This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.
Using Locus-specific Proteomics to Investigate Heterochromatin in *Trypanosoma brucei*

Tadhg Devlin

Thesis presented for the degree of

Doctor of Philosophy

University of Edinburgh

2022
Declaration

I declare that the content of this thesis is my own work and that all contributions and collaborations have been explicitly acknowledged in the text. No material presented in thesis has been submitted to any other university or for any other degree.

Tadhg Devlin
Abstract

The eukaryotic nucleus is classically divided into two broad categories: gene-poor heterochromatin and gene-rich euchromatin. In most model eukaryotes, heterochromatin is epigenetically defined by histone H3 lysine 9 methylation (H3K9me), and is rich in repetitive DNA sequences. *Trypanosoma brucei* is a kinetoplastid parasite that branched early in eukaryotic evolution, and is the causative agent of African sleeping sickness. Trypanosome histone proteins are divergent, and classical heterochromatin histone post-translational modifications (PTMs) such as H3K9me are absent. As a result, the proteins and histone PTMs that define heterochromatin in *T. brucei* are unknown. Using transcription activator-like effector (TALE) DNA-binding proteins, we have developed a system to purify proteins associated with repetitive DNA sequences, which are candidate heterochromatin regions. We designed TALEs to bind telomeres, centromeres, transposable elements, and the approximately 100 transcriptionally silent minichromosomes, as these repetitive sequences are likely to form heterochromatin. In combination with label-free quantitative mass spectrometry, we hoped this approach would identify previously unknown heterochromatin-associated proteins. Results obtained with telomeres provided proof-of-principle that this was a viable strategy for purifying chromatin-associated proteins in *T. brucei*, identifying both known and novel telomere-interacting proteins. This included the discovery of a potential role for two zinc-finger proteins at telomeres, in addition to their previously characterised role in post-transcriptional gene regulation. Expansion of the technology to minichromosomes identified that these chromosomes may assemble a kinetochore, in contrast to the prevailing model in the literature. Overall, this work shows that locus-specific proteomics can be a useful tool for investigating trypanosomatid chromatin biology, and could be applied to other genomic loci of interest in future.
Lay summary

The word ‘genome’ refers to all of the genetic information, or DNA, inside a cell. In multicellular organisms like humans, every cell has the same genome, despite the fact that many of our cells look and behave very differently. For example, the cells that make up your skin are very different to the cells that make up your brain. If they all have the same genetic information, how is it that they end up being so different? This is achieved by only using certain parts of that information in certain cells. So if the genome was a textbook made of 20 different chapters, to make a brain you might use the information from chapters 1, 3, and 11, but to make skin you might use chapters 2, 5, and 15. In a brain cell, the information found in chapters other than 1, 3, and 11 is not needed, and so it needs to be silenced. Genes (information) are silenced by being physically compacted. This is achieved by wrapping the DNA around proteins called histones, similar to wrapping thread around a spool. These histone spools can be stacked on top of one another to compact the DNA even further. This compact, wrapped up structure of DNA and proteins is called heterochromatin. Inside cells, there are special proteins which wrap up unwanted DNA so it can be silenced. These proteins for making heterochromatin are shared between lots of different organisms, including between humans and yeast. However, the proteins that make heterochromatin in the single celled parasite Trypanosoma brucei are completely unknown. They do not have any of the proteins that make heterochromatin in other organisms, but we know they do it because we can see it when we look down the microscope. Therefore, there must be some proteins involved in making heterochromatin in this parasite. The aim of my project was to find out what they are.

There are certain DNA sequences that do not contain any genes but do contain lots of repetitive ‘junk’ DNA. In other organisms, these sequences are usually packaged into heterochromatin. I developed a system to purify these sequences and any proteins stuck to them from T. brucei cells, with the hope of identifying proteins that caused them to compact into heterochromatin. The system worked well on some sequences and not well on others, but overall we found some interesting proteins at these sequences that we had not expected to find. When I investigated two of those proteins further, I found that they seemed to have two very different jobs inside cells – one to do with a certain repetitive sequence and one to do with something completely different.
Acknowledgements

The biggest thank you of all goes to Dr Tania Auchynnikava, without whom this thesis would not exist. She provided training, supervision, and mentorship in an environment where all were sorely lacking.

Next in line is Dr Desislava Staneva, who provided technical help both inside and outside the lab. She is the only other person who will really understand.

A huge thank you goes to Julie Young for preparing many gallons of HMI-9 for me. Without her, none of these experiments would have been possible.

Thank you to Professor Keith Matthews for ensuring I got over the line, and for his help in securing a PIPS placement.

Thank you to members of the Matthews and Allshire labs both past and present for their input along the way. Special mentions go to Sharon White, Balazs Szoor, and Mathieu Cayla.

I would like to thank my family and friends for their support, which got me through the dark times. Special mentions go to Lizi Hegarty, Alex McDonnell, Alex Morgan, Helen Feord, Ben Craske, Paul Reddin (RIP) and Oisín Moran – I could not have done it without you.

Finally, I would like to thank Aisling McGowan, whose love gave me a reason to keep going.
Never forget how you were treated, and never let it happen again.
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>5mC</td>
<td>5-methylcytosine</td>
</tr>
<tr>
<td>ABC</td>
<td>Ammonium bicarbonate</td>
</tr>
<tr>
<td>ACN</td>
<td>Acetonitrile</td>
</tr>
<tr>
<td>Ago1</td>
<td>Protein argonaute 1</td>
</tr>
<tr>
<td>BES</td>
<td>Bloodstream expression site</td>
</tr>
<tr>
<td>BSF</td>
<td>Bloodstream form</td>
</tr>
<tr>
<td>bp</td>
<td>Base pair</td>
</tr>
<tr>
<td>BSA</td>
<td>Bovine serum albumin</td>
</tr>
<tr>
<td>CDS</td>
<td>Coding sequence</td>
</tr>
<tr>
<td>CENP-A</td>
<td>Centromere protein A</td>
</tr>
<tr>
<td>ChIP</td>
<td>Chromatin immunoprecipitation</td>
</tr>
<tr>
<td>CIR147</td>
<td>Chromosome internal repeats 147</td>
</tr>
<tr>
<td>CRISPR</td>
<td>Clustered regularly interspaced short palindromic repeats</td>
</tr>
<tr>
<td>DABCO</td>
<td>1, 4-diazabicyclo[2.2.2]octane</td>
</tr>
<tr>
<td>DAPI</td>
<td>4′,6-diamidino-2-phenyindole</td>
</tr>
<tr>
<td>DMP</td>
<td>Dimethyl pimelimidate</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>DTT</td>
<td>Dithiothreitol</td>
</tr>
<tr>
<td>ECL</td>
<td>Enhanced chemiluminescence</td>
</tr>
<tr>
<td>EDTA</td>
<td>Ethylenediaminetetraacetic acid</td>
</tr>
<tr>
<td>EGTA</td>
<td>Egtazic acid</td>
</tr>
<tr>
<td>ES</td>
<td>Expression site</td>
</tr>
<tr>
<td>ESAG</td>
<td>Expression site-associated gene</td>
</tr>
<tr>
<td>ESB</td>
<td>Expression site body</td>
</tr>
<tr>
<td>FACS</td>
<td>Fluorescence-activated cell sorting</td>
</tr>
<tr>
<td>FASP</td>
<td>Filtered-aided sample preparation</td>
</tr>
<tr>
<td>FACT</td>
<td>Facilitates chromatin transcription</td>
</tr>
<tr>
<td>G1</td>
<td>Gap 1 phase</td>
</tr>
<tr>
<td>G2</td>
<td>Gap 2 phase</td>
</tr>
<tr>
<td>GFP</td>
<td>Green fluorescent protein</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>H3K9me</td>
<td>Histone H3 lysine 9 methylation</td>
</tr>
<tr>
<td>HAT</td>
<td>Histone acetyltransferase or Human African trypanosomiasis</td>
</tr>
<tr>
<td>HDAC</td>
<td>Histone deacetylase</td>
</tr>
<tr>
<td>HP1</td>
<td>Heterochromatin protein 1</td>
</tr>
<tr>
<td>HRP</td>
<td>Horseradish peroxidase</td>
</tr>
<tr>
<td>IAA</td>
<td>Iodoacetamide</td>
</tr>
<tr>
<td>IP</td>
<td>Immunoprecipitation</td>
</tr>
<tr>
<td>JBP1</td>
<td>J binding protein 1</td>
</tr>
<tr>
<td>JBP2</td>
<td>J biosynthesis protein 2</td>
</tr>
<tr>
<td>kb</td>
<td>Kilobase</td>
</tr>
<tr>
<td>kDa</td>
<td>Kilodalton</td>
</tr>
<tr>
<td>kDNA</td>
<td>Kinetoplast DNA</td>
</tr>
<tr>
<td>KKT</td>
<td>Kinetoplastid kinetochore protein</td>
</tr>
<tr>
<td>LC-MS/MS</td>
<td>Liquid chromatography with tandem mass spectrometry</td>
</tr>
<tr>
<td>MES</td>
<td>Metacyclic expression site or 2-(N-morpholino)ethanesulfonic acid</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td>PAD</td>
<td>Protein associated with differentiation</td>
</tr>
<tr>
<td>PAGE</td>
<td>Polyacrylamide gel electrophoresis</td>
</tr>
<tr>
<td>PBS</td>
<td>Phosphate-buffered saline</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>PCF</td>
<td>Procyclic form</td>
</tr>
<tr>
<td>PTM</td>
<td>Post-translational modification</td>
</tr>
<tr>
<td>PTU</td>
<td>Polycistronic transcription unit</td>
</tr>
<tr>
<td>RHS</td>
<td>Retrotransposon hot spot</td>
</tr>
<tr>
<td>RBP</td>
<td>RNA binding protein</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>RNAi</td>
<td>RNA interference</td>
</tr>
<tr>
<td>RNAP</td>
<td>RNA polymerase</td>
</tr>
<tr>
<td>S</td>
<td>Synthesis phase</td>
</tr>
<tr>
<td>SDS</td>
<td>Sodium dodecyl sulfate</td>
</tr>
<tr>
<td>SET</td>
<td>Suppressor of variegation 3-9, enhancer of zeste and trithorax</td>
</tr>
<tr>
<td>SIF</td>
<td>Stumpy induction factor</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>siRNA</td>
<td>Small interfering RNA</td>
</tr>
<tr>
<td>SL</td>
<td>Spliced leader</td>
</tr>
<tr>
<td>SLACS</td>
<td>Spliced leader-associated conserved sequence</td>
</tr>
<tr>
<td>SSR</td>
<td>Strand-switch region</td>
</tr>
<tr>
<td>TAE</td>
<td>Tris-acetate-EDTA</td>
</tr>
<tr>
<td>TBE</td>
<td>Tris-borate-EDTA</td>
</tr>
<tr>
<td>TE</td>
<td>Tris-EDTA</td>
</tr>
<tr>
<td>TFA</td>
<td>Trifluoroacetic acid</td>
</tr>
<tr>
<td>TLF</td>
<td>Trypanosome lytic factor</td>
</tr>
<tr>
<td>TSR</td>
<td>Transcription start region</td>
</tr>
<tr>
<td>TRITC</td>
<td>Tetramethylrhodamine</td>
</tr>
<tr>
<td>TTR</td>
<td>Transcription termination region</td>
</tr>
<tr>
<td>UTR</td>
<td>Untranslated region</td>
</tr>
<tr>
<td>VSG</td>
<td>Variant surface glycoprotein</td>
</tr>
<tr>
<td>WT</td>
<td>Wild type</td>
</tr>
<tr>
<td>YFP</td>
<td>Yellow fluorescent protein</td>
</tr>
</tbody>
</table>
Table of contents

Title page I
Declaration of originality II
Abstract III
Lay summary IV
Acknowledgements V
Abbreviations VII
Table of contents X

Chapter 1: Introduction

1.1 Trypanosomes 1
   1.1.2 Trypanosoma brucei 2

1.2 Genome Organisation in Trypanosoma brucei 3
   1.2.1 Variant surface glycoprotein genes 4
   1.2.2 Telomeres 5
   1.2.3 Centromeres 5
   1.2.4 Transposable Elements 6

1.3 Chromatin 8
   1.3.1 Histones and histone variants in T. brucei 9
   1.3.2 DNA modifications in T. brucei 11

1.4 Heterochromatin in T. brucei 13
   1.4.1 Candidate heterochromatin regions in the T. brucei genome 14
      1.4.1.1 Telomeres 15
      1.4.1.2 Centromeres 15
      1.4.1.3 Transposable elements 15
      1.4.1.4 Minichromosomes 16
      1.4.1.5 VSG and procyclin loci 16
1.4.2 Histones and histone variants 17
1.4.3 Chromatin Remodellers and histone chaperones 17
1.4.4 Writers, Readers, and Erasers 19
1.4.5 RNA interference 19

1.5 Tools for the exploration of trypanosome heterochromatin:
Transcription activator-like effector DNA binding proteins 21

1.6 Locus-specific proteomics 22
1.6.1 Locus-specific proteomics in Trypanosoma brucei 26

1.7 Project aims 27

Chapter 2: Materials and methods 28
2.1 Molecular cloning and standard DNA methods 28
2.1.1 Small scale plasmid preparations 28
2.1.2 Ethanol precipitation of DNA 28
2.1.3 Polymerase Chain Reaction (PCR) 28
2.1.4 DNA Agarose gel electrophoresis 29
2.1.5 DNA gel extraction 29
2.1.6 Plasmid restriction digests 29
2.1.7 Bacterial transformations 29
2.1.8 Colony PCR 30
2.1.9 Gibson Assembly 30
2.1.10 Sanger sequencing 30
2.1.11 TALEs target sequence design 31

2.2 Standard Trypanosome methods 31
2.2.1 Trypanosome cell culture 31
2.2.2 Trypanosome genomic DNA extraction 32
2.2.3 Trypanosome protein extraction for Western blotting 32
2.2.4 Trypanosome transfections 33
2.2.5 N-terminal protein tagging 33
2.3 Standard protein methods 34
2.3.1 Western blotting 34
2.3.2 Silver Staining 34
2.4 Immunofluorescence microscopy 34
2.5 Chromatin Immunoprecipitation (ChIP) methods 35
2.5.1 Chromatin Immunoprecipitation (ChIP) 35
2.5.2 Quantitative PCR (qPCR) 37
2.6 Immunoprecipitation followed by Mass Spectrometry Analysis 38
2.6.1 Crosslinking of antibody to dynabeads 38
2.6.2 Protein immunoprecipitation from whole cell lysates 38
2.6.3 FASP Digestion 39
2.6.4 Stage Tip Preparation 39
2.6.5 Mass spectrometry data analysis – raw data 39
2.7 Southern blotting 40

Chapter 3: Design, Cloning, and Characterisation of Transcription Activator-like Effectors (TALEs) Directed to Bind Repetitive Regions of the T. brucei Genome 42
3.1 Introduction 42
3.2 Selection of TALEs target sequences for binding different repetitive regions of the T. brucei genome 42
3.3 Cloning of the TALE expression vector using multiple cloning methodologies 43
3.3.1 Using annealed oligonucleotide cloning to insert the La protein NLS into the CATpHD449 Δ1-354 PAD1 3′UTR plasmid 45
3.3.2 Cloning the TALE domain using the Musunuru/Cowan Lab TALEN kit 47
3.3.3 Cloning the trypanosome TALE expression vector using Gibson Assembly 49
3.4 Western blot analysis detects TALE protein expression

3.5 Sub-cellular localisation of TALE proteins by immunofluorescence microscopy

   3.5.1 Immunofluorescence of Telomere-TALE shows a punctate distribution similar to that reported for telomere proteins

   3.5.2 Immunolocalisation of Centromere-TALE shows localisation in two primary puncta

   3.5.3 Immunofluorescence microscopy of Ingi-TALE detects a punctate distribution pattern throughout the nucleus

   3.5.4 Minichromosome-TALE localisation is similar to that reported for minichromosomes

3.6 Cells expressing the exogenous TALE proteins have a normal growth rate

3.7 Chromatin immunoprecipitation (ChIP)-qPCR with TALE proteins shows variable enrichment at target loci

3.8 Telomere-TALE expression causes decrease in telomere length

3.9 Discussion

   3.9.1 Western analysis shows higher TALE expression levels than expected

   3.9.2 Immunofluorescence shows distinct localisations for all TALE proteins

   3.9.3 TALE expression does not interfere with cell growth

   3.9.4 ChIP-qPCR shows variation in target enrichment for the different TALE proteins

   3.9.5 Telomere length may be decreased in Telomere-TALE cells

   3.9.6 Conclusions

Chapter 4: Locus-specific proteomics using TALEs

4.1 Introduction

4.2 Proteomic identification of telomere-associated proteins

   4.2.1 TRF - a positive control for telomere-binding proteins
4.2.2 Telomere-TALE identifies known and novel telomere-binding proteins 80

4.3 Proteomic identification of centromere-associated proteins 82
   4.3.1 KKT2 associates with kinetochore, mitotic, and chromatin proteins 82
   4.3.2 The Centromere-TALE failed to significantly enrich any proteins 88

4.4 Proteomic investigation of the TALE protein designed to bind Ingi transposable elements 90
   4.4.1 The Ingi-TALE failed to significantly enrich any chromatin proteins 90

4.5 Proteomic identification of proteins associated with the minichromosome 177 bp repeat 92
   4.5.1 Minichromosome-TALE enriches for known centromeric proteins 92
   4.5.2 KKT2 associates with both minichromosomal 177 bp repeats and CIR147 centromere repeats 95

4.6 Discussion 96
   4.6.1 Affinity purification of endogenous telomere-binding protein TRF 96
   4.6.2 Affinity purification of synthetic telomere-binding protein Telomere-TALE 97
   4.6.3 Affinity purification of endogenous kinetochore-centromere protein KKT2 98
   4.6.4 Affinity purification of synthetic centromere protein Centromere-TALE 99
   4.6.5 Affinity purification of synthetic Ingi transposable element-binding protein Ingi-TALE 100
   4.6.6 Affinity purification of synthetic minichromosomal 177 bp repeat-binding protein Minichromosome-TALE 101
   4.6.7 Conclusions 104

Chapter 5: Characterisation of the role of ZC3H39 and ZC3H40 at Telomeres 105
5.1 Introduction 105
5.2 YFP-tagged ZC3H39/40 localizes to the cytoplasm in bloodstream form parasites 106
5.3 Chromatin Immunoprecipitation of ZC3H39/40 shows no enrichment on telomeric repeats

5.4 ZC3H39 associates with ZC3H40 and RNA editing components

5.5 ZC3H40 and ZC3H39 associate with the same core proteins

5.6 YFP-tagged UMSBP2 exhibits an exclusively nuclear localisation

5.7 UMSBP2 associates with the same core proteins as ZC3H39/40 as well as multiple chromatin factors

5.8 Investigation of the impact of ZC3H39 and ZC3H40 knockout on telomere protein composition

5.8.1 Immunoprecipitation of TRF from wild type cells provides a control for telomere protein composition

5.8.2 Immunoprecipitation of TRF in $\Delta$ZC3H39$\Delta$ZC3H40 cells shows a potential decrease in chromatin association

5.9 Telomere length in $\Delta$ZC3H39$\Delta$ZC3H40 cells does not change substantially relative to wild type

5.10 Discussion

5.10.1 ZC3H39 and ZC3H40 immunofluorescence show localisation consistent with literature

5.10.2 ChIP-qPCR of ZC3H39 and ZC3H40 shows no enrichment of either protein on telomeric TTAGGG repeats

5.10.3 ZC3H39 associates with multiple RNA editing proteins

5.10.4 ZC3H40 appears to form a complex or complexes with ZC3H39 and RNA editing factors

5.10.5 Immunoprecipitation of UMSBP2 shows distinct and overlapping interactome compared with ZC3H39 and ZC3H40

5.10.6 Does loss of ZC3H39 and ZC3H40 alter telomere protein
composition? 137

5.10.6.1 TRF shows weak interaction with some proteins that associate with ZC3H39/49, in addition to known telomeric proteins 137

5.10.6.2 TRF immunoprecipitation in ΔZC3H39ΔZC3H40 cells shows decreased enrichment of chromatin factors 137

5.10.7 Telomere length appears unchanged in ΔZC3H39ΔZC3H40 cells 138

Chapter 6: Discussion 140

Appendices 145

Appendix A 145

Primers used in this study

Appendix B 147

Chemical solutions

(i) Miscellaneous 147

(ii) ChIP buffers 148

References 150
1. Introduction

1.1 Trypanosomes

Trypanosomes are unicellular flagellated kinetoplastid parasites that belong to the eukaryotic supergroup Excavata, and are evolutionarily distant from the main eukaryotic lineage of Opisthokonta (Akiyoshi & Gull, 2013). Kinetoplastids such as *Trypanosoma brucei*, *Trypanosoma cruzi*, and *Leishmania spp.* represent a large disease burden in the developing world. For example, *T. brucei* is the causative agent of Human African Trypanosomiasis (HAT) and the wasting disease ‘nagana’ in livestock. HAT is caused by the two subspecies *Trypanosoma brucei rhodesiense* and *Trypanosoma brucei gambiense*, which show differences in their geographical distribution throughout sub-Saharan Africa (Figure 1.1). The subspecies *Trypanosoma brucei brucei* does not cause human disease owing to the presence of a trypanosome lytic factor in human serum, which provides innate immunity to certain kinetoplastid parasites (Capewell et al., 2015; Samanovic et al., 2009).

Figure 1.1. The geographical distribution of human infections with different *Trypanosoma brucei* subspecies. *Trypanosoma brucei rhodesiense* infections (blue) are seen in eastern and southern Africa, whereas *Trypanosoma brucei gambiense* infections (red) are seen in western and central Africa. Figure from Büscher et al. (Büscher et al., 2017).
1.1.2 *Trypanosoma brucei*

*T. brucei* is transmitted by the bite of the tsetse fly (*Glossina*), and undergoes a complex lifecycle between the insect vector and mammalian host (Figure 1.2). The different stages of the lifecycle are associated with changes in cellular morphology, metabolism, and gene expression (Fenn & Matthews, 2007). Some of these changes are accompanied by alterations in nuclear architecture and chromatin organisation, the evidence for which will be discussed later. The two most well characterised lifecycle forms are the procyclic form and slender bloodstream form, and so essentially all of the literature cited herein will relate to these two forms. In addition, the subspecies *Trypanosoma brucei brucei* is used in almost all cases, and use of any other subspecies will be specified in the text.

![Figure 1.2. The lifecycle of *Trypanosoma brucei*. Slender forms proliferate rapidly in the host bloodstream, differentiating to cell-cycle arrested stumpy forms upon accumulation of a density-dependent signal known as stumpy-inducing factor (SIF). After a Tsetse fly blood meal, stumpy forms differentiate into procyclic forms in the Tsetse midgut. Procyclic forms migrate to the salivary glands via the proventriculus, differentiating to epimastigotes. These cells go on to form metacyclic forms, which are mammalian infective. Metacyclic forms are transmitted to a new mammalian host via the bite of the Tsetse, where they differentiate to slender forms, and the cycle repeats. Figure from Silvester et al. (Silvester et al., 2017).
1.2. Genome Organisation in *Trypanosoma brucei*

African trypanosomes differ from other eukaryotes in several aspects of their gene expression and genome organisation. The nuclear genome of *T. brucei* is comprised of 11 pairs of megabase chromosomes, 1-5 intermediate chromosomes, and approximately 100 minichromosomes (El-Sayed et al., 2000). Most housekeeping genes are found on the megabase chromosomes where they are organised into directional gene clusters that are transcribed as polycistronic transcription units by RNA polymerase II (RNAPII) (Imboden et al., 1987; Muhich & Boothroyd, 1988). Directional gene clusters are separated by either convergent or divergent strand switch regions. Monocistronic RNAs are generated by trans-splicing of a capped spliced leader (SL) RNA, or so-called mini exon, to the 5’ UTR and polyadenylation of the 3’ UTR to make monocistronic RNAs (C. E. Clayton, 2016; Liang et al., 2003).

The mitochondrial genome is comprised of a network of circular DNAs known as mini- and maxicircles, which form the kinetoplast (Englund et al., 1982). The approximately 50 maxicircles contain genes encoding electron transport chain components and other mitochondrial proteins (Aphasizheva et al., 2014). The thousands of minicircles contain guide RNA (gRNA) genes. These gRNA genes are required because the majority of the maxicircle genes are encrypted, and require insertion and/or deletion of uridines in order to become open reading frames (Benne et al., 1986). This RNA editing requires a host of specialized proteins, some of which are discussed in Chapter 5.
1.2.1 Variant surface glycoprotein genes

Transcription of some protein coding genes by RNA polymerase I (RNAPI) is another unusual feature of gene expression in African trypanosomes. These include procyclin genes, encoding the major surface protein of procyclic forms, and variant surface glycoprotein (VSG) genes, which encode the surface coat proteins that cover bloodstream and metacyclic form T. brucei cells, mediating antigenic variation and immune evasion. The sub-telomeres of the megabase chromosomes and intermediate chromosomes house the approximately 15 bloodstream expression sites (BESs) for the VSG (Hertz-Fowler et al., 2008), as well as the metacyclic expression sites (MESs) which house the metacyclic VSG (mVSG) genes. Interestingly, VSG genes are expressed monoallelically. This requires all but one of the thousands of VSG genes to be silenced, including those in expression sites, and activation of a new VSG gene requires its translocation of the active expression site to a sub-nuclear structure known as the expression site body (ESB) (Navarro & Gull, 2001). This allelic exclusion is mediated by a VSG exclusion (VEX) protein complex comprised of VEX1 and VEX2 (Faria et al., 2019; Glover et al., 2016).
BESs contain extensive repeats both upstream of their promoter (50 bp repeats) and immediately upstream of the VSG gene (70 bp repeats) (Figure 1.3), the latter facilitating gene conversion and replacement of the active VSG with a new VSG from another location by homologous recombination (G. Hovel-Miner et al., 2016). The minichromosomes are believed to act as a reservoir for additional VSG genes, and will be discussed in more detail later (Gull et al., 1998).

1.2.2. Telomeres

Similar to humans, trypanosome telomeres are comprised of tandem (TTAGGG)$_n$ repeats which extend for approximately 15 kb on average (Blackburn & Challoner, 1984; Dreesen & Cross, 2008; G. A. Hovel-Miner et al., 2012a; van der Ploeg et al., 1984). In all eukaryotes, telomere ends are protected from recognition by the DNA damage response machinery by the binding of specialized proteins (de Lange, 2001). In mammals, telomeres are bound by a 6 protein complex known as shelterin, comprised of TRF1, TRF2, POT1, TIN2, TPP1 and RAP1 (de Lange, 2005). In \textit{T. brucei}, proteins with the same core functions as TRF2, RAP1, and TIN2 have been identified (Jehi et al., 2014; B. Li et al., 2005; Yang et al., 2009). Additional telomere-associated proteins identified in \textit{T. brucei} such as telomere-associated protein 1 (TelAP1) (Reis et al., 2018) do not share sequence similarity with the remaining shelterin components.

1.2.3. Centromeres

Centromeres were first identified in \textit{T. brucei} megabase chromosomes using etoposide-mediated topoisomerase-II cleavage (Obado et al., 2007a). Interestingly, no site-specific accumulation of topoisomerase-II was seen on the intermediate chromosomes or the minichromosomes. This contributed to the hypothesis that these chromosomes segregate via a different mechanism to the megabase chromosomes, which is discussed in detail in Section 4.6.6 (Ersfeld & Gull, 1997). \textit{T. brucei} centromeres are 20 to 120 kb in size, approximately similar to those of fission yeast (Echeverry et al., 2012; Steiner et al., 1993). All centromeric regions contain various degenerate retrotransposable elements, and are comprised of AT-rich repeats (Echeverry et al., 2012; Obado et al., 2007a). In addition, they are adjacent to arrays of rRNA genes in several cases. These repeat sequences have been organised into 4 classes.
by Obado et al. (Table 1.1) (Obado et al., 2007a). Enrichment of several kinetoplastid kinetochore (KKT) proteins by ChIP-seq at these loci later confirmed them to be true centromeres (Akiyoshi & Gull, 2014b). In contrast, centromeric sequences in two related trypanosomatids, *T. cruzi* and *Leishmania major*, are GC-rich and non-repetitive (Garcia-Silva et al., 2017; Obado et al., 2005, 2007a).

Table 1.1. Centromeric repeats on *T. brucei* megabase chromosomes can be divided into four classes based on their differing repeat sequences.

<table>
<thead>
<tr>
<th>Class</th>
<th>Chromosome number</th>
<th>Repeat length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class I</td>
<td>4, 5, 8, 9, 10, 11</td>
<td>147</td>
</tr>
<tr>
<td>Class II</td>
<td>2, 7</td>
<td>30</td>
</tr>
<tr>
<td>Class III</td>
<td>1, 6</td>
<td>30</td>
</tr>
<tr>
<td>Class IV</td>
<td>3</td>
<td>120</td>
</tr>
</tbody>
</table>

1.2.4. Transposable Elements

Similar to other eukaryotes, the *T. brucei* genome contains a variety of transposable elements (TEs). However these elements comprise less than 5% of the genome, in contrast to roughly 40% for humans and nearly 85% for some species (Bringaud et al., 2008; Cordaux & Batzer, 2009; Wegrzyn et al., 2013). The *T. brucei* genome contains several types of class 1 TEs which are predicted to transpose via reverse transcription of an RNA intermediate, but no class II TEs which transpose via a DNA intermediate. All of these class I TEs belong to either the CRE or the *ingi* clade. The CRE clade contains the site-specific retroposon SLACS (Spliced Leader Associated Conserved Sequence), which inserts into the spliced leader RNA genes discussed earlier. The *ingi* clade contains the non-site-specific retroposons *ingi*, RIME, TbSIDER, and DIRE. These elements all contain a conserved 79 base-pair (bp) sequence at their N-terminus. Although not site-specific, these retroposons are highly over-represented in strand-switch regions, and to a lesser extent subtelomeric regions, than within directional gene clusters (Bringaud et al., 2008). TEs are also abundant at the centromere regions of the 11 megabase chromosomes, as mentioned above. The most obvious interpretation of this localisation is that the only other sites available for insertion are within directional gene clusters, where they are more likely to be mutagenic and therefore selected against. This logic
seems to hold true for the *T. cruzi* genome, but not for the *L. major* genome. In *L. major* there is a much higher density of TEs within directional gene clusters (approximately 50X), and these elements are found within the 3'UTRs of genes, where they contribute to post-transcriptional regulation of gene expression (Bringaud et al., 2007a).

The retrotransposon hot spot (RHS) gene family is an exception to the lack of site-specificity in TE insertion. This multigene family clusters at the sub-telomeres of megabase chromosomes, and encodes mostly nuclear proteins, which appear to be involved in transcription (Florini et al., 2019). Approximately 60% of this TE family are pseudogenes, in many cases a direct result of the site-specific insertion of Ingi/RIME elements in their N-terminal regions.

### Table 1.2. Transposable elements found in *T. brucei*. Copy number figures are per haploid genome. LTR = long terminal repeats.

<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
<th>Copy Number</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Non site-specific retroposons (Ingi clade)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ingi</td>
<td>5.25 kb</td>
<td>115</td>
<td>(Kimmel et al., 1987; Murphy et al., 1987)</td>
</tr>
<tr>
<td>RIME (Ribosomal Mobile Element)</td>
<td>0.5 kb</td>
<td>86</td>
<td>(Hasan et al., 1984)</td>
</tr>
<tr>
<td>TbSIDER (Short Interspersed DEgenerated Retroposon)</td>
<td>0.57 kb</td>
<td>26</td>
<td>(Bringaud et al., 2007b)</td>
</tr>
<tr>
<td>DIRE (Degenerate <em>ingi</em>/L1Tc-Related Elements)</td>
<td>5 kb</td>
<td>73</td>
<td>(Ghedin et al., 2004)</td>
</tr>
<tr>
<td><strong>LTR Retrotransposons</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VIPER (Vestigal InterPosEd Retroelement)</td>
<td>4.5 kb</td>
<td>26</td>
<td>(Lorenzi et al., 2006)</td>
</tr>
<tr>
<td>SIRE (Short Interspersed Repetitive Element)</td>
<td>0.43 kb</td>
<td>10</td>
<td>(Berriman et al., 2005)</td>
</tr>
<tr>
<td><strong>Site-specific retroposons (CRE clade)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLACS</td>
<td>6.7 kb</td>
<td>4</td>
<td>(Aksoy, 1991)</td>
</tr>
</tbody>
</table>
1.3 Chromatin

Like all eukaryotes, the nuclear genome in *T. brucei* is packaged in chromatin, where the DNA is complexed with structural and regulatory proteins. The core unit of chromatin is the nucleosome, which consists of 146 bp of DNA wrapped around a histone octamer comprised of two copies each of histones H2A, H2B, H3, and H4, as well as the ‘linker’ histone H1 (McGhee & Felsenfeld, 1980). Post-translational modification of histones, particularly of the N-terminal tails that protrude from the core particle, can alter chromatin structure as well as gene expression. For example, di- and tri-methylation of lysine 9 on histone H3 (H3K9me2/3) is associated with compact, transcriptionally ‘inactive’ heterochromatin, while loosely packed, transcriptionally active euchromatin is enriched in H3K4me. Prior to the characterisation of these PTMs, heterochromatin and euchromatin had been defined cytologically, with dense heterochromatin being found primarily at the nuclear periphery (Passarge, 1979; Rae & Franke, 1972).

In addition to low levels of transcription, heterochromatin is characterized by its enrichment for repetitive sequences, distinct replication timing, and compact structure during interphase (Allshire & Madhani, 2017). Heterochromatin is also characterized by specific proteins such as H3K9me2/3-binding protein heterochromatin protein 1 (HP1) (James & Elgin, 1986; Janssen et al., 2018). Heterochromatin can be further subdivided into constitutive and facultative heterochromatin. Constitutive heterochromatin is generally comprised of repetitive elements such as telomeres, centromeres, and transposable elements, and is important for preventing genetic instability by silencing the latter. Facultative heterochromatin is cell-type or condition specific, and can be found on the inactive X chromosome, for example.
Chromatin can also be remodelled, for example by replacing core histones with histone variants, some of which have important cellular functions. One such example is the H3 variant centromere protein A (CENP-A) which replaces H3 in some nucleosomes within centromeric repeats, where it forms the foundation of the kinetochore in many eukaryotes (McKinley & Cheeseman, 2016).

Chromatin is remodelled by the action of chromatin remodelling enzymes, which use ATP hydrolysis to move nucleosomes along the chromatin fibre, as well as eject, destabilize, or restructure them as required (Clapier & Cairns, 2009). These processes are necessary for allowing access to DNA for transcription, replication, and repair, and for creating specialized regions of chromatin (Clapier & Cairns, 2009). Histone chaperone complexes are also important for many of these functions (Hammond et al., 2017). For example, a specialized histone chaperone called HJURP (Holliday Junction-Recognizing Protein) delivers the aforementioned CENP-A to centromeres (Dunleavy et al., 2009; Foltz et al., 2009).

1.3.1 Histones and histone variants in *T. brucei*

Although the histone cores of trypanosomes are conserved with those of higher eukaryotes, the sequences of the N-terminal tails are highly divergent (Galanti et al., 1998; Janzen, Fernandez, et al., 2006; Mandava et al., 2007b). As a result, many of the conserved modifications such as H3K9me seen in other eukaryotes are either absent or not recognized by existing antibody reagents due to differences in surrounding residues (see Figure 1.5 below). Conversely, several unusual modifications have been identified, including methylation of N-terminal alanine residues on H2A, H2B, and H4, and hyper-acetylation of the H2A C-terminus (Janzen, Fernandez, et al., 2006). Although most modifications have not been characterised, some have been implicated in transcription initiation (Siegel et al., 2009; Wright et al., 2010) and cell cycle regulation (Janzen, Hake, et al., 2006).

The ‘linker’ histone H1 is also divergent, and appears to contain just the C-terminal domain of H1 from higher eukaryotes, lacking the globular domain important for interaction with the nucleosome (Grüter & Betschart, 2001; Y. B. Zhou et al., 1998). This has led to the hypothesis that trypanosome H1 may contribute to chromatin compaction in a distinct manner to that seen in other eukaryotes, and will be discussed later in Section 1.4.2.
Fig 1.5. Multiple sequence alignment of core histone termini of *T. brucei* with those of common model eukaryotes. Residues conserved among the species are highlighted in black. Vertical arrows represent important regions of homology such as H3K9 as mentioned in the text. Acetylation, methylation and ubiquitination sites of *T. brucei* histones are indicated by the symbol *. Red = acetylation; blue = methylation; green = ubiquitination. Figure from Mandava et al. (Mandava et al., 2007a).
T. brucei has four histone variants: H2AZ, H2BV, H3V, and H4V. H2A.Z and H2BV are enriched at transcription start sites, while H3V and H4V are found at transcription termination sites (Reynolds et al., 2016; Schulz et al., 2016; Siegel et al., 2009). H3V appears to work in concert with base J (discussed in Section 1.3.2 below) to terminate transcription and prevent read through at convergent strand switch regions (Reynolds et al., 2016; Schulz et al., 2016). H3V is also enriched at repetitive loci such as telomeres, Chromosome Internal Repeats of 147 bp (CIR147 repeats), ingi TE’s, MC 177 bp repeats, and silent BESs (Akiyoshi & Gull, 2014a; Lowell & Cross, 2004; Müller et al., 2018). Although enriched at CIR147 repeats, sequence analysis would suggest that this H3V is not homologous to the centromere-defining variant CENP-A seen in higher eukaryotes, and it is not required for chromosome segregation or cell viability (Lowell & Cross, 2004). H3V knockout cells show an increase in small interfering RNAs (siRNAs) from all known siRNA loci, which suggests it does not just terminate transcription, but might have a role in transcriptional silencing more generally (Reynolds et al., 2016). This role does not appear to be restricted to RNAPII transcription, as deletion of H3V also shows an increase in the expression of silent BESs (Reynolds et al., 2016; Schulz et al., 2016). The Chromosome Conformation Capture technique Hi-C, which identifies which genomic sequences are physically close to one another within the nucleus, showed that deletion of H3V causes telomere clustering and an increase in interaction frequency between silent BESs (Müller et al., 2018). Deletion of both H3V and H4V results in an opening of the chromatin structure at silent BESs as measured by assay for transposase accessible chromatin followed by sequencing (ATAC-seq) (Müller et al., 2018). These results further strengthen the hypothesis that H3V could have a role in transcriptional silencing and heterochromatin formation in T. brucei.

1.3.2 DNA modifications in T. brucei

Another feature of chromatin is that the DNA itself can be modified in order to alter gene expression and chromatin structure. In many prokaryotic and eukaryotic systems, DNA can be methylated on cytosine residues of CG dinucleotides (5meC), and in higher eukaryotes this is associated with heterochromatin and transcriptional repression (Jones, 2012). N6-methyladenine (6mA) is another, less common, DNA modification which has been implicated in the regulation of TEs in both Drosophila melanogaster and mice (Wu et al., 2016; G. Zhang et al., 2015).

The T. brucei genome encodes a putative DNA methyltransferase (DNMT) which is expressed in both bloodstream and procyclic forms, and low levels (~0.01%) of 5meC were
identified in both lifecycle forms by immunoblotting and immunoprecipitation with an anti-5meC antibody as well as mass spectrometry (Militello et al., 2008). The immunoprecipitated DNA was cloned and sequenced, finding 5meC at 65 loci, with a slight bias towards retrotransposon hotspot loci (see Section 1.2.4) and VSG pseudogenes. However, the low levels of DNA methylation present make it unclear how important this modification is in *T. brucei*. Unpublished bisulphite sequencing data from the Allshire lab failed to identify any highly methylated loci in either bloodstream or procyclic forms. The reason for the discrepancy is unclear. Tagging of the putative DNMT with mNeonGreen at either terminus by the TrypTag initiative gave a cytoplasmic localisation (Dean et al., 2017). N-terminal tagging with YFP in the bloodstream form also found a cytoplasmic localisation (Dr Roberta Carloni, Allshire/Matthews labs, unpublished). ChIP-seq analysis found no enrichment on chromosomes, and immunoprecipitation followed by mass spectrometry failed to identify any specific interacting partners, and no nuclear or chromatin-associated proteins (Dr Roberta Carloni, unpublished). Taken together these data suggest that this protein does not methylate specific nuclear DNA loci, and may well function as a cytoplasmic methyltransferase acting on an unknown substrate. In support of this, *L. donovani* also has no detectable DNA methylation, despite expressing a DNMT (Cuypers et al., 2020).

On the other hand, the nuclear DNA of *T. brucei* does contain a kinetoplastid-specific modified base, β-d-glucopyranosyloxymethyluracil, or base J (J. Gomers-ampt et al., 1991; Gomers-Ampt et al., 1993; J. H. Gomers-ampt et al., 1993). This modification is reported to be exclusive to bloodstream form parasites, and is located at telomeres, 50 bp repeats upstream of VSG promoters, 70 bp repeats in the VSG expression sites, 177 bp repeats on minichromosomes, spliced leader RNA genes, 5S RNA genes, and transcription termination sites (Cliffe et al., 2009; van Leeuwen et al., 1996, 1997, 2000). Base J appears to be important for preventing read-through transcription at convergent SSRs (Schulz et al., 2016). Interestingly, no base J was found at RIME or SLACS transposable elements, and so seems to only be associated with tandem repeat elements (van Leeuwen et al., 2000).

Base J is synthesised and maintained by J-binding protein 1 (JBP1), JBP2, and a β-glucosyltransferase (Adamopoulos et al., 2019; Bullard et al., 2014; Cliffe et al., 2009; Dipaolo et al., 2005; Kieft et al., 2007). JBP1 and JBP2 appear to form part of a complex enriched at transcription start regions (Stanewa et al., 2021), which is somewhat at odds with the reported distribution of base J mentioned above.
Figure 1.6. Base J is synthesised by modification of thymidine residues in duplex DNA. Thymine is first hydroxylated by JBP1 or JBP2 to form hydroxymethyldeoxyuridine (HOMedU), then glucosylated to form J. Figure adapted from Borst and Sabatini, 2008 (Borst & Sabatini, 2008).

1.4. Heterochromatin in *T. brucei*

Transmission electron microscopy analysis shows that there are areas within the *T. brucei* nucleus that differ in electron density, with electron dense ‘heterochromatic’ regions associated mainly with the nuclear periphery (Rout & Field, 2001a). Comparison of slender and procyclic nuclei showed marked differences in the amount of electron dense chromatin present, with more extensive heterochromatin regions visible in the slender forms (Figure 1.7 below). This could be consistent with lifecycle-specific gene expression requirements.
Figure 1.7. Differences in heterochromatin between Procyclic and Bloodstream from *T. brucei* are visible by electron microscopy. Uranyl acetate stained isolated nuclei. The procyclic form nucleus is rounded, with a large nucleolus, and has small regions of heterochromatin distributed throughout the nucleus and at the nuclear envelope (excluding the areas surrounding the nuclear pores). The bloodstream form nucleus is irregular in shape, has a smaller nucleolus, and large regions of heterochromatin that are predominantly associated with the NE. Adapted from Figure 3, Rout & Field, 2001 (Rout & Field, 2001b).

1.4.1. Candidate heterochromatin regions in the *T. brucei* genome

The sequences present within the electron dense chromatin visible in Figure 1.7 are unknown. Based on the composition of heterochromatin in other eukaryotes, the following regions are likely candidates due to their repetitive structure: telomeres, centromere repeats (Obado et al., 2007b, 2011), 177bp repeats present on the minichromosomes (Sloof et al., 1983), and the transposable elements discussed in Section 1.2, as well as 50 bp repeats upstream of VSG expression site promoters, and 70bp repeats upstream of VSG genes (Hertz-Fowler et al., 2008). Procyclin genes could also be packaged in heterochromatin in BSF where they are silenced. There is experimental evidence to support the idea that at least some of these sequences are localised at the nuclear periphery (discussed below), and therefore potentially packaged in heterochromatin.
1.4.1.1. Telomeres

Telomeres are packaged in constitutive heterochromatin in the nuclei of most eukaryotes (Janssen et al., 2018). Fluorescence *in situ* hybridisation (FISH) for the telomere repeat sequence and immunofluorescence (IF) analysis of *T. brucei* telomere proteins TbTRF, TbRAP1, TbTIF2, TelAP1, and UMSBP2 shows that telomeres have a peripheral position within the nucleus, which is compatible with their localization in the peripheral heterochromatin seen in Figure 1.7 (Jehi et al., 2014; Klebanov-Akopyan et al., 2018; B. Li et al., 2005; Reis et al., 2018; Yang et al., 2009). In addition, Telomeric Repeat-containing RNAs (TERRA), which have been shown to be involved in heterochromatin formation in human cells (Deng et al., 2009), have been identified in *T. brucei* (Damasceno et al., 2017a), which is consistent with the potential packaging of trypanosome telomeres in heterochromatin.

The presence of a telomere position effect (TPE) is another piece of evidence suggesting telomeres may be packaged in heterochromatin in *T. brucei* (Horn & Cross, 1995). TPE refers to a phenomenon whereby genes proximal to telomeres are transcriptionally silenced due to the spread of heterochromatin from telomeres into sub-telomeres, and appears to be broadly conserved in eukaryotes (Ottaviani et al., 2008).

1.4.1.2. Centromeres

Pericentromeric regions surrounding centromeres are heterochromatic in many eukaryotes. Heterochromatin and related proteins have been shown to be vital for centromere function in other organisms (discussed extensively in Janssen et al., 2018). As described in Section 1.2.3, *T. brucei* centromeres are enriched in repetitive sequences. These sequences are targeted by the RNA interference (RNAi) pathway, which is involved in heterochromatin formation in other eukaryotes (discussed in Section 1.4.4). However, centromeres are not localized at the nuclear periphery during interphase (Akiyoshi & Gull, 2014b).

1.4.1.3. Transposable elements

Transposable elements are potentially hazardous to genome stability, and are consequently packaged in heterochromatin in many eukaryotic genomes (Janssen et al., 2018). As outlined
in Section 1.4.4, the various transposable elements in the *T. brucei* genome are targeted by the RNAi pathway, which is important for heterochromatin formation in other eukaryotes. The subcellular localisations of the various transposable elements have not been reported, so it is unknown whether they can be found in the heterochromatin at the nuclear periphery.

### 1.4.1.4. Minichromosomes

The roughly 100 minichromosomes are another prime candidate for forming heterochromatin. These chromosomes contain no housekeeping genes and are transcriptionally silent (Weiden et al., 1991; Wickstead et al., 2004). The majority of these chromosomes is comprised of repetitive DNA, as visualised schematically in Figure 1.3. Minichromosomes are also localized to the nuclear periphery in G1 phase of the cell cycle in both procyclic and slender forms (Chung et al., 1990; Ersfeld & Gull, 1997). Integration of a tetracycline-inducible ectopic gene expression cassette into the minichromosomal 177 bp repeats gives better regulation of protein expression than when integrated into the rDNA spacer locus, which might be due to the minichromosomes having a silent chromatin state less permissive to gene expression (Wickstead et al., 2002).

### 1.4.1.5. VSG and procyclin loci

Silent VSG ESs also follow a similar localization pattern to telomeres, with a combination of FISH and GFP-*lacI* tagging showing they are found at the nuclear periphery in procyclic forms, but lack a specific localisation within the nucleus in slender forms (Chaves et al., 1998; Landeira & Navarro, 2007). It is tempting to speculate that the ESs would have greater mobility within the nucleus in bloodstream forms to allow for VSG switching. Further, the active ES has been shown to migrate to the nuclear periphery upon differentiation of slender to procyclic forms in culture (Landeira & Navarro, 2007). The lamin-like protein NUP-1 also influences telomere positioning, and knockdown in BSF cells results in de-repression of procyclin, VSG genes, and the SL genes (DuBois et al., 2012). The subcellular localization of procyclin genes in slender forms has not been explored, but this de-repression suggests they too might be silenced by moving to the nuclear periphery in bloodstream forms. These data bolster the idea that tethering a specific locus to the nuclear periphery is important for its transcriptional repression. It is worth noting however that the VSG and procyclin loci are transcribed by RNAPI,
and there is no evidence RNAPII transcription can be silenced by localizing to the nuclear periphery in trypanosomes.

### 1.4.2. Histones and histone variants

Histones and histone variants could also have a role in heterochromatin formation and function in *T. brucei*. Knockdown of H1 in slender forms leads to loss of electron dense chromatin, an increase in nuclease sensitivity, VSG de-repression, and increased VSG switching (Povelones et al., 2012). Interestingly, this increase in nuclease sensitivity was not seen in procyclic forms. The authors posit that this could either be due to H1 playing different roles in these two lifecycle forms, or simply due to procyclic cells having less electron dense chromatin to begin with. Further work analysing nascent transcripts found that H1 depletion increases transcripts from VSG and procyclin genes 2 to 6-fold, but had no effect on the levels of RNAPII transcripts (Pena et al., 2014).

During interphase, H3V is concentrated at the nuclear periphery (Lowell & Cross, 2004), and as mentioned in Section 1.3.1, it is known to be enriched on all of the sequences discussed in Section 1.4.1. Further, FISH, Hi-C, and assays for transposase-accessible chromatin using sequencing (ATAC-seq) showed that telomeres clustered, silent BESs interacted with one another more frequently, and chromatin structure was more open at BES promoters in cells lacking H3V (Müller et al., 2018). These findings suggest H3V plays a role in silencing sub-telomeric regions of the *T. brucei* genome.

### 1.4.3. Chromatin Remodellers and histone chaperones

As mentioned previously, chromatin remodelers and histone chaperones are important for creating specialized regions of chromatin in other eukaryotes. Some of the chromatin remodelers and histone chaperones characterised in *T. brucei* will be discussed below, with a focus on possible connections to heterochromatin formation or function.

Imitation SWItch (ISWI) functions as the catalytic subunit of a diverse family of chromatin remodelers with roles in transcription, DNA replication, and chromosome organization (Clapier & Cairns, 2009; Corona & Tamkun, 2004). *T. brucei* possesses a divergent ISWI chromatin remodelling complex. Immunoprecipitation of TAP-tagged ISWI identified a 4 subunit complex comprised of ISWI, nucleoplasmin-like protein (NLP), regulator
of chromosome condensation 1-like protein (RCCP) and phenylalanine/tyrosine-rich protein (FYRP) (T. Stanne et al., 2015). NLP, RCCP, and FYRP are not homologous to ISWI partners identified in other eukaryotes. A combination of ChIP-slot blot and ChIP-qPCR experiments found ISWI enrichment on minichromosome 177 bp repeats, telomeric repeats, 50 bp repeats from the VSG BESs, a silent VSG gene, and some strand-switch regions where RNAPII transcription starts and terminates. Enrichment was also found at the RNAPI-transcribed rDNA, VSG, and procyclin loci. However, ISWI was also enriched at the tubulin gene array, suggesting a general role in transcription in *T. brucei*, which is unsurprising given its role in other eukaryotes. On the other hand, knockdown of ISWI resulted in increased transcript levels from the rDNA, silent VSG, and procyclin loci but not the tubulin gene array, indicating a role for ISWI in the control of RNAPI - but not RNAPII - transcription in bloodstream forms (Hughes et al., 2007; T. Stanne et al., 2015; T. M. Stanne et al., 2011). The genome wide distribution of ISWI using ChIP-seq has not been reported, and therefore it is unknown whether ISWI is present at the other putative heterochromatin loci in *T. brucei*.

Anti-silencing function 1 (Asf1) is a histone chaperone which is involved in depositing and removing H3-H4 heterodimers in various cellular processes, including DNA replication, DNA damage repair, transcription, and silencing (Mousson et al., 2007). There are two isoforms of Asf1 in *T. brucei*, Asf1A and Asf1B. Asf1A has a predominantly cytoplasmic localisation, but concentrates in the nucleus during S-phase, whereas Asf1B is nuclear throughout the cell cycle (Pascoalino et al., 2014). Knockdown of Asf1A causes de-repression of a reporter gene integrated at a silent VSG ES at all stages of the cell cycle, suggesting it has a role in ES silencing independent of its role in DNA replication (Alsford & Horn, 2012).

Chromatin Assembly Factor-1 (CAF-1) is a highly conserved histone H3-H4 chaperone, which deposits H3-H4 onto newly synthesized DNA during DNA replication (Panne et al., 2018). CAF-1 also functions in DNA repair, heterochromatin formation, and heterochromatin maintenance in other eukaryotes (Cheng et al., 2019; Gaillard et al., 1997; H. Huang et al., 2010; S. Huang et al., 2007; Linger & Tyler, 2005a, 2005b; Nabatiyan et al., 2006; Quivy et al., 2008; Roelens et al., 2017). In *T. brucei*, loss of CAF-1b results in de-repression of silent VSG ESs, and the CAF-1 complex as a whole was recently shown to associate with the VSG allelic exclusion complex discussed in Section 1.2.1 (Alsford & Horn, 2012; Faria et al., 2019). If silent VSG ESs are packaged in heterochromatin, then this would also implicate CAF-1 in heterochromatin maintenance in trypanosomes.
1.4.4. Writers, Readers, and Erasers

The fact that the *T. brucei* genome encodes numerous proteins containing domains that potentially bind histone modifications (‘Readers’), modify histones (‘Writers’), or remove histone modifications (‘Erasers’) is another obvious source of heterochromatin formation. These include a chromodomain proteins, bromodomain proteins, histone acetyltransferases (HATs), histone deacetylases (HDACs), and su(var)3–9, enhancer of zeste, trithorax (SET) domain proteins among others, and are discussed elsewhere (Figueiredo et al., 2009). Several of these proteins have been linked with some of the histone modifications discussed previously (Kawahara et al., 2008; Siegel et al., 2007; Yang et al., 2017), and some have been implicated in telomeric silencing (Alsford et al., 2007; Q.-P. Wang et al., 2010). A screen of approximately 70 of these proteins did not identify any of them as being highly enriched on the repetitive sequences discussed in Section 1.4.1 above (Staneva et al., 2021).

1.4.5. RNA interference

Another potential source of heterochromatin formation in *T. brucei* is via RNA-mediated recruitment of silencing factors as seen in other systems such as *Schizosaccharomyces pombe* and *Arabidopsis thaliana* (Martienssen & Moazed, 2015). In *S. pombe*, the argonaute protein Ago1 forms part of the RNA-induced transcriptional silencing (RITS) complex (Verdel et al., 2004), where it binds single-stranded siRNAs produced by the action of the RNA-dependent RNA polymerase complex (RDRC) (Motamedi et al., 2004) and Dicer. The RITS uses these siRNAs to target nascent heterochromatin transcripts and to recruit the Clr4 methyltransferase complex (CLRC) to induce H3K9me and heterochromatin formation (K. Zhang et al., 2008). The RITS is then further recruited in a positive feedback loop by binding H3K9me chromatin via its chromodomain (Sadaie et al., 2004). In *A. thaliana*, this process is broadly similar, with the important addition that the DNA at the siRNA-targeted loci is methylated, in addition to the histones being post-translationally modified (Martienssen & Moazed, 2015).

*T. brucei* has a single Argonaute protein, TbAGO1 (Shi et al., 2004), and two Dicers, one cytoplasmic and one nuclear (Patrick et al., 2009a; Shi et al., 2006) and a fully functional RNAi pathway. The major siRNA producing loci in *T. brucei*, in order of decreasing abundance, are the ingi and SLACS TE’s (Djikeng et al., 2001), CIR147 repeats (Patrick et al., 2009a), inverted repeats, and convergent transcription units (Tschudi et al., 2012a). Strangely, no siRNAs mapped to tandem repeats other than the CIR147 repeats. This would suggest that the RNAi pathway is not involved in silencing other tandem repeats such as the 50 bp repeats.
and 70 bp repeats found within the VSG ES. Consistent with this, ES silencing is normal in AGO1 knockout cells (Janzen, van Deursen, et al., 2006a). The absence of siRNAs from silent VSG genes and VSG expression sites from AGO1 immunoprecipitates argues against its involvement in silencing one of the primary candidate heterochromatin regions of the *T. brucei* genome (Tschudi et al., 2012b). It is important to note that all of the work referenced from the Ullu lab has been performed on the subspecies *T. brucei rhodesiense*, which could have differences to the *T. b. brucei* subspecies. It is also worth noting that Ago1-associated small RNA reads generated by Tschudi et al. were only mapped to the 11 megabase chromosomes. One might expect the 177bp inverted repeats on the minichromosomes to also be targeted by the RNAi machinery, but no investigation of this has been reported.

AGO1 epitope tagged at the N-terminus appears to be exclusively cytoplasmic, and ChIP and immunoprecipitation experiments show no detectable association with chromatin (Shi et al., 2004; Staneva et al., 2021). This suggests that if AGO1 is involved in heterochromatin formation in *T. brucei*, then it likely has an indirect role compared to that seen in *S. pombe* and plants.

The presence of siRNAs from the CIR147 centromere repeats also suggests that the endogenous RNAi system might be linked to the assembly of centromeric heterochromatin, as in *S. pombe* (Allshire & Madhani, 2017; Buscaino et al., 2010). Consistent with this, AGO1 is important for faithful segregation of large and minichromosomes, and knockout leads to a strong growth defect in procyclic form cells (Durand-Dubief et al., 2007; Durand-Dubief & Bastin, 2003). However, this growth defect was not seen in bloodstream form cells (Janzen, van Deursen, et al., 2006b).

On the other hand, the RNAi machinery is not conserved among all trypanosomatid species, which argues against its possible importance in heterochromatin formation in this class of organisms as a whole. RNAi is present in the related African trypanosome *Trypanosoma congoense* (Inoue et al., 2002) and *L. brefiliensis* (Peacock et al., 2007), but not in *T. cruzi*, *L. major*, *L. donovani* or *L. infantum* (DaRocha et al., 2004; Robinson & Beverley, 2003).
1.5. Tools for the exploration of trypanosome heterochromatin: Transcription activator-like effector DNA binding proteins

Transcription activator-like (TAL) effectors are DNA binding proteins found in plant pathogenic *Xanthomonas* bacteria (Figure 1.8). These proteins are injected into plant cells where they enter the nucleus, bind to specific promoter sequences, and activate transcription of specific plant genes to aid their virulence (Kay et al., 2007; Römer et al., 2007). DNA recognition is mediated by tandem repeats of approximately 34 amino acids, with each repeat binding one of the four DNA nucleotides (Boch et al., 2009; Moscou & Bogdanove, 2009). Each repeat unit, which forms a two-helix bundle, contains 2 variable amino acid residues called repeat-variable di-residues (RVD’s) which are responsible for nucleotide binding and recognition, with different pairs of RVD’s binding different nucleotides (Figure 1.8, panel B). For example, H-G binds thymine, N-N binds guanine and so on (Moscou & Bogdanove, 2009).

**Figure 1.8.** TAL effectors form a right-handed superhelix around the major groove of DNA and make base-specific contacts with the sense strand. (A) The structure of the DNA binding region of the TAL effector PthXo1 from the rice pathogen *Xanthomonas oryzae* in complex with its target site. Each colour represents one of the tandem repeats. Each repeat binds a different DNA nucleotide depending on two variable amino acid residues termed repeat-variable di-residues (RVD’s). (B) Examples of different RVD pairs and the corresponding nucleotides they bind. For example, H-D binds cytosine via hydrogen bonds formed between the amino group of cytosine and the hydroxyl group of aspartate. Figure adapted from Mak et. al (A. N.-S. Mak et al., 2012).
Based on this modular structure, it was quickly realized that repeat units with known RVD pairs could be placed in tandem to bind any DNA sequence of interest, effectively allowing the design of bespoke DNA binding proteins. Researchers took advantage of this, fusing these domains to FokI restriction endonuclease domains to create the genome editing tools transcription activator-like effector nucleases (TALENs). TALENs have been used for genome editing in a variety of organisms, including yeast, plants, zebrafish, rats, and human cells (Bedell et al., 2012; Cermak et al., 2011; Ding et al., 2013; P. Huang et al., 2011; T. Li et al., 2011, 2012; Tesson et al., 2011). Other applications of TAL effectors (TALEs) that were rapidly developed included designing customizable transcription factors (Boch et al., 2009; Geißler et al., 2011; Morbitzer et al., 2010; F. Zhang et al., 2011) and visualizing specific chromatin loci within the cell (Fujimoto et al., 2016; Miyanari et al., 2013). TALEs have also been used for locus-specific protein enrichment approaches which will be discussed in Section 1.5.

TALEs have since been superseded by CRISPR-Cas proteins for genome editing due to their relative simplicity and speed of design and use. Design of a new TALE domain requires extensive cloning, whereas targeting a new locus using CRISPR proteins simply requires a new guide RNA. This is particularly advantageous when pursuing high throughput approaches. However, TALENs may be superior for some applications such as editing within heterochromatic regions (Jain et al., 2021).

1.6. Locus-specific proteomics

In genome biology, it is well known that the protein composition of a particular genomic locus can vary greatly, and the proteins present at a particular locus heavily influence its function. For example, a promoter sequence will be bound by various transcription factors and contain histones with particular PTMs that contribute to its successful transcription into RNA. In order to understand the function of a particular locus, researchers currently use techniques such as chromatin immunoprecipitation coupled to next generation sequencing (ChIP-seq) to map the genome-wide distribution of particular histone PTMs and chromatin proteins, attempting to correlate these activities with nuclear processes such as transcription. Combined with affinity purification coupled to mass spectrometry (AP-MS), these methodologies have expanded our knowledge of genome biology hugely. However, these techniques require prior knowledge of the particular protein or PTM, and depend on the availability of antibodies against that factor or the use of epitope-tagged proteins. Further, each factor needs to be assayed individually, and the data generated is genome-wide as opposed to being locus specific. For example, an
AP-MS experiment using a chromatin reader as bait identifies proteins associated with that reader regardless of its location within the nucleus, but the reader may interact with different proteins depending on the locus in question.

When determining which proteins were bound by a specific DNA sequence, researchers traditionally used *in vitro* DNA affinity pulldowns (Kadonaga & Tjian, 1986). More recently, this technique has been expanded to utilise methylated DNA and nucleosomes containing specific histone PTMs (Bartke et al., 2010; Mittler et al., 2009; Spruijt et al., 2013; Vermeulcn et al., 2010). These *in vitro* approaches are useful, but cannot completely recapitulate the *in vivo* chromatin environment.

The limitations of these *in vitro* approaches has led to extensive attempts at performing locus-specific chromatin purifications. Various different approaches have been used (Figure 1.9 below). The primary methodologies used are: immunoprecipitation of locus-specific factors, genetic engineering, engineered DNA-binding proteins, and oligonucleotide probes. Each of these is discussed below. Additional approaches based on proximity biotinylation and a genomics-based method have also been developed, but will not be discussed here (Gao et al., 2018; Korthout et al., 2018; Myers et al., 2018; Qiu et al., 2019; Schmidtmann et al., 2016). A more extensive discussion of the challenges involved and critique of approaches used can be found elsewhere (Gauchier et al., 2020).
Figure 1.9. Different approaches have been used to purify locus-specific chromatin. (a) Capture using labelled oligonucleotides. (b) Capture using dCas9 proteins. (c) Immunoprecipitation of factors specific to the locus of interest. (d) Capture using epitope-tagged TALE proteins. (e) Genomic engineering-based methods. Figure adapted from Gauchier et. al (Gauchier et al., 2020).

Immunoprecipitation of locus-specific factors can be performed for a small number of loci in the cell such as telomeres and centromeres, which are bound by cohorts of highly specific proteins not found elsewhere. However, this approach is of limited value, as most loci are bound by proteins found at many different sites genome-wide.

Genetic engineering-based approaches modify the locus of interest by inserting specialised DNA sequences such as Tet operator (TetO) or LexA repressor sequences. Expression of the epitope-tagged TetR and LexA proteins in these cells then allows for the
enrichment of the locus and its associated proteins. Genetic engineering-based approaches have been used to investigate the composition of the PHO5 gene promoter, the chicken β-globin HS4 insulator, the GAL1 promoter region in S. cerevisiae, the human fetal-γ-globin gene, and others (Byrum et al., 2012; Fujita & Fujii, 2011; Griesenbeck et al., 2003; Hamperl et al., 2014; Pourfarzad et al., 2013). However, the fact that the locus itself is modified in such approaches raises the question of whether the results identified truly reflect the composition of the unmodified locus.

Approaches based on engineered DNA-binding proteins utilise either TALEs or catalytically inactive Cas9 (dCas9) proteins. In most cases, an epitope tag is used to purify the engineered DNA-binding protein along with its associated target locus. TALEs have been used to purify the GAL1 promoter from S. cerevisiae, telomeres in mouse and human cells, and human pericentromeric SATIII repeats (Byrum et al., 2013; Fujita et al., 2013; Witte et al., 2020). The dCas9 approach has also been used to purify the GAL1 promoter in yeast, the IRF-1 locus from human cells, telomeres, single-copy loci from the human β-globin cluster, and the histone locus from Drosophila cells (Fujita & Fujii, 2013; Liu et al., 2017; Tsui et al., 2018; Waldrip et al., 2014).

Labelled oligonucleotide probes specific for a locus of interest can be hybridized to crosslinked, denatured chromatin, and then used for purification of the locus. Oligonucleotide-based approaches have been used to purify human telomeres, mouse pericentromeric heterochromatin, a ribosomal RNA promoter, human alpha-satellite repeats, and multiple loci in S. cerevisiae, including the ENO2 and GAL1 promoter regions (Buxton et al., 2017; Déjardin & Kingston, 2009; Guillen-Ahlers et al., 2016; Ide & Dejardin, 2015; Kennedy-Darling et al., 2014; Saksouk et al., 2014). The lack of genetic engineering requirements makes such methodologies relatively straightforward, but the requirement to use crosslinked samples may lead to the presence of excessive background signal in some cases.

All of these methods are faced with the same challenges: 1) chromatin solubilisation and preparation; 2) locus-capture; and 3) enrichment of the target locus relative to background signal. Achieving the perfect buffer conditions to stabilise all chromatin components in vitro cannot practically be achieved, so certain interactions will inevitably be lost. Formaldehyde crosslinking can be used to stabilise interactions, but is thought to increase background signal and makes solubilisation more difficult. Capturing a specific locus is challenging due to the fact that it may only be present in small amounts relative to the rest of the genome, and it will share similarities with all other loci. For example, histones will be present at all loci, and transcription factors and RNA polymerase components would presumably be present at all promoters and across gene bodies. Depending on the copy number of the locus in question,
and the overall size of the genome of the organism, the signal-to-noise ratio in these approaches can be very high. For example, purifying a 1kb locus from a 1 gigabase genome represents 0.0001% of that genome. The remaining 99.9999% of the sample will share DNA and histones with the target locus, and a large percentage will share certain other chromatin proteins such as readers or writers of histone modifications. All of these factors combine to make this a formidable biochemical challenge.

1.6.1. Locus-specific proteomics in *Trypanosoma brucei*

None of the approaches outlined in Figure 1.9 have been reported in *T. brucei*. However, *T. brucei* has a much more condensed genome than many eukaryotes due to its unusual genome organisation as discussed in Section 1.2. Theoretically, this substantially decreases the locus enrichment challenge. For example, a 1 kb locus represents ~0.0028% of the genome. Further, the *T. brucei* genome contains many repetitive loci of unknown protein composition such as the putative heterochromatin loci discussed in Section 1.4.1. At present, DNA affinity pulldown or performing ChIP-seq on all chromatin proteins are the only techniques available for identifying proteins enriched at all of these loci. Therefore, locus-specific approaches represent one of the only methodologies that can be used for characterising the proteins bound to these loci *in vivo*. Characterising the proteomes of these repetitive sequences will be key to understanding the formation and function of heterochromatin in *T. brucei*.

In order to address the question of what proteins are present on putative heterochromatin sequences in *T. brucei* using locus-specific proteomics, it was first necessary to decide which approach to employ. As mentioned above, and discussed in detail elsewhere (Gauchier et al., 2020), each of the approaches outlined in Figure 1.9 have strengths and weaknesses, and no technique is completely superior to others. In addition, the skillsets and experience available in the laboratory influence which approach is most practical for this specific project. Using TALE DNA binding proteins allowed me to capitalise on the immunoprecipitation and mass spectrometry expertise of Dr Tania Auchynnikava from the Allshire lab.
1.7. Project aims

The objective of this study was to use TALEs to identify proteins enriched at putative heterochromatin sequences in the *Trypanosoma brucei* genome. To address this question, the following aims had to be achieved:

1. Generate TALEs to bind different repetitive sequences in the *T. brucei* genome, and demonstrate their ability to bind those sequences.

2. Use the TALEs generated in aim 1 to purify and identify proteins associated with those repetitive sequences.

3. Characterise the role(s) of any novel protein(s) identified at those repetitive sequences.
Chapter 2: Materials & Methods

2.1 Molecular cloning and standard DNA methods

2.1.1 Small scale plasmid preparations

Single colonies from bacterial transformation plates were grown in 4 ml LB supplemented with the relevant antibiotic(s) at 37°C overnight with shaking. Plasmids were isolated using a Monarch Plasmid Miniprep Kit (NEB) following the manufacturers’ instructions. Plasmid DNA concentration was quantified using a NanoDrop Spectrophotometer (Thermo Fisher Scientific). Plasmids were stored at -20°C.

2.1.2 Ethanol precipitation of DNA

Ethanol precipitation was used to concentrate or remove impurities from various DNA samples. 0.1 volumes of 3 M sodium acetate pH 5.2 and 2.5-3 volumes of ice-cold 100% ethanol were added to the DNA. Samples were precipitated at -80°C for 1 h or at -20°C overnight. Samples were centrifuged at 16,000 g for 15 min at 4°C and supernatant was discarded. DNA was washed once with ice-cold 70% ethanol. Samples were centrifuged again at 16,000 g for 15 min at 4°C, supernatant was discarded, and the DNA pellet was air-dried for 5 min at room temperature. DNA was resuspended in the desired volume of TE buffer (0.1 M Tris-Cl pH 8.0) or dH2O. DNA concentration was measured using a NanoDrop (Thermo Fisher Scientific).

2.1.3 Polymerase Chain Reaction (PCR)

PCR amplifications of DNA were carried out in 0.2 ml PCR tubes (STARLAB) in an MJ Research PTC-225 Gradient Thermal Cycler (Marshall Scientific). PCR reactions where the product would be utilised for cloning were performed using Phusion High-Fidelity DNA Polymerase (New England Biolabs (NEB)) in most cases. All primers (Appendix A) were purchased from Integrated DNA Technologies (IDT). PCR amplifications used to confirm the presence/absence of a DNA fragment in a sample were carried out with Taq DNA Polymerase (Roche). All PCR reactions were set up following the manufacturer’s recommendations.
2.1.4 DNA Agarose gel electrophoresis

Agarose powder was dissolved by boiling in 1X TBE (89 mM Tris-HCl, 89 mM Boric acid, 2 mM EDTA) or TAE buffer (40 mM Tris-HCl, 20 mM acetic acid, 1 mM EDTA). TAE was used when fragments were to be excised and purified from the gel for cloning or transfection. TBE was used when gels were used to detect the presence or absence of particular bands. The solution was cooled to approximately 60°C before adding 0.3 µg/ml ethidium bromide (Sigma-Aldrich). Gels were poured into a mould with ends sealed by masking tape. Gel combs were used to make appropriate wells for sample loading. After the gel had solidified, the tape and comb were removed and it was transferred to a gel tank filled with 1X TBE or 1X TAE buffer, as appropriate. DNA samples were resuspended in Gel Loading Dye (NEB) and loaded in the gel wells alongside either DNA ladders made in-house or a commercial 1 kb ladder (Promega) that allowed estimation of the fragment sizes. Gels were run at 100-120 V for time sufficient to resolve the desired fragments (30-60 minutes in most cases).

2.1.5 DNA gel extraction

DNA fragments which required gel purification were purified from the agarose gel using the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel) following the manufacturer’s instructions. The resulting sample concentration was measured by NanoDrop (Thermo Fisher Scientific). Where necessary, eluted DNA was ethanol precipitated.

2.1.6 Plasmid restriction digests

Plasmids were digested with restriction enzymes from NEB following the manufacturer’s instructions. In most cases, digestions were performed in CutSmart Buffer (NEB) at 37°C for 1 h using high fidelity (HF) enzymes.

2.1.7 Bacterial transformations

Stable competent E. coli cells (NEB) were stored at -80°C and thawed on ice immediately before transformation. Approximately 1-2 µl of a plasmid ligation of Gibson Assembly reactions were added to 50 µl competent cells and incubated for 30 min on ice. Cells were then heat-shocked at 42°C for 30 s and placed back on ice for 5 min. 950 µl of Stable outgrowth media (NEB) was added to the cells which were then incubated at 37°C for 1 hour with shaking. Finally, the cells were plated onto LB selective plates containing appropriate antibiotics and
incubated at or 30°C for 24 hours. Plates were then removed from the incubator and the success of the transformations was assessed by the presence/absence of colonies.

2.1.8 Colony PCR

Bacterial colonies were picked using a sterile pipette tip and resuspend in 50 µL sterile dH₂O. 2.5 µL of this bacterial suspension was added to 22.5 µL of the PCR mastermix (Roche reagents) outlined below.

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>dH₂O</td>
<td>18.375 µL</td>
</tr>
<tr>
<td>PCR buffer + Mg</td>
<td>2.5 µL</td>
</tr>
<tr>
<td>10 mM dNTPs</td>
<td>0.5 µL</td>
</tr>
<tr>
<td>Forward primer (10 µM)</td>
<td>0.5 µL</td>
</tr>
<tr>
<td>Reverse primer (10 µM)</td>
<td>0.5 µL</td>
</tr>
<tr>
<td>HotStart Taq polymerase</td>
<td>0.125 µL</td>
</tr>
</tbody>
</table>

2.1.9 Gibson Assembly

All Gibson Assembly primers were designed using SnapGene so that following PCR, 15-25 bp homology between adjacent fragments would be generated. Fragments were generated by PCR using Phusion polymerase (NEB). Fragment concentrations were measured using a Nanodrop, and Gibson assembly was set up in a 1:3 (vector:insert1:insert2) molar ratio, using 75 ng vector DNA. Fragments were added to 10 μl Gibson Assembly MasterMix (NEB) in a 20 μL reaction and incubated for 30 min at 50°C. 2 μl Gibson Assembly product was transformed into 50 μl of Stable Competent E. coli (NEB).

2.1.10 Sanger sequencing

BigDye Terminator Cycle sequencing kit (Applied Biosystems – Thermo Scientific) was used for Sanger sequencing. Reactions were set up as follows.
Reagent Volume:

- 150-300 ng plasmid DNA
- Primer (3.2 μM) 1 μL
- Big Dye Mix 2 μL
- H₂O Up to 10 μL

PCR programme:

- 1x 95°C for 5 min
- 95°C for 30 sec
- 55°C for 15 sec
- 64°C for 4 min x25

Samples were then submitted to Edinburgh Genomics for sequencing.

2.1.11 TALEs target sequence design

All TALE target sites were designed to bind 15 bp sequences which began after a thymine, as required for the Musunuru/Cowan lab TALEN kit (Ding et al., 2013). Minichromosomal 177 bp repeat sequences were obtained from Bill Wickstead (University of Nottingham). A consensus sequence of the conserved N-terminal 79 bp of Ingi transposable elements was obtained from Frederic Bringaud (University of Bordeaux). CIR147 repeat sequences were obtained from publications (Obado et al., 2007a; Patrick et al., 2009b). TALE target sequence preferences are not well understood. Therefore, candidate 15 bp sequences were picked from the middle of the target sequences. BLAST searches were used to identify exact matches in the TREU927 reference genome. Candidate sequences with more than one exact match were discarded. Candidate sequences with only one exact match were selected.

2.2 Standard Trypanosome methods

2.2.1 Trypanosome cell culture

Trypanosoma brucei brucei Lister 427 bloodstream form monomorphic cells were used for all experiments in this project. Cell lines were grown at 37°C and 5% CO₂ in HMI-9 medium supplemented with 10% Fetal Calf Serum (Gibco) passed through a 0.22 μm filter, 1%
Penicillin-Streptomycin (Gibco) and selective drug(s) when required (Hirumi & Hirumi, 1989). Cell cultures were maintained below 3 x 10^6 cells/ml. Glycerol stocks were made by mixing 500 µl trypanosome cell culture with 500 µl of HMI-9 containing 14% glycerol (7% glycerol final concentration). Cryovials were then incubated overnight at -80°C in a polystyrene box before being moved to a box in the -80°C freezer. Glycerol stocks were thawed by placing at 37°C and adding the contents of the cryovial to 5ml HMI-9 medium supplemented with 30% FCS without selective drugs. Cells were then incubated for 2-3 days at 37°C and 5% CO₂ until fully recovered. Selective drug(s) (if required) were only added after this point.

<table>
<thead>
<tr>
<th>Name</th>
<th>Concentration (µg/ml)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blasticidin</td>
<td>5</td>
</tr>
<tr>
<td>Phleomycin</td>
<td>2.5</td>
</tr>
<tr>
<td>G418</td>
<td>2.5</td>
</tr>
<tr>
<td>Hygromycin</td>
<td>1</td>
</tr>
<tr>
<td>Puromycin</td>
<td>0.5</td>
</tr>
</tbody>
</table>

2.2.2 Trypanosome genomic DNA extraction

Approximately 1 x 10^8 cells were harvested by centrifugation at 1000 g, 10 min. Cells were then washed once with 500 µl 1X PBS (137 mM NaCl, 2.7 mM KCl, 10 mM Na₂HPO₄, KH₂PO₄) and pelleted by centrifugation at 1000 g, 10 min. Genomic DNA was then extracted using the DNeasy Blood & Tissue kit (Qiagen) following the manufacturer’s instructions. DNA samples were then ethanol precipitated and resuspended in TE buffer, 10 mM Tris-HCl pH 8, or ddH₂O depending on the application. DNA concentrations were quantified using a NanoDrop (Thermo Fisher Scientific).

2.2.3 Trypanosome protein extraction for Western blotting

Approximately 4 x 10^7 cells were harvested by centrifugation at 1000 g, 10 min. Cells were the washed once with 500 µl 1X PBS and resuspended in 75 µl 1X PBS + 25 µl 4X NuPAGE LDS Sample Buffer (Thermo Fisher Scientific) to give a final concentration of 4x10^6 cells per 10 ul.
Samples were then boiled at 95°C for 5 min to ensure cells were dead before removal from the CAT3 facility. In some cases samples were then subjected to sonication using a Diagenode Bioruptor for 10 cycles, 30s ON/30s OFF at 4°C on high setting to shear the DNA and reduce the viscosity to aid loading on gels.

2.2.4 Trypanosome transfections

3-5 x 10^7 cells were harvested per transfection by centrifugation at 1000 g, 10 min. Cells were washed once with 5 ml media (HMI-9 or SDM-79) and pelleted again by centrifugation at 1000 g, 10 min before resuspending in 100 µl ice-cold TbBSF transfection buffer (Schumann Burkard et al., 2011), and transferring to an electroporation cuvette (Ingenio). 10-20 µl of DNA for transfection containing 1-5 µg DNA was added to the cuvette. Cells were mixed with DNA by pipetting up and down, and electroporated in the Amaza Nucleofector II (Lonza) using the X-001 programme for bloodstream cells (Burkard et al., 2007). A “no DNA” mock transfection was always performed in parallel as a negative control. Electroporated bloodstream cells were added to 30 ml HMI-9 medium and two 10-fold serial dilutions were performed. 1 ml of transfected cells were plated per well on 24-well plates (1 plate per serial dilution) and incubated at 37°C and 5% CO_2 for a minimum of 6 hours before adding 1 ml media containing 2X concentration of the required selective drug per well. A positive control was also performed by adding media containing no selective drug to 8 wells of the control transfection plate. Cells were only judged to be true transfectants once the mock-transfected cells were dead, after which they were moved to 6-well plates.

2.2.5 N-terminal protein tagging

Proteins were tagged N-terminally at the endogenous locus with YFP using the pPOTv4 plasmid (Dean et al., 2015). Homologous recombination templates were generated using fusion PCR. An approximately 1700 bp region of the pPOTv4 plasmid containing a blasticidin resistance cassette and a YFP tag was amplified. Next, 300-500 bp fragments homologous to the end of the 5' UTR of each gene and the beginning of the coding sequence of each gene were amplified. These 300-500 bp fragments contained short overlapping sequences facilitating annealing to either end of the pPOTv4 fragment. Fusion PCR products were transfected into bloodstream form parasites as described above. Cell lines obtained after drug selection were tested for fusion protein expression by Western blotting and for correct integration into the genome by PCR.
2.3 Standard protein methods

2.3.1 Western blotting

Whole cell protein extracts were separated by polyacrylamide gel electrophoresis (PAGE). Samples were run on NuPAGE Bis-Tris Mini Gels (Thermo Fisher Scientific) for 30-60 min in a Mini Gel Tank (Thermo Fisher Scientific) in 1X NuPAGE MES Running Buffer at 200 V. Following PAGE, proteins were transferred onto nitrocellulose membranes in a Mini Blot Module (Thermo Fisher Scientific) at 20 V for 1 h. Prior to blocking, membranes were stained with Ponceau S (Sigma-Aldrich) to assess efficiency of protein transfer. Membranes were then blocked in 5% milk in PBS-T (PBS + 0.05% tween) at room temperature for 30 min, shaking. After blocking, membranes were incubated with mouse anti-GFP (Roche) (1:1000 in 5% milk in PBS-T) at 4°C overnight on a lab rocker. The next morning, membranes were washed 3 x 10 min with PBS-T on a lab shaker and incubated with Horseradish peroxidase (HRP)-conjugated anti-mouse secondary antibody (1:2500 in 5% milk in PBS-T) at room temperature for 1 h, shaking. Membranes were washed again 3 x 10 min with PBS-T and incubated with Amersham ECL Prime Western Blotting Detection Reagent (GE Healthcare) following manufacturer’s instructions. Proteins were visualised using Amersham Hyperfilm ECL (GE Healthcare). Films were then scanned onto the lab computer and labelled in PowerPoint.

2.3.2 Silver Staining

Protein extracts were separated by polyacrylamide gel electrophoresis (PAGE) on NuPAGE Bis-Tris Mini Gels (Thermo Fisher Scientific). Samples were run for 30-45 min in a Mini Gel Tank (Thermo Fisher Scientific) in 1X NuPAGE MES Running Buffer at 160 V. Silver staining of SDS-PAGE gels was then performed using the SilverQuest Silver Staining Kit (Thermo Fisher Scientific). Silver stained gels were placed in a plastic pocket and scanned onto the lab computer before labelling in PowerPoint.

2.4 Immunofluorescence microscopy

$1 \times 10^6$ parasites were harvested by centrifugation at 1900 g for 6 min. Supernatant was discarded, cells were washed in 500 µl cold 1X PBS and centrifuged at 1900 g for 6 min. Cells were then fixed by resuspending in 125 µl cold 1X PBS + 125 µl 8% paraformaldehyde and
incubating on ice for 10 min, flicking the tube periodically. Fixed samples were centrifuged at 1900 g for 6 min, supernatant discarded, and cells resuspended in 130 µl 0.1 M glycine (in 1X PBS) before incubating overnight at 4°C. The following day, cells were centrifuged at 1900 g for 6 min, supernatant discarded, and resuspended in 200 µl 1X PBS. Wells were drawn on poly-lysine slides using a hydrophobic pen and 10 µl cells were added to each well. Slides were incubated for 1 h at room temperature in a humid chamber to allow parasites to stick to the wells. Excess PBS was removed using a vacuum pump and cells were permeabilised by adding 0.1% Triton X-100 (in 1X PBS) to the wells for 2 min. Slides were washed with 1X PBS for 5 min on a lab rocker. Slides were then blocked in 2% BSA (in 1X PBS) for 45 min at room temperature inside a humid chamber. Excess blocking solution was removed using a vacuum pump and the cells were incubated with a 1:250 dilution (in 2% BSA) of rabbit anti-GFP (Thermo Fisher Scientific) primary antibody, for 45 min at room temperature inside a humid chamber. The slides were then washed 3 x 5 min with 1X PBS on a lab rocker. The cells were incubated with 1:500 dilution (in 2% BSA) of Alexa Fluor-568 or -488 conjugated goat anti-rabbit (Thermo Fisher Scientific) secondary antibody for 45 min at room temperature inside a humid chamber. From here on, the slides were kept in the dark as much as possible to prevent photobleaching of the secondary antibody signal. Slides were washed 2 x 5 min with 1X PBS on a lab rocker, incubated with 1µg/ml 4',6-diamidino-2-phenylindole (DAPI) for 2 min and washed again 2 x 5 min with 1X PBS. Coverslips were mounted onto the slides by adding a few drops of mowiol + DABCO (Appendix B) and sealing with nail polish. Slides were left to dry overnight at room temperature in the dark before visualisation under the microscope (Zeiss Axio Imager).

2.5 Chromatin Immunoprecipitation (ChIP) methods

2.5.1 Chromatin Immunoprecipitation (ChIP)

3-5 x 10^8 cells were harvested per ChIP sample by centrifugation at 1500 g, 10 min. Supernatant was discarded and cell pellets were resuspended in 40 ml HMI-9 media at room temperature. Cells were then fixed by adding 4 ml formaldehyde solution (Appendix B) to each sample (1% final formaldehyde concentration) and incubating for 20 min at room temperature, inverting the tube every 3-5 min. Fixation was quenched by the addition of 2.5 ml of room temperature 2 M glycine, followed by centrifugation at 1500 g for 20 min at 4°C. Supernatant was discarded and cell pellets were resuspended in 30 ml cold 1X PBS by pipetting up and
down and centrifuged at 3500 g for 20 min at 4°C. Supernatant was discarded and cell pellets were resuspended by vigorous vortexing in 2.5 ml Lysis buffer 1 (Appendix B), supplemented just before use with EDTA-free protease inhibitor tablets (Roche). Samples were incubated at 4°C for 10 min on a tube rotator then centrifuged at 3500 g for 20 min at 4°C. Supernatant was discarded and pellets were resuspended by vigorous vortexing in 2.5 ml Lysis buffer 2 (Appendix B), supplemented just before use with EDTA-free protease inhibitor tablets (Roche) and incubated at room temperature for 10 min on a lab rocker. Samples were then centrifuged at 3500 g for 20 min at 4°C. Supernatant was discarded and pellets were resuspended in 300 µl Lysis buffer 3 (Appendix B), supplemented just before use with 20% SDS (0.2% final concentration) and EDTA-free protease inhibitor tablets (Roche), and transferred to 1.5 ml microfuge tubes. Chromatin was fragmented by sonication in a Bioruptor (Diagenode) sonicator at 4°C for 30 cycles (30 s ON / 30 s OFF per cycle, high setting). After sonication, 900 µl Lysis buffer 3, supplemented with protease inhibitors but not SDS, was added to each sample. The samples were then centrifuged at 16,000 g for 10 min at 4°C and the resulting supernatants were transferred to new 1.5 ml microfuge tubes. At this stage, 10 µl was taken from each sample and frozen overnight at -80°C – this was an input sample. The remaining sample was subjected to immunoprecipitation overnight rotating at 4°C with 10 µg rabbit anti-GFP antibody (Thermo Fisher Scientific) and 50 µl Protein G Dynabeads (washed twice with Lysis buffer 3 prior to use). The following day, beads were collected using a magnet and the supernatant was discarded. Beads were first washed with 900 µl Lysis buffer 3 on a magnet. Beads were then washed with 900 µl Wash buffer 1 (Appendix B) rotating for 10 min at 4°C, followed by a wash with 900 µl Wash buffer 2 (Appendix B) rotating for 10 min at room temperature. Beads were then resuspended in 900 µl TE buffer (Appendix B) and transferred to a new microfuge tube to remove unspecific DNA from the sample bound to the previous tube. 200 µl Elution buffer (Appendix B) was added to the beads to remove immunoprecipitated material. 200 µl Elution buffer was also added to the input samples previously placed at -80°C. All samples were then incubated overnight at 65°C with shaking at 1000 rpm to aid elution and reverse crosslinking of DNA and proteins. The next day, the beads were separated from the eluate on a magnet and discarded. 10 µl DNAse-free RNAse (500 µg/ml) (Roche) was added per eluate and incubated at 37°C for 2 hours shaking at 1000 rpm. 8 µl Proteinase K (10 mg/ml) was then added per sample and incubated at 55°C for 2 hours shaking at 1000 rpm. The DNA was purified using a QIAquick PCR Purification Kit (Qiagen) and eluted in 60 µl buffer EB.
2.5.2 Quantitative PCR (qPCR)

ChIP samples were diluted as follows in dH$_2$O to give comparable amounts of DNA: inputs diluted 1:100; IPs diluted 1:20. All qPCRs were performed in 384-well plates (Roche) with 10 μL of PCR reaction per well. Reagent volumes for each well were as follows:

- 5 μL LightCycler 480 SYBR Green Master Mix (Roche)
- 0.5 μL Forward primer
- 0.5 μL Reverse primer
- 1 μL H$_2$O
- 3 μL Template DNA

Different qPCR cycling conditions were used for reactions using the telomere repeat primers (PCR programme 2). All other qPCRs used the same conditions (PCR programme 1).

PCR programme 1:
- 1x 95°C for 5 min
- 95°C for 5 sec  x40 cycles
- 55°C for 12 sec
- 72°C for 20 sec

PCR programme 2:
- 1x 95°C for 2 min
- 95°C for 20 sec  x40 cycles
- 60°C for 30 sec
- 72°C for 30 sec

Data were analysed using the LightCycler 480 Software (Roche). All ChIP enrichments were calculated as % DNA immunoprecipitated at the locus of interest.
2.6 Immunoprecipitation followed by Mass Spectrometry Analysis

2.6.1 Crosslinking of antibody to dynabeads

All crosslinking steps were performed at room temperature. 1 ml Protein G Dynabeads (Thermo Fisher Scientific) were washed 2 x in 1 ml Lysis buffer (50 mM Tris-Cl pH 8, 150 mM KCl, 0.2% NP-40). One vial of mouse anti-GFP antibody (Roche) was resuspended in 1 ml Lysis buffer by pipetting up and down and added to the beads. Beads were rotated for 30 min to allow antibody to bind to the protein G ligand. Beads were washed 2 x in 1 ml Lysis buffer by pipetting up and down and added to 10 ml Borate buffer (40 mM Boric acid, 40 mM sodium tetraborate decahydrate) supplemented with 50 mg dimethyl pimelimidate (DMP) (Thermo Fisher Scientific). Only DMP opened within one month of the experiment was used. Beads were rotated for 30 min, washed once with 10 ml Lysis buffer, and finally resuspended in 2 ml Lysis buffer. Antibody-crosslinked beads were stored at 4°C for several months and used as required.

2.6.2 Protein immunoprecipitation from whole cell lysates

3-5 x 10^8 cells were harvested per immunoprecipitation by centrifugation at 1500 g for 10 min. Cells were washed once with cold PBS and centrifuged at 1500 g for 10 min. Cell pellets were resuspended in 200 µl cold Lysis buffer (50 mM Tris-Cl pH 8, 150 mM KCl, 0.2% NP-40), supplemented just before use with EDTA-free protease inhibitor tablets (Roche). Samples were sonicated in a Bioruptor (Diagenode) sonicator at 4°C for 3 cycles (12 s ON / 60 s OFFy, high setting). Lysates were then clarified by centrifugation at 16000 g for 5 min at 4°C. The soluble fraction (supernatant) was transferred to a new tube and used for the immunoprecipitation. Pellets were discarded. 120 µl of antibody crosslinked beads were added per sample, and samples were rotated for 1 h at 4°C. Samples were then washed 3x with 1 ml Lysis buffer (supplemented just before use with EDTA-free protease inhibitor tablets (Roche)) by gently pipetting up and down. Immunoprecipitated proteins were then eluted using 60 µl of RapiGest surfactant (Waters) (0.1% in 100 mM Tris-HCl pH 8.0) at 55°C for 15 min. Only RapiGest resuspended within one week of the experiment was used. The majority of the sample (45-50 µl) was frozen immediately at -80°C, with a portion (10-15 µl) kept for analysis by silver stain and western blotting if necessary.
2.6.3 FASP Digestion

DTT was added to each sample to a final concentration of 25 mM. The samples were incubated at 95°C for 5 min then allowed to cool to room temperature. Samples were mixed with 100 µl 8 M Urea in 0.1 M Tris-HCl pH 8.0 and loaded onto Vivakon spin column 30K cartridges. Columns were centrifuged at 13,500 g for 10-15 min. 100 µl 0.05 M IAA (in 8 M Urea in 0.1 M Tris-HCl pH 8.0) was added to each column before shaking at 600 rpm for 1 min at room temperature and incubated for 20 min at room temperature in the dark. Columns were centrifuged at 13,500 g for 10-15 min. 100 µl 8 M Urea in 0.1 M Tris-HCl pH 8.0 was added to each column before centrifuging at 13,500 g for 10-15 min. Columns were then washed twice by adding 100 µl ABC buffer (50 mM Ammonium bicarbonate) and centrifuging at 13,500 g for 10-15 min. ABC buffer was always prepared fresh using MS grade water. Each sample was digested with 0.5 µg MS Grade Pierce Trypsin Protease (Thermo Fisher Scientific) in 100 µl ABC buffer overnight at 37°C.

2.6.4 Stage Tip Preparation

After overnight trypsin digestion, the columns were centrifuged at 13,500 g for 15 min, then washed with 100 µl ABC buffer and centrifuged again at 13,500 g for 15 min. Samples were acidified to pH 2-3 (assessed with pH indicator paper) with ~ 10 µl 10% TFA. Stage tips (Rappsilber et al., 2007) were prepared as follows. Three C-18 paper discs were cut from a membrane using a modified syringe and placed in a p200 pipette tip. Stage tips were conditioned by sequential passing through centrifugation at 1000 g of 50 µl methanol, 50 µl 80% ACN in 0.1% TFA, and 50 µl 0.1% TFA, in this order. The samples containing tryptic peptides were then loaded onto the tips at 1000 g for 15 min. Stage tips were washed with 75 µl 0.1% TFA and frozen at -20°C until they were to be eluted. Elution of peptides from stage tips and injection into the Orbitrap Fusion Lumos Trivid mass spectrometer (ThermoFisher Scientific) were performed by Dr Tania Auchynnikava.

2.6.5 Mass spectrometry data analysis – raw data

Peptides were resuspended in 0.1%TFA for LC MS/MS. Peptides were separated using RSLC Ultimate3000 system (Thermo Scientific) fitted with an EasySpray column (50 cm; Thermo Scientific) utilising 2-40-95% nonlinear gradients with solvent A (0.1% formic acid) and solvent B (80% acetonitrile in 0.1% formic acid). The EasySpray column was directly coupled to an
Orbitrap Fusion Lumos (Thermo Scientific) operated in DDA mode. “TopSpeed“ mode was used with 3 s cycles with standard settings to maximize identification rates: MS1 scan range - 350-1500 mz, RF lens 30%, AGC target 4.0e5 with intensity threshold 5.0e3 , filling time 50 ms and resolution 60000, monoisotopic precursor selection and filter for charge states 2-5. HCD (27%) was selected as fragmentation mode. MS2 scans were performed using Ion Trap mass analyser operated in rapid mode with AGC set to 2.0e4 and filling time to 50ms. The resulting shot-gun data were processed using Maxquant 1.3.8. and visualized using Perseus 1.6.0.2 (Tyanova et al., 2016).

2.7 Southern blotting

2 µg of genomic DNA per sample was digested with MboI (NEB) and AluI (NEB) in CutSmart buffer (NEB) at 37°C overnight following the manufacturer’s instructions. Samples were then run on a 0.5% agarose gel for approximately 21 hours at 25 V. A DNA ladder comprising a 1:1 combination of Lambda DNA/HindIII ladder and 1 kb ladder (Promega) was run either side of the samples. The gel was then imaged on a UV gel dock with a ruler placed alongside to facilitate labelling the molecular weight marker at the end of the experiment. The gel was depurinated in 0.25 M HCl for 30 min with shaking. Depurination solution was discarded and denaturation solution (0.5 M NaOH/1.5 M NaCl) was added and the gel was incubated for 30 min, shaking. Denaturation solution was discarded and neutralization solution (1.5 M NaCl/0.5 M Tris-HCl, pH 8) was added and the gel was incubated for 30 min, shaking. DNA was then transferred to a nylon membrane (Amersham) by capillary transfer in 20X SSC (3M NaCl, 0.3 M Tri-Sodium Citrate pH 7). After transfer, gel lanes were marked with a pencil and the gel discarded. The DNA was crosslinked to the membrane in a UV-crosslinker at a setting of 120 mJ cm\(^{-2}\) on both sides then patted dry on Whatman paper. The membrane was placed in a hybridisation tube containing 25 ml of pre-hybridisation solution (5× SSC, 0.1% Sarkosyl, 0.04% SDS) and incubated in a hybridisation oven for 2 hours at 65°C. While incubating, the (TTAGGG)\(_3\) probe was labelled with DIG using the DIG Oligonucleotide 3′-End Labelling Kit (Roche). The pre-hybridisation solution was discarded and the membrane was incubated in 25 ml hybridisation solution containing 20 µl of DIG-labelled telomere probe overnight at 65 °C in a rotating hybridisation oven. The next day, the hybridisation solution was discarded and the membrane was washed 3x in wash buffer 1 (2× SSC/0.1% (wt/vol) SDS) for 15 min, shaking. Membranes were then washed in 2× SSC for 15 min, shaking. The membrane was then placed in a heat-sealable pouch with 60 ml of blocking buffer (1× maleic acid buffer/1% (wt/vol) blocking reagent) and incubated for 30 min at room temperature, shaking. Blocking buffer was then drained and 40 ml of fresh blocking buffer containing 2 µl of anti-DIG-AP
antibody (Roche) was added and incubated for 30 min at room temperature, shaking. Antibody solution was discarded and the membrane was washed 2x in wash buffer 2 (1× maleic acid buffer/0.3% (vol/vol) Tween 20) for 15 min, shaking. The membrane was then equilibrated in 50 ml AP buffer (0.1 M Tris-HCl, pH 8/0.1 M NaCl) for 2 min before placing it in a fresh heat-sealable pouch with 4 ml AP buffer containing 40 µl CDP-Star (Roche). A 10 ml serological pipette was rolled over the sealed pouch for 5 min. DNA was then visualised using Amersham Hyperfilm ECL (GE Healthcare). Films were then scanned onto the lab computer and labelled in PowerPoint.
Chapter 3: Design, Cloning, and Characterisation of Transcription Activator-like Effectors (TALEs) Designed to Bind Repetitive Regions of the \textit{T. brucei} Genome

3.1 Introduction

Before any locus-specific proteomics could be performed, synthetic TALE proteins needed to be designed and expressed in \textit{T. brucei}, and tested for their ability to bind their target sequences with specificity. Firstly, TALE fusion protein constructs that could be utilised for proteomics but also visualisation were designed, in conjunction with a suitable trypanosome expression vector for these exogenous proteins. After confirming their expression and any growth phenotypes associated with TALE expression, microscopic visualisation was used to provide an indication as to their ability to bind the expected target sequences. Finally, chromatin immunoprecipitation (ChIP) experiments were performed to verify the binding of the different TALE proteins to their different target sequences.

3.2 Selection of TALEs target sequences for binding different repetitive regions of the \textit{T. brucei} genome

The Musunuru/Cowan lab TALEN cloning kit (https://www.addgene.org/kits/musunuru-cowan-talen/) was used to generate TALE proteins. This kit allows TALE proteins to be made that are capable of binding any 15 bp DNA sequence as long as it is immediately preceded by a thymidine (Ding et al., 2013). Therefore, 15 bp target sequences needed to be selected for each of the four repetitive elements I planned to target, namely telomeric TTAGGG repeats, centromeric chromosome internal repeats of 147 bp (CIR147), ingi transposable elements, and minichromosomal 177 bp repeats. First, the sequences of the four different classes of repeats were obtained. 15 bp sequences unique to each of those repeats were then identified by performing BLAST searches against the TREU 927 reference genome. Sequences with no exact matches elsewhere in the genome were selected. In the case of the Telomere-TALE, only one possible sequence was available within the constraints due to the repeat length being only 6 bp and the requirement for target sequences to begin after a thymidine. More detail on target sequence selection is available in Section 2.1.11.
Table 3.1. 15 base pair target sequences of the four TALE constructs.

<table>
<thead>
<tr>
<th>Sequence Name</th>
<th>15 bp target sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Telomeric (TTAGGG)$_n$ repeats</td>
<td>AGGGTTAGGGTTAGG</td>
</tr>
<tr>
<td>CIR147 repeats</td>
<td>GCCGACCACCTCAAC</td>
</tr>
<tr>
<td>Ingi transposable elements</td>
<td>GCAAGCTTGTGCATG</td>
</tr>
<tr>
<td>Minichromosomal 177 bp repeats</td>
<td>GCAGCTTAACGCTA</td>
</tr>
</tbody>
</table>

3.3 Cloning of the TALE expression vector using multiple cloning methodologies

In order to generate TALE proteins capable of binding different repetitive DNA sequences in the trypanosome genome, we first needed to consider: 1) the structure of the fusion proteins to be expressed in trypanosomes; 2) whether the construct would be inducible or constitutively expressed; and 3) how to achieve appropriate protein levels.

The final fusion protein would require a nuclear localisation signal (NLS), the TALE DNA binding domain, and at least one epitope tag suitable for use in immunoblotting, immunofluorescence, and immunoprecipitation in both native and denaturing conditions. The NLS from the *T. brucei* La protein was utilised, as this had been characterised and shown to be capable of localising GFP to the nucleus when placed at its N-terminus (Marchetti et al., 2000). Various epitope tags were available from the suite of pPOT trypanosome tagging vectors (Dean et al., 2015), and the 3xTy-YFP-3xTy tag was chosen to allow the use of both anti-GFP and anti-BB2 antibodies in follow-up analyses. Since the proteins were exogenous and would be expressed from a plasmid, there was no obvious reason to tag the construct at the N-terminus over the C-terminus, as the 3′UTR would not be disrupted as can occur with endogenous locus tagging. Therefore, it was decided that the final fusion protein would be structured as follows: La protein NLS-TALE DNA binding domain-3xTy-YFP-3xTy.

There are arguments for and against inducible or constitutive expression of the TALE proteins. Expression of synthetic DNA-binding proteins might be harmful to the cell, particularly when binding regions such as centromeres, where TALE binding could potentially disrupt kinetochore formation or function, negatively affecting cell division. If this were the case, then inducible expression might be preferable, as this would limit any potential negative effects of
TALE expression to a narrower time window, for example just prior to the use of cells for analysis, but not during the primary growth phase. However, inducible expression systems are often difficult to regulate precisely, and low-level expression in the absence of inducer is commonly observed, as is protein overexpression. Further, many of the available inducible constructs are designed to integrate into the minichromosomal 177 bp repeats, which were one of the target sequences for the TALE approach. Using one of those expression vectors may have lead to alterations in minichromosome protein composition, for example by recruiting RNA polymerase and other transcription-associated proteins. Due to these concerns, it was decided not to utilise an inducible system. On the other hand, low-level, constitutive expression might be possible through use of particular 3’UTRs. If expression levels were low enough, any negative impacts of TALE expression might be avoidable. The primary concern would then be whether sufficiently high levels were present to facilitate enrichment of the locus compared with background for use in proteomics experiments. Vectors allowing integration into the beta-tubulin locus were available in the Matthews lab, and so one of these, the CATpHD449 plasmid, was chosen for TALE fusion protein expression (Mayho et al., 2006). This plasmid was a derivative of the pHD 449 plasmid published by Dr Christine Clayton’s lab at the University of Heidelberg (Biebinger et al., 1997).

In order to achieve low-level expression, we took advantage of the fact that protein expression levels are strongly influenced by 3’UTRs in trypanosomes (C. Clayton, 2013; Kolev et al., 2014; Kramer & Carrington, 2011). Thus, we used the 3’UTR from a protein known to be expressed at very low levels in proliferating bloodstream forms for TALE fusion protein expression. The 3’UTR chosen was a heavily truncated version of the PAD1 3’UTR, with bases 1-354 deleted from the full length PAD1 3’UTR. This Δ1-354 PAD1 3’UTR was shown to give 7.13% of the protein levels seen with a truncated aldolase 3’UTR in a chloramphenicol acetyltransferase reporter assay in monomorphic 427 parasites (Paula McGregor, PhD thesis). This 3’UTR was used as a starting point, in the knowledge that if lower expression levels were required, other 3’UTRs giving lower expression generated for use in that thesis were available. Codon optimisation is also known to influence protein expression levels in trypanosomes (Jeacock et al., 2018), but this approach would have been impractical in this instance, as new, codon-altered versions of the TALE cloning plasmids would have been extremely expensive to generate.
3.3.1 Using annealed oligonucleotide cloning to insert the La protein NLS into the CATpHD449 Δ1-354 PAD1 3'UTR plasmid

The CATpHD449 Δ1-354 PAD1 3'UTR plasmid contained HindIII and Ndel restriction sites between the Actin 5'UTR and the beginning of the chloramphenicol resistance gene. Digestion of the plasmid with these enzymes would allow for an insertion between the Actin 5'UTR and any subsequent open reading frame. Therefore, oligonucleotides encoding the La protein NLS with overhangs complementary to those generated by HindIII and Ndel digestion were designed. These oligonucleotides were annealed, phosphorylated, and ligated into the linearised CATpHD449 Δ1-354 PAD1 3'UTR plasmid. Transformants were subjected to small scale plasmid preparation, followed by Sanger sequencing to verify correct integration and sequence identity. A schematic of the construction process is presented in Figure 3.1 below.
Figure 3.1. Inserting the La protein NLS into CATpHD449 PAD1 3′UTR Δ1-354 by annealed oligonucleotide cloning. (A) Schematic of the cloning process. Oligonucleotides encoding the La protein NLS were annealed and ligated between the HindIII and NdeI sites (labelled) of CATpHD449 PAD1 3′UTR Δ1-354. Oligonucleotide sequences can be found in Appendix A. (B) Plasmid map of the final CATpHD449 PAD1 3′UTR Δ1-354 NLS plasmid. Beta Tubulin — β-tubulin intergenic region targeting sequence. CmR — chloramphenicol-resistance gene. BleoR — bleomycin-resistance gene. AmpR — Ampicillin-resistance gene. Ori — origin of replication. Images taken from SnapGene.
3.3.2 Cloning the TALE domain using the Musunuru/Cowan Lab TALEN kit

As discussed in Section 1.5, TALE proteins bind their target DNA sequence using multiple DNA-binding motifs placed in series, each motif containing the necessary repeat variable di-residues (RVD’s) to bind a particular DNA base. The plasmid library contains inserts, termed multimers, which are capable of binding specific 4- and 3-bp sequences. These multimers are joined together using the Golden Gate Assembly cloning methodology to generate a full length TALE domain. This methodology utilises type IIS restriction enzymes, which cut DNA outside of their recognition sequence, allowing the ordered assembly of multiple inserts into a single vector simultaneously. The repetitive nature of the central DNA-binding domain of TALE proteins prohibits the use of PCR-based methods such as Gibson Assembly for this cloning step.

In order to generate a TALE domain that binds the 15 bp sequence TCAGTCAGTCAGTCA, for example, one must digest the following 4 plasmids from the kit: I-TACG, II-TCAG, III-TCAG, and IV-TCA. After digestion with the type IIS enzyme BsmBI, the overhangs generated are specific to each multimer, meaning that there is only one possible orientation in which they can be ligated together. Digestion of the pTAL_GFP plasmid from the kit with this same enzyme generates overhangs complementary to those found on the 5’ end of I-TCAG and the 3’ end of IV-TCA only. Ligation will generate a full length TALE domain comprised of the invariant N- and C-terminal regions of the TALE protein (contained in the pTAL_GFP plasmid), with a central variable domain comprised of the now ligated multimers I-IV. A schematic representation of this process is shown in Figure 3.2. The pTAL_GFP plasmid contains additional protein domains and sequences which are utilised for genome editing in human cells, but these were not required for this project.

<table>
<thead>
<tr>
<th>Sequence Name</th>
<th>1st Tetramer plasmid</th>
<th>2nd Tetramer plasmid</th>
<th>3rd Tetramer plasmid</th>
<th>4th Trimer plasmid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Telomeric (TTAGGG)_n repeats</td>
<td>I-AGGG</td>
<td>II-TTAG</td>
<td>III-GGTT</td>
<td>IV-AGG</td>
</tr>
<tr>
<td>CIR147 centromere repeats</td>
<td>I-GCAG</td>
<td>II-CGTT</td>
<td>III-GTGC</td>
<td>IV-ATG</td>
</tr>
<tr>
<td>Ingi transposable elements</td>
<td>I-GCCG</td>
<td>II-GCCA</td>
<td>III-CCTC</td>
<td>IV-AAC</td>
</tr>
<tr>
<td>Minichromosomal 177 bp repeats</td>
<td>I-GCGC</td>
<td>II-AGTT</td>
<td>III-AACG</td>
<td>IV-CTA</td>
</tr>
</tbody>
</table>

Table 3.2. Plasmids from the Musunuru/Cowan lab kit encoding tetramer/trimer pieces required for cloning TALE domains to bind the sequences specified in Table 3.1.
Figure 3.2. The Musunuru/Cowan lab TALEN kit allows cloning of TALE domains that bind any 15 bp DNA sequence. (A) Schematic of the Telomere-TALE fusion protein. The 15 bp target sequence on telomere repeats is highlighted in red. NLS – nuclear localisation signal. YFP – yellow fluorescent protein. Black bar – glycine-serine linker region. Image made using Biorender.com. (B) Schematic of the TALEs cloning workflow used in this study. TALE domains are first constructed using the Musunuru/Cowan lab TALEN kit followed by transformation into repeat-tolerant competent cells and sequencing. TALE domains are then amplified by PCR and inserted into the plasmid from Section 3.1 using Gibson Assembly. Figure modified from Ding et al using Biorender.com (Ding et al., 2013).
Step 2 in Figure 3.2 above proved to be problematic, with efforts at simultaneously ligating all four fragments into the vector failing repeatedly. As a result, the four fragments were ligated in the absence of vector, and the resulting product gel purified (see example, Figure 3.3 below). This larger single insert, comprised of multimers I-IV ligated together, was then successfully ligated into the linearised pTAL_GFP plasmid.

Transformation of the plasmid generated in Step 2 also presented difficulties. Transformation into DH5α cells gave colonies containing collapsed TALE domains, that is, TALE domains that were several hundred base pairs too short in length. However, transformation into NEB Stable cells (https://www.neb.com/products/c3040-neb-stable-competent-e-coli-high-efficiency#Product%20Information) resulted in colonies containing the correct insert size.

Figure 3.3. Example of the TALE multimer ligation in the absence of vector. The inserts digested from plasmids I, II, and III are approximately 410 bp, while the insert digest from plasmid IV is approximately 260 bp, giving a full ligation product of approximately 1500 bp. Partial ligation products of lower molecular weight containing some combination of three fragments or less ligated together are visible below the correct band. The band at 1500 bp was excised and gel purified before ligating into the pTAL-GFP destination vector. Promega 1 kb ladder, 1% TAE agarose.

3.3.3 Cloning the trypanosome TALE expression vector using Gibson Assembly

Due to the absence of compatible restriction sites, Gibson Assembly was used to generate the final trypanosome expression vector. The backbone of the plasmid generated in Section 3.3.1, the TALE domain from Section 3.3.2, and the 3xTy-YFP-3xTy tag from pPOTv6 were all amplified by PCR, and assembled using Gibson Assembly (Figure 3.4, A).
Figure 3.4. Generation of the final Telomere-TALE expression vector using Gibson Assembly. (A) Schematic of the cloning process. The relevant sections of each plasmid were amplified using PCR, then assembled using Gibson Assembly. See Section 2.1.9 for more information on this process. Primer sequences can be found in Appendix A. (B) Plasmid map of the final CATpHD449 PAD1 3’UTR Δ1-354 NLS Telomere-TALE Ty-YFP-Ty plasmid. Linearising the plasmid via digestion at the NotI site (labelled) allows integration into the beta-tubulin locus. Images taken from SnapGene.
The Gibson Assembly product also needed to be transformed into NEB Stable cells, as transformation into DH5α cells gave colonies containing collapsed TALE domains. Plasmid clones were Sanger sequenced using forward and reverse primers annealing within the aldolase 5’ UTR and PAD1 3’UTR respectively, in addition to the 4 primers used to sequence the TALE domain in Section 3.2.2. In all cases, the resulting plasmid sequences were as designed. One plasmid clone was selected for each TALE, linearised using NotI digestion, and transfected into bloodstream form monomorphic 427 trypanosomes (Section 2.2.4).

![Figure 3.5. TALEs Gibson Assembly cloning process. (A) PCR products used for Gibson Assembly (red boxes) were excised from the gel, purified, and incubated with the Gibson Assembly mastermix (NEB). (B) Colony PCR of bacterial colonies generated after transformation with Gibson Assembly reaction. Red boxes – correct insert. Green box – collapsed TALE domain. Empty lanes – no TALE domain/unsuccessful PCR.](image)

3.4 Western blot analysis detects TALE protein expression

Multiple parasite clones were retained from each transfection and tested by western blotting to assay whether the TALE proteins were being expressed at sufficient levels. As mentioned in Section 3.3, the Δ1-354 PAD1 3’UTR used on this construct was predicted to give very low protein expression levels compared with a truncated aldolase 3’UTR (Paula McGregor PhD Thesis), and so there were concerns that the protein would not be expressed at sufficient levels for detection. Western blotting would also highlight any potential issues with the cloning of the TALE protein.
Figure 3.6. Western blots confirming TALE protein expression. Samples were run on 10-12% gels in MES buffer at 200 V for 60 min. Membranes were incubated with monoclonal mouse anti-GFP primary antibody and anti-mouse HRP secondary antibody. The expected molecular weight of each TALE protein is approximately 110 kDa. The YFP positive control is YFP-SET15 obtained from Dr Roberta Carloni. (A) Telomere-TALE. 5x10^6 cells/lane, 5 second exposure. (B) Centromere-TALE. 5x10^6 cells/lane, 5 second exposure. (C) Ingi-TALE. 5x10^6 cells/lane, 5 second exposure. (D) Minichromosome-TALE. 2x10^6 cells/lane, 10 second exposure. Samples sonicated after lysis.

All clones tested for each of the four TALE cell lines expressed the desired TALE protein. In all cases, the protein was easily detectable, and had expression levels similar to the YFP-positive control SET15. The molecular weight appeared to be lower than the approximate 110 kDa expected in panels A, B, and D, and perhaps slightly higher in panel C. The reason for this discrepancy is unclear, but may reflect the migration characteristics of the respective TALE proteins.
Some smearing at high molecular weight can be seen in each of the blots from Figure 3.6 (A, C, D). This was likely due to high molecular weight DNA causing increased levels of viscosity of the samples. This issue appeared to be solved by sonicating samples (Figure 3.6, B).

One clone of each cell line was chosen and used for all of the remaining subsequent analyses. Clones used for all remaining experiments in this thesis were:

- Telomere-TALE clone D1
- Centromere-TALE clone 5
- Ingi-TALE clone 5
- Minichromosome-TALE clone 3

3.5 Sub-cellular localisation of TALE proteins by immunofluorescence microscopy

The next step was to investigate the subcellular localisation of the different TALE proteins by immunofluorescence. This would give some indication as to whether they were exhibiting a pattern consistent with correctly binding their target loci. The subcellular localisations of telomeres, centromeres, and minichromosomes have all been determined previously (discussed below), and so comparisons were readily available for expected TALE localisations for these loci. However, the localisation of ingi repeats has not been reported. As discussed in Section 1.2.4, these elements are primarily localised in the sub-telomeres and at strand-switch regions. Therefore, one would expect to see this TALE distributed throughout the nucleus at many different locations.
3.5.1 Immunofluorescence of Telomere-TALE shows a punctate distribution similar to that reported for telomere proteins

Based on reported fluorescence in situ hybridization (FISH) data and immunofluorescence of various telomere proteins, telomeres are known to exhibit a punctate localisation pattern, with puncta localised throughout the nucleus, often at the nuclear periphery (B. Li et al., 2005; Reis et al., 2018; Yang et al., 2009). The distribution observed for the Telomere-TALE was consistent with this (Figure 3.7). However, without co-staining of telomere repeats or telomere-binding proteins, one cannot conclude that this TALE is binding telomeres based on these images alone.

Figure 3.7. Telomere-TALE Immunofluorescence. Bloodstream form cells were fixed in 4% PFA and proteins were immunolocalised using a polyclonal anti-GFP primary antibody and an Alexafluor-568 labelled secondary antibody (pink in merged panel). Nuclear and kinetoplast DNA were stained with DAPI (blue in merged panel). DAPI and TRITC (Tetramethylrhodamine-5-(and 6)-isothiocyanate) refer to fluorophores used for excitation.
Repeated efforts were made to demonstrate co-localisation of the Telomere-TALE and endogenous telomere proteins using two different approaches: 1) Co-staining in cells expressing YFP-TRF (cell line provided by Dr Roberta Carloni) and Telomere-TALE-mCherry; and 2) Co-staining Telomere-TALE-YFP cells with anti-TRF or anti-TelAP1 antibodies provided by Dr Christian Janzen (University of Wurzburg). Both approaches suffered with background signal issues, and neither yielded sufficiently clear images (data not shown).

3.5.2 Immunolocalisation of Centromere-TALE shows localisation in two primary puncta

No FISH analysis has been reported for CIR147 centromere repeats in *T. brucei*, but the localisation of kinetochore proteins can be used as a proxy for centromere localisation. FISH has been performed for the centromere of chromosome 3 only, as the repeats at this centromere differ substantially in sequence to those on the other megabase chromosomes (Akiyoshi & Gull, 2014b). In G1, kinetochores appear as a cluster of dots in the centre of the nucleus, while centromere 3 appears as two dots in the centre of the nucleus (Akiyoshi & Gull, 2014b).
Figure 3.8. The Centromere-TALE localises in two primary puncta in the nucleus of 1K1N cells. Bloodstream form cells were fixed in 4% PFA and proteins were immunolocalised using a polyclonal anti-GFP primary antibody and an Alexafluor-568 labelled secondary antibody (pink in merged panel). Nuclear and kinetoplast DNA were stained with DAPI (blue in merged panel). DAPI and TRITC refer to fluorophores used for excitation.

The Centromere-TALE clearly formed a pair of dots in the nucleus, albeit with substantial background signal (Figure 3.8). This distribution is relatively consistent with what has been reported previously, but as with the Telomere-TALE, this cannot be taken as definitive evidence of localisation at CIR147 centromere repeats.

Limited attempts at staining the Centromere-TALE and the kinetochore protein KKT2 in the same cell line were made. No antibodies were available against any kinetochore proteins. Thus, cells expressing YFP-KKT2 (gifted by Dr Desislava Staneva) and Centromere-TALE-mCherry (instead of YFP) were generated and analysed by immunofluorescence microscopy, allowing co-localisation of the respective signals to be assessed (Figure 3.9). The resulting images revealed that there was partial co-localisation of the two proteins in metaphase cells. However, the background signal from the TALE protein was very high, and attempts at optimisation did not yield any improvement in the quality of images (data not shown). Therefore, this co-localization approach was not pursued further, as ChIP
experiments were expected to provide stronger evidence for the Centromere-TALE binding to centromeric DNA.

Figure 3.9. KKT2 and Centromere-TALE Co-Immunofluorescence. (A) Cartoons of the two epitope-tagged proteins visualised in panel B. Cartoons generated using Biorender.com. (B) Cells were fixed in 4% PFA. YFP-KKT2 was stained with polyclonal anti-GFP primary antibody and Alexa Fluor-488 secondary antibody (green in merged panel). The Centromere-TALE-mCherry was stained with anti-BB2 primary antibody and Alexa Fluor-568 secondary antibody (pink in merged channel); this stained the Ty-tags flanking the mCherry in the construct. Nuclear and kinetoplast DNA were stained with DAPI (blue in merged panel).
3.5.3 Immunofluorescence microscopy of Ingi-TALE detects a punctate distribution pattern throughout the nucleus

As discussed above, the localisation of Ingi clade retroposons within the nucleus has not been reported. However, their concentration at sub-telomeres and at strand switch regions in the genome would anticipate a punctate distribution throughout the nucleus, including peripheral dots due to their proximity to telomeres.

![Immunofluorescence microscopy of Ingi-TALE](image)

Figure 3.10. Ingi-TALE Immunofluorescence shows the protein distribution throughout the nucleus, with a strong central focus in each cell. Bloodstream form cells were fixed in 4% PFA and proteins were immunolocalised using a polyclonal anti-GFP primary antibody and an Alexafluor-568 labelled secondary antibody (pink in merged panel). Nuclear and kinetoplast DNA were stained with DAPI (blue in merged panel). DAPI and TRITC refer to fluorophores used for excitation.

The Ingi-TALE exhibited a distribution relatively consistent with expectations (Figure 3.10). Most cells seemed to show one particularly strong cluster in the nucleus, with small puncta throughout the remainder of the nucleus. Similar to the other TALEs, this does not provide definitive evidence of localisation at its target sequence. However, the distribution being consistent with expectations provided sufficient evidence to proceed with ChIP experiments.
3.5.4 Minichromosome-TALE localisation is similar to that reported for minichromosomes

As mentioned in Section 3.5, the distribution of minichromosomes in the nucleus has been reported by various research groups. The use of FISH probes that hybridise to the minichromosomal 177 bp repeats has provided information about the movement of these chromosomes throughout the cell cycle, which can be used for comparison with the Minichromosome-TALE (Figure 3.11). Minichromosomes are known to reside at the periphery of the nucleus during interphase, migrate to the centre of the nucleus in metaphase, and move toward the spindle poles as mitosis progresses (Ersfeld & Gull, 1997).

![Figure 3.11. Minichromosome-TALE immunolocalisation shows similar movement throughout the cell cycle as that reported for minichromosomes.](image)

**Figure 3.11.** Minichromosome-TALE immunolocalisation shows similar movement throughout the cell cycle as that reported for minichromosomes. Bloodstream form cells were fixed in 4% PFA and proteins were immunolocalised using a polyclonal anti-GFP primary antibody and an Alexafluor-488 labelled secondary antibody (green in merged panel). Nuclear and kinetoplast DNA were stained with DAPI (blue in merged panel). DAPI and FITC refer to fluorophores used for excitation.

The pattern of the primary puncta observed for the Minichromosome-TALE (Figure 3.11) are consistent with their enrichment on minichromosomes. However, as with the other TALEs,
there was a large amount of background signal, suggesting that: 1) a significant amount of the protein may remain unbound to its target sequence; 2) a significant amount of the protein may be loosely bound to its target sequence; 3) a significant amount of the protein may be bound non-specifically to other similar sequences; or 4) some combination of the above.

3.6 Cells expressing the exogenous TALE proteins have a normal growth rate

TALEs have never been reported to have been expressed in trypanosomes previously. Therefore, growth assays were performed to determine the effects, if any, of TALE expression on cell growth. There were concerns that for telomeres and centromeres in particular the TALE could have a negative impact on the functioning of these critical regions of the genome. For example, the binding of the Centromere-TALE might affect the ability of kinetochore proteins to bind centromeres, negatively affecting chromosome segregation. Similarly, the Telomere-TALE may compete with the endogenous TTAGGG repeat binding protein, TRF, to bind telomeres. It was unclear whether the presence of TALEs bound to minichromosomal 177 bp repeats or Ingi retroelements might have a negative impact, as the roles of these sequences within the cell, if any, are unknown.
Figure 3.12. Growth profiles of TALE cell lines show TALE expression is well tolerated for each protein. Error bars represent SD from the mean of 3 independent experiments using 3 biological replicates each. WT = wild type. (A) Telomere-TALE. (B) Minichromosome-TALE. (C) Centromere-TALE. (D) Ingi-TALE.

All four cell lines showed no decrease in growth rate compared with the parent WT cell line. In fact, all TALE-expressing cell lines appeared to grow faster than the WT cell line, contrary to expectation. Although further clones have not been analysed to determine whether this was consistent for each TALE cell line, this may reflect the accelerated growth sometimes seen with transgenic parasite lines after selection, where cells displaying the most accelerated growth rate emerge first from the population. Overall, however, the analysis revealed that strong growth effects were not observed with any TALE-expressing cell line, and that further analysis of interactions of the TALE proteins could be pursued.

3.7 Chromatin immunoprecipitation (ChIP)-qPCR with TALE proteins shows variable enrichment at target loci

In order to assess the specificity of the binding of TALE proteins to their target loci, ChIP assays were performed for each protein. These assays revealed that there was good enrichment of both the telomere and centromere TALEs at their respective target loci, with minimal signal from the histone H3 and actin B negative control loci (panels A and B). Enrichment of the Minichromosome-TALE (panel C) at its target locus was also good, but in this case signal from the three negative control loci used was much higher than in the Telomere- or Centromere-TALE experiments, suggesting a higher level of off-target binding than seen for either of those constructs. This situation was even more extreme in the case of the Ingi-TALE (panel D), where enrichment at the target locus compared to the negative control loci was poor. The error bars also showed high levels of variation between experiments, making it impossible to conclude if the Ingi-TALE construct was truly enriched at its target sequence compared with background.
Figure 3.13. ChIP-qPCR of TALE cell lines shows varying degrees of enrichment of TALE proteins at target loci. (A) Telomere-TALE (B) Centromere-TALE (C) Minichromosome-TALE (D) Ingi-TALE. Data are mean ± standard deviation from three biological replicates. X-axis labels represent primers designed for each locus. All immunoprecipitations were performed using polyclonal anti-GFP antibodies as outlined in Section 2.5.
3.8 Telomere-TALE expression causes decrease in telomere length

As discussed in Section 3.5, binding of the Telomere-TALE to the (TTAGGG)$_n$ repeats may disrupt binding of TRF to these repeats, and hence telomere function. TRF is important for telomere end protection (B. Li et al., 2005), and so disruption of the TRF-telomere interaction could cause telomere shortening over time. Alternatively, the telomeres might elongate, in response to the additional TTAGGG binding protein being present within the cell.

Telomere length was assayed by Southern analysis (Figure 3.14). In this assay, genomic DNA is digested using restriction enzymes with 4-base pair recognition sequences that are not found in the telomeric repeats. This leaves telomeres intact but digests the remainder of the genome into small fragments due to high prevalence of these restriction enzyme sites outside of telomeres. After electrophoretic separation of DNA and transfer to a nylon membrane, telomeres are detected using a short probe complementary to the telomeric repeats labelled at its 3’ end with a digoxigenin (DIG). DIG can then be detected chemiluminescently using an ascorbate peroxidase-labelled anti-DIG antibody. The molecular weight of the resulting bands is assessed compared to a DNA ladder of known molecular weight, and substantial differences in telomere size between samples can easily be distinguished relative to the ladder.
Figure 3.14. Telomere length analysis of Telomere-TALE cells. (A) Flow diagram outlining the main experimental steps. The two restriction enzymes used for digestion of genomic DNA (italics) and their recognition sites are shown in red and blue respectively. DIG – Digoxigenin; AP – Ascorbate peroxidase. (B) Southern blot showing telomere length in wild type and Telomere-TALE cell lines. 2 µg of genomic DNA was digested with AluI and MboI, which recognise 4bp sequences not found within the TTAGGG repeats. Membranes were hybridised with a (TTAGGG)₃ probe labelled with DIG-dUTP at its 3’ end before chemiluminescent detection with CDP-Star. (C) Ethidium bromide stained 0.5% TBE agarose gel. Images are the same as those used in Chapter 5, Section 5.9, but the lanes containing the ΔZC3H39, ΔZC3H40, and ΔZC3H39ΔZC3H40 samples have been cropped out.
Figure 3.14 shows a difference between telomeres of WT cells and those of the Telomere- TALE cells. *T. brucei* telomere repeat arrays have previously been characterised as being approximately 15 kb in length on average (G. A. Hovel-Miner et al., 2012a). The WT digested lane in panel A shows two distinct bands of approximately 23 kb and 15 kb, as well as a smear migrating down the gel decreasing in intensity from 15 kb to 3 kb, consistent with reported telomere lengths for this strain (Dreesen & Cross, 2008). In contrast, the Telomere-TALE lane shows a broad smear ranging from approximately 6 kb down to 1 kb. Panel B, comprising the ethidium bromide stained digested DNA of each cell line, does not show any large differences in DNA loading or digestion profile of the respective samples. Therefore, it appears that the expression of the Telomere-TALE causes severe shortening of telomeres.

Due to time pressure and the focus of the project on using the TALE proteins for affinity purification experiments, no investigation into the possible mechanisms behind this difference in telomere length has been performed.

### 3.9 Discussion

This chapter showcased the successful design, cloning, expression, and basic characterisation of cell lines expressing TALE proteins designed to bind telomeres, centromeres, ingi transposable elements, and minichromosomes. The cloning strategy outlined in Section 3.3 successfully generated epitope-tagged TALE proteins capable of binding different DNA sequences within the cell. The results from this chapter provided sufficient evidence that these cell lines could be utilised for attempts at purifying proteins associated with three of the four loci. These attempts are outlined in the next chapter.

#### 3.9.1 Western analysis shows higher TALE expression levels than expected

In all four cases, the expressed TALE proteins appeared to be expressed at high levels, with very low exposure times providing saturating levels of signal. In all cases, the number of cell equivalents loaded on the gel was in the standard range used in the lab (i.e. 2-5 x 10^6 cells), suggesting reasonably high protein expression levels. Despite the fact that the 3'UTR used for these constructs had been characterised as giving very low levels of expression compared to an aldolase 3'UTR, it is possible that the protein turnover kinetics of the TALEs could differ substantially from that of the chloramphenicol acetyltransferase used in the reporter assay.
characterising this truncated PAD1 3′UTR. Moreover, their accumulation in the nucleus and their DNA binding could combine to a decreased turnover compared with a cytosolic protein such as chloramphenicol acetyl transferase.

3.9.2 Immunofluorescence shows distinct localisations for all TALE proteins

The cellular distribution of the respective TALE proteins was assayed to determine whether they exhibited the expected cellular localisation. This revealed that all four cell lines showed subcellular localisations that were clearly distinct from one another, and broadly consistent with what would be expected for proteins bound to these target sequences. However, subcellular localisation can only be taken as an indication of potential correct binding, and not evidence of it. Although more detailed locational information could be generated by, for example, FISH combined with TALE immunofluorescence, it was decided not to pursue this due to time constraints. It was reasoned that progressing to ChIP experiments would yield results more rapidly, as experience and expertise relating to this were available in the lab, whereas they were not available for FISH experiments. Further, ChIP analysis would be a superior experiment to FISH, as it was expected to provide stronger evidence of correct localisation of the TALE (that is, it shows the protein is within formaldehyde crosslinking distance of its target sequence), as well as providing quantitative data.

3.9.3 TALE expression does not interfere with cell growth

The increased growth rate of the TALE-expressing cell lines compared to WT cells was unexpected. One could reasonably expect binding of the TALE protein at centromeres or telomeres to negatively affect cell growth. TALE binding to minichromosomes is less likely to interfere with cell growth, since these chromosomes do not carry essential housekeeping genes. Their hypothesised role of expanding the reservoir of available VSG genes to aid in immune evasion would not be required in cultured cells, and so any possible negative influence on that process would not arise in culture unless perturbation of mitotic events occurred through the binding of the TALE protein. Similarly, Ingi clade transposable elements are not known to have any cellular function in *T. brucei*, and therefore any possible impacts of TALE proteins binding to these elements are difficult to predict. Overall, reasons for the apparently increased growth of the TALE expressing cells are unclear, although further
controls would be needed to establish if this was linked to the TALE protein expression. For example, after transfection the selection for transformed cells can isolate cell lines with faster overall growth rate compared with wild type parents unrelated to their expression of exogenous or tagged proteins. To eliminate this possibility, TALEs unable to bind any genomic DNA could be analysed as a negative control, or TALE expression could be prevented by CRISPR-Cas9 excision of the construct to determine if growth rates reduced to wild type levels. Further independent clones could also be analysed to assess whether the phenotype was consistent, and inducible expression could be used to assess if the growth rate was dependent on the expression level of the different TALE proteins.

3.9.4 ChIP-qPCR shows variation in target enrichment for the different TALE proteins

All four TALE proteins showed enrichment at their target locus, but showed varying degrees of background signal from negative control loci. The Telomere- and Centromere-TALEs showed high levels of enrichment for their target loci, with very low enrichment of the negative control loci. The Minichromosome-TALE also showed good enrichment at its target locus compared with the negative control loci, but these showed higher levels of enrichment than seen with the Telomere- and Centromere-TALEs. This suggests higher levels of off-target binding by this TALE than the Telomere- and Centromere-TALEs. Similarly, the Ingi-TALE showed enrichment at its target, but high levels of enrichment of the Actin B and Histone H3 genes as well, suggesting even higher levels of off-target binding than for the Minichromosome-TALE. The exact reasons for this variation in enrichment at different target sites relative to negative control loci are unclear, but factors that are likely to influence this will be discussed below.

TALE proteins have general DNA-binding activity due to their affinity for the negatively charged backbone of DNA and the thymidine found before its target sequence (A. N. S. Mak et al., 2012; Rinaldi et al., 2017). Further, there is some flexibility seen in base recognition from different repeat variable di-residues, and a small number of base mismatches can be tolerated (Juillerat et al., 2014; Rogers et al., 2015). The combination of these factors suggests certain sequences may be superior to others in terms of their ability to support high-specificity TALE binding. It is possible that the sequences chosen for the Ingi- and Minichromosome-TALE are not bound as tightly as those of the Telomere- and Centromere-TALE.
Another factor that could be influencing off-target binding is the copy number of the locus within the cell. The different TALE proteins should be expressed at similar levels due to the presence of the same 3’UTR on each protein, but there are very large differences in copy number between the different target sequences, particularly the Ingi elements compared with the other three sequences (discussed in detail in Chapter 4, Section 4.6). Therefore, when all available TALE binding sites are occupied, the remaining TALE molecules not bound to the target site will differ in abundance for each protein in each cell line. In the case of a low copy number target site, for example the Ingi elements, this means large amounts of TALE protein in the nucleus that remain free to bind off-target sequences at low levels. This means that in the case of a low copy number sequence, the fraction of the TALE protein being immunoprecipitated in a ChIP experiment that is bound to its target site is much lower than for a high copy number sequence. This of course assumes similar levels of protein breakdown within the different cell lines, which is reasonable to assume based on the protein levels detected (Figure 3.6). An approach to reduce this phenomenon would be potentially to use a system allowing titratable expression of the TALE protein, for example regulated through doxycycline inducible expression. Although not considered optimal here, for those TALEs with low genomic copy number targets such an approach could reduce the background interactions observed.

3.9.5 Telomere length may be decreased in Telomere-TALE cells

The finding that Telomere-TALE expression appears to cause telomere shortening is an interesting observation but requires further exploration. For example, this conclusion would be strengthened by the addition of a control probe for a chromosomal internal locus such as centromere repeats. However, this would require digestion with different restriction enzymes, as the 4-base cutters used in the experiment would likely digest many other loci. Telomere length could also be measured using an alternative method such as qPCR. No investigation of possible mechanisms for this potential telomere shortening phenotype were explored, but possibilities are discussed below.

If Telomere-TALE binding interferes with the binding of endogenous telomere proteins, then perhaps the levels of these proteins at telomeres could be assayed by ChIP-qPCR in WT vs Telomere-TALE cells. Similarly, protein levels for known telomere-associated proteins could be compared using quantitative western analysis or quantitative mass spectrometry. However, based on other results (Section 4.2.2), it is clear that occupancy of telomere proteins on telomeres remains high in the Telomere-TALE cell line, as they were highly enriched in Telomere-TALE immunoprecipitation preparations. Further, it has been shown that loss of
TRF does not alter telomere length, but affects the G-overhang structure (B. Li et al., 2005). Therefore, the possible telomere length phenotype seen in this experiment is not consistent with loss of TRF alone.

The apparent telomere-shortening phenotype seen here is interesting given the fact that expression of the Telomere-TALE did not cause any decrease in growth rate compared to wild-type cells. If progressive telomere loss were to occur, one would expect to see a profound growth defect. However, these parasites were routinely cultured for hundreds of generations without any apparent decrease in growth rate. Trypanosomes do not appear to have a telomere-loss checkpoint (Glover et al., 2007), but one would expect progressive loss of terminal DNA to eventually have negative consequences. This may suggest that the cells stabilize telomere length in some fashion, as has been reported after loss of telomerase (Dreesen & Cross, 2006). Further, trypanosome telomeres are known to progressively elongate by approximately 10 bp per population doubling (Dreesen & Cross, 2008; Horn et al., 2000), which could contribute to a possible telomere length stabilization effect.

Shorter telomeres have been associated with increased VSG switch frequency, so it would be interesting to determine rates of VSG switching in the Telomere-TALE cell line (G. A. Hovel-Miner et al., 2012b). This could be explored using antibodies against known VSG proteins combined with immunostaining and flow cytometry.

3.9.6 Conclusions

Overall, the analyses presented showed that synthetic TALEs could be expressed in trypanosomes, and appear to be well tolerated. It also demonstrated that TALEs can successfully bind a target sequence in trypanosomes. This not only paves the way for the locus-specific proteomics (Chapter 4), but also for their potential future use in genome editing, tethering assays, locus visualisation, transcriptional modulation and more.
Chapter 4: Locus-specific proteomics using TALEs allows identification of proteins enriched at specific chromosomal locations

4.1 Introduction

Chapter 3 focussed on generating TALE-expressing cell lines and assaying whether the TALEs bound their target locus. All cell lines successfully expressed the TALE constructs and achieved intended enrichment of the target locus in quantitative ChIP experiments. The next step was to utilise these cell lines for affinity purification experiments in order to determine if our strategy of using the TALEs to purify a particular genomic locus was feasible.

All experiments performed in this chapter used approximately equivalent cell numbers (ranging from 3-5 x 10⁸ cells per replicate) and the same protocol for each TALE unless otherwise stated. All immunoprecipitation experiments were performed in triplicate. An aliquot of each replicate was analysed by SDS-PAGE and silver stained in order to assay whether IP had been successful, and if replicates were comparable before digesting and injecting into the mass spectrometer.

In order to validate the immunoprecipitation protocol, positive control experiments were performed using the telomere repeat (TTAGGG)_n-binding factor TRF and the kinetochore protein KKT2 (Akiyoshi & Gull, 2014b; B. Li et al., 2005). These proteins have known interaction partners, and are known to bind two of the loci of interest for this project – telomeres and centromeres. Therefore, immunoprecipitations performed with these proteins could inform us as to whether proteins associated with these loci could be enriched using our conditions, or if troubleshooting and optimisation would be needed to develop a more robust protocol.
Figure 4.1. Workflow for affinity purification coupled to liquid chromatography tandem mass spectrometry (LC-MS/MS) of YFP-tagged control proteins and TALEs. For each experiment, triplicate cultures were grown for each cell line. Cells were sonicated briefly to aid lysis and shear chromatin. Dynabeads crosslinked to anti-GFP monoclonal antibody were used to immunoprecipitate YFP-tagged TALEs and control proteins. Figure generated using BioRender (Biorender.com).
4.2 Proteomic identification of telomere-associated proteins

4.2.1 TRF - a positive control for telomere-binding proteins

TRF is the most well characterised telomere-associated protein in *T. brucei*, and has been shown to directly bind the telomeric TTAGGG repeats (B. Li et al., 2005). This means it can be used as a positive control of an endogenous telomere-binding protein for comparison with the synthetic telomere-binding protein, the Telomere-TALE. Therefore, affinity purification coupled to liquid chromatography tandem mass spectrometry (LC-MS/MS) of YFP-TRF was performed in order to generate a comparison data set to use as a positive control for the Telomere-TALE. As expected this identified proteins previously identified as interacting with TRF in the literature, as well as other known telomere-associated proteins. Several proteins not previously reported to be present at telomeres were also identified. These included the DNA repair proteins MRE11 and RAD50, HDAC3, PIP5Pase1, PP1, and the Zinc Finger proteins ZC3H39 and ZC3H40. Proteins of interest will be discussed further below. A full list of significantly enriched proteins is provided in Table 4.1.

![Figure 4.2. Silver stained SDS-PAGE gel of IP samples from WT and YFP-TRF](image)

Approximately 15% of each sample was run on a 10-12% gel, 195 V, 50 mins, MES buffer, SeeBluePlus2 protein standard. Proteins were stained using the SilverQuest Silver Staining kit. The gel clearly shows that the replicates from each cell line are similar, and therefore can be compared to one another. WT = wild-type.
As mentioned above, samples were subjected to SDS-PAGE and silver staining prior to digestion and injection into the mass spectrometer (example in Figure 4.2 above). This served two purposes: 1) It identified whether proteins were present, giving rough estimates of the total amount of protein present and the protein complexity of the samples; and 2) Allowed us to determine whether the replicates generated from each cell line were comparable to one another, and each other. This is important for quantitation, and allows firm conclusions to be able to be drawn from the experiments. For example, you would expect the majority of the proteins in wild-type and tagged cell lines to be the same, as many proteins will adhere to the dynabeads and antibody used during the immunoprecipitation process. However, if large differences in protein abundance or composition were visible by silver staining, one should be concerned, as this would likely reflect technical differences in sample preparation, rendering your samples incomparable to one another. This point holds for variation both between, and within, sample groups. Since statistical analysis will be one of the parameters utilised to infer whether proteins are enriched in a particular sample (see all volcano plots in this thesis), minimising variation between replicates is key to avoiding a large number of false positive protein identifications.
Figure 4.3. YFP-TRF associates with predominantly telomeric proteins. Significance thresholds (dotted lines) used to determine significantly enriched proteins are $p < 0.01$ (Y-axis) and $\log_2 > 2$ (X-axis). The data shown are for one experiment based on 3 biological replicates. Proteins with similar known functions have been grouped by colour. Proteins of particular interest are labelled where possible. A full list of the proteins enriched in this plot can be found in Table 4.1 below. The bait protein is in red. Proteins previously reported to be telomere-associated are in blue. Histones and histone variants are in green. The zinc finger proteins ZC3H39 & ZC3H40 are in beige. Proteins involved in DNA repair are in orange. Known chromatin-associated proteins are in purple. Proteins outside of these chosen categories are in black. Proteins below the significance thresholds are grey.
Table 4.1. Proteins enriched in YFP-TRF pulldown. All proteins above the significance thresholds defined in Figure 4.3 are listed below along with their UniProt ID, EuPathDB gene code (Tb927 genome), and name as listed in the TriTrypDB. Proteins are colour coded as in Figure 4.3. Proteins have been ranked from highest to lowest log\(_2\) fold change.

<table>
<thead>
<tr>
<th>UniProt ID</th>
<th>EuPathDB Gene ID</th>
<th>Name</th>
<th>Log2FC</th>
<th>-log student's t-test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q388W1</td>
<td>Tb927.10.12850</td>
<td>ttagg binding factor (TRF)</td>
<td>11.4488</td>
<td>4.208880733</td>
</tr>
<tr>
<td>Q383U2</td>
<td>Tb927.11.9870</td>
<td>Telomere-associated protein 1 (TelAP1)</td>
<td>10.6092</td>
<td>5.278258895</td>
</tr>
<tr>
<td>Q385L3</td>
<td>Tb927.11.5550</td>
<td>DNA polymerase theta (POLQ)</td>
<td>10.5817</td>
<td>6.221766373</td>
</tr>
<tr>
<td>Q38FD7</td>
<td>Tb927.9.4000</td>
<td>hypothetical protein, conserved</td>
<td>10.5448</td>
<td>4.051895146</td>
</tr>
<tr>
<td>Q57XY9</td>
<td>Tb927.3.1560</td>
<td>TRF-interacting Factor 2 (TIF2)</td>
<td>10.2832</td>
<td>4.450465761</td>
</tr>
<tr>
<td>Q586Z9</td>
<td>Tb927.6.4330</td>
<td>hypothetical protein, conserved</td>
<td>10.1356</td>
<td>5.978846837</td>
</tr>
<tr>
<td>Q387A4</td>
<td>Tb927.11.370</td>
<td>repressor activator protein 1 (RAP1)</td>
<td>9.88125</td>
<td>4.290541391</td>
</tr>
<tr>
<td>Q38BZ3</td>
<td>Tb927.10.2520</td>
<td>PrimPol-like protein 2 (PPL2)</td>
<td>8.67295</td>
<td>3.881820487</td>
</tr>
<tr>
<td>Q586G9</td>
<td>Tb927.2.2190</td>
<td>histone deacetylase 3 (HDAC3)</td>
<td>8.20092</td>
<td>4.031833993</td>
</tr>
<tr>
<td>Q385E2</td>
<td>Tb927.11.6270</td>
<td>inositol-1,4,5-trisphosphate (IP3) 5-phosphatase, putative (PIPSase1)</td>
<td>7.49135</td>
<td>3.993118123</td>
</tr>
<tr>
<td>Q38F55</td>
<td>Tb927.9.2420</td>
<td>hypothetical protein, conserved</td>
<td>5.93654</td>
<td>4.987620643</td>
</tr>
<tr>
<td>Q38Q5</td>
<td>Tb927.11.16920</td>
<td>V5G exclusion protein 1 (VEX1)</td>
<td>5.47485</td>
<td>4.237426163</td>
</tr>
<tr>
<td>Q387K7</td>
<td>Tb927.10.15350</td>
<td>histone H3 variant V (H3V)</td>
<td>4.85346</td>
<td>4.510553883</td>
</tr>
<tr>
<td>Q384L0</td>
<td>Tb927.11.8090</td>
<td>protein phosphatase 1, putative</td>
<td>4.20716</td>
<td>3.610388838</td>
</tr>
<tr>
<td>Q384L8</td>
<td>Tb927.11.8210</td>
<td>RAD50 DNA repair-like protein</td>
<td>4.15672</td>
<td>4.218622248</td>
</tr>
<tr>
<td>Q38F86</td>
<td>Tb927.9.2520</td>
<td>microtubule-associated protein (GB4)</td>
<td>4.0132</td>
<td>4.173457763</td>
</tr>
<tr>
<td>Q38D9</td>
<td>Tb927.9.11850</td>
<td>structural maintenance of chromosome 1, putative (SMC1)</td>
<td>3.9452</td>
<td>2.125245561</td>
</tr>
<tr>
<td>Q586P4</td>
<td>Tb927.2.4390</td>
<td>endo/exonuclease Mre11 (MRE11)</td>
<td>3.91469</td>
<td>2.928979474</td>
</tr>
<tr>
<td>Q388B1</td>
<td>Tb927.10.14950</td>
<td>2C3H40</td>
<td>3.90935</td>
<td>3.812881147</td>
</tr>
<tr>
<td>Q57UB5</td>
<td>Tb927.5.3510</td>
<td>structural maintenance of chromosome 3, putative (SMC3)</td>
<td>3.84727</td>
<td>3.053298911</td>
</tr>
<tr>
<td>Q385S1</td>
<td>Tb927.11.4970</td>
<td>predicted WD40 repeat protein</td>
<td>3.57369</td>
<td>2.833037559</td>
</tr>
<tr>
<td>Q386I3</td>
<td>Tb927.11.3280</td>
<td>Kinesin-13 5, putative (KIN13-5)</td>
<td>3.40996</td>
<td>6.280101072</td>
</tr>
<tr>
<td>P96938</td>
<td>Tb927.11.13640</td>
<td>Thymine dioxygenase JB1 (JB1)</td>
<td>2.95776</td>
<td>3.271147403</td>
</tr>
<tr>
<td>Q57XR0</td>
<td>Tb927.7.6900</td>
<td>sister chromatid cohesion protein 1 (SCC1)</td>
<td>2.8331</td>
<td>2.471948884</td>
</tr>
<tr>
<td>Q581B8</td>
<td>Tb927.8.3980</td>
<td>hypothetical protein, conserved</td>
<td>2.8056</td>
<td>3.201087359</td>
</tr>
<tr>
<td>Q57YX3</td>
<td>Tb927.7.2820</td>
<td>histone H2A, putative (H2A)</td>
<td>2.75212</td>
<td>3.582616618</td>
</tr>
<tr>
<td>Q583S0</td>
<td>Tb927.6.2640</td>
<td>importin alpha subunit, putative (KAP60)</td>
<td>2.65931</td>
<td>2.845828085</td>
</tr>
<tr>
<td>Q38C4</td>
<td>Tb927.10.560</td>
<td>40S ribosomal proteins S11, putative</td>
<td>2.64863</td>
<td>4.613852386</td>
</tr>
<tr>
<td>Q388B1</td>
<td>Tb927.10.14930</td>
<td>2C3H39</td>
<td>2.51104</td>
<td>4.279543564</td>
</tr>
<tr>
<td>Q38D1</td>
<td>Tb927.9.10920</td>
<td>kinetoplastid kinetochore protein 3 (KKT3)</td>
<td>2.46033</td>
<td>4.242899334</td>
</tr>
<tr>
<td>Q496C8</td>
<td>Tb927.11.18680</td>
<td>thymine light chain LCB, putative (DYNNL1)</td>
<td>2.30273</td>
<td>3.719110548</td>
</tr>
<tr>
<td>Q57Z31</td>
<td>Tb927.5.4170</td>
<td>histone H4, putative (H4)</td>
<td>2.23819</td>
<td>3.142592984</td>
</tr>
<tr>
<td>Q382Q9</td>
<td>Tb927.11.12720</td>
<td>hypothetical protein, conserved</td>
<td>2.23404</td>
<td>3.148211644</td>
</tr>
<tr>
<td>Q38FS5</td>
<td>Tb927.9.2390</td>
<td>hypothetical protein, conserved</td>
<td>2.1791</td>
<td>3.63580409</td>
</tr>
<tr>
<td>Q582M3</td>
<td>Tb927.7.6360</td>
<td>histone H2A variant Z (H2AZ)</td>
<td>2.09233</td>
<td>3.82065516</td>
</tr>
<tr>
<td>Q389T1</td>
<td>Tb10.406.0330</td>
<td>histone H2B, putative (H2B)</td>
<td>2.06898</td>
<td>3.67438916</td>
</tr>
<tr>
<td>Q387T4</td>
<td>Tb927.11.2250</td>
<td>conserved protein, unknown function</td>
<td>2.06776</td>
<td>2.268126833</td>
</tr>
<tr>
<td>E4GYX7</td>
<td>Tb927.1.2430</td>
<td>histone H3, putative (H3)</td>
<td>2.01944</td>
<td>3.18035497</td>
</tr>
</tbody>
</table>
Summary of published information on the identified proteins with respect to telomere association:

- **Telomere proteins (blue).** TRF-interacting factor 2 (TIF2), Telomere-associated protein 1 (TelAP1), DNA polymerase theta (POLQ), Tb927.6.4330, and Tb927.9.4000 have previously been shown to interact with TRF in bloodstream-form cells, while PrimPol-like protein 2 (PPL2) bound a telomeric-repeat oligonucleotide incubated with procyclic-form cell lysate (Reis et al., 2018). Repressor activator protein 1 (RAP1) has been shown to interact with TRF via yeast two-hybrid assay, and to bind the telomeric repeats (Cestari et al., 2019; Yang et al., 2009). Tb927.6.4330 partially co-localises with TRF by immunofluorescence, and loss of Tb927.6.4330 results in de-repression of a telomeric reporter gene (Glover et al., 2016). Identification of all known telomere proteins suggested that this was a very robust protocol for purifying telomere-associated proteins.

- **Histones (green).** As discussed in Chapter 1, the H3V has been shown to be present at telomeres (Lowell & Cross, 2004; Siegel et al., 2009), and deletion of H3V causes telomere clustering and increased interactions between repressed BESs (Müller et al., 2018). The presence of the histone H2A variant, H2AZ, was unexpected. Although H2AZ was originally shown by ChIP-slot blot to immunoprecipitate small amounts of telomeric repeats (Lowell et al., 2005), it has since been characterised as localizing to RNA polymerase II transcription start regions (Siegel et al., 2009). Siegel et al. do not discuss any H2AZ being present at telomeres. The presence of H2AZ at telomeres could potentially facilitate the transcription of telomeric repeat-containing RNA (TERRA), which are known to be produced from the active ES (Nanavaty et al., 2017).

- **DNA repair proteins (orange).** MRE11 and RAD50 are members of the MRN complex (MRE11-RAD50-NBS1) involved in DNA double-strand break repair (Syed & Tainer, 2018). Neither MRE11 nor RAD50 have been shown to interact with TRF or telomeres in *T. brucei* previously. Previous work on MRE11 identified a small amount of co-localisation with telomeric repeats in procyclic but not bloodstream forms via combined FISH and immunofluorescence (Tan et al., 2002). It is unclear why the third member of the complex, NBS1, was absent from these preparations. The MRN complex is known to be involved in telomere maintenance in other eukaryotes, and so may play a similar role in *T. brucei* (Syed & Tainer, 2018).
Chromatin-associated proteins.

- VEX1: VEX1 is involved in allelic exclusion (discussed in Section 1.2.1), and is enriched at telomeres by ChIP-seq, especially at the telomere associated with the active ES (Faria et al., 2019).

- HDAC3: There is no evidence in the literature linking HDAC3 to telomeres directly. However, RNAi knockdown of HDAC3 led to de-repression at VSG expression site promoters in both bloodstream- and procyclic-forms (Q.-P. Wang et al., 2010), suggesting a role at sub-telomeric regions at least. In contrast, ChIP-seq demonstrated that HDAC3 was also found to associate with RNAPII transcription-start regions (Staneva et al., 2021). However, immunoprecipitations did show HDAC3 interacting with several telomere proteins (TRF, TIF2, TelAP1, and RAP1), as well as BDF2, which has also been shown to be required for silencing telomeric ES-associated VSG genes (Schulz et al., 2015; Staneva et al., 2021).

- PIP5Pase: PIP5Pase has previously been shown to be involved in control of transcription at VSG expression sites, and to bind to the 70 bp repeats upstream of VSG genes, but not to telomeres (Cestari et al., 2019; Cestari & Stuart, 2015). Immunoprecipitation experiments identified that PIP5Pase interacts with RAP1, PP1, and several other proteins to form a complex. No interaction with TRF was identified in that study.

- JBP1: As discussed in Section 1.3.2, JBP1 binds base J, which is enriched at telomeres in bloodstream form cells (Cross et al., 1999; van Leeuwen et al., 1996, 1997, 1998). Therefore, the presence of JBP1 in these preparations is consistent with the literature.

- Cohesin: Cohesin is a highly conserved four-subunit complex (SMC1, SMC3, SCC1 and SCC3) that mediates sister chromatid cohesion by forming a ring structure around sister chromatids (Peters & Nishiyama, 2012). All four subunits are present in *T. brucei* (Gluenz et al., 2008). The role of cohesin at trypanosome telomeres has not been investigated, but partial depletion of the SCC1, SMC1, or SCC3 subunits led to an increase in *in situ* VSG switching, suggesting a role for cohesin in regulating these telomere-proximal genes (Landeira et al., 2009). Exactly what function cohesin performs at telomeres remains unclear. Cohesin is known to be present at telomeres in *S. cerevisiae*, and that it has a role in
transcriptional silencing at the mating type loci (Losada, 2007), and so could be performing a similar silencing function in *T. brucei*.

- **ZC3H39 and ZC3H40 (beige).** ZC3H39 and ZC3H40 were unexpected TRF interactors. Both proteins have been shown to bind one-another, localise to the cytoplasm, and bind transcripts encoding components of the electron transport chain (Trenaman et al., 2019). The only reported evidence linking these proteins to telomeres is that knockdown of either causes de-repression of silent VSG genes, suggesting a role in silencing telomere-adjacent genes.

- The role of PP1 at trypanosome telomeres has never been investigated. PP1 was recently shown to be part of a complex involved in transcription termination in *Leishmania tarentolae* (Kieft et al., 2020). Since this complex is involved in binding base-J, which has previously been shown to be found at telomeres (discussed above), perhaps PP1 could be part of a complex performing a similar function at *T. brucei* telomeres.
4.2.2 Telomere-TALE identifies known and novel telomere-binding proteins

Based on the results from the previous section, it was clear that: a) we had a protocol capable of purifying telomere-associated proteins; and b) we had a good control data set to use as a comparison for the Telomere-TALE. Encouragingly, affinity purification coupled to liquid chromatography tandem mass spectrometry (LC-MS/MS) of the Telomere-TALE identified nearly all known telomere-associated proteins in *T. brucei*: TRF, TIF2, TelAP1, Tb927.6.4430, Tb927.9.4000, PPL2, RPA1, POLQ (highlighted in blue in Table 4.2). In addition, large amounts of histones and histone variants were identified, as well as DNA replication proteins. The zinc finger proteins ZC3H39 and ZC3H40 identified in the TRF pulldown experiment were also identified here, providing further evidence for their presence at telomeres. Proteins of interest will be discussed further below. A full list of significantly enriched proteins is provided in Table 4.2.

Figure 4.4. Telomere-TALE associates with nearly all known telomeric proteins. Significance thresholds (dotted lines) used to determine significantly enriched proteins are $p < 0.01$ (Y-axis) and $\log_2 > 2$ (X-axis). The data shown are for one experiment based on 3 biological replicates. Proteins with similar known functions have been grouped by colour. Proteins of particular interest are labelled where possible. A full list of the proteins enriched in this plot can be found in Table 4.2. The bait protein is in red. Proteins previously seen at telomeres in the literature are in blue. Histones and histone variants are in green. The zinc finger proteins ZC3H39 & ZC3H40 are in beige. Replication protein A subunits are in purple. Proteins outside of these chosen categories are in black. Proteins below the significance thresholds are grey.
Table 4.2. Proteins enriched in Telomere-TALE pulldown

All proteins above the significance thresholds defined in Figure 4.4 are listed below along with their UniProt ID, EuPathDB gene code (Tb927 genome), and name as listed in the TriTrypDB. Proteins are colour coded as in Figure 4.4. Proteins have been ranked from highest to lowest log₂ fold change. Q9U6Y5 (red) is present as a marker for the Telomere-TALE protein, as TALE constructs generated in this study are not present in the UniProt database, whereas peptides belonging to the YFP tag on the C-terminus of the TALE construct can be identified.

<table>
<thead>
<tr>
<th>UniProt ID</th>
<th>EuPathDB Gene ID</th>
<th>Name</th>
<th>Log2FC</th>
<th>-log student’s t-test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q9U6Y5</td>
<td>N/A</td>
<td>GFP-like fluorescent chromoprotein FP506</td>
<td>8.8613</td>
<td>3.063786083</td>
</tr>
<tr>
<td>Q388W1</td>
<td>Tb927.10.12850</td>
<td>ttaggg binding factor (TRF)</td>
<td>6.7432</td>
<td>3.226333819</td>
</tr>
<tr>
<td>Q388Z3</td>
<td>Tb927.10.2520</td>
<td>PrimPol-like protein 2 (PPL2)</td>
<td>6.3149</td>
<td>2.807303273</td>
</tr>
<tr>
<td>Q586Z9</td>
<td>Tb927.6.4330</td>
<td>hypothetical protein, conserved</td>
<td>6.1248</td>
<td>3.307845196</td>
</tr>
<tr>
<td>Q38FD7</td>
<td>Tb927.9.4000</td>
<td>hypothetical protein, conserved</td>
<td>5.9467</td>
<td>3.628801566</td>
</tr>
<tr>
<td>Q384B5</td>
<td>Tb927.11.9130</td>
<td>Replication factor A protein 1 (RPA1)</td>
<td>5.5355</td>
<td>2.314519654</td>
</tr>
<tr>
<td>Q387X7</td>
<td>Tb927.10.15350</td>
<td>histone H3 variant V (H3V)</td>
<td>5.4883</td>
<td>3.047932825</td>
</tr>
<tr>
<td>Q385L3</td>
<td>Tb927.11.5550</td>
<td>DNA polymerase theta (POLQ)</td>
<td>5.4379</td>
<td>2.589853101</td>
</tr>
<tr>
<td>Q57XY9</td>
<td>Tb927.3.1560</td>
<td>TRF-Interacting Factor 2 (TIF2)</td>
<td>5.3667</td>
<td>2.325508549</td>
</tr>
<tr>
<td>Q383U2</td>
<td>Tb927.11.9870</td>
<td>Telomere-associated protein 1 (TelAP1)</td>
<td>5.2677</td>
<td>2.444235729</td>
</tr>
<tr>
<td>Q388B1</td>
<td>Tb927.10.14950</td>
<td>ZC3H40</td>
<td>5.0889</td>
<td>2.283377436</td>
</tr>
<tr>
<td>Q57W67</td>
<td>Tb927.8.3060</td>
<td>cytosolic leucyl aminopeptidase, putative</td>
<td>5.0647</td>
<td>2.446232073</td>
</tr>
<tr>
<td>Q388B3</td>
<td>Tb927.10.14930</td>
<td>ZC3H39</td>
<td>4.7887</td>
<td>3.156785781</td>
</tr>
<tr>
<td>Q57Z9N</td>
<td>Tb927.5.1700</td>
<td>replication factor A 28 kDa subunit, putative</td>
<td>4.7386</td>
<td>2.339763346</td>
</tr>
<tr>
<td>Q57Z31</td>
<td>Tb927.5.4170</td>
<td>histone H4, putative (H4)</td>
<td>4.6054</td>
<td>2.352120302</td>
</tr>
<tr>
<td>Q57YA3</td>
<td>Tb927.7.2820</td>
<td>histone H2A, putative (H2A)</td>
<td>4.4442</td>
<td>2.89107652</td>
</tr>
<tr>
<td>Q581C6</td>
<td>Tb927.8.4060</td>
<td>Flagellum adhesion protein 2 (FLA2)</td>
<td>4.1701</td>
<td>2.324318672</td>
</tr>
<tr>
<td>Q383E9</td>
<td>Tb927.11.11290</td>
<td>heat shock protein 70, putative (HSP70.c)</td>
<td>4.1493</td>
<td>2.029719092</td>
</tr>
<tr>
<td>Q580T8</td>
<td>Tb927.3.5370</td>
<td>hypothetical protein, conserved</td>
<td>4.0842</td>
<td>2.755902335</td>
</tr>
<tr>
<td>Q389T1</td>
<td>Tb10.406.0330</td>
<td>histone H2B, putative (H2B)</td>
<td>3.8917</td>
<td>2.700154209</td>
</tr>
<tr>
<td>Q587H6</td>
<td>Tb927.2.2670</td>
<td>histone H4 variant (H4V)</td>
<td>3.8333</td>
<td>2.584457227</td>
</tr>
<tr>
<td>Q38DK0</td>
<td>Tb927.9.11940</td>
<td>replication factor A protein 3, putative (RPA3)</td>
<td>3.7248</td>
<td>2.213610165</td>
</tr>
<tr>
<td>Q4GYX7</td>
<td>Tb927.1.2430</td>
<td>histone H3, putative (H3)</td>
<td>3.5533</td>
<td>2.467187132</td>
</tr>
<tr>
<td>Q57Z2H2</td>
<td>Tb927.7.5170</td>
<td>60s ribosomal protein L25</td>
<td>2.9864</td>
<td>2.452607813</td>
</tr>
<tr>
<td>Q38A76</td>
<td>Tb927.10.8980</td>
<td>hypothetical protein, conserved</td>
<td>2.7161</td>
<td>2.325461225</td>
</tr>
<tr>
<td>Q384T0</td>
<td>Tb927.11.7350</td>
<td>Histone H2B variant V (H2BV)</td>
<td>2.6365</td>
<td>2.230096296</td>
</tr>
<tr>
<td>Q582M3</td>
<td>Tb927.7.6360</td>
<td>histone H2A variant Z (H2AZ)</td>
<td>2.5668</td>
<td>2.108527957</td>
</tr>
<tr>
<td>Q381S2</td>
<td>Tb927.11.15130</td>
<td>DnaJ homolog, putative (J7)</td>
<td>2.4464</td>
<td>2.403973336</td>
</tr>
<tr>
<td>Q38EA4</td>
<td>Tb927.9.8720</td>
<td>fructose-1,6-bisphosphatase (FBPase)</td>
<td>2.0381</td>
<td>2.188624619</td>
</tr>
</tbody>
</table>
Summary of published information on the identified proteins with respect to telomere association:

- **Telomere proteins (blue).** All of the known telomere-associated proteins seen in the TRF immunoprecipitation except for RAP1 were identified here. As mentioned above, RAP1 has been shown to bind to telomeric repeats both in vivo and in vitro (Afrin, Gaurav, et al., 2020), so the reasons for its absence here are unclear.

- **Histones (green).** Similar to TRF, the most enriched histone protein in the Telomere-TALE immunoprecipitation was H3V. This is consistent with previously described data discussed in Section 4.2.1. Interestingly, the histone variants H2BV and H4V were seen here but not in the TRF immunoprecipitation. These variants have previously been characterised as being present at RNAPII transcription start- and termination-regions, respectively, with small amounts of H4V also being seen at telomeres (Siegel et al., 2009). Loss of H4V has also been shown to increase VSG switching and to alter both local and global chromatin structure, albeit to a lesser extent than loss of H3V (Müller et al., 2018). The identification of H2BV is not necessarily surprising, as H2AZ has been characterised as dimerizing only with H2BV and not canonical H2B (Lowell et al., 2005). Similar data have not been reported for H3V and H4V, and so it is possible that these variants do not exclusively dimerize with one another.

- **ZC3H39 & ZC3H40 (beige).** Strikingly, ZC3H39 and ZC3H40 were also enriched in this experiment, bolstering the idea that they are genuine telomere-associated proteins.

- **Replication protein A (purple).** Replication protein A (RPA) is a heterotrimeric, single-stranded DNA-binding protein required for DNA replication, recombination, and repair (Wold, 1997). RPA-1 has been shown to bind to single-stranded telomeric-repeats in vitro in both *T. cruzi* and *Leishmania amazonensis* (Neto et al., 2007; Pavani et al., 2018). In *T. brucei*, the 28 kDa subunit of RPA bound a telomere-repeat oligonucleotide incubated with procyclic-form cell lysate (Reis et al., 2018). RPA is involved in telomere maintenance in *S. cerevisiae* and *S. pombe* (Luciano et al., 2012; Schramke et al., 2003; Smith et al., 2000), and is also thought to have a role at trypanosome telomeres (Fernandes et al., 2020).
4.3 Proteomic identification of centromere-associated proteins

4.3.1 KKT2 associates with kinetochore, mitotic, and chromatin proteins

As with telomeres, I also used an endogenous control protein for comparison with the synthetic Centromere-TALE – the kinetoplastid kinetochore protein 2 (KKT2) (Akiyoshi & Gull, 2014b). Hence, a comparison data set of known Centromere-associated proteins to use as a positive control for the Centromere-TALE was generated by performing affinity purification coupled to liquid chromatography tandem mass spectrometry (LC-MS/MS) of YFP-KKT2. This identified 22 kinetochore proteins (21 KKT proteins and KKIP5), as well as 3 members of the cohesin complex and other known regulators and mediators of mitosis. Histones and several nuclear pore complex (NPC) proteins were also identified. Proteins of interest will be discussed further below. A full list of significantly enriched proteins is provided in Table 4.3.

Figure 4.5. YFP-KKT2 associates with kinetochore and mitotic proteins. Significance thresholds (dotted lines) used to determine significantly enriched proteins are $p < 0.01$ (Y-axis) and $\log_2 > 2$ (X-axis). The data shown are for one experiment based on 3 biological replicates. Proteins with similar known functions have been grouped by colour. Proteins of particular interest are labelled where possible. A full list of the proteins enriched in this plot can be found in Table 4.3. The bait protein is in red. Kinetochore proteins (KKTs and KKIPs) are in orange. Histones and histone variants are in green. Other known mitotic proteins are in purple. Members of the cohesin complex are in blue. Proteins outside of these chosen categories are in black. Proteins below the significance thresholds are grey.
Table 4.3. Proteins enriched in YFP-KKT2 pulldown

All proteins above the significance thresholds defined in Figure 4.5 are listed below along with their UniProt ID, EuPathDB gene code (Tb927 genome), and name as listed in the TriTrypDB. Proteins are colour coded as in Figure 4.5. Proteins have been ranked from highest to lowest log₂ fold change.

<table>
<thead>
<tr>
<th>UniProt ID</th>
<th>EuPathDB Gene ID</th>
<th>Name</th>
<th>Log2FC</th>
<th>-log student’s t-test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q383M7 Tb927.11.10520</td>
<td>kinetoplastid kinetochore protein 2 (KKT2)</td>
<td>9.2203</td>
<td>4.277352252</td>
<td></td>
</tr>
<tr>
<td>Q38D1 Tb927.9.10920</td>
<td>kinetoplastid kinetochore protein 3 (KKT3)</td>
<td>8.0767</td>
<td>3.954524052</td>
<td></td>
</tr>
<tr>
<td>Q5845 Tb927.4.5110</td>
<td>kinetoplastid kinetochore protein 8 (KKT8)</td>
<td>7.9766</td>
<td>6.25506769</td>
<td></td>
</tr>
<tr>
<td>Q57V0 Tb927.7.2110</td>
<td>kinetoplastid kinetochore protein 11 (KKT11)</td>
<td>7.792</td>
<td>5.427693154</td>
<td></td>
</tr>
<tr>
<td>Q57N9 Tb927.8.1150</td>
<td>kinetoplastid kinetochore protein 9 (KKT9)</td>
<td>6.5559</td>
<td>3.619342416</td>
<td></td>
</tr>
<tr>
<td>Q5806 Tb927.8.1680</td>
<td>kinetoplastid kinetochore protein 12 (KKT12)</td>
<td>6.5324</td>
<td>4.343411806</td>
<td></td>
</tr>
<tr>
<td>Q38U0 Tb927.11.12410</td>
<td>kinetoplastid kinetochore protein 10 (KKT10)</td>
<td>6.1368</td>
<td>3.511583235</td>
<td></td>
</tr>
<tr>
<td>Q57U8 Tb927.5.3510</td>
<td>structural maintenance of chromosome 3 , putative (SMC3)</td>
<td>5.8876</td>
<td>3.64964231</td>
<td></td>
</tr>
<tr>
<td>Q38D9 Tb927.9.11850</td>
<td>structural maintenance of chromosome 1 , putative (SMC1)</td>
<td>5.8863</td>
<td>3.429803018</td>
<td></td>
</tr>
<tr>
<td>Q38A3 Tb927.10.6330</td>
<td>kinetoplastid kinetochore protein 1 (KKT1)</td>
<td>5.8817</td>
<td>3.695545032</td>
<td></td>
</tr>
<tr>
<td>Q58C6 Tb927.6.1210</td>
<td>kinetoplastid kinetochore protein 6 (KKT6)</td>
<td>5.7045</td>
<td>3.77365838</td>
<td></td>
</tr>
<tr>
<td>Q3887 Tb927.11.1030</td>
<td>kinetoplastid kinetochore protein 7 (KKT7)</td>
<td>5.6352</td>
<td>2.366280277</td>
<td></td>
</tr>
<tr>
<td>Q38A7 Tb927.11.8220</td>
<td>aurora B kinase (AUK1)</td>
<td>5.4704</td>
<td>6.05662731</td>
<td></td>
</tr>
<tr>
<td>Q57V4 Tb927.9.11850</td>
<td>structural maintenance of chromosome 1 , putative (SMC1)</td>
<td>5.427693154</td>
<td>2.519457046</td>
<td></td>
</tr>
<tr>
<td>Q38A7 Tb927.11.8220</td>
<td>aurora B kinase (AUK1)</td>
<td>5.4704</td>
<td>6.05662731</td>
<td></td>
</tr>
<tr>
<td>Q587D3 Tb927.6.4670</td>
<td>MORN repeat-containing protein 1 (MORN1)</td>
<td>4.6014</td>
<td>2.86029993</td>
<td></td>
</tr>
<tr>
<td>Q587D3 Tb927.6.4670</td>
<td>MORN repeat-containing protein 1 (MORN1)</td>
<td>4.6014</td>
<td>2.86029993</td>
<td></td>
</tr>
<tr>
<td>Q387E7 Tb927.11.1030</td>
<td>kinetoplastid kinetochore protein 7 (KKT7)</td>
<td>5.6352</td>
<td>2.366280277</td>
<td></td>
</tr>
<tr>
<td>Q57V4 Tb927.9.11850</td>
<td>structural maintenance of chromosome 1 , putative (SMC1)</td>
<td>5.427693154</td>
<td>2.519457046</td>
<td></td>
</tr>
<tr>
<td>Q38A7 Tb927.11.8220</td>
<td>aurora B kinase (AUK1)</td>
<td>5.4704</td>
<td>6.05662731</td>
<td></td>
</tr>
<tr>
<td>Q587D3 Tb927.6.4670</td>
<td>MORN repeat-containing protein 1 (MORN1)</td>
<td>4.6014</td>
<td>2.86029993</td>
<td></td>
</tr>
<tr>
<td>Q387E7 Tb927.11.1030</td>
<td>kinetoplastid kinetochore protein 7 (KKT7)</td>
<td>5.6352</td>
<td>2.366280277</td>
<td></td>
</tr>
<tr>
<td>Q38A7 Tb927.11.8220</td>
<td>aurora B kinase (AUK1)</td>
<td>5.4704</td>
<td>6.05662731</td>
<td></td>
</tr>
<tr>
<td>Q587D3 Tb927.6.4670</td>
<td>MORN repeat-containing protein 1 (MORN1)</td>
<td>4.6014</td>
<td>2.86029993</td>
<td></td>
</tr>
<tr>
<td>Q387E7 Tb927.11.1030</td>
<td>kinetoplastid kinetochore protein 7 (KKT7)</td>
<td>5.6352</td>
<td>2.366280277</td>
<td></td>
</tr>
<tr>
<td>Q38A7 Tb927.11.8220</td>
<td>aurora B kinase (AUK1)</td>
<td>5.4704</td>
<td>6.05662731</td>
<td></td>
</tr>
<tr>
<td>Q587D3 Tb927.6.4670</td>
<td>MORN repeat-containing protein 1 (MORN1)</td>
<td>4.6014</td>
<td>2.86029993</td>
<td></td>
</tr>
</tbody>
</table>
Summary of published information on the identified proteins with respect to centromere association and/or roles in mitosis:

- **Kinetochore proteins (orange).** KKTs 1-20 and KKT25 were identified in this experiment, representing almost all known kinetochore proteins (Akiyoshi & Gull, 2014b; Nerusheva et al., 2019; Nerusheva & Akiyoshi, 2016). This validates the protocol in terms of its suitability for purifying centromere-associated proteins, providing an excellent positive control data set for comparison with affinity purification experiments performed with the Centromere-TALE.

- **Cohesin (blue).** As mentioned in Section 4.2.1, the cohesin ring mediates sister chromatid cohesion. During mitosis, at the metaphase to anaphase transition, the SCC1 subunit is cleaved by the protease separase, resulting in opening of the cohesin ring and initiation of anaphase (Uhlmann et al., 1999). This process is conserved in trypanosomes (Bessat & Ersfeld, 2009; Gluenz et al., 2008) and so the presence of cohesin is not unexpected for immunoprecipitation of a centromere/kinetochore protein.

- **Known mitotic proteins (purple).** The chromosomal passenger complex (CPC) is a key regulator of mitosis in eukaryotes, and in metazoa is comprised of four subunits: Aurora B kinase, INCENP, survivin, and borealin (Carmena et al., 2012). *T. brucei* has an Aurora B kinase homologue, AUK1, but no homologues of the other subunits found in metazoa. AUK1 has previously been shown to be required for mitotic spindle formation, chromosome segregation, and cytokinesis in both bloodstream and procyclic forms, and to be a key component of a novel, tripartite, kinetoplastid CPC (Z. Li et al., 2008; Z. Li & Wang, 2006; Tu et al., 2006). Specifically, AUK1 was shown to interact with CPC1 and KIN-B via yeast two-hybrid, and to form a complex with CPC1, KIN-B, and two additional proteins not seen in this preparation, CPC2 and KIN-A (Z. Li et al., 2008). KIN-B is localized primarily at the spindle during anaphase, while CPC1 and AUK1 localize to the spindle during metaphase and anaphase (Z. Li et al., 2008). Similar to AUK1, knockdown of either CPC1 or KIN-B results in defects in spindle formation, as well as mitotic and cytokinetic arrest (Z. Li et al., 2008).

- **Histones (green).** Core histones H2A, H2B, and H4 were identified along with the histone H3 variant, suggesting nucleosomes comprising these four histones are present at centromeres. This is consistent with the enrichment of H3V at centromeres shown by ChIP-seq (Akiyoshi & Gull, 2014b). The role of H3V at centromeres in *T.
*brucei* is unclear. As discussed in Section 1.3, many eukaryotes have a centromere-specific histone H3 variant called CENP-A or CENH3. However, trypanosome H3V is not centromere specific – it is highly enriched at telomeres and transcription termination regions (Siegel et al., 2009). Any potential role for H3V in centromere or kinetochore function has not been investigated.

- **Chromatin associated proteins (not coloured in Figure 4.5).**

  - **Condensin (CND):** Similar to cohesin, condensin forms a ring structure around DNA, and is involved in genome organization and chromatin compaction (Kalitsis et al., 2017). All five subunits are present in *T. brucei* (Gluenz et al., 2008), but have not been characterised. Condensin has been shown to accumulate at specific regions of the genome in other eukaryotes, primarily centromeres, telomeres, and transcription start- and termination-regions (Kalitsis et al., 2017). Therefore, its association with KKT2 is consistent with what is observed in other organisms. It is unclear why only CND2 and CND3, and not the other subunits of the condensin ring, were enriched in this experiment.

  - **FACT:** Facilitates chromatin transcription (FACT) is a histone chaperone complex involved in transcription, DNA replication, and DNA repair in many eukaryotes (Gurova et al., 2018). In *T. brucei*, FACT appears to be comprised of just two subunits, Spt16 and POB3. Knockdown of either subunit lead to de-repression of silent VSG ES's (Denninger et al., 2010; Denninger & Rudenko, 2014). The genome wide distribution of FACT in *T. brucei* has not been reported, and so these data are the first evidence to suggest FACT is enriched at trypanosome centromeres/kinetochores. However, knockdown of FACT has been shown to result in a partial G2/M cell cycle arrest, and disruption of minichromosome segregation in both bloodstream and procyclic forms as assayed by FISH (Denninger et al., 2010). Unfortunately, the impact of FACT knockdown on megabase chromosome segregation was not investigated in that study, and so no conclusions can be drawn about FACT playing a more general role in chromosome segregation in *T. brucei*. FACT has been shown to have roles in chromosome segregation in other eukaryotes. For example, in *S. pombe*, FACT is required for accurate chromosome segregation (Lejeune et al., 2007), and in chicken cells FACT is required for proper deposition of CENP-A at centromeres (Okada et al., 2009).
• **Nuclear pore proteins (not coloured in Figure 4.5).** The enrichment of nuclear pore proteins with KKT2 suggests a role for the nuclear periphery in centromere/kinetochore function in *T. brucei*. This would be consistent with what is known in other eukaryotes, where nuclear pore proteins have been shown to perform roles in kinetochore function, spindle assembly, and cytokinesis (Chatel & Fahrenkrog, 2011). Mlp2, also known as Nup92, is a nuclear basket protein that localizes along the mitotic spindle and to the spindle attachment sites at the poles of dividing nuclei in mitosis (DeGrasse et al., 2009; Holden et al., 2014). Nup64 is an FG-repeat NUP (DeGrasse et al., 2009). Nup152 is an outer ring nuclear pore protein (Obado et al., 2016). NUP-1 is the trypanosome lamin analogue (DuBois et al., 2012).
4.3.2 The Centromere-TALE failed to significantly enrich any proteins

Affinity purification coupled to liquid chromatography tandem mass spectrometry (LC-MS/MS) of the Centromere TALE failed to enrich for any known centromere proteins or even known chromatin proteins, e.g. histones. This was surprising given the positive results obtained in the ChIP experiments discussed in Section 3.7.

![Image](image_url)

**Figure 4.6. Centromere-TALE failed to interact with kinetochore proteins, histones, or chromatin associated proteins under initial conditions.** Attempt 1. Significance thresholds (dotted lines) used to determine significantly enriched proteins are $p < 0.01$ (Y-axis) and $\log_2 > 2$ (X-axis). The data shown are for one experiment based on 3 biological replicates. The bait protein is in red. The protein labelled in black is Tb927.11.6390, a putative START domain containing protein. This protein contains a lipid binding domain and was identified by the TrypTag project (tryptag.org) as a weakly cytoplasmic protein. It is most likely a contaminant. Proteins below the significance thresholds are grey.

The contrast between the ChIP-qPCR results and the results shown in Figure 4.6 above led us to hypothesise that perhaps the majority of the Centromere-TALE protein within the nucleus was not bound to the CIR147 repeats, and therefore the majority of the protein that was being immunoprecipitated would not identify any chromatin bound proteins. Since the immunoprecipitation step of our protocol is performed on the soluble fraction of chromatin...
(insoluble material is removed via centrifugation), we reasoned we might be losing some of the chromatin-bound Centromere-TALE to the insoluble pellet. Therefore, the experiment was repeated with double the amount of sonication (6 cycles instead of 3), as well as double the amount of anti-GFP coupled dynabeads in an attempt to: 1) increase the solubility of the chromatin; and 2) increase the ability to bind any chromatin-associated Centromere-TALE. However, this modified protocol also failed to identify any significantly enriched proteins.

Figure 4.7. Centromere-TALE failed to interact with kinetochore proteins, histones, or chromatin associated proteins under revised conditions. Attempt 2. Significance thresholds (dotted lines) used to determine significantly enriched proteins are p < 0.01 (Y-axis) and log2 > 2 (X-axis). The data shown are for one experiment based on 3 biological replicates. The bait protein is in red. No other proteins were significantly enriched in this experiment. Proteins below the significance thresholds are grey.
4.4 Proteomic investigation of the TALE protein designed to bind Ingi transposable elements

4.4.1 The Ingi-TALE failed to significantly enrich any chromatin proteins

In contrast to telomeres and centromeres, it is unknown whether transposable elements (TEs) are bound by specific proteins in *T. brucei*. As discussed in Section 1.3.2, trypanosomes do not have KRAB-ZFPs or DNA methylation, two of the primary mechanisms used to silence TEs in other eukaryotes. As in many other eukaryotes, TEs are known to be silenced post-transcriptionally by the RNA interference machinery in trypanosomes. However, as mentioned in Section 1.4.5, TbAGO1 is cytoplasmic (Shi et al., 2004; Staneva et al., 2021), and is therefore unlikely to participate directly in the generation of repressive chromatin over these dispersed repetitive elements, in contrast to the role of Argonaute proteins in other eukaryotes. This does not exclude the possibility that other, yet to be identified proteins could mediate chromatin or DNA modifications at these loci. To explore this possibility, affinity purification coupled to liquid chromatography tandem mass spectrometry (LC-MS/MS) of the Ingi-TALE was performed. Unfortunately, there was no enrichment of histones or other chromatin-associated or nuclear proteins in these affinity purifications.
Figure 4.8. Ingi-TALE failed to interact with histones or chromatin associated proteins. Significance thresholds (dotted lines) used to determine significantly enriched proteins are $p < 0.01$ (Y-axis) and log2 $> 2$ (X-axis). The data shown are for one experiment based on 3 biological replicates. A full list of the proteins enriched in this plot can be found in Table 4.4. Enriched proteins are in black. Proteins below the significance thresholds are grey.

Table 4.4. Proteins enriched in Ingi-TALE pulldown

All proteins above the significance thresholds defined in Figure 4.8 are listed below along with their UniProt ID, EuPathDB gene code (Tb927 genome), and name as listed in the TriTrypDB. Proteins are colour coded as in Figure 4.8. Proteins have been ranked from highest to lowest log2 fold change.

<table>
<thead>
<tr>
<th>Protein ID</th>
<th>Gene ID</th>
<th>Name</th>
<th>Student’s T-test difference</th>
<th>$-\log$ student’s T-test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>D6XFL0</td>
<td>Tb927.4.2740</td>
<td>p25-alpha, putative</td>
<td>12.12444878</td>
<td>3.419975009</td>
</tr>
<tr>
<td>Q9NDH8</td>
<td>Tb427.10.2890</td>
<td>enolase</td>
<td>2.973535538</td>
<td>3.606722715</td>
</tr>
</tbody>
</table>

The overall enrichment of proteins with the Ingi-TALE was very low, with even the bait protein being below the significance thresholds set (not labelled in Figure 4.8). Further, there were no histones or other chromatin proteins present, suggesting a poor purification. The two proteins that were enriched are likely contaminants. For example, enolase is a glycolytic enzyme. Based on these results, and time constraints, it was decided not to perform any more work on the Ingi-TALE. Extensive optimisation will be required to determine if the Ingi-TALE can successfully affinity purify proteins associated with trypanosome TEs. Possible reasons for the lack of protein enrichment and potential optimisation strategies are discussed in Section 4.6.5.
4.5 Proteomic identification of proteins associated with the minichromosome 177 bp repeat

4.5.1 Minichromosome-TALE enriches for known centromeric proteins

To identify proteins associated with the 177 bp repeats on minichromosomes, affinity purification coupled to liquid chromatography tandem mass spectrometry (LC-MS/MS) of the Minichromosome-TALE was carried out, identifying eight kinetochore proteins as well as three cohesin subunits. Proteins of interest will be discussed further below. A full list of significantly enriched proteins is provided in Table 4.5.

Figure 4.9. Minichromosome-TALE associates with kinetochore and cohesin components. Significance thresholds (dotted lines) used to determine significantly enriched proteins are \( p < 0.01 \) (Y-axis) and \( \log_2 > 2 \) (X-axis). The data shown are for one experiment based on 3 biological replicates. Proteins with similar known functions have been grouped by colour. Proteins of particular interest are labelled where possible. A full list of the proteins enriched in this plot can be found in Table 4.5 below. The bait protein is in red. Kinetochore proteins (KKTs) are in orange. Histones and histone variants are in green. Members of the cohesin complex are in blue. Proteins outside of these chosen categories are in black. Proteins below the significance thresholds are grey.
Table 4.5. Proteins enriched in Minichromosome-TALE pulldown

All proteins above the significance thresholds defined in Figure 4.9 are listed below along with their UniProt ID, EuPathDB gene code (Tb927 genome), and name as listed in the TriTrypDB. Proteins are colour coded as in Figure 4.9. Proteins have been ranked from highest to lowest log$_2$ fold change. Q9U6Y5 (red) is present as a marker for the Minichromosome-TALE protein, as TALE constructs generated in this study are not present in the UniProt database, whereas peptides belonging to the YFP tag on the C-terminus of the TALE construct can be identified.

<table>
<thead>
<tr>
<th>UniProt ID</th>
<th>EuPathDB Gene ID</th>
<th>Name</th>
<th>Log2FC</th>
<th>-log student’s t-test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q9U6Y5</td>
<td>N/A</td>
<td>GFP-like fluorescent chromoprotein FP506</td>
<td>9.0692</td>
<td>4.131934367</td>
</tr>
<tr>
<td>Q38D71</td>
<td>Tb927.9.10920</td>
<td>kinetoplastid kinetochore protein 3 (KKT3)</td>
<td>7.6791</td>
<td>3.689561467</td>
</tr>
<tr>
<td>Q57U85</td>
<td>Tb927.5.3510</td>
<td>structural maintenance of chromosome 3, putative (SMC3)</td>
<td>6.4017</td>
<td>3.804779413</td>
</tr>
<tr>
<td>Q383M7</td>
<td>Tb927.11.10520</td>
<td>kinetoplastid kinetochore protein 2 (KKT2)</td>
<td>5.7343</td>
<td>3.521908632</td>
</tr>
<tr>
<td>Q38D91</td>
<td>Tb927.9.11850</td>
<td>structural maintenance of chromosome 1, putative (SMC1)</td>
<td>5.6662</td>
<td>2.138427427</td>
</tr>
<tr>
<td>Q387X7</td>
<td>Tb927.10.15350</td>
<td>histone H3 variant V (H3V)</td>
<td>5.0751</td>
<td>2.410668306</td>
</tr>
<tr>
<td>Q57X80</td>
<td>Tb927.7.6900</td>
<td>Sister chromatid cohesion protein 1 (SCC1)</td>
<td>4.3201</td>
<td>3.006011603</td>
</tr>
<tr>
<td>Q57Y3A</td>
<td>Tb927.7.2820</td>
<td>histone H2A, putative (H2A)</td>
<td>4.154</td>
<td>2.689127251</td>
</tr>
<tr>
<td>Q389T1</td>
<td>Tb10.406.0330</td>
<td>histone H2B, putative (H2B)</td>
<td>4.0503</td>
<td>2.565934889</td>
</tr>
<tr>
<td>Q584U5</td>
<td>Tb927.4.5110</td>
<td>kinetoplastid kinetochore protein 8 (KKT8)</td>
<td>3.9494</td>
<td>2.209608088</td>
</tr>
<tr>
<td>Q57Z31</td>
<td>Tb927.5.4170</td>
<td>histone H4, putative (H4)</td>
<td>3.8468</td>
<td>2.483721266</td>
</tr>
<tr>
<td>Q57V91</td>
<td>Tb927.8.1150</td>
<td>kinetoplastid kinetochore protein 9 (KKT9)</td>
<td>3.774</td>
<td>3.029867048</td>
</tr>
<tr>
<td>Q57Y00</td>
<td>Tb927.7.2110</td>
<td>kinetoplastid kinetochore protein 11 (KKT11)</td>
<td>3.6047</td>
<td>2.229481403</td>
</tr>
<tr>
<td>Q38A3X</td>
<td>Tb927.10.6330</td>
<td>kinetoplastid kinetochore protein 1 (KKT1)</td>
<td>3.4432</td>
<td>3.003182145</td>
</tr>
<tr>
<td>Q387E7</td>
<td>Tb927.11.1030</td>
<td>kinetoplastid kinetochore protein 7 (KKT7)</td>
<td>3.4323</td>
<td>2.449891606</td>
</tr>
<tr>
<td>Q587H6</td>
<td>Tb927.2.2670</td>
<td>histone H4 variant (H4V)</td>
<td>3.2615</td>
<td>2.167638597</td>
</tr>
<tr>
<td>Q382U0</td>
<td>Tb927.11.12410</td>
<td>kinetoplastid kinetochore protein 10 (KKT10)</td>
<td>3.202</td>
<td>2.191484053</td>
</tr>
<tr>
<td>Q388W1</td>
<td>Tb927.10.12850</td>
<td>ttagg binding factor (TRF)</td>
<td>3.0103</td>
<td>2.316529188</td>
</tr>
<tr>
<td>Q4GYX7</td>
<td>Tb927.1.2430</td>
<td>histone H3, putative (H3)</td>
<td>2.9983</td>
<td>2.240069478</td>
</tr>
<tr>
<td>Q383E9</td>
<td>Tb927.11.11290</td>
<td>heat shock protein 70, putative (HSP70.c)</td>
<td>2.9841</td>
<td>2.153645241</td>
</tr>
<tr>
<td>Q383U2</td>
<td>Tb927.11.9870</td>
<td>Telomere-associated protein 1 (TelAP1)</td>
<td>2.9149</td>
<td>2.005065359</td>
</tr>
</tbody>
</table>

Summary of published information on the identified proteins with respect to minichromosome association:

- **Kinetochore proteins (orange).** The most highly enriched kinetochore proteins were KKT3 and KKT2, which are constitutively present at the kinetochore, and are hypothesised to directly bind centromeric DNA (Akiyoshi & Gull, 2014b). Both of these proteins were shown by Akiyoshi and Gull via ChIP-seq to be enriched on a ‘synthetic’ minichromosome comprised of 177 bp repeats. KKT8, KKT9, KKT10, and KKT11 localize at kinetochores from S-phase to metaphase, and KKT1 and KKT7 are found at the kinetochore from S-phase to anaphase. RNAi knockdown of KKT10/19 (due to their similarity both proteins were targeted by the same construct) was shown to cause minichromosome segregation defects (Akiyoshi & Gull, 2014b). These data provide the only evidence available connecting any known kinetochore proteins to...
minichromosome segregation. ChIP-seq of KKT10/19 could identify whether either protein is bound to 177 bp repeats.

- **Cohesin (blue).** The same three cohesin subunits identified in the TRF and KKT2 pulldowns were also enriched in this experiment: SMC1, SMC3, and SCC1. Association of cohesin with *T. brucei* minichromosomes has not been reported. Based on FISH analysis of the 177 bp repeats, RNAi knockdown of SMC3 does not appear to affect the segregation of minichromosomes (Bessat & Ersfeld, 2009). Any potential roles for SMC1 or SCC1 in *T. brucei* minichromosome segregation have not been investigated further here or elsewhere.

- **Histones (green).** Similar to the Telomere TALE, large amounts of histones and histone variants were identified. These include the histone H3 variant, which is known to be found at minichromosomes (Akiyoshi & Gull, 2014b; Lowell & Cross, 2004). There is no published evidence for the presence of H4V on minichromosomes, but it has not been investigated. ChIP-seq data are available for H4V (Müller et al., 2018), and could be analysed for the enrichment of reads binding the 177 bp repeats relative to input sequences.

- **Telomere proteins (not coloured in Figure 4.9).** Surprisingly, the telomere proteins TRF and TelAP1 were also enriched in the Minichromosome-TALE affinity purification. No evidence of association of either of these proteins with minichromosomes has been reported in the literature, but it would be reasonable to assume that they are bound to telomeres on minichromosomes as well as megabase chromosomes. Potential explanations as to why these proteins were identified in this experiment are discussed in Section 4.6.6.
4.5.2 KKT2 associates with both minichromosomal 177 bp repeats and CIR147 centromere repeats

The ChIP-seq data generated by Akiyoshi and Gull showing the enrichment of KKT2 on 177 bp repeats was performed in procyclic form cells. It is not known whether KKT2 is enriched on 177 bp repeats in bloodstream form cells. To validate my finding that KKT2 was present on 177 bp repeats in bloodstream forms, ChIP-qPCR was performed with YFP-KKT2 using primers for the CIR147 centromere repeats (positive control), 177 bp repeats, and two negative control loci (H3 genes and Actin B).

Figure 4.10. YFP-KKT2 is enriched on minichromosome 177 bp repeats as well as CIR147 centromere repeats. Data are mean ± standard deviation from three biological replicates. X-axis labels represent primers designed for each locus. CIR147 primers detect centromere repeats from megabase chromosomes 4, 5, 8, 9, 10, and 11.

High levels of enrichment of YFP-KKT2 on 177 bp repeats were detected, comparable to those seen on CIR147 repeats from the megabase chromosome centromeres, bolstering my proteomics findings from the previous section. This level of enrichment suggests large amounts of KKT2 are indeed bound to 177 bp repeats. It is therefore reasonable to conclude that kinetochores may assemble on the 177 bp repeats of minichromosomes.
4.6 Discussion

4.6.1 Affinity purification of endogenous telomere-binding protein TRF

Overall, this experiment was very successful, identifying all known telomere proteins from the literature, as well as many new TRF-interactor proteins not reported previously. These novel interactors included MRE11, RAD50, Cohesin, HDAC3, PIP5Pase, VEX1, and JBP1, all of which have plausible telomere connections based on information in the literature and/or their roles in other eukaryotes. However, the roles of many of these proteins at telomeres have not been explored, and much work remains to be done in the future to elucidate their functions.

There were several proteins among the less enriched (log2FC < 4) which were seen in this and subsequent repeats of the TRF pulldown (data not shown) that will be discussed here due to their repeated appearance in these experiments. These included the importin alpha subunit KAP60, the kinetochore protein KKT3, the microtubule-associated protein GB4, and the dynein light chain LC8 (DYNLL1).

KAP60 has been shown to interact with the NLS of RAP1 by yeast two-hybrid assay (Afrin, Kishmiri, et al., 2020), and to be enriched in nuclear pore complex preparations (DeGrasse et al., 2009). As discussed in Section 1.3, tethering genes to the nuclear periphery is associated with transcriptional repression in other eukaryotes. However, tethering to the NPC has been shown to mediate both gene activation and repression depending on the locus (Buchwalter et al., 2018). Trypanosome telomeres are known to reside primarily at the nuclear periphery, but there is no evidence in the literature of telomeres interacting with nuclear pore complexes (NPCs). How exactly trypanosome telomeres are tethered to the nuclear periphery is unknown. Based on this data, one could hypothesise that this is mediated via the NPC.

The association of KKT3 with TRF has never been reported in the literature. As discussed in Section 4.3.1, interactions between the kinetochore and NPC are seen in other eukaryotes. Therefore, it is possible that the KKT3-TRF interaction, if real, occurs at the nuclear periphery. This could occur at anaphase, when the kinetochores move to the spindle poles at the periphery of the new daughter nuclei (Akiyoshi & Gull, 2014b).

GB4 is a microtubule-associated protein which appears to bind microtubules at their plus end, and is located almost exclusively at the posterior cell tip (Rindisbacher et al., 1993). It was originally considered a contaminant in these preparations, but was repeatedly seen here and in experiments performed for Chapter 5. GB4 has never been reported to be associated with telomeres, and any potential role it may have is unclear.
Similar to GB4, DYNLL1 was originally considered a contaminant, but due to repeated appearance in TRF pulldowns, further literature investigation was performed. Although presumably part of the dynein motor, DYNLL1 is also a subunit of a kinetoplastid-specific RNAPI transcription factor, class I transcription factor A (CITFA), and knockdown of DYNLL1 results in cell cycle defects, a decrease in VSG mRNA and rRNA levels, and cell death (Brandenburg et al., 2007; Kirkham et al., 2015; Nguyen et al., 2012). These data are consistent with a role for DYNLL1 at sub-telomeric promoters, but not necessarily telomere repeats. It is also interesting that no other members of the CITFA complex were identified.

4.6.2 Affinity purification of synthetic telomere-binding protein Telomere-TALE

The Telomere-TALE purification identified all bar one known telomere protein, showing that this approach can efficiently purify telomeres. This proof of principle gave us the confidence to expand the approach to other loci of interest. As discussed in Section 1.6, similar approaches have been used to purify telomeres in other organisms, but this is the first time any locus-specific purification has been attempted, or successfully implemented, in trypanosomes.

In comparison to a published data set which characterised proteins bound by a telomere repeat oligonucleotide in vitro, the Telomere-TALE gave a much more extensive list of proteins (Reis et al., 2018). Such oligo-based approaches cannot purify the histones associated with a locus, whereas the Telomere-TALE identified large amounts of histones. Although not performed in this project, analysis of histone modifications is key to understanding the chromatin environment, and these data indicate that TALEs could be a useful tool for investigating the histone modifications present at telomeres and other loci. Very little is known about the genome-wide distribution of histone modifications in T. brucei, with the exception of a small number of modifications discussed in Section 1.3.1, so this technology could be a useful tool in the future.

Many of the chromatin-associated proteins identified in the TRF pulldown were absent from this experiment. These included the MRN complex proteins, VEX1, PIP5Pase, and Cohesin. The reasons for this are unclear, as the presence of core telomere proteins and RPA subunits in the Telomere-TALE purification suggest an ability to purify non-histone chromatin-associated proteins.

The presence of ZC3H39 and ZC3H40 in both this pulldown and the TRF pulldown is both novel and unexpected, and suggests that these zinc finger proteins may indeed bind
telomeres or telomere-associated proteins. Based on the RNAi phenotype discussed in Section 4.2.1, they could be performing a silencing role at telomeres.

4.6.3 Affinity purification of endogenous kinetochore-centromere protein KKT2

Of the known kinetochore components in *T. brucei* (as determined under native conditions), only KKIP1, KKT22, KKT23, and KKT24 were absent. KKIP1 localizes at kinetochores from S-phase to anaphase (D’Archivio & Wickstead, 2017). KKT22 and KKT23 are constitutively localized at the kinetochore, while KKT24 is localized at the kinetochore from S-phase to anaphase (Nerusheva et al., 2019). Why exactly these kinetochore components were absent when all other known kinetochore proteins (excluding the remaining KKIP/KOK proteins, which may or may not be *bona fide* kinetochore proteins) were present is unclear. However, KKT22-24 were not identified in the original N-terminal tagging and immunoprecipitation screen performed by Akiyoshi and Gull, and were identified subsequently via immunoprecipitation of a C-terminally tagged KKT3. Therefore it is possible that the N-terminal YFP tag used on KKT2 in this experiment somehow interferes with KKT22-24 interactions in a similar manner. Similarly, KKIP1 was not identified in the aforementioned screen either, and has never been seen to interact with KKT proteins without using chemical crosslinking. Thus, its absence could be due to tag-interference effects as well. The other possibility is spatial, i.e. KKIP1 is thought to bridge the inner and outer kinetochore complexes, and may be too distant from KKT2 in the 3-dimensional architecture of such a large macromolecular complex to interact directly (Brusini et al., 2021).

The presence of KKIP5 but absence of other KKIP proteins is interesting. KKIP5 is only localized to kinetochores from S-phase to metaphase, and according to Zhou et al was only visible by immunofluorescence in approximately 12% of cells under normal conditions (Q. Zhou et al., 2019). They found this level increased in response to DNA damage, and went on to characterise the role of KKIP5 in modulating a DNA damage-induced metaphase checkpoint in *T. brucei*. Given its supposed scarcity, the identification of KKIP5, but absence of other KKT and KKIP proteins that have more consistent kinetochore localization, warrants discussion. It suggests either that KKIP5 is a more central component of kinetochores than previously reported, or that DNA damage had somehow accumulated in these cultures either during growth or during the centrifugation and washing steps prior to cell lysis. Since mass spectrometry is orders of magnitude more sensitive than immunofluorescence, it is possible that small amounts of KKIP5 are present at kinetochores in all cells, and that this is not detectable by immunofluorescence. It seems unlikely that the cells could have accumulated
abnormal levels of DNA damage in culture while simultaneously continuing to grow exponentially, and no DNA damaging agents were added to the media, for example, through using phleomycin for selection.

4.6.4 Affinity purification of synthetic centromere protein Centromere-TALE

The failure of the Centromere-TALE to enrich for chromatin proteins such as histones, let alone the expected kinetochore proteins, could be due to several factors. The most likely are: 1) suboptimal TALE expression levels; 2) Insufficient locus enrichment; and 3) locus targeting issues.

As discussed briefly in Section 4.3.2, there was concern that the centromere TALE was in a large excess inside the cell relative to the number of binding sites, causing the majority of the protein to be unbound to its target site. This would result in immunoprecipitating primarily the unbound fraction, as this would be in the majority, and would also not be subjected to any potential solubilisation issues that the chromatin bound fraction might face. Doubling the amount of sonication to improve chromatin solubilisation, as well as doubling the amount of anti-GFP-crosslinked dynabeads, failed to give the improved results. However, it seems unlikely that centromeres per se are difficult to purify, as KKT2 and associated chromatin proteins were successfully purified using the same protocol. Several cell lines containing different 3′UTRs which might give lower protein levels were generated, but quantitative measurements of protein levels between these and the original construct were never performed, and so these cell lines were never carried forward for ChIP and proteomics analysis.

For issues relating to insufficient locus enrichment, one key factor is the copy number of the locus being purified. The average centromere length, based on restriction endonuclease mapping, ranges from 20-120 kb (Echeverry et al., 2012). If we estimate an average centromere to be 70 kb x 6 chromosomes with CIR147 repeats x 2 alleles = 820 kb of CIR147 repeats (as discussed in Section 1.2.3, centromeres 4, 5, 8, 9, 10, and 11 are comprised of CIR147 repeats. The other centromeres are comprised of different repeats). In contrast, there is approximately 244 x 10kb = 2440 kb of telomeric sequence available for binding by a TALE. There is also a theoretical difference between the two when looking at the number of possible TALE binding sites within a given length; e.g. 147 bp of centromere sequence only contains one TALE binding site, whereas 147 bp of telomere sequence could theoretically bind 7 TALE proteins due to it having a 6bp repeat length. Thus, 820kb of CIR147 repeat could bind ~5,500
TALE molecules, whereas 2440kb of telomeric repeats could bind ~122,000 TALE molecules, a greater than 22 fold difference in theoretical occupancy. These estimates would be affected by the presence of endogenous centromere/telomere binding proteins and other chromatin-associated proteins, but the discrepancies between the two loci in terms of these factors are highly unlikely to outweigh the more than 20 fold difference in hypothetical TALE binding capacity.

Locus targeting issues are another possibility. In the ChIP-qPCR data shown in Section 3.7, the Centromere-TALE is enriched on CIR147 repeats. However, the ChIP was performed using formaldehyde crosslinking, whereas the immunoprecipitation performed prior to MS analysis was performed under native conditions. It is possible that, for unknown reasons, this TALE only bound its locus weakly, and that other 15 bp target sequences from centromeres might perform better.

It was identified after the completion of the project that the 15 bp target sequence of the Centromere-TALE contained one mismatch to the CIR147 repeats on chromosome 8, and 2 mismatches to the repeats present on the other megabase chromosomes. This occurred due to a discrepancy between the genome sequence available at the time of the original analysis (TREU 927 isolate) and the genome sequence of the *T. brucei* isolate actually used in this research (Lister 427). These mismatches may explain the poor performance of the Centromere-TALE in this immunoprecipitation experiments, particularly in light of the absence of any chemical crosslinking, as mismatches are known to reduce TALE binding efficiency. It is possible that formaldehyde crosslinking stabilised weak interactions of the Centromere-TALE with CIR147 repeats, and this is why the ChIP-qPCR results from Chapter 3 were positive. However, the lack of crosslinking in the experiments performed for this Chapter could have led to a loss of binding strength, and therefore poor purification for proteomics.

### 4.6.5 Affinity purification of synthetic Ingi transposable element-binding protein Ingi-TALE

As with the Centromere-TALE, copy number could be an important factor influencing the failure of the Ingi-TALE to identify chromatin-associated proteins. The 79 bp signature to which the TALE protein is designed to bind is only present approximately 600 times within the cell (Bringaud et al., 2008). This figure may be larger in the more complete Tb427 genome from 2018, but a repeat of this analysis has not been performed. This means that the signal-to-
noise ratio for purifying proteins from this locus is hundreds of times lower than for telomeres, for example. As a result, purifying this locus is objectively much more challenging.

Another possibility that could explain the lack of identified proteins is that there are simply no proteins which specifically bind TEs in *T. brucei*. Given the fact that there are only a handful of TEs within the genome that are fully functional, this could well be the case. That is, siRNA-mediated silencing may be sufficient to silence these, and the cells never required the evolution of more sophisticated chromatin-mediated silencing. The high density of TEs at transcription start regions would likely be refractory to chromatin-mediated silencing, as this could negatively impact transcription of nearby polycistrons.

The fact that the bait protein was so poorly enriched suggests that the immunoprecipitation in the Ingi-TALE cell line was not very efficient, as even if no proteins do interact with the locus, one would expect enrichment of the bait protein to be higher. This could potentially be an issue of chromatin-solubilization. Additional cycles of sonication, or nuclease treatment, could improve this.

### 4.6.6 Affinity purification of synthetic minichromosomal 177 bp repeat-binding protein Minichromosome-TALE

The identification of kinetochore proteins associating with minichromosomes is a novel finding, and runs counter to the primary model of minichromosome segregation in the literature. Minichromosomes are hypothesised to segregate via a different mechanism to the megabase chromosomes. This model, termed the lateral stacking model, posits that megabase chromosomes are segregated via conventional end-on attachments to spindle microtubules, but minichromosomes segregate via side-on attachments to non-kinetochore spindle microtubules, with motor proteins transferring them to the poles of the two daughter nuclei (Gull et al., 1998). This model arose in response to several observations: (i) the large discrepancy between the number of kinetochores and the number of chromosomes; (ii) the large discrepancy between the number of spindle microtubules and the number of chromosomes; and (iii) differences in the movement of minichromosomes and megabase chromosomes throughout the cell cycle (Ersfeld & Gull, 1997; Gull et al., 1998). The FISH analysis performed by Ersfeld and Gull established that minichromosomes appeared to be segregated equally between daughter cells, and so simpler segregation mechanisms, such as random dispersion to daughter cells, were disregarded. Later work concluded that
minichromosomes were indeed segregated with fidelity over many generations (Wickstead et al., 2003).

Although far from a complete repertoire of kinetochore proteins were identified in this experiment, the proteins identified cover a range of potential functions based on their differences in localisation at the kinetochore throughout the cell cycle, lending weight to the argument that minichromosomes may assemble a full kinetochore. It is unclear whether the absence of the remaining KKT and KKIP proteins from this experiment is due merely to a lack of enrichment, or to genuine absence of certain proteins from minichromosome kinetochores. Although it is possible that minichromosomes could assemble a simplified kinetochore compared with megabase chromosomes, perhaps lacking proteins involved in regulation of checkpoints for example, the former hypothesis seems more reasonable. ChIP-qPCR and ChIP-seq experiments such as that shown in Section 4.5.2 could be performed for all kinetochore proteins to assay whether they are bound to minichromosomal 177 bp repeats. Akiyoshi and Gull state that their finding of KKT2 and KKT3 enrichment on 177 bp repeats could just be due to spatial proximity and not genuine DNA binding, but this could be questioned given the use of a short-range crosslinker such as formaldehyde (Akiyoshi & Gull, 2014b).

If minichromosomes do indeed function as centromeres, and assemble kinetochores, then the association of cohesin with minichromosomes seen in this experiment would be expected. ChIP-seq data is available for SCC1 (Müller et al., 2018) and could be analysed for the presence of reads binding the 177 bp repeats. It is also worth noting that depletion of separase led to mis-segregation of minichromosomes, suggesting a role for cohesin in their segregation (Bessat & Ersfeld, 2009). In budding yeast, it has been shown that cohesin-based cohesion of minichromosomes is superior for circular minichromosomes as compared with linear minichromosomes, and this is likely due to spindle forces causing cohesin to slide away from the centromere and off the small minichromosomes (Farcas et al., 2011; Ivanov & Nasmyth, 2005). If we were to apply this model to trypanosomes, then we would expect minichromosomes to segregate earlier than megabase chromosomes in mitosis, as they would be pulled apart by spindle forces earlier than the megabase chromosomes due to a loss of cohesion. This could be tested by generating a synthetic circular minichromosome of a similar size to endogenous minichromosomes, comprising primarily 177 bp repeats, but also containing a unique marker sequence such as a LacO array to facilitate its identification from the endogenous minichromosome population. Therefore, movement of bulk minichromosomes could be monitored via FISH for the 177 bp repeats or minichromosome-specific VSG genes, while movement of the synthetic chromosome could be monitored via FISH for the LacO array or a LacI-GFP construct. Analysis of chromosome movements
throughout the cell cycle could then be analysed in cohesin knockdowns, and with the synthetic minichromosome in a circular or linear form. All of these experiments would also have to be performed with a minichromosome containing centromere repeats from the megabase chromosomes, e.g. CIR147 bp repeats, in order to draw any conclusions about the influence of cohesin and cohesion on megabase chromosomes.

As mentioned in Section 4.5.1, the identification of telomere proteins in the minichromosome-TALE pulldown was a surprise. Since minichromosomes have telomeric repeats, it is possible that one TALE-bound DNA fragment could contain both 177 bp repeats and TTAGGG repeats, with KKT proteins and cohesin bound to the 177 bp repeats, and telomeric proteins bound to the TTAGGG repeats. Cell lysates are only subjected to a small amount of sonication (see Section 2.6.2), and so chromatin would not be expected to be sheared to small fragment sizes. The other possibilities are that: (i) TRF and TelAP1 bind the 177 bp repeats (ii) TRF and TelAP1 interact with one (or more) of the other proteins present (iii) a significant amount of the 177TALE is binding off-target to telomeric repeats.

Although it is possible that TRF and TelAP1 bind the 177 bp repeats, it seems unlikely given both the highly specific sequence at telomeres and the specialized role of telomere proteins within the cell. TRF in particular, which binds the TTAGGG repeats, would have to accept several mismatches in its DNA-binding myb-domain in order to bind a different sequence to any appreciable amount. It is unknown whether TelAP1 binds telomeric repeats directly, or is merely a binding partner of TRF.

As seen in Section 4.2.1, the kinetochore protein KKT3 was identified in the TRF pulldown, and this interaction was seen repeatedly (data not shown). If this interaction is genuine, then it is possible that TRF and TelAP1 are immunoprecipitating with the Minichromosome-TALE due to an interaction with KKT3.

The 15 bp sequence targeted by the Telomere-TALE is AGGGTTAGGGTTAGG, while the 15 bp sequence targeted by the Minichromosome-TALE is GCCGAGTTAACGCTA. The longest stretch of identical sequence between these two target sites is 4 bp (underlined), but the surrounding residues differ substantially. Further, the absence of cytosine residues in the telomere repeats would require all four of the cytosine-binding RVDs of the TALE to bind to other bases. One investigation of TALE mismatch tolerance found that even 3 mismatched RVDs led to a severe decrease in TALE binding, and 6 mismatches dropped binding efficiency down to 1% of a TALEN containing no mismatches for the same sequence (Juillerat et al., 2014). For the Telomere-TALE to bind the Minichromosome-TALE target sequence would require 11 mismatches to be tolerated, which appears unlikely based on the analysis of Juillerat et al.
The size of the 177bp repeat region on the roughly 100 minichromosomes is 20-80 kb based on restriction mapping and Sanger sequencing of 17 minichromosomes by Wickstead et al (Wickstead et al., 2004). If we approximate an average of 50 kb, then there is 5000 kb of 177bp repeat within the cell. This gives a theoretical maximum of ~28,000 binding sites for the TALE protein, placing it between telomeres and CIR147 repeats in terms of total number of TALE binding sites. When looked at through this lens, we can paint a picture of diminishing success with these experiments as we move from highest to lowest number of TALE binding sites.

4.6.7. Conclusions

The two major, novel findings of the work in this chapter were the identification of ZC3H39 and ZC3H40 interacting with telomeres, and the presence of kinetochore proteins on the 177 bp repeats of minichromosomes. Due to time limitations imposed by a mandatory 3-month industrial placement, and later the COVID19 pandemic and associated restrictions, I decided to focus on just one of these for the remainder of project. In consultation with my thesis committee, it was decided that characterising the role of ZC3H39 and ZC3H40 at telomeres might be more tractable, and should be the focus of the remainder of my PhD.
Chapter 5: Characterisation of the role of ZC3H39 and ZC3H40 at Telomeres

5.1 Introduction

In Chapter 4, immunoprecipitation coupled to mass spectrometry analysis of the four TALE cell lines and two control proteins (TRF, KKT2) had varying degrees of success, with the Telomere- and Minichromosome-TALEs both generating interesting data sets. Many novel interacting proteins were also identified in the control experiments with TRF and KKT2, but these proteins were not the primary focus of this project, and so the majority of these novel interactions have not been explored further. However, one novel interaction, that of ZC3H39 and ZC3H40 with both the Telomere-TALE and TRF, will be explored further in this chapter.

The identification of ZC3H39 and ZC3H40 at telomeres via two separate immunoprecipitation experiments of distinct telomere-binding proteins, and their absence from other TALE and YFP-tagged protein immunoprecipitations, suggested that this interaction was not due to non-specific binding or adsorption of the proteins to TALE proteins or YFP tags in general. Their identification with the Telomere-TALE and TRF, both of which are DNA-binding proteins, suggests some fraction of ZC3H39 and ZC3H40 has a close proximity to telomeric chromatin. ZC3H39 and ZC3H40 are RNA-binding proteins, therefore it is possible that they bind transcripts or TERRA RNA transcribed from the VSG ESs, for example (Damasceno et al., 2017b).
Multiple data sets were available to analyse in order to investigate this question (Trenaman et al., 2019). These included crosslinking immunoprecipitation (CLIP)-seq data, where RNAs bound to ZC3H39 and ZC3H40 were sequenced, and whole transcriptome data from cells where ZC3H39 and/or ZC3H40 were knocked out/down. Dr Marcel Lafos (Allshire Lab) searched for reads aligning to telomeric repeats in all of these data sets in order to ask whether: 1) Do ZC3H39 and ZC3H40 bind TERRA?; and 2) Does loss of ZC3H39 and/or ZC3H40 alter the level of TERRA compared to wild-type cells? In their original analyses performed by Trenaman et al., the filtering steps would have excluded reads mapping to repetitive sequences, and so these questions were not addressed. The analyses performed by Dr Lafos did not identify any abundance of telomere reads in the CLIP-seq data, or changes in reads mapping to telomeres in the RNA-seq data (data not shown). This suggested that it was unlikely ZC3H39 and ZC3H40 bind TERRA RNAs, or regulate their levels.

Since loss of ZC3H39 and ZC3H40 following RNAi led to de-repression of telomere-proximal silent VSG ESs (Trenaman et al., 2019), and both proteins appear to associate with telomeres (Chapter 4), it was reasoned that: 1) ChIP-qPCR using primers for the telomeric repeats might show enrichment of these proteins at telomeres; and 2) Immunoprecipitation of these proteins may identify TRF or other telomere proteins. Therefore, cell lines expressing epitope tagged versions of each protein were generated and carried forward to explore these possibilities.

5.2 YFP-tagged ZC3H39/40 localizes to the cytoplasm in bloodstream form parasites

YFP-ZC3H39 and YFP-ZC3H40 cell lines were generated in order to investigate the subcellular localisation of both proteins. Proteins were tagged at the endogenous locus as in Staneva et al. (Staneva et al., 2021). This had been investigated previously by the Horn lab as well as the Tryptag initiative (Dean et al., 2017; Trenaman et al., 2019), however it was important to verify my own constructs to determine if they generated a localisation that was consistent with that observed in other studies. In particular, I wanted to determine if there was any protein present in the nucleus, as suggested by the results of my TRF and Telomere-TALE proteomics analysis.
Figure 5.1. Immunofluorescence of ZC3H39 and ZC3H40 indicates that each protein has a cytosolic distribution. Proteins were immunolocalised in bloodstream form trypanosomes using a polyclonal anti-GFP primary antibody and an Alexafluor-488 labelled secondary antibody (green in merged panel). Nuclear and kinetoplast DNA were stained with DAPI (blue in merged panel). YFP-TRF was used as a control and a marker of telomeres. Images chosen are representative of the cell population as a whole. DAPI and FITC (Fluorescein isothiocyanate) refer to fluorophores used for excitation.

The observed subcellular distribution of ZC3H39 and ZC3H40 (Figure 5.1) was consistent with that seen elsewhere, and did not show any similarity to the distribution of the known telomere-binding protein TRF. It is tempting to speculate that some of the protein may be localised to the extreme nuclear periphery, however the resolution of these images does not allow such a conclusion to be drawn.

Extensive attempts were made to co-stain ZC3H39 and ZC3H40 with either TRF or TelAP1 using antibodies generated against these proteins (provided by Dr Christian Janzen, University of Wurzburg). However, issues with a high background signal plagued these experiments, making the data uninterpretable (data not shown).
Some basic attempts were also made to wash out cytosolic proteins prior to fixation and antibody staining, with the intention to remove cytosolic ZC3H39 and ZC3H40, whilst retaining any nuclear ZC3H39 and ZC3H40 in place for staining. However, these efforts at detergent extraction did not reveal any obvious nuclear staining in the ZC3H39 and ZC3H40 cells, and so were abandoned (data not shown).

5.3 Chromatin Immunoprecipitation of ZC3H39/40 shows no enrichment on telomeric repeats

Despite the lack of any detectable nuclear localisation of ZC3H39 and ZC3H40 in the previous section, ChIP-qPCR was performed on both YFP-tagged ZC3H cell lines to assess if ZC3H39 and ZC3H40 were associated with telomeric repeats. This is a more sensitive assay than immunofluorescence, and could theoretically identify small amounts of ZC3H39 and ZC3H40 bound to telomere repeats. YFP-TRF was used as a positive control for a telomere-binding protein.

Figure 5.2. Quantitative Chromatin Immunoprecipitation of YFP-ZC3H39, YFP-ZC3H40, YFP-TRF, and untagged (WT) cells shows no enrichment of ZC3H39 or ZC3H40 at telomeric repeats. Data are mean ± standard deviation from three biological replicates. X-axis labels represent primers designed to detect association with each locus. Cells were crosslinked in 1% formaldehyde prior to lysis, sonication, immunoprecipitation, and DNA purification as outlined in Section 2.5.
No enrichment of either ZC3H39 or ZC3H40 was detected on telomeric repeats, in contrast to the known telomere binding protein TRF (Figure 5.2). If the association of ZC3H39 and ZC3H40 with telomeres is indirect, then crosslinking would likely be less efficient than for a protein that directly binds the telomeric repeats. Therefore, the formaldehyde concentration was increased from 1% to 2% to assay whether increased levels of crosslinking might be required in order to allow ZC3H39 and ZC3H40 to stably associate with DNA. However, even under these conditions, no enrichment of either ZC3H protein with TTAGGG telomeric repeats was detected (data not shown), although it remains formally possible that association with other DNA sequences occurs.

5.4 ZC3H39 associates with ZC3H40 and RNA editing components

Despite the lack of enrichment of telomeric repeats with ZC3H39 and ZC3H40 (Figure 5.2), it was decided that identifying proteins interacting with ZC3H39 might identify TRF or other known telomere proteins. Such analysis might provide evidence supporting, or not, the telomeric association of these proteins. Affinity purification coupled to liquid chromatography tandem mass spectrometry (LC-MS/MS) of YFP-ZC3H39 identified a diverse set of proteins as interactors, particularly components of the trypanosome RNA editing machinery. The top interactors (log2FC >6) were ZC3H40, heterogeneous nuclear ribonucleoproteins F/H homologue (HNRNPH/F), mitochondrial RNA binding complex 1 subunit (MRB1590), universal minicircle sequence binding protein 2 (UMSBP2), the RNA-binding protein DRBD2, multiple RNA Editing Substrate Binding Complex (RESC) components, and additional RNA editing proteins. The RESC is a large multiprotein complex consisting of three distinct modules: 1) the guide-RNA binding complex (GRBC); 2) the RNA editing mediator complex (REMC); and 3) the polyadenylation mediator complex (PAMC) (Aphasizheva & Aphasizhev, 2016). The RESC functions in concert with a separate, catalytic complex called the RNA Editing Core Complex (RECC) to form the RNA-editing holoenzyme (Aphasizheva & Aphasizhev, 2016). Proteins of interest are discussed further below. A full list of significantly enriched proteins is provided in Table 5.1.

None of the core telomeric proteins known in the literature were identified, but several proteins with some weak links to telomeres were found among the top interactors. These included UMSBP2, DRBD2, and RGG2; the evidence for their telomere association is discussed below.
Figure 5.3. YFP-ZC3H39 associates primarily with the RNA Editing machinery. Significance thresholds (dotted lines) used to determine significantly enriched proteins are $p < 0.01$ (Y-axis) and $\log_2 > 2$ (X-axis). The data shown are for one experiment based on 3 biological replicates. Proteins of particular interest are labelled where possible. A full list of the proteins enriched in this plot can be found in Table 5.1. The bait protein is in red. Components of the RNA Editing Substrate Binding Complex (RESC) are in blue. RNA editing proteins not part of the core RESC are in green. Proteins above the significance thresholds outside of those chosen categories are in black. Proteins below the significance thresholds are grey.
Table 5.1. Proteins enriched in YFP-ZC3H39 pulldown. All proteins above the significance thresholds defined in Figure 5.3 are listed below along with their UniProt ID, EuPathDB gene code (Tb927 genome unless unavailable), and name as listed in the TriTrypDB. The bait protein is highlighted in red. Proteins are colour coded as in Figure 5.3. Proteins have been ranked from highest to lowest log₂ fold change.

<table>
<thead>
<tr>
<th>UniProt ID</th>
<th>EuPathDB Gene ID</th>
<th>Name</th>
<th>Log2FC</th>
<th>-log student's T-test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q388B1</td>
<td>Tb927.10.14950</td>
<td>ZC3H40</td>
<td>12.5139</td>
<td>8.07938874</td>
</tr>
<tr>
<td>Q388B3</td>
<td>Tb927.10.14930</td>
<td>ZC3H39</td>
<td>11.7155</td>
<td>5.67838184</td>
</tr>
<tr>
<td>Q586X0</td>
<td>Tb927.2.3880</td>
<td>heterogeneous nuclear ribonucleoproteins F/H homologue (HNRNPH/F)</td>
<td>8.07445</td>
<td>5.67318669</td>
</tr>
<tr>
<td>Q57F2</td>
<td>Tb927.3.1590</td>
<td>mitochondrial RNA binding complex 1 subunit (MRB1590)</td>
<td>7.96064</td>
<td>3.964316071</td>
</tr>
<tr>
<td>Q388B0</td>
<td>Tb927.10.6060</td>
<td>universal minicircle sequence binding protein 2 (UMS6P2)</td>
<td>7.9345</td>
<td>6.05084597</td>
</tr>
<tr>
<td>Q38D54</td>
<td>Tb927.9.13990</td>
<td>RNA-binding protein, putative (DRBD2)</td>
<td>7.5901</td>
<td>4.24196768</td>
</tr>
<tr>
<td>Q57Y20</td>
<td>Tb927.7.6410</td>
<td>RNA-editing mediator complex 5 (REMC5)</td>
<td>7.16385</td>
<td>6.31152606</td>
</tr>
<tr>
<td>Q57WL3</td>
<td>Tb927.8.6170</td>
<td>RNA-editing mediator complex protein 5 (REMC5A)</td>
<td>6.92022</td>
<td>5.316376022</td>
</tr>
<tr>
<td>Q389P7</td>
<td>Tb927.10.10830</td>
<td>RGG-containing protein 2 (RGG2)</td>
<td>6.10163</td>
<td>4.25552548</td>
</tr>
<tr>
<td>Q381M2</td>
<td>Tb927.11.15640</td>
<td>mitochondrial RNA binding complex 1 subunit (MERS1)</td>
<td>6.07895</td>
<td>4.69354681</td>
</tr>
<tr>
<td>Q57X7</td>
<td>Tb927.7.2570</td>
<td>guide RNA associated protein, GAP2 (GRBC1)</td>
<td>5.82279</td>
<td>5.185609562</td>
</tr>
<tr>
<td>Q38980</td>
<td>Tb927.10.12330</td>
<td>ZC3H34</td>
<td>5.60205</td>
<td>4.884078946</td>
</tr>
<tr>
<td>Q388E8</td>
<td>Tb927.10.14550</td>
<td>ATP-dependent RNA helicase HEL67</td>
<td>5.49634</td>
<td>3.08512819</td>
</tr>
<tr>
<td>Q586X1</td>
<td>Tb927.2.3800</td>
<td>guide RNA associated protein 1 (GRBC2)</td>
<td>5.47906</td>
<td>4.20347507</td>
</tr>
<tr>
<td>Q586P0</td>
<td>Tb927.2.6070</td>
<td>mitochondrial RNA binding complex 1 subunit (MRB6070)</td>
<td>5.44504</td>
<td>3.666052522</td>
</tr>
<tr>
<td>Q38D2</td>
<td>Tb927.9.10770</td>
<td>polyadenylate-binding protein 2 (PABP2)</td>
<td>5.24869</td>
<td>3.21503342</td>
</tr>
<tr>
<td>P50098</td>
<td>Tb927.10.15310</td>
<td>Inosine-5’-monophosphate dehydrogenase</td>
<td>5.09357</td>
<td>2.254648753</td>
</tr>
<tr>
<td>Q387Y0</td>
<td>Tb927.11.14400</td>
<td>hypothetical protein, conserved</td>
<td>5.05706</td>
<td>4.74191855</td>
</tr>
<tr>
<td>Q385C5</td>
<td>Tb927.11.6440</td>
<td>hypothetical protein, conserved</td>
<td>5.05186</td>
<td>4.676217123</td>
</tr>
<tr>
<td>Q386F4</td>
<td>Tb927.11.2250</td>
<td>conserved protein, unknown function</td>
<td>4.60036</td>
<td>3.84895615</td>
</tr>
<tr>
<td>Q5725</td>
<td>Tb927.8.6170</td>
<td>GTP reductase</td>
<td>4.3563</td>
<td>4.48698263</td>
</tr>
<tr>
<td>Q57F6</td>
<td>Tb927.8.750</td>
<td>nuclear RNA-binding protein, putative (Nopp44/46-2)</td>
<td>4.2413</td>
<td>4.036125184</td>
</tr>
<tr>
<td>Q38A3</td>
<td>Tb927.10.7910</td>
<td>RNA editing complex protein, putative (MERS3)</td>
<td>4.1961</td>
<td>3.38586031</td>
</tr>
<tr>
<td>Q381D2</td>
<td>Tb927.11.16550</td>
<td>ZC3H34</td>
<td>4.16039</td>
<td>3.663579618</td>
</tr>
<tr>
<td>Q38C7</td>
<td>Tb927.11.14100</td>
<td>Double RNA-binding domain containing protein 4 (DRBD4)</td>
<td>4.10416</td>
<td>3.668047647</td>
</tr>
<tr>
<td>Q580H2</td>
<td>Tb927.8.1510</td>
<td>ATP-dependent RNA helicase DBP2B, putative</td>
<td>3.89499</td>
<td>4.040174712</td>
</tr>
<tr>
<td>Q584Q1</td>
<td>Tb927.6.2580</td>
<td>RNA-binding protein, putative (TRRM2)</td>
<td>3.59125</td>
<td>3.12959008</td>
</tr>
<tr>
<td>Q587B0</td>
<td>Tb927.6.4440</td>
<td>RNA-binding protein 42 (RNA-binding motif protein 42) (RBP42)</td>
<td>3.4725</td>
<td>2.672450731</td>
</tr>
<tr>
<td>Q57G6</td>
<td>Tb927.3.3300</td>
<td>FA219</td>
<td>3.41984</td>
<td>3.57197588</td>
</tr>
<tr>
<td>Q583X9</td>
<td>Tb927.4.3020</td>
<td>ATP-dependent DEAH-box RNA helicase, putative</td>
<td>3.3922</td>
<td>2.603102365</td>
</tr>
<tr>
<td>Q57WL2</td>
<td>Tb927.8.8180</td>
<td>RNA editing mediator complex protein 4 (REMC4)</td>
<td>3.38897</td>
<td>2.70905419</td>
</tr>
<tr>
<td>Q587E3</td>
<td>Tb927.6.4770</td>
<td>protein mkt1, putative</td>
<td>3.37324</td>
<td>2.68581909</td>
</tr>
<tr>
<td>Q583X8</td>
<td>Tb927.11.9510</td>
<td>nucleic acid binding protein, putative</td>
<td>3.09863</td>
<td>4.62875078</td>
</tr>
<tr>
<td>Q38C87</td>
<td>Tb927.10.1640</td>
<td>ZC3H30</td>
<td>2.89072</td>
<td>4.69093556</td>
</tr>
<tr>
<td>Q38A3Z</td>
<td>Tb927.10.6130</td>
<td>hypothetical protein, conserved</td>
<td>2.36634</td>
<td>2.582950842</td>
</tr>
<tr>
<td>Q58717</td>
<td>Tb927.2.3100</td>
<td>kinetoplast polyadenylation/uridylation factor 1 (KPAF1)</td>
<td>2.29779</td>
<td>3.82931997</td>
</tr>
<tr>
<td>Q584T3</td>
<td>Tb927.6.2230</td>
<td>RGG protein (RGG1)</td>
<td>2.20449</td>
<td>2.718438962</td>
</tr>
</tbody>
</table>
Summary of reported information on the identified proteins with respect to telomere association (where applicable):

- **HNRNPH/F** proteins are involved in alternative splicing in humans, and bind TERRA RNAs in mice (de Silanes et al., 2010; Gautrey et al., 2015). The HNRNPH/F homologue in *T. brucei* is found in both the nucleus and the cytoplasm in BSFs, and appears to play a role in splicing and the stability of RNAs containing a purine-rich motif in their UTRs (Gupta et al., 2013). The fact that HNRNPH/F binds a purine-rich motif means it is possible that, similar to its function in mice, it could also bind TERRA RNAs in trypanosomes.

- **UMSBP2** has been shown to co-localize with *T. brucei* telomeres, and bound the single stranded telomeric G-strand *in vitro* (Klebanov-Akopyan et al., 2018). Further, knockdown of UMSBP2 led to telomere clustering, the generation of DNA damage at telomeres, and a large increase in C-rich 5’ telomeric overhangs (Klebanov-Akopyan et al., 2018). Knockdown also led to the accumulation of cells with enlarged nuclei and 1K0N cells (zoids), suggesting some influence on mitotic regulation (Milman et al., 2007).

- **DRBD2** was shown to associate with TelAP1 when it was overexpressed in PCFs (Reis et al., 2018). However, the protein was seen to be cytoplasmic by the TrypTag initiative when tagged at either terminus in PCFs (Dean et al., 2017). Interestingly, DRBD2 interacts with RHS2; RHS2 is essential, co-localizes with RNAPII via ChIP-seq, is required for proper transcription, and may be involved in RNA export from the nucleus (Florini et al., 2019). This could provide a link between ZC3H39 and transcription or mRNA export, but RHS2 was not identified as being associated with ZC3H39.

- **RNA editing substrate binding complex (RESC) components (blue).**
  - **RGG2:** RGG2 is a member of the REMC subcomplex of the RESC (Aphasizheva et al., 2014; Fisk et al., 2008). RGG2 is an essential protein important for the processivity of RNA editing, which binds poly(G) and poly(U) RNA (Ammerman et al., 2010; Fisk et al., 2008; Foda et al., 2012; Travis et al., 2019). Localisation of the C-terminally tagged protein by TrypTag indicated a kinetoplast localization, consistent with its role in mitochondrial RNA editing (Dean et al., 2017). RGG2 was shown to associate with the telomere protein TelAP1 when it was overexpressed in procyclic forms (Reis et al., 2018). There are no other reports of RGG2...
associating with telomere proteins in the literature. The preference of RGG2 for poly(G) RNA raises the possibility that this protein could bind TERRA RNA, but any role at telomeres has to be reconciled with its lack of nuclear localisation. However, the protein was only tagged at the C-terminus by TrypTag, so it is possible that the N-terminally tagged protein, or the endogenous untagged protein, would have a different localisation.

- REMC5, REMC5A, and REMC4: Similar to RGG2, these proteins are found in the RNA editing mediator complex (REMC) module of the RESC (Aphasizheva et al., 2014). Yeast 2-hybrid and pulldown experiments predict interaction between REMC4 and GRBC1/2, GRBC5, and GRBC6, suggesting a key role in bridging the REMC and GRBC modules of the RESC (Aphasizheva et al., 2014).

- GRBC1/GRBC2: Both proteins are components of the GRBC module of the RESC, and have been shown to bind one another and gRNA in vitro (Aphasizheva et al., 2014; Aphasizheva & Aphasizhev, 2010).

- Additional RNA editing proteins not part of the core RESC (green).
- MERS1: mitochondrial edited mRNA stability factor 1 (MERS1) is a hydrolase, and, along with MERS2 and MERS3 (also identified in this experiment), forms the pyrophosphohydrolase complex, or PPsome, which removes pyrophosphates from the 5’ end of mRNAs (Sement et al., 2018). MERS1 is known to associate with guide RNA binding complex protein 1 (GRBC1) and GRBC2 from the GRBC subcomplex of the RESC, and is important for stability of edited mRNAs (Sement et al., 2018; Weng et al., 2008).

- KPAF1: Kinetoplast polyadenylation/uridylation factor (KPAF) 1 is required for A/U tailing of edited mRNAs, and interacts with the PAMC module of the RESC (Aphasizheva et al., 2011, 2014).

- MRB1590/6070: MRB1590 and MRB6070 have been identified in some purifications of RESC/MRB1 complexes and not others, and so are considered to interact with this complex dynamically (Acestor et al., 2009; Ammerman et al., 2011, 2012; Kafková et al., 2012; Weng et al., 2008). Many of these interactions are abrogated upon RNase treatment, which favours this hypothesis. Knockdown of MRB1590 impacts the editing of several transcripts, including the mRNA of ATP
synthase subunit 6 (Shaw et al., 2015). MRB1590 and MRB6070 maintain their interaction upon RNase treatment, suggesting they may form a complex (Shaw et al., 2015).

- **RGG1**: RGG1 was shown to associate strongly with GRBC1 and 2, as well as other RESC components and RNA editing accessory proteins (Hashimi et al., 2008). Consistent with this, knockdown of RGG1 leads to a decrease in edited mRNAs, suggesting a role for RGG1 in the editing process or stabilization of edited transcripts (Hashimi et al., 2008).

### 5.5 ZC3H40 and ZC3H39 associate with the same core proteins

In order to validate the interaction between ZC3H39 and ZC3H40, affinity purification coupled to liquid chromatography tandem mass spectrometry (LC-MS/MS) of YFP-ZC3H40 was performed. This identified ZC3H39 as expected, as well as the same core set of interacting proteins as ZC3H39: HNRNPH/F, UMSBP2, DRBD2, the same RNA Editing Substrate Binding Complex (RESC) components, and almost all of the same additional RNA editing proteins. In terms of differences between ZC3H39 and ZC3H40, RGG1 and MRB6070 were absent here, while additional RESC subunits, RNA editing associated helicase 2 (REH2), and MRB1820 not seen in the ZC3H39 immunoprecipitation were present. A selection of the proteins not identified in the ZC3H39 will be discussed further below. Due to the large number of proteins identified, only a small selection was chosen for discussion. A full list of significantly enriched proteins is provided in Table 5.2.
Figure 5.4. YFP-ZC3H40 associates with the same core proteins as ZC3H39. Significance thresholds (dotted lines) used to determine significantly enriched proteins are $p < 0.01$ (Y-axis) and $\log_2 > 3$ (X-axis). The data shown are for one experiment based on 3 biological replicates. Proteins of particular interest are labelled where possible. A full list of the proteins enriched in this plot can be found in Table 5.2. The bait protein is in red. Components of the RNA Editing Substrate Binding Complex (RESC) are in blue. RNA editing proteins not part of the core RESC are in green. Proteins above the significance thresholds outside of those chosen categories are in black. Proteins below the significance thresholds are grey.
Table 5.2. Proteins enriched in YFP-ZC3H40 pulldown. All proteins above the significance thresholds defined in Figure 5.4 are listed below along with their UniProt ID, EuPathDB gene code (Tb927 genome unless unavailable), and name as listed in the TriTrypDB. The bait protein is highlighted in red. Proteins are colour coded as in Figure 5.4. Proteins have been ranked from highest to lowest log₂ fold change.

<table>
<thead>
<tr>
<th>UniProt ID</th>
<th>EuPathDB Gene ID</th>
<th>Name</th>
<th>Log2FC</th>
<th>-log student's T-test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q388B1</td>
<td>Tb927.10.14950</td>
<td>ZC3H40</td>
<td>12.8309</td>
<td>9.1564822</td>
</tr>
<tr>
<td>Q388B3</td>
<td>Tb927.10.14930</td>
<td>ZC3H39</td>
<td>12.0565</td>
<td>5.767625379</td>
</tr>
<tr>
<td>Q586A0</td>
<td>Tb927.2.3880</td>
<td>heterogeneous nuclear ribonucleoproteins F/H homologue (HNRNPH/F)</td>
<td>8.97936</td>
<td>5.955746663</td>
</tr>
<tr>
<td>Q38D54</td>
<td>Tb927.9.13990</td>
<td>RNA-binding protein, putative (DRBD2)</td>
<td>8.67557</td>
<td>4.39879496</td>
</tr>
<tr>
<td>Q38B00</td>
<td>Tb927.10.6060</td>
<td>universal microcircle sequence binding protein 2 (UMS8P2)</td>
<td>8.46065</td>
<td>6.524843702</td>
</tr>
<tr>
<td>O57V20</td>
<td>Tb927.4.4160</td>
<td>RNA editing mediator complex protein 5 (REMC5)</td>
<td>7.99625</td>
<td>6.698239882</td>
</tr>
<tr>
<td>O57W3</td>
<td>Tb927.8.8170</td>
<td>RNA editing mediator complex protein 5 (REMC5A)</td>
<td>7.87226</td>
<td>5.50506532</td>
</tr>
<tr>
<td>O57K7</td>
<td>Tb927.7.2570</td>
<td>guide RNA associated protein, GAP2 (GRBC1)</td>
<td>7.46699</td>
<td>6.06314932</td>
</tr>
<tr>
<td>O381M2</td>
<td>Tb927.11.15640</td>
<td>mitochondrial RNA binding complex 1 subunit (MERS1)</td>
<td>7.2697</td>
<td>5.10156815</td>
</tr>
<tr>
<td>O57F2</td>
<td>Tb927.3.1550</td>
<td>mitochondrial RNA binding complex 1 subunit (MRB1590)</td>
<td>7.26624</td>
<td>3.80833531</td>
</tr>
<tr>
<td>O389P7</td>
<td>Tb927.10.10830</td>
<td>RGG-containing protein 2 (RGG2)</td>
<td>7.42052</td>
<td>4.62090657</td>
</tr>
<tr>
<td>O58K61</td>
<td>Tb927.2.3800</td>
<td>guide RNA associated protein 1 (GRBC2)</td>
<td>7.11672</td>
<td>5.027978567</td>
</tr>
<tr>
<td>Q388E8</td>
<td>Tb927.10.14550</td>
<td>ATP-dependent RNA helicase HEL67</td>
<td>6.92449</td>
<td>3.453257511</td>
</tr>
<tr>
<td>O38D2</td>
<td>Tb927.9.10770</td>
<td>polyadenylate-binding protein 2 (PABP2)</td>
<td>6.86733</td>
<td>3.70001837</td>
</tr>
<tr>
<td>Q38B00</td>
<td>Tb927.10.6060</td>
<td>universal microcircle sequence binding protein 2 (UMS8P2)</td>
<td>8.46065</td>
<td>6.524843702</td>
</tr>
<tr>
<td>Q586A0</td>
<td>Tb927.2.3880</td>
<td>heterogeneous nuclear ribonucleoproteins F/H homologue (HNRNPH/F)</td>
<td>8.97936</td>
<td>5.955746663</td>
</tr>
<tr>
<td>Q38D54</td>
<td>Tb927.9.13990</td>
<td>RNA-binding protein, putative (DRBD2)</td>
<td>8.67557</td>
<td>4.39879496</td>
</tr>
<tr>
<td>Q38B00</td>
<td>Tb927.10.6060</td>
<td>universal microcircle sequence binding protein 2 (UMS8P2)</td>
<td>8.46065</td>
<td>6.524843702</td>
</tr>
<tr>
<td>Q586A0</td>
<td>Tb927.2.3880</td>
<td>heterogeneous nuclear ribonucleoproteins F/H homologue (HNRNPH/F)</td>
<td>8.97936</td>
<td>5.955746663</td>
</tr>
<tr>
<td>Q38D54</td>
<td>Tb927.9.13990</td>
<td>RNA-binding protein, putative (DRBD2)</td>
<td>8.67557</td>
<td>4.39879496</td>
</tr>
<tr>
<td>Q38B00</td>
<td>Tb927.10.6060</td>
<td>universal microcircle sequence binding protein 2 (UMS8P2)</td>
<td>8.46065</td>
<td>6.524843702</td>
</tr>
<tr>
<td>Q586A0</td>
<td>Tb927.2.3880</td>
<td>heterogeneous nuclear ribonucleoproteins F/H homologue (HNRNPH/F)</td>
<td>8.97936</td>
<td>5.955746663</td>
</tr>
<tr>
<td>Q38D54</td>
<td>Tb927.9.13990</td>
<td>RNA-binding protein, putative (DRBD2)</td>
<td>8.67557</td>
<td>4.39879496</td>
</tr>
<tr>
<td>Q38B00</td>
<td>Tb927.10.6060</td>
<td>universal microcircle sequence binding protein 2 (UMS8P2)</td>
<td>8.46065</td>
<td>6.524843702</td>
</tr>
<tr>
<td>Q586A0</td>
<td>Tb927.2.3880</td>
<td>heterogeneous nuclear ribonucleoproteins F/H homologue (HNRNPH/F)</td>
<td>8.97936</td>
<td>5.955746663</td>
</tr>
<tr>
<td>Q38D54</td>
<td>Tb927.9.13990</td>
<td>RNA-binding protein, putative (DRBD2)</td>
<td>8.67557</td>
<td>4.39879496</td>
</tr>
<tr>
<td>Q38B00</td>
<td>Tb927.10.6060</td>
<td>universal microcircle sequence binding protein 2 (UMS8P2)</td>
<td>8.46065</td>
<td>6.524843702</td>
</tr>
<tr>
<td>Q586A0</td>
<td>Tb927.2.3880</td>
<td>heterogeneous nuclear ribonucleoproteins F/H homologue (HNRNPH/F)</td>
<td>8.97936</td>
<td>5.955746663</td>
</tr>
<tr>
<td>Q38D54</td>
<td>Tb927.9.13990</td>
<td>RNA-binding protein, putative (DRBD2)</td>
<td>8.67557</td>
<td>4.39879496</td>
</tr>
<tr>
<td>Q38B00</td>
<td>Tb927.10.6060</td>
<td>universal microcircle sequence binding protein 2 (UMS8P2)</td>
<td>8.46065</td>
<td>6.524843702</td>
</tr>
<tr>
<td>Q586A0</td>
<td>Tb927.2.3880</td>
<td>heterogeneous nuclear ribonucleoproteins F/H homologue (HNRNPH/F)</td>
<td>8.97936</td>
<td>5.955746663</td>
</tr>
<tr>
<td>Q38D54</td>
<td>Tb927.9.13990</td>
<td>RNA-binding protein, putative (DRBD2)</td>
<td>8.67557</td>
<td>4.39879496</td>
</tr>
<tr>
<td>Q38B00</td>
<td>Tb927.10.6060</td>
<td>universal microcircle sequence binding protein 2 (UMS8P2)</td>
<td>8.46065</td>
<td>6.524843702</td>
</tr>
<tr>
<td>Q586A0</td>
<td>Tb927.2.3880</td>
<td>heterogeneous nuclear ribonucleoproteins F/H homologue (HNRNPH/F)</td>
<td>8.97936</td>
<td>5.955746663</td>
</tr>
<tr>
<td>Q38D54</td>
<td>Tb927.9.13990</td>
<td>RNA-binding protein, putative (DRBD2)</td>
<td>8.67557</td>
<td>4.39879496</td>
</tr>
</tbody>
</table>
Summary of reported information on selected proteins identified as associating with ZC3H40 and not ZC3H39:

- **Additional RESC subunits**: PAMC1 and 4, REMC1, 2, and 3, and GRBC4, 5, and 6 are all known subunits of the RESC (Aphasizheva & Aphasizhev, 2016).

- **REH2**: the REH2 helicase is an essential protein, which has been shown to bind gRNAs, multiple subunits of the RESC, and even the RECC at low levels (Hashimi et al., 2009; Hernandez et al., 2010). Therefore, its identification here along with many other RNA editing components is consistent with previous findings.

- **MRB1820**: MRB1820, also known as RGG3, has previously been identified as interacting with GRBC1 and 2, RGG1, and REH2 (Hashimi et al., 2008; Hernandez et al., 2010; McAdams et al., 2015). RGG3/MRB1820 is essential in PCFs, and required for stability of certain highly edited transcripts (McAdams et al., 2015). It is considered to play a regulatory role, as protein-protein interactions seen in the literature have varied between different laboratories, with the protein only interacting with the RESC to variable degrees.
La protein: In other eukaryotes, the La protein binds the 3’ termini of various small RNAs, particularly those transcribed by RNAPIII, protecting them from degradation by exonucleases (Wolin & Cedervall, 2003). In humans, La protein has been shown to interact with telomerase and influence telomere length (Ford et al., 2001). La protein family members have also been shown to be a subunit of telomerase in ciliates and fission yeast (Aigner et al., 2000, 2003; Aigner & Cech, 2004; Möllenbeck et al., 2003; Pérez-Moscoso et al., 2018; Witkin & Collins, 2004). In T. brucei, a role for La protein in telomere biology has not been investigated, but it does appear to play a role in ribosome biogenesis (Shan et al., 2019).
5.6 YFP-tagged UMSBP2 exhibits an exclusively nuclear localisation

In order to confirm the interaction of ZC3H39 and ZC3H40 with UMSBP2, it was decided to perform the reciprocal immunoprecipitation experiment with this protein, i.e. to identify UMSBP2-interacting proteins. Firstly, a cell line expressing a YFP-tagged allele of UMSBP2 at the endogenous locus was generated as in Staneva et al. (Staneva et al., 2021), and immunolocalisation was performed to assay if the N-terminally tagged protein localised as expected within the cell.

Figure 5.5. YFP-UMSBP2 is localized throughout the nucleus. The protein were immunolocalised in bloodstream form trypanosomes using a polyclonal anti-GFP primary antibody and an Alexafluor-488 labelled secondary antibody (green in merged panel). Nuclear and kinetoplast DNA were stained with DAPI (blue in merged panel). Images chosen are representative of the cell population as a whole. DAPI and FITC refer to fluorophores used for excitation.
The nuclear localisation of YFP-UMSBP2 seen here is consistent with that seen elsewhere (Dean et al., 2017; Klebanov-Akopyan et al., 2018). However, the distribution here is less punctate than that seen elsewhere, and the protein seems to be found almost everywhere inside the nucleus apart from the nucleolus. This difference may be correct, or may simply reflect an issue of staining and resolution.

5.7 UMSBP2 associates with the same core proteins as ZC3H39/40 as well as multiple chromatin factors

Since the majority of the proteins found to interact with ZC3H39/40 were not telomeric, affinity purification coupled to liquid chromatography tandem mass spectrometry (LC-MS/MS) of YFP-UMSBP2 was performed. Since UMSBP2 has been characterised as having a role at telomeres (Section 5.4), this would: 1) verify the interaction of UMSBP2 with ZC3H39/40; and 2) theoretically only purify any nuclear fraction of ZC3H39/40, as UMSBP is only present in the nucleus. However, this analysis identified the same set of top interacting proteins as those found to associate with ZC3H39/40, bolstering the idea that these proteins are in some form of complex. Indeed, this experiment identified a more extensive list of RNA editing substrate binding complex (RESC) subunits and additional RNA editing proteins, but also identified several chromatin-associated proteins, and proteins previously identified at telomeres either in the literature or my own data provided in this thesis. None of the core telomeric proteins highlighted in blue in Sections 4.2.1, 4.2.2, and 5.8 were identified (TRF, TIF2, RAP1, TelAP1, Tb927.6.4330, Tb927.9.4000, PPL2, POLQ). As in Section 5.5, only a small selection of the large number of proteins identified will be discussed here. A full list of significantly enriched proteins is provided in Table 5.3.
Figure 5.6. YFP-UMSBP2 associates with the same core proteins as ZC3H39 and ZC3H40. Significance thresholds (dotted lines) used to determine significantly enriched proteins are $p < 0.01$ (Y-axis) and $\log_2 > 5$ (X-axis). The data shown are for one experiment based on 3 biological replicates. Proteins of particular interest are labelled where possible. A full list of the proteins enriched in this plot can be found in Table 5.3. The bait protein is in red. The zinc finger proteins ZC3H39 & ZC3H40 are in beige. Components of the RNA Editing Substrate Binding Complex (RESC) are in blue. RNA editing proteins not part of the core RESC are in green. Chromatin-associated proteins are in purple. Proteins above the significance thresholds outside of those chosen categories are in black. Proteins below the significance thresholds are grey.
### Table 5.3. Proteins enriched in YFP-UMSBP2 pulldown.

All proteins above the significance thresholds defined in Figure 5.6 are listed below along with their UniProt ID, EuPathDB gene code (Tb927 genome unless unavailable), and name as listed in the TriTrypDB. Proteins are colour coded as in Figure 5.6. Proteins have been ranked from highest to lowest log$_2$ fold change.

<table>
<thead>
<tr>
<th>Protein ID</th>
<th>EuPathDB Gene ID</th>
<th>Name</th>
<th>Log2FC</th>
<th>-log student’s T-test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q586X0</td>
<td>Tb927.2.3880</td>
<td>heterogeneous nuclear ribonucleoproteins F/H homologue (HNRNPH/F)</td>
<td>11.8367</td>
<td>6.168000577</td>
</tr>
<tr>
<td>Q57Y20</td>
<td>Tb927.4.4160</td>
<td>RNA editing mediator complex protein 5 (REMC5)</td>
<td>9.72916</td>
<td>5.079535269</td>
</tr>
<tr>
<td>Q389B0</td>
<td>Tb927.10.12330</td>
<td>ZC3H34</td>
<td>9.56614</td>
<td>4.916641114</td>
</tr>
<tr>
<td>Q389P7</td>
<td>Tb927.10.10830</td>
<td>RGG-containing protein 2 (RGG2)</td>
<td>9.39738</td>
<td>4.80032021</td>
</tr>
<tr>
<td>Q57X7</td>
<td>Tb927.7.2570</td>
<td>guide RNA associated protein, GAP2 (GRBC1)</td>
<td>9.2063</td>
<td>4.910404641</td>
</tr>
<tr>
<td>Q57W3</td>
<td>Tb927.9.8170</td>
<td>RNA editing mediator complex protein 5 (REMC5A)</td>
<td>9.24874</td>
<td>4.505765833</td>
</tr>
<tr>
<td>Q57G7</td>
<td>Tb927.5.2080</td>
<td>GMP reductase</td>
<td>8.45544</td>
<td>5.71347741</td>
</tr>
<tr>
<td>Q57W2</td>
<td>Tb927.8.8180</td>
<td>RNA editing mediator complex protein 4 (REMC4)</td>
<td>8.53</td>
<td>4.056675006</td>
</tr>
<tr>
<td>Q57F2</td>
<td>Tb927.3.1550</td>
<td>mitochondrial RNA binding complex 1 subunit (MRB1550)</td>
<td>8.77821</td>
<td>4.67355313</td>
</tr>
<tr>
<td>Q586P0</td>
<td>Tb927.2.6070</td>
<td>mitochondrial RNA binding complex 1 subunit (MRB6070)</td>
<td>8.17137</td>
<td>5.942151554</td>
</tr>
<tr>
<td>Q389H3</td>
<td>Tb927.10.97910</td>
<td>RNA editing complex protein, putative (MERS1)</td>
<td>7.96437</td>
<td>6.158888335</td>
</tr>
<tr>
<td>Q585B3</td>
<td>Tb927.6.1470</td>
<td>hypothetical protein, conserved</td>
<td>7.93442</td>
<td>4.007907067</td>
</tr>
<tr>
<td>Q381U0</td>
<td>Tb927.11.14950</td>
<td>zinc finger protein 2 (ZFP2)</td>
<td>7.82921</td>
<td>5.07990908</td>
</tr>
<tr>
<td>Q57G8</td>
<td>Tb927.10.2240</td>
<td>RNA binding protein</td>
<td>7.80397</td>
<td>5.27099341</td>
</tr>
<tr>
<td>Q57G8</td>
<td>Tb927.9.1500</td>
<td>RNA editing associated helicase 2 (REH2)</td>
<td>7.74454</td>
<td>3.89140346</td>
</tr>
<tr>
<td>Q586X1</td>
<td>Tb927.2.3800</td>
<td>guide RNA associated protein 1 (GRBC2)</td>
<td>7.68426</td>
<td>3.894925377</td>
</tr>
<tr>
<td>Q389H2</td>
<td>Tb927.3.1550</td>
<td>mitochondrial RNA binding complex 1 subunit (MRB1550)</td>
<td>8.77821</td>
<td>4.67355313</td>
</tr>
<tr>
<td>Q581T1</td>
<td>Tb927.4.1500</td>
<td>RNA editing associated helicase 2 (REH2)</td>
<td>7.74454</td>
<td>3.89140346</td>
</tr>
<tr>
<td>Q381U0</td>
<td>Tb927.11.14950</td>
<td>zinc finger protein 2 (ZFP2)</td>
<td>7.82921</td>
<td>5.07990908</td>
</tr>
<tr>
<td>Q57G8</td>
<td>Tb927.10.2240</td>
<td>RNA binding protein</td>
<td>7.80397</td>
<td>5.27099341</td>
</tr>
<tr>
<td>Q57G8</td>
<td>Tb927.9.1500</td>
<td>RNA editing associated helicase 2 (REH2)</td>
<td>7.74454</td>
<td>3.89140346</td>
</tr>
<tr>
<td>Q586X1</td>
<td>Tb927.2.3800</td>
<td>guide RNA associated protein 1 (GRBC2)</td>
<td>7.68426</td>
<td>3.894925377</td>
</tr>
<tr>
<td>Q58800</td>
<td>Tb927.10.6060</td>
<td>universal minicircle sequence binding protein 2 (UMSBP2)</td>
<td>7.62966</td>
<td>3.77380342</td>
</tr>
<tr>
<td>Q581T1</td>
<td>Tb927.4.1500</td>
<td>RNA editing associated helicase 2 (REH2)</td>
<td>7.74454</td>
<td>3.89140346</td>
</tr>
<tr>
<td>Q586X1</td>
<td>Tb927.2.3800</td>
<td>guide RNA associated protein 1 (GRBC2)</td>
<td>7.68426</td>
<td>3.894925377</td>
</tr>
<tr>
<td>Q58800</td>
<td>Tb927.10.6060</td>
<td>universal minicircle sequence binding protein 2 (UMSBP2)</td>
<td>7.62966</td>
<td>3.77380342</td>
</tr>
<tr>
<td>Q581T1</td>
<td>Tb927.4.1500</td>
<td>RNA editing associated helicase 2 (REH2)</td>
<td>7.74454</td>
<td>3.89140346</td>
</tr>
<tr>
<td>Q586X1</td>
<td>Tb927.2.3800</td>
<td>guide RNA associated protein 1 (GRBC2)</td>
<td>7.68426</td>
<td>3.894925377</td>
</tr>
<tr>
<td>Q58800</td>
<td>Tb927.10.6060</td>
<td>universal minicircle sequence binding protein 2 (UMSBP2)</td>
<td>7.62966</td>
<td>3.77380342</td>
</tr>
<tr>
<td>Q581T1</td>
<td>Tb927.4.1500</td>
<td>RNA editing associated helicase 2 (REH2)</td>
<td>7.74454</td>
<td>3.89140346</td>
</tr>
<tr>
<td>Q586X1</td>
<td>Tb927.2.3800</td>
<td>guide RNA associated protein 1 (GRBC2)</td>
<td>7.68426</td>
<td>3.894925377</td>
</tr>
<tr>
<td>Q58800</td>
<td>Tb927.10.6060</td>
<td>universal minicircle sequence binding protein 2 (UMSBP2)</td>
<td>7.62966</td>
<td>3.77380342</td>
</tr>
<tr>
<td>Q581T1</td>
<td>Tb927.4.1500</td>
<td>RNA editing associated helicase 2 (REH2)</td>
<td>7.74454</td>
<td>3.89140346</td>
</tr>
<tr>
<td>Q586X1</td>
<td>Tb927.2.3800</td>
<td>guide RNA associated protein 1 (GRBC2)</td>
<td>7.68426</td>
<td>3.894925377</td>
</tr>
<tr>
<td>Q58800</td>
<td>Tb927.10.6060</td>
<td>universal minicircle sequence binding protein 2 (UMSBP2)</td>
<td>7.62966</td>
<td>3.77380342</td>
</tr>
<tr>
<td>Q581T1</td>
<td>Tb927.4.1500</td>
<td>RNA editing associated helicase 2 (REH2)</td>
<td>7.74454</td>
<td>3.89140346</td>
</tr>
<tr>
<td>Q586X1</td>
<td>Tb927.2.3800</td>
<td>guide RNA associated protein 1 (GRBC2)</td>
<td>7.68426</td>
<td>3.894925377</td>
</tr>
<tr>
<td>Q58800</td>
<td>Tb927.10.6060</td>
<td>universal minicircle sequence binding protein 2 (UMSBP2)</td>
<td>7.62966</td>
<td>3.77380342</td>
</tr>
</tbody>
</table>
Summary of reported information on the proteins identified in this experiment that were not identified as associating with ZC3H40 and/or ZC3H39:

- Replication factor A (RPA): RPA was identified in the Telomere-TALE immunoprecipitation in Section 4.2.2, where the evidence for its potential role at trypanosome telomeres was discussed. This interaction with UMSBP2, which appears to have a role at telomeres, provides a further potential link between RPA and telomeres.

- Retrotransposon hotspot protein 2 (RHS2): RHS2 was discussed in Section 5.4, due to its interaction with DRBD2. The presence of RHS2 suggests a possible link between UMSBP2 and transcription, which is in stark contrast to the presence of mitochondrial RNA editing proteins, which are highly represented in these purifications.

- DNA topoisomerase II (TOP2): The T. brucei genome encodes multiple type II topoisomerase genes, which encode nuclear and mitochondrial proteins, respectively (Kulikowicz & Shapiro, 2006). The protein identified here is the mitochondrial topoisomerase, which is important for kDNA topology and replication (Lindsay et al., 2008; Z. Wang & Englund, 2001). Its presence in the interactome of this nuclear protein is one of many perplexing findings from this analysis.

- Phytanoyl-CoA dioxygenase (PhyH), putative (MRB7260): MRB7260 is considered an accessory factor to the RESC, as it is not identified in all RESC purifications. It is required for multiple processes in the RNA editing pathway in PCFs and BSFs, and for proper RESC assembly (McAdams et al., 2018).
5.8 Investigation of the impact of ZC3H39 and ZC3H40 knockout on telomere protein composition

In order to assay whether ZC3H39 or ZC3H40 influenced the protein composition of telomeres in any way, the telomeric repeat binding protein TRF was N-terminally tagged at the endogenous locus with YFP, and immunoprecipitation experiments were performed in a double knockout cell line where both copies of ZC3H39 and ZC3H40 had been deleted. This double knockout cell line was provided by Dr Anna Trenaman from Dr David Horn’s laboratory at the University of Dundee.

In order to draw any proper conclusions from this experiment, additional controls were required. Therefore, samples were prepared from WT, YFP-TRF, ΔZC3H39ΔZC3H40, and YFP-TRF + ΔZC3H39ΔZC3H40 cells. Direct comparison with an existing WT vs YFP-TRF data set was not possible, as differences between the samples could be due to differences in the performance of the mass spectrometer over time, different batches of antibody, different buffers and other variables, and not due to the loss of ZC3H39 and ZC3H40.

5.8.1 Immunoprecipitation of TRF from wild type cells provides a control for telomere protein composition

Comparison of an untagged control with YFP-TRF enriched all of the same top interactors identified previously (Section 4.2.1), including ZC3H39 and ZC3H40, as expected. Several additional chromatin associated proteins, including the CAF-1 complex, FACT complex, and replication factor proteins were identified. Interestingly, several proteins never seen in the previous TRF pulldown, but seen in the ZC3H39/ZC3H40/UMSBP2 immunoprecipitates were enriched. These included two of the top interactors from those experiments, HNRNPH/F and DRBD2, as well as polyadenylate-binding protein 2 (PABP2). Proteins of interest are discussed further below. A full list of significantly enriched proteins is provided in Table 5.4.
Figure 5.7. YFP-TRF associates with the same core proteins seen in Section 4.2.1. Significance thresholds (dotted lines) used to determine significantly enriched proteins are $p < 0.01$ (Y-axis) and $\log_2 > 2$ (X-axis). The data shown are for one experiment based on 3 biological replicates. Proteins with similar known functions have been grouped by colour. Proteins of particular interest are labelled where possible. A full list of the proteins enriched in this plot can be found in Table 5.4. The bait protein is in red. Proteins previously seen at telomeres in the literature are in blue. Histones and histone variants are in green. The zinc finger proteins ZC3H39 & ZC3H40 are in beige. Proteins involved in DNA repair are in orange. Known chromatin-associated proteins are in purple. Proteins outside of these chosen categories are in black. Proteins below the significance thresholds are grey.
Table 5.4. Proteins enriched in YFP-TRF pulldown in WT cells. All proteins above the significance thresholds defined in Figure 5.7 are listed below along with their UniProt ID, EuPathDB gene code (Tb927 genome unless unavailable), and name as listed in the TriTrypDB. Proteins are colour coded as in Figure 5.7. Proteins highlighted in yellow were previously identified in the ZC3H39, ZC3H40, and UMSBP2 pulldowns in Sections 5.4, 5.5, and 5.7. Proteins have been ranked from highest to lowest log₂ fold change.

<table>
<thead>
<tr>
<th>UniProt ID</th>
<th>EuPathDB Gene Code</th>
<th>Name</th>
<th>Log2FC</th>
<th>-log student’s T-test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q57XY9</td>
<td>Tb927.3.1560</td>
<td>TRF-Interacting Factor 2 (TIF2)</td>
<td>11.3839</td>
<td>4.359255809</td>
</tr>
<tr>
<td>Q38SUJ2</td>
<td>Tb927.11.9870</td>
<td>Telomere-associated protein 1 (TelAP1)</td>
<td>11.2197</td>
<td>4.714768002</td>
</tr>
<tr>
<td>Q38FD7</td>
<td>Tb927.9.4000</td>
<td>hypothetical protein, conserved</td>
<td>11.1556</td>
<td>4.494574254</td>
</tr>
<tr>
<td>Q38SL5</td>
<td>Tb927.11.5550</td>
<td>DNA polymerase theta (POLQ)</td>
<td>10.9333</td>
<td>6.666513679</td>
</tr>
<tr>
<td>Q38WJ1</td>
<td>Tb927.10.12850</td>
<td>ilagg binding factor (TIF)</td>
<td>10.7333</td>
<td>5.057488609</td>
</tr>
<tr>
<td>Q586CD80</td>
<td>Tb927.6.4330</td>
<td>hypothetical protein, conserved</td>
<td>10.8362</td>
<td>5.057488609</td>
</tr>
<tr>
<td>Q387L4</td>
<td>Tb927.11.370</td>
<td>repressor activator protein 1 (RAP1)</td>
<td>9.55161</td>
<td>4.494574254</td>
</tr>
<tr>
<td>Q38BZ3</td>
<td>Tb927.10.2520</td>
<td>PrimPol-like protein 2 (PPL2)</td>
<td>8.76455</td>
<td>5.386096499</td>
</tr>
<tr>
<td>Q385L3</td>
<td>Tb927.11.5550</td>
<td>DNA polymerase theta (POLQ)</td>
<td>10.9333</td>
<td>6.666513679</td>
</tr>
<tr>
<td>Q385E2</td>
<td>Tb927.11.370</td>
<td>repressor activator protein 1 (RAP1)</td>
<td>9.55161</td>
<td>4.494574254</td>
</tr>
<tr>
<td>Q38BZ3</td>
<td>Tb927.10.2520</td>
<td>PrimPol-like protein 2 (PPL2)</td>
<td>8.76455</td>
<td>5.386096499</td>
</tr>
<tr>
<td>Q385L3</td>
<td>Tb927.11.5550</td>
<td>DNA polymerase theta (POLQ)</td>
<td>10.9333</td>
<td>6.666513679</td>
</tr>
<tr>
<td>Q385E2</td>
<td>Tb927.11.370</td>
<td>repressor activator protein 1 (RAP1)</td>
<td>9.55161</td>
<td>4.494574254</td>
</tr>
<tr>
<td>Q38BZ3</td>
<td>Tb927.10.2520</td>
<td>PrimPol-like protein 2 (PPL2)</td>
<td>8.76455</td>
<td>5.386096499</td>
</tr>
</tbody>
</table>

**Note:** The table includes proteins that were highlighted in yellow in previous sections, indicating they were previously identified. Proteins have been ranked based on their log₂ fold change, with the highest fold change listed first. The table also includes the UniProt ID, EuPathDB gene code (with Tb927 genome unless unavailable), name, log₂ fold change, and -log student’s T-test p-value for each protein.
A brief summary of chromatin-associated proteins identified in this analysis but not seen in the previous YFP-TRF immunoprecipitation (Section 4.2.1) is provided below:

- **Replication factor A (RPA):** The identification of replication factor A subunits associating with telomeres in this experiment is consistent with their identification in the Telomere-TALE immunoprecipitation in Section 4.2.2, and bolsters that finding.

- **FACT:** As mentioned in Section 4.3.1, the FACT histone chaperone complex has been shown to play a role in maintaining repression of silent VSG ESs. It has been shown by ChIP-qPCR to bind silent VSG ESs, but any association with telomere repeats was not assayed in that study (Denninger et al., 2010).

- **CAF-1:** As discussed in Section 1.4.3, CAF1 is a histone chaperone complex associated with DNA replication. CAF-1 has not been reported to be found at telomeres, but its association with VEX1 – which has – could provide a mechanism for telomeric localisation.

- **La protein:** This protein was seen to interact with ZC3H40 and UMSBP2 in Sections 5.5 and 5.7, respectively. As discussed in Section 5.5, the La protein and its family members have characterised roles at telomeres in other eukaryotes, but no investigation of this has been reported for *T. brucei*.
5.8.2 Immunoprecipitation of TRF in ΔZC3H39ΔZC3H40 cells shows a potential decrease in chromatin association

Immunoprecipitation of YFP-TRF from double knockout cells lacking both ZC3H39 and ZC3H40 showed no loss of any of the known telomere components seen to associate with TRF in wild-type cells, but did seem to have much lower enrichment of chromatin-associated proteins, including histones. There was also no enrichment of HNRNPH/F, DRBD2, or either of the PABP proteins seen in the control immunoprecipitation.

Figure 5.8. YFP-TRF associates with the same core proteins in ΔZC3H39ΔZC3H40 cells as in wild type cells. Significance thresholds (dotted lines) used to determine significantly enriched proteins are p < 0.01 (Y-axis) and log2 > 2 (X-axis). The data shown are for one experiment based on 3 biological replicates. Proteins with similar known functions have been grouped by colour. Proteins of particular interest are labelled where possible. A full list of the proteins enriched in this plot can be found in Table 5.5. The bait protein is in red. Proteins previously seen at telomeres in the literature are in blue. Proteins involved in DNA repair are in orange. Known chromatin-associated proteins are in purple. Proteins outside of these chosen categories are in black. Proteins below the significance thresholds are grey.
Table 5.5. Proteins enriched in YFP-TRF pulldown in ΔZC3H39ΔZC3H40 cells. All proteins above the significance thresholds defined in Figure 5.8 are listed below along with their UniProt ID, EuPathDB gene code (Tb927 genome unless unavailable), and name as listed in the TriTrypDB. Proteins are colour coded as in Figure 5.8. Proteins have been ranked from highest to lowest log2 fold change.

<table>
<thead>
<tr>
<th>UniProt ID</th>
<th>EuPathDB Gene ID</th>
<th>Name</th>
<th>Log2FC</th>
<th>-log student’s T-test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q57XY9</td>
<td>Tb927.3.1560</td>
<td>TRF-Interacting Factor 2 (TIF2)</td>
<td>11.9234</td>
<td>4.90507562</td>
</tr>
<tr>
<td>Q383U2</td>
<td>Tb927.11.9870</td>
<td>Telomere-associated protein 1 (TelAP1)</td>
<td>11.4642</td>
<td>4.404852883</td>
</tr>
<tr>
<td>Q586Z9</td>
<td>Tb927.6.4330</td>
<td>hypothetical protein, conserved</td>
<td>11.1945</td>
<td>4.917540704</td>
</tr>
<tr>
<td>Q385L3</td>
<td>Tb927.11.5550</td>
<td>DNA polymerase theta (POLQ)</td>
<td>11.0316</td>
<td>4.657467565</td>
</tr>
<tr>
<td>Q38F07</td>
<td>Tb927.9.4000</td>
<td>hypothetical protein, conserved</td>
<td>10.9012</td>
<td>4.461757731</td>
</tr>
<tr>
<td>Q388W1</td>
<td>Tb927.10.12850</td>
<td>ttaggg binding factor (TRF)</td>
<td>10.4773</td>
<td>4.642163633</td>
</tr>
<tr>
<td>Q388Z3</td>
<td>Tb927.10.2520</td>
<td>PrimPol-like protein 2 (PPL2)</td>
<td>9.58538</td>
<td>4.502057797</td>
</tr>
<tr>
<td>Q387L4</td>
<td>Tb927.11.370</td>
<td>repressor activator protein 1 (RAP1)</td>
<td>8.83003</td>
<td>3.605028362</td>
</tr>
<tr>
<td>Q586G9</td>
<td>Tb927.2.2190</td>
<td>histone deacetylase 3 (HDAC3)</td>
<td>8.13906</td>
<td>4.067838558</td>
</tr>
<tr>
<td>Q385E2</td>
<td>Tb927.11.6270</td>
<td>inositol-1,4,5-trisphosphate (IP3) 5-phosphatase, putative (PIP5Pase1)</td>
<td>7.28106</td>
<td>3.686758864</td>
</tr>
<tr>
<td>Q586G9</td>
<td>Tb927.2.4580</td>
<td>UNC119</td>
<td>5.40422</td>
<td>4.083514152</td>
</tr>
<tr>
<td>Q38FR6</td>
<td>Tb927.9.2520</td>
<td>microtubule-associated protein (GB4)</td>
<td>5.23713</td>
<td>2.305112153</td>
</tr>
<tr>
<td>Q384A0</td>
<td>Tb927.11.8090</td>
<td>protein phosphatase 1, putative</td>
<td>4.96828</td>
<td>2.742731789</td>
</tr>
<tr>
<td>Q382Y9</td>
<td>Tb927.3.2460</td>
<td>hypothetical protein, conserved</td>
<td>4.80197</td>
<td>2.497835332</td>
</tr>
<tr>
<td>Q57W25</td>
<td>Tb927.5.800</td>
<td>casein kinase 1, isoform 2 (CK1.2)</td>
<td>4.6376</td>
<td>2.784349375</td>
</tr>
<tr>
<td>Q384B5</td>
<td>Tb927.11.9130</td>
<td>Replication factor A protein 1 (RPA1)</td>
<td>3.4911</td>
<td>2.445239426</td>
</tr>
<tr>
<td>Q38DO9</td>
<td>Tb927.9.11850</td>
<td>structural maintenance of chromosome 1, putative (SMC1)</td>
<td>3.47529</td>
<td>3.867296781</td>
</tr>
<tr>
<td>Q8T8P1</td>
<td>Tb427.02.4390</td>
<td>endo/exonuclease Mre11 (MRE11)</td>
<td>3.25832</td>
<td>2.483270352</td>
</tr>
<tr>
<td>Q38F55</td>
<td>Tb927.9.2420</td>
<td>hypothetical protein, conserved</td>
<td>2.31636</td>
<td>2.002891077</td>
</tr>
<tr>
<td>B3GVQ2</td>
<td>Tb427.BES56.8</td>
<td>Expression site-associated gene 8 (ESAG8) protein</td>
<td>2.33195</td>
<td>2.36047883</td>
</tr>
<tr>
<td>Q18A89</td>
<td>Tb927.10.8850</td>
<td>A distinct subfamily of CDD/CDA-like deaminases, putative</td>
<td>2.01349</td>
<td>2.12383224</td>
</tr>
</tbody>
</table>

The additional cross comparisons of WT vs ΔZC3H39ΔZC3H40, and WT vs YFP-TRF + ΔZC3H39ΔZC3H40 were also performed, and will be discussed briefly here (data not shown). WT vs ΔZC3H39ΔZC3H40 showed one highly enriched protein in each sample. These were Variant surface glycoprotein MITAT 1.1 and Variant surface glycoprotein MITAT 1.2, respectively. This indicates a change in expressed VSG between the WT and ΔZC3H39ΔZC3H40 cells, which is consistent with previous findings (Trenaman et al., 2019). The WT vs YFP-TRF + ΔZC3H39ΔZC3H40 also showed the enrichment of Variant surface glycoprotein MITAT 1.2 with the bait protein. The bait protein samples were depleted of expression site-associated gene 7 (ESAG7), which is consistent with the observed VSG-switching phenotype.
5.9 Telomere length in ΔZC3H39ΔZC3H40 cells does not change substantially relative to wild type

As discussed previously, loss of ZC3H39 and/or ZC3H40 results in an increase in VSG switching (Trenaman et al., 2019). One variable known to affect the frequency of VSG switching is telomere length (G. A. Hovel-Miner et al., 2012a). Therefore, I investigated whether loss of either, or both, of the ZC3H proteins would cause changes in telomere length compared to wild type cells as assayed by Southern blot. Cell lines where both alleles of ZC3H39 and/or ZC3H40 were deleted were generously gifted by Dr Anna Trenaman from Dr David Horn’s laboratory at the University of Dundee. The experimental procedure was performed exactly as in Section 3.9.

Bulk telomere signal (strong band, Figure 5.9, B) does not appear to change significantly in terms of average molecular weight (MW) in the ΔZC3H39, ΔZC3H40, or ΔZC3H39ΔZC3H40 cell lines compared to wild type (WT) cells. However, all bands do not appear to be exactly the same as one another. For example, WT telomeres appear as two distinct high MW bands, followed by a smear migrating down to approximately 4 kb. This appears to be similar in ΔZC3H39 cells, albeit with the bands being somewhat less distinct. In ΔZC3H40 cells, there are not two distinct bands, and some stronger signal is visible at approximately 15 and 10 kb that cannot be seen in either WT or ΔZC3H39 cells. ΔZC3H39ΔZC3H40 cells display a stronger signal than any of the other lanes, which appears to extend over a larger MW range than those seen for the other lanes. However, it is possible that the increased signal intensity seen in the ΔZC3H39ΔZC3H40 lane is due to a higher concentration of DNA being loaded in this lane compared with the other lanes. Despite the fact that DNA concentrations were measured using a Nanodrop spectrophotometer prior to restriction digestion, this remains a possibility. The ethidium bromide stained agarose gel (panel C, Figure 5.9) does not appear to show a higher concentration of DNA in the ΔZC3H39ΔZC3H40 lane compared with other wells, but this is much less sensitive than the chemiluminescence used in panel B.
Figure 5.9. Telomere length analysis of ∆ZC3H39, ∆ZC3H40, and ∆ZC3H39∆ZC3H40 cells. (A) Flow diagram outlining the main experimental steps. The two restriction enzymes used for digestion of genomic DNA (italics) and their recognition sites are shown in red and blue respectively. DIG – Digoxigenin; AP – Ascorbate peroxidase. (B) Southern blot detecting telomeres. 2 µg of genomic DNA was digested with AluI and MboI, which recognise 4bp sequences not found within the TTAGGG repeats. Membranes were hybridised with a (TTAGGG)₆ probe labelled with DIG-dUTP at its 3’ end before chemiluminescent detection with CDP-Star. (C) Ethidium bromide stained TBE 0.5% agarose gel. Samples were run for ~21 hours, 25 V, overnight. DNA ladder is 1:1 combination of Lambda DNA/HindIII ladder and Promega 1 kb ladder. Images are the same as those used in Section 3.9, but the lane containing the Telomere TALE sample has been cropped out. Kb – kilobase pairs.
Overall, the evidence provided above would suggest that no substantial changes in telomere length are seen after loss of ZC3H39 or ZC3H40, or both ZC3H39 and ZC3H40. However, there is sufficient evidence to say that the question warrants further investigation. Strategies to further address whether ZC3H39 and/or ZC3H40 contribute to telomere length regulation are discussed in Section 5.10.7.

5.10 Discussion

5.10.1 ZC3H39 and ZC3H40 immunofluorescence show localisation consistent with literature

The absence of any discernible nuclear signal in the immunofluorescence of ZC3H39 or ZC3H40, while consistent with the literature and TrypTag, is contrary to their identification at telomeres in the Telomere-TALE and TRF immunoprecipitations in Chapter 4. Mass spectrometry is orders of magnitude more sensitive than immunofluorescence, but it is still discouraging not to identify any clear nuclear signal.

An alternative approach to determine the subcellular localisation of ZC3H39 and ZC3H40 would be to perform subcellular fractionation followed by western analysis of the different fractions. This would be a more sensitive assay, as the lysate of millions of cells would be used, as opposed to observing individual cells in immunofluorescence microscopy. However, such approaches can be problematic due to contamination between fractions.

Alternative microscopy-based approaches could also be investigated. For example, detergent extraction protocols could be employed (Dean & Sunter, 2020). Limited attempts were made (Section 5.2), but the procedure was not optimised and would be worth further consideration in the future.
5.10.2 ChIP-qPCR of ZC3H39 and ZC3H40 shows no enrichment of either protein on telomeric TTAGGG repeats

It is unclear how to reconcile the presence of ZC3H39 and ZC3H40 at telomeres in the TRF and Telomere-TALE immunoprecipitation experiments with the complete absence of any enrichment at telomeric repeats by ChIP. However, this does not mean that these proteins are not present at telomeres, as there are many reasons why this experiment might not show enrichment of TTAGGG repeats with ZC3H39 and ZC3H40. Some of these possibilities are discussed below.

If ZC3H39/40 are indeed present at telomeres, they appear to be playing a dual role in addition to their regulation of respiratory protein RNAs. Based on the immunofluorescence data (Section 5.2), the proteins are predominantly cytosolic, and any nuclear fraction, if it exists, clearly represents a very small fraction. This could mean that in a ChIP experiment, the antibody is binding primarily the cytosolic ZC3H39/40, and there is no binding capacity remaining to bind the much smaller nuclear fraction. This could be investigated by either increasing antibody concentration, or isolating nuclei from cells before crosslinking, and performing the ChIP experiment on isolated crosslinked nuclei instead, thereby removing any cytoplasmic proteins.

Secondly, it is possible that ZC3H39 and ZC3H40 bind somewhere upstream of the TTAGGG repeats. Using primers designed to bind various sites upstream of the TTAGGG repeats, such as VSG genes, ESAGs, or 50 bp repeats could assess this. Performing ChIP-seq could provide more detailed information on the genomic location(s) of these proteins, if they do indeed bind chromatin.

5.10.3 ZC3H39 associates with multiple RNA editing proteins

The identification of RNA editing proteins associating with ZC3H39 is a novel finding. ZC3H39 and ZC3H40 were previously shown to bind transcripts encoding proteins belonging to the electron transport chain (ETC), all of which were transcribed from the nuclear genome.

Although no core telomere-associated proteins were identified as interacting with ZC3H39, several proteins with some evidence connecting them to telomere biology were identified. DRBD2 and RGG2 were only identified as interacting with the telomere protein TelAP1 when it was overexpressed (Reis et al., 2018). However, the evidence for association of UMSBP2 with telomeres is much stronger (Klebanov-Akopyan et al., 2018). The presence
of UMSBP2 creates a possible link between ZC3H39 and ZC3H40 and the single-stranded telomere ends, and requires further investigation.

The majority of the proteins identified as associating with ZC3H39 are cytosolic, but among the top interactors, UMSBP2 and HNRNPH/F are found in the nucleus. The association of these nuclear proteins with ZC3H39 is consistent with the hypothesis that ZC3H39 may have a nuclear role, possibly at telomeres.

5.10.4 ZC3H40 appears to form a complex or complexes with ZC3H39 and RNA editing factors

The number of proteins identified as interacting with ZC3H40 was greater than that identified as interacting with ZC3H39. Note that the X-axis significance threshold used for the ZC3H40 immunoprecipitation data (Figure 5.4) was 3, while that used for the ZC3H39 immunoprecipitation data (Figure 5.3) was 2. Without this change, the number of proteins identified for ZC3H40 was so large that it made the analyses more difficult to present the data in a useful manner.

More importantly, the top interacting proteins identified (log2FC > 7) were the same as those identified for ZC3H39, suggesting the existence of a complex, or multiple complexes, comprising ZC3H39, ZC3H40, UMSBP2, HNRNPH/F, DRBD2, and RNA-editing factors, or some combination of these. Exactly what the functions of these putative complexes might be is unclear. They comprise a collection of proteins with differing annotated subcellular localisations (nuclear, cytosolic, mitochondrial) that have never been reported to associate with one another.

5.10.5. Immunoprecipitation of UMSBP2 shows distinct and overlapping interactome compared with ZC3H39 and ZC3H40
The number of proteins identified as interacting with UMSBP2 was extremely large. As a result, the X-axis significance threshold used for Figure 5.6 was 5. This reduced the number of significantly enriched proteins down to a manageable number for analysis. It remains to be determined whether this large number of proteins represents a reproducible, accurate UMSBP2 interactome.

UMSBP2 is primarily nuclear (Klebanov-Akopyan et al., 2018). Therefore, this represents the third distinct immunoprecipitation performed using a nuclear bait protein that has identified ZC3H39 and ZC3H40 as interactors (UMSBP2, TRF, Telomere-TALE). In contrast to the Telomere-TALE and TRF, UMSBP2 appears to have a more widespread distribution within the nucleus than just at telomeres. No ChIP-qPCR or ChIP-seq analysis was performed on UMSBP2, but this could provide more clues as to its possible functions. UMSBP2 binds the single-stranded telomeric G-strand \textit{in vitro}, but a more thorough investigation of its binding sequence preferences was not performed (Klebanov-Akopyan et al., 2018).

The detection of RPA subunits and RHS2 as associating with UMSBP2 is suggestive of a wider role for UMSBP2 within the nucleus. It is possible that UMSBP2 acts as a single-stranded nucleic acid binding protein during replication and/or transcription. This would be consistent with the cell cycle arrest seen upon RNAi knockdown (Milman et al., 2007).

On the other hand, why exactly mitochondrial RNA editing and cytosolic proteins comprise the bulk of the UMSBP2 interactome seems contradictory with respect to its nuclear localisation and its reported role in telomere biology (Klebanov-Akopyan et al., 2018). The identification of the mitochondrial DNA topoisomerase II (TOP2mt) is another example of this. TOP2mt shares only 36.61% and 35.78% identities with the nuclear DNA topoisomerases IIα and IIβ respectively (blastp search, BLOSUM62 matrix). Therefore, it seems unlikely that TOP2mt was identified mistakenly by the MaxQuant software due to similarity of tryptic peptides between it and either of the nuclear topoisomerases.

5.10.6. Does loss of ZC3H39 and ZC3H40 alter telomere protein composition?
5.10.6.1. TRF shows weak interaction with some proteins that associate with ZC3H39/49, in addition to known telomeric proteins

The immunoprecipitation of TRF in wild-type cells, which was performed nominally as a control for the same experiment in $\Delta ZC3H39\Delta ZC3H40$ cells, identified several proteins not seen in the TRF immunoprecipitation experiment shown in Section 4.2.1. This included the chromatin proteins RPA, FACT, and CAF, all of which have some plausible connection to telomere biology (discussed in Section 5.8). Strikingly, several proteins identified in the ZC3H39, ZC3H40, and UMSBP2 immunoprecipitations were identified as interacting with TRF in wild-type cells. This included HNRNPH/F and DRBD2 (two of the top interactors from those experiments), as well as the polyadenylate binding proteins PABP1 and PABP2. Enrichment levels of HNRNPH/F, DRBD2, and PABP1/2 was much lower than that of core telomere proteins (Log2FC ~3 vs ~10), but similar to that of several core histones and other chromatin associated proteins. Lower enrichment suggests a weak or indirect interaction. The number of proteins identified in this TRF immunoprecipitation was nearly double that identified in Section 4.2.1. The greater number of proteins above the significance threshold here probably explains the absence of all of these proteins from previous experiments.

5.10.6.2. TRF immunoprecipitation in $\Delta ZC3H39\Delta ZC3H40$ cells shows decreased enrichment of chromatin factors

In contrast to wild-type cells, the TRF immunoprecipitation from $\Delta ZC3H39\Delta ZC3H40$ cells enriched relatively few proteins above the significance threshold. All of the core telomere proteins (TRF, TIF2, RAP1, TelAP1, Tb927.4.9000, Tb927.6.4330, PPL2, POLQ) were identified, but histones and several chromatin-associated proteins identified previously were not detected in this double mutant.

Two proteins that were identified as interacting with TRF in both wild-type and $\Delta ZC3H39\Delta ZC3H40$ immunoprecipitations, but not previously, were UNC119 and Casein kinase 1, isoform 2 (CK1.2). UNC119 has been characterised as a component of the lipidated protein intraflagellar transport (LIFT) pathway in *T. brucei*, where it binds myristoylated proteins (Pandey et al., 2020). However, it seems unlikely that its only role is at the flagellum, as the protein appears to be highly concentrated in the nucleus when tagged at either terminus (TrypTag) (Dean et al., 2017). As myristoylation is often important for membrane targeting, it
is possible that some telomere proteins could be myristoylated to facilitate anchoring to the nuclear envelope. Note that this protein was also detected in the ZC3H40 and UMSBP2 immunoprecipitations from wild-type cells (Sections 5.5 and 5.7, respectively).

CK1.2 is essential in bloodstream forms and knockdown results in enlarged, multinucleated cells before death, suggesting a role in cell cycle regulation (Urbaniak, 2009). This would be consistent with its known roles in other eukaryotes (Knippschild et al., 2005). The identification of CK1.2 as associating with TRF suggests it may phosphorylate telomere proteins.

5.10.7. Telomere length appears unchanged in \(\Delta\)ZC3H39\(\Delta\)ZC3H40 cells

The experiment performed in Section 5.9 did not identify any clear differences in telomere length in \(\Delta\)ZC3H39, \(\Delta\)ZC3H40, or \(\Delta\)ZC3H39\(\Delta\)ZC3H40 cells relative to wild-type cells. However, small potential differences were identified between wild-type and \(\Delta\)ZC3H40 and \(\Delta\)ZC3H39\(\Delta\)ZC3H40 cells, but there are several caveats that prevent strong conclusions being drawn.

Firstly, the strong signal seen in the \(\Delta\)ZC3H39\(\Delta\)ZC3H40 lane may be due to a higher concentration of DNA, despite the fact that input DNA concentrations – measured using a Nanodrop spectrophotometer – were matched between lanes. The Qubit fluorometer could be used to calculate DNA concentrations prior to restriction digestion, as this provides much more accurate data than the Nanodrop instrument. The question of DNA input comparability between lanes could also be assessed by using a control oligo probe for another genomic locus as commonly performed in other protein and nucleic acid blotting approaches. However, this locus would need to remain intact after digestion of DNA with 4-base cutter restriction enzymes.

The potential differences in telomere length identified in Figure 5.9 were very small, and could therefore benefit from increased electrophoretic separation of DNA fragments. Alternatively, qPCR-based methods of telomere length measurement exist, and could provide quantitative data (Cawthon, 2002, 2009). This may be superior to Southern blotting for identifying subtle changes in telomere length.

Further, there could be differences in telomeres between wild-type cells and the knockout strains due to the unique history of the cell lines. Although all parasites used belong to the 427 monomorphic strain, the wild-type strain used originates from a Matthews lab stock,
while the knockout cell lines originate from a Horn lab stock. These cell lines have been grown independently for a number of years, and telomeres may differ between them. Therefore, repeating the experiment using the Horn lab wild-type parent cell line may provide a more suitable control.
Heterochromatin, which is enriched in repetitive sequences, is an integral feature of eukaryotic chromosomes that is necessary for critical genome functions such as centromere function, controlling cell-type specific transcription, and maintaining genome stability by silencing transposable elements (Allshire & Madhani, 2017; Janssen et al., 2018). Heterochromatin formation and function is mediated by specific histone post-translational modifications such as histone H3K9 methylation, and proteins such as HP1 that bind these modifications. These factors are crucial for the creation of heterochromatin domains and the recruitment and regulation of various proteins to heterochromatin (Canzio et al., 2013; Eskeland et al., 2007; Swenson et al., 2016).

Trypanosome histones are divergent, and histone modifications such as H3K9 methylation, which characterise heterochromatin in other eukaryotes, are absent. Similarly, the specialised proteins that evolved to ‘write’, ‘read’, and ‘erase’ these modifications are absent from the genomes of trypanosomes. However, electron micrographs of *T. brucei* nuclei clearly show the presence of electron dense heterochromatin localised primarily to the nuclear periphery. Therefore, mechanisms of heterochromatin formation and function appear to exist, but the proteins mediating these processes are yet to be characterised. Further, the *T. brucei* genome contains abundant repetitive sequences such as telomere repeats, centromere repeats, transposable elements, and transcriptionally silent minichromosomes; such repetitive sequences would be packaged in heterochromatin in other eukaryotes. Current models of heterochromatin formation and function are based on observations of common model organisms belonging to the Opisthokonta supergroup of Eukaryota (Allshire & Madhani, 2017). Identifying the mechanisms of heterochromatin formation and function in organisms that belong to other eukaryotic supergroups, such as trypanosomes, would provide greater understanding of the mechanisms of heterochromatin formation and function in eukaryotes as a whole. The identification of novel proteins involved in these processes could also lead to the development of drugs for the treatment of human and animal disease.

This study aimed to develop a system to identify proteins associated with repetitive sequences predicted to be packaged in heterochromatin in the *T. brucei* genome. I sought to address this question by using the specificity of TALE proteins to purify different repetitive sequences and their associated proteins, with the ultimate goal of identifying proteins required for heterochromatin formation and function in *T. brucei*.

In the first stage of the project, I generated four different monomorphic bloodstream from cell lines expressing synthetic TALEs designed to bind four different repetitive sequences from the *T. brucei* genome. The TALEs targeted 15 bp sequences found within the telomere
repeats, CIR147 centromere repeats, Ingi clade transposable elements, and minichromosomal 177 bp repeats, respectively. TALEs had never been reported to be expressed in trypanosomes, and there was some concern over their potential toxicity to the cell. In all cases, however, expression of the exogenous TALE proteins was well tolerated. Chromatin immunoprecipitation (ChIP) was used to assess the ability of each TALE to enrich its target sequence, and this revealed variable levels of enrichment depending on the target site. The variable levels of enrichment could be explored further by performing ChIP-seq, which would provide information about TALE binding genome-wide, and could identify possible off-target binding sites of the different TALE proteins. Determining the degree of off-target binding would be beneficial both by strengthening the data from Chapters 3 and 4 of this thesis, but also providing insight as to the suitability of TALEs for the various applications mentioned below.

Developing a system for expressing synthetic TALEs in *T. brucei*, and demonstrating that they can enrich a target sequence, paves the way for the future use of TALEs for multiple applications, including as enrichment tools for proteomics as outlined in this work. For example, TALEs could also be explored as tools for genome editing, transcriptional modulation, and protein tethering assays.

Immunoprecipitation coupled to label-free quantitative mass spectrometry analysis (IP-MS) of the four different TALEs gave varying results. The Telomere-TALE immunoprecipitation enriched an overlapping and independent set of proteins compared with the known telomere-binding protein TRF. In addition to identifying nearly all proteins previously identified at telomeres in *T. brucei* (Reis et al., 2018), the Telomere-TALE also enriched the zinc finger proteins ZC3H39 and ZC3H40, which have previously been characterised as post-transcriptional regulators of respiratory chain protein transcripts (Trenaman et al., 2019). The Minichromosome-TALE enriched multiple kinetochore proteins and cohesin subunits, which suggested minichromosomes may assemble a kinetochore and segregate via a more conventional mechanism than previously thought (Ersfeld & Gull, 1997; Gull et al., 1998). The Centromere- and Ingi-TALEs failed to consistently enrich any proteins significantly, despite efforts at optimisation. Immunoprecipitations performed with TRF and the kinetochore protein KKT2, which were performed primarily as controls, also identified many previously unreported interaction partners for these proteins.

Overall, these analyses showed that TALEs can be used to enrich proteins associated with certain repetitive sequences without the need for chemical crosslinking. In future, this technology could be expanded to other repetitive sequences such as the 70 bp repeats present in VSG expression sites or the procyclin loci to identify proteins involved in the
regulation of lifecycle-specific genes. In addition, the failure of the Centromere-TALE to enrich any known centromere proteins should be investigated further. For example, TALEs designed to bind different 15 bp sequences within the CIR147 repeats might be more successful. The IP-MS data set generated for YFP-KKT2 in this study provides an excellent comparison data set for identifying novel centromere-associated proteins. The Ingi-TALE could also be optimised using a similar approach. Very little has been reported on transposable elements in \textit{T. brucei} in the last decade. Consequently, identifying any proteins uniquely associated with these elements would be very beneficial for the community. This is also true for the study of \textit{T. brucei} minichromosomes, which have not been studied extensively for more than 15 years. The results of the Minichromosome-TALE IP-MS provides evidence to suggest these chromosomes assemble a kinetochore, a hypothesis which should be investigated more thoroughly.

Since ZC3H39 and ZC3H40 were enriched in both the TRF and Telomere-TALE IP-MS analyses, I decided to focus the final stage of the project on elucidating whether ZC3H39 and ZC3H40 function in telomere biology in \textit{T. brucei}. Epitope-tagging of each protein followed by immunolocalisation did not identify any detectable nuclear fraction, and ChIP analysis did not show enrichment of telomeric TTAGGG repeats with either protein. IP-MS analysis of both ZC3H39 and ZC3H40 did not enrich any known telomere-related proteins other than UMSBP2 (Klebanov-Akopyan et al., 2018). Unexpectedly, these analyses identified a host of proteins involved in mitochondrial RNA editing, among others. Based on these data, ZC3H39 and ZC3H40 may be members of a complex also containing UMSBP2, DRBD2, HNRNPH/F, and several proteins belonging to the RNA editing substrate binding complex (RESC).

Although these novel interactions prompted many questions about possible roles of ZC3H39 and ZC3H40 in RNA editing, these were not pursued further. I decided to maintain focus on the original investigation of potential roles for ZC3H39 and ZC3H40 in telomere biology. UMSBP2 was the only protein with a strong telomere connection enriched in the ZC3H39 and ZC3H40 IP-MS analyses, and so I decided to attempt to verify this interaction. Tagging and IP-MS of UMSBP2 enriched ZC3H39, ZC3H40, and the other putative complex members identified in the ZC3H39 and ZC3H40 IP-MS analyses. In contrast to the nuclear localisation of UMSBP2, mitochondrial RNA editing proteins were also highly enriched in this analysis. The only nuclear proteins identified as associating with UMSBP2 suggest connections to DNA replication and transcription. These analyses point to UMSBP2 playing multiple roles within the cell, which remain to be elucidated.

Analysis of the TRF interactome in \(\Delta ZC3H39\Delta ZC3H40\) cells identified a possible decrease in the ability of TRF to bind chromatin. Telomere length analysis in \(\Delta ZC3H39\),
ΔZC3H40, and ΔZC3H39ΔZC3H40 cells did not identify any clear changes in telomere length compared to wild-type cells, but the experiment could benefit from an additional control.

Chapter 5 failed to identify the role played by ZC3H39 and ZC3H40 at T. brucei telomeres, but did identify many novel protein-protein interactions that deserve further investigation. For example, the existence of a putative complex comprising ZC3H39, ZC3H40, HNRNPH/F, UMSBP2, DRBD2, and RESC components should be investigated further. Additional epitope-tagging and immunoprecipitations could be used to confirm all of these interactions. The presence of RESC components leads to the hypothesis that this complex might play a role in RNA editing. Therefore, cell lines where individual members of the putative complex are knocked down or knocked out could be used to investigate how perturbing the complex affects RNA editing. For example, if this complex was required for U-insertion/deletion editing of certain transcripts, loss of complex function might lead to the accumulation of nonsense transcripts.

Regarding the role of ZC3H39 and ZC3H40 at T. brucei telomeres, future directions are unclear. The telomere length analyses discussed above should be repeated using additional controls discussed in Chapter 5 to ensure there is indeed no change in telomere length after loss of ZC3H39 and ZC3H40. The decrease in enrichment of chromatin factors seen in the TRF IP-MS analysis performed in ΔZC3H39ΔZC3H40 cells could be due to changes in chromatin structure. Loss of ZC3H39 and ZC3H40 results in increased VSG-switching (Trenaman et al., 2019), which could be associated with changes in chromatin structure. Changes in chromatin structure are seen in ΔH3VΔH4V cells (Müller et al., 2018), which also show increased levels of VSG-switching compared to wild-type cells. Similar to Muller et al., this could be assayed by performing Hi-C analysis in ΔZC3H39ΔZC3H40 cells to determine if their loss leads to increased interaction frequency between VSG expression sites located on different chromosomes. Similarly, assay for transposase accessible chromatin (ATAC)-seq could be performed to monitor potential changes in chromatin structure at the promoters of silent VSG expression sites. Another possibility is that ZC3H39 and ZC3H40 affect the function of the recently characterised VEX complex (Faria et al., 2019) in some way, perhaps negatively altering VEX localisation or activity.

Overall, this work successfully developed a new methodology for purifying proteins associated with repetitive sequences in T. brucei, providing both proof-of-principle and novel application of the methodology to a sequence of previously unknown protein composition. The identification of multiple novel protein-protein and protein-DNA interactions have yielded interesting insights and hypotheses which can be explored further by the lab in years to come.
Appendices

Appendix A
### Primers used in this study

<table>
<thead>
<tr>
<th>Sequence</th>
<th>General Primers</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAATACGACTCACTATAGGG</td>
<td>T7 promoter forward</td>
</tr>
<tr>
<td>CAGGAACAGCTATGAC</td>
<td>M13 reverse</td>
</tr>
<tr>
<td>GTG TCG CCC TCG AAC TTC</td>
<td>YFP.REV</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sequence</th>
<th>General Tryp Primers</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCTCTAACAGGCACGGAAG</td>
<td>Aldolase 5’UTR.FOR</td>
</tr>
<tr>
<td>AGTAAGGAGATCGCTCTG</td>
<td>PAD1 3’UTR.REV</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sequence</th>
<th>La protein NLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGCTACATGC CGCGGCAC ACAACACCAACTTATATCGTTAAATT</td>
<td>Modified La protein NLS.FOR (Gibson)</td>
</tr>
<tr>
<td>TACCTCGCGCGAGCGCTTGTGCGCGCATGT</td>
<td>Modified La protein NLS.REV</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sequence</th>
<th>TALEs Gibson Assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td>GGTAGTTAAGATCCGAACACCAACTTATATCGTTAAATT</td>
<td>Δ1-354 NLS Vector.FOR (Gibson)</td>
</tr>
<tr>
<td>CAAATCTACTACCTCGCGCGAGCGC</td>
<td>Δ1-354 NLS Vector.REV (Gibson)</td>
</tr>
<tr>
<td>CGCGAGGTAGTAGATTGAGAATCTTTGGGTGATATTCACAGC</td>
<td>TALE domain.FOR (Gibson)</td>
</tr>
<tr>
<td>CCAGAAGCTGCTGCACACTCGATGTGATG</td>
<td>TALE domain.REV (Gibson)</td>
</tr>
<tr>
<td>AGTGGGAGGTCTTTGTGATGTGTTTCCCGG</td>
<td>Ty-YFP-Ty fragment.FOR (Gibson)</td>
</tr>
<tr>
<td>TGTTCGGATCTTAAACTACCCGATCCTGTACGATCCCTTACGG</td>
<td>Ty-YFP-Ty fragment.REV (Gibson)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sequence</th>
<th>TALEs sequencing</th>
</tr>
</thead>
<tbody>
<tr>
<td>GGCCAGTTGCTGAAGATCG</td>
<td>TALseq+</td>
</tr>
<tr>
<td>CGCTACAAGATGATCATTAGTG</td>
<td>TALseq-</td>
</tr>
<tr>
<td>TCACAGCAGCAAGGAAAAG</td>
<td>TALseq.FOR</td>
</tr>
<tr>
<td>CCACCTCGATGTGATGTACCGCTTACGG</td>
<td>TALseq.REV</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sequence</th>
<th>qPCR primers</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCTTGCTGGGACACCGTTT</td>
<td>Ingi1.FOR</td>
</tr>
<tr>
<td>GCCGGGATTTCTCTGCAACAT</td>
<td>Ingi1.REV</td>
</tr>
<tr>
<td>ATGCAATATGTAAGGTTT</td>
<td>CiR147 repeats.FOR</td>
</tr>
</tbody>
</table>
### Chemical solutions

Appendix B
(i) Miscellaneous

1X PBS pH 7.4
137 mM NaCl
2.7 mM KCl
10 mM Na$_2$HPO$_4$
1.8 mM KH$_2$PO$_4$

Mowiol + DABCO
10% (w/v) Mowiol 4-88 reagent (Calbiochem)
25% (v/v) Glycerol
100 mM Tris-HCl pH 8.5
DABCO 2.5% (v/v)

TbBSF transfection buffer
90 mM NaPO$_4$
5 mM KCl
0.15 mM CaCl$_2$
50 mM HEPES-KOH pH 7.3

20X SSC
3 M NaCl
0.3 M Sodium citrate

Maleic acid buffer
100 mM Maleic acid
150 mM NaCl

(ii) ChIP buffers

Formaldehyde solution
50 mM HEPES-KOH pH 7.5
100 mM NaCl
1 mM EDTA
0.5 mM EGTA
8% (v/v) Formaldehyde

Lysis buffer 1
50 mM HEPES-KOH pH 7.5
140 mM NaCl
1 mM EDTA
10% (v/v) Glycerol
0.5 % (v/v) NP-40
0.25% (v/v) Triton X-100

Lysis buffer 2
10 mM Tris-HCl pH 8.0
200 mM NaCl
1 mM EDTA
0.5 mM EGTA

Lysis buffer 3
50 mM HEPES-KOH pH 7.5
140 mM NaCl
1 mM EDTA
1% (v/v) Triton X-100
0.1% (v/v) sodium deoxycholate

Wash buffer 1
50 mM HEPES-KOH pH 7.5
500 mM NaCl
1 mM EDTA
1% (v/v) Triton X-100
0.1% (w/v) sodium deoxycholate
**Wash buffer 2**
10 mM Tris-HCl pH 8.0
250 mM LiCl
0.5% (v/v) NP-40
0.5% (w/v) sodium deoxycholate
1 mM EDTA

**TE buffer**
10 mM Tris-HCl pH 8.0
1 mM EDTA

**Elution buffer**
50 mM Tris-HCl pH 8.0
10 mM EDTA
1% (v/v) SDS

---

**References**


Kimmel, B. E., ole-MoiYoi, O. K., & Young, J. R. (1987). Ingi, a 5.2-kb dispersed sequence element from Trypanosoma brucei that carries half of a smaller mobile element at either
end and has homology with mammalian LINEs. *Molecular and Cellular Biology, 7*(4), 1465–1475.


Passarge, E. (1979). Emil Heitz and the concept of heterochromatin: longitudinal chromosome differentiation was recognized fifty years ago. American Journal of Human Genetics, 31(2), 106. /pmc/articles/PMC1685768/?report=abstract


Wright, J. R., Siegel, T. N., & Cross, G. A. M. (2010). Histone H3 trimethylated at lysine 4 is enriched at probable transcription start sites in Trypanosoma brucei. *Molecular and
https://doi.org/10.1038/nature17640


https://doi.org/10.1038/nbt.1775

https://doi.org/10.1016/J.CELL.2015.04.018

https://doi.org/10.1038/nsmb.1406

https://doi.org/10.1093/NAR/GKZ476