Bailey, T. L. and Elkan, C. (1994) 'Fitting a mixture model by expectation maximization to discover motifs in biopolymers.', *Proceedings. International Conference on Intelligent Systems for Molecular Biology*, 2, pp. 28–36.

Balk, B. *et al.* (2013) 'Telomeric RNA-DNA hybrids affect telomere-length dynamics and senescence', *Nature Structural & Molecular Biology*. Nature Publishing Group, 20(10), pp. 1199–1205.

Barry, J. D. (1997) 'The relative significance of mechanisms of antigenic variation in African trypanosomes', *Parasitology Today*, 13(6), pp. 212–218.

Barry, J. D. and McCulloch, R. (2001) 'Antigenic variation in trypanosomes: enhanced phenotypic variation in a eukaryotic parasite.', *Advances in parasitology*, 49, pp. 1–70.

Bartossek, T. *et al.* (2017) 'Structural basis for the shielding function of the dynamic trypanosome variant surface glycoprotein coat', *Nature Microbiology*, 2(11), pp. 1523–1532.

Basu, U. *et al.* (2011) 'The RNA exosome targets the AID cytidine deaminase to both

strands of transcribed duplex DNA substrates.', *Cell*. Elsevier, 144(3), pp. 353–63.

Becherel, O. J. *et al*. (2013) 'Senataxin Plays an Essential Role with DNA Damage Response Proteins in Meiotic Recombination and Gene Silencing', *PLoS Genetics*. Edited by P. J. McKinnon. Public Library of Science, 9(4), p. e1003435.

Beckers, A. *et al*. (1981) 'Experimental infections of laboratory rodents with recently isolated stocks of Trypanosoma brucei gambiense. 1. Parasitological investigations.', *Zeitschrift fur Parasitenkunde (Berlin, Germany)*, 64(3), pp. 285–96.

Belanger, K. G. and Kreuzer, K. N. (1998) 'Bacteriophage T4 initiates bidirectional DNA replication through a two-step process.', *Molecular cell*, 2(5), pp. 693–701.

Benson, G. (1999) 'Tandem repeats finder: a program to analyze DNA sequences', *Nucleic Acids Research*. Oxford University Press, 27(2), pp. 573–580.

Benz, C. *et al*. (2017) 'Cell cycle synchronisation of Trypanosoma brucei by centrifugal counter-flow elutriation reveals the timing of nuclear and kinetoplast DNA replication', *Scientific Reports*. Nature Publishing Group, 7(1), p. 17599.

Berens, R. L., Krug, E. C. and Marr, J. J. (1995) 'Purine and Pyrimidine Metabolism', in *Biochemistry and Molecular Biology of Parasites*. Elsevier, pp. 89–117.

Berg, J. M. (Jeremy M. *et al*. (2002) *Biochemistry Section 25.4Key Steps in Nucleotide Biosynthesis Are Regulated by Feedback Inhibition*. W.H. Freeman.

Bernards, A. *et al*. (1981) 'Activation of trypanosome surface glycoprotein genes involves a duplication-transposition leading to an altered 3′ end', *Cell*, 27(3), pp. 497–505.

Berriman, M. (2005) 'The Genome of the African Trypanosome Trypanosoma brucei', *Science*, 309(5733), pp. 416–422.

Besnard, E. *et al*. (2012) 'Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs.', *Nature structural & molecular biology*, 19(8), pp. 837–44.

Bianchi, A. and Shore, D. (2007) 'Early Replication of Short Telomeres in Budding Yeast', *Cell*. Cell Press, 128(6), pp. 1051–1062.

Bierhoff, H. *et al*. (2010) 'Noncoding Transcripts in Sense and Antisense Orientation Regulate the Epigenetic State of Ribosomal RNA Genes', *Cold Spring Harbor Symposia on Quantitative Biology*, 75(0), pp. 357–364.

Boitz, J. M. *et al*. (2012) 'Purine salvage in Leishmania: complex or simple by design?', *Trends in parasitology*. NIH Public Access, 28(8), pp. 345–52.

Bonnet, A. *et al*. (2017) 'Introns Protect Eukaryotic Genomes from Transcription-Associated Genetic Instability', *Molecular Cell*, 67(4), p. 608–621.e6.

Boothroyd, C. E. *et al*. (2009) 'A yeast-endonuclease-generated DNA break induces antigenic switching in Trypanosoma brucei.', *Nature*, 459(7244), pp. 278–81.

Boque-Sastre, R. *et al*. (2015) 'Head-to-head antisense transcription and R-loop formation promotes transcriptional activation.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of

Sciences, 112(18), pp. 5785–90.

Brandenburg, J. *et al.* (2007) 'Multifunctional class I transcription in Trypanosoma brucei depends on a novel protein complex.', *The EMBO journal*, 26(23), pp. 4856–66.

Bringaud, F. *et al.* (2002) 'A new, expressed multigene family containing a hot spot for insertion of retroelements is associated with polymorphic subtelomeric regions of Trypanosoma brucei.', *Eukaryotic cell*. American Society for Microbiology (ASM), 1(1), pp. 137–51.

Brogna, S. and Ashburner, M. (1997) 'The Adh-related gene of Drosophila melanogaster is expressed as a functional dicistronic messenger RNA: multigenic transcription in higher organisms.', *The EMBO journal*. European Molecular Biology Organization, 16(8), pp. 2023–31.

Bubeck, D. *et al.* (2011) 'PCNA directs type 2 RNase H activity on DNA replication and repair substrates', *Nucleic Acids Research*, 39(9), pp. 3652–3666.

Büscher, P. *et al.* (2017) 'Human African trypanosomiasis', *The Lancet*, pp. 2397–2409.

Calderano, S. G. *et al.* (2015) 'Single molecule analysis of Trypanosoma brucei DNA replication dynamics.', *Nucleic acids research*, 43(5), pp. 2655–65.

Caljon, G. *et al.* (2016) 'The Dermis as a Delivery Site of Trypanosoma brucei for Tsetse Flies', *PLOS Pathogens*. Edited by N. C. Peters. Public Library of Science, 12(7), p. e1005744.

Callejas, S. *et al.* (2006) 'Hemizygous subtelomeres of an African trypanosome chromosome may account for over 75% of chromosome length.', *Genome research*, 16(9), pp. 1109–18.

Campbell, D. A., Thomas, S. and Sturm, N. R. (2003) 'Transcription in kinetoplastid protozoa: why be normal?', *Microbes and infection*, 5(13), pp. 1231–40.

Camps, M. and Loeb, L. A. (2005) 'Critical role of R-loops in processing replication blocks.', *Frontiers in bioscience : a journal and virtual library*, 10, pp. 689–98.

Capewell, P. *et al.* (2016) 'The skin is a significant but overlooked anatomical reservoir for vector-borne African trypanosomes', *eLife*, 5.

Carles-Kinch, K. and Kreuzer, K. N. (1997) 'RNA-DNA hybrid formation at a bacteriophage T4 replication origin 1 1Edited by M. Gottesmann', *Journal of Molecular Biology*, 266(5), pp. 915–926.

Castellano-Pozo, M. *et al.* (2013) 'R loops are linked to histone H3 S10 phosphorylation and chromatin condensation.', *Molecular cell*. Elsevier, 52(4), pp. 583–90.

Castillo-Acosta, V. M. *et al.* (2012) 'Increased uracil insertion in DNA is cytotoxic and increases the frequency of mutation, double strand break formation and VSG switching in Trypanosoma brucei.', *DNA repair*, 11(12), pp. 986–95.

Castillo-Acosta, V. M. *et al.* (2012) 'Trypanosomes lacking uracil-DNA glycosylase are hypersensitive to antifolates and present a mutator phenotype', *The International Journal of Biochemistry & Cell Biology*, 44(9), pp. 1555–1568.

Cayrou, C. *et al*. (2011) 'Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features.', *Genome research*, 21(9), pp. 1438–49.

Ceron, C. R. *et al*. (1979) 'Purine metabolism in trypanosomatids.', *The Journal of protozoology*, 26(3), pp. 479–83.

Cerritelli, S. M. *et al*. (2003) 'Failure to Produce Mitochondrial DNA Results in Embryonic Lethality in Rnaseh1 Null Mice', *Molecular Cell*, 11(3), pp. 807–815.

Cerritelli, S. M. and Crouch, R. J. (2009) 'Ribonuclease H: the enzymes in eukaryotes.', *The FEBS journal*, 276(6), pp. 1494–505.

Chan, Y. A. *et al*. (2014) 'Genome-Wide Profiling of Yeast DNA:RNA Hybrid Prone Sites with DRIP-Chip', *PLoS Genetics*. Edited by M. Snyder. Public Library of Science, 10(4), p. e1004288.

Chan, Y. A., Hieter, P. and Stirling, P. C. (2014) 'Mechanisms of genome instability induced by RNA-processing defects.', *Trends in genetics : TIG*, 30(6), pp. 245–53.

Chang, D. D. and Clayton, D. A. (1985) 'Priming of human mitochondrial DNA replication occurs at the light-strand promoter.', *Proceedings of the National Academy of Sciences of the United States of America*, 82(2), pp. 351–5.

Chang, D. D., Hauswirth, W. W. and Clayton, D. A. (1985) 'Replication priming and transcription initiate from precisely the same site in mouse mitochondrial DNA.', *The EMBO journal*, 4(6), pp. 1559–67.

Chang, E. Y.-C. *et al*. (2017) 'RECQ-like helicases Sgs1 and BLM regulate R-loop-associated genome instability', *The Journal of Cell Biology*, 216(12), pp. 3991–4005.

Chaves, I. *et al*. (1998) 'Subnuclear localization of the active variant surface glycoprotein gene expression site in Trypanosoma brucei.', *Proceedings of the National Academy of Sciences of the United States of America*, 95(21), pp. 12328–33.

Chaves, I. *et al*. (1999) 'Control of variant surface glycoprotein gene-expression sites in Trypanosoma brucei', *The EMBO Journal*, 18(17), pp. 4846–4855.

Checchi, F. *et al*. (2008) 'Estimates of the duration of the early and late stage of gambiense sleeping sickness', *BMC Infectious Diseases*. BioMed Central, 8(1), p. 16.

Chédin, F. (2016) 'Nascent Connections: R-Loops and Chromatin Patterning.', *Trends in genetics : TIG*. NIH Public Access, 32(12), pp. 828–838.

Chen, J.-M. *et al*. (2007) 'Gene conversion: mechanisms, evolution and human disease', *Nature Reviews Genetics*. Nature Publishing Group, 8(10), pp. 762–775.

Chen, K. *et al*. (2015) 'The Overlooked Fact: Fundamental Need for Spike-In Control for Virtually All Genome-Wide Analyses.', *Molecular and cellular biology*. American Society for Microbiology (ASM), 36(5), pp. 662–7.

Chen, L. *et al*. (2017) 'R-ChIP Using Inactive RNase H Reveals Dynamic Coupling of R-loops with Transcriptional Pausing at Gene Promoters', *Molecular Cell*. Cell Press, 68(4), p. 745–757.e5.

Chen, P. B. *et al*. (2015) 'R loops regulate promoter-proximal chromatin architecture and cellular differentiation', *Nature Structural & Molecular Biology*.

Nature Publishing Group, 22(12), pp. 999–1007.

Chen, Y.-Z. *et al.* (2004) 'DNA/RNA helicase gene mutations in a form of juvenile amyotrophic lateral sclerosis (ALS4).', *American journal of human genetics*. Elsevier, 74(6), pp. 1128–35.

Chon, H. *et al.* (2009) 'Contributions of the two accessory subunits, RNASEH2B and RNASEH2C, to the activity and properties of the human RNase H2 complex.', *Nucleic acids research*, 37(1), pp. 96–110.

Chon, H. *et al.* (2013) 'RNase H2 roles in genome integrity revealed by unlinking its activities.', *Nucleic acids research*, 41(5), pp. 3130–43.

Clayton, C. (2016) 'Gene expression in Kinetoplastids', *Current Opinion in Microbiology*, 32, pp. 46–51.

Clayton, C. E. (2014) 'Networks of gene expression regulation in Trypanosoma brucei', *Molecular and Biochemical Parasitology*, 195(2), pp. 96–106.

Cliffe, L. J. *et al.* (2010) 'Two thymidine hydroxylases differentially regulate the formation of glucosylated DNA at regions flanking polymerase II polycistronic transcription units throughout the genome of Trypanosoma brucei.', *Nucleic acids research*, 38(12), pp. 3923–35.

Conover, H. N. *et al.* (2015) 'Stimulation of Chromosomal Rearrangements by Ribonucleotides', *Genetics*, 201(3), pp. 951–961.

Conway, C. *et al.* (2002) 'Two pathways of homologous recombination in Trypanosoma brucei.', *Molecular microbiology*, 45(6), pp. 1687–700.

Cornelio, D. A. *et al.* (2017) 'Both R-loop removal and ribonucleotide excision repair activities of RNase H2 contribute substantially to chromosome stability', *DNA Repair*. Elsevier, 52, pp. 110–114.

Costantino, L. and Koshland, D. (2015) 'The Yin and Yang of R-loop biology.', *Current opinion in cell biology*. NIH Public Access, 34, pp. 39–45.

Cristini, A. *et al.* (2018) 'RNA/DNA Hybrid Interactome Identifies DXH9 as a Molecular Player in Transcriptional Termination and R-Loop-Associated DNA Damage', *Cell Reports*, 23(6), pp. 1891–1905.

Cross, G. A. (1975) 'Identification, purification and properties of clone-specific glycoprotein antigens constituting the surface coat of Trypanosoma brucei.', *Parasitology*, 71(3), pp. 393–417.

Cross, G. A. M., Kim, H.-S. and Wickstead, B. (2014) 'Capturing the variant surface glycoprotein repertoire (the VSGnome) of Trypanosoma brucei Lister 427.', *Molecular and biochemical parasitology*, 195(1), pp. 59–73.

Crow, Y. J. *et al.* (2006) 'Mutations in genes encoding ribonuclease H2 subunits cause Aicardi-Goutières syndrome and mimic congenital viral brain infection', *Nature Genetics*, 38(8), pp. 910–916.

Crow, Y. J. *et al.* (2015) 'Characterization of human disease phenotypes associated with mutations in *TREX1* , *RNASEH2A* , *RNASEH2B* , *RNASEH2C* , *SAMHD1* , *ADAR* , and *IFIH1*', *American Journal of Medical Genetics Part A*. Wiley-Blackwell, 167(2), pp. 296–312.

Cusanelli, E. and Chartrand, P. (2015) ‘Telomeric repeat-containing RNA TERRA: a noncoding RNA connecting telomere biology to genome integrity.’, *Frontiers in genetics*, 6, p. 143.

D’Archivio, S. and Wickstead, B. (2017) ‘Trypanosome outer kinetochore proteins suggest conservation of chromosome segregation machinery across eukaryotes.’, *The Journal of cell biology*. Rockefeller University Press, 216(2), pp. 379–391.

Dang, H. Q. and Li, Z. (2011) ‘The Cdc45·Mcm2–7·GINS Protein Complex in Trypanosomes Regulates DNA Replication and Interacts with Two Orc1-like Proteins in the Origin Recognition Complex’, *Journal of Biological Chemistry*, 286(37), pp. 32424–32435.

Das, A. *et al.* (2006) ‘Biochemical characterization of Trypanosoma brucei RNA polymerase II’, *Molecular and Biochemical Parasitology*. Elsevier, 150(2), pp. 201–210.

Dasgupta, S., Masukata, H. and Tomizawa, J. (1987) ‘Multiple mechanisms for initiation of ColE1 DNA replication: DNA synthesis in the presence and absence of ribonuclease H.’, *Cell*. Elsevier, 51(6), pp. 1113–22.

David Barry, J. and McCulloch, R. (2001) *Advances in Parasitology Volume 49*, *Advances in Parasitology*. Elsevier (Advances in Parasitology).

Denninger, V. *et al.* (2010) ‘The FACT subunit TbSpt16 is involved in cell cycle specific control of VSG expression sites in Trypanosoma brucei’, *Molecular Microbiology*, 78(2), pp. 459–474.

Denninger, V. and Rudenko, G. (2014) ‘FACT plays a major role in histone dynamics affecting VSG expression site control in Trypanosoma brucei.’, *Molecular microbiology*. Wiley-Blackwell, 94(4), pp. 945–62.

Devlin, R. *et al.* (2016) ‘Mapping replication dynamics in *Trypanosoma brucei* reveals a link with telomere transcription and antigenic variation’, *eLife*. eLife Sciences Publications Limited, 5, p. e12765.

Dobson, R. *et al.* (2011) ‘Interactions among Trypanosoma brucei RAD51 paralogues in DNA repair and antigenic variation.’, *Molecular microbiology*, 81(2), pp. 434–56.

Dreesen, O. and Cross, G. A. M. (2006) ‘Consequences of telomere shortening at an active VSG expression site in telomerase-deficient Trypanosoma brucei.’, *Eukaryotic cell*, 5(12), pp. 2114–9.

Dreesen, O., Li, B. and Cross, G. A. M. (2007) ‘Telomere structure and function in trypanosomes: a proposal.’, *Nature reviews. Microbiology*, 5(1), pp. 70–5.

DuBois, K. N. *et al.* (2012) ‘NUP-1 Is a Large Coiled-Coil Nucleoskeletal Protein in Trypanosomes with Lamin-Like Functions’, *PLoS Biology*. Edited by T. Misteli, 10(3), p. e1001287.

Dumelie, J. G. and Jaffrey, S. R. (2017) ‘Defining the location of promoter-associated R-loops at near-nucleotide resolution using bisDRIP-seq’, *eLife*. eLife Sciences Publications Limited, 6, p. e28306.

Duncker, B. P., Chesnokov, I. N. and McConkey, B. J. (2009) ‘The origin recognition complex protein family.’, *Genome biology*. BioMed Central, 10(3), p. 214.

Dunn, K. and Griffith, J. D. (1980) ‘The presence of RNA in a double helix inhibits

its interaction with histone protein.', *Nucleic acids research*, 8(3), pp. 555–66.

Echeverry, M. C. *et al.* (2012) 'Centromere-associated repeat arrays on Trypanosoma brucei chromosomes are much more extensive than predicted', *BMC Genomics*, 13(1), p. 29.

Egan, B. *et al.* (2016) 'An Alternative Approach to ChIP-Seq Normalization Enables Detection of Genome-Wide Changes in Histone H3 Lysine 27 Trimethylation upon EZH2 Inhibition', *PLOS ONE*. Edited by Z. Qin. Public Library of Science, 11(11), p. e0166438.

Epshtein, A., Potenski, C. and Klein, H. (2016) 'Increased spontaneous recombination in RNase H2- deficient cells arises from multiple contiguous rNMPs and not from single rNMP residues incorporated by DNA polymerase epsilon', *Microbial Cell*, 3(6), pp. 248–254.

Fadda, A. *et al.* (2014) 'Transcriptome-wide analysis of trypanosome mRNA decay reveals complex degradation kinetics and suggests a role for co-transcriptional degradation in determining mRNA levels.', *Molecular microbiology*. Wiley-Blackwell, 94(2), pp. 307–26.

Fairlamb, A. H. and Horn, D. (2018) 'Melarsoprol Resistance in African Trypanosomiasis.', *Trends in parasitology*. Elsevier, 34(6), pp. 481–492.

Fenn, K. and Matthews, K. R. (2007) 'The cell biology of Trypanosoma brucei differentiation.', 10(6), pp. 539–46.

Ferguson, M. A. *et al.* (1988) 'The glycosylphosphatidylinositol membrane anchor of Trypanosoma brucei variant surface glycoprotein.', *Biochemical Society transactions*, 16(3), pp. 265–8.

Figueiredo, L. M. and Cross, G. A. M. (2010) 'Nucleosomes Are Depleted at the VSG Expression Site Transcribed by RNA Polymerase I in African Trypanosomes', *Eukaryotic Cell*, 9(1), pp. 148–154.

Figueiredo, L. M., Janzen, C. J. and Cross, G. A. M. (2008) 'A histone methyltransferase modulates antigenic variation in African trypanosomes.', *PLoS biology*. Public Library of Science, 6(7), p. e161.

Franco, J. *et al.* (2012) 'Monitoring the use of nifurtimox-eflornithine combination therapy (NECT) in the treatment of second stage gambiense human African trypanosomiasis', *Research and Reports in Tropical Medicine*. Dove Press, 3, p. 93.

Franco, J. R. *et al.* (2014) 'Epidemiology of human African trypanosomiasis.', *Clinical epidemiology*. Dove Press, 6, pp. 257–75.

Freudenreich, C. H. (2018) 'R-loops: targets for nuclease cleavage and repeat instability', *Current Genetics*.

Gaidamakov, S. A. *et al.* (2005) 'Eukaryotic RNases H1 act processively by interactions through the duplex RNA-binding domain', *Nucleic Acids Research*, 33(7), pp. 2166–2175.

Gall, J. G. and Pardue, M. L. (1969) 'Formation and detection of RNA-DNA hybrid molecules in cytological preparations.', *Proceedings of the National Academy of Sciences of the United States of America*, 63(2), pp. 378–83.

Gan, W. *et al.* (2011) 'R-loop-mediated genomic instability is caused by impairment

of replication fork progression.', *Genes & development*. Cold Spring Harbor Laboratory Press, 25(19), pp. 2041–56.

Garcia-Silva, M. *et al*. (2017) 'Identification of the centromeres of *Leishmania major* : revealing the hidden pieces', *EMBO reports*, 18(11), pp. 1968–1977.

Ginger, M. L. *et al*. (2002) 'Ex vivo and in vitro identification of a consensus promoter for VSG genes expressed by metacyclic-stage trypanosomes in the tsetse fly.', *Eukaryotic cell*. American Society for Microbiology, 1(6), pp. 1000–9.

Ginno, P. A. *et al*. (2012) 'R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters.', *Molecular cell*, 45(6), pp. 814–25.

Ginno, P. A. *et al*. (2013) 'GC skew at the 5' and 3' ends of human genes links R-loop formation to epigenetic regulation and transcription termination.', *Genome research*. Cold Spring Harbor Laboratory Press, 23(10), pp. 1590–600.

Giordani, F. *et al*. (2016) 'The animal trypanosomiases and their chemotherapy: a review.', *Parasitology*. Cambridge University Press, 143(14), pp. 1862–1889.

Glover, L. *et al*. (2007) 'Deletion of a trypanosome telomere leads to loss of silencing and progressive loss of terminal DNA in the absence of cell cycle arrest.', *Nucleic acids research*. Oxford University Press, 35(3), pp. 872–80.

Glover, L. *et al*. (2016) 'VEX1 controls the allelic exclusion required for antigenic variation in trypanosomes', *Proceedings of the National Academy of Sciences*, 113(26), pp. 7225–7230.

Glover, L., Alsford, S. and Horn, D. (2013) 'DNA break site at fragile subtelomeres determines probability and mechanism of antigenic variation in African trypanosomes.', *PLoS pathogens*, 9(3), p. e1003260.

Glover, L. and Horn, D. (2012) 'Trypanosomal histone γH2A and the DNA damage response.', *Molecular and biochemical parasitology*, 183(1), pp. 78–83.

Glover, L., Jun, J. and Horn, D. (2011) 'Microhomology-mediated deletion and gene conversion in African trypanosomes.', *Nucleic acids research*, 39(4), pp. 1372–80.

Glover, L., McCulloch, R. and Horn, D. (2008) 'Sequence homology and microhomology dominate chromosomal double-strand break repair in African trypanosomes.', *Nucleic acids research*, 36(8), pp. 2608–18.

Godoy, P. D. d. M. *et al*. (2009) 'Trypanosome Prereplication Machinery Contains a Single Functional Orc1/Cdc6 Protein, Which Is Typical of Archaea', *Eukaryotic Cell*, 8(10), pp. 1592–1603.

Gowrishankar, J., Leela, J. K. and Anupama, K. (2013) 'R-loops in bacterial transcription: their causes and consequences.', *Transcription*. Taylor & Francis, 4(4), pp. 153–7.

Graf, F. E. *et al*. (2013) 'Aquaporin 2 Mutations in Trypanosoma brucei gambiense Field Isolates Correlate with Decreased Susceptibility to Pentamidine and Melarsoprol', *PLoS Neglected Tropical Diseases*. Edited by E. Matovu. Public Library of Science, 7(10), p. e2475.

Graf, M. *et al*. (2017) 'Telomere Length Determines TERRA and R-Loop Regulation through the Cell Cycle.', *Cell*. Elsevier, 170(1), p. 72–85.e14.

Graham, S. V and Barry, J. D. (1995) 'Transcriptional regulation of metacyclic variant surface glycoprotein gene expression during the life cycle of Trypanosoma brucei.', *Molecular and cellular biology*. American Society for Microbiology (ASM), 15(11), pp. 5945–56.

Groh, M. and Gromak, N. (2014) 'Out of balance: R-loops in human disease.', *PLoS genetics*. Public Library of Science, 10(9), p. e1004630.

Guirnalda, P. *et al*. (2007) 'Anti-Trypanosoma brucei activity in Cape buffalo serum during the cryptic phase of parasitemia is mediated by antibodies', *International Journal for Parasitology*, 37(12), pp. 1391–1399.

Günzl, A. *et al*. (1997) 'Transcription of the Trypanosoma brucei spliced leader RNA gene is dependent only on the presence of upstream regulatory elements.', *Molecular and biochemical parasitology*, 85(1), pp. 67–76.

Günzl, A. *et al*. (2003) 'RNA polymerase I transcribes procyclin genes and variant surface glycoprotein gene expression sites in Trypanosoma brucei.', *Eukaryotic cell*, 2(3), pp. 542–51.

Hagblom, P. *et al*. (1985) 'Intragenic recombination leads to pilus antigenic variation in Neisseria gonorrhoeae.', *Nature*, 315(6015), pp. 156–8.

El Hage, A. *et al*. (2010) 'Loss of Topoisomerase I leads to R-loop-mediated transcriptional blocks during ribosomal RNA synthesis.', *Genes & development*, 24(14), pp. 1546–58.

El Hage, A. *et al*. (2014) 'Genome-wide distribution of RNA-DNA hybrids identifies RNase H targets in tRNA genes, retrotransposons and mitochondria.', *PLoS genetics*, 10(10), p. e1004716.

Hajduk, S. and Ochsenreiter, T. (2010) 'RNA editing in kinetoplastids.', *RNA biology*, 7(2), pp. 229–36.

Halász, L. *et al*. (2017) 'RNA-DNA hybrid (R-loop) immunoprecipitation mapping: an analytical workflow to evaluate inherent biases.', *Genome research*. Cold Spring Harbor Laboratory Press, 27(6), pp. 1063–1073.

Hamperl, S. *et al*. (2017) 'Transcription-Replication Conflict Orientation Modulates R-Loop Levels and Activates Distinct DNA Damage Responses', *Cell*, 170(4), p. 774–786.e19.

Hamperl, S. and Cimprich, K. A. (2014) 'The contribution of co-transcriptional RNA:DNA hybrid structures to DNA damage and genome instability.', *DNA repair*, 19, pp. 84–94.

Hamperl, S. and Cimprich, K. A. (2016) 'Conflict Resolution in the Genome: How Transcription and Replication Make It Work', *Cell*. Cell Press, 167(6), pp. 1455–1467.

Hartley, C. L. and McCulloch, R. (2008) 'Trypanosoma brucei BRCA2 acts in antigenic variation and has undergone a recent expansion in BRC repeat number that is important during homologous recombination.', *Molecular microbiology*, 68(5), pp. 1237–51.

Heinz, S. *et al*. (2010) 'Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell

identities.', *Molecular cell*, 38(4), pp. 576–89.

Herrera-Moyano, E. *et al*. (2014) 'The yeast and human FACT chromatin-reorganizing complexes solve R-loop-mediated transcription-replication conflicts.', *Genes & development*. Cold Spring Harbor Laboratory Press, 28(7), pp. 735–48.

Hertz-Fowler, C. *et al*. (2008) 'Telomeric expression sites are highly conserved in Trypanosoma brucei.', *PloS one*, 3(10), p. e3527.

Hesslein, D. and Campbell, A. (1997) 'Molecular cloning and expression of a ribonuclease H from the kinetoplastid, Trypanosoma brucei1Note: The nucleotide sequence data of the approximately 2.37 kb genomic fragment carrying the gene discussed in this paper has been submitted to Genbank™ under', *Molecular and Biochemical Parasitology*, 86(1), pp. 121–126.

Higgins, M. K. *et al*. (2013) 'Structure of the trypanosome haptoglobin-hemoglobin receptor and implications for nutrient uptake and innate immunity.', *Proceedings of the National Academy of Sciences of the United States of America*, 110(5), pp. 1905–10.

Hiller, B. *et al*. (2012) 'Mammalian RNase H2 removes ribonucleotides from DNA to maintain genome integrity.', *The Journal of experimental medicine*. Rockefeller University Press, 209(8), pp. 1419–26.

Hirumi, H. and 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.', *The Journal of parasitology*, 75(6), pp. 985–9.

Ho, H. H. *et al*. (2006) 'Ordered assembly of the duplicating Golgi in Trypanosoma brucei.', *Proceedings of the National Academy of Sciences of the United States of America*, 103(20), pp. 7676–81.

Holmes, P. (2013) 'Tsetse-transmitted trypanosomes – Their biology, disease impact and control', *Journal of Invertebrate Pathology*. Academic Press, 112, pp. S11–S14.

Hong, X., Cadwell, G. W. and Kogoma, T. (1996) 'Activation of stable DNA replication in rapidly growing Escherichia coli at the time of entry to stationary phase.', *Molecular microbiology*, 21(5), pp. 953–61.

Horn, D. (2004) 'The molecular control of antigenic variation in Trypanosoma brucei.', *Current molecular medicine*, 4(6), pp. 563–76.

Horn, D. and Cross, G. A. M. (1995) 'A developmentally regulated position effect at a telomeric locus in Trypanosoma brucei', *Cell*. Cell Press, 83(4), pp. 555–561.

Hoshina, S. *et al*. (2013) 'Human origin recognition complex binds preferentially to G-quadruplex-preferable RNA and single-stranded DNA.', *The Journal of biological chemistry*, 288(42), pp. 30161–71.

Hovel-Miner, G. *et al*. (2016) 'A Conserved DNA Repeat Promotes Selection of a Diverse Repertoire of Trypanosoma brucei Surface Antigens from the Genomic Archive.', *PLoS genetics*, 12(5), p. e1005994.

Hovel-Miner, G. A. *et al*. (2012) 'Telomere length affects the frequency and mechanism of antigenic variation in Trypanosoma brucei.', *PLoS pathogens*. Public Library of Science, 8(8), p. e1002900.

Hraiky, C., Raymond, M. A. and Drolet, M. (2000) 'RNase H overproduction corrects

a defect at the level of transcription elongation during rRNA synthesis in the absence of DNA topoisomerase I in Escherichia coli.', *The Journal of biological chemistry*, 275(15), pp. 11257–63.

Huang, J. and Van, der P. 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.', 10(12), pp. 3877–85.

Huertas, P. and Aguilera, A. (2003) 'Cotranscriptionally Formed DNA:RNA Hybrids Mediate Transcription Elongation Impairment and Transcription-Associated Recombination', *Molecular Cell*, 12(3), pp. 711–721.

Huppert, J. L. and Balasubramanian, S. (2005) 'Prevalence of quadruplexes in the human genome', *Nucleic Acids Research*, 33(9), pp. 2908–2916.

Hutchinson, S., Glover, L. and Horn, D. (2016) 'High-resolution analysis of multi-copy variant surface glycoprotein gene expression sites in African trypanosomes', *BMC Genomics*. BioMed Central, 17(1), p. 806.

Itaya, M. (1990) 'Isolation and characterization of a second RNase H (RNase HII) of Escherichia coli K-12 encoded by the rnhB gene.', *Proceedings of the National Academy of Sciences of the United States of America*, 87(21), pp. 8587–91.

Itoh, T. and Tomizawa, J. (1980) 'Formation of an RNA primer for initiation of replication of ColE1 DNA by ribonuclease H.', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 77(5), pp. 2450–2454.

Jackson, A. P. *et al.* (2016) 'Kinetoplastid Phylogenomics Reveals the Evolutionary Innovations Associated with the Origins of Parasitism.', *Current biology : CB*. Elsevier, 26(2), pp. 161–172.

Jacobs, R. T. *et al.* (2011) 'Benzoxaboroles: a new class of potential drugs for human African trypanosomiasis', *Future Medicinal Chemistry*, 3(10), pp. 1259–1278.

Jacobs, R. T. *et al.* (2011) 'SCYX-7158, an Orally-Active Benzoxaborole for the Treatment of Stage 2 Human African Trypanosomiasis', *PLoS Neglected Tropical Diseases*. Edited by J. Keiser. Public Library of Science, 5(6), p. e1151.

Jamonneau, V. *et al.* (2012) 'Untreated Human Infections by Trypanosoma brucei gambiense Are Not 100% Fatal', *PLoS Neglected Tropical Diseases*. Edited by J. M. Ndung'u. Public Library of Science, 6(6), p. e1691.

Janzen, C. J. *et al.* (2006) 'Selective di- or trimethylation of histone H3 lysine 76 by two DOT1 homologs is important for cell cycle regulation in Trypanosoma brucei.', *Molecular cell*, 23(4), pp. 497–507.

Jenjaroenpun, P. *et al.* (2015) 'QmRLFS-finder: a model, web server and stand-alone tool for prediction and analysis of R-loop forming sequences', *Nucleic Acids Research*. Oxford University Press, 43(W1), pp. W527–W534.

Jensen, B. C. *et al.* (2014) 'Extensive stage-regulation of translation revealed by ribosome profiling of Trypanosoma brucei', *BMC Genomics*. BioMed Central, 15(1), p. 911.

Jensen, R. E. and Englund, P. T. (2012) 'Network News: The Replication of Kinetoplast DNA', *Annual Review of Microbiology*, 66(1), pp. 473–491.

Jeong, H.-S. *et al.* (2004) 'RNase H2 of Saccharomyces cerevisiae is a complex of

three proteins.', *Nucleic acids research*, 32(2), pp. 407–14.

Johnson, P. J., Kooter, J. M. and Borst, P. (1987) 'Inactivation of transcription by UV irradiation of T. brucei provides evidence for a multicistronic transcription unit including a VSG gene', *Cell*. Elsevier, 51(2), pp. 273–281.

Jones, N. G. *et al.* (2014) 'Regulators of Trypanosoma brucei cell cycle progression and differentiation identified using a kinome-wide RNAi screen.', *PLoS pathogens*, 10(1), p. e1003886.

Juven-Gershon, T. *et al.* (2008) 'The RNA polymerase II core promoter - the gateway to transcription.', *Current opinion in cell biology*. NIH Public Access, 20(3), pp. 253–9.

Kabeche, L. *et al.* (2018) 'A mitosis-specific and R loop-driven ATR pathway promotes faithful chromosome segregation.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 359(6371), pp. 108–114.

Keating, J. *et al.* (2015) 'Human African trypanosomiasis prevention, treatment and control costs: A systematic review', *Acta Tropica*. Elsevier, 150, pp. 4–13.

Kim, H.-S. and Cross, G. A. M. (2010) 'TOPO3alpha influences antigenic variation by monitoring expression-site-associated VSG switching in Trypanosoma brucei.', *PLoS pathogens*, 6(7), p. e1000992.

Kim, H.-S. and Cross, G. A. M. (2011) 'Identification of Trypanosoma brucei RMI1/BLAP75 homologue and its roles in antigenic variation.', *PloS one*, 6(9), p. e25313.

Kind, B. *et al.* (2014) 'Altered spatio-temporal dynamics of RNase H2 complex assembly at replication and repair sites in Aicardi–Goutières syndrome', *Human Molecular Genetics*, 23(22), pp. 5950–5960.

Kind, J. and van Steensel, B. (2010) 'Genome–nuclear lamina interactions and gene regulation', *Current Opinion in Cell Biology*, 22(3), pp. 320–325.

Kinner, A. *et al.* (2008) 'Gamma-H2AX in recognition and signaling of DNA double-strand breaks in the context of chromatin.', *Nucleic acids research*. Oxford University Press, 36(17), pp. 5678–94.

Kobil, J. (2000) 'Trypanosoma brucei RNase HI requires its divergent spacer subdomain for enzymatic function and its conserved RNA binding motif for nuclear localization', *Molecular and Biochemical Parasitology*, 107(1), pp. 135–142.

Kobil, J. H. and Campbell, A. G. (2000) 'Functional analysis of the domain organization of Trypanosoma brucei RNase HI.', *Biochemical and biophysical research communications*, 270(2), pp. 336–42.

Kochiwa, H., Tomita, M. and Kanai, A. (2007) 'Evolution of ribonuclease H genes in prokaryotes to avoid inheritance of redundant genes.', *BMC evolutionary biology*. BioMed Central, 7, p. 128.

Kolev, N. G. *et al.* (2010) 'The Transcriptome of the Human Pathogen Trypanosoma brucei at Single-Nucleotide Resolution', *PLoS Pathogens*. Edited by S. M. Beverley. Public Library of Science, 6(9), p. e1001090.

Kolev, N. G., Günzl, A. and Tschudi, C. (2017) 'Metacyclic VSG expression site promoters are recognized by the same general transcription factor that is required

for RNA polymerase I transcription of bloodstream expression sites', *Molecular and Biochemical Parasitology*. Elsevier, 216, pp. 52–55.

Kong, D. and DePamphilis, M. L. (2001) 'Site-specific DNA binding of the Schizosaccharomyces pombe origin recognition complex is determined by the Orc4 subunit.', *Molecular and cellular biology*, 21(23), pp. 8095–103.

Kyes, S. A., Kraemer, S. M. and Smith, J. D. (2007) 'Antigenic variation in Plasmodium falciparum: gene organization and regulation of the var multigene family.', *Eukaryotic cell*. American Society for Microbiology (ASM), 6(9), pp. 1511–20.

De Kyvon, M.-A. L.-C. *et al.* (2016) 'Congenital Trypanosomiasis in Child Born in France to African Mother.', *Emerging infectious diseases*. Centers for Disease Control and Prevention, 22(5), pp. 935–7.

Landeira, D. *et al.* (2009) 'Cohesin regulates VSG monoallelic expression in trypanosomes.', *The Journal of cell biology*, 186(2), pp. 243–54.

Landeira, D. and Navarro, M. (2007) 'Nuclear repositioning of the *VSG* promoter during developmental silencing in *Trypanosoma brucei*', *The Journal of Cell Biology*, 176(2), pp. 133–139.

Lang, K. S. *et al.* (2017) 'Replication-Transcription Conflicts Generate R-Loops that Orchestrate Bacterial Stress Survival and Pathogenesis', *Cell*, 170(4), p. 787–799.e18.

De Lange, T. *et al.* (1983) 'Telomere conversion in trypanosomes.', *Nucleic acids research*, 11(23), pp. 8149–65.

Lee, D. Y. and Clayton, D. A. (1998) 'Initiation of mitochondrial DNA replication by transcription and R-loop processing.', *The Journal of biological chemistry*. American Society for Biochemistry and Molecular Biology, 273(46), pp. 30614–21.

Lee, S. J. (1991) 'Expression of growth/differentiation factor 1 in the nervous system: conservation of a bicistronic structure.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 88(10), pp. 4250–4.

Lenhard, B., Sandelin, A. and Carninci, P. (2012) 'Metazoan promoters: emerging characteristics and insights into transcriptional regulation', *Nature Reviews Genetics*. Nature Publishing Group, 13(4), pp. 233–245.

Leong, W. Y. *et al.* (2017) 'Spliceosomal Mutations Induce R Loop-Associated ATR Signaling', *Blood*, 130(Suppl 1).

Li, X. and Manley, J. L. (2005) 'Inactivation of the SR Protein Splicing Factor ASF/SF2 Results in Genomic Instability', *Cell*, 122(3), pp. 365–378.

Liang, X.-H., Xu, Y.-X. and Michaeli, S. (2002) 'The spliced leader-associated RNA is a trypanosome-specific sn(o) RNA that has the potential to guide pseudouridine formation on the SL RNA.', *RNA (New York, N.Y.)*. Cold Spring Harbor Laboratory Press, 8(2), pp. 237–46.

Liang, X. -h. *et al.* (2003) 'trans and cis Splicing in Trypanosomatids: Mechanism, Factors, and Regulation', *Eukaryotic Cell*, 2(5), pp. 830–840.

Lin, Y. *et al.* (2010) 'R loops stimulate genetic instability of CTG.CAG repeats.',

*Proceedings of the National Academy of Sciences of the United States of America*, 107(2), pp. 692–7.

Liu, A. Y. C. *et al.* (1983) 'The transposition unit of variant surface glycoprotein gene 118 of Trypanosoma brucei', *Journal of Molecular Biology*, 167(1), pp. 57–75.

Liu, B. *et al.* (2005) 'Fellowship of the rings: the replication of kinetoplast DNA.', *Trends in parasitology*, 21(8), pp. 363–9.

Lombraña, R. *et al.* (2015) 'R-loops and initiation of DNA replication in human cells: a missing link?', *Frontiers in genetics*, 6, p. 158.

López-Estraño, C., Tschudi, C. and Ullu, E. (1998) 'Exonic sequences in the 5' untranslated region of alpha-tubulin mRNA modulate trans splicing in Trypanosoma brucei.', *Molecular and cellular biology*, 18(8), pp. 4620–8.

Love, M. I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*. BioMed Central, 15(12), p. 550.

van Luenen, H. G. A. M. *et al.* (2012) 'Glucosylated hydroxymethyluracil, DNA base J, prevents transcriptional readthrough in Leishmania.', *Cell*. Elsevier, 150(5), pp. 909–21.

Luke, B. *et al.* (2008) 'The Rat1p 5' to 3' exonuclease degrades telomeric repeat-containing RNA and promotes telomere elongation in Saccharomyces cerevisiae.', *Molecular cell*, 32(4), pp. 465–77.

Lukes, J. *et al.* (2002) 'Kinetoplast DNA network: evolution of an improbable structure.', *Eukaryotic cell*. American Society for Microbiology, 1(4), pp. 495–502.

Lutje, V., Seixas, J. and Kennedy, A. (2010) 'Chemotherapy for second-stage Human African trypanosomiasis', in Lutje, V. (ed.) *Cochrane Database of Systematic Reviews*. Chichester, UK: John Wiley & Sons, Ltd, p. CD006201.

MacLean, L. M. *et al.* (2010) 'Focus–Specific Clinical Profiles in Human African Trypanosomiasis Caused by Trypanosoma brucei rhodesiense', *PLoS Neglected Tropical Diseases*. Edited by P. Büscher. Public Library of Science, 4(12), p. e906.

Maduike, N. Z. *et al.* (2014) 'Replication of the Escherichia coli chromosome in RNase HI-deficient cells: multiple initiation regions and fork dynamics.', *Molecular microbiology*, 91(1), pp. 39–56.

Maere, S., Heymans, K. and Kuiper, M. (2005) 'BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks', *Bioinformatics*. Oxford University Press, 21(16), pp. 3448–3449.

Mair, G. *et al.* (2000) 'A new twist in trypanosome RNA metabolism: cis-splicing of pre-mRNA.', *RNA (New York, N.Y.)*. Cold Spring Harbor Laboratory Press, 6(2), pp. 163–9.

Mandelboim, M. *et al.* (2002) 'On the Role of Exon and Intron Sequences in *trans* - Splicing Utilization and cap 4 Modification of the Trypanosomatid *Leptomonas collosoma* SL RNA', *Journal of Biological Chemistry*, 277(38), pp. 35210–35218.

Marahrens, Y. and Stillman, B. (1992) 'A yeast chromosomal origin of DNA replication defined by multiple functional elements.', *Science (New York, N.Y.)*, 255(5046), pp. 817–23.

Marande, W., Lukes, J. and Burger, G. (2005) 'Unique mitochondrial genome structure in diplonemids, the sister group of kinetoplastids.', *Eukaryotic cell*. American Society for Microbiology, 4(6), pp. 1137–46.

Marcello, L. and Barry, J. D. (2007) 'Analysis of the VSG gene silent archive in Trypanosoma brucei reveals that mosaic gene expression is prominent in antigenic variation and is favored by archive substructure.', *Genome research*, 17(9), pp. 1344–52.

Maree, J. P. and Patterton, H.-G. (2014) 'The epigenome of Trypanosoma brucei: a regulatory interface to an unconventional transcriptional machine.', *Biochimica et biophysica acta*, 1839(9), pp. 743–50.

Marques, C. A. *et al.* (2016) 'Diverged composition and regulation of the Trypanosoma brucei origin recognition complex that mediates DNA replication initiation.', *Nucleic acids research*, 44(10), pp. 4763–84.

Martínez-Calvillo, S. *et al.* (2003) 'Transcription of Leishmania major Friedlin chromosome 1 initiates in both directions within a single region.', *Molecular cell*, 11(5), pp. 1291–9.

Martínez-Calvillo, S. *et al.* (2004) 'Transcription initiation and termination on Leishmania major chromosome 3.', *Eukaryotic cell*. American Society for Microbiology (ASM), 3(2), pp. 506–17.

Martinez, P. *et al.* (2010) 'Mammalian Rap1 controls telomere function and gene expression through binding to telomeric and extratelomeric sites', *Nature Cell Biology*. Nature Publishing Group, 12(8), pp. 768–780.

Masukata, H. and Tomizawa, J. (1990) 'A mechanism of formation of a persistent hybrid between elongating RNA and template DNA.', *Cell*, 62(2), pp. 331–8.

Matthews, K. R. (2005) 'The developmental cell biology of Trypanosoma brucei.', *Journal of cell science*, 118(Pt 2), pp. 283–90.

Matthews, K. R. and Gull, K. (1997) 'Commitment to differentiation and cell cycle re-entry are coincident but separable events in the transformation of African trypanosomes from their bloodstream to their insect form.', *Journal of cell science*, 110 ( Pt 20), pp. 2609–18.

Matthews, K. R., Tschudi, C. and Ullu, E. (1994) 'A common pyrimidine-rich motif governs trans-splicing and polyadenylation of tubulin polycistronic pre-mRNA in trypanosomes.', *Genes & development*, 8(4), pp. 491–501.

Mayer, C. *et al.* (2006) 'Intergenic Transcripts Regulate the Epigenetic State of rRNA Genes', *Molecular Cell*, 22(3), pp. 351–361.

McCulloch, R. and Barry, J. D. (1999) 'A role for RAD51 and homologous recombination in Trypanosoma brucei antigenic variation.', *Genes & development*. Cold Spring Harbor Laboratory Press, 13(21), pp. 2875–88.

McCulloch, R., Morrison, L. J. and Hall, J. P. J. (2015) 'DNA Recombination Strategies During Antigenic Variation in the African Trypanosome', *Microbiology Spectrum*.

McCulloch, R., Rudenko, G. and Borst, P. (1997) 'Gene conversions mediating antigenic variation in Trypanosoma brucei can occur in variant surface glycoprotein

expression sites lacking 70-base-pair repeat sequences.', *Molecular and cellular biology*, 17(2), pp. 833–43.

Melville, S. E. *et al.* (1998) 'The molecular karyotype of the megabase chromosomes of Trypanosoma brucei and the assignment of chromosome markers.', *Molecular and biochemical parasitology*, 94(2), pp. 155–73.

Melville, S. E., Gerrard, C. S. and Blackwell, J. M. (1999) 'Multiple Causes of Size Variation in the Diploid Megabase Chromosomes of African Trypanosomes', *Chromosome Research*. Kluwer Academic Publishers, 7(3), pp. 191–203.

Meyer, A. *et al.* (2016) 'Past and Ongoing Tsetse and Animal Trypanosomiasis Control Operations in Five African Countries: A Systematic Review', *PLOS Neglected Tropical Diseases*. Edited by J. Raper. Public Library of Science, 10(12), p. e0005247.

Michaeli, S. (2011) '*Trans* -splicing in trypanosomes: machinery and its impact on the parasite transcriptome', *Future Microbiology*, 6(4), pp. 459–474.

Militello, K. T. *et al.* (2008) 'African trypanosomes contain 5-methylcytosine in nuclear DNA.', *Eukaryotic cell*. American Society for Microbiology, 7(11), pp. 2012–6.

Mischo, H. E. *et al.* (2011) 'Yeast Sen1 helicase protects the genome from transcription-associated instability.', *Molecular cell*. Elsevier, 41(1), pp. 21–32.

Misra, S. *et al.* (2005) 'A type II ribonuclease H from Leishmania mitochondria: an enzyme essential for the growth of the parasite.', *Molecular and biochemical parasitology*, 143(2), pp. 135–45.

Munday, J. C. *et al.* (2014) 'Trypanosoma brucei aquaglyceroporin 2 is a high-affinity transporter for pentamidine and melaminophenyl arsenic drugs and the main genetic determinant of resistance to these drugs', *Journal of Antimicrobial Chemotherapy*, 69(3), pp. 651–663.

Murphy, W. J., Watkins, K. P. and Agabian, N. (1986) 'Identification of a novel Y branch structure as an intermediate in trypanosome mRNA processing: evidence for trans splicing.', *Cell*, 47(4), pp. 517–25.

Nadel, J. *et al.* (2015) 'RNA:DNA hybrids in the human genome have distinctive nucleotide characteristics, chromatin composition, and transcriptional relationships', *Epigenetics & Chromatin*. BioMed Central, 8(1), p. 46.

Nakama, M. *et al.* (2012) 'DNA-RNA hybrid formation mediates RNAi-directed heterochromatin formation', *Genes to Cells*, 17(3), pp. 218–233.

Nanavaty, V. *et al.* (2017) 'Trypanosoma brucei RAP1 maintains telomere and subtelomere integrity by suppressing TERRA and telomeric RNA:DNA hybrids.', *Nucleic acids research*. Oxford University Press, 45(10), pp. 5785–5796.

Narayanan, M. S. *et al.* (2011) 'NLP is a novel transcription regulator involved in VSG expression site control in Trypanosoma brucei', *Nucleic Acids Research*, 39(6), pp. 2018–2031.

Narayanan, M. S. and Rudenko, G. (2013) 'TDP1 is an HMG chromatin protein facilitating RNA polymerase I transcription in African trypanosomes.', *Nucleic acids research*, 41(5), pp. 2981–92.

Navarro, M. and Gull, K. (2001) ‘A pol I transcriptional body associated with VSG mono-allelic expression in Trypanosoma brucei.’, *Nature*, 414(6865), pp. 759–63.

Nguyen, H. D. *et al.* (2017) ‘Functions of Replication Protein A as a Sensor of R Loops and a Regulator of RNaseH1’, *Molecular Cell*, 65(5), p. 832–847.e4.

Nguyen, T. N. *et al.* (2014) ‘Promoter occupancy of the basal class I transcription factor A differs strongly between active and silent VSG expression sites in Trypanosoma brucei’, *Nucleic Acids Research*, 42(5), pp. 3164–3176.

Nilsson, D. *et al.* (2010) ‘Spliced Leader Trapping Reveals Widespread Alternative Splicing Patterns in the Highly Dynamic Transcriptome of Trypanosoma brucei’, *PLoS Pathogens*. Edited by M. Parsons. Public Library of Science, 6(8), p. e1001037.

Norris, S. J. (2014) ‘vls Antigenic Variation Systems of Lyme Disease Borrelia: Eluding Host Immunity through both Random, Segmental Gene Conversion and Framework Heterogeneity.’, *Microbiology spectrum*. NIH Public Access, 2(6).

Nowotny, M. *et al.* (2005) ‘Crystal structures of RNase H bound to an RNA/DNA hybrid: substrate specificity and metal-dependent catalysis.’, *Cell*. Elsevier, 121(7), pp. 1005–16.

Nowotny, M. *et al.* (2008) ‘Specific recognition of RNA/DNA hybrid and enhancement of human RNase H1 activity by HBD.’, *The EMBO journal*. European Molecular Biology Organization, 27(7), pp. 1172–81.

O’Connell, K., Jinks-Robertson, S. and Petes, T. D. (2015) ‘Elevated Genome-Wide Instability in Yeast Mutants Lacking RNase H Activity.’, *Genetics*, 201(3), pp. 963–75.

Odiit, M., Kansiime, F. and Enyaru, J. C. (1997) ‘Duration of symptoms and case fatality of sleeping sickness caused by Trypanosoma brucei rhodesiense in Tororo, Uganda.’, *East African medical journal*, 74(12), pp. 792–5.

Ogbadoyi, E. *et al.* (2000) ‘Architecture of the Trypanosoma brucei nucleus during interphase and mitosis.’, *Chromosoma*, 108(8), pp. 501–13.

Ohle, C. *et al.* (2016) ‘Transient RNA-DNA Hybrids Are Required for Efficient Double-Strand Break Repair.’, *Cell*. Elsevier, 167(4), p. 1001–1013.e7.

Ohshima, K. *et al.* (1996) ‘TTA.TAA triplet repeats in plasmids form a non-H bonded structure.’, *The Journal of biological chemistry*, 271(28), pp. 16784–91.

Ohtani, N. *et al.* (1999) ‘Molecular diversities of RNases H.’, *Journal of bioscience and bioengineering*, 88(1), pp. 12–9.

Orlando, D. A. *et al.* (2014) ‘Quantitative ChIP-Seq normalization reveals global modulation of the epigenome.’, *Cell reports*. Elsevier, 9(3), pp. 1163–70.

Özer, Ö. and Hickson, I. D. (2018) ‘Pathways for maintenance of telomeres and common fragile sites during DNA replication stress.’, *Open biology*. The Royal Society, 8(4).

Pan, X. *et al.* (2010) ‘Transcription of AAT•ATT Triplet Repeats in Escherichia coli Is Silenced by H-NS and IS1E Transposition’, *PLoS ONE*. Edited by D. T. Kirkpatrick, 5(12), p. e14271.

Pan, X. *et al.* (2014) ‘R-loop structure: the formation and the effects on genomic

stability.', *Yi chuan = Hereditas*, 36(12), pp. 1185–94.

Parajuli, S. *et al.* (2017) 'Human ribonuclease H1 resolves R-loops and thereby enables progression of the DNA replication fork.', *The Journal of biological chemistry*. American Society for Biochemistry and Molecular Biology, 292(37), pp. 15216–15224.

Patrick, K. L. *et al.* (2009) 'Distinct and overlapping roles for two Dicer-like proteins in the RNA interference pathways of the ancient eukaryote Trypanosoma brucei', *Proceedings of the National Academy of Sciences*, 106(42), pp. 17933–17938.

Paull, T. T. and Deshpande, R. A. (2014) 'The Mre11/Rad50/Nbs1 complex: Recent insights into catalytic activities and ATP-driven conformational changes', *Experimental Cell Research*, 329(1), pp. 139–147.

Pays, E. *et al.* (1983) 'At least two transposed sequences are associated in the expression site of a surface antigen gene in different Trypanosome clones', *Cell*, 34(2), pp. 359–369.

Pays, E. *et al.* (1985) 'Telomeric reciprocal recombination as a possible mechanism for antigenic variation in trypanosomes.', *Nature*, 316(6028), pp. 562–4.

Perry, K. L., Watkins, K. P. and Agabian, N. (1987) 'Trypanosome mRNAs have unusual "cap 4" structures acquired by addition of a spliced leader', *Proceedings of the National Academy of Sciences*, 84(23).

Picard, F. *et al.* (2014) 'The spatiotemporal program of DNA replication is associated with specific combinations of chromatin marks in human cells.', *PLoS genetics*, 10(5), p. e1004282.

Pokatayev, V. *et al.* (2016) 'RNase H2 catalytic core Aicardi-Goutières syndrome-related mutant invokes cGAS-STING innate immune-sensing pathway in mice.', *The Journal of experimental medicine*. Rockefeller University Press, 213(3), pp. 329–36.

Povelones, M. L. *et al.* (2012) 'Histone H1 Plays a Role in Heterochromatin Formation and VSG Expression Site Silencing in Trypanosoma brucei', *PLoS Pathogens*. Edited by E. Ullu. Public Library of Science, 8(11), p. e1003010.

Powell, W. T. *et al.* (2013) 'R-loop formation at Snord116 mediates topotecan inhibition of Ube3a-antisense and allele-specific chromatin decondensation.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 110(34), pp. 13938–43.

Priotto, G. *et al.* (2009) 'Nifurtimox-eflornithine combination therapy for second-stage African Trypanosoma brucei gambiense trypanosomiasis: a multicentre, randomised, phase III, non-inferiority trial', *The Lancet*, 374(9683), pp. 56–64.

Proudfoot, C. and McCulloch, R. (2005) 'Distinct roles for two RAD51-related genes in Trypanosoma brucei antigenic variation.', *Nucleic acids research*, 33(21), pp. 6906–19.

Quinlan, A. R. and Hall, I. M. (2010) 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics*. Oxford University Press, 26(6), pp. 841–842.

Rach, E. A. *et al.* (2011) 'Transcription Initiation Patterns Indicate Divergent

Strategies for Gene Regulation at the Chromatin Level', *PLoS Genetics*. Edited by J. D. Lieb. Public Library of Science, 7(1), p. e1001274.

Ramirez, F. *et al.* (2014) 'deepTools: a flexible platform for exploring deep-sequencing data', *Nucleic Acids Research*, 42(W1), pp. W187–W191.

Reijns, M. A. M. *et al.* (2012) 'Enzymatic removal of ribonucleotides from DNA is essential for mammalian genome integrity and development.', *Cell*, 149(5), pp. 1008–22.

Reuner, B. *et al.* (1997) 'Cell density triggers slender to stumpy differentiation of Trypanosoma brucei bloodstream forms in culture', *Molecular and Biochemical Parasitology*, 90(1), pp. 269–280.

Reynolds, D. *et al.* (2014) 'Regulation of transcription termination by glucosylated hydroxymethyluracil, base J, in Leishmania major and Trypanosoma brucei.', *Nucleic acids research*. Oxford University Press, 42(15), pp. 9717–29.

Reynolds, D. *et al.* (2016) 'Histone H3 Variant Regulates RNA Polymerase II Transcription Termination and Dual Strand Transcription of siRNA Loci in Trypanosoma brucei', *PLOS Genetics*. Edited by L. Figueiredo. Public Library of Science, 12(1), p. e1005758.

Richard, P. and Manley, J. L. (2017) 'R Loops and Links to Human Disease', *Journal of Molecular Biology*, 429(21), pp. 3168–3180.

Rico, E. *et al.* (2013) 'Bloodstream form pre-adaptation to the tsetse fly in Trypanosoma brucei', *Frontiers in Cellular and Infection Microbiology*. Frontiers, 3, p. 78.

Rippe, K. and Luke, B. (2015) 'TERRA and the state of the telomere', *Nature Structural & Molecular Biology*. Nature Publishing Group, 22(11), pp. 853–858.

Robinson, D. R. and Gull, K. (1991) 'Basal body movements as a mechanism for mitochondrial genome segregation in the trypanosome cell cycle', *Nature*, 352(6337), pp. 731–733.

Robinson, N. P. *et al.* (2002) 'Inactivation of Mre11 does not affect VSG gene duplication mediated by homologous recombination in Trypanosoma brucei.', *The Journal of biological chemistry*, 277(29), pp. 26185–93.

Rocha, G. *et al.* (2004) 'Possible cases of sexual and congenital transmission of sleeping sickness', *The Lancet*, 363(9404), p. 247.

Roth, C. *et al.* (1989) 'Active late-appearing variable surface antigen genes in Trypanosoma equiperdum are constructed entirely from pseudogenes.', *Proceedings of the National Academy of Sciences of the United States of America*, 86(23), pp. 9375–9.

Rout, M. P. and Field, M. C. (2001) 'Isolation and characterization of subnuclear compartments from Trypanosoma brucei. Identification of a major repetitive nuclear lamina component.', *The Journal of biological chemistry*, 276(41), pp. 38261–71.

Roy, D. *et al.* (2010) 'Competition between the RNA transcript and the nontemplate DNA strand during R-loop formation in vitro: a nick can serve as a strong R-loop initiation site.', *Molecular and cellular biology*. American Society for Microbiology,

30(1), pp. 146–59.

Roy, D. and Lieber, M. R. (2009) ‘G clustering is important for the initiation of transcription-induced R-loops in vitro, whereas high G density without clustering is sufficient thereafter.’, *Molecular and cellular biology*, 29(11), pp. 3124–33.

Rudenko, G. *et al.* (1995) ‘A ribosomal DNA promoter replacing the promoter of a telomeric VSG gene expression site can be efficiently switched on and off in T. brucei.’, *Cell*, 83(4), pp. 547–53.

Rudenko, G. *et al.* (1996) ‘Telomere exchange can be an important mechanism of Variant Surface Glycoprotein gene switching in Trypanosoma brucei’, *Molecular and Biochemical Parasitology*, 80(1), pp. 65–75.

Ryan, K. A. and Englund, P. T. (1989) ‘Replication of kinetoplast DNA in Trypanosoma equiperdum. Minicircle H strand fragments which map at specific locations.’, *The Journal of biological chemistry*, 264(2), pp. 823–30.

Santos-Pereira, J. M. and Aguilera, A. (2015a) ‘R loops: new modulators of genome dynamics and function’, *Nature Reviews Genetics*. Nature Publishing Group, 16(10), pp. 583–597.

Santos-Pereira, J. M. and Aguilera, A. (2015b) ‘R loops: new modulators of genome dynamics and function’, *Nature Reviews Genetics*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 16(10), pp. 583–597.

Sanz, L. A. *et al.* (2016) ‘Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals’, *Molecular Cell*, 63(1), pp. 167–178.

Saxowsky, T. T. *et al.* (2003) ‘Trypanosoma brucei has two distinct mitochondrial DNA polymerase beta enzymes.’, *The Journal of biological chemistry*. American Society for Biochemistry and Molecular Biology, 278(49), pp. 49095–101.

Schmitz, K.-M. *et al.* (2010) ‘Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes’, *Genes & Development*, 24(20), pp. 2264–2269.

Schneider, C. A., Rasband, W. S. and Eliceiri, K. W. (2012) ‘NIH Image to ImageJ: 25 years of image analysis’, *Nature Methods*, 9(7), pp. 671–675.

Schroeder, J. W. *et al.* (2015) ‘Ribonucleotides in bacterial DNA.’, *Critical reviews in biochemistry and molecular biology*. NIH Public Access, 50(3), pp. 181–93.

Schulz, D. *et al.* (2016) ‘Base J and H3.V Regulate Transcriptional Termination in Trypanosoma brucei.’, *PLoS genetics*, 12(1), p. e1005762.

Schwartz, S. and Ast, G. (2010) ‘Chromatin density and splicing destiny: on the cross-talk between chromatin structure and splicing.’, *The EMBO journal*. EMBO Press, 29(10), pp. 1629–36.

Schwede, A. *et al.* (2011) ‘The VSG C-terminal domain is inaccessible to antibodies on live trypanosomes.’, *Molecular and biochemical parasitology*, 175(2), pp. 201–4.

Sequeira-Mendes, J. *et al.* (2009) ‘Transcription initiation activity sets replication origin efficiency in mammalian cells.’, *PLoS genetics*, 5(4), p. e1000446.

Shah, J. S. *et al.* (1987) ‘The 5’ flanking sequence of a Trypanosoma brucei variable

surface glycoprotein gene.', *Molecular and biochemical parasitology*, 24(2), pp. 163–74.

Shannon, P. *et al.* (2003) 'Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks', *Genome Research*, 13(11), pp. 2498–2504.

Sheader, K. *et al.* (2005) 'Variant surface glycoprotein RNA interference triggers a precytokinesis cell cycle arrest in African trypanosomes.', *Proceedings of the National Academy of Sciences of the United States of America*, 102(24), pp. 8716–21.

Shuda, K. *et al.* (2009) 'Aurora kinase B modulates chromosome alignment in mouse oocytes', *Molecular Reproduction and Development*, 76(11), pp. 1094–1105.

Siegel, T. N. *et al.* (2009) 'Four histone variants mark the boundaries of polycistronic transcription units in Trypanosoma brucei.', *Genes & development*, 23(9), pp. 1063–76.

Siegel, T. N. *et al.* (2010) 'Genome-wide analysis of mRNA abundance in two life-cycle stages of Trypanosoma brucei and identification of splicing and polyadenylation sites', *Nucleic Acids Research*, 38(15), pp. 4946–4957.

Siegel, T. N. *et al.* (2011) 'Gene expression in Trypanosoma brucei: lessons from high-throughput RNA sequencing.', *Trends in parasitology*, 27(10), pp. 434–41.

Simarro, P. P. *et al.* (2010) 'The Atlas of human African trypanosomiasis: a contribution to global mapping of neglected tropical diseases', *International Journal of Health Geographics*. BioMed Central, 9(1), p. 57.

Simarro, P. P. *et al.* (2012) 'Update on field use of the available drugs for the chemotherapy of human African trypanosomiasis.', *Parasitology*, 139(7), pp. 842–6.

Skourti-Stathaki, K., Kamieniarz-Gdula, K. and Proudfoot, N. J. (2014) 'R-loops induce repressive chromatin marks over mammalian gene terminators.', *Nature*. Europe PMC Funders, 516(7531), pp. 436–9.

Skourti-Stathaki, K. and Proudfoot, N. J. (2014) 'A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression.', *Genes & development*. Cold Spring Harbor Laboratory Press, 28(13), pp. 1384–96.

Skourti-Stathaki, K., Proudfoot, N. J. and Gromak, N. (2011) 'Human Senataxin Resolves RNA/DNA Hybrids Formed at Transcriptional Pause Sites to Promote Xrn2-Dependent Termination', *Molecular Cell*. Cell Press, 42(6), pp. 794–805.

Sollier, J. *et al.* (2014) 'Transcription-coupled nucleotide excision repair factors promote R-loop-induced genome instability.', *Molecular cell*. NIH Public Access, 56(6), pp. 777–85.

Sollier, J. and Cimprich, K. A. (2015) 'Breaking bad: R-loops and genome integrity.', *Trends in cell biology*.

Sparks, J. L. *et al.* (2012) 'RNase H2-initiated ribonucleotide excision repair.', *Molecular cell*, 47(6), pp. 980–6.

Spieth, J. *et al.* (1993) 'Operons in C. elegans: Polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions', *Cell*. Cell Press, 73(3), pp. 521–532.

Stanne, T. M. *et al.* (2011) 'TbISWI regulates multiple polymerase I (Pol I)-transcribed loci and is present at Pol II transcription boundaries in Trypanosoma brucei', *Eukaryotic Cell*.

Stanne, T. M. and Rudenko, G. (2010) 'Active VSG Expression Sites in Trypanosoma brucei Are Depleted of Nucleosomes', *Eukaryotic Cell*, 9(1), pp. 136–147.

Stein, H. and Hausen, P. (1969) 'Enzyme from calf thymus degrading the RNA moiety of DNA-RNA Hybrids: effect on DNA-dependent RNA polymerase.', *Science (New York, N.Y.)*, 166(3903), pp. 393–5.

Stirling, P. C. *et al.* (2012) 'R-loop-mediated genome instability in mRNA cleavage and polyadenylation mutants', *Genes & Development*, 26(2), pp. 163–175.

Stork, C. T. *et al.* (2016) 'Co-transcriptional R-loops are the main cause of estrogen-induced DNA damage.', *eLife*. eLife Sciences Publications, Ltd, 5.

Sturm, N. R., Fleischmann, J. and Campbell, D. A. (1998) 'Efficient trans-splicing of mutated spliced leader exons in Leishmania tarentolae.', *The Journal of biological chemistry*, 273(30), pp. 18689–92.

Sun, Q. *et al.* (2013) 'R-loop stabilization represses antisense transcription at the Arabidopsis FLC locus.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 340(6132), pp. 619–21.

Swallow, B. M. (1999) 'Impacts of trypanosomiasis on African agriculture'.

Tadokoro, T. and Kanaya, S. (2009) 'Ribonuclease H: molecular diversities, substrate binding domains, and catalytic mechanism of the prokaryotic enzymes', *FEBS Journal*. Wiley/Blackwell (10.1111), 276(6), pp. 1482–1493.

Tan, K. S. W., Leal, S. T. G. and Cross, G. A. M. (2002) 'Trypanosoma brucei MRE11 is non-essential but influences growth, homologous recombination and DNA double-strand break repair.', *Molecular and biochemical parasitology*, 125(1–2), pp. 11–21.

Tetley, L. *et al.* (1987) 'Onset of expression of the variant surface glycoproteins of Trypanosoma brucei in the tsetse fly studied using immunoelectron microscopy.', *Journal of cell science*, 87 ( Pt 2), pp. 363–72.

Thomas, M., White, R. L. and Davis, R. W. (1976) 'Hybridization of RNA to double-stranded DNA: formation of R-loops.', *Proceedings of the National Academy of Sciences of the United States of America*, 73(7), pp. 2294–8.

Thomas, S. *et al.* (2009) 'Histone acetylations mark origins of polycistronic transcription in Leishmania major.', *BMC genomics*. BioMed Central, 10, p. 152.

Thon, G. *et al.* (1990) 'Trypanosome variable surface glycoproteins: composite genes and order of expression.', *Genes & development*, 4(8), pp. 1374–83.

Thon, G., Baltz, T. and Eisen, H. (1989) 'Antigenic diversity by the recombination of pseudogenes.', *Genes & development*, 3(8), pp. 1247–54.

Tian, M. and Alt, F. W. (2000) 'Transcription-induced cleavage of immunoglobulin switch regions by nucleotide excision repair nucleases in vitro.', *The Journal of biological chemistry*. American Society for Biochemistry and Molecular Biology, 275(31), pp. 24163–72.

Tiengwe, C. *et al.* (2012) 'Genome-wide analysis reveals extensive functional

interaction between DNA replication initiation and transcription in the genome of Trypanosoma brucei.', *Cell reports*, 2(1), pp. 185–97.

Tiwari, K. and Dubey, V. K. (2018) 'Fresh insights into the pyrimidine metabolism in the trypanosomatids', *Parasites & Vectors*. BioMed Central, 11(1), p. 87.

Tomson, B. N. and Arndt, K. M. (2013) 'The many roles of the conserved eukaryotic Paf1 complex in regulating transcription, histone modifications, and disease states', *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1829(1), pp. 116–126.

Torreele, E. *et al.* (2010) 'Fexinidazole – A New Oral Nitroimidazole Drug Candidate Entering Clinical Development for the Treatment of Sleeping Sickness', *PLoS Neglected Tropical Diseases*. Edited by M. Boelaert. Public Library of Science, 4(12), p. e923.

Tschudi, C. *et al.* (2012) 'Small interfering RNA-producing loci in the ancient parasitic eukaryote Trypanosoma brucei', *BMC Genomics*, 13(1), p. 427.

Tuduri, S. *et al.* (2009) 'Topoisomerase I suppresses genomic instability by preventing interference between replication and transcription.', *Nature cell biology*, 11(11), pp. 1315–24.

Turchi, J. J. *et al.* (1994) 'Enzymatic completion of mammalian lagging-strand DNA replication.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 91(21), pp. 9803–7.

Turner, C. M. *et al.* (1988) 'An estimate of the size of the metacyclic variable antigen repertoire of Trypanosoma brucei rhodesiense.', *Parasitology*, 97 ( Pt 2), pp. 269–76.

Turner, C. M., Aslam, N. and Dye, C. (1995) 'Replication, differentiation, growth and the virulence of Trypanosoma brucei infections.', *Parasitology*, 111 ( Pt 3), pp. 289–300.

Ulbert, S., Chaves, I. and Borst, P. (2002) 'Expression site activation in Trypanosoma brucei with three marked variant surface glycoprotein gene expression sites.', *Molecular and biochemical parasitology*, 120(2), pp. 225–35.

Vanhamme, L. *et al.* (2000) 'Differential RNA elongation controls the variant surface glycoprotein gene expression sites of Trypanosoma brucei.', *Molecular microbiology*, 36(2), pp. 328–40.

Vanhollebeke, B. and Pays, E. (2010) 'The trypanolytic factor of human serum: many ways to enter the parasite, a single way to kill', *Molecular Microbiology*. Blackwell Publishing Ltd, 76(4), pp. 806–814.

Vashee, S. *et al.* (2003) 'Sequence-independent DNA binding and replication initiation by the human origin recognition complex.', *Genes & development*. Cold Spring Harbor Laboratory Press, 17(15), pp. 1894–908.

Vassella, E. *et al.* (1997) 'Differentiation of African trypanosomes is controlled by a density sensing mechanism which signals cell cycle arrest via the cAMP pathway.', *Journal of cell science*, 110 ( Pt 2, pp. 2661–71.

Vink, C., Rudenko, G. and Seifert, H. S. (2012) 'Microbial antigenic variation mediated by homologous DNA recombination.', *FEMS microbiology reviews*. The

Oxford University Press, 36(5), pp. 917–48.

Wahba, L. *et al*. (2011) 'RNase H and Multiple RNA Biogenesis Factors Cooperate to Prevent RNA:DNA Hybrids from Generating Genome Instability', *Molecular Cell*, 44(6), pp. 978–988.

Wahba, L. *et al*. (2016) 'S1-DRIP-seq identifies high expression and polyA tracts as major contributors to R-loop formation.', *Genes & development*. Cold Spring Harbor Laboratory Press, 30(11), pp. 1327–38.

Wahba, L., Gore, S. K. and Koshland, D. (2013) 'The homologous recombination machinery modulates the formation of RNA–DNA hybrids and associated chromosome instability', *eLife*. eLife Sciences Publications Limited, 2, p. e00505.

Wahba, L. and Koshland, D. (2013) 'The Rs of Biology: R-Loops and the Regulation of Regulators', *Molecular Cell*. Cell Press, 50(5), pp. 611–612.

Wedel, C. *et al*. (2017) 'GT-rich promoters can drive RNA pol II transcription and deposition of H2A.Z in African trypanosomes', *The EMBO Journal*, 36(17), p. e201695323.

Wellinger, R. E., Prado, F. and Aguilera, A. (2006) 'Replication fork progression is impaired by transcription in hyperrecombinant yeast cells lacking a functional THO complex.', *Molecular and cellular biology*. American Society for Microbiology, 26(8), pp. 3327–34.

Wheeler, R. J. *et al*. (2013) 'Cytokinesis in Trypanosoma brucei differs between bloodstream and tsetse trypomastigote forms: implications for microtubule-based morphogenesis and mutant analysis.', *Molecular microbiology*, 90(6), pp. 1339–55.

Wheeler, R. J., Gluenz, E. and Gull, K. (2013) 'The Limits on Trypanosomatid Morphological Diversity', *PLoS ONE*. Edited by Z. Li. Public Library of Science, 8(11), p. e79581.

White, R. L. and Hogness, D. S. (1977) 'R loop mapping of the 18S and 28S sequences in the long and short repeating units of Drosophila melanogaster rDNA.', *Cell*, 10(2), pp. 177–92.

Wickstead, B., Ersfeld, K. and Gull, K. (2004) 'The small chromosomes of Trypanosoma brucei involved in antigenic variation are constructed around repetitive palindromes.', *Genome research*, 14(6), pp. 1014–24.

Williams, J. S., Gehle, D. B. and Kunkel, T. A. (2017) 'The role of RNase H2 in processing ribonucleotides incorporated during DNA replication', *DNA Repair*, 53, pp. 52–58.

Wilusz, C. J., Wormington, M. and Peltz, S. W. (2001) 'The cap-to-tail guide to mRNA turnover', *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, 2(4), pp. 237–246.

Wimberly, H. *et al*. (2013) 'R-loops and nicks initiate DNA breakage and genome instability in non-growing Escherichia coli.', *Nature communications*, 4, p. 2115.

Wongsurawat, T. *et al*. (2012) 'Quantitative model of R-loop forming structures reveals a novel level of RNA-DNA interactome complexity.', *Nucleic acids research*. Oxford University Press, 40(2), p. e16.

Woodward, R. and Gull, K. (1990) 'Timing of nuclear and kinetoplast DNA

replication and early morphological events in the cell cycle of Trypanosoma brucei.', *Journal of cell science*, 95 ( Pt 1), pp. 49–57.

Wright, J. R., Siegel, T. N. and Cross, G. A. M. (2010) 'Histone H3 trimethylated at lysine 4 is enriched at probable transcription start sites in Trypanosoma brucei.', *Molecular and biochemical parasitology*, 172(2), pp. 141–4.

Wu, L. and Hickson, I. D. (2003) 'The Bloom's syndrome helicase suppresses crossing over during homologous recombination', *Nature*, 426(6968), pp. 870–874.

Xu, B. and Clayton, D. A. (1996) 'RNA-DNA hybrid formation at the human mitochondrial heavy-strand origin ceases at replication start sites: an implication for RNA-DNA hybrids serving as primers.', *The EMBO journal*. European Molecular Biology Organization, 15(12), pp. 3135–43.

Xu, W. *et al.* (2017) 'The R-loop is a common chromatin feature of the Arabidopsis genome', *Nature Plants*. Nature Publishing Group, 3(9), pp. 704–714.

Yang, X. *et al.* (2009) 'RAP1 is essential for silencing telomeric variant surface glycoprotein genes in Trypanosoma brucei.', *Cell*, 137(1), pp. 99–109.

Yu, K. *et al.* (2003) 'R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells', *Nature Immunology*. Nature Publishing Group, 4(5), pp. 442–451.

Yu, T.-Y., Kao, Y. and Lin, J.-J. (2014) 'Telomeric transcripts stimulate telomere recombination to suppress senescence in cells lacking telomerase.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 111(9), pp. 3377–82.

Zang, C. *et al.* (2009) 'A clustering approach for identification of enriched domains from histone modification ChIP-Seq data', *Bioinformatics*, 25(15), pp. 1952–1958.

Zhang, Y. *et al.* (2008) 'Model-based Analysis of ChIP-Seq (MACS)', *Genome Biology*, 9(9), p. R137.

Zheng, L.-L. *et al.* (2013) 'Comparative transcriptome analysis of small noncoding RNAs in different stages of Trypanosoma brucei.', *RNA (New York, N.Y.)*. Cold Spring Harbor Laboratory Press, 19(7), pp. 863–75.

Zimmer, A. D. and Koshland, D. (2016) 'Differential roles of the RNases H in preventing chromosome instability.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 113(43), pp. 12220–12225.

Zomerdijk, J. C. *et al.* (1990) 'The promoter for a variant surface glycoprotein gene expression site in Trypanosoma brucei.', *The EMBO journal*, 9(9), pp. 2791–801.