



# THE UNIVERSITY *of* EDINBURGH

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.

**Complexity and dynamics of  
kinetoplast DNA in the sleeping  
sickness parasite  
*Trypanosoma brucei***

Sinclair Cooper

School of Biological Sciences  
Thesis submitted for a degree of Doctor of  
Philosophy

University of Edinburgh  
2016

# 1. Preface

## 1.1 Abstract

The mitochondrial genome (kinetoplast or kDNA) of *Trypanosoma brucei* is highly complex in terms of structure, content and function. It is composed of two types of molecules: 10-50 copies of identical ~23-kb maxicircles and 5,000-10,000 highly heterogeneous 1-kb minicircles. Maxicircles and minicircles form a concatenated network that resembles chainmail. Maxicircles are the equivalent of mitochondrial DNA in other eukaryotes, but 12 out of the 18 protein-coding genes encoded on the maxicircle require post-transcriptional RNA editing by uridylyate insertion and removal before a functional mRNA can be generated. The 1-kb minicircles make up the bulk of the kDNA content and facilitate the editing of the maxicircle-encoded mRNAs by encoding short guide RNAs (gRNAs) that are complementary to blocks of edited sequence. It is estimated that there are at least hundred classes of minicircle, each class encoding a different set of gRNAs. At each cycle of cell division the contents of the kDNA genome must be faithfully copied and segregated into the daughter cells. Mathematical modelling of kDNA replication has shown that failure to segregate evenly will eventually result in loss of low copy number minicircle classes from the population. Depending on the type of minicircle that is lost this can result in immediate parasite death or, if the loss occurred in the bloodstream stage, render the cells unable to complete the canonical life-cycle in the tsetse fly vector.

In order to investigate minicircle complexity and replication in *T. brucei* further we i) first established the true complexity of the kDNA genome using a Illumina short read sequencing and a bespoke assembly pipeline, ii) annotated the minicircles to establish the editing capacity of the cells, iii) analysed expression levels of predicted gRNA gene cassettes using small RNA data, and iv) carried out a long term time course

to measure how kDNA complexity changes over time and compared this to preliminary model predictions. The structure of this thesis follows these steps.

Using these approaches, 365 unique and complete minicircle sequences were assembled and annotated, representing 99% of the minicircle genome of the differentiation competent (i.e. transmission competent) *T. brucei* strain AnTat90.13. These minicircles encode 593 canonical gRNAs, defined as having a match in the known editing space, and a further 558 non-canonical gRNAs with unknown function. Transcriptome analysis showed that the non-canonical gRNAs, like the canonical set, have 3' U-tails and showed the same length distribution. Canonical and non-canonical sets differ, however, in their sense to anti-sense transcript ratios.

*In vitro* culturing of bloodstream form *T. brucei* for ~500 generations resulted in loss of ~30 minicircle classes. After incorporating parameters for network size and minicircle diversity determined above, model fitting to longitudinal kDNA complexity data will provide estimations for the fidelity of kDNA segregation. The refined mathematical model for kDNA segregation will permit insight into time constraints for transmissibility during chronic infections due to progressive minicircle loss. It also has the potential to shed light on the selective pressures that may have led to the apparent co-evolution of the concatenated kDNA network structure and parasitism in kinetoplastids.



## 1.2 Lay Summary

Trypanosomatids are small unicellular parasites that cause disease on a global scale. The success of these parasites is intrinsically linked to their cell biology, many facets of which have evolved to become highly unusual and sometimes bizarre. This makes trypanosomes interesting research targets for therapeutic purposes as well as for investigating their fascinating and unique cell biology.

The focus of this thesis is the mitochondrion of the sleeping sickness parasite *Trypanosoma brucei* which lives in the blood stream of its mammalian hosts. If untreated these parasites cause debilitating disease followed by a coma and eventually death. It is transmitted from one mammalian host to another by blood feeding flies. The mitochondrion is an internal structure found in nearly all complex cells that is often described as the 'power plant' of the cell. However the mitochondrion is much more than just a 'power plant', especially in trypanosomes where the processes in the mitochondrion are linked to the parasite being transmitted from one host to another. This is in part because of how the parasite stores information in the mitochondrion. The genetic information for running the trypanosome mitochondrion is stored in a unique structure called the 'kinetoplast' and is expressed in a unique way called 'RNA-editing'. The kinetoplast is made up of two types of circular DNA molecules, large ones called 'maxicircles' and thousands of smaller ones called 'minicircles'. The maxicircle encoded information is stored in a form that is incomplete. Minicircles house DNA sequences that allow the cell to fill in the missing information via RNA-editing. RNA-editing is an elaborate process and is essential to the survival of the cell. It is known that maintaining the information stored on the minicircles is essential to allow the cell to be transmitted via its insect vector. The maintenance of the minicircles is dependent on the minicircles duplicating and segregating properly when the cell divides. Defects in the

minicircle replication process result in loss of minicircles.

The primary focus of this thesis is to investigate the genetic information contained within the kinetoplast. The information that is encoded on the maxicircle is known, however the minicircles are much more complex and the true complexity is still to be elucidated. Minicircles produce small RNA molecules called 'guide RNAs', which carry out the only known function of minicircles, to direct the process of RNA editing. The guide RNAs provide the missing information required to edit the maxicircle RNAs. The number of minicircles is not known, nor is the the number of guide RNAs they produce. In order to truly understand the mitochondrion in trypanosomes we need to know what the protein coding information is as well as how it is generated. This requires knowledge of the minicircle repertoire.

The second aim is to use the information about how many minicircles there are and information about how they change over time to make predictions about how they are replicated. This will allows us to investigate how the complexity of the kinetoplast is linked to the life-cycle of the parasite.

Taken together these two aims will begin to give us an insight into why the mitochondrion in these parasites has evolved to be so different from all other organisms.

### **1.3 Declaration**

I declare that all material presented in this thesis is my own work, unless otherwise stated. The work has not been submitted for any other degree previously.

Sinclair Cooper

### **1.4 Experimental contribution**

Transfections detailed in Section 3.2.1 were carried out with assistance from Caroline Dewar. Mouse infections for cell differentiation, as detailed in Section 3.2.1 were performed by Caroline Dewar. Library preparation and sequencing of libraries constructed from TREU 667 cells (Section 3.2.6) was carried out in the lab of Dr. Torsten Ochsenreiter.

## 1.5 Acknowledgements

Many people have helped me along the way in the completion of this thesis. Firstly thanks to my two supervisors Dr. Achim Schnauffer and Dr. Nick Savill who have provided invaluable advice and stimulating discussions which have played a critical role fuelling my interest in the topic. Also for giving me the freedom to explore and learn on my own.

Thanks also go to the other members of the Schnauffer lab. To past members who were there when I started, Matt Gould and Laura Jeacock many thanks for teaching me the basics and getting me started in the world of trypanosomes. Special thanks to Caroline Dewar for expertise and donating her time to helping me generate cell lines. Also to Al for pointing me in the right direction at the very very beginning. Newer members of the lab Claudia, Migla, Karolina, Marios and Julie for providing a vibrant and fun atmosphere for doing science and a special mention to Stefan who was one of the good ones, and a great scientist.

Of course to my family to whom I am forever indebted for their support in practically every aspect of my life. Edinburgh based friends and those much further away both climbers and non-climbers for the weekends and holidays where the main concerns are “will we stay dry today?” and “what shall we eat tonight?” .

# Table of Contents

1. Preface.....	2
1.1 Abstract.....	2
1.2 Lay Summary.....	4
1.3 Declaration.....	6
1.4 Experimental contribution.....	6
1.5 Acknowledgements.....	7
2. Introduction.....	11
2.1 Kinetoplastids.....	11
2.2 Trypanosomes.....	11
2.2.1 Human African Trypanosomiasis.....	12
2.2.2 Nagana.....	13
2.2.3 Surra, Dourine and non-tsetse transmitted trypanosomiasis.....	14
2.2.4 Prevention and treatment of trypanosome infection.....	15
2.2.5 Trypanosomes as a model organism for the study of mitochondrial biology .....	17
2.3 Cell biology.....	18
2.3.1 The life cycle of <i>T. brucei</i> .....	18
2.3.2 Nuclear Genome.....	21
2.3.3 Transcription.....	22
2.3.4 VSGs.....	22
2.3.5 Glycosomes.....	23
2.4 The <i>T. brucei</i> mitochondrion.....	24
2.4.1 The mitochondrial genome of kinetoplastids; a very brief history.....	27
2.4.2 kDNA structure in <i>T. brucei</i> .....	29
2.4.2.1 Maxicircles.....	31
2.4.2.2 Minicircles.....	34
2.4.2.3 Mitochondrial transcription and RNA processing.....	36
2.4.3 RNA editing.....	40
2.4.3.1 Mechanisms of RNA editing.....	41
2.4.3.2 Mis-editing and alternative editing.....	47
2.4.4 Evolution of RNA editing.....	48
2.4.5 kDNA replication.....	50
2.4.6 Living without kDNA.....	52
2.5 The role of computational analyses of kDNA in the Next Generation Sequencing era.....	54
2.6 Aims.....	57
3. Complexity of kDNA.....	58
3.1 Introduction to project.....	58
3.2 Methods.....	58
3.2.1 Trypanosome cell culture and differentiation.....	58
3.2.2 Purification of kDNA.....	59
3.2.3 Sequencing of kDNA and quality control of reads.....	60
3.2.4 Processing sequencing data.....	61
3.2.4.1 Maxicircle Assembly.....	61
3.2.4.2 Minicircle Assembly.....	62
3.2.4.3 Assembling minicircles from publicly available short read data.....	64
3.2.4.4 Identification of gRNA genes.....	65

i. Identification of gRNA genes in assembled minicircles by alignment to the fully edited sequences.....	66
ii. Identification of gRNA genes from short reads predicted by alignment to fully edited sequences.....	67
iii. Identification of candidate gRNA genes in assembled minicircles by nucleotide bias.....	67
3.2.5 Modelling gRNA distribution.....	69
3.2.6 Small RNA preparation and sequencing.....	70
3.2.6.1 Library preparation.....	70
3.2.7 Small RNA library quality control and processing.....	71
3.2.7.1 Identification of gRNAs from small RNA reads.....	72
3.2.7.2 U-tail analysis.....	73
3.2.7.3 gRNA conservation analysis.....	73
3.3 Results.....	75
3.3.1 Establishing parameters for gRNA identification by mapping to fully edited mRNAs.....	75
3.3.2 Minicircle assembly.....	76
3.3.3 Maxicircle and minicircle copy numbers.....	80
3.3.4 Annotation.....	82
3.3.4.1 CSB and inverted repeat identification.....	82
3.3.4.2 Annotating canonical gRNA genes on minicircles.....	85
3.3.4.3 Prediction of non-canonical gRNAs by nucleotide bias.....	87
3.3.5 Small RNA analysis.....	92
3.3.5.1 Mapping small RNAs to assembled and annotated minicircles.....	96
3.3.5.2 Life cycle analysis of gRNA expression.....	101
3.3.5.3 gRNA 5' sequences from gRNAs predicted using small RNA data..	106
3.3.5.4 Sense vs Anti-sense gRNAs.....	107
3.3.5.5 U-tail analysis.....	109
3.3.5.6 gRNA conservation.....	115
3.3.6 Modelling gRNA distribution.....	117
3.3.7 Analysis of kDNA from publicly available short read DNA data.....	118
3.3.7.1 4.6.2 Assembly of minicircles from publicly available short read data.....	119
3.3.7.2 Comparing minicircles from different cell lines.....	123
3.3.7.3 Loss of editing capacity in monomorphic cell lines.....	125
3.4 Discussion.....	127
3.4.1 Minicircle assembly and annotation.....	127
3.4.2 Small RNA analyses.....	133
3.4.3 Comparative kDNA complexity.....	139
3.4.4 gRNA and minicircle conservation.....	140
3.4.5 Modelling gRNA distributions amongst minicircles.....	141
4. An experimental and in silico approach to analysing kDNA segregation.....	142
4.1 Summary of research question and approach.....	142
4.2 Materials and Methods.....	144
4.2.1 Cell line generation and culturing protocol.....	144
4.2.2 Time course culturing protocol.....	145
4.2.2.1 Pilot time course.....	145
4.2.2.2 Long term time course.....	145
4.2.3 kDNA purification.....	146
4.2.4 Restriction digest.....	146

4.2.5 Sequencing of kDNA.....	146
4.2.6 Sequence analysis.....	147
4.2.7 Modelling minicircle replication.....	147
4.3 Results.....	149
4.3.1 Pilot time-course.....	149
4.3.1.1 Sample QC and sequencing.....	149
4.3.1.2 Minicircle loss in WT <i>T. brucei</i> within ~300 generations of in vitro culturing.....	151
4.3.1.3 Minicircle copy number distributions change during in vitro culture	156
4.3.1.4 The kDNA-independent cells lost kDNA rapidly under bottlenecked culturing conditions.....	158
4.3.1.5 Pilot time course conclusions.....	160
4.3.2 Long term time course.....	161
4.3.2.1 Sample QC and sequencing.....	163
4.3.2.2 Complexity analysis.....	163
4.3.2.3 Copy number distribution.....	166
4.3.3 Model fitting.....	167
4.3.3.1 Preliminary simulations.....	168
4.4 Discussion.....	172
4.5 Outlook.....	173
5. Conclusion.....	176
6. Appendix.....	182
6.1 Additional minicircle annotation statistics.....	182
6.2 PCR validation of assembled minicircles.....	183
6.3 gRNA depth plots.....	185
6.4 A mutant cell line has minicircles enriched for A6 and RPS12 gRNAs.....	198
6.5 <i>T. b. gambiense</i> minicircle complexity.....	199
6.6 Long term time-course samples.....	200
6.7 Core chromosomal regions.....	202
7. Bibliography.....	203

## 2. Introduction

### 2.1 Kinetoplastids

Kinetoplastids are early branching eukaryotes belonging to the phylum *Euglenozoa*. They are defined by the presence of a large mass of DNA located in their singular mitochondrion, called the kinetoplast. Members of this group can be free living or parasitic. *Trypanosomatida* are an order of the group which are of particular interest as many species in this group are obligate parasites of humans and domesticated animals. These parasites cause significant disease burden but also present interesting model organisms for the study of mitochondrial biology in early branching eukaryotes.

The variety and geographical distribution of kinetoplastid parasites which infect both humans and animals is large; *Leishmania spp*, *Trypanosoma cruzi* and *Trypanosoma vivax* are all within this group and share the unusual mitochondrial organisation we see in *Trypanosoma brucei* which is the organism of interest in this study.

### 2.2 Trypanosomes

*Trypanosoma brucei* and its subspecies are important parasites of both humans and animals. These flagellated protozoa are the causative agents of Human African Trypanosomiasis (HAT or sleeping sickness) in humans and Nagana in livestock. They are extracellular parasites and are found predominantly in the bloodstream of their mammalian hosts. Both the human and animal diseases place a significant burden on those in affected areas. People who are most affected depend on agriculture for subsistence, live rurally and have high levels of exposure to the tsetse fly (*Glossina*). The tsetse is the only transmission vector for *T. brucei* and its sub-species (barring *T.*



*evansi* and *T. equiperdum*) and these parasites are part of a group referred to as the salivarian trypanosomes (this group includes *T. congolense* and *T. vivax*) (Jackson, 2014). This dependence on transmission by the tsetse has localised *T. brucei* infection to areas which are infested with the insect, this area is often referred to as the tsetse belt. The tsetse belt spans 36 countries across sub-Saharan Africa (<http://www.who.int/mediacentre/factsheets/fs259/en/>). With fewer than 12,000 new cases a year and 50,000-70,000 currently infected individuals, HAT infection is not a significant disease in a global context (Brun et al., 2010) and is considered to be a neglected tropical disease (Hotez et al., 2007; Fèvre et al., 2008; Hotez and Kamath, 2009). Animal African Trypanosomiasis (AAT) however has high disease burden and approximately three million cattle die every year as a result of AAT infection (<http://www.fao.org/ag/againfo/programmes/en/paat/disease.html>). Despite there being relatively few cases of HAT infection when compared to other diseases, those affected by both human and animal *T. brucei* infection are poor and rely on good physical health (either for themselves or the livestock that they depend on) to make a living, this further compounds poverty and slows development (Wilson et al. 1963). This is reflected in the estimated economic losses from cattle infection alone, with 1-1.2 billion US dollars lost annually in sub-Saharan Africa.

### 2.2.1 Human African Trypanosomiasis

There are two subspecies of *Trypanosoma brucei* that are infective to humans, *T. b. gambiense* and *T. b. rhodesiense*, each of which has a different rates of infection, pathology, vector (different subspecies of *Glossina*) and geographical distribution (Simarro et al., 2010). The common factor between *T. b. gambiense* and *T. b. rhodesiense* infection is lethality if left untreated as a result of their ability to resist lysis

by normal human serum (NHS). This is achieved by neutralising the ionic pore-forming activity of apolipoprotein L1 (ApoL-1) which enters non-NHS resistant trypanosomes via various pathways (one of which is the haptoglobin-hemoglobin receptor of the parasite) (Pays et al., 2014). The mechanisms by which this occurs is different for *T. b. gambiense* and *T. b. rhodesiense* (Uzureau et al., 2013).

*T. b. gambiense* accounts for over 90% of all HAT cases and is distributed around central and West Africa. Foci include riverine savannah, forests and mangroves. The pathology of this subspecies is chronic and an infection can persist for up to three years. Stage one of the infection is characterised by acute to sub-acute febrile illness. Stage two is characterised by sleep disturbance, neurological and psychiatric disorders due to invasion of the central nervous system (CNS) by parasites which have crossed the blood brain barrier (BBB) (Steinmann et al., 2015).

*T. b. rhodesiense* infection, in contrast to *T. b. gambiense*, is zoonotic and results in an acute form of the disease. Symptoms appear more quickly, within a few months or weeks of the initial infection. The disease progresses rapidly and stage two of the disease is also defined by CNS invasion. *T. b. rhodesiense* infection is localised to eastern and southern Africa where animals are thought to be the primary reservoir.

In addition to the two subspecies of *T. brucei* that typically infect humans there have been multiple cases of unusual infections by species which are generally considered non-infective (Truc et al. 2013). Whilst these cases usually correspond to defects in the host immune system there have been recent reports that at least some strains of *T. lewisi* are resistant to lysis by human serum (Lun et al. 2015).

### 2.2.2 Nagana

The common disease caused by infection with tsetse-transmitted *Trypanosoma* species

in animals is known as Nagana or Animal African Trypanosomiasis (AAT). AAT infection primarily affects domesticated livestock and is caused by three species of *Trypanosoma*, *T. b. brucei*, *T. congolense* and *T. vivax*. AAT infection can cause rapid and visible deterioration of the condition of livestock, namely lethargy, weight loss and weakness. The pathology and outcome of AAT infection is dependent on the host-parasite interaction, i.e. how resistant the cattle are to infection (termed trypanotolerance) and the sub-species of parasite with which the animal is infected. More traditional breeds of cattle that were introduced to the tsetse belt earlier as domesticated livestock (such as humpless taurine cattle in West Africa) show higher levels of trypanotolerance (Connor, 1994; Yaro et al., 2016).

### 2.2.3 Surra, Dourine and non-tsetse transmitted trypanosomiasis

Although it was previously stated that salivarian trypanosome infection is restricted to the tsetse belt there are some subspecies of *T. brucei* that have escaped this region and are propagated either by mechanical action of biting insects or by sexual transmission in horses (Lai et al., 2008).

The most geographically widespread pathogenic trypanosomatid is *T. evansi*, the causative agent of surra. As *T. evansi* does not (or cannot) complete the conventional trypanosomatid life-cycle inside the tsetse vector (see section 2.4.6) it is transmitted mechanically through a wide variety of potential vectors, including biting insects, vampire bats, sucking insects etc. (Desquesnes et al., 2013). It is thought that the most important of these transmission pathways are biting insects from (but not restricted to) the tabanid family. The variety of potential transmission vectors and modes of transmission (horizontal, vertical, iatrogenic and per-oral) has allowed this parasite to leave the tsetse belt in Africa and infect a wide variety of hosts globally:

camels in northern Africa and the middle east, buffaloes in Asia, horses in Brazil, to name but a few (Njiru et al., 2006; Desquesnes et al., 2013). There have also been reports of rare cases of humans being infected with *T. evansi* (Truc et al., 2013). Due to the variety of potential hosts available for *T. evansi*, the pathology associated with infection is highly variable, however anaemia appears to be a major component (Eyob and Matios, 2013). This variety of hosts is also reflected in the fact that *T. evansi* has evolved from *T. brucei* more than once (Carnes et al., 2015).

*Trypanosoma equiperdum* causes dourine which is a disease of equidae. *T. equiperdum* is another sub-species of *T. brucei* that has lost the ability to complete the canonical life-cycle (Lai et al. 2008; ). It is primarily transmitted sexually, although there are reports of vertical transmission (Desquesnes et al., 2013). Dourine causes severe pathology in horses. Donkeys can be carriers but do not appear to exhibit pathology.

Mechanical transmission by biting flies is also the standard route of transmission for *T. vivax* in South America (Desquesnes & Dia 2003).

#### 2.2.4 Prevention and treatment of trypanosome infection

As with all infectious diseases, prevention is the preferred method of reducing disease incidence in an at-risk population. The gold standard of disease prevention and eventually elimination is vaccination. There is currently no vaccine against trypanosome infection. The fact that many animals and some sub-populations of humans have evolved some form of trypanotolerance indicates that it is possible for the host immune system to halt and/or mitigate the terminal outcome of a trypanosome infection. Despite this there have been no successful vaccines as yet (La Greca and

Magez, 2011). As a result the only current option for prevention of trypanosome infection is a two pronged approach of reducing the tsetse fly vector population and comprehensive disease surveillance programmes. Tsetse control measures include either ground or aerial insecticide spraying, insecticide treated animals, baited traps or the release of sterile insects ([http://www.who.int/trypanosomiasis\\_african/vector\\_control/en/](http://www.who.int/trypanosomiasis_african/vector_control/en/)). All of these approaches require significant coordination from governments and local authorities across vast areas of land in order prevent re-infestation. Using sterile insects to produce a sterile population of tsetse has had some small scale success in a closed system (in the case of an island) (Vreysen M., et al. 2000). Disease surveillance is an important counterpart to any prevention programme (Horstmann, 1974), the development and use of rapid diagnostic tests (such as the card agglutination test for trypanosomiasis, or CATT (Mitashi et al., 2012)) will be key in any effort to control and eradicate trypanosome infection.

There are limited options for treatment of HAT, and diagnosis tends to be late in the infection, thus most chemotherapy must be targeted to stage II of the infection. The available drugs are not ideal for multiple reasons, including: high species specificity, the requirement for hospitalisation for treatment, and reports of resistance becoming more common. The two primary drugs for treatment of *T. b. gambiense* and *T. b. rhodesiense* (pentamidine and suramin respectively) are ineffective against stage II of either infection. Each drug has high specificity for both species and stage of parasite which they are effective against, thus requiring accurate diagnosis before treatment (Babokhov et al., 2013). Melarsoprol is a more effective drug (as it can kill both stages in *T. b. rhodesiense* and *T. b. gambiense*) however it is toxic and results in ~5% of patients dying post-treatment (Legros et al., 2002). There are also several new drug

combination strategies such as NECT (nifurtimox-eflornithine combination treatment) which have a high cure rate and low toxicity, however the potential for resistance is high due to the low doses used (Babokhov et al., 2013). Drug resistance in trypanosomes is a perennial problem and the mechanism by which parasites become resistant is generally associated with reduced drug uptake (Barrett et al., 2011). Cases of multi-drug resistance are becoming common (Barrett et al., 2011; Baker et al., 2013).

### 2.2.5 Trypanosomes as a model organism for the study of mitochondrial biology

Aside from the previously mentioned global public health and development rationale behind the scientific study of trypanosomatids there are many unusual aspects of their cell biology which make them an intriguing model organism for laboratory study. Notable and unusual aspects of trypanosomatid biology are their ability to perform trans-splicing of RNA (Preußner et al., 2012), possession of glycosomes (Michels et al., 2006), and acidocalcisomes (Rohloff et al., 2004) as well as the way in which they evade the host immune response via VSG switching (MacGregor et al., 2012).

Another characteristic and essential facet of trypanosome biology is the kinetoplast which is unique in many aspects of its structure, function, regulation, replication and maintenance (Schneider, 2001; Lukes et al., 2002). This uniqueness makes kinetoplast associated proteins and pathways appealing drug targets. In addition to, and in spite of their uniqueness, the systems associated with the kinetoplast can provide information regarding eukaryotic evolution given the estimated distance from the last common eukaryotic ancestor (LECA) (Embley and Hirt, 1998). They are of the supergroup Excavata which is far removed from other eukaryotes with well-studied

mitochondrial biology (mostly organisms of the supergroup Opisthokonta). It is accepted that a single event gave rise to the endosymbiotic relationship between alpha-proteobacteria and mitochondria-lacking host cell (van der Giezen, 2011). It has subsequently been proposed that the alpha-proteobacteria came already with a double membrane and genes which probably came from a previous endosymbiotic event (Gray et al., 1999; Gray, 2014). The kinetoplast gives us a window into an alternative outcome from this endosymbiotic event, in which a complex mitochondrial genome has evolved along with unique gene expression pathways. Conserved features between the mitochondria of other higher eukaryotes and early branching organisms such as kinetoplastids can give us information about origins, function and even insights into mitochondrial diseases of higher eukaryotes.

These aspects coupled with the relative ease with which large numbers of trypanosomes can be cultivated in the lab, the extensive characterisation of their cell and molecular biology and the availability of genetic manipulation tools makes them a particularly appealing model system.

## 2.3 Cell biology

The complexities of the trypanosomatid life-style as well as their proximity to LECA, have resulted in many aspects of the cell biology of trypanosomes that are unique. This section will briefly describe some facets of the cell biology of *T. brucei* that are unusual, particularly pertinent to the research questions addressed in this thesis or both.

### 2.3.1 The life cycle of *T. brucei*

*T. brucei* undergoes distinct changes in morphology, termed pleomorphism, which have been well characterised (Figure 2.1) (Vickerman, 1985; Vickerman et al., 1988). The morphological changes are accompanied by changes in gene expression profile and cell

cycle (Matthews, 2005).

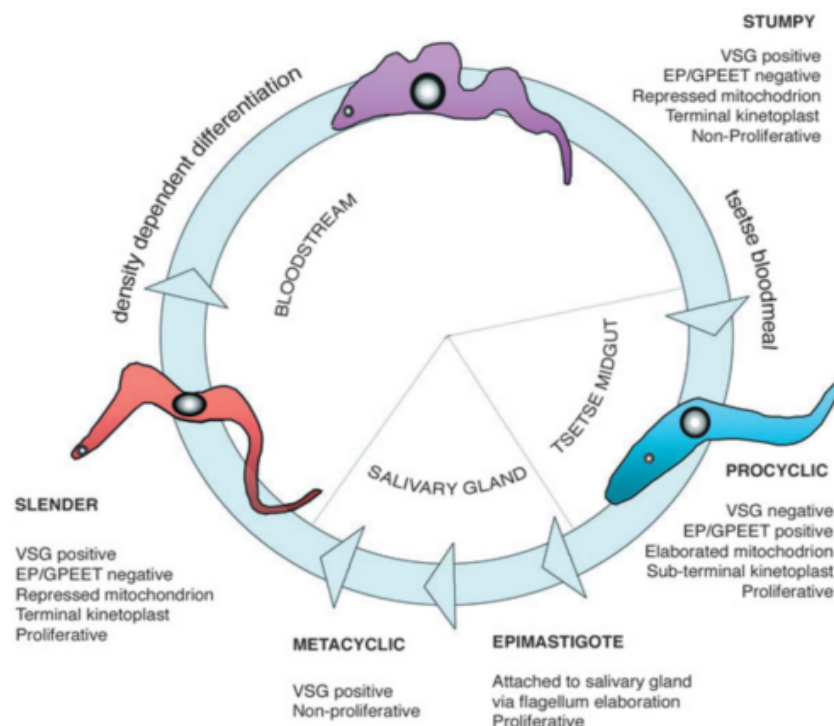


Figure 2.1 Life cycle progression in *T. brucei* (Matthews 2005).

The mammalian 'long slender' blood-stream form (BSF) parasites are proliferative and coated with a dense layer of stage-specific Variant Surface Glycoprotein (VSG) which serves as the primary method by which they evade the host immune response. The kinetoplast is located at the posterior end and the mitochondrion is in a repressed state. As the cell numbers increase BSF cells begin to transition to quiescent 'short stumpy' forms (via an intermediate stage) (Matthews, 1999). This differentiation to a form pre-adapted for transmission is triggered by a quorum sensing type mechanism, and is associated with the release of an, as yet, unidentified molecule (or set of molecules) called stumpy induction factor (SIF) (Vassella et al., 1997). The purpose of the formation of non-dividing cells at high parasitaemia is two fold; first, to prevent a high parasite burden from killing the host, thus prolonging infection time and second, to generate cells that can survive in the insect vector (Turner et al., 1995). Cells that have



lost the ability to form a short stumpy morphology in a density-dependent fashion are termed “monomorphic”. They will divide exponentially making them extremely virulent (Morrison, 2011). This has the effect of reducing the length of time that the host is infected and reduce the chances of transmission to a tsetse. Stumpy cells are partially adapted to life in the tsetse with some mitochondrial up-regulation. Proteomic analysis has shown an increase in abundance of all respiratory chain complexes compared to the slender BSF stage (Gunasekera et al., 2012). However they have a short life span of ~3 days (Turner et al., 1995). Stumpy forms express VSGs; however, they do not undergo antigenic variation (Pérez-Morga et al., 2001). A small number of BSF cells do not differentiate to stumpy forms and instead undergo antigenic variation to generate a sub-population of dividing cells that the host immune system has yet to mount an antibody response to.

The identity of SIF is not known, it is likely a small molecule (<500 Da) (Vassella et al., 1997). Incubating BSFs with cell permeable analogues of cyclic adenosine monophosphate (cAMP) have the effect of generating a “stumpy-like” phenotype (Laxman et al., 2006). The key genes involved in the BSF-to-stumpy transition have been elucidated in a genome wide screen (Mony et al., 2013). The order and importance (relative to one another) of the genes identified in this screen still need to be established.

Stumpy forms and BSFs are taken up by the tsetse fly vector in a blood meal. The BSFs die as they are not adapted to survive in the midgut and the stumpy forms differentiate into procyclic form (PCF) parasites. The transition from stumpy to PCF cells takes less than 24 hours, this transition can be reproduced *in vitro* by the addition of citrate or cis-aconitate in conjunction with a change in temperature (from 37°C to 27°C) (Engstler and Boshart, 2004). Procyclic cells are proliferative, VSG negative and

have a branched mitochondrion with the kinetoplast positioned more centrally. After 10 days of proliferation in the midgut the PCF parasites escape through the peritrophic membrane to the proventriculus of the fly and establish as proventricular PCFs. The cells then differentiate again into non-dividing proventricular mesocyclics and migrate to the salivary glands where they attach to the microvilli via their flagellum. They then divide as epimastigotes before again undergoing cell cycle arrest, disengaging from the microvilli and becoming metacyclics. The metacyclic cells express VSG and are infective to mammals. When the tsetse takes its next blood meal they are injected along with anti-coagulants in the salivary gland, into the mammalian host (Rotureau and Van Den Abbeele, 2013).

### 2.3.2 Nuclear Genome

The megachromosomes of the nuclear genome of *T. brucei* were sequenced and annotated more than a decade ago (Berriman et al., 2005). These efforts revealed new features of the genome and, arguably more importantly, paved the way for development of new tools which have opened many new avenues of research for the trypanosome biology community (Forrester and Hall, 2014). For example, RNA interference libraries (Alford et al., 2011), host/parasite co-evolution studies (Capewell et al., 2015) and comparisons to related pathogens such as *Leishmania major* (Ivens et al., 2005). In addition to these advances the genome sequencing effort was able to confirm the polycistronic nature of gene expression in trypanosomes (Siegel et al., 2011) again highlighting the unique cell biology of these parasites.

### 2.3.3 Transcription

Transcription in kinetoplastids shows many variations from the generally accepted norms in other eukaryotes (Campbell et al., 2003). The role of RNA polymerases (RNAPs) in kinetoplastids is much more flexible than in other organisms. For example RNAP I is responsible for transcribing major developmentally regulated surface proteins, procyclins, mammalian VSGs as well as ribosomal RNAs, in contrast to its role in other eukaryotes where its sole purpose is to transcribe rRNA genes (Lee and Van der Ploeg, 1997). The expression of most protein coding genes is not regulated at the level of transcription initiation. Instead expression levels of the majority of protein coding genes may be in part crudely regulated by the number of gene copies present, this is reflected by the organisation of some genes in multi-copy tandem arrays (Clayton, 2002). Regulation of gene expression also occurs at the mRNA processing level and a plethora of mechanisms exist to control the stability of an mRNA transcript. The dominant mechanism is related to the 3'-untranslated regions (UTRs) which modulate RNA degradation and translation efficiency (McNicoll et al., 2005).

### 2.3.4 VSGs

VSGs are a set of surface proteins which are a key factor in the parasite's ability to survive in the (highly hostile) mammalian bloodstream. VSGs are tightly packed on the cell surface of the BSF parasites and mask the invariant proteins on the cell surface that are required for essential activities, such as nutrient uptake (Schwede and Carrington, 2010). VSGs represent over 95% of exposed cell surface proteins and function to mitigate the effect of the host immune antibody response. The proteins are recycled via the flagellar pocket and also help to clear host antibody bound to the surface by the effect of hydrodynamic flow when moving through the host bloodstream (Engstler et al.,

2007). It has been shown that stumpy cells have a faster rate of endocytosis allowing them to clear surface-bound antibodies faster. This makes them more resistant to antibody mediated lysis, and probably contributes to why stumpy forms become the dominant cell type at peak parasitaemia.

The mechanisms of VSG gene rearrangement and expression provide an efficient method by which the parasites undergo clonal antigenic variation (Turner, 1999). This allows the parasite to switch its VSG protein coat to another isoform which in turn will illicit a new host antibody response. The expression of these different VSG genes must be regulated to ensure that only one isoform is expressed at any one time. To ensure this is the case, VSGs can only be expressed when a complete VSG gene cassette is positioned within an expression site (ES) (Borst, 1986). The mechanisms of VSG expression can involve DNA rearrangements, replacing the old VSG with a new one, or control at a transcript level (for a detailed reviews see (Taylor and Rudenko, 2006) and (Morrison et al., 2009)). However, recent evidence is showing that the concept of only one VSG per population at any one time is not quite true (McCulloch and Field, 2015). As the infection progresses, the diversity of potential VSGs is increased through the generation of mosaic VSGs, whereby portions of VSG genes recombine to produce new isoforms and the number of potential novel VSGs is vast (Hall et al. 2013).

### 2.3.5 Glycosomes

Glycosomes are peroxisome-like membrane bound organelles that are unique to and ubiquitous amongst kinetoplastids and diplomonids (Morales et al., 2016). They contain many enzymes of the glycolytic pathway and serve to sequester the process of glycolysis away from the cytosol and prevent the toxic accumulation of hexose

phosphates (Bakker et al., 2000). The redox balance in the glycosome in BSF parasites is maintained via Gly-3P/DHAP shuttling and glycerol-3-phosphate dehydrogenase/TAO activity (Michels et al., 2006).

## 2.4 The *T. brucei* mitochondrion

The *Trypanosoma brucei* life-cycle is well characterised and involves complex changes in gene expression, which in turn result in changes in morphology, metabolic pathways and cell cycle. The tight regulation of its single mitochondrion is recognised as a key player in the progression through the life cycle of *T. brucei* (Figure 2.2) (Vickerman, 1965; Matthews, 1999; Timms et al., 2002). The transition from *T. brucei* living in the blood-stream of a mammalian host to living in the midgut of a tsetse fly involves a switch in mitochondrial function. This change in mitochondrial activity reflects a switch from cytosolic glycolysis as the primary source of ATP generation (Bakker et al., 1997; Michels et al., 2006) to mitochondrial oxidative and substrate phosphorylation.

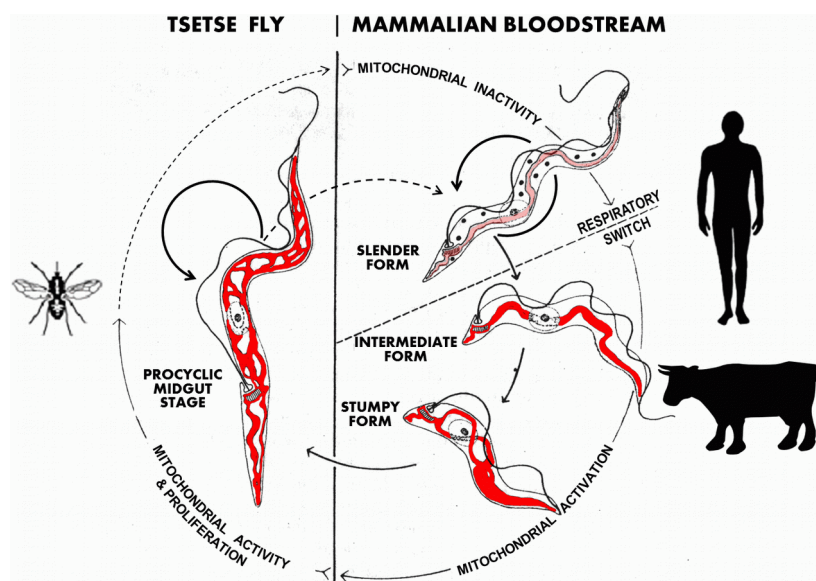


Figure 2.2: *Trypanosoma brucei* life cycle and mitochondrial activity (Vickerman 1965, modified by Achim Schnauffer).

The proliferative BSF parasites produce little or no ATP in the mitochondrion (Mazet et al., 2013); short stumpy cells however are thought to begin to produce some mitochondrially derived ATP (Bienen et al., 1993). They survive primarily on the abundant glucose found in the blood stream which they metabolise via glycolysis. This functional repression of mitochondrial ATP production in BSF parasites is reflected by the morphological repression of the BSF mitochondria which lacks a branched structure and cristae (as shown in Figure 2.2). The first seven of the ten steps of the glycolytic pathway in BSF *T. brucei* are carried out in a peroxisome-like organelle, termed the glycosome. Under aerobic conditions this process generates two ATP from one glucose molecule. The remaining steps occur in the cytosol where pyruvate is the final product, the majority of which is secreted and some of which is transported to the mitochondrion for acetate production (Visser and Opperdoes, 1980; Mazet et al., 2013). The action of aldolase during glycolysis generates two carbon intermediates, glyceraldehyde-3-phosphate (G-3-P) and dihydroxyacetone phosphate (DHAP). DHAP is converted to G-3-P which serves to regenerate glycosomal  $\text{NAD}^+$  and maintains the metabolism of G-3-P to pyruvate. G-3-P leaves the glycosome and via the action of glycerol-3-phosphate dehydrogenase (G3PDH) in the mitochondrion is oxidised to DHAP which is shuttled back to the glycosome where it is converted to G-3-P and the process of glycolysis is continued (Michels et al., 2006). In the mitochondrion, the action of G3PDH generates two electrons which are accepted by oxygen via the action of the trypanosome alternative oxidase (TAO). The TAO is a plant-like alternative oxidase essential for aerobic respiration in BSF cells as it is the only terminal oxidase (Chaudhuri et al., 1995). BSF mitochondria lack most of the key components required to carry out canonical mitochondrial function, namely the components of the (tricarboxylic acid) TCA cycle and enzymes required for oxidative phosphorylation

therefore the mitochondrial membrane potential is generated by an alternative method (van Hellemond et al., 2005). In order to maintain a mitochondrial membrane potential the function of complex V is reversed from being an ATP synthase to being an ATPase (Schnauffer et al., 2005; Brown et al., 2006). The consequence of this adaptation for the components of the respiratory chain in the mitochondrion is outlined in Figure 2.3.

Little is known about metabolic processes in short stumpy parasites due to the difficulties in obtaining sufficient numbers of these quiescent cells which have a short lifespan (Turner et al., 1995). It is known however that the transition from short stumpy to PCF involves the production of a more elaborate and “classical” respiratory chain that includes complexes III and IV. There are changes in the abundance of mRNAs and protein in short-stumpy and PCF parasites when compared to BSF cells (Priest and Hajduk, 1994; Gunasekera et al., 2012).

PCF cells have a fully operational respiratory chain and survive by metabolising proline, scavenged from the tsetse fly midgut, via oxidative phosphorylation. There are several more distinct lifecycle stages found in the insect vector reflective of the migration through various tissues with varying pH. Subsequent stages are important for preparing the parasite for re-infection into a new mammalian host (Vickerman et al., 1988). In the absence of glucose, proline is the primary carbon source which is used to generate  $\alpha$ -ketoglutarate. This in turn feeds a partially functioning TCA cycle (van Weelden et al., 2003) producing succinate and ATP. Succinate is then oxidised by complex II, which functions in a canonical electron transport chain, which contributes to oxidative phosphorylation (Bringaud et al., 2012) (Figure 2.3). The TAO in PCF parasites provides a less efficient alternative pathway for reducing oxygen as it bypasses several proton pumping steps which serve to build the proton gradient. Complex V functions as a canonical ATP synthase in PCFs and is powered by the

proton gradient generated by complexes III and IV (Tielens and Van Hellemond, 1998).

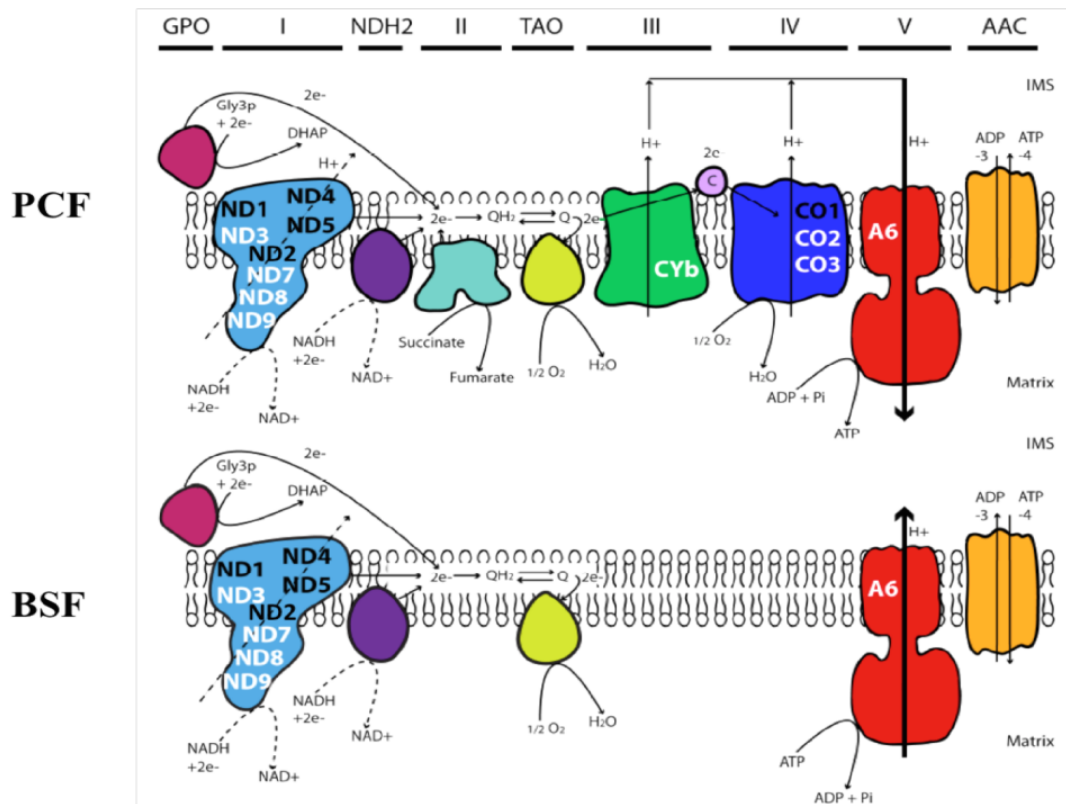


Figure 2.3 The respiratory chain of PCF and BSF *T. brucei*. Complexes I, III, IV and V contain kDNA encoded subunits. Edited subunits are in white, never-edited subunits are in black. GPO=glycerol-3-phosphate oxidase, NDH2=alternative complex I, TAO=trypanosome alternative oxidase, AAC=ADP/ATP antiporter, IMS=Intermembrane space (produced by Jeacock L. and Schnauffer A.).

#### 2.4.1 The mitochondrial genome of kinetoplastids; a very brief history

When compared to other eukaryotes, the mitochondrial genome of kinetoplastid species (from here on referred to as kDNA) is highly unusual both in terms of its structure and mechanisms by which its genomic content is expressed. The peculiarities of kDNA structure were first noticed in the 1960s shortly after the initial establishment of the importance of mitochondrial activity for progression in the trypanosome life cycle (Vickerman, 1965) and dozens of studies were subsequently carried out to investigate this intriguing structure. Out of these many studies only a few highlights are mentioned



in this short history, mostly pertaining to the pre-sequencing era of kDNA investigations. For example cell lysis experiments in *Leishmania tarentolae* (Simpson 1968) which compounded earlier studies in other organisms, (Nass and Nass, 1963) showed the physical attachment of the kDNA to the basal body, as well as initial suggestions that the kinetoplast was a tightly packed network of DNA molecules. In the same year, electron microscopy images, following isolation by caesium chloride fractionation from *T. cruzi* began to give indications of the contents of the kinetoplast (Riou and Delain, 1968), showing small concatenated circular molecules (termed minicircles). These findings were further reinforced and expanded in Simpson and da Silva, (1971) in which the authors used more electron microscopy to show complex networks of concatenated minicircles. Simpson and da Silva (1971) also made initial estimations of the total number of minicircles per cell using DNA melting and annealing curves and estimated molecular weight of minicircles compared to the total amount of DNA present per cell. The estimations predicted  $1.3 \times 10^4$  minicircles per cell. Indications of larger circular molecules associated with the kDNA were observed later (Steinert and Van Assel, 1975) and eventually given the term maxicircle (Kleisen et al., 1976). Subsequent studies proposed, correctly, that the maxicircle was likely to be analogous to the mitochondrial DNA of the cell (Fairlamb et al., 1978). Furthermore in Stuart, (1979) the author estimated the number of maxicircles and minicircles present in each kDNA genome and proposed 45 and 5500 molecules per cell, respectively. Based on restriction digest profile analyses Stuart (1979) also postulated that maxicircles were probably mostly identical in sequence whereas minicircles were heterogeneous as had been previously assumed. After these early studies highlighted intriguing differences in the mitochondrial genome of kinetoplastid organisms when compared to other eukaryotes, various labs began to study them in detail. New

molecular biology tools and the advent of DNA sequencing would prove to be key in determining of the function of kDNA.

#### 2.4.2 kDNA structure in *T. brucei*

As previously eluded to, kDNA is an extremely complex form of mitochondrial DNA. Unlike most other eukaryotes, there is only one large mitochondrion per cell in kinetoplastids and therefore only one mitochondrial genome which makes up around 10% of the total DNA content of the cell. The physical structure of kDNA can be seen clearly in transmission electron microscopy (TEM) images of the cell (Figure 2.4A). The concatenated circular molecules are condensed into a disc-like structure closely associated with, and physically attached to, the basal body at the posterior of the cell. Due to the high density of the disc-like kDNA, it is relatively easy to isolate from lysed cells by means of low speed centrifugation (Shapiro et al., 1999). Subsequent treatment of the isolated kDNA using topoisomerase allows visualisation by electron microscopy of the two different types of molecule present in the kDNA network (Figure 2.4A) (Shapiro and Englund, 1995). The larger of the two species of molecule present is termed the maxicircle (large red ellipse, Figure 2.4A) and is catenated to a second species of smaller molecules, the minicircles (small red circle, Figure 2.4A) in the network (Shapiro, 1993). However selective removal of minicircles by linearisation using restriction enzymes suggests that maxicircles are present in an independent network within the larger kDNA network (Shapiro, 1993). The minicircles are catenated to approximately three other minicircles in a monolayer (similar to the structure of chain-mail) as predicted by network modelling in *C. fasciculata* (Chen et al., 1995b). The implications for the origins and replication machinery of such a complex structure involving two different networks of molecules in different

arrangements and different levels of concatenation are discussed in 2.4.5.

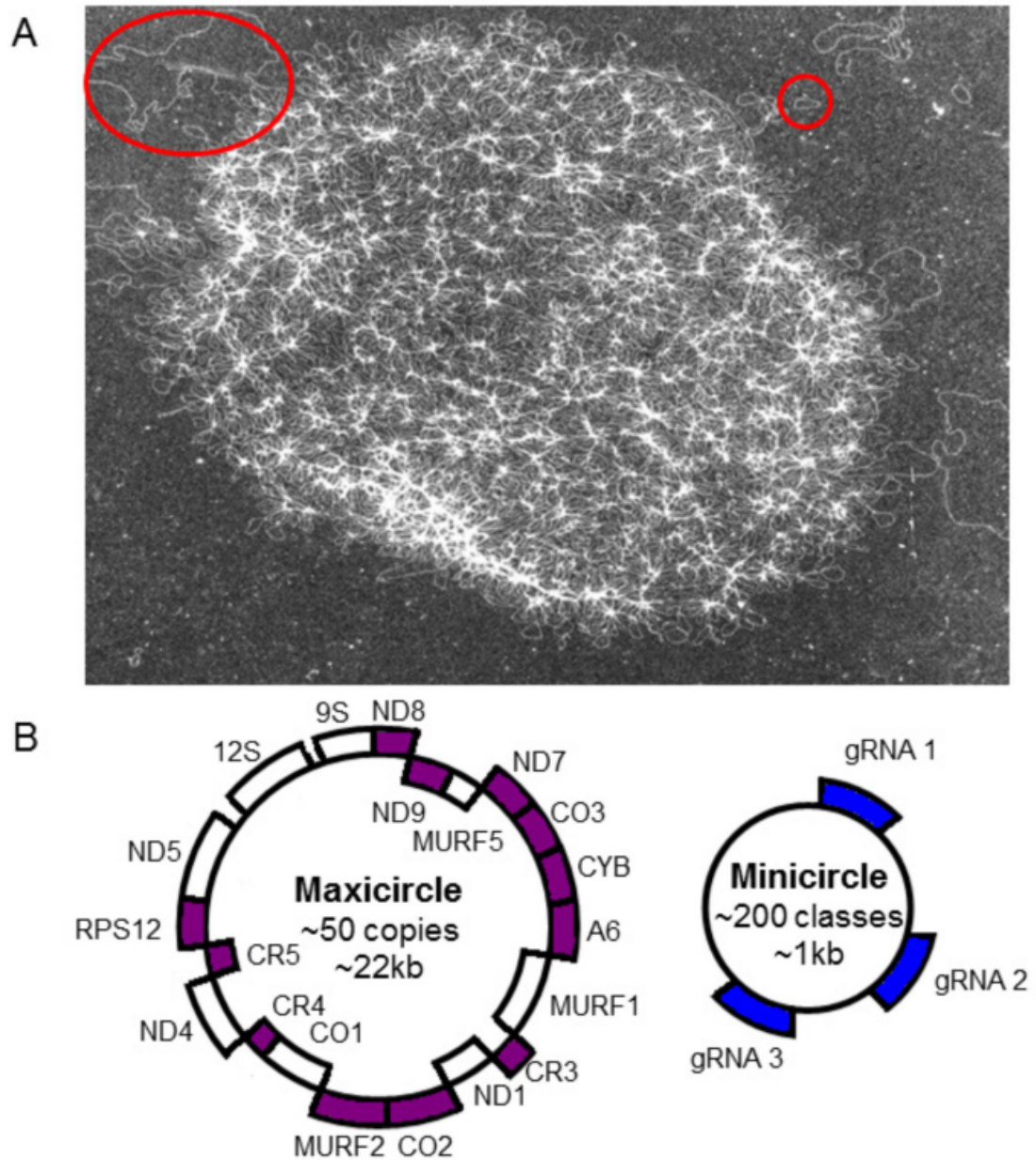


Figure 2.4: Summary of kDNA structure in *T. brucei*. (A) TEM of the kinetoplast of *T. brucei* with a maxicircle highlighted (large red ellipse) and a minicircle (small red circle) (source of the EM image unknown). (B) Maxicircle and minicircle schematic. The primary RNAs generated from the maxicircle that are coloured purple require post-transcriptional RNA editing before a functional protein can be generated. The gRNAs generated from the minicircle are coloured blue. (Produced by Jeacock L. and Schnauffer A.).

The observations referred to in this section and in section 2.4.1 were made across a variety of trypanosomatid species and serve to build up a good general picture of kDNA structure. From here on, the kDNA of *T. brucei* will be described (unless otherwise stated). The subspecies *T. b. brucei* is a good representative of a complex kDNA genome and is the model used for the entirety of this study, partially due to the ease with which we can safely culture large numbers of cells in the lab and also due to its close relationship to the subspecies that are human pathogens.

#### 2.4.2.1 Maxicircles

The maxicircle is a large ~22-25 kb molecule that encodes many protein subunits of the respiratory chain as well as a protein subunit and the two rRNA subunits of the mitochondrial ribosome (Figure 2.4B). It is analogous to the mitochondrial genome of other, more conventional, eukaryotes barring the fact that 12 out of the 18 genes on the maxicircle are cryptogenes. Cryptogenes do not contain all of the information required to generate a functional protein, many of the primary maxicircle transcripts contain stop codons and often the entire sequence is different from the fully edited version. The genes encoded by the maxicircle are detailed in Table 2.1. The sequence of the *T. brucei* EATRO 427 maxicircle is known (accession M94286) and analyses have been made of most of the transcripts. The first gene to be discovered to be a cryptogene was the COII gene (Benne et al. 1986). The fact that the COII gene was conserved across various trypanosome species, coupled with partial homology to the essential COII subunit of complex IV, implied that it had a function. However, compared to canonical COII genes, the trypanosomatid sequences were found to contain an internal frameshift that would presumably result in a non-functional protein. RNA analysis confirmed

addition of four non-templated uridylyl (U) nucleotides in the mRNA. This initial discovery was followed up by further studies showing small scale editing of several mitochondrially encoded genes. These were editing events in which the addition of a few non-templated nucleotides resulted in the creation of start codons in apocytochrome b (Cyb) (Feagin et al., 1988) and the repair of internal frameshifts in cytochrome oxidase II (COII) (Shaw et al., 1989). Subsequent to these discoveries of small scale RNA editing events more substantial cases of pan-edited mitochondrial transcripts became apparent: the previously elusive COIII, a subunit of Complex IV, was found to be present in the mitochondrial genome of *T. brucei* (Shaw et al., 1988) albeit in a near unrecognisable form prior to RNA-editing. The true extent of pan-editing in the *T. brucei* mitochondrial genome quickly became apparent in the following years (Table 2.1).

Full mapping of the maxicircle pre-mRNAs shows strikingly that the non-variable, protein coding, region is quite compact (~17 kb) with many overlapping genes (Koslowsky and Yahampath, 1997). The variable region on the maxicircle is found between the 3' end of the ND5 gene and the 5' end of the 12s rRNA genes. Its size varies between kinetoplastid species (Maslov et al., 1984) and even between isolates of *T. brucei*. It is made up of rapidly evolving repetitive sequences (Borst et al. 1982), which present difficulties for modern conventional short read sequencing. Sloof et al. 1992 and Myler et al. 1993 used a cloning and Sanger sequencing approach to analyse the sequence of the variable region of *T. brucei* strains Lister 427 and EATRO 164, finding that it was approximately ~8 and ~7kb long, respectively, and made up of two sections of tandem repeats and one section of non-repetitive DNA. The authors speculated (along with others) that the variable region is the origin of replication for the maxicircle and Myler et al. (1993) found a putative Topoisomerase II binding site.

<b>Mitochondrial transcript</b>	<b>Respiratory complex/function</b>	<b>U insertions/deletions</b>	<b>Length of primary RNA</b>	<b>Length of edited mRNA</b>	<b>Citation</b>
ND1	Complex I	Not edited	960	-	
ND3	Complex I	210/13	268	465	Read et al. 1994
ND4	Complex I	Not edited	1314	-	
ND5	Complex I	Not edited	1773	-	Read et al. 1992
ND7	Complex I	553/89	783	1246	Koslowsky et al. 1990
ND8	Complex I	259/46	361	574	Souza et al. 1992
ND9	Complex I	345/20	322	647	Souza et al. 1993
Cyb	Complex III	34/0	1118	1152	Feagin et al. 1987
COI	Complex IV	Not edited	1650	-	
COII	Complex IV	4/0	659	663	Benne et al. 1986
COIII	Complex IV	547/41	463	970	Feagin et al. 1988
A6	Complex V	447/28	401	820	Bhat et al. 1990
RPS12	Ribosomal protein s12	132/28	221	325	Marchal et al. 1993
MURF1	Unknown function	Not edited	1343	-	
MURF2	Unknown function	26/4	1091	1108	Feagin & Stuart 1988
MURF5	Unknown function	Not Edited	234	-	
CR3	Unknown function	148/13	164	300	Stuart et al. 1997
CR4	Unknown function	325/40	283	568	Corell at al. 1994
9S rRNA	SSU ribosomal RNA	3'-oligo uridylation	610	-	Adler et al. 1991
12S rRNA	LSU ribosomal RNA	3'-oligo uridylation	1149	-	Adler et al. 1991

*Table 2.1 Maxicircle encoded genes in T. brucei. Genes which are edited are indicated*

as are functions if known.

#### 2.4.2.2 Minicircles

Minicircles are small (~1kb) molecules which encode non-coding 'guide' RNAs (gRNAs) that guide the insertion or deletion of uridine residues required to modify eleven of the maxicircle generated precursor mRNAs (the guide RNA for editing of COII is encoded in cis in the 3' region of the COII mRNA). Minicircles are heterogeneous and *T. brucei* kDNA has been estimated to contain 5,000-10,000 of these molecules (Stuart, 1979; Steinert and Van Assel, 1980). Steinert and Van Assel (1980) estimated the number of classes present in the *T. brucei* kDNA to be around 250. Each class can be present in highly variable copy number; from just single copies of a given class to 100s. The total size of the network and valence has been recently addressed using two computational modelling methods (Michieletto et al., 2014) and (Diao et al., 2015). Both studies proffered differing theories to explain why the kDNA network has a fairly regulated size.

Whilst the minicircles are heterogeneous, there are some features that are conserved within a population, and even some motifs that are conserved across species. Conserved sequence blocks (CSBs) are short hyper conserved sequences found in a larger ~100 bp conserved sequence region (CSR) thought to be involved in binding the replication machinery for minicircles (Abu-Elneel et al., 1999; Shlomai, 2004) (for more detail on kDNA replication see section 2.4.5). CSB1, CSB2 and CSB3 are 10, 8 and 12 bp sequences found within the CSR, inter-spaced by more variable regions. CSB3 is the most highly conserved of these three sequences (Ray, 1989). It is conserved across various trypanosomatid species and is also referred to as the Universal Minicircle Sequence (UMS) (Ntambi and Englund, 1985), making it useful

for identification of minicircles. The sequences between the CSBs can vary. However the number of nucleotides between them is highly stable and the total size of the CSR is highly conserved (Ray, 1989). CSRs are a common feature of minicircles from all kinetoplastids. However the structure, frequency and sequence varies from one species to another (Ryan et al., 1988; Ray, 1989; Yurchenko et al., 1999; Thomas et al., 2007).

Outside the conserved region, minicircles have a single DNA bend region and 2-4 gRNAs each flanked by 18 bp inverted repeats (Pollard et al., 1990; Pollard and Hajduk, 1991; Corell et al., 1993) (the organisation of minicircles varies across other species of trypanosomatid). Section 2.4.2.3 discusses the composition of gRNAs in detail and section 2.4.3 details their function. The inverted repeats and the bend region are of unknown function. Analysis of DNA bend regions in other eukaryotes suggests that the bend region plays some role in the initiation of transcription (Schroth et al., 1992), although more likely is that the bend region is important for organisation of minicircles within the network (Jensen and Englund, 2012). The 18 bp inverted repeats (Jasmer and Stuart 1986) flank most, if not all, gRNA genes and are usually around 110 bp apart. This positioning either side of the gRNA genes and almost always being ~32 nt upstream of the gRNA 5' RYAYA motif is suggestive of the inverted repeats being important for the expression of gRNAs in *T. brucei* (Pollard et al., 1990). However, the true function of the inverted repeats is not known, and the proposal that they may function as gRNA promoters has not yet been experimentally verified.

The process of RNA editing is essential and dependent on the complexity of the kDNA genome. This is because the editing process (detailed in section 2.4.3) requires complete and overlapping coverage of all editing events by a gRNA (which are approximately 60 nt in length; the details of gRNA structure are described in 2.4.2.3). Given this, the true complexity of the *T. brucei* kDNA network remains uncertain.



Historically, as mentioned above, the number of sequence classes in *T. brucei* is cited as being from 80-100 (Hong and Simpson, 2003) to 200-250 (Stuart, 1979; Steinert and Van Assel, 1980). However even these conservative estimates of kDNA complexity would produce far more gRNAs than is required simply to cover all editing events. Assuming 3-4 gRNAs per minicircle and 200 classes of minicircle this would produce ~600-800 gRNAs. Given ~3000 editing events in total, the minimum number of 60 nt gRNAs required to fully cover each editing event once is approximately ~200 gRNAs (Corell et al., 1993). The levels of redundancy are large, however they can be partly explained by the fact that different minicircles will encode gRNAs that vary in sequence but still mediate exactly the same editing events (Riley et al., 1994). It has also been suggested that another reason for the large amounts of gRNA redundancy in *T. brucei* is because it has multiple gRNAs per minicircle (Savill & Higgs 2000). This may lead to an increase in minicircle diversity (up to the maximum network size as dictated either by physical constraints or network-lattice related mechanisms) and a decrease in the number of functional gRNAs per minicircles.

#### 2.4.2.3 Mitochondrial transcription and RNA processing

A nuclear encoded mitochondrial RNA polymerase (mtRNAP) bears some homology to a T7 phage RNA polymerase and is probably responsible for transcribing both the mini- and maxicircles in *T. brucei* (Harris et al., 1990b; Grams et al., 2002; Hashimi et al., 2009). The transcription of the major strand of the maxicircle begins ~1.2 kb upstream of the 12s rRNA gene; a transcription start site on the minor strand is yet to be determined. Similar to maxicircles it appears that both strands of the minicircle are transcribed (Aphasizheva & Aphasizhev 2010; Suematsu et al. 2016). However, as yet, no definitive minicircle promoter sequence has been identified.

All polycistronic transcription of genetic information, almost by definition, requires post-transcriptional processing to some extent before a functional transcript is produced. The gRNAs and mRNAs from the mitochondria of kinetoplastids are no exception and RNA editing itself could be described as an elaborate post-transcriptional processing event. The full complexities of RNA processing in kinetoplastids (summarised in Figure 2.5) are still, after many decades of study, only just becoming apparent.

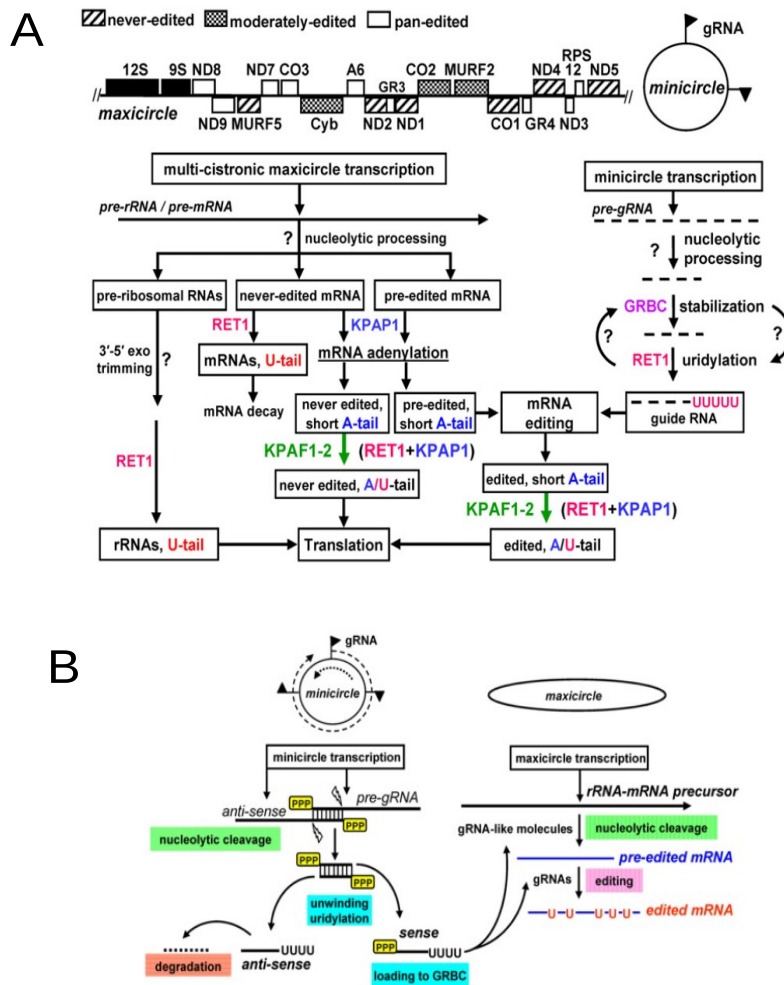


Figure 2.5: Mitochondrial RNA processing in trypanosomes. A: Polycistronic transcripts are generated from the maxicircle, cleaved by an endonuclease and undergo 3' modification. 12 Transcripts are edited. Fully edited and mature never-edited transcripts are prepared for translation by extension of the 3' tail. B: Minicircles are also transcribed polycistronically, long pre-gRNAs are produced in both the sense and anti-sense orientation. Sense and anti-sense pre-gRNAs form a duplex at the 5' end, 3' ends are modified and trimmed to produce a gRNA duplex of ~60 nucleotides. 3' U-tails of 10-15 nt are then added. (Reproduced from Aphasizhev and Aphasizheva 2011)

The polycistronic RNAs produced from the maxicircle must undergo endonucleolytic and exonucleolytic processing to release a monocistronic pre-mRNA (Koslowsky and Yahampath, 1997). The molecules that facilitate this initial cleavage of mRNA transcripts are not known, however potential candidates are components of the editing

machinery that are RNase III-type endonucleases (Stuart et al., 2005; Carnes et al., 2008). Direct evidence for these endonucleases being involved in the processing of polycistronic mRNA precursors is lacking. Endonucleolytically processed maxicircle RNAs undergo 3' modification. The length of the 3' tail is important for the stability of the transcript and the type of 3' modification is different, depending on the origin of transcript. Most mRNA transcripts are firstly polyadenylated (Militello and Read, 2000) or, in the case of 12S and 9S rRNAs, poly-uridylylated at their 3' end (Adler et al., 1991). The 12S and 9S 3' modification is carried out by the action of an RNA editing terminal uridylyl transferase (TUTase) RET1 (Aphasizheva and Aphasizhev, 2010). In the case of mRNA, short, ~20 nt, U, A/U or A-tails (oligo-A/U<sub>20</sub>) are added to never-edited and pre-edited mRNAs (Hashimi et al., 2013). These short tails are required for maintenance of the transcript (Etheridge et al., 2008), however the exact nature of the protection given by the addition of these short tails is unclear. The addition of the A and U nucleotides is carried out by Kinetoplast Poly(A)-polymerase 1 (KPAP1) and RET1 respectively (Etheridge et al. 2008). Kao & Read (2005) showed that the addition of the oligo-A/U<sub>20</sub> tails had opposing effects on the stability of the transcript depending on its editing status. For example, the addition of oligo-A/U<sub>20</sub> tails to unedited RNAs targets them for degradation, however a small amount of RNA editing in conjunction with the addition of the oligo-A/U<sub>20</sub> tail stabilises the transcript. The role of proteins that interact with the RNAs to produce this editing induced switching effect mediated by oligo-A/U<sub>20</sub> addition is being investigated (Aphasizheva et al., 2011, 2016). Oligo-A/U<sub>20</sub> tails of never edited transcripts and fully-edited transcripts are finally extended to long (200-300 nt) hetero-polymer A/U tails by the action of RET and KPAP1. This tail extension is mediated by the action of pentatricopeptide repeat (PPR) proteins including kinetoplast

polyadenylation/uridylation factor 1 (KPAF1) which have been determined to be essential for the generation of mature mRNAs in the mitochondria (Mingler et al., 2006; Pusnik et al., 2007; Aphasizheva et al., 2011, 2016; Read et al., 2011).

Discovery of minicircle derived transcripts (Rohrer et al., 1987) coupled with the previous discovery of maxicircle derived cryptogenes gave initial clues about their possible function which came to light in 1990 (Blum et al., 1990; Sturm and Simpson, 1990). The poly-cistronic nature of minicircle transcription has been indicated by the presence of longer (~800 nt) precursor gRNA molecules (Grams et al., 2000; Aphasizheva and Aphasizhev, 2010). It is now also clear that the 5' ends of all gRNAs have a triphosphate group indicative of a lack of 5' processing (Pollard et al., 1990; Aphasizheva et al., 2014; Suematsu et al., 2016b). Precursors of gRNAs must subsequently undergo significant post-transcriptional processing to generate the short (~60 nt) mature gRNAs (Blum and Simpson, 1990). The mechanisms by which the precursor RNAs transcribed from minicircles are stabilised and processed to their functional form are now becoming more apparent (Suematsu et al. 2016). It has been suggested that both strands of the minicircle are transcribed, forming a long precursor gRNA duplex with overlapping 5' ends. The 3' ends are then processively degraded until a gRNA duplex of the correct length (~60 nt) is generated, the duplex is then re-uridyated and the anti-sense transcript is degraded. Suematsu et al (2016) implicate two enzymes as being the primary players in this processing machinery that they collectively term the mitochondrial 3' processome (Mpsome): the aforementioned RET1 and DSS1 (an exonuclease). These results are compelling and the key role of 3' uridylation by a TUTase enzyme bears some similarity to other small and microRNA biogenesis systems (Heo et al., 2012; Kim et al., 2015).

The characteristics of mature gRNAs are slowly becoming more concrete after

the initial low throughput biochemical and molecular biology approaches. As previously stated mature gRNAs are approximately ~60 nt long. They have a 3' oligo-U tail (Blum and Simpson, 1990) that is added post-transcriptionally (Aphasizhev et al., 2003). The initial observation that U-tails are generally between 5-25 nt in length is probably not universally true (Aphasizheva et al., 2014). The 5' end of the gRNA often contains an RYAYA motif (Pollard et al. 1990; Madej et al. 2008) or, less often, a tract of A nucleotides (Koslowsky et al., 2014). The function of the oligo-U tail is unclear. Initial theories were that the U-tail supplied the Us for insertion during the editing process via the formation of gRNA-mRNA chimeras (Blum et al., 1991). However *in vitro* systems for RNA editing firmly established such chimeras as side products of editing and identified the pool of free UTP as a source for inserted Us (see below). Furthermore, it was subsequently shown that the U-tails are not used up during editing, and it was proposed that the gRNA U-tail may serve to help stabilise the gRNA-mRNA interaction by recognising the purine rich pre-edited sequence (McManus et al., 2000; Leung and Koslowsky, 2001).

### 2.4.3 RNA editing

RNA editing was first described in kinetoplastids (Benne et al., 1986), but has since been described in various forms in many other eukaryotes (Simpson and Emeson, 1996; Smith et al., 1997). The extensive post-transcriptional insertion and deletion of uridine residues for a long time appeared to be unique for mitochondrial RNA transcripts of kinetoplastids, but recently U insertion editing has been discovered in diplomonads and calcarean sponges (Lavrov et al., 2016; Moreira et al., 2016). In kinetoplastids the number of proteins known to be associated with this process is large and growing, and the production of gRNAs to facilitate the editing process is the only

known function of minicircles. In *T. brucei*, 12 out of the 18 maxicircle mRNA transcripts require RNA editing before a functional protein can be translated (Estévez & Simpson 1999). At the DNA level, many genes encode frameshifts and start/stop codons are missing (Benne et al., 1986) and for pan-edited genes entire sections of coding sequence are generated. Key findings in the history of the analysis of the maxicircle encoded transcripts are highlighted in section 2.4.2.1 and Table 1.1.

#### 2.4.3.1 Mechanisms of RNA editing

The process of RNA editing falls into two categories: U-insertion editing and U-deletion editing. However, the basic process is very similar (Figure 2.6). It firstly involves hybridisation of the 5' end of the gRNA, the so-called anchor region, to the pre-edited mRNA sequence downstream (with respect to the maxicircle mRNA) of the first editing site. Mismatches in the gRNA-mRNA duplex result in loops being generated (Leung and Koslowsky, 2001). If the loop is on the gRNA side of the duplex, Us are inserted into the pre-edited mRNA to match guiding A or G nucleotides in the gRNA until a proper duplex is formed and the loop is removed. If the loop is on the mRNA side of the duplex then Us are removed until the duplex can be extended. This process continues iteratively until the gRNA-mRNA duplex is complete with stable Watson-Crick base pairing and G-U base pairing characteristic of RNA-RNA interactions.

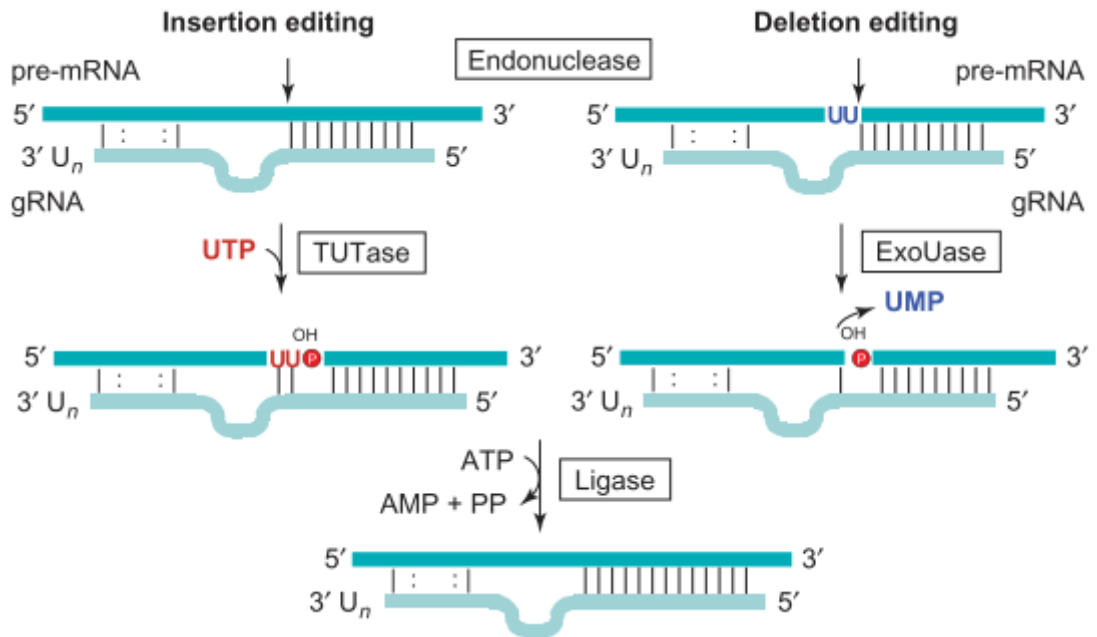


Figure 2.6: Events pertaining to insertion and deletion editing catalysed by the activity of the RNA editing core complex (RECC). Reproduced from Stuart et al. 2005. An endonuclease catalyses the breaking of the pre-edited mRNA at a position dictated by mismatches in the gRNA-mRNA duplex. U residues are either added, by a TUTase, or excised by an ExoUase. The breaks in the edited pre-mRNA strand are closed in an ATP dependent reaction by a ligase.

The insertion and deletion processes are catalysed by a multi-enzyme complex termed the RNA editing core complex (RECC or 20S editosome), which catalyses a nuclease-ligation sequence of reactions. The steps are as follows: i) cleavage at the mismatch position by an RNA-editing endonuclease (REN), ii) addition or deletion of uridylyls catalysed by an RNA editing terminal uridylyl transferase (TUTase) or a U-specific exonuclease (ExoUase) respectively, and finally iii) an ATP-dependent ligation catalysed by an RNA editing ligase (REL). This cleavage-ligation sequence of reactions was initially confirmed by reproducing individual steps of the process *in vitro* using synthetic mRNAs and gRNAs (Kable et al., 1996; Seiwert et al., 1996). This was followed by the identification of the individual enzymes required to carry out these processes, beginning with the REL (Schnauffer et al., 2001; Aphasizhev and Aphasizheva, 2014; Read et al., 2015). There are three isoforms of the RECC,



associated with distinct cleavage specificities for insertion, deletion and *cis*-editing (as found in the editing of COII, whereby the gRNA is encoded in the 3' end of the primary transcript) (Carnes et al., 2008, 2011). Each of these isoforms, (insertion, deletion and *cis*), has a different endonuclease associated, (REN1, REN2 and REN3 respectively). Whilst it has been shown that the RECC is the complex involved in catalysing the key enzymatic editing reactions, it has also been shown that it is associated with other complexes which are implicated in substrate binding, stabilisation and processing. Figure 2.7 shows a current view of the complex interactions associated with the RNA editing process. One complex that is assumed to interact with the RECC in an RNA dependent manner is termed the mitochondrial RNA-binding complex 1 (MRB1), it is also known as the RNA editing substrate binding complex (RESC) or the Guide RNA Binding Complex (GRBC). These discrepancies in nomenclature are in part due to three labs discovering the MRB1 complex at the same time in 2008 (Hashimi et al., 2008; Panigrahi et al., 2008; Weng et al., 2008). Taken together these studies identified a set of interacting proteins, some of which had predicted RNA binding domains, subsequent RNAi knock-down studies were able to further characterise the various proteins associated with the MRB1 complex. Guide RNA associated proteins (GAP1 and GAP2) were the first associated subunits to be analysed revealing their role in the stabilisation of gRNAs (Harris et al., 1990b; Weng et al., 2008; Hashimi et al., 2009). GAP1/2 RNAi had an effect on the production of edited mRNAs which require *trans*-acting gRNAs (COII was unaffected). These two proteins were also found to be co-dependent for stability (Hashimi et al., 2009). Yeast 2-hybrid analyses then identified non-transient and RNase independent interactions defining this stable complex as the MRB1 which within it contains the GAP1/2 proteins (Ammerman et al., 2012; Aphasizheva et al., 2014). The overlap between yeast 2-hybrid screens and the earlier

pull down studies in 2008 also identified a further complex and protein associated with the MRB1. TbrGG2 is the founding member of an eponymous sub-complex, and its ablation was previously shown to have an effect on RNA editing (Fisk et al., 2008). The TbrGG2 subcomplex interacts with the MRB1 in a partially RNase dependent manner (Ammerman et al., 2012). MRB1 component MRB10130 was also found to interact weakly with a large number of proteins from both the MRB1 and TbrGG2 complexes, its structure is predicted to be similar to that of the armadillo family of proteins and it is speculated that it is probably involved in complex organisation (Ammerman et al., 2012; Read et al., 2015). For a comprehensive list of proteins associated with either MRB1 or the TbrGG2 complexes see (Hashimi et al., 2013; Read et al., 2015; Aphasizheva and Aphasizhev, 2016) and an outline of the nomenclature see (Simpson et al., 2010). In addition to the proteins that have been found to be directly associated with the MRB1 complex or the TbrGG2 complex. The RNA editing helicases (REH) 1 and 2 are gRNA binding proteins that have some bearing on gRNA stability (Hernandez et al., 2010) (much like the GAP1/2 proteins) and a role in RNA unwinding and could perhaps interact with the MRB via RNA base pairing (Kumar et al., 2016).

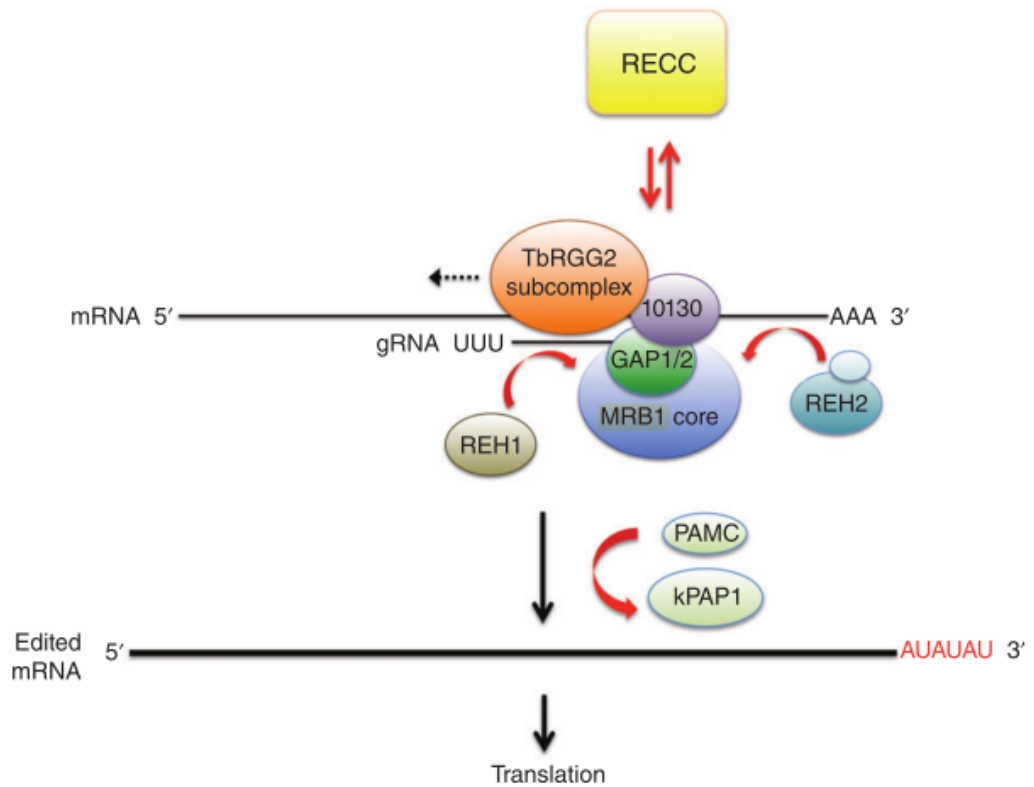


Figure 2.7: Current view of complex interactions during the RNA editing process Taken from Read et al. (2016). The MRB1 complex, is comprised of GAP1/2, which binds gRNAs, the TbRGG2 subcomplex which facilitates the 3'-5' progression, MRB10130 which facilitates complex organisation and other proteins. RNA helicases (REH1/2) promote RNA association with the MRB1 complex. Appropriate isoforms of the RECC (insertion/deletion/cis) associate transiently with the MRB1 complex. Complexes kPAP1 and PAMC add an A/U-tail, which is dependent on the completion of editing.

The exact functions of many of the proteins in these complexes are still being investigated and high throughput RNA sequencing is beginning to give insights into how complex function relates to RNA structure and stability and how this might impact regulation of expression. Aphasizheva et al. in 2014 used combination of high throughput RNA-seq, proteomic, genetic and functional studies in order to try and pull together and replicate information from various previous studies about the RNA-editing complexes. The authors were able to mostly confirm the combined role of the MRB1, TbRGG2 and MRB10130 in the initiation, progression and completion (via poly(A)<sub>20</sub>

addition) of editing. They also substantiated earlier reports (Corell et al., 1993; Savill and Higgs, 2000; Madej et al., 2007, 2008a) that gRNAs (as defined by a match in the editing space; from here on I will refer to these as canonical gRNAs) only represent a fraction of the total small RNA transcripts in the mitochondrion. This calls into question how one defines gRNAs, and what constitutes a functional vs a non-functional gRNA. It has been noted that the criteria by which matches to the editing space are defined is very close to statistical noise (von Haeseler et al., 1992). Seemingly, gRNAs with no match in the editing space still bear some of the hallmarks of a canonical-gRNA, notably 3' U-tails. The trend for gRNAs with 10-15 nt U-tails to be associated with components of the MRB1 complexes but not with the RECC suggests that perhaps the role of gRNA presentation to the RECC by the MRB1 complex is U-tail dependent. Madina et al. 2014 carried out a similar study (with different subunits of the MRB1 complex, pull-downs of the REH2 helicase and and the MRB3010 protein (a protein within the MRB1 core) followed by deep sequencing) which suggested that the MRB3010 protein is required for the initiation of RNA editing at specific blocks.

Taken together, there are a number of complex processes involved in successful editing of maxicircle transcripts in kinetoplastids, and much progress has been made in elucidating the key players and order of events. Many questions are still outstanding, however. What are the functions for the remaining MRB1 subunits which are apparently essential to the editing process, but the function has yet to be determined? How is life-cycle specific editing achieved? Are gRNAs partly involved in differential expression of maxicircle transcripts? What is the role, if any, of non-canonical gRNAs? Is the initiation of RNA editing the key rate limiting step? What triggers the addition of long (A/U)<sub>~200</sub> tails to fully edited and never edited mRNAs?

#### 2.4.3.2 Mis-editing and alternative editing

So called 'mis-editing' of maxicircle transcripts is a phenomenon in which the junction regions between fully edited and unedited regions are incorrectly edited, it was noticed early on in the investigations into RNA editing (Sturm et al., 1992; Maslov et al., 1994). It has been proposed as an integral part of the guiding process (postulating that editing is effectively random and that gRNAs serve to stabilise correctly edited transcripts) (Weiner and Maizels, 1990; Koslowsky et al., 1991). This theory was questioned following elucidation of RNA editing machinery, which suggests more precise insertion or deletion of nucleotides. However, recent high throughput methods are showing that the appearance of mis-editing and editing intermediates are quite frequent (Simpson et al., 2016; David et al., 2015) suggesting a role in the editing process. The process by which these errors in the RNA editing arise is unclear, however it is likely caused by mis-annealing of gRNA anchors to incorrect mRNA transcripts, facilitated by the low complexity seen in the anchor regions of many identified gRNAs (Koslowsky et al., 2014). The repair of mis-edited transcripts is essential and is perhaps part of the explanation for the levels of redundancy we see in gRNA populations.

Alternative editing is a proposed mechanism by which kinetoplastids may generate protein diversity (Ochsenreiter et al., 2008b) and it is recognised as a common feature of other organisms that carry out other forms of RNA editing (Rosenthal, 2015). A single example has been demonstrated to produce a stable protein (Ochsenreiter and Hajduk, 2006) which is proposed to be involved in kDNA maintenance (Ochsenreiter, Anderson, et al. 2008). The authors used low throughput methods to identify potential alternatively edited transcripts for a number of mRNA species. New high throughput sequencing of both small RNAs (Madina et al., 2014) and longer RNAs (Simpson et al., 2016) (Jeacock et al. 2016 – in revision) in conjunction with novel computational algorithms, are providing more cases whereby alternative open reading frames (ORFs)

leading to increases protein diversity could be generated. Analyses of this nature combined with RNAi of RNA editing subunits of interest and/or unknown function could shed light on editing progression.

#### 2.4.4 Evolution of RNA editing

The acquisition of a mitochondria-type organelle is generally accepted to have been a single event (Archibald, 2015). Mitochondrial RNA editing mechanisms on the other hand have seemingly arisen later in the eukaryotic lineage and have appeared multiple times. Looking across the spectrum of eukaryotes that carry out mitochondrial RNA editing, we see many different, phylogenetically isolated, mechanisms (Gott and Emeson, 2000; Gray, 2003; Lavrov et al., 2016; Moreira et al., 2016). There are several theories for why such a complex system for U insertion/deletion editing evolved in kinetoplastids, which are mostly speculative at this stage. One proposed reason behind why editing persists has been postulated as being related to life-cycle complexity. The implication being that a multi-host, parasitic, life-cycle gives a selective advantage to organisms with segregated genomes (Speijer, 2006). The fact that some free living species of euglenozoa, for example *Bodo saltans*, have segregated genomes and carry out RNA editing (Blom et al., 1998), calls into question the universality of this argument. Despite this, it is still true that organisms within this phyla that are parasitic have more elaborate RNA editing patterns than non-parasitic organisms. Another theory is that RNA editing is “on the way out”, and rather than being selected for (as suggested by Speijer (2006) and by Arts and Benne, (1996)), is in the process of slowly being selected against as reverse transcription converts edited mRNAs into cDNA, and steadily replaces cryptogenes with functional ones via recombination. This theory was suggested as the crown species of the phylum seem to undergo less editing than earlier

branching species, with editing being increasingly restricted to the 5' end of several transcripts (Figure 2.8) (Arts and Benne, 1996; Simpson et al., 2000). The ability for RNA editing to generate protein diversity has also been suggested as a potential reason for the evolution and persistence of RNA editing (as described in section 2.4.3.2), however conclusive evidence for this is so far lacking, and the negative effects of having one coding region for two genes may outweigh any benefits. Constructive neutral evolution is a process by which complexity arises by chance with no selective advantage given, becomes essential and through a ratchet-like process becomes ever more complex (Stoltzfus, 2012). This has been proposed by Lukes et al. (2009) to be the method by which the complexity we see in the kDNA genomes of trypanosomatids has come to be. It has been pointed out that there are some discrepancies with this theory: i) accumulated neutral changes can render them costly rendering them no longer neutral ii) the ratchet-like process cannot be reversible, however the cDNA-recombination theory suggests that it could be reversed in trypanosomatids (Speijer, 2010).

None of these theories taken together offer any convincing explanation as to how the RNA editing process and associated genome complexity came to be, nor why it persists. This remarkable complexity and diversity in biological systems is a trend repeatedly seen across protists (Lukes et al., 2009) and highlights how early lineages of life have developed complex and diverse mechanisms to solve the problem of living. This is especially true for the diversity of mitochondrial genomes in euglenozoa (Lukes et al., 2005; Faktorová et al., 2016).

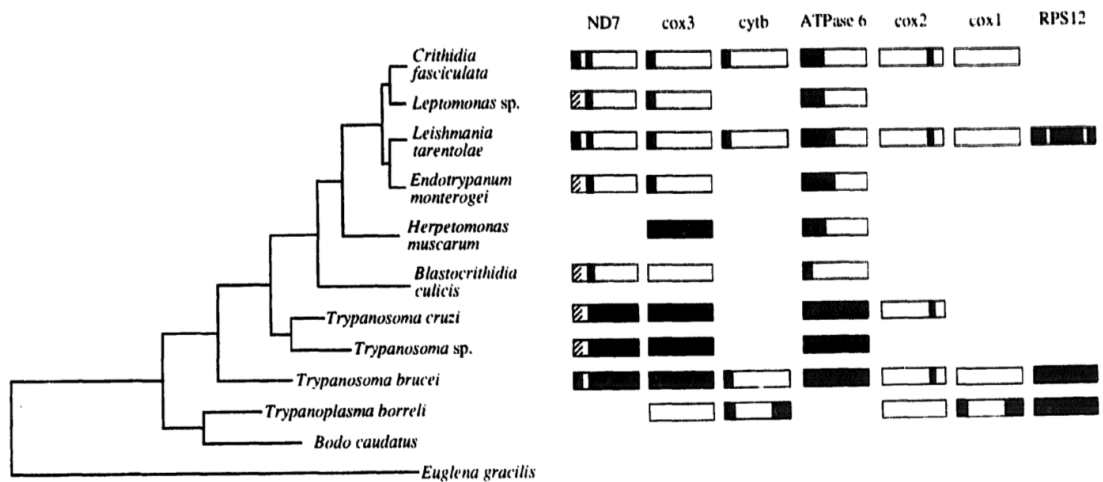


Figure 2.8: Phylogeny of RNA editing in kinetoplastids based on 18S rRNA sequences. Editing patterns are shown on the right. Edited regions are shown in black, unedited are shown in white. Taken from Arts and Benne (1996).

#### 2.4.5 kDNA replication

Trypanosomes replicate by binary fission and the replication of kDNA occurs during a distinct period of the cell cycle, at around the time of the nuclear S phase (Woodward and Gull, 1990). The complexity of the kDNA network in *T. brucei*, with its thousands of linked circular molecules, poses an interesting question for genome replication and segregation. There are upwards of 100 proteins implicated in the maintenance and replication of kDNA (Jensen and Englund, 2012). The minicircles and maxicircles are covalently closed, arranged tightly together, and are physically attached to the flagella via the Tripartite Attachment Complex (TAC). Briefly, the steps required for the replication of each of the 5,000-10,000 molecules are as follows: i) release from the kDNA disc by topoisomerase II (TopoII) (Wang and Englund, 2001), ii) replication via the formation of  $\theta$  structures, probably involving binding of the UMSBP via the CSB region (Abu-Elneel et al., 1999), as well as DNA polymerases (Pols) 1B and 1C (Bruhn et al., 2011) and a primase (Hines and Ray, 2010, 2011), iii) segregation of daughter minicircles to the antipodal sites of the kDNA disc (Liu et al., 2005), iv) primers are



removed, gaps filled by Pol  $\beta$  and nicks sealed by DNA ligase  $\kappa\beta$  (Saxowsky et al., 2003; Sinha et al., 2004), v) minicircles are then reattached to the periphery of the network by Topo II (Liu et al., 2005; Jensen and Englund, 2012). The mechanism of segregation and repositioning of the newly replicated minicircles at the antipodal sites of the kDNA disc is not known. Newly replicated minicircles are characterised by a nick in sequence as shown in *C. fasciculata* (Pérez-Morga and Englund, 1993) which must eventually be closed in the final step. It has also been shown, using similar nucleotide labelling techniques, that the mechanism for segregation is different in *T. brucei* (and some other flagellates) from most other trypanosomatids (Perez-Morga and Englund, 1993; Guilbride and Englund, 1998). Most trypanosomatids seemingly distribute newly replicated minicircles in a ring-like structure around the periphery of the rotating kDNA ring, 180° apart. *T. brucei* does not rotate the kDNA disc and newly synthesised minicircles build a dumbbell-like structure, with the centre of the disc being emptied as the antipodal sites increase in mass (Liu et al., 2005; Jensen and Englund, 2012). Eventually all the minicircles are replicated, and the kDNA disc is split. The mechanisms that control this are not known, although it is becoming clear that the TAC plays an important role in the segregation mechanism. Protein p166 is the first component of the TAC found in an RNAi screen for proteins whose knock-down was associated with loss of kDNA (Zhao et al., 2008). RNAi induction of p166 resulted in some cells with double sized kDNA networks and others with completely ablated kDNA. AEP-1 (as mentioned in section 2.4.3.2) was found to be part of the TAC, and expression of a truncated version of AEP-1 resulted in a similar phenotype to p166 knock-down (Ochsenreiter et al., 2008a). Two more TAC proteins have been identified: Mab22 (Bonhivers et al., 2008) and TAC40 (Schnarwiler et al., 2014), both of which are also essential for proper kDNA segregation.

The question of how the replicated minicircles are evenly segregated was addressed using a mathematical modelling method (Savill and Higgs, 1999). Experimental evidence in *C. fasciculata* showed that the presence of few highly over-represented (major classes) and many low copy number (minor) classes could be accounted for by segregating the minicircles in a random way (as suggested by (Maslov and Simpson, 1992; Thiemann et al., 1994)). The model also suggested that random segregation of this manner would eventually result in minor minicircle classes being lost. This fits with data from two lab strains of *L. tarentolae* (Thiemann et al., 1994; Simpson et al., 2000, 2015).

#### 2.4.6 Living without kDNA

As referred to in section 2.2.3, there are naturally occurring (sub)species of *T. brucei* which have reduced (dyskinetoplastic; Dk) or completely lost their kDNA (akinetoplastic, Ak) (Schnauffer et al., 2002). These cells are trapped in the mammalian host and cannot survive in the insect vector, as they likely lack the ability to express the full repertoire of genes required to efficiently metabolise carbon sources other than glucose. For example the absence of complexes I, III, IV and V results in a non-functional electron transport chain. This results in the parasites only being able to be transmitted mechanically, although it should be noted that this has increased the geographical range of some of these parasites markedly. The occurrence of Dk and Ak parasites is usually accompanied by specific point mutations in the  $\gamma$  subunit of the ATP synthase complex, which obviate the need of the bloodstream *T. brucei* to express the mitochondrially encoded subunit of the  $F_1F_0$ -ATPase (Schnauffer et al., 2005; Lai et al., 2008; Dean et al., 2013) This removal of the selective pressure to retain kDNA results in either a kDNA genome which is Dk, as in the case of *T. evansi*, which in most

examples has lost the maxicircle and has only one, or just a few, classes of minicircle remaining (Borst et al., 1987) or Ak. Lai et al. (2008) and Carnes et al. (2015) both suggest that *T. evansi* and *T. equiperdum* should be regarded as subspecies of *T. brucei* (i.e. *T. b. evansi* and *T. b. equiperdum*) and that multiple independent events have led to appearance of these new sub-species. Lun et al. (2010) proposed a series of events that would lead to naturally occurring species of Ak and Dk trypanosomes in the wild: i) mutations in proteins required for proper segregation of the kDNA leads to asymmetric kDNA division and loss of editing capacity; ii) this is followed by reduced ability to survive in the insect and further loss of kDNA complexity (as only the A6 subunit and RPS12 are required) iii) acquisition of a compensatory mutation ablating the requirement of any kDNA entirely. This is a plausible scenario, however the order of events could easily be reversed, with the acquisition of a compensatory mutation first, removing the selective pressure to retain kDNA and resulting in quick loss of minicircles and maxicircles in chronically infected hosts. This is discussed by Schnauffer (2010), in which the various mechanisms that have been proposed by Lun et al (2010) and others (Jensen et al., 2008) are evaluated. It is proposed that genetic exchange of kDNA in the tsetse fly is the method by which complexity is recouped (Gibson et al., 1997; Lai et al., 2008). It is conceivable, however, that the selective pressure of being required to express the majority of mitochondrially encoded transcripts could be sufficient to maintain kDNA complexity.

Ak and Dk mutant cell lines have been generated in the laboratory by serial treatment with sub-lethal doses of intercalating dyes (Stuart, 1971). Subsequently it was shown that introducing a single point mutation in the  $\gamma$  subunit of the ATP synthase complex, as found in laboratory induced or naturally occurring Ak and Dk parasites, was sufficient to permit total independence from kDNA (Dean et al., 2013). In the case

of *T. brucei* an L262P mutation in the C terminal part of the  $\gamma$  subunit causes a functional uncoupling of the  $F_1F_0$ . This dispenses with the requirement of kDNA as the mutant parasites no longer depend on a functional  $F_0F_1$  ATPase to generate a mitochondrial membrane potential (Schnauffer et al., 2005). Despite the fact that there are no longer any substrates available for the RNA editing process, the nuclearly encoded RNA editing machinery and other redundant proteins are still produced and imported into the mitochondrion (Schnauffer et al., 2002; Lai et al., 2008; Paris et al., 2011; Carnes et al., 2015).

It seems that kDNA and the generation of PCFs in *T. brucei* are intrinsically linked. However, the order by which events occur in the wild is unclear. Are Ak/Dk mutants a result of a loss of the ability to complete the life-cycle, or is the loss of tsetse infectivity quickly followed by the loss of kDNA complexity? Also if Dk/Ak parasites occur so often and are seemingly so effective in still being transmitted, why is the complex tsetse lifecycle maintained along with its reliance on an energy intensive system of mitochondrial gene expression?

## **2.5 The role of computational analyses of kDNA in the Next Generation Sequencing era**

The inherent complexity of the kDNA genome and its functional products has in the past, limited the resolution at which the genome and its' gene products could be investigated using traditional molecular biology methods. Next generation sequencing (NGS), in conjunction with information gleaned from previous molecular and biochemical investigation, is providing new insights into the true complexity contained within the kDNA genome.

When this study was first conceived in 2012, there had yet to be any comprehensive high throughput studies of the kDNA genome or transcriptome in any kinetoplastid. Some progress was being made in several labs. Sets of annotated minicircle sequences from various *T. brucei* strains were generated using low throughput methods and had been compiled and compared (Hong and Simpson, 2003; Ochsenreiter et al., 2007a). The first shotgun sequencing approach for minicircle assembly was carried out in *T. cruzi* (Thomas et al., 2007) using reads from a whole genome sample. Similarly, some transcriptome studies that focussed on gRNAs, whilst informative, were of limited scope due to the low throughput methods used (Madej et al. 2008a; Madej et al. 2007). They were able to confirm the general characteristics of gRNAs, as well as give indications of the scale of non-canonical gRNAs which have no match in the editing space. However, a full set of overlapping gRNAs had not yet been determined for any editing cascade in *T. brucei*. This left many open questions. What is the true complexity of the kDNA genome? How many gRNAs are there and how are they transcribed? Does regulation of guide RNAs contribute to life-cycle specific patterns of RNA-editing or generate increased protein diversity via alternative editing? Why is the kDNA genome in *T. brucei* so complex? How is this complexity maintained, and does it vary over time?

The first kinetoplastid to have both its kDNA genome and transcriptome sequenced was *Leishmania tarentolae* (Simpson et al., 2015). In addition to that study small RNA data sets, either on their own (Koslowsky et al., 2014; Kirby et al., 2016) or in conjunction with molecular biology methods for investigating protein function (Aphasizheva et al., 2014; Madina et al., 2014; Suematsu et al., 2016), have since been published and provide insights into global gRNA characteristics. This new abundance of large data sets will allow global comparative, and perhaps even phylogenetic, studies

of kDNA. The data from these studies are discussed in detail along-side the data presented in this thesis.

Mathematical modelling has been an essential tool in developing and testing hypotheses for why the kDNA genome has such a complex structure and how that relates to its function and the life-cycle of the parasite (Chen et al., 1995a; Hermann et al., 1997; Savill and Higgs, 1999, 2000; Simpson et al., 2000; Michieletto et al., 2014; Diao et al., 2015). The scope for using mathematical modelling approaches in conjunction with the large amounts of data currently available is still to be fully realised. Combinations of high dimensional data (genome, transcriptome and proteome) and computational analyses will give insight into many currently unanswered questions as well as highlighting gaps in our current knowledge.

## 2.6 Aims

The aims of this thesis are primarily pertaining to the structure, content and replication of the kDNA genome in *T. b. brucei*.

### 1. What is the true complexity of the kDNA genome in *T. brucei*?

More specifically how many unique minicircles are present in the differentiation competent AnTat90.13 strain of *T. brucei*? How many gRNAs do they encode and how is this reflected in coverage of the editing space?

### 2. How random is the segregation of the kDNA genome?

We aim to use mathematical modelling in conjunction with a long term time course analysis of kDNA minicircle complexity in *T. brucei* to finally ascertain how random the partial random segregation of minicircles during replication is.

## 3. Complexity of kDNA

### 3.1 Introduction to project

The true complexity of the kDNA genome in *T. brucei* in a single strain has not yet been established. To that end, short read sequencing has been used to assemble a close to complete set of minicircle and maxicircle contigs for *T. brucei* AnTat90.13. Detailed analyses of the sequences have then been used to annotate conserved features and gRNA genes. In addition, an analysis of kDNA complexity in other laboratory strains of *T. brucei* has been carried out.

### 3.2 Methods

#### 3.2.1 Trypanosome cell culture and differentiation

Differentiation competent *T. b. brucei* BSF strain Antat 1.1 90.13 (AnTat90.13) (Engstler and Boshart, 2004) cells were transfected (Macgregor et al., 2013) with a WT copy of the ATP synthase gamma, whilst a parallel culture was transfected with a mutant copy of ATP synthase gamma with the L262P point mutation (Dean et al., 2013) (dispensing with the requirement for kDNA in the BSF). Cells were maintained in HMI-9 (Hirumi and Hirumi, 1989) media supplemented with 10% FBS at a cell density no greater than  $10^6$  cells per ml.

In order to generate a procyclic form (PCF) equivalent of the WT BSF AnTat90.13 cell line, a single mouse was infected with BSF AnTat90.13 cells and harvested at peak parasitaemia. Mouse blood was then added to a flask containing 6 mM cis-aconitate in HMI-9 (supplemented with 10% FBS) and incubated at 37°C for 24 hours. After the 24 hour incubation period, the mouse blood had coagulated and settled in the bottom of the flask. The parasite cells were then collected from the top layer, centrifuged (3000g for 5 minutes), washed with SDM-80 (Lamour et al.,



2005) media and maintained in SDM-80 at 27°C at a cell density no greater than  $1 \times 10^7$  cells per ml.

### 3.2.2 Purification of kDNA

Kinetoplast DNA was essentially purified as described (Fairlamb et al., 1978; Pérez-Morga and Englund, 1993) using the following modifications, with helpful advice from Michele Klingbeil. Cells ( $3 \times 10^8$  cells for a typical preparation) were spun down in 50 ml Falcon tubes at 1760 g for 10 minutes, washed in PBS-G (137 mM NaCl, KCl 2.7 mM, 10 mM Na<sub>2</sub>HPO<sub>4</sub>, 1.8 mM KH<sub>2</sub>PO<sub>4</sub>, 55 mM glucose, pH 7.4) and then resuspended in 1 ml NET-100 (100 mM NaCl, 100 mM EDTA, 10 mM Tris-HCl, pH 8.0) solution. The samples were then spun for 5 minutes at 3000 rpm and the supernatant removed. Cells were then lysed by adding 870 µl NET 100, 10 µl Proteinase K (20 mg/ml; Invitrogen) and 100 µl of 10% (w/v) SDS solution before being passed through a 23 gauge needle twice. Samples were then incubated at 56°C overnight and subsequently treated with RNase A (10 µl of 10 mg/ml) at 37°C for 15 minutes. Each sample was split into two 500 µl volumes, loaded onto 700 µl of 20% sucrose cushion and spun at 20,000 g for 60 minutes. The supernatant was then removed leaving behind ~50 µl which was then resuspended in 250 µl TE buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA) and the two samples recombined and loaded onto a second 700 µl 20% sucrose cushion. The samples were centrifuged (20,000 g 60 mins), and kDNA purified from the bottom fraction using a standard ethanol precipitation method and dissolved in H<sub>2</sub>O.

For total DNA preparations DNA was purified using a phenol-chloroform based method; it was found that column based purification methods enriched for nuclear DNA. Briefly, one volume of phenol:chloroform:isoamyl alcohol (25:24:1) was

added to a cell pellet ( $3 \times 10^8$  cells for typical preparation) and agitated with a vortex for 20 seconds. Samples were then centrifuged at 16,000 x g for 5 minutes and the upper aqueous phase removed and purified using an ethanol precipitation method.

### 3.2.3 Sequencing of kDNA and quality control of reads

In order to generate a reference quality minicircle assembly multiple samples were pooled to ensure good representation of even minor sequence classes. The samples used are summarised in Table 3.2. They were all derived from an AnTat9013 (with a WT ATP synthase gamma replacement) strain and its equivalent which has been transfected with the L262P single point mutation in the gamma subunit of the ATP synthase complex (Dean et al., 2013). There are multiple samples for the WT strain (See section 4.2.2.1 for a description of timecourse samples). Minicircles assembled from all time points after T0 and from the mutant cell line were found to be subsets of the first WT time point, they could therefore be safely pooled in order to increase overall depth and increase the chances of assembling very low copy number minicircles.

Libraries were constructed from purified kDNA (unless otherwise stated) using the Illumina TruSeq DNA Sample Prep Kit or the TruSeq Nano DNA Sample Prep Kit to generate 300 bp paired end Illumina MiSeq reads with a ~550 bp insert (Edinburgh Genomics Ashworth Laboratories, West Mains Road, Kings Buildings, Edinburgh, EH93FL). In total, six samples were generated in this way, two of these six samples had genomic DNA from an akinetoplastic (AK) strain (whereby the AnTat90.13 cell line with the L262P gamma mutation was treated with sub-lethal concentrations of ethidium bromide until the kDNA becomes ablated) added in order to bulk the DNA for library preparation (Table 3.2:I,II). In addition to these six samples, two whole genome samples were generated, initially for comparison of kDNA coverage for the different

DNA preparation methods, but also for normalisation of minicircle and maxicircle copy numbers (for which average nuclear genome depth can be used as a standard). Total read counts and estimation of kDNA/minicircle coverage (calculated using the minicircle conserved 12mer, CSB3) are summarised in Table 3.2. Samples were quality checked and trimmed using fastqc (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) and Trimmomatic (Bolger et al., 2014) respectively.

### 3.2.4 Processing sequencing data

Scripts used for processing sequencing data in this study are available at <https://github.com/sinclaircooper/bioinformatics>. This includes scripts for general bioinformatics tasks as well as minicircle and kDNA specific programmes. Each script includes all the information required for running. Where samples were aligned to the nuclear genome, the *T. b. brucei* Lister 427 genome was used (obtained from <http://www.genedb.org/>). Where it is specified that the reads were mapped to the core chromosomal regions, coordinates from Table 6.3 from Appendix section 6.7 were used.

#### 3.2.4.1 Maxicircle Assembly

Contigs, generated using velvet (Zerbino and Birney, 2008) which were likely to be maxicircle fragments were identified by ublast (Edgar, 2010) to the published maxicircle sequence (M94286.1). A 20 kbp fragment was identified. The missing ~3 kbp from this contig is the variable region, which is difficult to assemble using conventional short reads. Attempts were made to close the gap and circularise the AnTat90.13 maxicircle using a long range PCR approach, this was not successful. Pre-mRNA coding regions were identified and annotated using ublast and custom scripts.

#### 3.2.4.2 Minicircle Assembly

Various assemblies were performed using Velvet (Zerbino and Birney, 2008) with various kmer values and coverage cut-off values (coverage cut-off was deemed to be important because there are likely to be large variations in the copy number for each of the classes of minicircle (Maslov and Simpson, 1992). These parameters were initially identified as being important using simulated data sets of minicircles, whereby an artificial kinetoplast genome was generated by taking a set of previously published minicircles and multiplying them to resemble variations in copy number. This analysis revealed coverage cut-off to be a key parameter for data sets with highly variable copy number. It should be noted that there is no generally accepted or applicable coverage cut-off for minicircle assembly, and this must be optimised for each data set, reflecting differences in library preparation, cell lines, kDNA coverage and kDNA complexity.

Two alternative assemblies were selected from a range of assemblies generated with parameter sweeps whereby the coverage cut-off was varied. The two assemblies were selected based on the criteria that one had a high number of diverse but incomplete contigs and another with a smaller number of complete (~1 kb) contigs. The first assembly was performed with a coverage cut-off of 10 to elucidate low coverage contigs and the second was performed with a coverage cut-off of 60. These assemblies produced 565 and 452 contigs respectively. These were then merged using the CAP3 (Huang and Madan, 1999) assembly algorithm using the default parameters.

An artefact of the way in which assembly algorithms like Velvet handle Illumina type short read data from circular molecules is that fully assembled contigs have the ends duplicated. The result of this can be seen in Figure 3.2A (Raw assembly plot) whereby the main peak of contigs is at 1.1-1.3 kbp, rather than the expected ~1 kbp. These duplicated ends can be used to test for contigs derived from circular

sources. The circularity test (similar in approach to (Jørgensen et al., 2014)) involves slicing all contigs in half and passing these pairs of halves through the CAP3 assembly algorithm; this was carried out using a bash batch processing script so that each pair of halves were processed separately (avoiding the possibility of any chimeric contigs being re-assembled if, for example, the contigs happened to be sliced in a conserved region). For a handful of cases the output of the circularity test was checked manually by mapping back reads and checking that paired reads mapped across the region where the contigs had been re-assembled.

Minicircles sequences that passed the circularity test were checked for the presence of the CSB3 12mer (GGGGTTGGTGT) (Ray, 1989) using a custom python script. There were two sets of contigs which did not pass both the circularity test and the CSB3 test. Sequences which passed the the circularity test, were the correct size (1 kbp) and failed the CSB3 test were closely inspected; they were found to have one of two variations of the CSB3 motif (GGGGTTGATGT/AGGGTTGGTGT). Another sub-set of contigs were those that failed the circularity test but contained the canonical CSB3 sequence; these sequences were not fully assembled (or the duplicated ends were too short to be reassembled by CAP3). Duplicate sequences were removed; any sequences that cluster with >95% identity were considered to be redundant using the “usearch -cluster\_fast” algorithm (Edgar, 2010).

A random selection of 13 minicircles were validated by PCR amplification from genomic DNA to confirm fidelity of the assembly process for the AnTat90.13 strain. A pair of specific primers was designed for each minicircle using the OligoPicker algorithm (Wang and Seed, 2003) (Appendix Table 6.1), and used to amplify from either total DNA or kDNA purified samples. OligoPicker allows specific design of primers by checking a list of sequences for which the primers must not match. For

example, for each of the 13 minicircles a pair of primers was designed and checked against the rest of the assembled minicircle data base for complementarity. Primers were also checked for complementarity to the published nuclear genome (*T. b. brucei* TREU 927) using ublast (Edgar, 2010). A 25 µl reaction was prepared for each primer set containing 5µl 5X PCR buffer (Promega), 0.2 µl dNTP (Promega), 1.5 µl MgCl, 0.5 µl of forward primer, 0.5 µl of reverse primer, 0.25 µl of goTaq polymerase (Promega), 12.05 µl H<sub>2</sub>O and 5 µl of purified genomic or kDNA (total DNA and kDNA purified template DNA are interchangeable and give the same results). The thermal cycle was as follows, denaturing at 94 °C for 5 minutes followed by a 30X loop of denaturing at 94°C for 2 minutes, annealing at 50°C for 1 minute and a 72°C extension for 1 minute 30 seconds. A final 8 minute extension at 72°C was carried out once the 30X loop was carried out. For each reaction, 5 µl was separated on a 1% agarose gel (1.5 g of agar, 150 ml TBE, 7.5 µl Ethidium bromide) at 100 V for 60 minutes. Gels were visualised and photographed under UV light. Amplified DNA was directly purified from the remaining PCR reaction (PCR cleanup kit, Qiagen) then either cloned into the pGEM T-Easy vector (Promega) and Sanger sequenced or directly Sanger sequenced (Edinburgh Genomics). This was also carried out for a random subset of minicircles assembled from other whole genome short read data sets obtained from the short read archive where the corresponding DNA or cell lines were readily available (Figure 3.3).

#### 3.2.4.3 Assembling minicircles from publicly available short read data

In addition to the reference set of minicircles generated from *T. brucei* AnTat90.13 minicircles were assembled from publicly available datasets. Data sets used in this study for minicircle assembly are shown in Table 3.12. Sample 388.5 is an ATP synthase L262P γ mutant generated by Dr. Caroline Dewar and is derived from Lister 427. The *T. brucei* EATRO 164 WT and AK data sets were used for optimisation of the

gRNA calling pipe (Table 3.1) and were generated from cells kindly provided by Ken Stuart, described in (Stuart, 1971). Whole genome Illumina libraries (50-bp paired-end) were generated for both the EATRO 164 WT and AK cell lines by Edinburgh Genomics.

#### 3.2.4.4 Identification of gRNA genes

Scripts for gRNA prediction are available from [https://github.com/sinclaircooper/gRNA\\_prediction](https://github.com/sinclaircooper/gRNA_prediction). gRNAs were predicted using three methods: prediction of gRNAs from assembled minicircles by alignment to fully edited sequences, prediction of the gRNAs directly from sequencing reads, again using alignment to fully edited sequences and prediction of gRNA regions by nucleotide bias. Canonical gRNA prediction has been described multiple times (von Haeseler et al., 1992; Simpson et al., 2003; Ochsenreiter et al., 2007b; Koslowsky et al., 2014). This approach was augmented by the addition of a permutation test to filter out the inevitable false positives generated in the use of large data sets (similar to Thomas et al. (2007)). My analysis was tested for specificity by using reads from two samples of the *T. brucei* EATRO 164 strain, for which there was a wild type data set and an equivalent akinetoplastic (AK) data set available (Stuart, 1971). Reads from the AK cell line were used as a negative control.

All data sets analysed in this study are available for inspection at a modified Jbrowse instance at [hank.bio.ed.ac.uk](http://hank.bio.ed.ac.uk). The web interface has tracks for all edited genes as well as drag and drop tracks for U-insertion and deletion events. Tracks for each data set are drag and drop and can be filtered by name in a search box. The track Antat9013\_DNA\_Cooper2016\_contigs shows the gRNAs called from assembled minicircles. Links to a Genbank instance of the corresponding minicircle are available upon clicking the the gRNA. Some tracks are presented with two colours for the called

gRNAs, grey bars indicate gRNAs with a mapping identity score of >85% and <90%, while Coloured bars indicate gRNAs with an identity score >=90%.

i. Identification of gRNA genes in assembled minicircles by alignment to the fully edited sequences

Assembled minicircles were chopped into 100 bp overlapping sequences (5 bp overlap). Each of these were aligned to a database of edited maxicircle genes. This edited database was constructed from the Lister 427 maxicircle (M94286) which was corrected in the non-edited bases using the maxicircle assembled in this study. Alignments were performed using the Smith-Waterman alignment algorithm, Water, from the Emboss suite of bioinformatics tools (Rice, 2000). A modified alignment matrix was used to allow G-U base pairing, characteristic of RNA-RNA interactions, including mRNA-gRNA base-pairing (Pollard et al., 1990). Each of the alignments were filtered for quality; the cut-off was set at 85% identity with no gaps and a minimum alignment length of 30 nucleotides. Each of the alignments that passed the quality filtering step were then subjected to a permutation test, whereby the query sequence was shuffled 300 times (this number of permutations was found to be sufficient for elimination of false positives and negatives) and each of these shuffled sequences re-aligned to the database of edited maxicircle mRNA. The alignment score (calculated from identity and length) for each of these shuffled sequences was then recorded to form a distribution of scores for a sequence of a given nucleotide composition. If the true alignment score fell within the top 1 percent of this distribution we considered the match to be non-random (i.e. the chance of the random match was less than 1%). Once a query sequence had passed both the above tests, it was then annotated as a gRNA gene on the minicircle from which it came. Annotations are stored in GFF3 flat file format (<https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md>) as well as Genbank format for ease of



analysis.

ii. Identification of gRNA genes from short reads predicted by alignment to fully edited sequences

Although we assembled an estimated 99% of minicircles (see section 3.3.2) there were still 1% that could not be assembled and were therefore missing. gRNAs on these missing minicircles cannot therefore be identified by the previous method. Instead we can directly map reads to the edited space to identify gRNAs on missing minicircles. Read sets were collapsed in order to reduce the size of the data sets being submitted for gRNA prediction. Only exact copies of any given read were collapsed at this stage, the number of times a given read appeared was appended to the read header. Collapsed reads were then mapped to the nuclear genome (*T. b. brucei* Lister 427 obtained from [www.genedb.org](http://www.genedb.org), version 4.2) using bowtie2 (Langmead and Salzberg, 2012) and the pre-edited maxicircle (using the Emboss “water” algorithm as described in the previous section, but with a standard DNA-DNA alignment scoring matrix). Sequences that either mapped to the nuclear genome using the bowtie2 default options or mapped to the pre-edited maxicircle with an identity higher than 85% and an alignment length greater than 30 nt were discarded. Reads that did not map to either of these data-bases were collected and then subjected to the same treatment as the sequences derived from minicircles. The resulting predicted gRNAs were collapsed one final time using a cluster based approach (Edgar, 2010) (usearch -cluster\_fast algorithm). Sequences that clustered with 95% identity or above were considered to be the same gRNA and the centroid from each cluster selected. The cluster size was then appended to the end of the gRNA ID to allow retrieval of gRNA copy number information. This approach was used for short DNA reads as well as small RNA reads.

iii. Identification of candidate gRNA genes in assembled minicircles by nucleotide bias

Guide RNAs that were predicted by the alignment method were annotated back onto the minicircles from which they originated. The coordinates from this method were used to generate an upstream training set by slicing 80 bp upstream plus 80 bp downstream from the 5' end of each gRNA sequence resulting in a set of 365 160 bp sequences. Similarly a downstream training set was generated by slicing 80bp upstream plus 80bp downstream from the 3' end of each gRNA sequence resulting in a different set of 365 160 bp sequences. The sequences in each training set were stacked and used to generate nucleotide frequencies for each position along the sequences. Let  $f_{i,j}^{(u)}$  and  $f_{i,j}^{(d)}$  be the frequencies of nucleotide  $i$  at position  $j$  in the upstream and downstream training sets respectively, with  $i \in \{A, C, G, T\}$  and where  $j \in [1, 160]$  is the position along the stacked sequences. The nucleotide frequencies were then used to generate upstream and downstream scoring vectors along each assembled minicircle. The upstream score at position  $k$  on a minicircle is the sum of the log nucleotide frequencies;

$$S_k^{(u)} = \sum_{j=1}^{160} \log(f_{n_{k-81+j}, j}^{(u)}) \text{ for } k = 81, \dots, N - 79$$

where  $N$  is the length of the minicircle and  $n_{k-81+j}$  is the nucleotide at position  $k-81+j$  on the minicircle. The downstream score at position  $k$  on a minicircle is given by

$$S_k^{(d)} = \sum_{j=1}^{160} \log(f_{n_{k-81+j}, j}^{(d)}) \text{ for } k = 81, \dots, N - 79$$

This gives a list of nucleotide bias scores for each position on each assembled minicircle. The peaks generated from the scoring vectors accurately match the start and stop positions of gRNAs identified by matching to the editing space and predict regions

on the minicircle that are likely to be gRNA genes based on nucleotide bias and position within a cassette. The two highest peaks that fell within a forward and reverse inverted repeat were chosen as the start and stop position for the putative gRNA. Note that this method will miss any gRNA gene candidates which are not flanked by inverted repeats, as a reliable statistical method for choosing multiple peaks across the length of the minicircle was not found. The approach is similar to a method described to identify gRNA genes in *T. cruzi* (Thomas et al., 2007) but dispenses with the requirement of using a Hidden-Markov model for the predictions.

### 3.2.5 Modelling gRNA distribution

After completion of the annotation process the distribution of gRNA genes amongst minicircles was known. These annotations were used to extract the number of possible gRNA cassettes. It was assumed that each cassette has the potential to contain a gRNA gene. The true distribution of gRNA genes amongst the minicircles was recorded for the gRNA genes of interest. The gRNA annotations were then randomised 1000X and each resulting gRNA distribution recorded. The random distributions were used to generate an average random distribution which was then compared to the true distribution. The two distributions (true and the randomised version) were plotted for A6 and RPS12 alone as well A6 and RPS12 combined. For each of these the output showed the chances of, for example, finding 1 A6 gRNA gene per minicircle, 2 A6 gRNA genes per minicircle or 3 A6 gRNA genes per minicircle for both the real and simulated distribution. A distribution which shows that it is extremely likely to find only one A6 and RPS12 gene per minicircle supports the hypothesis proposed by Speijer, (2006) (as this implies that A6 and RPS12 are more likely to be coupled to a non-essential gRNA gene) however if the true distribution matches the randomised

version there are other factors at play.

### 3.2.6 Small RNA preparation and sequencing

Crude mitochondrial fractions were prepared from BSF AnTat90.13, and from the same cell line after differentiation into PCF, by hypotonic lysis and differential centrifugation using approximately  $10^9$  cells, as described previously (Harris et al., 1990). Total RNA was isolated using Trizol reagent (Thermo Fisher) and the small fraction of <200bp was then size selected using the PureLink miRNA isolation kit (Thermo Fisher).

#### 3.2.6.1 Library preparation

Isolated small RNAs were then treated with a RNA 5' poly-phosphatase (Epicentre), as per the manufacturer's instructions, to degrade all 5' tri-phosphates, present at the 5' ends of gRNAs (Blum and Simpson, 1990) to mono-phosphates. Poly-phosphatase enzyme was removed using a phenol-chloroform extraction followed by sequential ligation of the Illumina 5' and 3' adapters, and subsequent generation of an Illumina small RNA library, following the manufacturers instructions.

A second pair of PCF and BSF small RNA libraries were also constructed from the *T. brucei* TREU 667 cell line. This pair of small RNA libraries were constructed using a different approach in an effort to enrich for gRNAs. First, a gRNA purification was carried out by incubating total RNA with a biotinylated oligoA. After separation using Streptavidin Magnetic Beads, the gRNA fraction was recovered and cDNA first strand synthesis performed using a tagged oligoA (with part of the Illumina P7 adapter sequence). The second strand was synthesized using RNaseH and DNA polymerase. The Illumina P5 adapter was then ligated. The resulting constructs were flanked with the P5 adapter at the 5' end and the P7 adapter at the 3' (oligoT) end, and therefore

retained the gRNA orientation. The library was completed by PCR which incorporated the index for multiplexing. This process was carried out by Dr. Torsten Ochsenreiter at the University of Bern.

### 3.2.7 Small RNA library quality control and processing

Barcode separated libraries were quality checked using fastqc and adapter trimmed using a paired end approach implemented using cutadapt version 1.9.1 (Martin, 2011). As the insert length for gRNAs is expected to be shorter than the length of the reads, Illumina adapters were searched for from both the 5' (GUUCAGAGUUCUACAGUCCGACGAUC) and 3' (TGGAATTCTCGGGTGCCAAGG) end of both the forward and reverse read. Untrimmed reads were discarded using the -discard\_untrimmed option. Reads that were shorter than 15 bp after adapter trimming were also discarded at this stage. If both the forward and reverse reads contained the 5' and 3' adapter they were then merged using PEAR (Zhang et al., 2014). Merged reads were then mapped to the nuclear genome (*T. b. brucei* Lister 427 obtained from [www.genedb.org](http://www.genedb.org), version 4.2) using bowtie2 with default settings (Langmead and Salzberg, 2012). Reads that mapped to the nuclear genome were discarded at this stage. Quality control statistics for small RNAs generated in this study are summarised in Table 3.6.

In addition to the small RNA libraries generated specifically for this study, previously published small RNA data sets from (Koslowsky et al., 2014; Madina et al., 2014; Suematsu et al., 2016) (Short read archive accession numbers SRR1041339, SRR2125761 and SAMN02795843, respectively) were used for comparison (Table 3.7). All of these small RNA data sets were generated from PCF cells. Koslowsky *et al.* (2014) used a stringent gel-based size selection method to isolate their gRNAs; a crude mitochondrial fraction was run on a denaturing 10%

polyacrylamide gel. A gRNA marker lane was run along side by 5' capping 10 µg of mtRNA with <sup>32</sup>P αGTP using the vaccinia capping enzyme. The gRNA marker lane was used to select and excise gRNAs in the size range 40-80 nt, small RNAs were then treated with a phosphatase and polynucleotide kinase to replace the 5' triphosphate with a mono-phosphate. The Illumina small RNA library prep protocol was then followed and a single end 75 bp library generated on the Illumina GAIIx platform. Suemetsu *et al.* (2016) used a similar approach to Koslowsky *et al.* (2014); mitochondrial fractions were fractionated on a 10%-30% glycerol gradient, RNA in the 35-75 nt size range was excised and processed with the ScriptMiner small RNA-Seq library prep kit (Epicentre). Single end 75 nt stranded Illumina reads were then generated. Madina *et al.* (2014) used a gel isolation method too, followed by treatment with a terminator 5' phosphate-dependent exonuclease, which will have the effect of degrading RNAs which are not primary transcripts (primary transcripts will have the 5' triphosphate). RNAs were then further processed using the Illumina small RNA library prep kit (with the addition of a phosphatase/PNK treatment step).

#### 3.2.7.1 Identification of gRNAs from small RNA reads

gRNAs were identified from the adapter trimmed and merged small RNAs using methods outlined in section 3.6.2. It was observed after calling gRNAs from the merged reads that a small proportion of reads (0.36% and 0.76% for the BSF and PCF sets respectively) still contained adapter sequence (probably due to low sequencing quality at the 3' end resulting in only part of the non-biological sequence being trimmed in the first instance). The erroneous presence of adapters after trimming in small RNA data sets was also observed by Suematsu *et al.* (2016). These erroneous sequences did not affect the ability of the gRNA calling pipe to detect gRNAs. Nevertheless they were re-trimmed using a custom script. This script has the potential to inadvertently trim off

the 3' end of the gRNA. The script uses a quality score based approach to trim the 3' ends but boosts the score of any T residue (assuming most gRNAs have an oligo-U tail, sequenced as Ts). Hence these sequences were not used for any U-tail analysis. Called, collapsed and clustered gRNAs were mapped back to the assembled and annotated minicircles (using bowtie2) to generate a high confidence set of gRNAs where both the DNA source and expressed RNA are known.

#### 3.2.7.2 U-tail analysis

U-tails for each gRNA were defined as any 3' Us up to the first non-U base. This may result in the over estimation of the length of U-tails in some cases, for example, when the 3' region of the gRNA is very U-rich. It may also result in the occasional under representation of some U-tails, for example, when a base in the U-tail is incorrectly called as another base, the U-tail read out for that particular read may be shortened. These scenarios are both rare and will have negligible effect on the overall pattern of 3' oligo-uridylation.

Using the above definition gRNA U-tails were counted and plotted as a proportion of total gRNAs identified from each small RNA read set in order to allow comparisons of dominant U-tail lengths. Additionally, the variability of U-tails for each gRNA species was calculated. For each gRNA gene, the U-tail lengths of the mapped reads were counted and binned into 5 categories (0, 1-5, 6-10, 11-15, >15); these are expressed as a percentage of total reads with a U-tail of a given length. This was carried out for small RNAs mapping to both canonical gRNA and non-canonical regions on the plus and minus strand.

#### 3.2.7.3 gRNA conservation analysis

gRNA conservation was compared across data sets with comparable numbers of unique

gRNAs. Read sets which had been clustered so as to only represent unique sequences (section 3.2.4.4(ii), Table 3.7) were merged into one file and re-clustered (usearch -cluster\_fast algorithm (Edgar, 2010)); gRNAs which were >95% identical were deemed to be the same. gRNAs were plotted in a binary heatmap, which was sorted such that the largest clusters (i.e., most conserved gRNAs across the data sets) are at the top, using a custom Python script.

A BLAST approach using the usearch/ublast algorithm was used to quantify gRNAs that are shared across cell lines. gRNAs were considered to be present in both sets if they were >90% identical with an expect value less than  $1 \times 10^{-11}$ . In order to represent the scenario where a gRNA appears to be present but does not satisfy the reciprocal scenario, both blast outputs are presented. For example, a gRNA which is shorter than normal (either due to exonuclease activity, or poor quality sequencing resulting in excessive quality trimming (the chances of which occurring have been mitigated as far as possible)), the shorter RNA will return a hit when being matched to its counterpart in the target database, however when the reciprocal is performed, the full length gRNA will not return a hit, as the gRNA in the target data set is too short. Trying to mitigate this outcome by performing 5' and 3' *in silico* trimming can result in non-specific gRNA matches.



### 3.3 Results

#### 3.3.1 Establishing parameters for gRNA identification by mapping to fully edited mRNAs

As outlined in the methods, our gRNA prediction pipeline was tested for specificity and sensitivity against *T. brucei* EATRO164 WT (positive control) and Ak (negative control) cell lines. These data sets were used together to tune the stringency of the filtering parameters, i.e., maximising coverage of the editing space in the wild type sample whilst minimising false positives in the Ak cell line (Figure 3.1, Table 3.1 and [hank.bio.ed.ac.uk](http://hank.bio.ed.ac.uk) or <http://tinyurl.com/zxbovov> with the appropriate tracks loaded). The percentage of total reads from the Ak cell line passing the alignment filters is 0.0002%, and can be considered to be the false positive rate of the gRNA calling pipeline. Using this method 99% of editing events were covered in the WT cell line compared to just 11% in the Ak cell line. The 11% of editing events covered in the Ak cell line seems high however it does not take into account the low coverage of these editing events when compared to the WT sample. A total of 7786 gRNAs were called for the positive control, compared to just 26 in the negative control, despite the initial number of reads for both data sets being approximately equivalent. This false positive to false negative balance was achieved using a cut-off of 85% identity and a minimum match length of 30 nt (in conjunction with the permutation test previously described) and was deemed to be an acceptable trade-off between sensitivity (i.e., obtaining full coverage of the editing space) and specificity (i.e., minimising false positives). These gRNA calling parameters are used throughout for this study. Other studies in which similar gRNA calling pipelines have been developed have not performed any such testing.

Cell Line	Total number of reads	Number of CSB3 reads	Number of non-genomic reads	Reads matching editing space
<i>T.brucei</i> EATRO 164 Ak	9174699	1078	2771927	26
<i>T.brucei</i> EATRO 164 WT	8088584	23313	2545447	7786

Table 3.1: gRNA calling pipeline specificity testing. Total genomic DNA was sequenced from *T. brucei* EATRO 164, from both a parental (WT) strain and a mutant (Ak) strain that lack kDNA. The Ak strain can be considered to be a negative control as it has no kDNA.

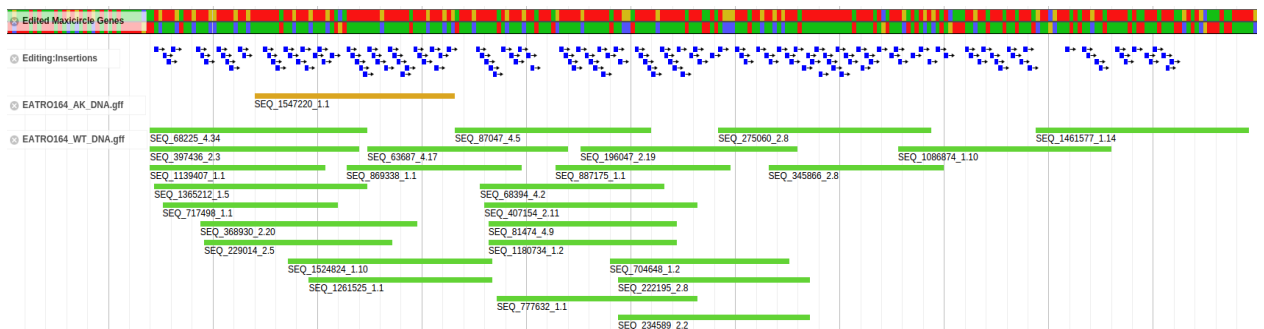


Figure 3.1 *T. brucei* EATRO 164 gRNA coverage predicted from DNA reads plotted using Jbrowse. A web based plot is available at <http://tinyurl.com/zxbovov>. Top to bottom: Nucleotide sequence for the edited transcript of CR3 (300 bp), nucleotides are coloured as follows; Red=T, Green=A, Blue=C, Yellow=G. Editing U insertion events; blue arrows indicate positions at which a U(T) was inserted. 164 AK: The akinetoplastic mutant *T. brucei* cell line showing close to no coverage, bar 1 false positive passing through the gRNA calling pipe (brown bar). 164 WT: full coverage of the editing space with gRNAs predicted from the WT *T. brucei* 164 (green bars).

### 3.3.2 Minicircle assembly

A reference set of minicircles was assembled using the short read datasets presented in table 3.2. The kDNA coverage is much increased for the kDNA purified data sets which did not have Ak genomic DNA added to increase the mass of kDNA being sequenced. Libraries were constructed using the Illumina TruSeq DNA prep kit for samples I-IV, samples V-VIII were processed using the Illumina TruSeq Nano Kit. The percentage of reads with the CSB3 12mer was used as a proxy measure for kDNA coverage.

Sample number	Sample description	Total read count	CSB3 containing reads	% of reads with CSB3
I	AKtotal+L262PkDNA	8121310	171721	2.11
II	Aktotal+WTKDNA	8704684	401686	4.61
III	WT_total	7578986	187645	2.48
IV	L262P_total	7284100	72389	0.99
V	T1_L262P_B	8936800	1854413	20.75
VI	T1_WT_A	12471650	2663097	21.35
VII	T6_WT_A	6358610	1504046	23.65
VIII	T6_WT_B	6605380	1288240	19.50

*Table 3.2: Estimations of kDNA coverage for 300 bp paired MiSeq reads generated for this study. Samples I and II are kDNA purifications with genomic DNA from an akinetoplasmic cell line added due to limitations in minimum mass of DNA required for use with Illumina library preparation kit. AK cell lines are generated by treating L262P transfected cell lines (Dean et al. 2013) with sub-lethal doses of ethidium bromide to ablate kDNA. Samples III and IV are total genomic DNA preparations, generated using phenol-chloroform extraction. Samples V-VIII are isolated kDNA samples, libraries for these samples were prepared using the Illumina TruSeq Nano kit which facilitates library prep using much lower masses of DNA.*

The product of merging two assemblies is presented in Figure 3.2A. In total 511 contigs were assembled, of these 503 were greater than 900 bp long. Figure 3.2B shows the size distribution of the circularised contigs. 473 contigs passed this test and of these, 471 fell within the expected 900-1.1 kb region. The other two contigs that passed this test were a bacterial contaminant (Accession number: CP007219) and a short region of repetitive satellite DNA (Accession: K00392). It should be noted that Figure 3.2A and B show the size distribution of contigs before any removal of duplicate sequences has been performed; removal of duplicate sequences can only occur once the contigs have been circularised; circularisation allows reorientation of contigs with CSB3 at the 5' end and allowed for easy and robust identification of redundant sequences. Figure 3.2C shows the contigs that have passed the circularity test and contain the minicircle conserved 12mer CSB3, 344 sequences in total. The set of contigs in Figure 3.2C have had all duplicate sequences removed and represent a set of fully circularised and unique

minicircles.

In total, we assembled 344 unique (<95% identity), fully circularised minicircles containing the conserved CSB3 12mer. In addition, a further 21 sequences, greater than 700 bp long, were identified that contained the CSB3 sequence, but were not fully assembled and could not be circularised, most probably because they had not been fully assembled. Mapping back reads that contain CSB3 to the 344 minicircles resulted in 93.7% aligning. This analysis is robust, as the reads for this data set are longer than the conserved sequence region (CSR, ~100 bp). If the reads were shorter they could map non-specifically, making assessments of what proportion of minicircles have been assembled less reliable (as occurs in other data sets using shorter reads). Mapping reads containing CSB3 back to the 21 non-circularised sequences resulted in 5.78% of reads aligning. Taking these together, we can conclude that of the minicircles present in this population of cells, we have assembled ~93.7% of minicircles completely and a further ~5.8% partially. This set of 365 minicircles was taken as the reference set used for copy number analysis and as a reference for observation in Section 4. These numbers compare favourably with previous attempts to fully assemble the kDNA from any kinetoplastid (Ochsenreiter et al., 2007b). A plot of the minicircle sequence lengths from the KISS database (Ochsenreiter et al., 2007b) is shown in Figure 3.2D. As this data set is a collection of minicircles from several strains and sources, one would expect it to be a good overall representation of the abundance and variety of minicircle classes, and their expected average size. However the distribution of contig sizes in the KISS data set is more heterogeneous than would be expected from a set of intact *T. brucei* minicircles.

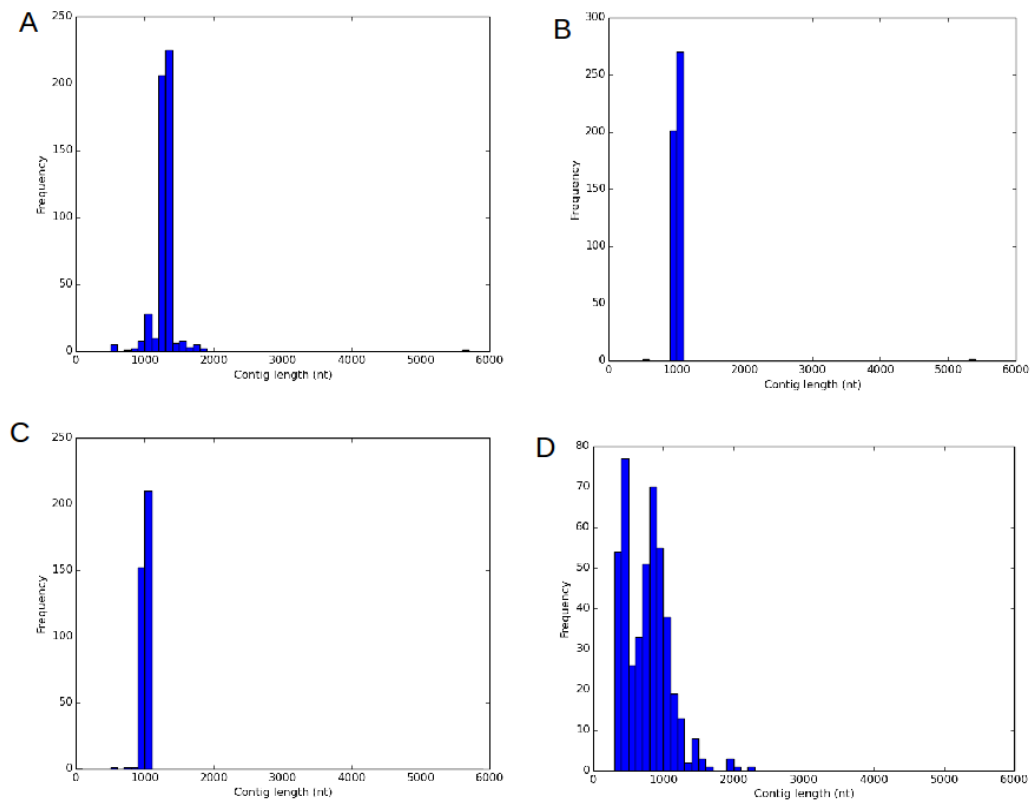


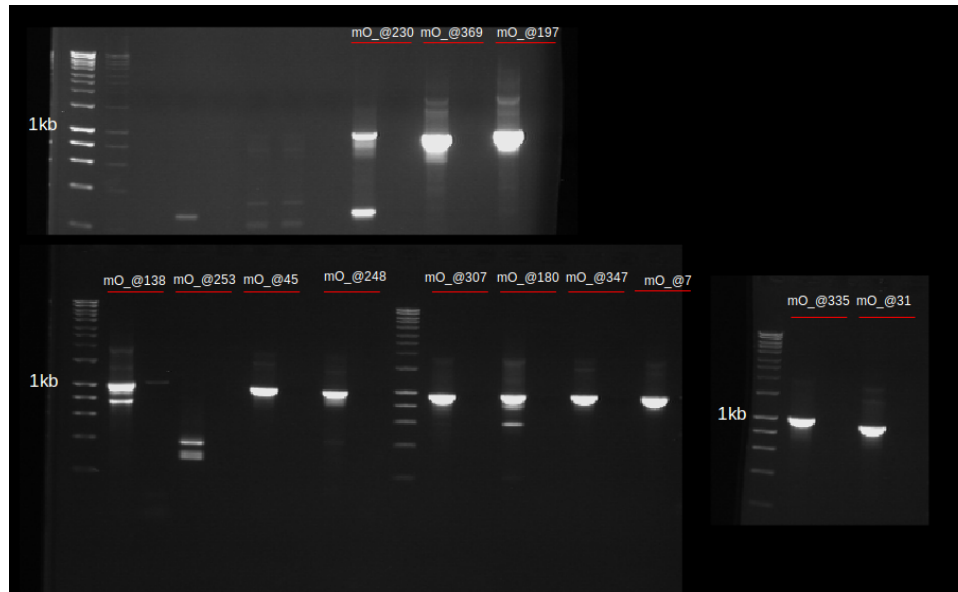
Figure 3.2 A: Assembly generated from 300 bp MiSeq reads. The main peak at 1.2-1.3 kb represents untrimmed minicircle sequences. Due to the method of merging assemblies some of the contigs in this set are duplicates.

B: Contigs from Figure 3.2:A that have passed the circularity test. The 1.2-1.3 kb peak from figure 3.2A is shifted down to ~1kb as the duplicated ends are trimmed during the process of testing for circularity. The single contig seen at approximately 500 bp blasts as *T. brucei* satellite DNA, which is highly repetitive and explains why it passes the circular test. The large contig at ~5300 bp is a bacterial contaminant (Accession CP007219).

C: Complete minicircles assembled from *T. brucei* AnTat 9013 reads. This set shows the circularised contigs that also contain the CSB3 12mer. This set contains 344 contigs in total, a further 21 contigs with a size greater than 500 bp were assembled (not shown) that also contained the CSB3 12mer. The set of minicircles shown in this plot represents ~93% of minicircles in the population of cells based on 93% of reads containing CSB3 mapping back. The mean length for the full set of minicircles (including non-circularised contigs) is 1004 nt, with a minimum length of 593 nt, maximum 1084 nt and standard deviation of 34 nt.

D: Contigs from the KISS database; generated from procyclic stage *T. brucei* TREU667 (also includes previously published minicircles). The spread of contig lengths is much greater than those assembled in this study, whilst the total number is comparable; 455 sequences in total with a mean size of ~760nt, maximum size of ~2200 nt, minimum size of 303 nt and an standard deviation of 317 nt.

A PCR approach was used to validate a random subset of 8 assembled minicircle contigs. The results of the PCR are shown in Figure 3.3, Primers used for validation are listed in table 6.1. All minicircles selected for this treatment were validated.



*Figure 3.3: Antat9013 minicircles were PCR amplified (band at 1 kb), bands were either excised from the gel and sequenced or PCR products directly Sanger sequenced. sequences were aligned to the original assembly for each minicircle. The minicircle ID is given by mO\_@#. Each pair of lanes shows the amplified minicircle and a negative control whereby water was used rather than template DNA.*

### 3.3.3 Maxicircle and minicircle copy numbers

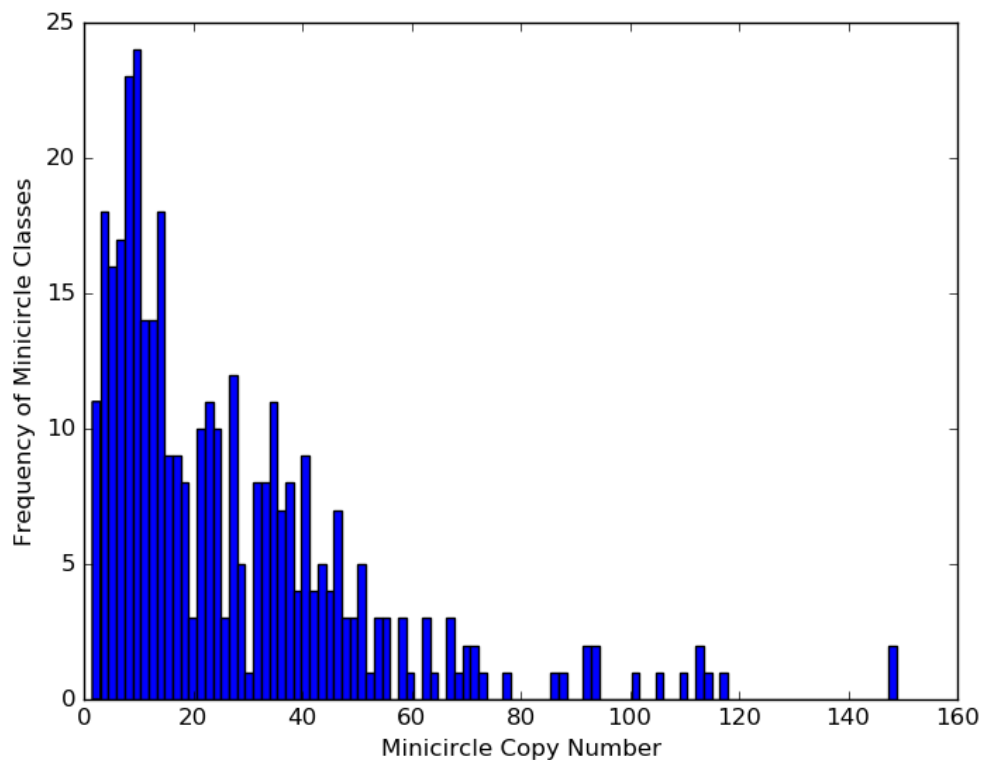
For the wild type sample prepared from total DNA (WT\_total, Table 3.2 ), the coverage of the genomic DNA can be used as a standard to estimate average maxicircle copy number in the population:

$$\text{Maxicircle copy number}_{\text{WT}} = \frac{\text{Maxicircle depth}_{\text{WT}}}{\text{Haploid genome depth}_{\text{WT}}}$$

If we assume that maxicircle copy number in kDNA enriched samples (where nuclear genome depth cannot be used as a standard) is the same as in the WT then we can estimate population average minicircle copy numbers in kDNA enriched samples:

$$\text{Population Average minicircle copy number} = \text{Maxicircle copy number}_{\text{WT}} \times \frac{\text{Minicircle depth}}{\text{Maxicircle depth}}$$

After mapping to the core chromosomal regions (excluding any highly repetitive telomeric regions), the average per nucleotide coverage for the haploid genome was calculated to be 23X. This was then used to approximately calculate the number of maxicircles present. Given that the per nucleotide coverage for the maxicircle was 700X we estimate that there were, on average, approximately 30 maxicircles per cell in the WT. The maxicircle copy number is then used to calculate copy numbers for individual minicircles and estimate the average number of minicircles per cell in this population. The minicircle class copy number distribution is shown in Figure 3.4. Mean copy number is 26X, but it is highly left skewed as predicted by theory (Savill and Higgs, 1999) with 95% of minicircles having an average copy number of between 1.5 and 71.2. The average frequency of minicircles per cell as calculated from the minicircles in Figure 3.4 is 9695, consistent with estimations of 5000-10,000 minicircles per *T. brucei* kinetoplast (Steinert and Van Assel, 1980). For samples where no standard is available (i.e. when analysing kDNA purified samples), the average total number of minicircles per cell is assumed to remain constant at 9695 minicircles per cell for a WT sample. Read coverage per minicircle is used to calculate what proportion of total kDNA reads cover a given minicircle; this can be used to estimate individual minicircle copy number in the absence of the genomic DNA to normalise against.



*Figure 3.4: Minicircle copy number distribution for WTtotal. The average number of maxicircles per cell was calculated to be 30 using the coverage of the nuclear genome as a standard. This was then used to calculate the copy number for each class of minicircle. Average number of minicircles per cell based on these calculations is 9695.*

### 3.3.4 Annotation

For all assembled minicircles, annotation of conserved features and gRNAs was made based on motif elucidation (based on kmer counting and using the MEME algorithm (Bailey et al., 2009)), literature (Ray, 1989) and gRNA prediction based on alignment or nucleotide bias.

#### 3.3.4.1 CSB and inverted repeat identification

Of the 344 sequences that had passed the circularisation test seven did not have an exact copy of the previously published CSB3 sequence, and instead had one of two related sequences which differed by one base (CSB3\_A8 and CSB\_A1 in Table 3.3 and



Figure 3.5). Meme (Bailey et al., 2009) motifs shown in Figure 3.5 were generated either from the set of unique CSBs (top row of blocks) in order to highlight sequence diversity or from the full set of CSBs extracted from minicircles (bottom row of blocks) thus highlighting the dominant sequences. CSB1 is made up of roughly equal proportions of two sequence variants as both the Memes are identical. The sequence diversity for CSB2 is much greater and there are many minor variants present in low copy number. The distances between CSB1,CSB2 and CSB3 motifs on minicircles are highly conserved (Figure 3.6); with a CSB1 to CSB2 distance of 19 nucleotides (SD=0.08 nt) and a CSB2 to CSB3 distance of 33 nt (SD=0.06 nt). The high degree of conservation in the CSR, especially the distances between the CSBs, is presumably important for the binding of the minicircle replication machinery.

Frequency	Motif Name	Motif
346	CSB1	[A G][G A T]GGGCGT[T G]C
271	CSB2	[T C]C[C A]CGT[T G A]C
341	CSB3	GGGGTTGGTGTGA
6	CSB3_A8	GGGGTTGATGTA
1	CSB3_A1	AGGGTTGGTGTGA
1226	Inverted_repeat_fwd	TAATA[G A]ATA
1117	Inverted_repeat_rev	TAT[T C]TATTA

Table 3.3: Breakdown of Conserved Sequence Blocks (CSBs) in 344 circularised minicircles. Some minicircles have more than one copy of CSB3.

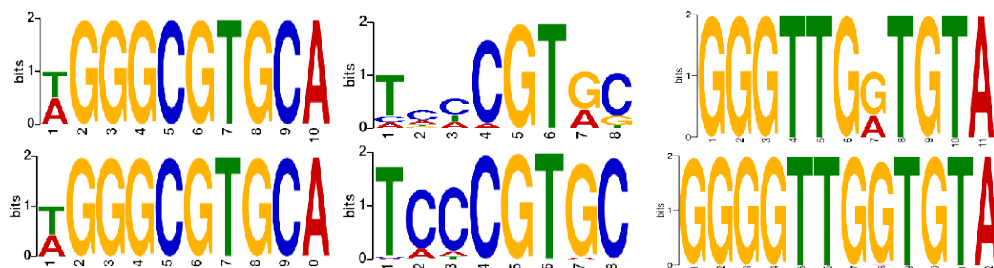


Figure 3.5: Meme generated motifs from either collapsed sequences (top; highlighting sequence diversity) or from the total set of CSBs (bottom; highlighting the dominant sequence). From left to right CSB1, CSB2, CSB3.

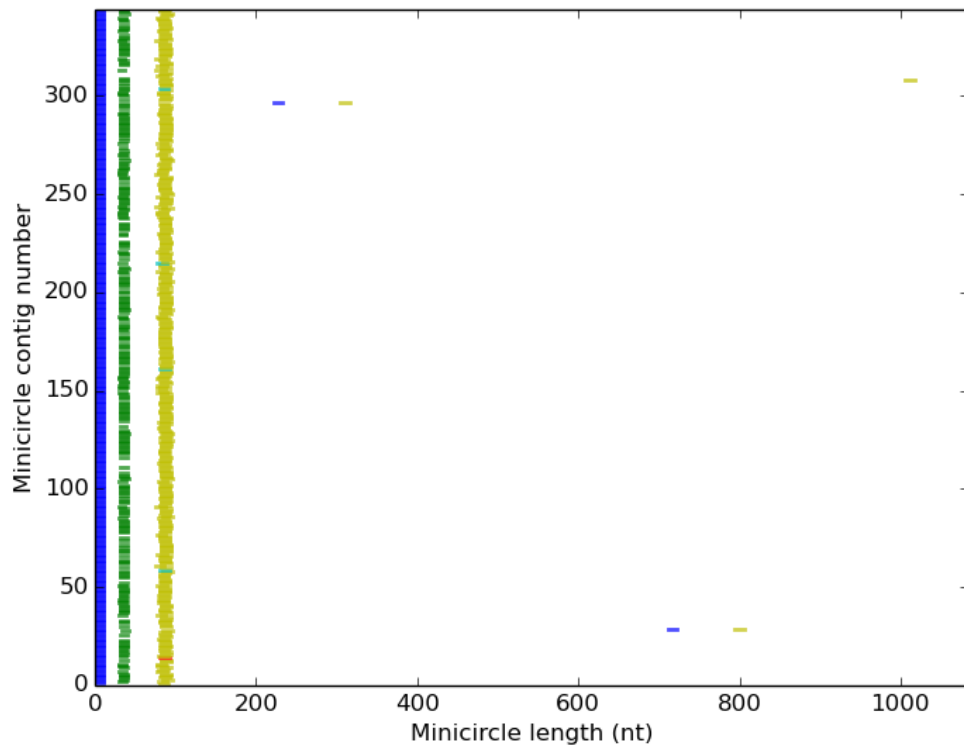


Figure 3.6: Circularised minicircles aligned with the 5' end of CSB1 set as the zero coordinate. Blue=CSB1, Green=CSB2, Yellow=CSB3, Cyan=CSB3\_A8, Red=CSB3\_A1.

Inverted repeats, thought to be important for gRNA expression (Pollard et al., 1990), were annotated using a regular expression pattern match/search approach developed based on literature (Hong and Simpson, 2003) (Table 3.3). The repeats were located based on sequences rather than looking for a forward repeat sequence and looking for a reverse counterpart. Of the 1172 forward, and 1046 reverse repeats identified, 1042 of them make up cassettes, i.e. a forward repeat followed by a reverse repeat. This equates to an average of 3.2 cassettes per minicircle. The mean distance between the forward and reverse repeat is 104 nt (Figure 3.7) (SD=10 nt) consistent with previously published observations (Pollard et al., 1990).

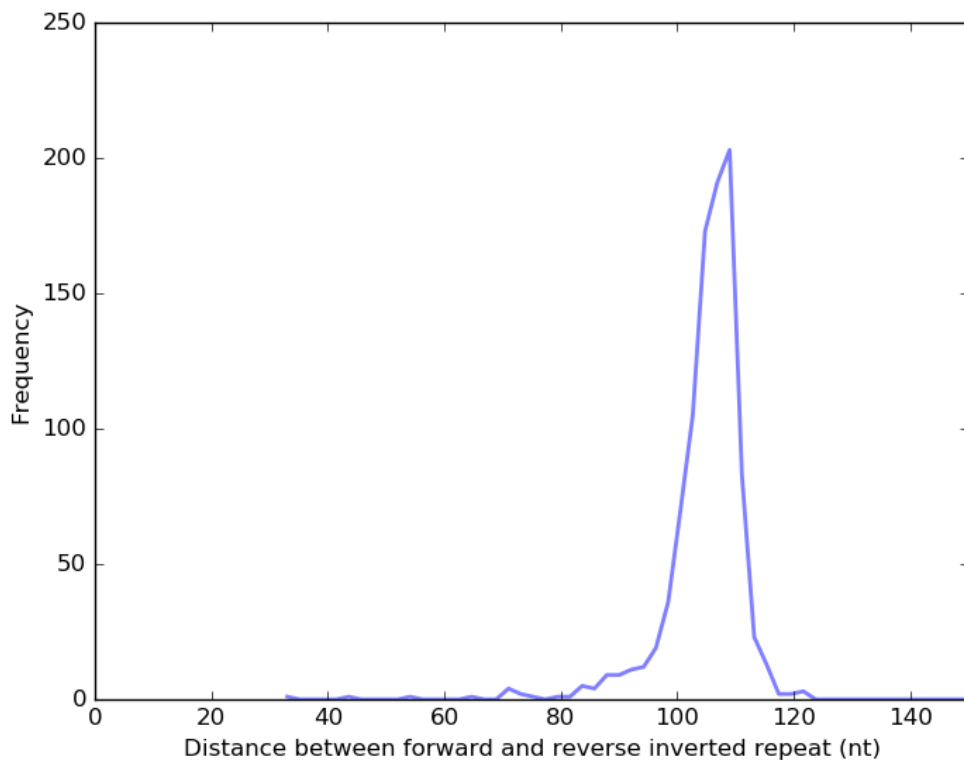


Figure 3.7: The distance between the forward and reverse inverted repeats plotted giving a distribution of cassette sizes. The mean distance is 104nt with an SD of 10.4 and a range of 32-243 nt.

### 3.3.4.2 Annotating canonical gRNA genes on minicircles

Assembled minicircles for AnTat90.13 were passed through the gRNA prediction by alignment pipeline (section 3.2.4.4.i), resulting in the prediction of 608 gRNA genes, covering 97% of the known editing space (see <http://tinyurl.com/z6dgy8d>, assembled minicircles can also be accessed here by clicking on links within the alignment description box). A breakdown of gRNAs identified for each edited gene is given in Table 3.4 (along with coverage generated from small RNA data sets presented in section 3.3.5). gRNAs predicted directly from reads cover 99.2% of all editing events, gRNA-edited mRNA mapping can be viewed in a modified JBrowse instance (Westesson et al., 2013) at <http://tinyurl.com/jrvfgkb>. In comparison, passing published *T. brucei* minicircles from the KISS database (Ochsenreiter et al., 2007b) through my

gRNA identification pipeline resulted in only 70% of the editing space being covered (see <http://tinyurl.com/jcz2ypu> for gRNAs predicted from the KISS database).

Edited sequence	Insertions/deletions	Number of gRNAs called from minicircles (Insertions covered/percentage of insertions covered)	gRNAs from DNA reads (Insertions covered/percentage of insertions covered)	gRNAs from BSF small RNA reads (Insertions covered/percentage of insertions covered)	gRNAs from PCF small RNA reads (Insertions covered/percentage of insertions covered)
a6	447/28	90 (428/96)	959 (447/100)	1482 (447/100)	616 (447/100)
co2	4/0	1 (0/0)	174 (4/100)	18 (4/100)	8 (4/100)
co3	547/41	151 (548/100)	1339 (548/100)	2206 (548/100)	958 (548/100)
cr3	148/13	24 (141/95)	285 (138/93)	608 (145/97)	218 (142/95)
cr4	148/13	57 (325/100)	944 (325/100)	1606 (325/100)	824 (325/100)
cyb	34/0	1 (31/91)	101 (34/100)	46 (34/100)	19 (34/100)
murf2	26/4	4 (0/0)	180 (26/100)	34 (26/100)	24 (26/100)
nd3	210/13	19 (178/89)	310 (200/100)	612 (200/100)	276 (200/100)
nd7	553/89	119 (547/99)	1094 (551/100)	2069 (552/100)	840 (552/100)
nd8	259/46	59 (243/94)	654 (259/100)	975 (259/100)	433 (259/100)
nd9	345/20	47 (328/96)	656 (334/98)	1308 (342/100)	624 (342/100)
rps12	132/28	34 (126/95)	231 (132/100)	713 (132/100)	287 (132/100)

*Table 3.4: Breakdown of gRNAs called from assembled minicircles, directly from DNA reads, small RNA reads (BSF and PCF). The numbers that are shown in this table are from the unique gRNAs called from each data set (defined by clustering). In parenthesis the number of editing events covered and the percentage of U-insertion editing events covered is shown for each data set and transcript (U-insertions covered/percentage of U-insertions covered).*

Despite the majority of editing events being covered by a gRNA predicted from an assembled minicircle there are still some gaps. Cyb for example has three U insertion events which are not covered by a gRNA. This suggests that a minicircle, which encodes this missing gRNA, has not been assembled. The missing gRNA is represented in the gRNAs called-from-reads set”. <http://tinyurl.com/gtutrjl> shows the two data sets side by side for the Cyb transcript. [mO @130 83](#) is the gRNA “called-from-assembled” minicircles, there are several sequences in the “called-from-reads” set which are similar. They are likely to represent polymorphisms in the minicircle. There is also one additional sequence (WT868525\_1.9) which covers the additional U insertions. In total 608 gRNA sequences have been predicted from the 365 minicircle sequences. Of these 608 gRNAs, 457 are in a cassette i.e., flanked by both forward and reverse inverted repeat. The other 151 gRNAs are found in regions where either a

forward repeat has been identified with no opposing reverse repeat, or vice versa. There are also 602 cassettes with no predicted gRNA present. It is notable that the cr3 edited sequence has gaps in the coverage for all of the data sets presented in Table 3.4, suggesting that AnTat90.13 cells may have lost editing capacity for this transcript. The function of cr3 is not known although this indicates a full length cr3 protein is not required for survival.

The nature of the predicted duplexes was investigated. It was found that the 5' end of the gRNA-mRNA duplexes have a greater proportion of Watson-Crick base pairs than the 3' end (Figure 6.1). This is similar to what has been reported by Koslowsky et al. (2014) and Kirby et al. (2016).

#### 3.3.4.3 Prediction of non-canonical gRNAs by nucleotide bias

Alignment of the 608 predicted gRNAs 5' to 3' (section 3.3.4.2) (with their upstream and downstream flanking 80 bp), revealed a nucleotide bias associated with gRNA cassettes (Figure 3.8, Figure 3.9). It was reasoned that this bias could be used to predict gRNA genes that could not be identified by alignment with the known edited sequence. “Upstream” and “downstream” scoring vectors based on gRNA cassette nucleotide bias were constructed (see Methods 3.2.4.4 iii). Running the scoring vectors over the length of each minicircle sequence accurately identified start and stop positions of canonical gRNAs predicted by alignment and, more importantly, putative gRNA-like genes (which we term non-canonical gRNA genes) in 'empty cassettes' (Figure 3.10, Figure 3.11). Using this method we detected 1039 regions matching the expected nucleotide bias for a gRNA. Of these 481 overlapped with canonical gRNA predictions (Table 3.5) leaving 558 non-canonical gRNA regions that do not have a match to the known editing space. Of the gRNAs predicted by alignment, 112 did not meet the stringent cut-off chosen for nucleotide-biased based prediction.

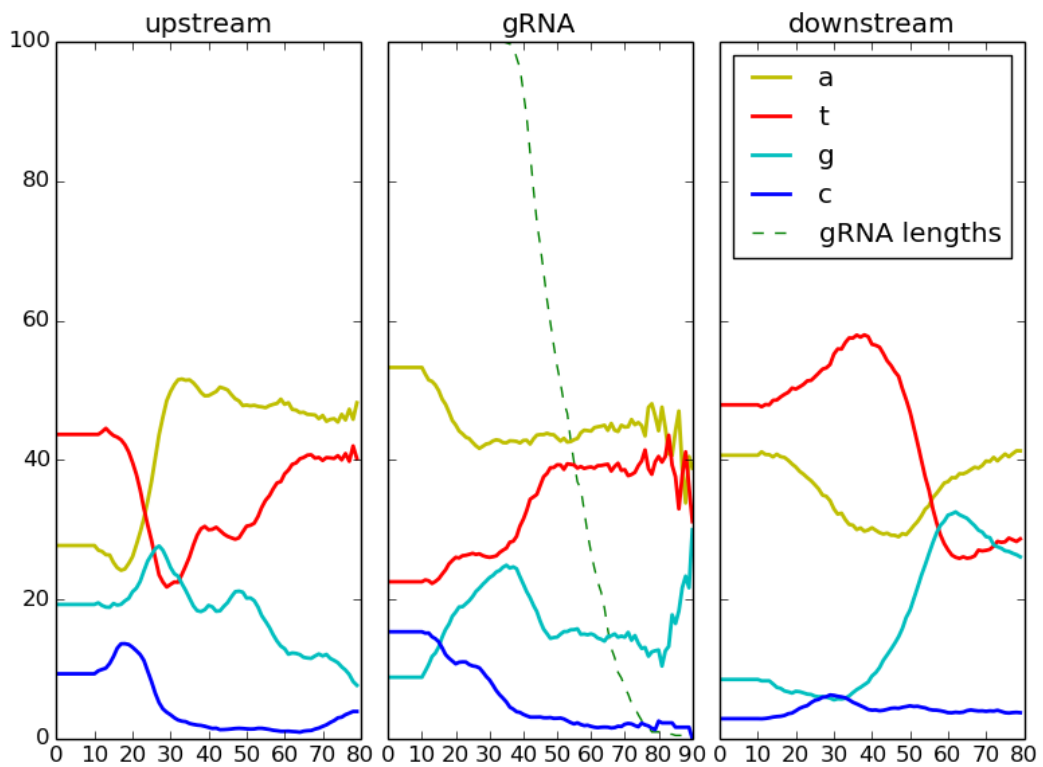


Figure 3.8: gRNAs predicted by alignment of minicircles to the known edited space show a strong nucleotide bias. gRNAs annotated on minicircles were aligned 5'-3' as well as 80bp upstream and downstream. gRNAs are A rich upstream and T rich downstream. gRNAs themselves are more A rich at the 5' end.

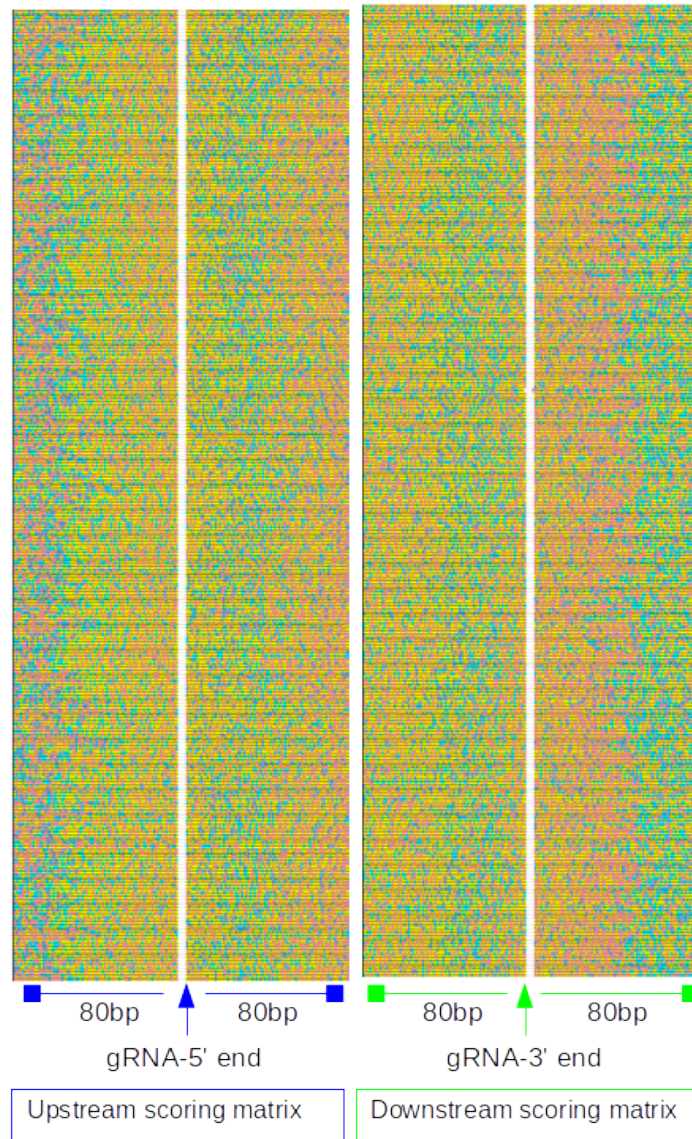


Figure 3.9: *gRNA nucleotide bias.* Coordinates generated from the Smith-Waterman approach were used to extract gRNAs as well as 80 bp upstream and downstream. The upstream-gRNA-downstream sequences were plotted in the Belvu alignment viewer; the colours are as follows, T:red,A:yellow,G:green,C:blue. The 5' end of the predicted gRNA-mRNA duplex are marked plus 80 bp upstream and downstream. These gRNAs plus the flanking regions were aligned and nucleotide frequencies calculated for each position; the “cassette” was broken in two giving upstream, gRNA start site probabilities and downstream, gRNA stop site probabilities. This method can be used to corroborate previously predicted gRNAs as well as highlight regions of minicircles that may contain gRNAs that do not match any of the genes in the known editing space.

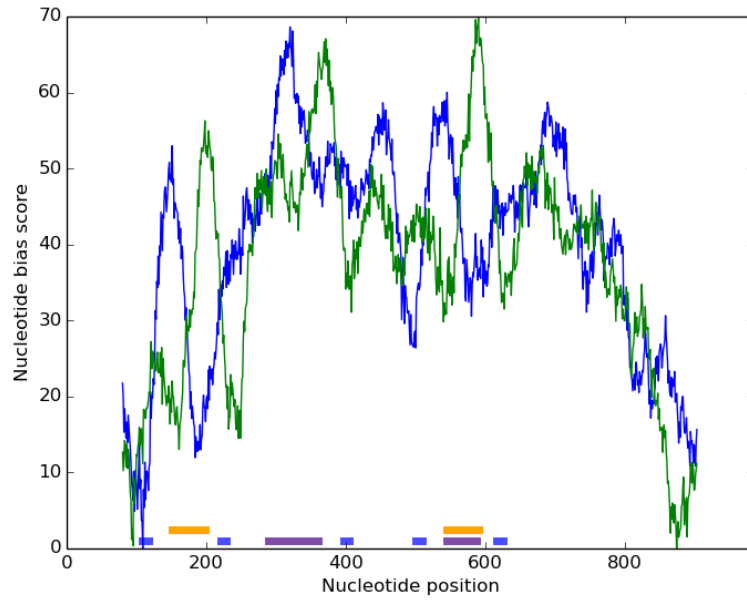


Figure 3.10: Nucleotide bias scoring matrix for minicircle [mO @32](#). Blue line: Computed scores for the upstream scoring matrix, Green line: computed scored for the downstream scoring matrix. Orange bar gRNA region predicted by nucleotide bias, Purple bar: ND7 gRNA predicted by alignment to the editing space.



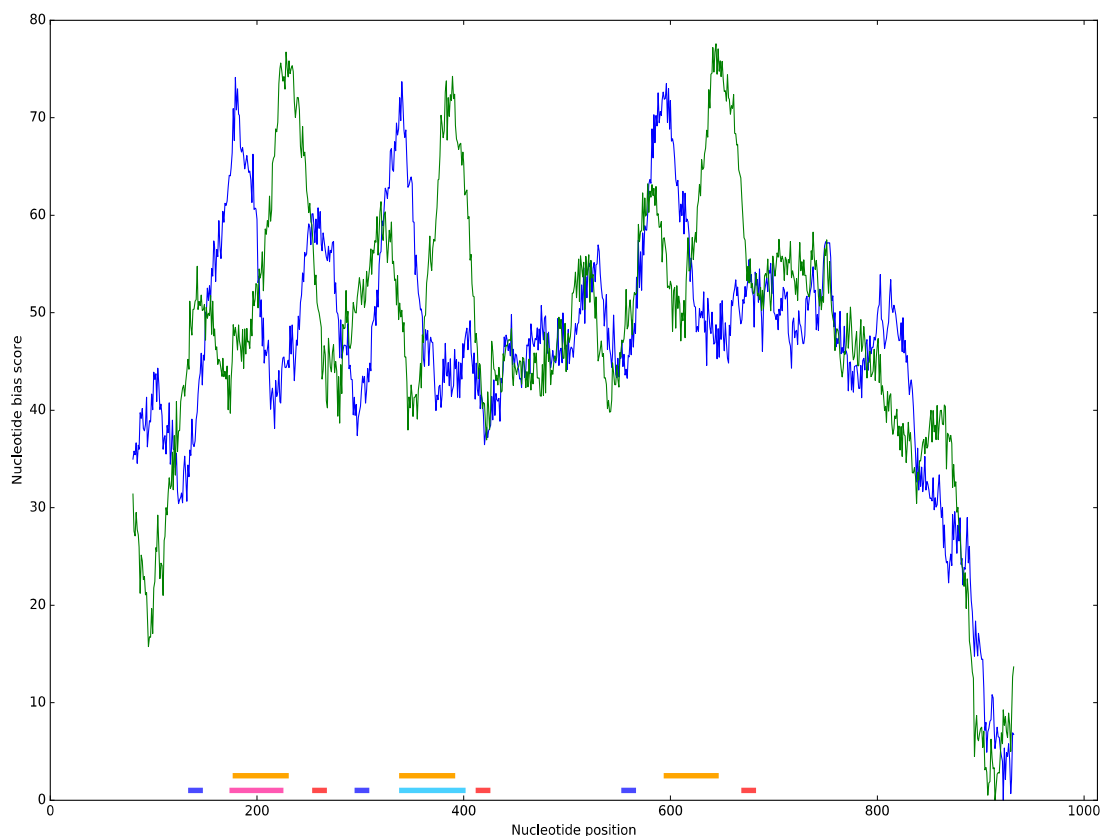
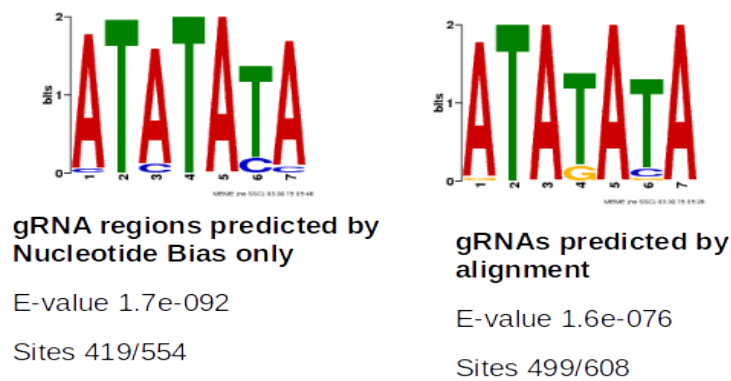


Figure 3.11: Nucleotide bias vector scores for representative minicircle (*mO @341*). Blue line; upstream scoring matrix, green line; downstream scoring matrix, orange bars; nucleotide bias start and stop predictions, pink bar; A6 gRNA, Light blue bar; co3 gRNA, short dark blue bars; inverted repeats.

Total gRNAs predicted from minicircles <b>1151</b>	Have the expected nucleotide composition <b>1039</b>	Have a match in the edited space <b>481</b>
		Do not have a match in the edited space <b>558</b>
	Do not have the expected nucleotide composition <b>112</b>	Predicted by alignment to the edited space only <b>112</b>

Table 3.5: gRNAs annotated on assembled Antat9013 minicircles either by alignment to edited maxicircle genes or nucleotide bias or both.

Canonical gRNAs have been shown to have an “RYAYA” motif at their 5' end (Madej et al., 2008b). Inspection of our canonical and non-canonical gRNA sets revealed a similar “ATATATA” motif for both sets (Figure 3.12). This motif is usually directly upstream of the region where the gRNA is predicted to align to the mRNA; 22% of predicted gRNAs have a 2bp overlap with the “ATATATA” motif.



*Figure 3.12: 5' regions of canonical and non-canonical gRNAs show a conserved ATATATA motif.*

### 3.3.5 Small RNA analysis

A breakdown of small RNA read mapping for the two pairs of mitochondrially enriched small RNA read sets generated in this study is shown in Table 3.6. The nuclear RNA contamination is significant, however upon comparison with other studies which have used a similar crude mitochondrial isolation method (Simpson et al., 2015), the genomic contamination levels are also quite variable, with 7-40% of small RNA reads

mapping to the nuclear genome. The output statistics from the gRNA calling pipeline (section 3.2.4.4(ii)) for the TREU677 and AnTat9013 data sets are shown in table 3.7 (as well as previously published small RNA data sets used for comparison). Full coverage of all editing events for all the data sets in Table 3.6 were predicted by alignment to the editing space. Interactive gRNA-mRNA duplexes are available at <http://hank.bio.ed.ac.uk>, and gRNA depth plots for each of the edited genes can be found in Appendix section 6.3.

The variability in the number of unique gRNAs called for each of the data sets is vast (Table 3.7) and the number of gRNAs is far more than is expected based on annotating gRNAs on minicircles. The source of this variability is hard to ascertain, it is possibly due to: true variability within minicircle classes, sequence variation caused by poor sequence quality of small RNAs or 3' or 5' heterogeneity of small RNAs causing poor collapsing.

Sample	Total reads	R1 with adapter	R2 with adapter	Too short (15<)	Passed Trim	Merged	Mapped to nuclear genome	Non genomic mapped to pre-edited maxicircle	Non genomic mapped to minicircle
AnTat90.1 3 BSF	79132567	78943722	77886440	25709924	53422643	52269312	52.74%	0.51%	13.65%
AnTat90.1 3 PCF	85827714	85735049	85317022	44942697	40885017	40541705	57.64%	0.63%	3.30%
TREU667 BSF	37497141	26911323	23846662	not filtered for size at this stage	22577393	22119548	63.63%	2.25%	3.21%
TREU667 PCF	33342736	17172629	15562411		14209601	13867894	54.83%	5.63%	10.96%

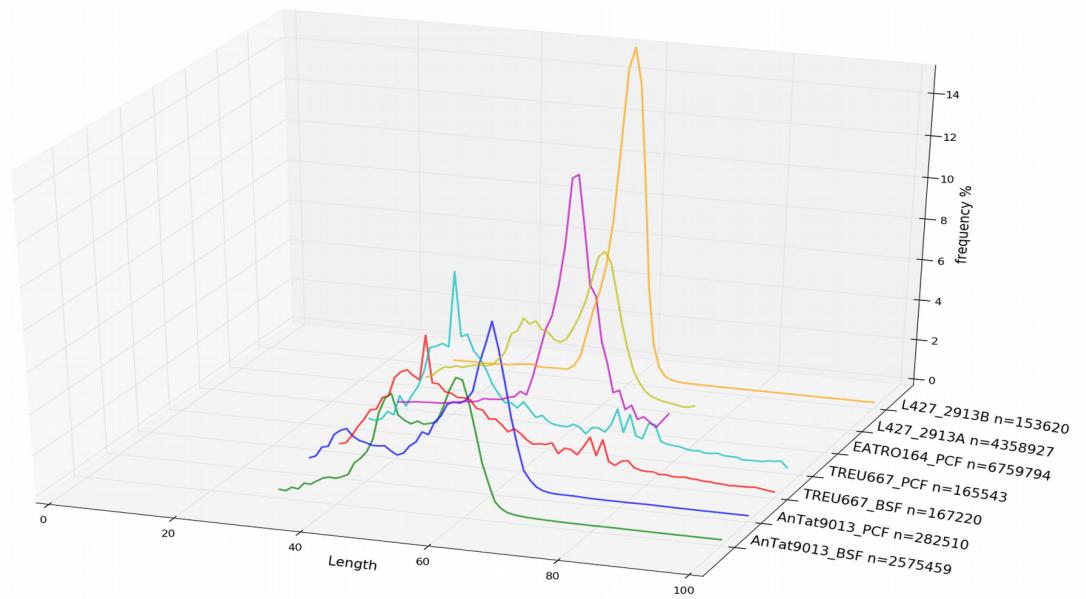
**Table 3.7: gRNA calling statistics for small RNA data sets used in this study.** Read sets where the Accession is provided were downloaded from the NCBI Short Read Archive. Methods for gRNA calling are outlined in section 3.2.4.4(ii). Briefly, paired reads are merged using PEAR, identical reads collapsed, reads that map to the pre-edited maxicircle genes discarded, reads that map to the edited maxicircle are then de-collapsed to give a true number of reads that map to the editing space.

Sample	Accession	Source	Merged	Collapsed (identical seqs)	Mapped to pre-edited	Match edited	Match edited (decollapsed)
AnTat90.13 BSF		This study	52,269,312	2,017,491	459,201	229,610	2,585,180
AnTat90.13 PCF		This study	40,541,705	1,323,516	296,786	49,804	284,720
TREU667 BSF		This study	22,119,548	2,940,554	996,790	21,455	61,058
TREU667 PCF		This study	13,867,894	2,048,572	537,771	21,969	75,405
EATRO164 PCF	SRR1041339	Koslowsky et al. 2014	Single end	3,401,211	423,503	348,195	984,299
L427_29.13A PCF	SRR2125761	Suematsu et al. 2016	Single end	7,619,099	2,449,687	845,437	3,956,121
L427_29.13B PCF	SAMN02795843	Madina et al. 2014	Single end	355,679	23,395	31,349	153,829

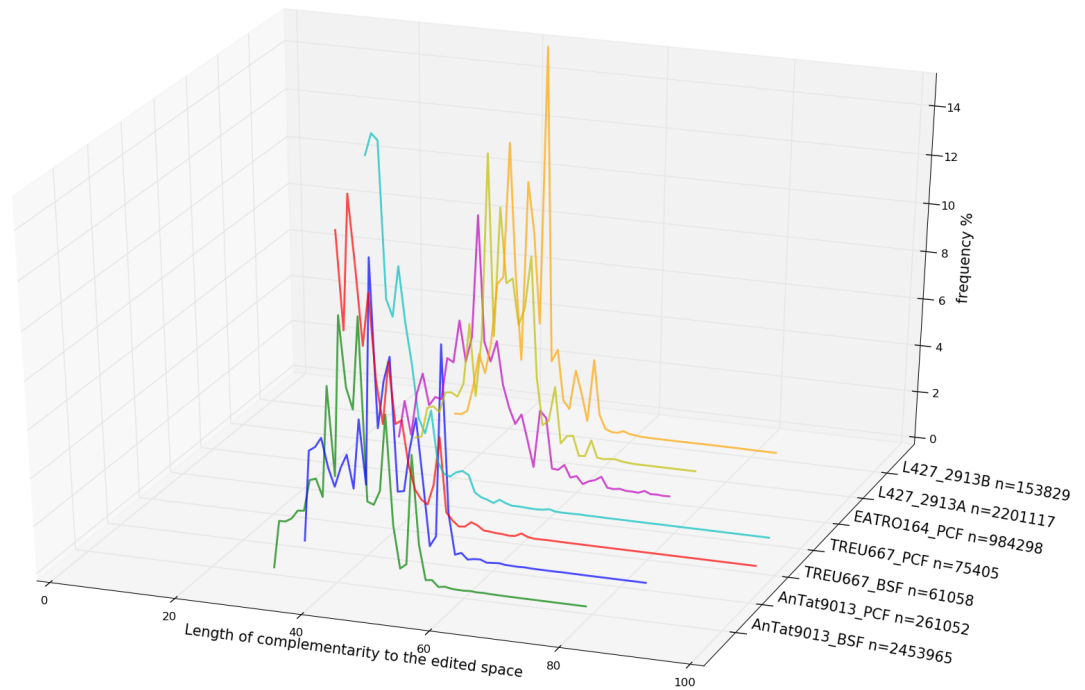
**Table 3.7: gRNA calling statistics for small RNA data sets used in this study.** Read sets where the Accession is provided were downloaded from the NCBI Short Read Archive. Methods for gRNA calling are outlined in section 3.2.4.4(ii). Briefly, paired reads are merged using PEAR, identical reads collapsed, reads that map to the pre-edited maxicircle genes discarded, reads that map to the edited maxicircle are then de-collapsed to give a true number of reads that map to the editing space.

In order to assess the fidelity of library construction and to investigate potential differences between gRNA data sets (both biological and experimental in origin) the lengths of the small RNAs which have a match in the editing space and the length of their complementarity to the edited mRNA were plotted. Lengths of adapter trimmed reads that have a match in the editing space for each of the samples from Table 3.7 are shown in Figure 3.13. The two sets of AnTat90.13 small RNA reads prepared for this study that have a match in the editing space show a generally similar size distribution, however there is a modest increase in the abundance of shorter gRNAs in the PCF set; this is probably due to experimental error rather than being biologically relevant. It was seen in bioanalyser traces for the cDNA generated from the Antat9013 PCF set that there were many short sequences likely to be adapter-adapter constructs. This has the overall effect of reducing the useful data generated from this library. The size distribution of the AnTat90.13 data sets and the 2913 PCF data sets are very similar (barring L427\_2913B, probably due to more stringent size selection from the denaturing gel), showing a major peak at ~60 bp as well as a minor peak at ~45 bp. The TREU667 BSF and PCF sets are significantly shorter overall, with a major peak at 45 nucleotides, this is probably due to the experimental method used to enrich for gRNAs in this set. Self priming may have resulted in the 3' trimming of many gRNAs.

To further quality control the small RNA samples the length distribution of complementarity to the edited mRNAs is shown in Figure 3.14, similar to the lengths of the reads which have a match in the editing space (Figure 3.13) the two TREU667 sets are shorter overall.



*Figure 3.13:* Length distribution for small RNAs which have a match in the editing space (sense only). Frequency is expressed as a percentage of total called gRNAs for each set. EATRO164 PCF, L427\_29.13A and L427\_29.13B are data downloaded from (Koslowsky et al. (2014), SRR1041339), (Suematsu et al. (2016), SRR2125761) and (Madina et al. (2014), SRR1293819) respectively.



*Figure 3.14* Length of complementarity to the edited space for small RNAs. Frequency is expressed as a percentage of total called gRNAs for each set. EATRO164 PCF, L427\_29.13A and L427\_29.13B are data downloaded from (Koslowsky et al. (2014), SRR1041339), (Suematsu et al. (2016), SRR2125761) and (Madina et al. (2014), SRR1293819) respectively.

### 3.3.5.1 Mapping small RNAs to assembled and annotated minicircles

Mapping of the Illumina sequenced small RNAs from AnTat90.13 cells to our set of assembled minicircles shows that both the canonical and non-canonical sets of predicted gRNAs are covered by stacks of reads in both the sense and anti-sense orientation. Figure 3.15 shows a representative minicircle (minicircle mO\_@32 as shown in Figure 3.10) . This minicircle has an example of each type of annotation and read mapping combination that are possible (as outlined in Table 3.5);

- 1) gRNA predicted by nucleotide bias only,
- 2) gRNA predicted by alignment to the editing space only,
- 3) gRNA predicted by both nucleotide bias and alignment to the edited space.

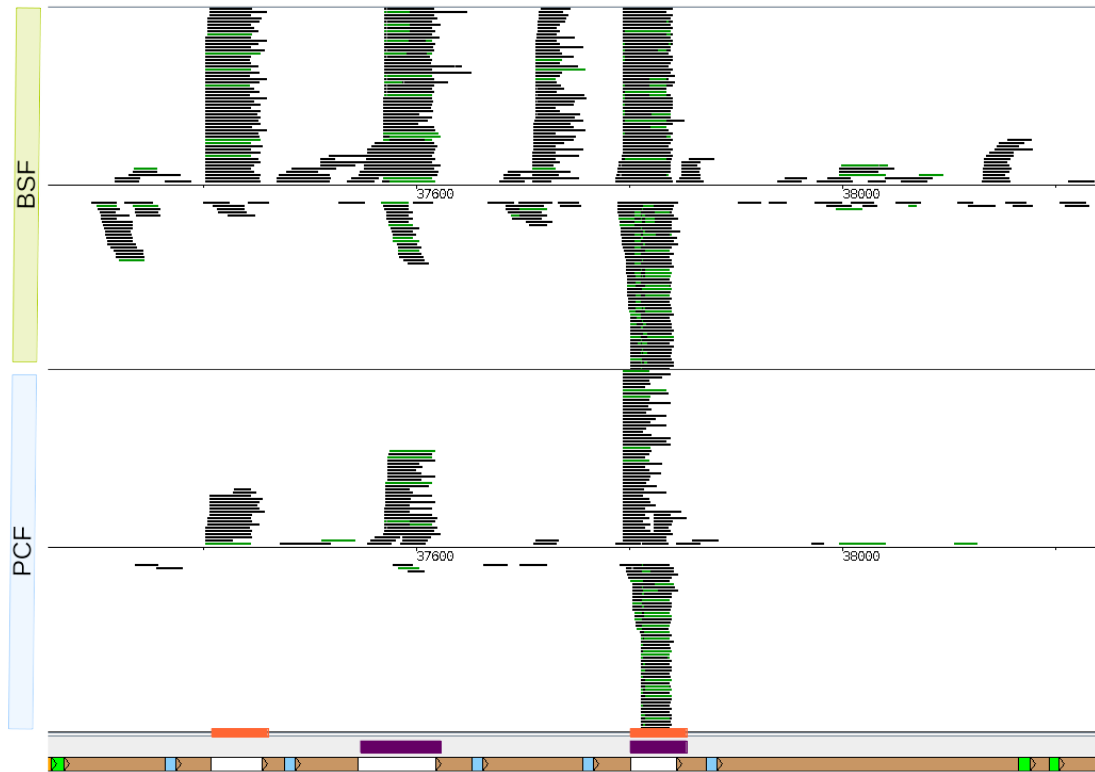


Figure 3.15: Direct mapping of non-nuclear, adapter and U-tail trimmed small RNAs to assembled minicircles (visualised in bamview (Carver et al., 2013)). Shown is representative minicircle ([mO @32](#), length=983bp) as it contains multiple examples of gRNA annotation methods as well as a stack of gRNAs which map to an unannotated region. The top panel shows the sense and anti-sense reads for the AnTat90.13 BSF and the bottom panel shows the AnTat90.13 PCF sample. Reads are coloured either black; indicating unique start and end positions or green indicating multiple reads with the same start and end positions. Purple Bars: ND7gRNAs predicted by alignment, Orange bars: predicted by nucleotide bias, green boxes from left to right, CSB3, CSB1, CSB2. Note the un-annotated stack of reads (~450bp).

Note the stack of reads mapping to a region (~450-500 bp) where neither annotation method predicted a gRNA gene. Closer inspection of the reads that map to this unannotated region shows that they have U-tails and therefore appear to represent gRNA-like transcripts. Cases in which stacks of reads with U-tails map to unannotated regions on minicircles are rare (<5 cases total). There are 63 cases in which annotated AnTat9013 gRNAs were not confirmed by transcriptome analysis from either

corresponding data set, i.e. mapping the PCF or BSF small RNAs for these 63 predicted gRNA genes does not generate a read depth greater than 5 across the length of the predicted gRNA gene. In 112 instances (Table 3.5), the alignment method correctly predicted a gRNA region where the nucleotide bias was below cut-off; closer inspection of 10 randomly selected cases showed noticeable nucleotide bias, which was below the the stringent cut-off chosen for this method. In Figure 3.15 the purple bar (~240 nt) represents a gRNA predicted by alignment (maps to ND7) but not by nucleotide bias (orange bar), although there are still strong peaks at the start and stop positions of the gRNA (Figure 3.10). Similarly the transcript produced from the ~450-500 bp region is marked by noticeable nucleotide bias (Figure 3.10), at least for the start position, however due to the close proximity to the neighbouring 3' gRNA the nucleotide bias peaks were probably lost as noise. The lack of detectable flanking inverted repeats also makes annotation difficult for the rare cases in which this occurs. Thus the majority of stacks of reads that map to the assembled set of minicircles showed noticeable nucleotide bias perhaps implying that nucleotide composition is important for the initiation of gRNA transcription and/or processing. It is conceivable that the 5' flanking nucleotide composition functions as a gRNA promoter, allowing binding of the transcription machinery, or as a recognition signal for the gRNA processing machinery (Suematsu et al. 2016).

Quantification of the number of predicted cassettes that have an average small RNA read depth greater than five is shown in Table 3.8. A further breakdown of strand specific gRNA coverage is shown in Table 3.9, note lack of coverage on the minus strand for the TREU667 pair; this is due to the method in which the gRNAs were purified. The TREU667 pair of samples were enriched for gRNAs by using an oligo-A bead purification approach, as anti-sense gRNAs have much shorter U-tails than their



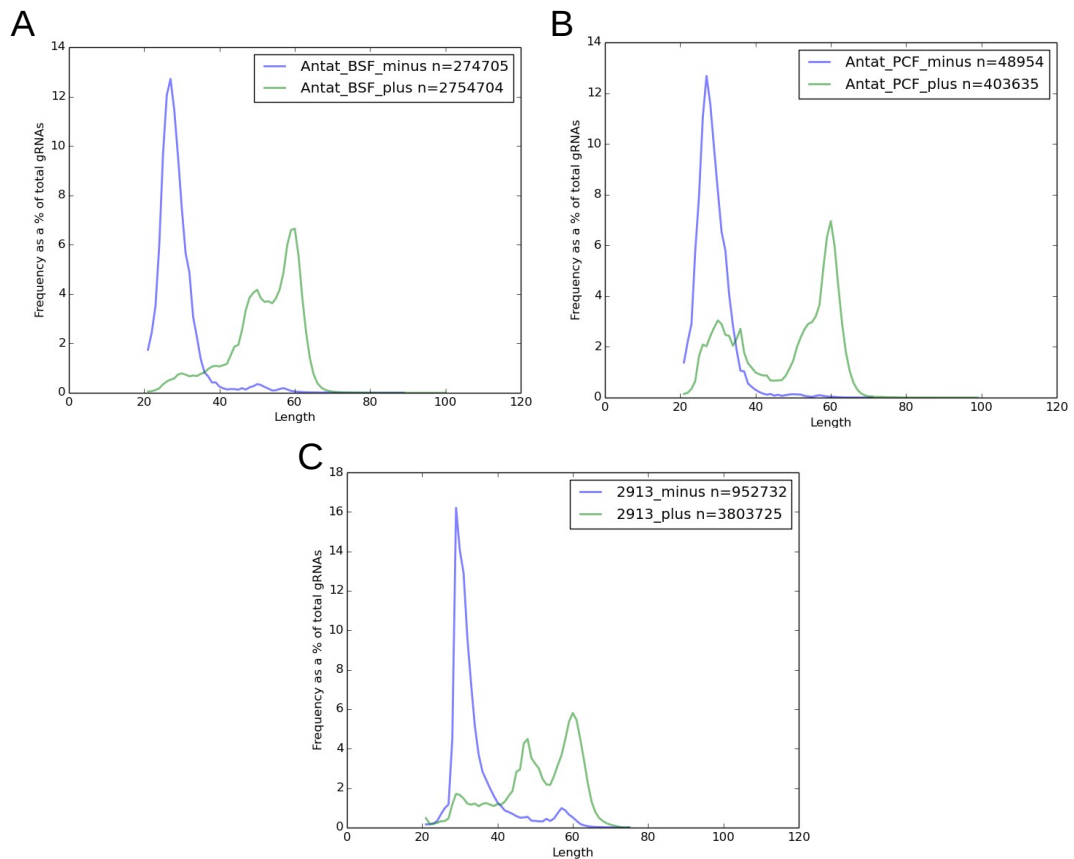
sense counterparts (see Figure 3.16) they will not be purified efficiently using this approach. There is also a lack of minus strand coverage for the EATRO164 PCF and L427\_29.13B reads, again this is due to the very stringent size selection methods used to isolate gRNAs in (Koslowsky et al., 2014) and (Madina et al., 2014). Figure 3.16 shows the size distribution of sense and anti-sense (U-tail intact) transcripts from the AnTat9013 and 2913 cells that mapped to assembled minicircles. The average size of sense transcripts mapping to gRNA genes ranges from 47-51 nt whereas the anti-sense transcripts have a average size ranging between 28-33 nt. Koslowsky et al., excised gRNAs from a denaturing gel in the size range of 40-80 nt explaining the lack of coverage for anti-sense transcripts. Similarly Madina et al., (2014) excised gRNAs from a denaturing gel from within a size range of 40-60 nt, again, this will select out the anti-sense gRNA transcripts which are shorter. Suematsu et al., (2016), selected RNAs from a size range of between 35-75 nt. This perhaps explains what they have a slightly higher mean for length of their anti-sense transcripts compared to data sets generated for this study (in which all small RNAs (<200 bp) were submitted for sequencing). It should be noted that reads obtained from public sources were not quality trimmed any further before mapping.

Cell line	Canonical cassettes (total=608) with small RNA read depth > 5	Non-canonical cassettes (total=558) with small RNA read depth >5
AnTat90.13 BSF	506	456
AnTat90.13 PCF	551	462
TREU667 BSF	157	124
TREU667 PCF	191	146
EATRO164 PCF	276	214
L427_29.13A PCF	180	136
L427_29.13B PCF	94	69

Table 3.8: U-tail trimmed small RNA reads mapped to assembled AnTat90.13 minicircles. gRNA cassettes with an average read depth > 5 were considered to be expressed. EATRO164 PCF, L427\_29.13A PCF and L427\_29.13B PCF are data downloaded from Koslowsky et al. (2014), (SRR1041339), Suematsu et al. (2016), (SRR2125761) and Madina et al. (2014), (SRX548086) respectively.

Cell line	Canonical cassettes (608) with small RNA read depth > 5		Non canonical cassettes (558) with small RNA read depth >5	
	Plus strand	Minus strand	Plus strand	Minus strand
AnTat90.13 BSF	500	468	437	404
AnTat90.13 PCF	519	365	419	303
TREU667 BSF	157	1	122	1
TREU667 PCF	189	3	144	5
EATRO164 PCF	276	40	210	46
L427_29.13A PCF	176	122	132	85
L427_29.13B PCF	94	0	67	2

Table 3.9: Strand specific gRNA expression. U-tail trimmed small RNA reads mapped to assembled minicircles. gRNA cassettes with an average read depth > 5 were considered to be expressed. EATRO164 PCF, L427\_29.13A and L427\_29.13B are data downloaded from Koslowsky et al. (2014), (SRR1041339), Suematsu et al. (2016), (SRR2125761) and Madina et al. (2014), (SRX548086) respectively.



*Figure 3.16: Length of small RNA reads for sense (green line) and anti-sense (blue line) which mapped to both canonical or non-canonical gRNAs. Reads mapping to anti-sense strands are shorter than sense reads. Frequencies are expressed as a percentage of total reads mapped. **A:** AnTat9013 BSF, **B:** AnTat9013 PCF, **C:** L427\_2913A (Suematsu et al. (2016)).*

### 3.3.5.2 Life cycle analysis of gRNA expression

In order to analyse the expression levels for the predicted gRNA cassettes the adapter trimmed reads were U-tail trimmed and mapped directly to the assembled and annotated minicircles. Using the AnTat9013 BSF and PCF small RNA data sets, matching transcripts could be identified for ~83% and ~91% of cassettes, respectively (gRNA cassettes with an average read depth >5 were considered expressed). For the other small RNA data sets the corresponding values ranged from ~15-45% (Table 3.8 and Table 3.9), indicating that gRNA sequences themselves are not highly conserved,

presumably because they evolve relatively quickly. Interestingly, initial comparison of the cassettes filled in the PCF vs BSF AnTat9013 sets showed that more gRNA cassettes were expressed in the PCF data set (Figure 3.17A). This was not an artefact caused by differences in sequencing depth, as overall gRNA coverage for BSF was ~9 times higher than for PCF (Tables 3.6 and 3.7). This observation of a higher number of gRNA genes being expressed in PCF than BSF also holds true for the TREU667 pair of samples (Figure 3.17B). PCF cells expressed more gRNAs for 8 (AnTat90.13) or 9 (TREU667) of the 12 edited genes than the BSF sample. Previous evidence suggested that there is no developmental regulation of gRNA expression (Koslowsky et al., 1992). This previous study was not a comprehensive analysis of total gRNA expression levels across the PCF and BSF life-cycle stages, and similarly the majority of gRNAs are indeed expressed in both life-cycle stages in our transcriptome data sets (explaining why they did not see any differences in their northern blots). However, high throughput NGS analysis allows a global and comprehensive analysis of which gRNAs are expressed in PCF and BSF life-cycle stages, and the observation of more gRNAs being expressed in PCF than BSF is consistent. This potentially points towards life cycle specific differential expression of gRNAs, or at least some difference in overall gRNA abundance. It should be noted, however, that complete editing capacity is maintained in both life-cycle stages despite the fact that some of the gRNAs expressed in PCF are not expressed in BSF.



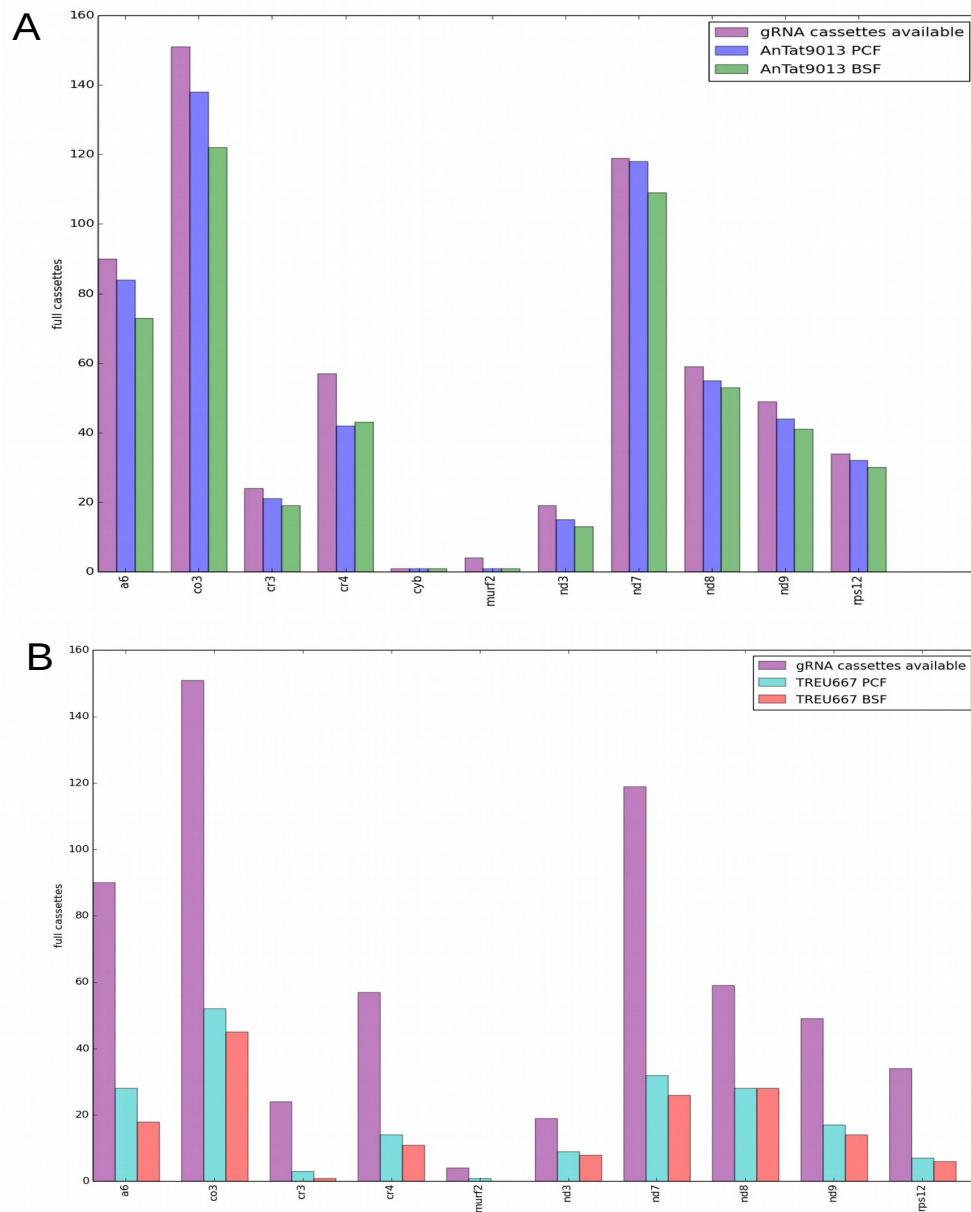


Figure 3.17 A: Number of expressed gRNA genes (average depth >5) after mapping U-tail stripped small RNAs generated from AnTat9013 PCF and BSF cells to assembled and annotated minicircles. A breakdown of the cassettes that are covered is shown in tables 8 and 9. B: Number of expressed cassettes (average depth >5) after mapping U-tail stripped small RNAs generated from TREU667 PCF and BSF cells to assembled and annotated minicircles. A breakdown of the cassettes that are covered is shown in tables 8 and 9.

When analysing the proportion of called gRNAs that code for edited genes the only notable difference for the Antat9013 gRNAs is the increase in PCF proportion of total gRNAs that have a match to edited RPS12 transcript (Figure 3.18A). The proportion of gRNAs coding for RPS12 increases from 17% to 25% between BSF and PCF respectively. This initially, perhaps, implied that RPS12 gRNA abundance is connected to mitochondrial up-regulation and life cycle progression, however the same pattern is not observed for the TREU667 pair of BSF and PCF samples (Figure 3.18B).

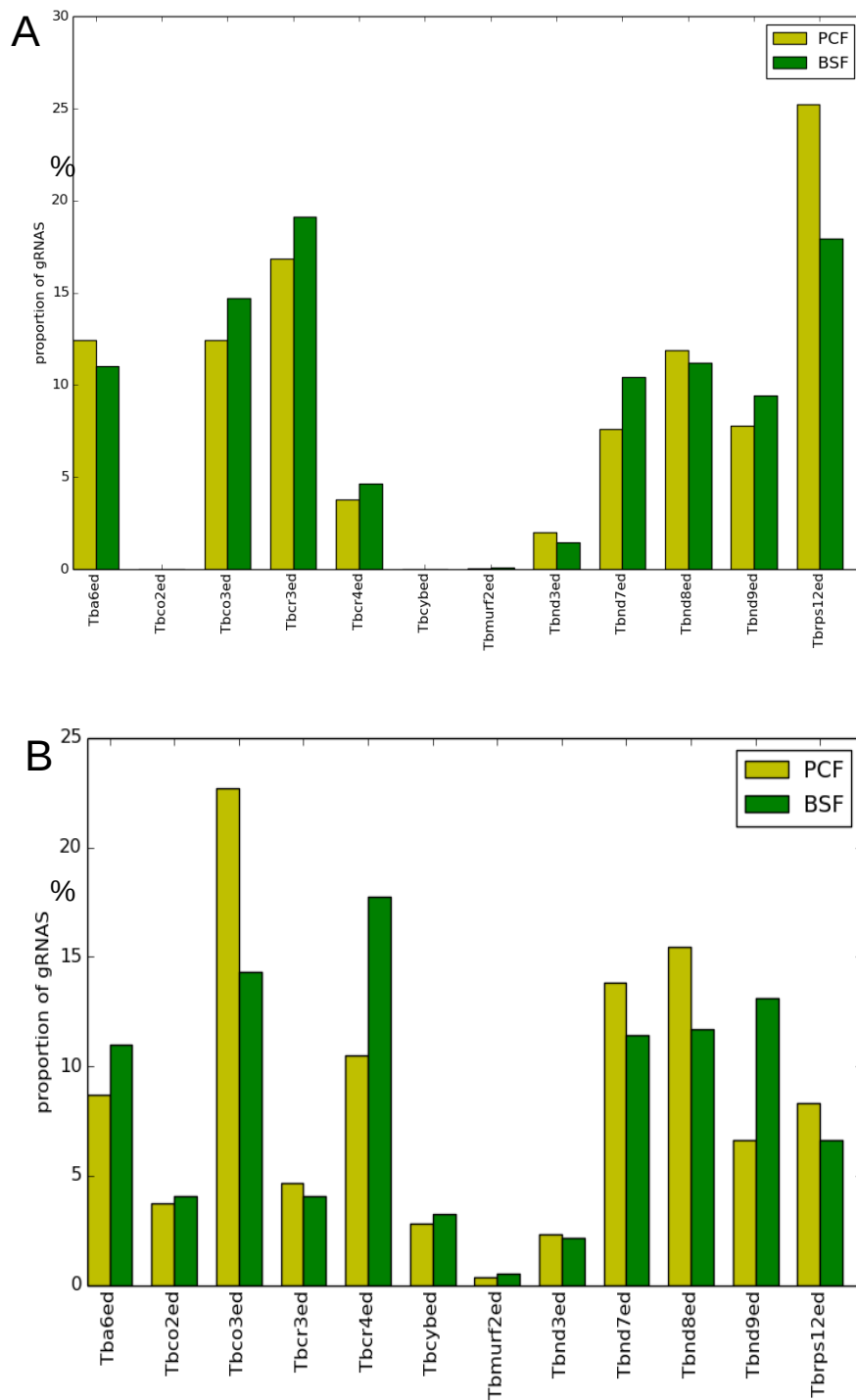
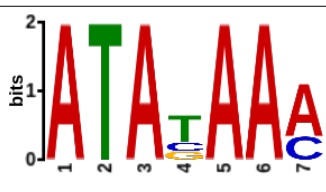
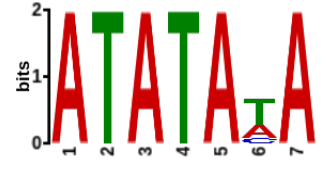


Figure 3.18: A: Proportions (%) of total Antat9013 gRNAs called by alignment to the editing space for each edited gene. B: Proportions (%) of total TREU 667 gRNAs called by alignment to the editing space for each edited gene



### 3.3.5.3 gRNA 5' sequences from gRNAs predicted using small RNA data

As previously described (Figure 3.5) the 5' RYAYA motif was searched for using the MEME suite (Bailey et al., 2009). In order to carry out the same analysis for gRNAs which were identified from small RNA data sets, the entire population of small RNA reads which have a match in the editing space were analysed using the DREME algorithm (Bailey, 2011). DREME facilitates elucidation of motifs from arbitrarily large sets of sequences. Similar motifs to those found in the gRNAs predicted from DNA sequences were found (Table 3.10). The difference in the major motif between the PCF and BSF samples is surprising.

Sample	MEME motif	E-value	Count
AnTat9013 BSF		8.6e-18351	134428/229610 (58%)
AnTat9013 PCF		2.9e-3485	24474/49804 (49%)

*Table 3.10* Analysis of 5' motifs for gRNAs called from small RNA data. Small RNAs with a match to edited mRNAs were U-tail stripped and directly piped into the DREME motif elucidation tool and motif with the smallest E-value (excluding the U-tail) is show in the table. Count is the number of times the motif was found/total canonical gRNAs submitted to DREME.

#### 3.3.5.4 Sense vs Anti-sense gRNAs

Manual inspection of stacks of small RNAs reads mapped to assembled and annotated minicircles (as in Figure 3.15) anecdotally suggested that there may be some differences in the ratio of sense to anti-sense reads for canonical and non-canonical gRNA regions. Figure 3.19 shows the per cassette ratio of sense to anti-sense transcripts for the data sets which had more than 100 gRNA genes expressed on both the plus and minus strand, as outlined in Table 3.9, thus, allowing large scale analysis of how sense and anti-sense transcripts relate to each other for each gRNA gene. A pattern of the canonical cassettes having a more equal ratio of sense to anti-sense reads is observed (an equal number of plus to minus transcripts would give a ratio of 1). Note that despite lack of gRNA coverage for the L427\_29.13A small-RNA data set a similar pattern of non-canonical cassettes having more sense transcripts than anti-sense transcripts is observed.

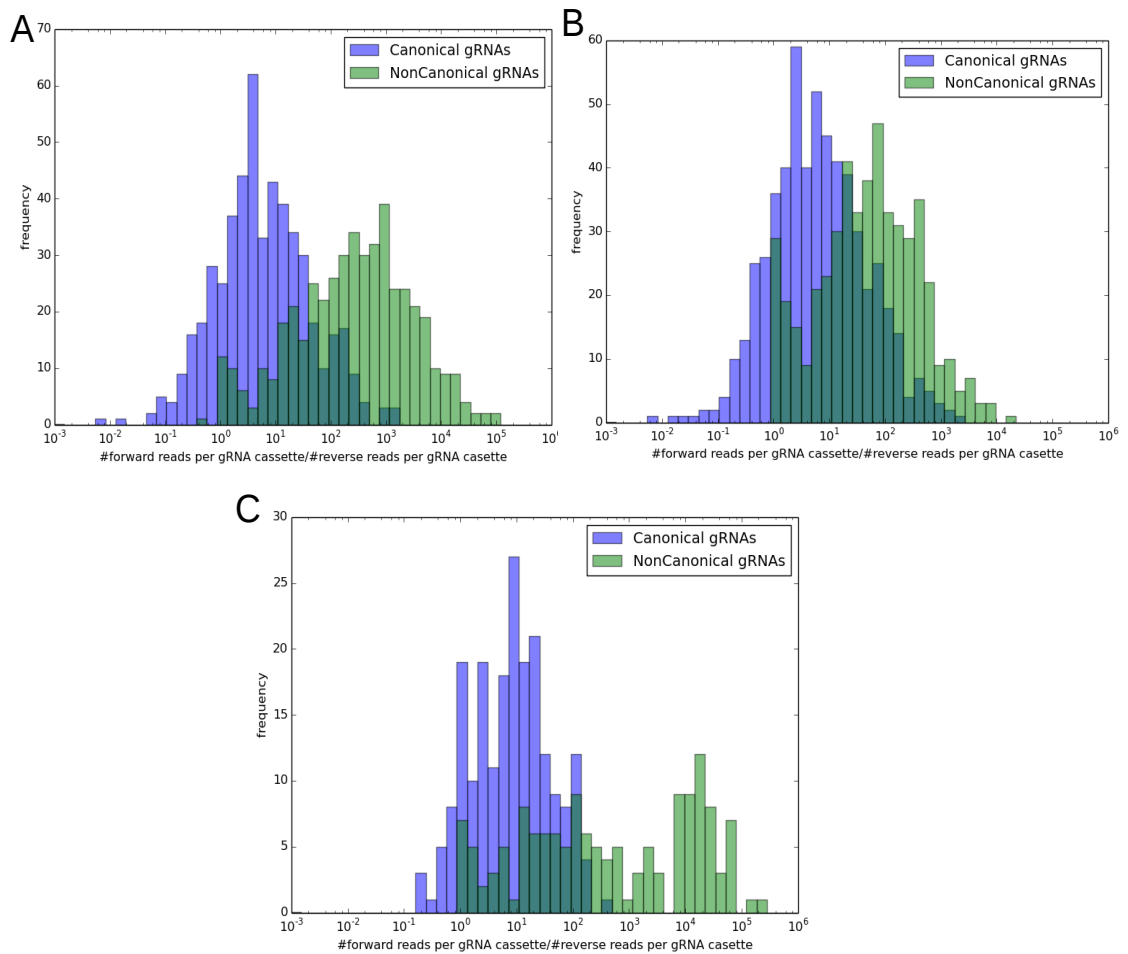
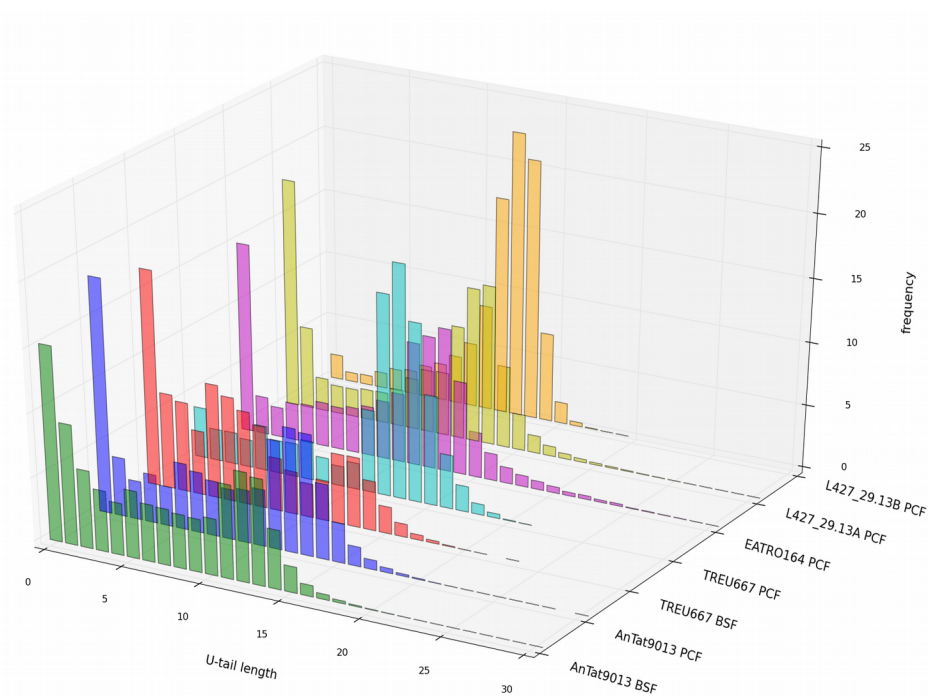


Figure 3.19: Ratio of per gRNA cassette sense to anti-sense transcripts. U-tail stripped small RNAs were mapped to annotated minicircles and the number of reads mapping to the plus strand within a predicted gRNA region were divided by the number mapped to the minus strand. Blue; canonical cassettes (have a match in the editing space). For cases in which there are no reads mapping to the minus strand the number of forward reads are divided by 1. Green; non-canonical cassettes (no-match in the editing space). **A:** Antat9013 BSF (Canonical mean=40, SD=2530; NonCanonical mean=2564, SD=8627, Kolmogorov-smirnof=0.629  $p=1.29e-85$ ), **B:** Antat9013 PCF (Canonical mean=45, SD=338; NonCanonical mean=343, SD=1312, Kolmogorov-smirnof=0.41,  $p=3.90e-41$ ), **C:** L427\_29.13A PCF (Canonical mean=25, SD=11298, NonCanonical mean=11296, SD=29795, Kolmogorov-smirnof=0.56,  $p=1.49e-24$ )

### 3.3.5.5 U-tail analysis

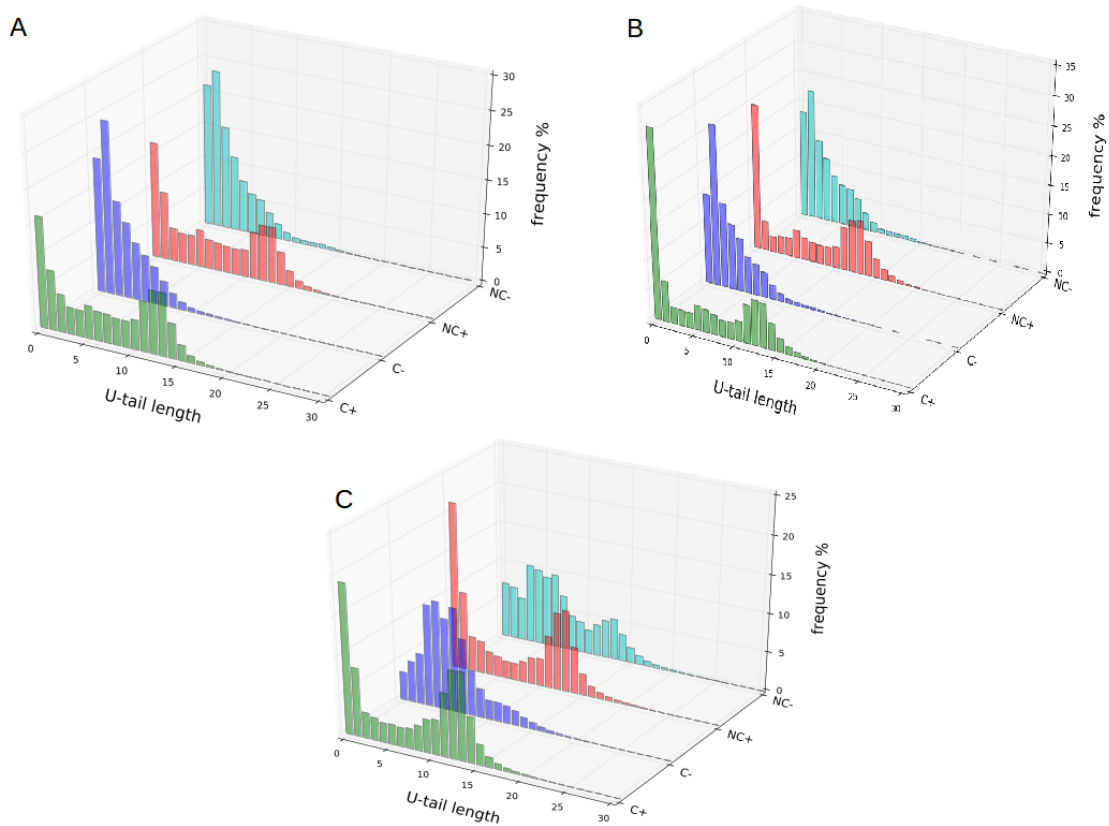
A recent analysis of gRNA processing (Suematsu et al., 2016) found a significant number of gRNAs that did not have a U-tail. This is also what is observed for gRNAs called from small RNA data sets analysed in this study. For all the data sets (except TREU667 PCF and L427\_29.13B) ~15-20% of called gRNAs had no U-tail or a single uridine at the 3'end (Figure 3.20). The profile of U-tail length looked similar for all the data sets analysed, with a major U-tail peak of between 10 and 15 nucleotides.



*Figure 3.20: U-tail length for small RNAs which have a match in the editing space. The frequencies of U-tail length are expressed as a percentage of total gRNAs called for each data set.*

The U-tail lengths for sense and anti-sense reads which mapped to the canonical and non-canonical cassettes were plotted for each of the samples which had more than 100 gRNA genes expressed (average read depth >5) for both sense and anti-sense transcripts (Figure 3.21). There are no differences between canonical and non-canonical U-tail lengths for either the sense or anti-sense mapped reads. There did however appear to be some subtle differences between the anti-sense U-tail lengths in

the AnTat90.13 samples generated in this study and the U-tail lengths in the anti-sense gRNAs from Suematsu et al. 2016). The U-tails for the anti-sense gRNAs from Suematsu show a main peak with a U-tail of length  $\sim 5$  nt, while the major U-tail length for the anti-sense transcripts from this study are at around  $\sim 2$  nt. This is probably a reflection of the size selection process used by Suematsu et al., by size selecting between 35-75 nt, the majority of anti-sense gRNAs with short or no U-tails will be lost (given that they have an average length of 28 nt).



*Figure 3.21: Utail plots for sense and anti-sense reads which mapped to either Canonical or Non-canonical cassettes. Cyan:(NC-) Anti-sense non-canonical reads, red:(NC+) Sense non-canonical reads, blue:(C-) anti-sense canonical reads, green:(C+) sense canonical reads. Frequencies are expressed as a percentage of total reads mapped. **A:** AnTat9013 BSF, **B:** AnTat9013 PCF, **C:** L427\_2913A.*

U-tail variability for each annotated gRNA gene (plots for both canonical and non-canonical, sense and anti-sense) are shown in Figures 3.22, 3.23, 3.24 and 3.25

(representative plots from the AnTat9013 BSF small RNA read sets are shown as all heat maps were similar). It was found that for the vast majority of gRNA genes, the corresponding transcripts have varying lengths of U-tails. The size distribution varies greatly between gRNAs and there is no apparent correlation between U-tail length and abundance (Figure 3.22). It was however, observed that reads mapping to the anti-sense strand have shorter U-tails overall. The dominant U-tail length for anti-sense gRNA reads is 1-5 nt in length (Figure 3.23). Sense gRNAs are more likely to have a range of U-tail lengths for each gRNA species, also noted for the non-canonical gRNAs. Non-canonical gRNAs show a difference in U-tail length between reads mapping to the plus and minus strand (Figure 3.24, Figure 3.25). We also see that there is no transcript-specific clustering for gRNAs which have a match in the editing space. What one might expect if U-tail length was important for gRNA stability is that gRNAs which are essential in BSF may be clustered together in the AnTat90.13 BSF samples but not in the PCF equivalent, but this is not observed.

Figures 3.22-3.23: Per gRNA cassette U-tail variability for reads mapping to gRNA cassettes (representative plot from Antat9013 BSF). gRNA genes are sorted according to percentage of reads with no U-tail. **3.22**: Canonical gRNAs sense strand, **3.23**: Canonical gRNAs minus strand. Left to right Panel 1: Per cassette U-tail length expressed as a percentage of reads mapped to each cassette, Panel 2: edited mRNA match for each given cassette or NB (cassette predicted by nucleotide bias only), Panel 3: per gRNA cassette read abundance.

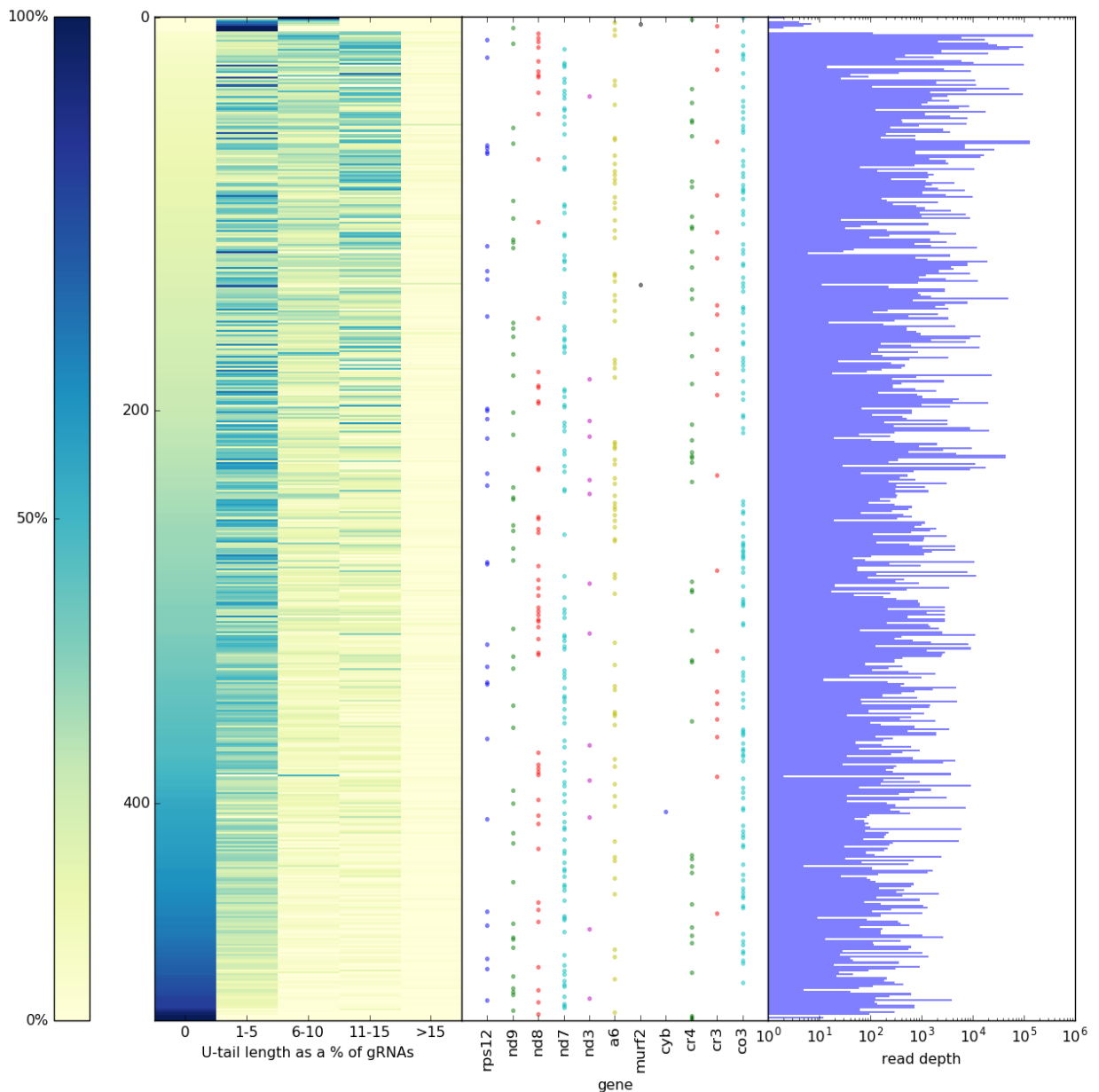


Figure 3.22: Per gRNA cassette U-tail variability for small RNA reads mapping to canonical cassettes on the sense strand.

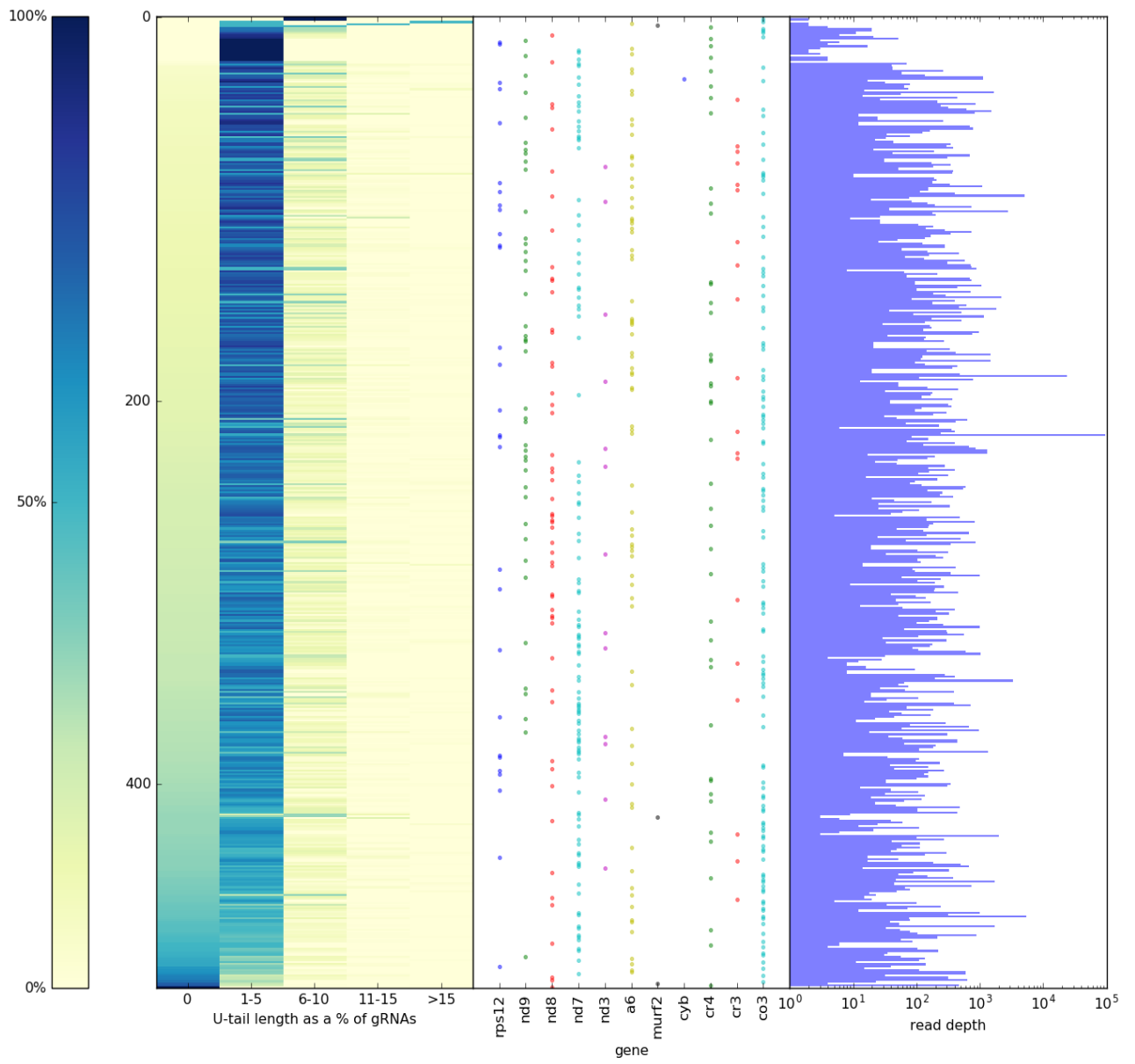


Figure 3.23: Per gRNA cassette U-tail variability for small RNA reads mapping to canonical cassettes on the anti-sense strand.



3.24: Non-canonical gRNAs plus strand, 3.25: Non-canonical gRNAs minus strand. Left to right, Panel 1: Per cassette U-tail length expressed as a percentage of reads mapped to each cassette, Panel 2: per gRNA cassette read abundance

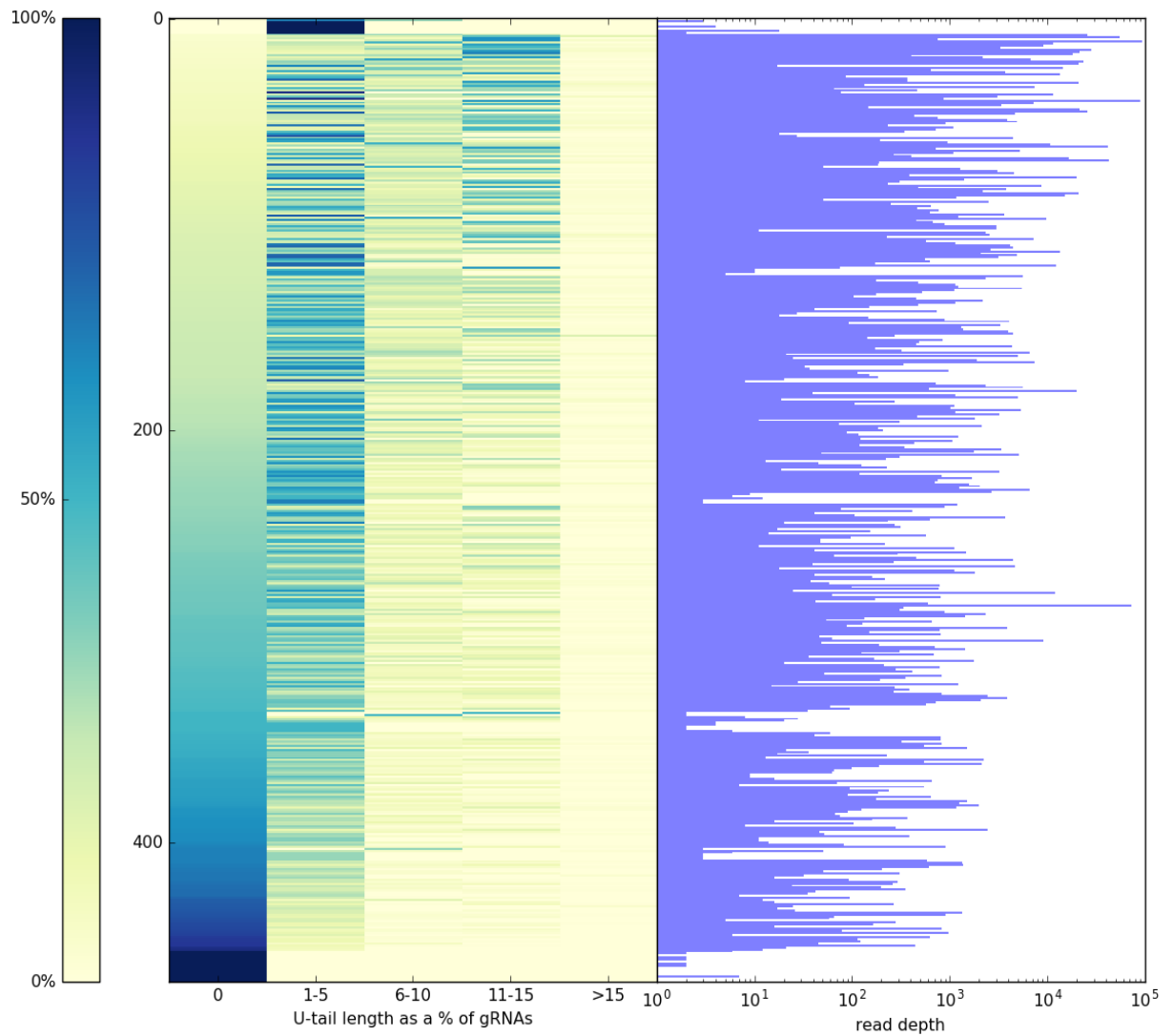
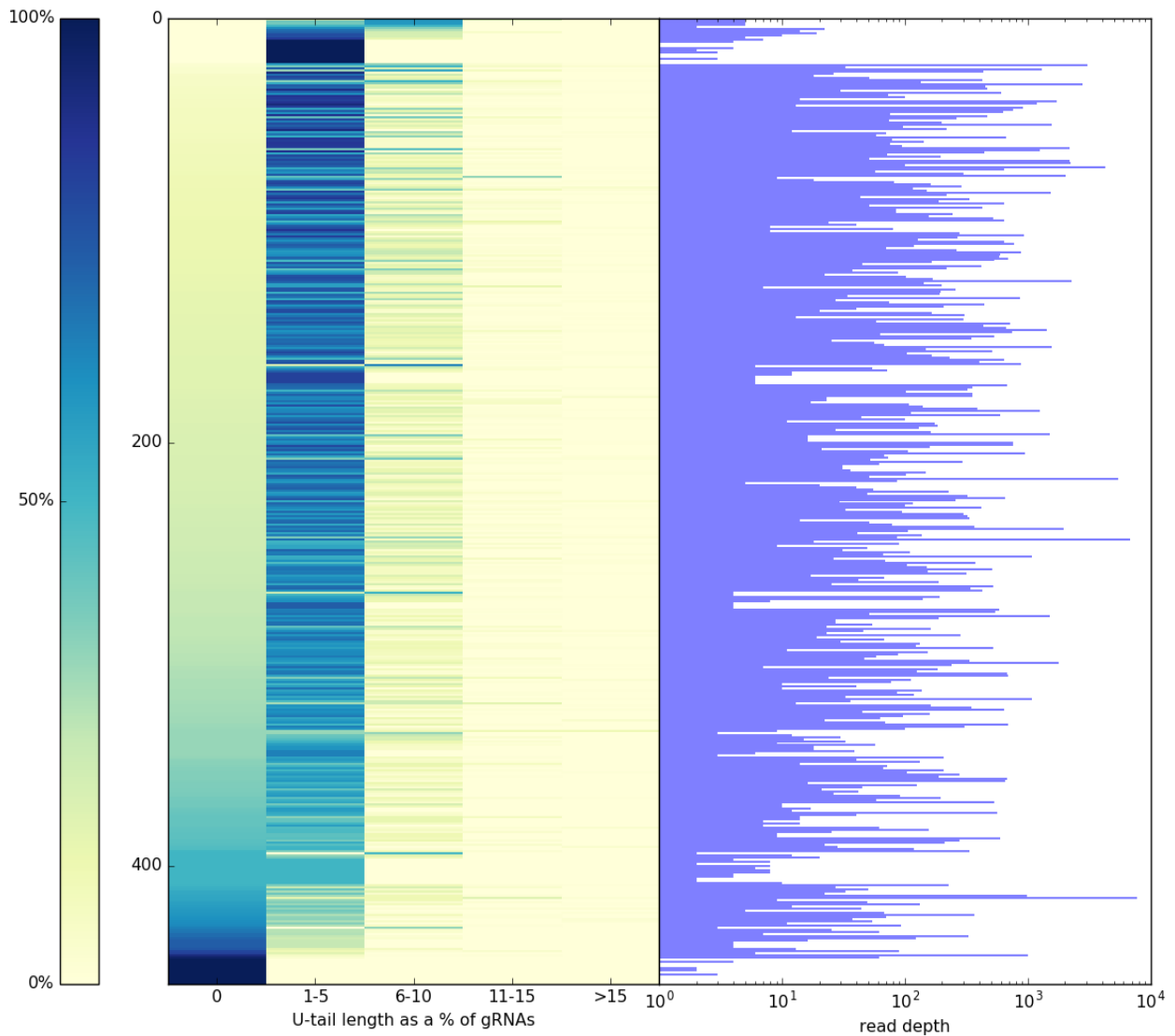


Figure 3.24: Per gRNA cassette U-tail variability for small RNA reads mapping to non-canonical cassettes on the sense strand.

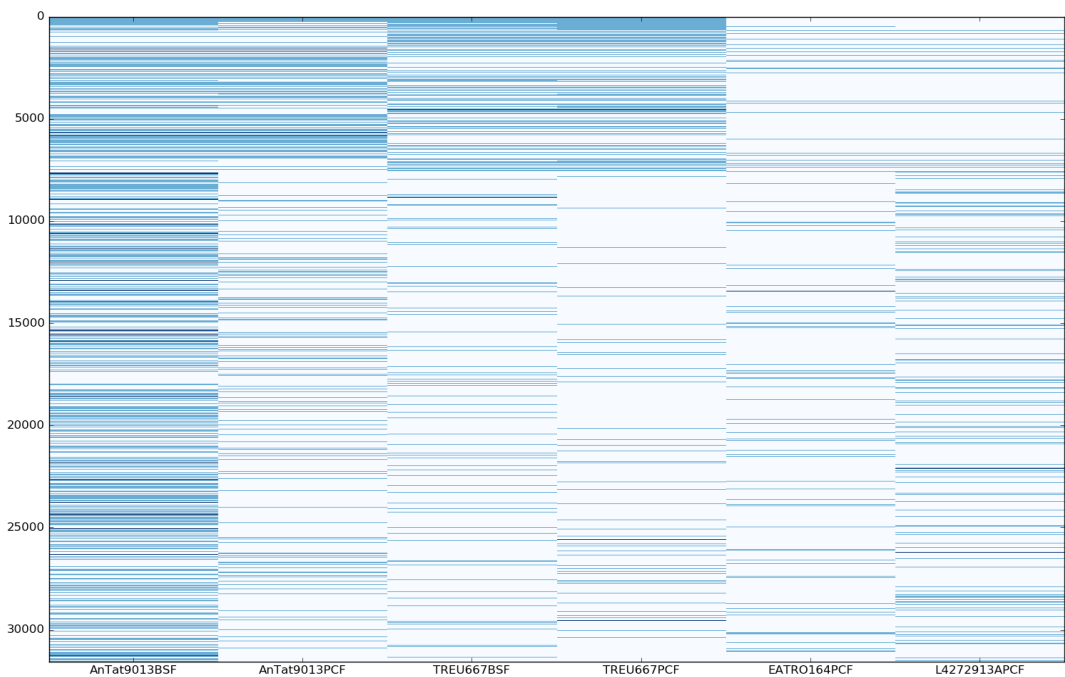


*Figure 3.25: Per gRNA cassette U-tail variability for small RNA reads mapping to non-canonical cassettes on the anti-sense strand.*

### 3.3.5.6 gRNA conservation

gRNAs which are shared across data sets are plotted in Figure 3.26. As expected samples which are the same cell line but from different life-cycle stages are similar. Table 3.11 shows the output from a BLAST comparison (-ublast (Edgar, 2010); options, evaluate 1e-11 and % i.d. 90% ) of the all of the gRNA data sets we have

available. The results show that the overlap between the different *T. b. brucei* data sets is very small. Note that the overlap is very dependent on the quality of the database that is being mapped to; for examples in our Antat9013 pair of samples the BSF data set has much higher kDNA coverage than the PCF (which also has shorter reads overall), thus when mapping the PCF set to the BSF set we get good overlap (as the entirety of the query will be present in the database gRNA). Whereas when the (usually longer) BSF reads are mapped to the PCF set the overlap is significantly reduced. This will be especially true for the comparing the TREU667 set to others (as they may have shortened 3' and 5' ends).



*Figure 3.26:* Cluster generated binary heatmap of gRNAs. gRNA sets were clustered together. Clusters are sorted by size (max size 6= gRNA is present in all samples). L4272913\_B is not included in this analysis.

Query	Database						
	AnTat9013BSF	AnTat9013PCF	Treu667BSF	Treu667PCF	164PCF	L427_2913APCF	L427_2913BPCF
AnTat9013BSF	12678	4831	225	292	1650	1253	1074
AnTat9013PCF	1604	5127	75	109	535	400	330
Treu667BSF	86	67	4088	1500	263	57	39
Treu667PCF	153	110	1552	3549	454	101	64
164PCF	470	414	232	387	2655	414	322
L427_2913APCF	844	689	228	177	1104	4483	1161
L427_2913BPCF	479	326	35	49	561	732	1465

*Table 3.11: BLAST approach for quantification of gRNA overlap between data sets. Databases of gRNAs which have been clustered using uclust (95%) identity are subjected to a pairwise blast analysis using ublast. BLAST of a database of gRNA classes against itself gives the total size of the database.*

### 3.3.6 Modelling gRNA distribution

To address the question of whether gRNA distribution is random or not a modelling approach was used to generate a set of minicircles which have a random distribution of gRNA genes. Random distributions were compared to real distributions to check if A6 and RPS12 gRNA distribution is random. Figure 3.27A shows a randomised distribution of A6 gRNAs compared to the true distribution. The overlap is shown in purple and suggested that the distribution of A6 gRNAs is random rather than determined. A distribution that would have supported the hypothesis in (Speijer, 2006) would have been to see more A6 gRNAs appearing as singlets (i.e. multiple A6 gRNAs per minicircle being very rare). More A6 gRNA singlets would mean more space for non-essential gRNAs to be coupled together with essential gRNAs. This random distribution of essential gRNAs is true for RPS12 (Figure 3.27B) and also when both are taken together (i.e. are A6 and RPS12 gRNAs more likely to be found separately or together), Figure 3.27C.

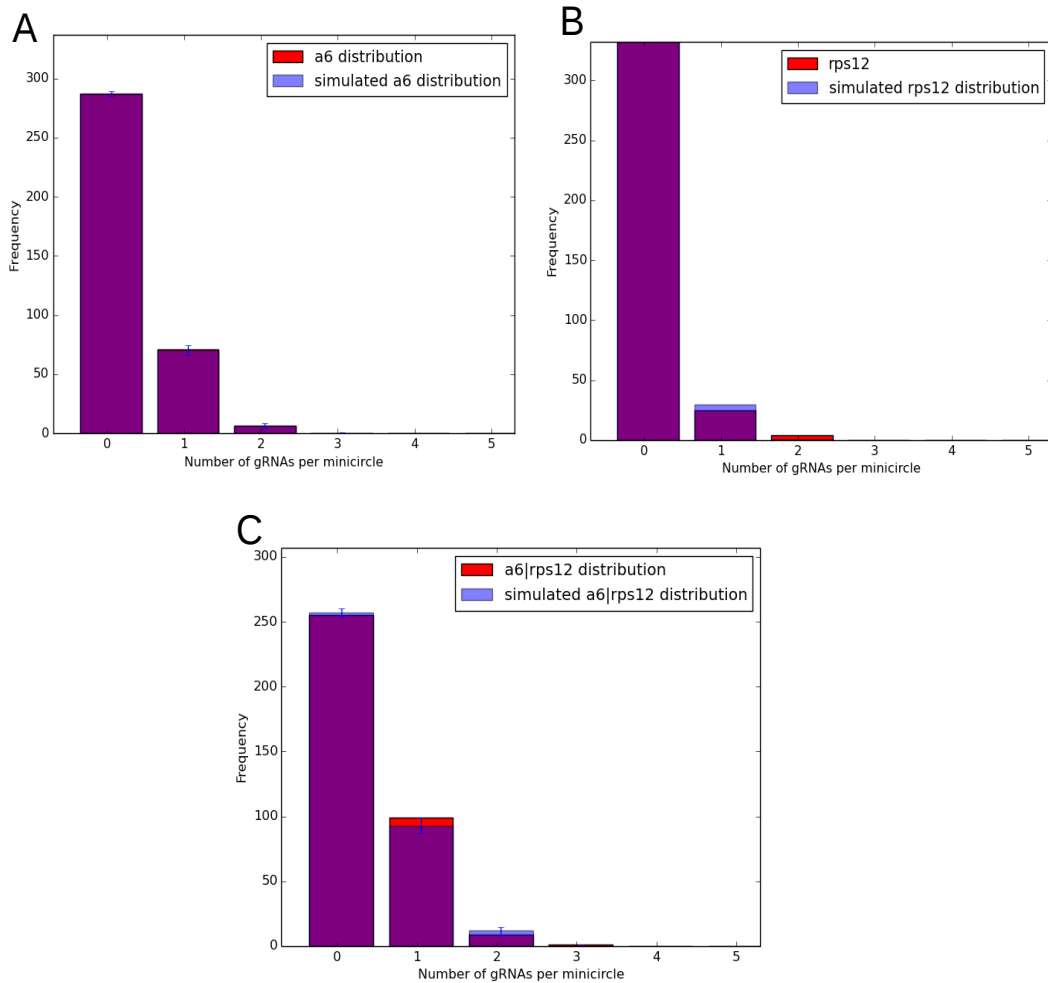


Figure 3.27: gRNA distribution amongst assembled AnTat90.13 minicircles. The distribution of gRNAs for a given gene of interest was recorded and compared to a randomised version of the same set of annotations, random distributions were generated from an average of 1000 permutations. Blue bars are the simulated distribution, red bars are the true distribution, the overlap is shown in purple, error bars are shown to highlight the variance in the simulated data set. A: distribution of A6 gRNAs amongst minicircles, B: distribution of RPS12 gRNAs amongst minicircles, C: distribution of A6 and RPS12 pairs on minicircles.

### 3.3.7 Analysis of kDNA from publicly available short read DNA data

In addition to the minicircles assembled for *T. brucei* Antat 9013 we have assembled minicircles from various publicly available whole genome short read data sets in which the kDNA has been sequenced incidentally rather than as the primary aim of the

sequencing initiative. Minicircles were assembled and annotated for *T. brucei* EATRO 164 (Stuart and Gelvin, 1980), Lister 427 (3 datasets related to 427 were used) and TREU 927. The datasets and the kDNA read statistics are shown in Table 3.12. *T. brucei* EATRO 164 was selected for this treatment as an equivalent AK sample was available and could be used as a negative control for optimisation of the gRNA prediction pipeline. WTSM is also a Lister 427 derived cell line from the lab of Keith Gull (Oxford). 388.5 is an L262Py mutant generated by Dr. C Dewar using a Lister 427 parental cell line.

sample	Accession/ (ref)	total reads	csb3 reads	% reads with csb3	read length	bases of CSB3 information	Freq of CSB3 read occurrence: 1 in N reads
TREU927	ERX000727	40424405	514909	1.27	76	39133084	78.5
427	ERX008998	33911957	379307	1.12	76	28827332	89.4
Lister 427-501	Cross et al., (2014)	25000000	174086	0.70	101	17582686	143.6
WTSM		21023505	119680	0.57	101	12087680	175.7
EATRO164	Stuart (1971)	12303508	38304	0.31	51	1953504	321.2
388.5 (427 derived)	Dewar C.	2426311	1046381	43.13	300	313914300	2.3

*Table 3.12: kDNA coverage statistics for read from publicly available datasets. Accession numbers are provided where available if no accession is available the publication reference is provided. The 388.5 cell line was generated by Dewar C.*

### 3.3.7.1 Assembly of minicircles from publicly available short read data

Contig length plots are shown for each of the data sets used (Figure 3.28). Interestingly despite differences in kDNA coverage for all of these datasets the samples which come from cell lines which are considered to be monomorphic (427 (Sanger centre), WTSM (Gull Lab) and 388.5) all consistently have fewer classes of minicircles than the cell lines which are thought to be differentiation competent (TREU 927, EATRO 164 (note low kDNA coverage from Table 3.12), AnTat9013, 427-501 (427 precursor). For information on the history of 427-501 and 427(New York) see (Cross et al., 2014).

The total number of minicircles expected to be in a sample can be extrapolated by extracting reads which contain the CSB3 12-mer, and mapping these

back to assembled contigs. The proportion of mapped CSB3 containing reads gives an indication of what proportion of minicircles have been sequenced with enough depth for a contig to be assembled (Table 3.13, Figure 3.29). It is notable that despite differences in kDNA coverage differentiation competent cell lines are consistently predicted to have ~400 minicircle classes compared to ~200 classes for monomorphic cell lines.

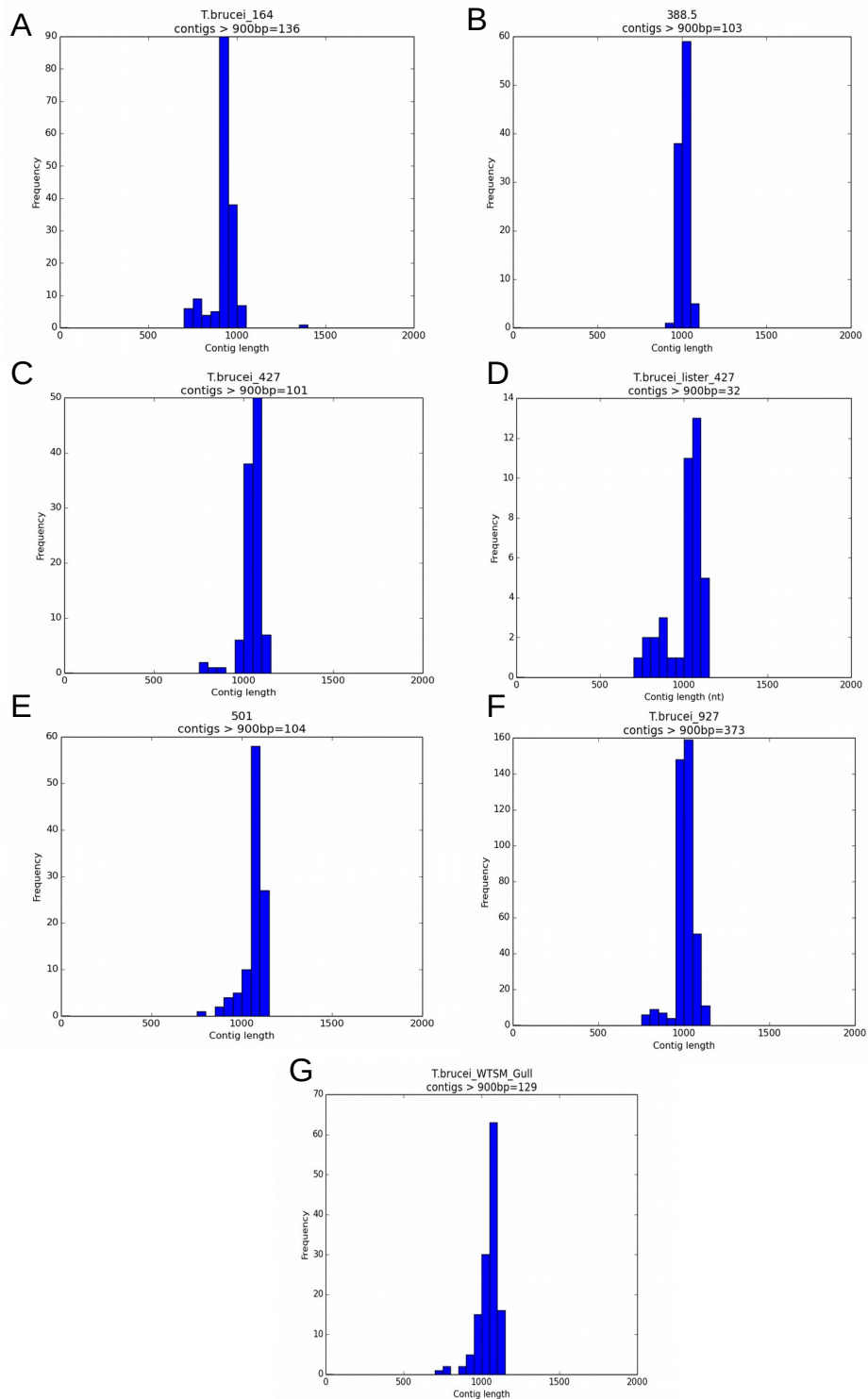


Figure 3.28: Size distribution of minicircles assembled from publicly available data sets. Minicircle assemblies using data either from publicly available sources; (NCBI SRA), TREU927 and Lister427. Or provided by George Cross; 501 and WTSM. Or datasets sequenced in this lab; 388.5 and 164. All data sets were assembled from reads that did not map to the nuclear genome using Velvet. Minicircle sequences were identified using the CSB3 sequence. 388.5 was assembled from 300bp paired end MiSeq reads (prepared from isolated kDNA) and thus could be subjected to the previously described test for circularity; 88 sequences passed the circularity test. A: TREU 164 minicircles, B: 388.5 (derived from 427), C: Lister\_427\_2913, D: Lister427, E: 501 (427 precursor), F: TREU\_927, G: WTSM.



Cell Line	Number of minicircles assembled	% of CSB3 reads mapping	Estimation of total number of minicircle classes	Ability to differentiate to procyclics
Antat9013	365	99	369	Yes
164	160	37	432	Yes
427	105	60	175	No
501 Cross	107	25	428	Yes
927	395	84	470	Yes
WTSM	134	63	213	No
388.5	103	83	157	No

Table 3.13: Number of minicircles assembled for each DNA data set used. For each data set, the reads with a CSB3 sequence were extracted and mapped back to the assembled contigs. The differentiation competency is a prediction based on anecdotal reports and was not performed in this study.

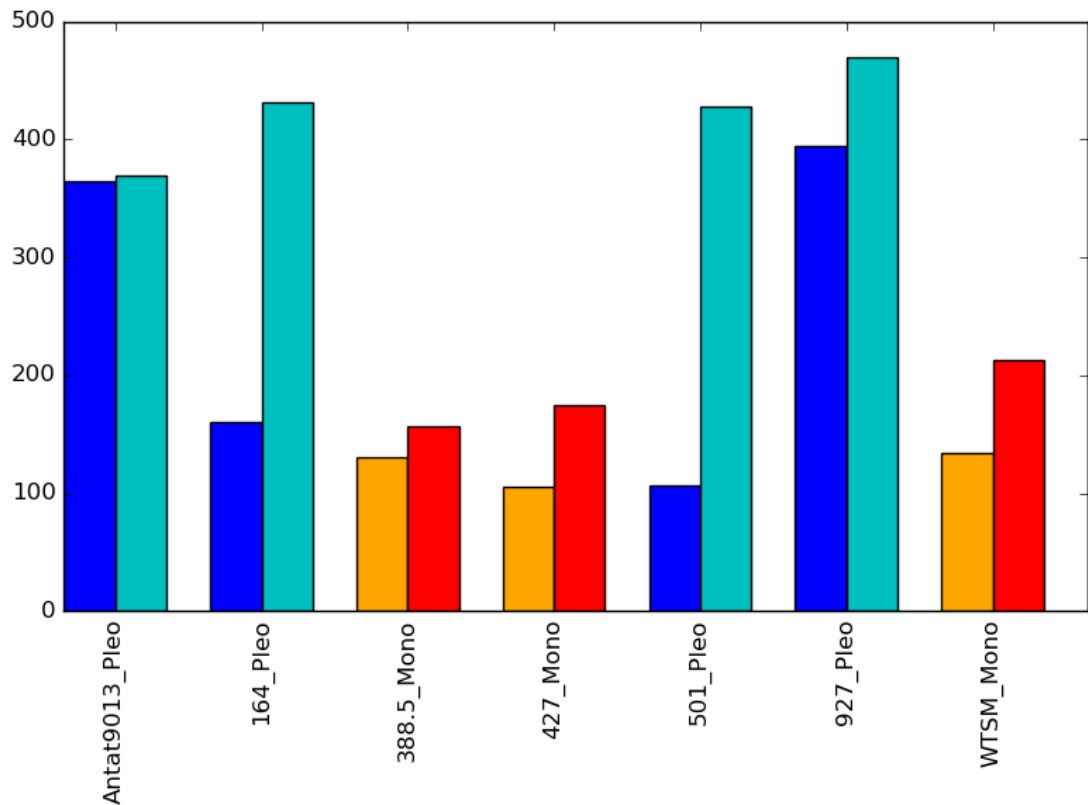


Figure 3.29: Number of minicircles (>700bp) assembled from publicly available data sets plus the total number predicted to be present based on the proportion of CSB3 containing reads that mapped. The proportion of CSB3 reads mapping was used to extrapolate and predict the total number of minicircle classes present in a population of cells. Dark blue: number of assembled minicircles in a differentiation competent cell line, Cyan: predicted number of minicircles in differentiation competent cell line, Orange: number of assembled minicircles for a monomorphic cell line, Red: predicted number of minicircles for a monomorphic cell line

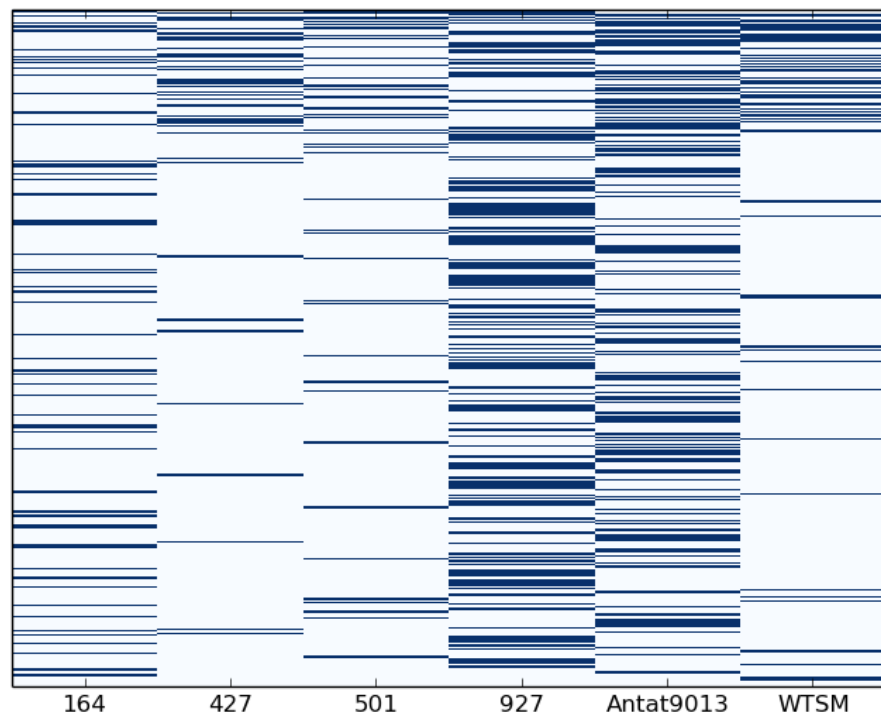
Sample 388.5 is a mutant cell line generated by Dr. Caroline Dewar and sequenced as part of a different project investigating kDNA requirements for life cycle progression (Dewar C. -in preparation). A notable feature of the minicircle classes assembled in this data set are that minicircles encoding A6 and RPS12 gRNAs are enriched in the population (Appendix 6.4, Figure 6.16). This perhaps suggests that the loss of minicircles in this cell line not been random and that gRNAs have been selected for which are required for survival. This is somewhat anecdotal, as we have not sequenced kDNA from the parental cell line so do not know if the selection occurred before mutation of the L262P  $\gamma$  mutation or has occurred as a result.

Similarly, as part of a different project, the kDNA complexity of *T. b. gambiense* type I isolates were investigated from whole genome samples. These samples were generated by Dr. Annette Macleod (University of Glasgow) as part of a project in which they show compelling evidence that the *T. b. gambiense* type I isolates are propagating asexually, and as a result harbour deleterious mutations via the action of Mullers ratchet (Macleod A. -under review). Logically, you would assume that these isolates would have low complexity kDNA genomes; as they have not had the kDNA complexity replenished via sexual recombination. This is what we see when we assemble minicircles and predict what the total complexity is via mapping back CSB3 reads (appendix 6.5, figure 6.17). Somewhat surprisingly 1 of these isolates also has a kDNA genome which is enriched for minicircles that code for A6 and RPS12 gRNAs (Figure 6.18).

#### 3.3.7.2 Comparing minicircles from different cell lines

A comparison of all minicircles assembled from the cell lines in Table 3.13 was carried out using a cluster based approach implemented using the `usearch -cluster_fast`

algorithm (Edgar, 2010). Contigs that clustered with an identity of 95% or above were considered to be the same minicircle. These clusters were then plotted using a binary heatmap to show presence or absence of a given minicircle class for each cell line (Figure 3.30). A more detailed breakdown of minicircle overlap using a global alignment approach (usearch -usearch\_global, similar to section 3.3.5.6) is shown in Table 3.14. Aligning a set of minicircles with itself gives the total number of minicircles per data set. The overlap of minicircles across the various data sets is very low.



*Figure 3.30: Cluster generated binary heatmap of minicircle classes (usearch uclust algorithm 95% identity). Dark blue bars show which samples have a minicircle present in a given cluster. Left to right: TREU164, Lister427, 501 (Lister 427 precursor), TREU927, Antat9013, WTSM.*

DB

Query	Antat9013	EATRO 164	388.5 (427 Edinburgh)	TREU 427 (New York)	501 (427 precursor)	TREU 927 (Sanger)	WTSM (New York)	KISS (TREU 667)
Antat9013	365	29	29	28	27	16	41	10
EATRO 164	31	160	11	10	7	16	11	11
388.5 (427 Edinburgh)	28	8	103	82	41	5	84	1
TREU 427 (New York)	36	8	87	105	37	6	82	1
501 (427 precursor)	41	6	44	38	107	5	51	6
TREU 927 (Sanger)	25	9	3	3	2	395	3	6
WTSM (New York)	58	10	88	81	48	7	134	2
KISS (TREU 667)	201	73	87	89	90	66	91	455

Table 3.14: Global alignment approach for comparing conserved minicircles. A query minicircle is considered to be present in a data base if it aligns with >95% identity over more than 700 bp of its length

### 3.3.7.3 Loss of editing capacity in monomorphic cell lines

Prediction of gRNA from existing short read data for monomorphic (Lister 427, accession:ERX008998) and differentiation competent (TREU 927, accession: ERX000727) cells line are presented at <http://tinyurl.com/jxvxodk>. The total editing space covered (i.e. percentage of editing events covered) by the 927 data set is 98% percent, compared to 93% in the 427 dataset. This is despite the 427 dataset having approximately a 3X greater kDNA coverage (based on maxicircle read depth, and % of reads containing CSB3). A breakdown of the number of editing events covered for each of these data sets is shown in Table 3.15. Notable is the complete absence of cytochrome b gRNAs in the monomorphic data set, as well as the relatively low coverage for many of the subunits of complex I (again bearing in mind that the Lister 427 data has approximately 3X greater kDNA coverage than the TREU 927 set). Figure 3.31 shows an example coverage plot for gRNA coverage of edited ND9 (interactive gRNA coverage plots are available at [hank.bio.ed.ac.uk](http://hank.bio.ed.ac.uk)).

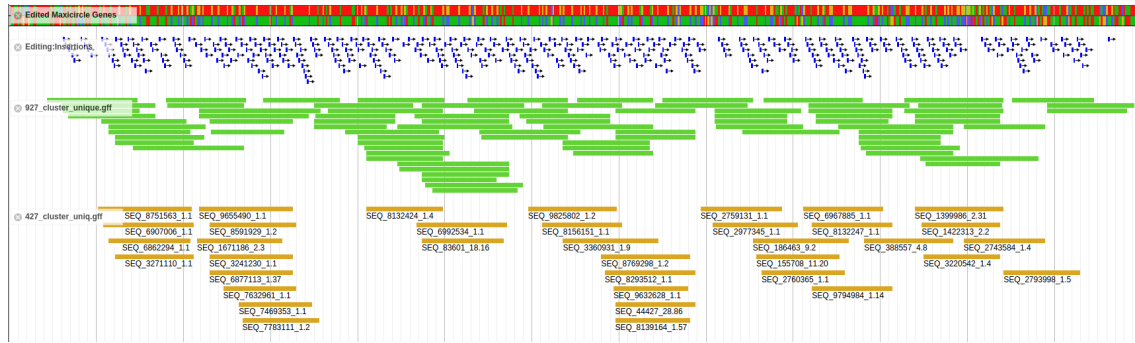


Figure 3.31: 427 monomorphic cell lines have gaps in the coverage of the editing space. Top to bottom: Edited nucleotide sequence for *T. brucei* ND9. Blue arrows indicate U-insertion events. Green bars show gRNA coverage for unique gRNAs predicted from differentiation competent 927 reads. Brown bars show unique gRNA coverage for monomorphic 427 reads.

Edited sequence	Insertions/deletions	Insertions covered/percentage of insertions covered	
		Lister 427	TREU927
a6	447/28	428/96.0	447/100.0
co2	4/0	0/0	0/0
co3	547/41	516/94.0	542/99.0
cr3	148/13	139/93.0	120/81.0
cr4	325/40	309/95.0	323/99.0
cyb	34/0	0/0	27/79.0
murf2	26/4	26/100.0	25/96.0
nd3	210/13	198/99.0	200/100.0
nd7	553/89	504/91.0	546/99.0
nd8	259/46	244/94.0	244/94.0
nd9	345/20	297/87.0	342/100.0
rps12	132/28	132/100.0	131/99.0

Table 3.15: Editing space coverage for monomorphic (Lister 427) and pleomorphic (TREU 927) cell lines. gRNAs were called directly from short DNA reads. These sets were then used to measure the number of editing events covered by each data set.

## 3.4 Discussion

### 3.4.1 Minicircle assembly and annotation

The true complexity contained within the kDNA genome of the bloodstream stage of a fully differentiation competent *T. brucei* cell line (AnTat90.13) has now been established (it should be noted that results for other strains and snapshots in time will vary). We have determined the complete sequences for 365 unique minicircles, representing at least 99% of the total number of unique minicircles in this strain (based on the number of CSB-3 reads captured). This number is somewhat consistent with some previous estimations of the number of minicircle classes per cell (Steinert and Van Assel, 1980), but considerably higher than others (Stuart, 1979). This is a significant improvement on previous sequencing attempts (Hong and Simpson, 2003; Ochsenreiter et al., 2007a). It should be noted that in both of these studies comparative sequence analyses were carried out across minicircles isolated from several strains of *T. brucei*, hence this is the first study of the kDNA complexity in a single strain. Steinert & Van Assel (1980) measured the renaturation kinetics of the *T. brucei* kDNA and estimated that there are between 200 and 300 distinct sequence classes. The question of overall complexity in *T. brucei* has been a conundrum since these early studies indicated that the overall kDNA heterogeneity in this species is much greater than that observed in other trypanosomatids. The complexity far outweighs the requirements based purely on the provision that all editing events must be covered (Hong & Simpson 2003). Mapping back CSB3 containing reads to the assembled set of minicircle contigs resulted in a 99% alignment rate, suggesting that three or four minicircle classes were either not assembled, or did not have a sufficiently conserved CSR sequence which could be used to identify them. It is possible that the 1% of CSB3 containing reads which did not map to the assembled set of minicircles could also be maxicircle derived

(Sloof et al., 1992), filtering for maxicircle derived CSB3 containing reads would be able to verify this.

The size distribution of the minicircles assembled in this study is very tight. Considering the fully circularised minicircles only, they have a mean length of 1005 bp and a standard deviation of just 21.5 bp (Figure 3.2D). The reason for why minicircles in *T. brucei* are ~1 kbp in the first place is somewhat perplexing. Is the size and small variation in the size related to the replication or transcription machinery or is it an intrinsic component of keeping the network intact? Minicircle sizes vary amongst trypanosomatids, ranging from ~2.5 kbp in *C. fasciculata* (Birkenmeyer et al., 1985) and ~0.9 kbp in *L. tarentolae* (Simpson et al., 2015) to ~0.5 kbp in *T. vivax* (Borst et al., 1985; Greif et al., 2015). These variations in size could be reflections of recombination events where by two minicircles combined at regions of homology (for example in a conserved gRNA region) to form a larger minicircles, as proposed by Sugisaki & Ray (1987). This is supported by larger minicircles having multiple CSB sequences. The fact that the smallest minicircle is found in *T. vivax*, which is thought to have diverged early from the salivarian clade (Cortez et al., 2006), is also perhaps supportive for this in the case of the salivarian trypanosomes.

Analyses of the CSR amongst the 365 assembled minicircle classes showed that CSB3 is the most conserved of the three hyper conserved sequences. For the 365 sequences, 93% of them had an exact copy of the CSB3 12-mer (GGGGTTGGTGTA) as described (Ray, 1989; Hong and Simpson, 2003) with two minor sequence variants CSB3\_A8 and CSB3\_A1 present (Table 3.3). CSB1 was also found to be highly conserved, however two sequence variants were found to be present in roughly equal proportions (Figure 3.5), giving a CSB1 consensus of WGGGCGTGCA. Whilst there was found to be a dominant CSB2 sequence, it was

much more diverse than the flanking CSB1 and CSB3. This suggests that CSRs involvement in binding minicircle replication machinery is perhaps primarily dependent on CSB1/3 and is also perhaps reliant on the distance between these two hyper conserved regions, given that the space between them is also highly conserved. However, it has been shown that UMSBP only requires the CSB3 sequence to bind to minicircles and recruit other elements of the replication machinery (Abu-Elneel et al., 1999). Thus the role of the other elements of the CSR is unclear. In *C. fasciculata* it has been shown that the binding of the UMSBP to the CSB3 sequence induces a conformational change in the minicircle DNA (Onn et al., 2006). It is possible that this change then allows the binding of other proteins (for example a primase (Hines and Ray, 2010)), which in turn interact with other elements in the CSR. In addition some minicircle classes had multiple copies of CSBs. Two minicircle classes had an almost complete additional CSR sequence, with an extra CSB1 and CSB3 motif; a third minicircle class had an extra CSB3 motif. These findings are possibly evidence for a recombination event that could give rise to multiple CSRs per minicircle. Recombination of minicircles is an aspect that needs to be investigated further.

The annotation of the imperfect inverted repeat sequences, thought to be important for gRNA transcription (Pollard et al., 1990) revealed 1042 “perfect” cassettes, whereby the forward repeat was followed by a reverse repeat and 134 “broken” cassettes containing only one of the repeat sequences. The distance between inverted repeats in “perfect” cassettes is  $104 \pm 10.4$  nt (Figure 3.7). Given 365 minicircle classes, the number of cassettes per minicircle is 3.2. This was roughly consistent with earlier predictions of three gRNAs per minicircle (Pollard et al., 1990; Hong and Simpson, 2003).

Annotation of the gRNA genes by aligning the minicircle sequences to



the edited maxicircle transcripts resulted in 608 gRNA genes being detected, 457 of those within “perfect cassettes” of inverted repeats. Manual inspection of a few of the remaining 151 canonical gRNA genes found one repeat motif, either upstream or downstream, but with no counterpart. Thus, more than half the number of cassettes (as defined by an inverted repeat) remained ‘empty’. Nevertheless the gRNAs that were detected covered nearly all known editing sites in the 11 trans-edited mRNAs (canonical ‘editing space’), with 97% of all U-insertion and deletion events covered by at least one gRNA. Based on previously annotated minicircles, there is some precedence for the presence of gRNAs with no known function (Hong and Simpson, 2003; Madej et al., 2008b; Ochsenreiter et al., 2008b; Simpson et al., 2015). In the case of *L. tarentolae* (Simpson et al., 2015), where each minicircle only encodes one gRNA, 10 out of the 114 minicircle sequences identified encoded putative gRNAs with no match to an edited transcript, in that and earlier work termed ‘orphan’ gRNAs (Simpson et al., 2000, 2015; Hajduk and Ochsenreiter, 2010). In the following, we refer to them as non-canonical gRNAs. One reason for not detecting a match in the editing space could be related to the stringency of the gRNA identification pipeline. However addressing this problem is difficult as it is currently unknown what defines a ‘functional’ gRNA. The observation that gRNA-mRNA matches are close to statistical noise (von Haeseler et al., 1992) further complicates gRNA detection. In the present work this problem was mitigated by tuning the gRNA detection algorithm using an Ak and WT cell line as negative and positive controls (see section 3.3.1). To further reduce the false positive rate a permutation test was developed. This tests each putative gRNA to ensure that the mapping was not non-specific mapping of low complexity reads by randomly shuffling the sequence 300 times and mapping the shuffled sequence to the edited sequence database; scores were recorded to give a distribution of permuted

scores. This was a similar approach to the permutation test used in Thomas et al. (2007). The true score was compared to the permuted scores to check if it is out-with the distribution of scores that random combinations of nucleotides with the same proportions of A, T, G and C achieve. This was found to significantly reduce the number of reads mapping to never edited regions, i.e. it reduced the false positive rate (see [hank.bio.ed.ac.uk](http://hank.bio.ed.ac.uk) or <http://tinyurl.com/jekxqry> for EATRO 164, Ak/WT gRNA mapping). This tuning of the gRNA identification pipe using kDNA positive and negative samples is a novel approach that has not been used by any other gRNA detection methods (Koslowsky et al., 2014; Simpson et al., 2015; Suematsu et al., 2016) and is especially useful for the elimination of false positives from large data sets, for example millions of small RNA reads (discussed below).

Given the extensive development, testing and fine tuning of the gRNA calling pipe to eliminate false positives and negatives. As well as the previous cases where minicircles with no canonical gRNA genes have been found, we conclude that the fact that we only detect half as many canonical gRNA genes as we might expect (based on the number of minicircle classes and gene cassettes) was not down to the gRNA identification method. We considered that the presence of inverted repeats alone might be an unreliable indicator for the presence of a gRNA gene and performed a thorough sequence analysis of canonical gRNA genes. This led to the identification of significant nucleotide bias in the gRNA sequence as well as in the regions immediately upstream and downstream, and to the development of a novel gRNA identification pipeline that was independent of homology to known edited sequence. Notably, the 151 canonical gRNA genes not flanked by inverted repeats (as mentioned above) all appeared to have the same strong upstream and downstream nucleotide bias as the 457 canonical gRNA genes found within such a cassette. Applied to the complete minicircle

set this pipeline confirmed 558 putative non-canonical gRNA genes within the “empty” cassettes defined by inverted repeats. Thus, the number of such non-canonical gRNAs is apparently much higher in *T. brucei* than in *L. tarentolae*, this could be related to gRNA redundancy and having multiple gRNAs per minicircle (Savill and Higgs, 2000). Further investigation of the non-canonical gRNA genes showed that they not only shared the same nucleotide bias characteristics as canonical genes, but also shared the same ATATATA initiation/anchor motif at the predicted 5' end (Figure 3.12), similar to the previously described RYAYA motif (Pollard et al., 1990). The fact that the putative non-canonical gRNA genes shared many characteristics with the canonical led to the investigation of whether they give rise to stable transcripts, which is discussed in the next section.

It was also noted that there were other regions outside of cassettes which had a significant nucleotide bias indicating the potential presence of additional gRNA genes. Further optimisation of the algorithms required to be able to reliably predict putative non-canonical gRNA genes in the absence of inverted repeats as the signal-to-noise ratio of nucleotide bias can be low. Such optimisation efforts promise to be rewarding, however, as this will allow global and unbiased prediction of putative gRNA genes in other kDNA sequencing data. The nucleotide composition of minicircles is known to be AT-rich, and AnTat90.13 minicircles have a GC content of just 27% compared to the nuclear genome, which is 50% GC (Berriman et al. 2005), and maxicircles, which have a GC content of 24%. The di-nucleotide frequency of minicircles has been observed to generate very different profiles for various species of trypanosomatids (de Oliveira Ramos Pereira & Brandão 2013) and could conceivably be used to isolate minicircle reads from whole genome data (Greif et al. 2015). The nucleotide composition of gRNAs and the flanking sequence has not been studied in

detail before now. It has not been tested if the nucleotide bias scoring vectors generated for *T. brucei* AnTat90.13 in this study would work for other trypanosomatids such as *Leishmania spp.*. The scoring vectors were used directly to identify gRNA genes in other strains of *T. b. brucei* (data not shown) and found to be accurate for the prediction of gRNAs in these cases also. Given the ubiquity of the nucleotide bias across the majority of gRNAs identified (the nucleotide bias is more consistently present than the inverted repeats) it is conceivable that the upstream and downstream nucleotide composition is important for the binding of the transcription machinery and may in fact constitute the long-sought gRNA promoters. Testing the scoring vectors generated in this study against more phylogenetically distant trypanosomatids will give information about how conserved the gRNA nucleotide bias is.

### 3.4.2 Small RNA analyses

Small RNAs were isolated from mitochondrially enriched fractions from AnTat90.13 BSF and, after *in vitro* differentiation, PCF cells. Similarly a pair of BSF and PCF samples were generated from *T. brucei* TREU 667 cells (provided by our collaborator T. Ochsenreiter). These two pairs were analysed along with three *T. brucei* PCF samples published by other labs: EATRO 164 from Koslowsky et al., (2014), and two different sets of Lister 427 samples from Suematsu et al., (2016) and (Madina et al., 2014). During preparation of this thesis analysis of gRNAs from EATRO 164 BSF was published by Kirby et al., (2016). This analysis will be discussed in the context of the present study, however the raw sequencing data could not be analysed with the pipelines developed for this thesis. Generation of these small RNA libraries used differing protocols that resulted in variations in size distributions and biases for coverage of the editing space and/or the coverage of minicircle gRNA cassettes. The

various protocols are briefly described in section 3.2.6.1, and basic read mapping statistics are detailed in Tables 3.6 and 3.7.

Based on the total number of canonical gRNAs annotated on minicircles was ~1000, it was expected that a similar number of unique gRNA classes would be present in the small RNA data sets. However the number of sequence classes after collapsing and clustering is far more than the number of annotated gRNAs, despite this the majority of gRNAs which have a match in the editing space also map to an annotated gRNA. Ascertaining the level at which this discrepancy between the number of unique gRNAs generated from small RNA data and the number of gRNAs annotated on minicircles is difficult. It could be down to sequencing errors in the small RNA reads generating false diversity, a problem with the cluster based classification method or a reflection of true heterogeneity of minicircles. These issues also have been encountered by others and are also affected by the questions of how gRNAs are defined and how different gRNA species are distinguished from each other. Given this conundrum the key questions when trying to define the number of gRNA classes are, how much internal variation should be accepted between related sequences of gRNA? And how much variation do you allow at the 5' or 3' prime ends? Koslowsky et al. (2014) opted for an approach whereby a class was defined as any small RNA that mapped within the same coordinates, allowing for some heterogeneity at the 3' end, gRNAs were then sorted into major and minor sequence variants based on abundance. The authors did not state the total number of unique classes they defined by this method. Madina et al. (2014) did not attempt to classify the number of unique gRNA classes in their data set, focussing on abundance of gRNAs covering transcripts of interest. Suematsu et al. (2016) classified their gRNAs based on sequence using a cluster based method which defined 60,441 gRNA species in their study. The approach

in this study is similar to Suematsu et al. (2016) and was to take small RNAs which have a match in the editing space and cluster them by 95% identity and then choose the centroid of the cluster (which is usually also the most abundant sequence) to represent the class. None of these approaches give a satisfactory method for which to classify the gRNAs. A possible source for the heterogeneity that has been seen in this study and others when attempting to classify gRNAs could be related to variations within what we have defined as a single minicircle class. There could be minor variants of minicircle sub-classes which in turn could give rise to more gRNAs than we have annotated gRNA genes to accommodate, this is supported by the fact that despite the number of clustered small RNA gRNAs being far beyond the number of annotated gRNAs the majority of these reads map to minicircles and was also observed by Simpson et al., (2015). Dissecting this in order to re-classify the number of minicircle classes we have defined by assembly and give a number of minicircle classes each with a number of sub-classes is potentially fraught with complications. It involves dissecting true minicircle sequence variants from sequencing errors, however it is possible and could account for the high number of gRNA classes we have identified. At the RNA level there are more variables that can account for the high number of clusters; variation in transcription start sites, 3' processing, errors in transcription and sequencing errors. Given these layers of potential variability it becomes clear that defining the total number of gRNA classes using transcriptome data alone is not sufficient, having an annotated set of minicircle sequences obviates the requirement for clustering as the major classes of gRNA are already annotated. This allows analyses of gRNA expression from minicircles with a set of high confidence gRNA classes where both DNA template and RNA transcript are known.

Mapping AnTat90.13 small RNA reads directly to the assembled

AnTat90.13 minicircles resulted in the majority of canonical and non-canonical gRNAs being covered at read depth >5 (Table 3.8). This was not the case for small RNA data sets from other sources, in fact the overlap was surprisingly low given that despite their different sources the gRNAs must carry out the same function. The data set with the highest coverage of AnTat90.13 minicircles was the EATRO164 PCF data set from (Koslowsky et al., 2014), in which only a third of the annotated gRNAs genes were covered. This is possibly evidence for the fast evolution of minicircles.

As reported in other studies, we found that both the sense and anti-sense strands of minicircles are transcribed, and at steady state the anti-sense transcripts are stable, however the canonical anti-sense transcripts are apparently more abundant than the non-canonical anti-sense transcripts. In fact, plotting the ratio of sense to anti-sense transcripts for both the canonical set and the non-canonical set showed that canonical gRNAs have an average sense to anti-sense transcript ratio of ten to one compared to the non-canonical gRNAs where the mean ratio is about 100 to one (Figure 3.19). The implication of how the ratio of sense to anti-sense transcripts relates to and perhaps affects the stability of mature gRNAs is unclear. What is clear, however, is that if both the sense and anti-sense transcripts are bi-directionally transcribed, initially resulting in equal numbers of precursor sense and anti-sense gRNAs (as described in (Suematsu et al. 2016)), then the non-canonical anti-sense transcripts are degraded more readily than the non-canonical anti-sense transcripts. It is not difficult to imagine a mechanism whereby functional gRNA duplexes which have a match in the editing space are sequestered into the editosome, protecting both sense and anti-sense transcripts from degradation, whilst non-canonical (potentially non-functional) gRNAs are not sequestered. However would require a portion of the sense side of the duplex to be exposed to allow probing of the pre-edited transcript. This would allow sequential

degradation of the non-canonical duplexes, first the anti-sense transcripts (potentially due to their shorter U-tails (Figure 3.21)) followed by the sense transcripts (partially protected from degradation by longer U-tails). Other than the ratio of sense to anti-sense transcripts the non-canonical gRNAs are similar to the canonical set in almost every other way, including U-tail length (Figure 3.24), dominant anchor motif and length. This implies that despite the fact that canonical gRNAs have a more equal stoichiometry of sense to anti-sense reads they still share the same upstream gRNA biogenesis pathways, involving bidirectional transcription and duplex trimming followed by re-uridylation. The mechanism by which the sense transcript is uridylylated more extensively than the anti-sense transcript remains a puzzle, or perhaps it's possible that the anti-sense transcripts are more prone to 3' exonuclease activity such as in miRNA processing which undergo asymmetrical degradation. How this data fits into the overall picture of gRNA processing (as reviewed in (Aphasizhev & Aphasizheva 2011)) is unclear. It also remains to be demonstrated that the anti-sense transcripts we and others have identified actually do form duplexes with gRNAs. They may be degraded RNAs left over from the bi-directional transcription process. However, they appear to be stable at steady state and have a consistent size, implying they are not in the process of being degraded. *In vivo* RNA cross linking experiments (such as in (Helwak et al., 2013)) have the capacity to answer many of the open questions surrounding the potential role of gRNA/antisense-gRNA duplexes in gRNA transcription and processing as well as the formation of gRNA/mRNA duplexes during the course of editing.

The mechanism by which stage specific editing of mitochondrially encoded RNAs is regulated is not clear. Hypotheses that it may be controlled by gRNA abundance have been questioned by northern blot analyses that showed gRNAs for



mRNAs with stage-specific editing to be present in both life cycle stages (Koslowsky et al. 1992). Kirby et al. (2016) have carried out the first NGS approach to answering these questions. They found that the abundance of gRNAs was greater in their PCF samples than in their BSF samples whilst the diversity remained the same; they also suggested that the initiator gRNAs are perhaps involved in differential editing. Our data does not support this: we found that the number of stably expressed gRNA classes was generally higher for the PCF than for the BSF (Figure 3.17). This trend holds true for the TREU 667 pair of samples also (initial mapping of the pair of data sets from Kirby et al. (2016) to AnTat90.13 minicircles shows the same trend (data not shown)). Initiator gRNAs were not quantified in this study. Analyses of the proportions of gRNAs that map to each of the edited transcripts suggested that there is a mild increase in the abundance of RPS12 gRNAs in the PCF vs BSF samples (Figure 3.18) this can also be seen in the gRNA depth profiles. Inspection of the plots in Figure 6.15 shows that the depth profile for the PCF vs the BSF samples for AnTat90.13 (A vs B) and TREU667 (C vs D) is different for the RPS12 transcript (inspection of depth plots for other transcripts shows a similar profile). All three of these read sets effectively represent biological replicates, however they do not agree with each other. The lack of replicates for each of these samples means that differences in sequencing preparation cannot be ruled out and true quantitative analyses cannot be made. A further caveat of the data presented in Kirby et al. (2016) is that the BSF and PCF samples have been cultured independently for ~30 years; additionally the BSF was grown (intermittently) in culture for ~10 years before a PCF equivalent was generated. Given the data presented in other studies (Simpson et al. 2015; Greif et al. 2015) and in Section 4 of this thesis it is highly likely that the PCF and BSF samples in Kirby et al. (2016) will have different patterns of editing as a result of diversification of minicircles due to

evolution and render any direct comparison of PCF to BSF samples moot. Taking our data and Kirby et al. (2016) at face value it appears that if there is a gRNA component to patterns of differential editing it is modest as no striking differences between the data sets have been observed.

### 3.4.3 Comparative kDNA complexity

Other trypanosomatids have long been suspected of having much reduced minicircle and gRNA complexity compared to *T. b. brucei* (Englund, 1979; Thomas et al., 2007), this was recently confirmed for *L. tarentolae* where a full assembly of minicircles (Simpson et al. 2015) identified only 114 classes. This comparatively and functionally excessive complexity of minicircles in *T. b. brucei* is perplexing and it seems reasonable to assume that it is related to the life-cycle of salivarian trypanosomes. However, the complexity of other salivarian trypanosomes, for example *T. b. gambiense*, *T. b. rhodesiense*, *T. vivax* and *T. congolense* has not been explored comprehensively (although there is some information available with regards to minicircle size and organisation (Borst et al., 1985)). A recent study showed low kDNA complexity in two related American strains of *T. vivax* (which probably left the tsetse belt ~150 years ago) (Greif et al., 2015) (~50 minicircle sequence classes were identified), however a comparative analysis to the number of minicircle in an African isolate was not made. This finding fitted the expectation that the strains which still proliferate in the insect vector and thus presumably require a functional electron transport chain (Jackson et al., 2015) will have a more complex kDNA genome. Comparative analysis of the minicircle complexity found in other strains of *T. brucei* (Section 3.3.7.1) shows that strains which are reported to be able to complete the canonical life-cycle through the tsetse are predicted to have almost twice the

complexity of ‘monomorphic’ strains that had been extensively cultured as BSF and had lost the ability for efficient differentiation (Figure 3.29). This taken together with preliminary data presented in this thesis that suggest limited minicircle complexity for *T. b. gambiense* Type I (Appendix 6.5) which, based on phylogenetic analyses, are clonal (Morrison et al., 2008) (and thought to be accumulating deleterious mutations as a result) supports the theory that the excessive kDNA complexity of *T. brucei* is related to the salivarian life-cycle and, at least in the long-term, requires genetic exchange in the tsetse (Gibson et al., 1997) for replenishment of minicircle complexity. The *T. b. gambiense* samples analysed in this study are predicted to have a complexity of just ~100 minicircles, whether this is a number typical of *T. b. gambiense* or is a result of an asexual life cycle has yet to be determined and could be addressed by analysing samples which have been placed as diverging earlier in the lineage of asexually propagating parasites. This relatively high minicircle complexity in *T. brucei* is probably a reflection of the amount of redundancy required to ensure that lethal gaps in the editing capacity do not appear given the expectation that the complexity of the kDNA genome is lost over time.

#### 3.4.4 gRNA and minicircle conservation

Having multiple data sets of both small RNAs representing gRNAs and minicircles allows a comparative analyses of the overlap between various strains. The overlap between gRNA data sets is very low (Table 3.11, Figure 3.26). This is also seen by Kirby et al. (2016) when they compare the gRNA overlap between two life-cycle stages of the same strain (which have proliferated in culture independently for an undefined number of generations). Likewise, the the overlap of minicircle sequence classes (Table 3.11 and Figure 3.30) is very low again highlighting how different the

complement of minicircles is among various strains of *T. brucei*. This is likely a result of minicircle sequences evolving quickly or changing sequence as a result of recombination events.

#### 3.4.5 Modelling gRNA distributions amongst minicircles

Having a complete set of annotated minicircles allowed us to ask questions about how complexity in the mitochondrial genome is maintained. It has been speculated that part of the reason that the kDNA genome exists in this segregated form is to intrinsically link the portions of the genome that are not required in the current life-cycle stage to portions of the genome that are essential for survival (Speijer, 2006). If true then one might expect that the genes essential for bloodstream form survival (A6 and RPS12) (Dean et al., 2013) would more likely be found on separate minicircles, thereby increasing the coupling of essential gRNAs to non-essential gRNAs. That is, we might expect the chance of finding a minicircle with multiple A6 and RPS12 gRNAs is lower than would be if gRNAs were distributed simply by chance.

This hypothesis was investigated using a modelling approach (Section 3.3.6) whereby the distribution of A6 and RPS12 gRNAs of AnTat90.13 was compared to a distribution in which the identified A6 and RPS12 gRNA genes were randomly distributed between minicircles. This modelling approach suggested that these gRNAs are distributed randomly. The above analysis was also performed with minicircle and gRNA data from a monomorphic cell line, with the same result (data not shown). Thus, our approach did not provide supporting evidence for Speijer's proposal. More studies are required, however, to substantiate evidence for or against this hypothesis. For example, it would be interesting to develop a mathematical model that compares the likelihood of gRNA loss for minicircle populations that have one vs. multiple gRNA

genes per minicircle. The next chapter describes a mathematical approach to kDNA dynamics that might facilitate such studies in the future.

## 4. An experimental and *in silico* approach to analysing kDNA segregation

### 4.1 Summary of research question and approach

The fidelity with which the kDNA genome is segregated in *T. brucei* has not been determined. Several authors have suggested that random segregation of replicated minicircles into daughter cells is responsible for the temporal variations in minicircle copy number that have been observed (Maslov and Simpson, 1992; Thiemann et al., 1994). The evolution of the kDNA network structure is thought to be perhaps related to improving on the random segregation of minicircles which would otherwise be free in the mitochondrial matrix (Lukes et al., 2002; Jensen and Englund, 2012). It has also been suggested that asymmetric kDNA division, whereby segregation of the genome actively favours one daughter over another, is the mechanism by which Ak/Dk parasites such as *T. evansi* and *T. equiperdum* have arisen in the wild (Lun et al., 2010). The loss of selective pressure to maintain kDNA complexity can also quickly result in a homogenised kDNA repertoire, as observed in Dk strains (Lai et al., 2008). This suggests that the presence of a kDNA network alone is not sufficient to maintain complexity over many generations (Simpson et al., 2000). The role of genetic exchange of minicircles in the tsetse fly has been proposed as essential to replenish kDNA heterogeneity (Gibson et al., 1997; Savill and Higgs, 1999).

The consequences of the random segregation hypothesis for kDNA composition, its temporal evolution and maintenance of complexity was studied using computer simulations (Savill and Higgs, 1999). It was found that this hypothesis was consistent with previously available data from *L. tarentolae*, which showed that lab adapted strains which had been grown in culture for long periods of time had lower

complexity than a more recently isolated strain (Gao et al., 2001). Random segregation of minicircles is able to explain several phenomena, including a small number of major and many minor classes of minicircle, the temporal fluctuations in minicircle copy number over short time periods and the loss of low copy number minicircles. The hypothesis predicted that over long time scales the average network size grew, cell viability is increased with larger networks and that redundant gRNAs will not persist in species with only one gRNA per minicircle.

The effects of varying the fidelity by which minicircles are inherited was also studied by the authors. They found that cell viability (as defined by loss of a minicircle class or from having too large a network) is reduced if kDNA is segregated asymmetrically, with one daughter cell receiving proportionately more minicircles than the other. They also found that deterministic segregation increases cell viability as the entire repertoire of genetic information is transmitted exactly at every generation. Evidence from studies which have previously been mentioned (Thiemann et al., 1994; Gao et al., 2001; Greif et al., 2015), in which it is observed that kDNA complexity can be lost strongly suggests that kDNA inheritance is not deterministic. Thus it is likely that trypanosome kDNA segregation lies somewhere on the spectrum of random minicircle segregation to deterministic.

This leads to many open questions pertaining to minicircle replication and segregation. Is deterministic kDNA segregation an expensive process or mechanistically impossible for trypanosomes to evolve? What mechanisms do trypanosomes use to depart from random segregation? Where on the segregation spectrum between random and deterministic do trypanosomes lie? The aim of the experiments detailed in this chapter are to finally answer this long standing question of how random the segregation process truly is and, having established that parameter, to,

produce a refined model of minicircle segregation (Savill and Higgs, 1999). The refined model will be able to make more accurate predictive estimates for how complexity changes over time and will even allow us to test the various hypotheses that have been proposed for why the complicated network structure evolved (Borst, 1991; Lukes et al., 2002).

To parametrise the model an *in vitro* evolution study approach was developed to measure kDNA complexity over time. A pilot study was initially carried out with two sets of parallel cultures of *T. brucei* AnTat90.13 (four samples in total). The pilot study had the following purposes i) a proof of concept that we can indeed observe variations in kDNA complexity *in vitro*, ii) to establish that sequencing purified kDNA would give sufficient read coverage for analysis, iii) to establish a practical time frame for measuring meaningful changes in kDNA complexity, iv) to investigate the effects of a mutated ATP synthase  $\gamma$  on kDNA complexity over time. Insight from the pilot study was then used to design a second longer time course experiment that was carried out for a period of 29 weeks, with kDNA samples from the start and end points sequenced.

## 4.2 Materials and Methods

### 4.2.1 Cell line generation and culturing protocol

Two transgenic cell lines based on differentiation competent *T. b. brucei* BSF strain Antat 1.1 90.13 (*TbAnTat*) (Engstler and Boshart, 2004) were used. One cell line, *TbAnTatL262P*, had one endogenous ATP synthase subunit  $\gamma$  allele replaced with an L262P mutant version (C. Dewar *et al.*, manuscript in preparation), dispensing with the requirement for kDNA in these cells (Dean et al., 2013a). An otherwise isogenic cell line, *TbAnTatWT*, expressed two wild type alleles of this gene and was therefore fully



dependent on expression of the mitochondrially encoded A6 and RPS12 genes. Both cell lines were competent for efficient differentiation into stumpy and procyclic forms (C. Dewar *et al.*, manuscript in preparation). Cells were maintained in HMI-9 medium (Hirumi & Hirumi 1989) supplemented with 10% FBS at a cell density no greater than  $10^6$  cells per ml.

#### 4.2.2 Time course culturing protocol

##### 4.2.2.1 Pilot time course

*TbAnTatWT* and *TbAnTatL262P* cells were cultured as described in (section 3.2.1) for 12 weeks in duplicate (replicates A and B). On a weekly basis 5-ml cultures were diluted 1:100 on Mondays, Wednesdays and Fridays. In two week intervals cultures were seeded into large culture vessels (300 ml of HMI-9) on Friday at a density that allowed growth to a total cell number of  $\sim 3 \times 10^8$  by the following Monday and harvested for kDNA purification the (section 3.2.2). Samples were numbered  $T_S1$ - $T_S6$  ( $T_S$  = short time course) according to which time point they were harvested at (Table 4.1). Four kDNA samples from early and late time points were selected for deep sequencing ( $T_S1WTA$ ,  $T_S1L262PB$ ,  $T_S6WTA$  and  $T_S6WTB$ , corresponding to weeks 2 and 12, respectively). All comparisons were made to reference sample  $T_{0WT}$  as defined in section 3.3.2, in which we assembled 365 minicircles. The duration of the experiment equated to approximately 250 generations. Generation time was taken to be approximately eight hours (based on growth curves from WT *AnTat90.13* cells carried out by others in the laboratory).

##### 4.2.2.2 Long term time course

A similar but longer time course ( $T_L$ ) as in 4.2.2.1 was carried out with *TbAnTatWT* cells in triplicate. Cells were cultured essentially as described in 4.2.2.1, but for a

period of 29 weeks. Every Friday cultures were transferred into large culture vessels (300 ml of HMI-9) and  $\sim 3 \times 10^8$  cells harvested for kDNA purification on the following Monday (see Table 6.2 for complete list of samples). Four kDNA samples from early and late time points were selected for deep sequencing ( $T_L0$ ,  $T_L28B$ ,  $T_L28C$  and  $T_L29A$ ;  $T0$  = time point zero; A, B, C correspond to the three replicates).

#### 4.2.3 kDNA purification

Kinetoplast DNA was essentially purified as described (Fairlamb et al., 1978; Pérez-Morga and Englund, 1993) and in detailed in section 3.2.2 with helpful advice from Michele Klingbeil. Pellets generated from the ethanol precipitation were dissolved in 15  $\mu$ l of  $H_2O$ .

#### 4.2.4 Restriction digest

Purified kDNA samples were digested with EcoRI restriction enzyme (Promega) in order to assess purity of the kDNA isolation process and ensure that sufficient material had been recovered for sequencing. 2.5  $\mu$ l of sample (the exact concentration was difficult to quantify due to the network nature of the kDNA) were combined with 2 units of EcoRI, 1  $\mu$ l of buffer H (Promega), 0.1  $\mu$ l of BSA (10 mg/ml) and volume adjusted to 10  $\mu$ l with  $H_2O$ . Reaction mixtures were incubated at 37°C for  $\sim 3$  hours. The entire 10  $\mu$ l sample was resolved on a Tris-borate EDTA (TBE) 0.8% (w/v) agarose gel (0.5  $\mu$ g/ml ethidium bromide) and visualised under ultraviolet light.

#### 4.2.5 Sequencing of kDNA

Sequencing of kDNA was carried out at Edinburgh Genomics. Sample DNA was fragmented to generate  $\sim 550$ -bp inserts and sequenced on an Illumina MiSeq to

generate 300-bp paired end reads or 100 bp paired end reads on an Illumina HiSeq for the long term time course. Reads were quality checked, adapter trimmed and quality trimmed using FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) and cutadapt (Martin, 2011), bases below Q15 were removed.

#### 4.2.6 Sequence analysis

Reads were mapped using bowtie2 (Langmead & Salzberg 2012), with default mapping parameters. Reads were mapped to minicircles assembled as described in chapter 3, the published *T. b. brucei* Lister 427 maxicircle (Sloof et al., 1992) (Accession: M94286) and the nuclear genome (*T. b. brucei* Lister 427, version 4 obtained from [www.genedb.org](http://www.genedb.org)) for which only the core chromosomal regions were mapped. Core regions were extracted using a table (Appendix 6.7, Table 6.3) kindly provided by Bernardo Foth (Wellcome Trust Sanger Institute, Hinxton).

#### 4.2.7 Modelling minicircle replication

The mathematical model of minicircle segregation used in Savill & Higgs (1999) was used as the basis for our model. We extended it as follows.

- (i) Population expansion and bottlenecking that matched our experimental protocol.
- (ii) A variable segregation probability  $p$ . This parameter is the probability that two sibling minicircles segregate into different daughter cells. If  $p = 0$ , both sibling minicircles always segregate into the same daughter cell. If  $p = 1$ , both sibling minicircles always segregate into different daughter cells. Setting  $p = 0.5$  recovers the random segregation model of Savill and Higgs 1999.

The model is defined as follows. Let  $C$  be the number of minicircle classes in a

population of cells. A cell in this population has  $M_i$  copies of minicircle class  $i$ . The total network size ( $T$ ) for this cell is the sum of all class copy numbers, i.e.:

$$T = \sum_{i=1}^C M_i.$$

On each generation each minicircle is replicated (see Introduction for how this occurs) resulting in a doubling of each class copy number, i.e,  $M_i \rightarrow 2M_i$ . The duplicated network must then be segregated into the two daughter cells, one daughter will receive  $m_i$  copies of class  $i$  the other will receive the remainder  $m'_i$  such that  $m_i + m'_i = 2M_i$ . There are several steps to calculating  $m_i$  and  $m'_i$ . We identify the two siblings as 1<sup>st</sup> and 2<sup>nd</sup>, and the two daughter cells they enter as A and B. The 1<sup>st</sup> sibling has a 50% chance of entering cell A and a 50% chance of entering cell B. There are  $M_i$  1<sup>st</sup> siblings and  $M_i$  2<sup>nd</sup> siblings. Therefore the number  $X$ , of 1<sup>st</sup> siblings entering cell A is binomially distributed with parameters  $M_i$  and 0.5. Let us say that  $X=x$  1<sup>st</sup> siblings entered cell A, and, therefore,  $M_i-x$  1<sup>st</sup> siblings entered cell B. If a 1<sup>st</sup> sibling entered cell A, the probability of the 2<sup>nd</sup> sibling entering cell A is  $1-p$ . Given that  $x$  1<sup>st</sup> siblings entered cell A, the number  $Y$ , of  $x$  2<sup>nd</sup> siblings entering cell A is binomially distributed with parameters  $x$  and  $1-p$ . If a 1<sup>st</sup> sibling entered cell B, the probability of the 2<sup>nd</sup> sibling entering cell A is  $p$ . Given that  $M_i-x$  1<sup>st</sup> siblings entered cell B, the number  $Z$ , of  $M_i-x$  2<sup>nd</sup> siblings entering cell A is binomially distributed with parameters  $M_i-x$  and  $p$ . Let  $Y=y$ , and  $Z=z$ , 2<sup>nd</sup> siblings entered cell A. Therefore the total number of minicircles of class  $i$  entering cell A is  $m_i=x+y+x$ , and the total number of minicircles entering cell B is  $m'_i=2M_i-(x+y+z)$ .

A bottlenecking function was added to the model to simulate the splitting of cells in the culturing protocol. The bottleneck function randomly selects a proportion of viable cells from the population (1/100 or 1/1000 when the number of generations that

have been run is a multiple of three), and carried them forward for subsequent division cycles. Cells that are not selected for propagation are deleted.

## 4.3 Results

### 4.3.1 Pilot time-course

A pilot time course study was carried out with two replicates each of cultures of *TbAnTatWT* and *TbAnTatL262P* (4 samples in total). The purpose of the pilot study was four-fold. i) As a proof of concept that we can observe variations in kDNA complexity over time *in vitro*; ii) to establish that the protocol for purification and sequencing of kDNA would give sufficient read coverage for analysis; iii) to establish a practical time frame for measuring meaningful changes in kDNA complexity; iv) to investigate the effects of complete kDNA-independence on kDNA complexity over time.

#### 4.3.1.1 Sample QC and sequencing

Restriction digest of kDNA purified samples has been shown to liberate the majority of minicircle classes from the kDNA network and generate a linearised minicircle band at 1 kb (Fairlamb et al., 1978). This approach can be used as a quality control measure for the kDNA purification process. A pure kDNA prep digested with EcoRI would be expected to have some of the undigested network remaining in the well, some higher molecular weight fragments (4.6 kb, 6.6 kb and 11 kb as calculated by performing an *in silico* digest using the AnTat90.13 maxicircle sequence assembled in 3.2.4.1) corresponding to digested maxicircle and linearised minicircle at 1 kb. Figure 4.1

shows an EcoRI digest for each of the fortnightly time points for this pilot study. For most samples the expected banding pattern for an EcoRI digest failed and was not repeated due to limitations in the amount of purified kDNA available for each time point. Products consistent with EcoRI-digested kDNA were observed for *TbAnTatWT* samples T<sub>s</sub>1WTA, T<sub>s</sub>5WTA, T<sub>s</sub>5WTB, T<sub>s</sub>6WTA and T<sub>s</sub>6WTB (lanes 1, 17, 18, 21 and 22). They show linearised minicircles at ~1 kb and a trio of higher molecular weight bands (6-10 kb) which are likely to be digested maxicircle as the sum of the sizes of these bands is approximately 22 kb. The boxed samples were sequenced and kDNA complexity analysed via NGS (see section 3.2.4.2). The presence of DNA in the wells for the remaining samples suggested that kDNA had been isolated but was resistant to digestion with EcoRI, perhaps because of contaminating inhibitors or because the DNA could not be dissolved after precipitation. Nonetheless it could be used as a qualitative proxy measure for the kDNA content of a cell line at a given time point. The boxed samples were sequenced and kDNA complexity analysed via NGS.

Interestingly, for the *TbAnTatL262P* samples (indicated by red numbers) presence of kDNA could only be confirmed up to week 2 (sample T<sub>s</sub>1L262PA in lane 3; specific EcoRI products were partially obscured by contaminating RNA or genomic DNA) or week 4 (sample T<sub>s</sub>2L262PB in lane 8). One two-week *TbAnTatL262P* sample was analysed by NGS (see section 3.2.4.2). For later time points, no specific bands or DNA in the well could be observed, suggesting that these kDNA-independent cells had lost their kDNA wholesale after a short culturing period. This was confirmed by analysis of DAPI (4',6-diamidino-2-phenylindole) stained cells via fluorescence microscopy (data not shown).

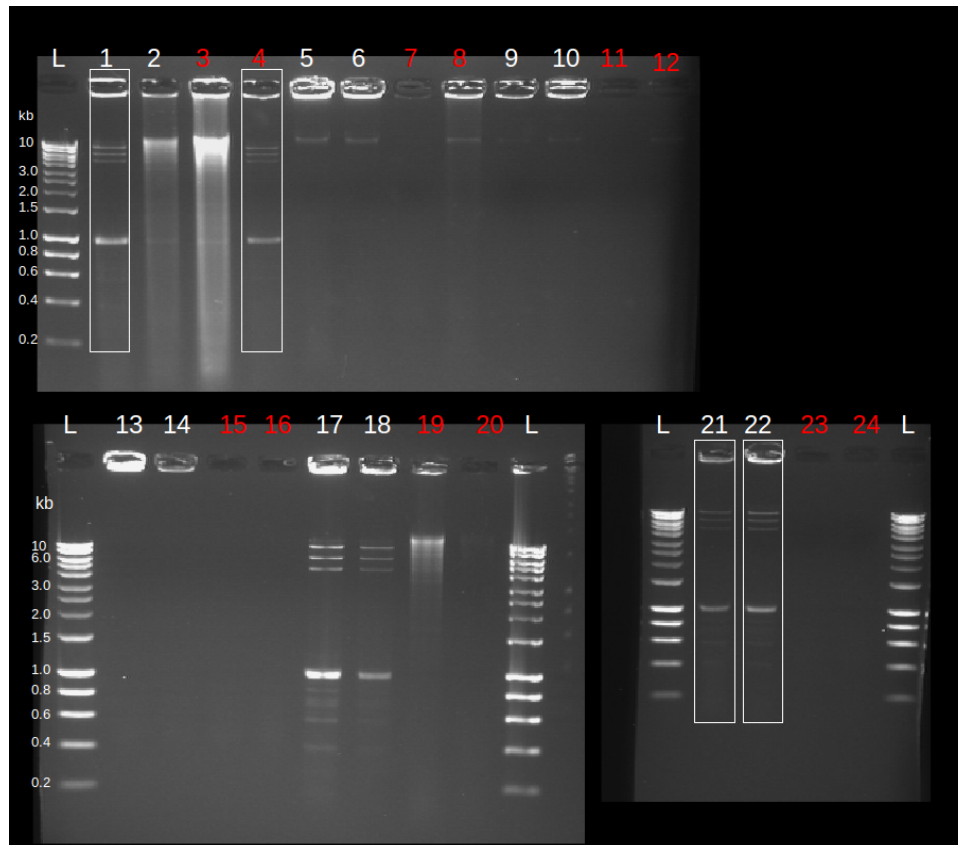


Figure 4.1 EcorI digested kDNA samples. L indicates hyper-ladder 1 (Bioline). Kb:kilo base pairs. Samples labelled in red are L262Py mutants. Samples loaded are as follows: 1:T<sub>s</sub>1WTA, 2:T<sub>s</sub>1WTB, 3:T<sub>s</sub>1L262PA, 4:T<sub>s</sub>1L262PB, 5:T<sub>s</sub>2WTA, 6:T<sub>s</sub>2WTB, 7:T<sub>s</sub>2L262PB, 8:T<sub>s</sub>2L262PA, 9:T<sub>2</sub>3WTA, 10:T<sub>2</sub>3WTB, 11:T<sub>2</sub>3L262PA, 12:T<sub>2</sub>3L262PB, 13:T<sub>2</sub> 4WTA, 14:T<sub>2</sub> 4WTB, 15:T<sub>2</sub> 4L262PA, 16:T<sub>2</sub> 4L262PB, 17:T<sub>2</sub>5WTA, 18:T<sub>2</sub>5WTB, 19: T<sub>2</sub>5L262PA, 20:T<sub>2</sub>5L262PB, 21: T<sub>2</sub>6WTA, 22: T<sub>2</sub>6WTB, 23:T<sub>2</sub>6L262PA, 24: T<sub>2</sub>6L262PB. Boxed samples were sequenced. Each set of 4 samples represents 2 weeks culturing, e.g. 1,2,3,4=time point 1.

#### 4.3.1.2 Minicircle loss in WT *T. brucei* within ~300 generations of *in vitro* culturing

Four samples from the short pilot time course were selected for NGS analysis (Table 4.1). Mapping kDNA samples from the two-week time points (T<sub>s</sub>1WTA and T<sub>s</sub>1L262PB) to the reference set of minicircles (T<sub>0</sub>WT) as generated in Section 3.3.2 resulted in full coverage of all assembled minicircles. The sequencing coverage of the kDNA can be estimated by checking the average depth of reads mapped to the maxicircle. The average depth of maxicircle coverage for all four samples ranged from

~10,000 to ~20,000 and was roughly proportional to the number of reads obtained for each sample (Table 4.1). For example, the T<sub>s</sub>1WTA sample produced 6.2 million reads and had a per nucleotide maxicircle depth of 19,601, giving an approximate #reads to maxicircle depth ratio of 300:1; this ratio was similar for all samples. This also indicated that the number of maxicircles was relatively stable during the ~250 generations (12 week) time course. Taking the number of maxicircles to be constant at 30 copies per network (as calculated in section 3.3.3) we can predict the kDNA coverage to be in the range of 333X-633X for a single copy molecule. Hence, a hypothetical minicircle present as a single copy molecule in 10% of the cells in a population would still be detected by more than 30 reads per nucleotide on average.

Sample	Approximate number of generations	Total number of reads (300 bp, paired end)	% of reads mapped to		
			minicircles	maxicircles (per nucleotide depth)	nuclear genome
T <sub>s</sub> 1WTA	42	6235825	84.07	10.05 (19601)	1.09
T <sub>s</sub> 1L262PB	42	4468400	70.27	11.04 (15152)	1.53
T <sub>s</sub> 6WTA	252	3302690	84.59	10.91 (10790)	0.56
T <sub>s</sub> 6WTB	252	3179305	71.47	11.05 (10519)	0.70

*Table 4.1:* Read mapping data for the pilot kDNA complexity time course. Percentage of reads mapped to reference minicircles and the published maxicircle. The maxicircle depth is the per nucleotide average.

Minicircles were considered to be present in the population if more than 60% of the minicircle was covered at a read depth greater than or equal to one. The 60% of the minicircle coverage cut-off was chosen to account for non-specific mapping of reads in the conserved region: if 300 bp paired end reads (with a ~550 bp insert) which map non-specifically to the conserved region of the minicircle are extended into variable region (through virtue of two minicircle classes being similar) they can cover up to



40% of the minicircle giving a false positive. However, altering this parameter to make it more or less sensitive to the detection of low copy had negligible effect on the number of minicircles counted as being lost (data not shown). Using this criterion samples T<sub>s</sub>6WTA and T<sub>s</sub>6WTB were found to have lost 5 and 25 of the original 365 minicircle classes from their populations, respectively (Table 4.2). A coverage plot for a representative minicircle, mO\_@261, which was deemed to have been lost from the cell population in sample T<sub>s</sub>6WTA but not in sample T<sub>s</sub>6WTB is shown in figure 4.2, a minicircle which is still present in all samples is shown for comparison. Although scored as 'present' in the latter sample according to the criteria discussed above, compared to the reference set and the earlier time points coverage was dramatically reduced in that sample as well. Close inspection of the single read that has mapped to [mO\\_@261](#) in figure 4.2 for sample T<sub>s</sub>6WTA shows that this read contains many mismatches and is probably non-specific read mapping. Four minicircles that were lost from both samples were all low copy number when measured at T0WT (Table 4.2). The average copy number for these lost minicircles at time point zero was 7.2 copies per cell, the average copy number for any given minicircle was 23 (as shown in section 3.3.3 figure 3.4). It should be noted however that the majority of minicircle classes have an average copy number between zero and twenty as outlined in an earlier chapter. In T<sub>s</sub>6WTA and T<sub>s</sub>6WTB the loss of gRNA classes as a result of minicircle loss would have no immediate effect on editing capacity of the cells as all these classes cover areas with redundancy. The loss of minicircles also appeared to be random with respect to the mRNAs affected: gRNA classes involved in editing of A6 and RPS12 (the two kDNA encoded mRNAs essential for BSF survival) were among those lost, along with gRNA classes involved in editing of transcripts that are only essential in the PCF insect stage. The number of gRNAs lost for any given mRNA was roughly

proportional to the total number of distinct gRNAs identified for that mRNA, with no apparent bias against gRNAs that are essential in the BSF stage (Table 4.2), although this has not been tested statistically.

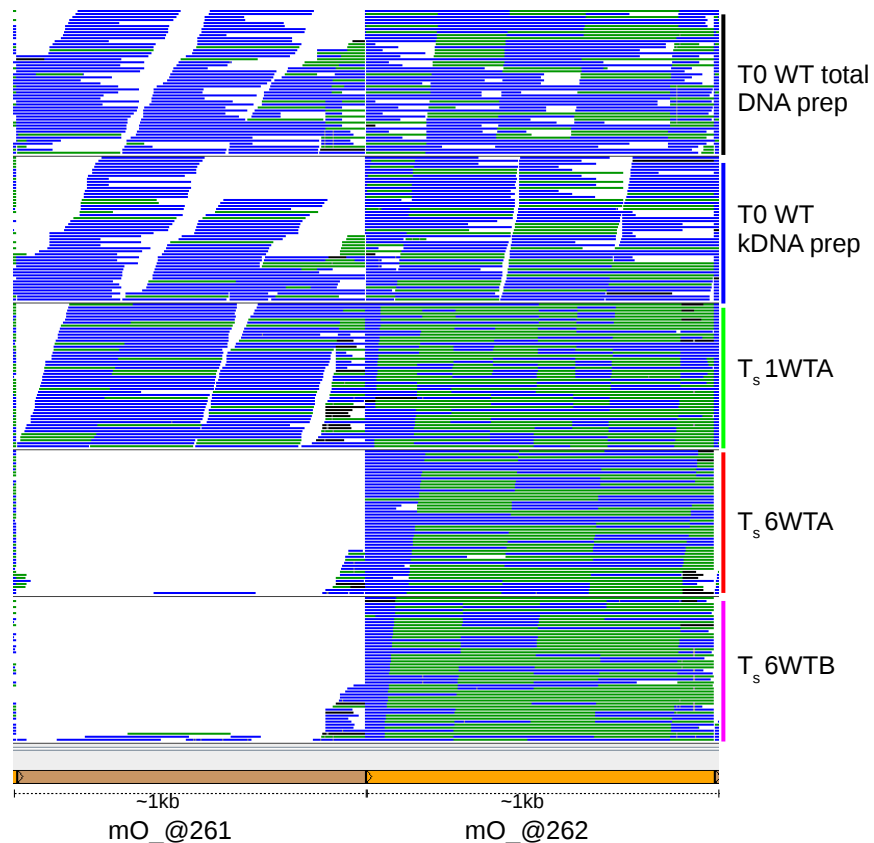


Figure 4.2: Minicircle loss during *in vitro* culturing of a WT strain of *AnTat90.13*. Reads directly mapped to assembled minicircles, Green bars indicate multiple reads with identical start and stop positions, Blue bars indicate reads with unique mapping positions. Based on our minimum coverage criteria (read depth  $\geq 1$  over  $>60\%$  of the contig length) minicircle [mO\\_@261](#) is considered to be lost in sample A but not sample B. Minicircle [mO\\_@262](#) is present in all six samples.

Lost from	Minicircle ID	Copy number at T0
Both samples	mO_@280	5.09
	mO_@321	7.73
	mO_@332	3.62
	mO_@339	2.59
T6 <sub>s</sub> WTB only	mO_@1	5.77
	mO_@112	8.02
	mO_@182	3.89
	mO_@239	5.68
	mO_@241	7.79
	mO_@264	15.08
	mO_@287	7.07
	mO_@304	4.04
	mO_@313	5.27
	mO_@323	3.93
	mO_@325	14.79
	mO_@330	2.94
	mO_@343	7.43
	mO_@37	9.04
	mO_@45	15.06
	mO_@66	10.59
	mO_@67	14.21
	mO_@7	5.39
	mO_@83	5.13
	mO_@86	9.00
mO_@94	6.73	
T6 <sub>s</sub> WTA only	mO_@261	1.98

*Table 4.2: Copy number at T0WT for minicircles lost at T6 for replicates A and B. The average copy number for these minicircles at T0 is 7.2, SD=3.9.*

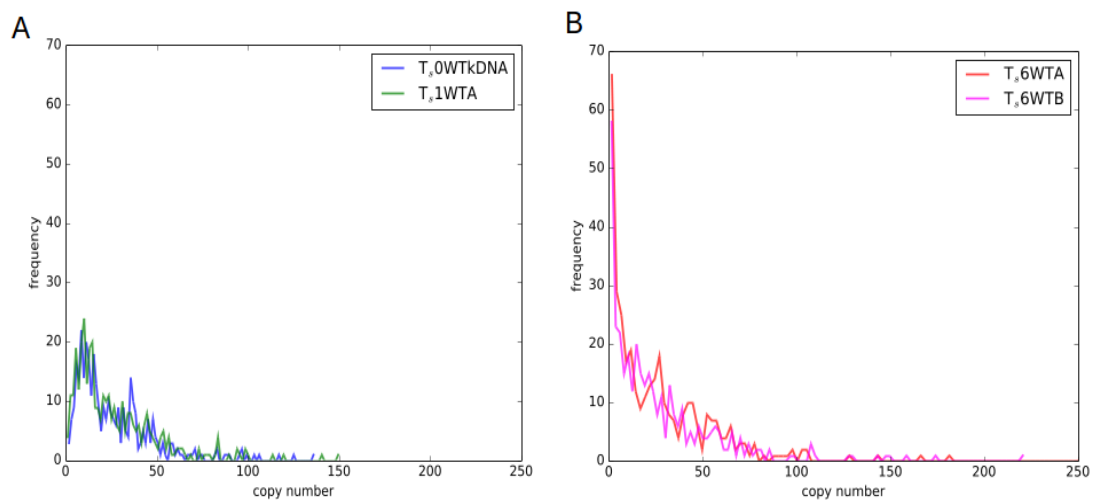
Transcript gRNA matches to	Number of gRNAs annotated (Percentage of total gRNAs annotated)	number of gRNAs lost (Percentage of lost gRNAs)	
		T <sub>s</sub> 6WTA	T <sub>s</sub> 6WTB
a6	48 (4.7)	2 (12.5)	3 (4.3)
co3	96 (9.3)	-	4 (5.7)
cr3	14 (1.4)	-	3 (4.3)
cr4	42 (4.1)	1 (6.25)	5 (7.14)
cyb	1 (0.1)	-	-
murf2	4 (0.4)	-	-
nd3	11 (1.1)	-	1 (1.4)
nd7	74 (7.2)	-	4 (5.7)
nd8	36 (3.5)	-	2 (2.8)
nd9	29 (2.8)	2 (12.5)	3 (4.3)
rps12	17 (1.7)	-	3 (4.3)
Predicted by nucleotide bias only	655 (63.7)	11 (68.75)	42 (60.0)
<b>Total gRNAs</b>	<b>1027</b>	<b>16</b>	<b>70</b>

Table 4.3: Proportions of gRNAs from minicircles that have been lost compared to the total number of gRNAs annotated to the reference minicircles.

#### 4.3.1.3 Minicircle copy number distributions change during *in vitro* culture

It was of interest to determine if the relative abundance of minicircle classes changed over the course of the pilot study. The depletion of nuclear DNA in the kDNA preparations meant that, in contrast to NGS analysis of total DNA extractions, an internal standard for determining the average number of minicircles and maxicircles per network was not available. Nonetheless, the relatively constant ratio of minicircle to maxicircle reads (Table 4.1; see also Table 4.4 below) suggested that the number of minicircle molecules per network did not change significantly over the course of the time course. Thus, assuming a constant average total number of 9695 minicircles per network (as calculated for the T0WT reference in Chapter 3.3.3), average copy numbers for each minicircle class were calculated by multiplying the proportion of

reads mapped to that class with 9695 and rounded to the nearest whole number. For each copy number, the cumulative frequency of minicircle classes per network with that copy number was determined and plotted to give a graphical representation of the copy number distribution (Figure 4.3). Copy number distributions were found to be markedly different between the early and late time-points (Figure 4.3).  $T_s1WTA$  was found to have distribution similar to that which was calculated from the reference samples ( $T_{0WTkDNA}$  and  $T_{0WTtotal}$ ). The late samples  $T_s6WTA$  and  $T_s6WTB$  after  $\sim 250$  generations of *in vitro* culture have a much more exaggerated distribution with the frequency of very minor classes being extremely high and only a few major classes.



*Figure 4.3: Minicircle copy number distributions for reference set and an early time point (A) and for late time points (B) in the pilot time course. Copy numbers were estimated by multiplying the total number of minicircles (9695 as calculated in Chapter 2) by the proportion of reads mapped per minicircle. A: Minicircle copy number distribution for the reference set (blue line) and sample  $T_s1WTA$  (green line). B: Minicircle copy number distribution for late time points  $T_s6$ .*

#### 4.3.1.4 The kDNA-independent cells lost kDNA rapidly under bottlenecked culturing conditions

The effect of kDNA-independence on maintenance of its complexity was also investigated. It was not possible to sequence late time points of *TbAnTatL262P* cells (kDNA-independent due to expression of an ATPase subunit  $\gamma$  with the L262P mutation; Dean et al., 2013) as these cells lost their kDNA spontaneously within four weeks of culturing (see Figure 4.1). Samples from two early time points in the culturing history of these cells were sequenced however; T0L262P, prepared directly (i.e.  $\sim$ 2 weeks, or  $\sim$ 40 generations) after clonal selection of L262P subunit  $\gamma$  transfectants, and T<sub>5</sub>1L262PA, prepared  $\sim$ 50 generations ( $\sim$ 2.5 weeks) later. The T0L262P sample was sequenced twice, a total DNA sample (T0L262P<sub>total</sub>) and a purified kDNA sample (T0L262P<sub>kDNA</sub>), described in section 3.2.4.2. For the total DNA sample the average number of mini- and maxicircles per cell could be estimated by comparing coverage to the diploid nuclear genome. The T0L262P<sub>total</sub> sample was calculated to have on average only 9 maxicircles and 2209 minicircles per cell. This indicates that within two weeks of transfection with the L262P  $\gamma$  mutation either (i) a substantial proportion of cells had lost their kDNA entirely, (ii) the kDNA network in many cells had shrunk, or (iii) a combination of both. In order to properly investigate this kinetoplast sizes must be quantified by DAPI staining under microscopy. These samples are available but need to be analysed.

The copy number distribution for the T0L262P<sub>total</sub> sample was determined and found to be already different from the T0WT copy number distribution (Figure 4.4). That these cells appear to be losing their kDNA rapidly makes estimating copy numbers from the kDNA purified samples difficult as we cannot assume that the total number of minicircles per network remains constant across time points.

It should be noted that whilst we observed these changes in the average size of the

network in a population and drastic changes in the copy number distribution of the L262P  $\gamma$  cell lines within a short time span, we did not detect any classes of minicircle which were completely absent.

For the kDNA independent cell lines where kDNA purified samples had been sequenced it is impossible to quantify copy numbers of minicircles and maxicircles or the total network size. This is because there is no nuclear DNA to compare to and the average size of the network is shrinking meaning that proportions of reads cannot be used to estimate minicircle copy number and the average maxicircle copy number cannot be assumed to be remaining constant. These samples (T0L262PkDNA and T<sub>s</sub>1L262PB) must be analysed by DAPI staining and microscopy in order to analyse network size.

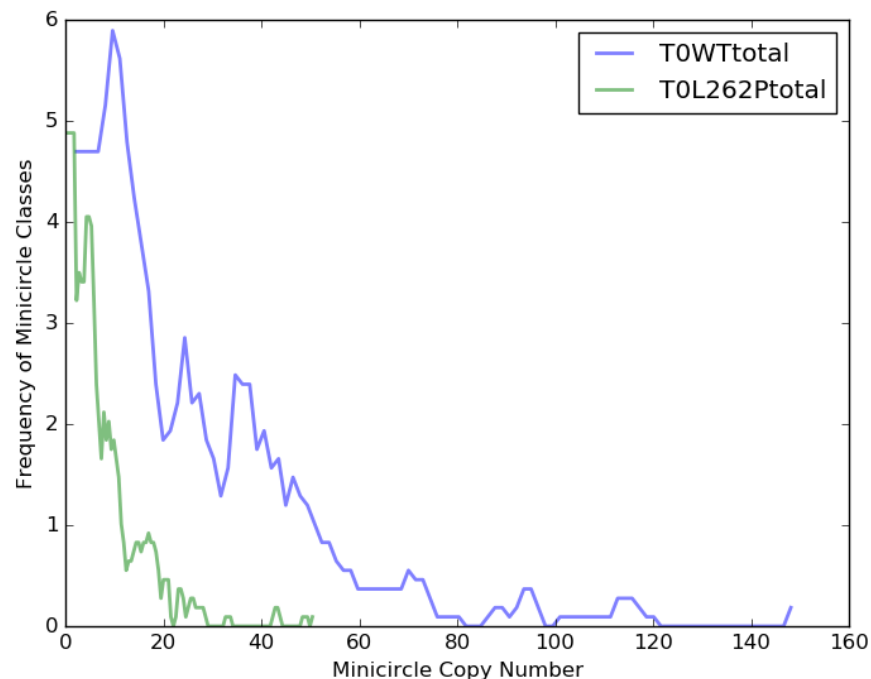


Figure 4.4: Minicircle copy number distributions were calculated for samples where the nuclear genome can be used as a standard. The average network size for the reference sample (T0WTtotal) was calculated to be 9695 minicircles and 30 maxicircles. The kDNA independent sample (T0L262Ptotal) was calculated to have a network size of 2209 minicircles and 9 maxicircles.

#### 4.3.1.5 Pilot time course conclusions

Taking these findings together, we were able to answer many of the questions we set out to with the pilot time course:

- i) We can observe changes in kDNA complexity over time in a population *in vitro*.
- ii) We were able to obtain sufficient reads from purified kDNA to assess minicircle loss with considerable sensitivity.
- iii) The time frame for the pilot time course was sufficient to observe loss of minicircle classes below the detection limit, however the number of minicircles lost was very different between the 2 parallel cultures. As expected, low copy number classes were lost first.
- iv) Removing kDNA dependence by introducing a compensatory mutation resulted in population-wide changes in kDNA abundance and, in two separate cultures, apparently complete loss within 100-150 generations (or 4-6 weeks) after selection of transfectants. In order to accurately quantitate these dynamics and the changes on a cellular level, however, total DNA samples would need to be sequenced to allow normalisation based on nuclear DNA, and microscopy would need to be used to investigate kDNA size per cell and the proportion of Dk/Ak cells in the population. In order to properly investigate the dynamics associated with this however total DNA samples would need to be sequenced to allow quantification of average network size and microscopy used to investigate the proportion of Dk/Ak cells.



### 4.3.2 Long term time course

Taking the results from the pilot time course into account we opted to run a longer time course at higher temporal resolution and with more stringent bottlenecks to increase the number of minicircles lost in a shorter time frame. A 29 week time course was carried out in triplicate, with purification of kDNA, preparation of samples for microscopy and preparation of cryo-stabilates on a weekly basis (Figure 4.5). The purpose of the slides for microscopy was to allow retrospective investigation of Ak/Dk proportions in the population and approximate quantitation of kDNA on a per cell basis. Additionally if cells are suspected to be differentiation incompetent after finding loss of gRNAs expected to be essential in the procyclic stage then the stabilates can be used to perform differentiation assays. The T0WT reference sample from the previous timecourse was expanded for kDNA isolation and the time point of isolation served as starting point for this long-term time course ( $T_{L0}$ ). At the same time the remaining culture was split into three parallel cultures that served as replicates (A, B and C).

This  $T_{L0}$  sample had therefore undergone two rounds of bottlenecks already; as described above, after parental cells had been transfected with a WT ATP synthase subunit  $\gamma$  replacement (to generate a cell line that was otherwise isogenic to the kDNA-independent L262P subunit  $\gamma$  cell line), a clonal line had been selected, expanded for sequencing of total DNA and isolated kDNA and cryo-stabilates had been prepared. The process of freezing and thawing the WT reference culture will likely have resulted in loss of some of the population, but this was not quantified.

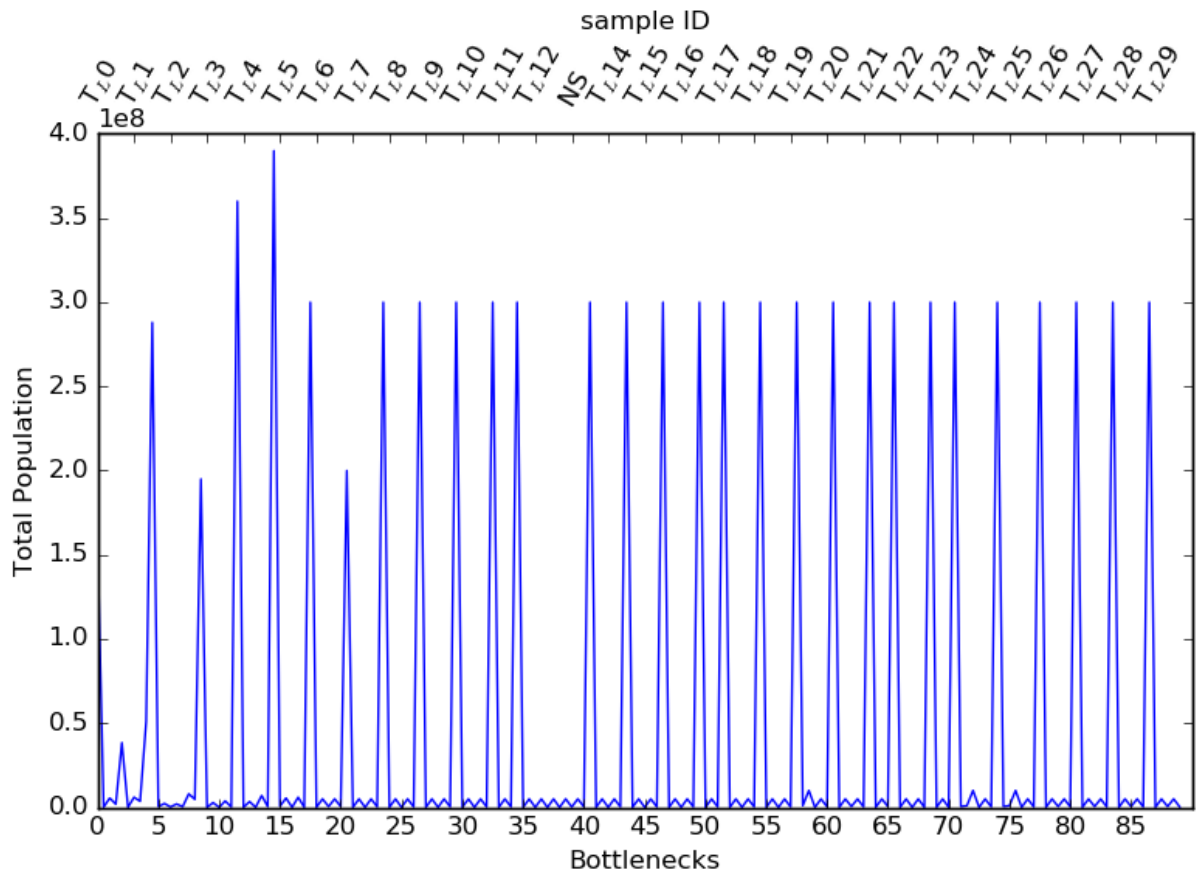


Figure 4.5 Sawtooth representation of cell growth and bottlenecks. Samples were expanded for kDNA purification one a week for 29 weeks. Sample IDs are indicated approximately above peaks from which they were derived. Number of time the cells were split, i.e. bottlenecks is shown in the x-axis. NS indicates no sample taken at week 13. The total population was calculated as the cell count multiplied by the volume of cells and was averaged across the three replicates (A,B,C).

Figure 4.5 shows a representation of the time course and bottlenecks in this experiment. The bottlenecks and numbers of cells shown in this plot are used to parametrise the bottleneck function in the model. In total 90 bottlenecks were carried out and of these 30 were carried out after the cells were bulked up for kDNA preparations (see section 6.6, table 6.2 for a list of samples generated).

#### 4.3.2.1 Sample QC and sequencing

At the end of this timecourse an aliquot of each kDNA prep was subjected to an EcoRI restriction digest and visualised on an ethidium bromide-stained agarose gel to assess the quality and quantity of kDNA purified for each sample (Appendix section 6.6, figure 6.19). In the first instance, kDNA complexity at the beginning and end of the timecourse were determined (sequencing at higher resolution was planned but not feasible within the time frame of this PhD thesis, see below). Four samples, an early timepoint (T<sub>L</sub>0, see above) and three late timepoints (T<sub>L</sub>28B, T<sub>L</sub>28C and T<sub>L</sub>29A), were sequenced using 100 bp paired end sequencing on an Illumina HiSeq. The two penultimate number 28 samples were chosen as the corresponding last time points, samples T<sub>L</sub>29B and T<sub>L</sub>29C, were of lesser quality (Appendix section 6.6, figure 6.19). A breakdown of the read mapping for these samples is shown in table 4.4.

Sample	Approximate number of generations	Number of reads (100 bp paired end)	% of reads mapped to		
			Minicircles	Maxicircles (per nucleotide depth)	Nuclear genome
T <sub>L</sub> 0	42	3347071	77.3	15.1 (5069)	0.8
T <sub>L</sub> 28B	609	3983308	81.8	11.7 (4646)	0.2
T <sub>L</sub> 28C	609	4274335	80.0	12.1 (5143)	1.1
T <sub>L</sub> 29A	630	4124299	78.0	11.4 (4709)	3.4

Table 4.4: Read mapping from the long term time-course.

#### 4.3.2.2 Complexity analysis

To determine minicircle complexity in the sequenced samples, reads were mapped to the 365 minicircle classes from the WT reference set and potential loss of classes assessed as before (a minicircle was considered to be present if more than 60% of the length has read depth  $\geq 1$ ). The T<sub>L</sub>0 sample had lost two minicircles, again indicative of

rapid loss of low copy number classes at the start of the time-course (data not shown). After ~600 generations of *in vitro* growth all three timecourse replicates had lost similar numbers of minicircle classes (T<sub>L</sub>28B: 28, T<sub>L</sub>28C: 26, T<sub>L</sub>29A: 33; Figure 4.6). Eight minicircles classes were lost from all three samples (Figure 4.6). In addition, the T<sub>L</sub>0 sample had lost two classes of minicircle compared to the T0WT reference. None of the classes that had been lost in any of the samples encoded non-redundant gRNAs and resulted in gaps in the editing sites covered. The gRNAs that have been lost from each of the samples can be viewed at [hank.bio.ed.ac.uk](http://hank.bio.ed.ac.uk) or at <http://tinyurl.com/gochbm5> where the appropriate samples have been pre-loaded. The samples are labelled as follows T0longtimecourse\_lost, T28Blongtimecourse\_lost, T28Clongtimecourse\_lost and T29Alongtimecourse\_lost. These tracks can be directly compared to the gRNA coverage presented in the AnTat9013\_DNA\_Cooper2016\_Contigs track.

As in the pilot time course, the number of gRNAs lost for each edited transcript is roughly proportional to the number of gRNAs present in the total population of minicircles for that transcript (Table 4.5).



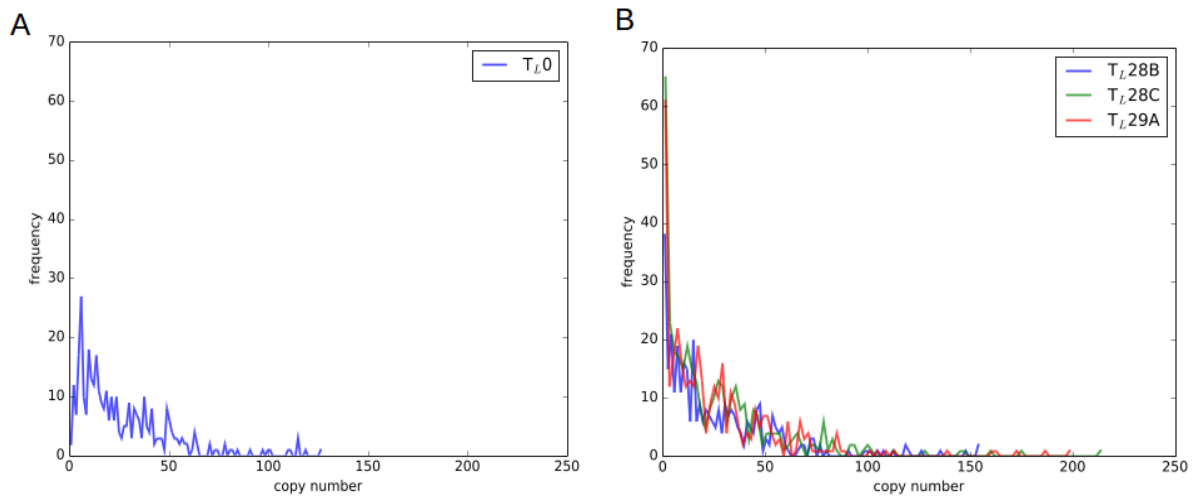
*Figure 4.6: Number of minicircles lost from late time points. Overlap of lost minicircle is shown as well as total number lost for each sample.*

Edited transcript the gRNA matches to	Number of annotated gRNAs (Percentage of total gRNAs annotated)	Number of gRNAs lost (percentage of gRNAs lost)			
		T <sub>L</sub> 0	T <sub>L</sub> 28B	T <sub>L</sub> 28C	T <sub>L</sub> 29A
a6	48 (4.7)	-	4 (5.6)	3 (3.9)	4 (4.4)
co3	96 (9.3)	-	4 (5.6)	6 (7.9)	8 (8.8)
cr3	14 (1.4)	-	1 (1.4)	-	3 (3.3)
cr4	42 (4.1)	-	5(6.9)	7 (9.2)	6 (6.7)
cyb	1 (0.1)	-	-	-	-
murf2	4 (0.4)	-	-	-	-
nd3	11 (1.1)	-	-	-	1 (1.1)
nd7	74 (7.2)	-	6 (8.3)	3 (3.9)	4 (4.4)
nd8	36 (3.5)	-	2 (2.8)	1 (1.3)	4 (4.4)
nd9	29 (2.8)	1 (16.7)	5 (6.9)	8 (10.5)	5 (5.6)
rps12	17 (1.7)	-	2 (2.8)	1 (1.3)	1 (1.1)
Predicted by nucleotide bias only	655 (63.7)	5 (83.3)	43 (59.7)	47 (61.8)	54 (60.0)
<b>Total gRNAs</b>	<b>1027</b>	<b>6</b>	<b>72</b>	<b>76</b>	<b>90</b>

*Table 4.5: Percentages of gRNAs from minicircles that have been lost compared to the total number of gRNAs annotated to reference minicircles. Time points are compared to annotations.*

#### 4.3.2.3 Copy number distribution

As with the pilot time course, comparing the minicircle copy number distribution of early and late kDNA samples showed a marked change in the distribution profile (Figure 4.7). Late time points have a much more skewed distribution with many low copy number minicircles and just a few high copy number classes, i.e. a few classes begin to dominate the population.



*Figure 4.7 Long term time course minicircle copy number distributions. Copy numbers were estimated by multiplying the total number of minicircles (9695 as calculated in Section 3.3.3) by the proportion of reads mapped per minicircle. A: Minicircle copy number distribution from early time point  $T_{L0}$ , B: Minicircle copy numbers from late time points,  $T_{L28B}$ ,  $T_{L28C}$  and  $T_{L29A}$ .*

### 4.3.3 Model fitting

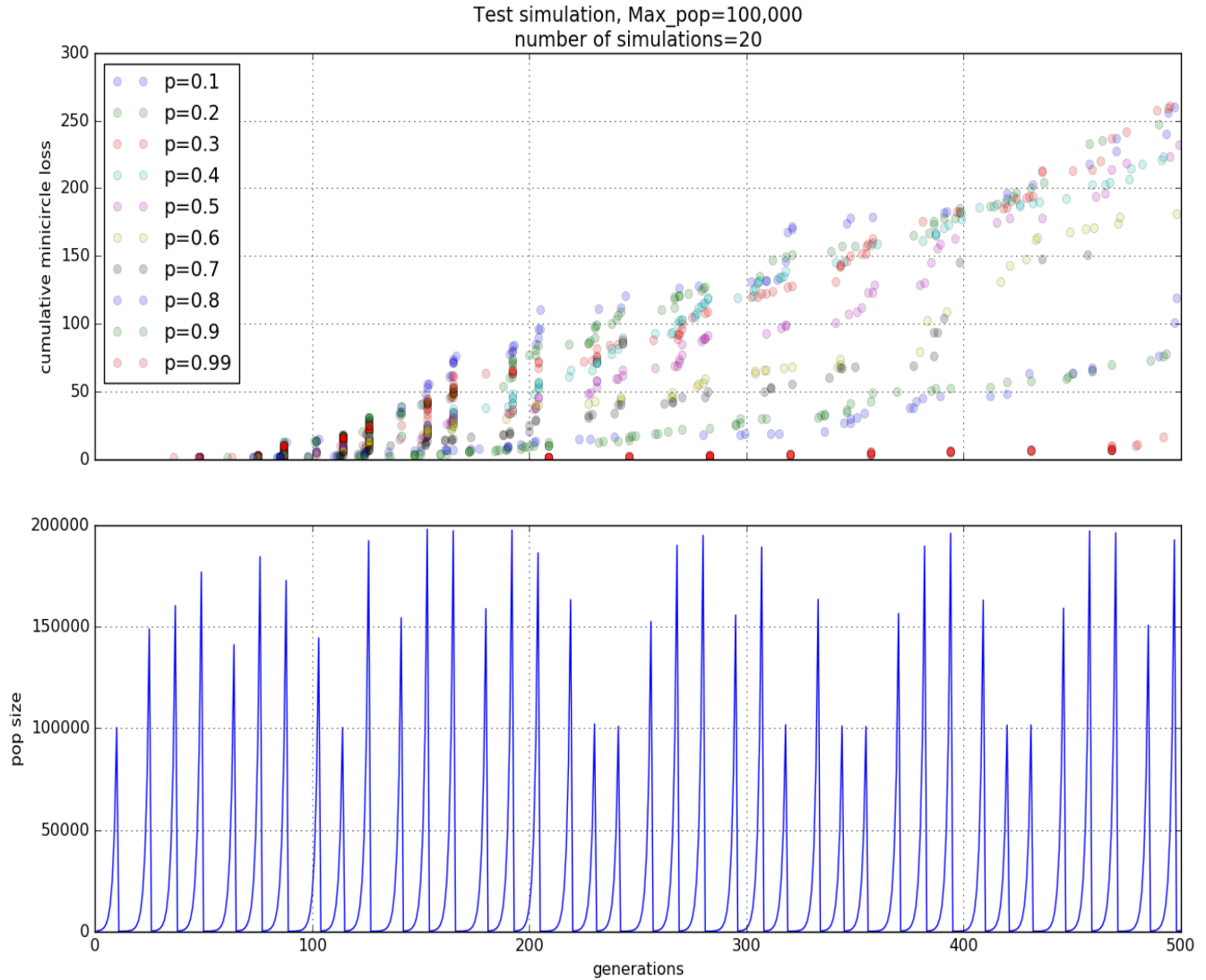
The final aim of Project 2 is to use the parameters obtained from assembly and annotation of the reference sample (Section 3) and from the time-course data to provide more accurate and precise parameter estimates for the model. The total complexity as defined by the number of minicircle classes in our reference set of minicircles (section 3.3.2) is 365. Estimates of how many of those minicircles are essential to facilitate minimal coverage of the editing space in BSF and PCF can be made (future work). The copy numbers of each of the minicircle classes from our reference set of minicircles (Section 3.3.3) is known. The final key parameter is the fidelity with which the kDNA genome is segregated. This can be determined by comparing simulations to the number of minicircles which are lost within a given number of generations, which has now been established.

#### 4.3.3.1 Preliminary simulations

Initial simulations, with the newly ascertained parameters, minicircle complexity, minicircle copy numbers and generations for minicircle loss, were run according to the conditions for the time-course. Parameter sweeps with the true population sizes and bottlenecks (as outlined in figure 4.5) will give an accurate segregation probability,  $p$ . Figure 4.8 shows a proof of concept simulation run with a low maximum population size of  $10^5$  cells (when the population is over  $10^5$  the model will run the bottleneck, hence the  $2 \times 10^5$  peak) which have been subjected to bottlenecks proportional to the ones carried out in the long term time-course. The simulation was run for 500 generations with a range of segregation probabilities ( $p$ ), with  $p=0.99$  being close to deterministic minicircle distribution,  $p=0.5$  being a random distribution and  $p=0.1$  being asymmetric distribution. For this small population size the rate of minicircles being lost is high even for a  $p$  which is close to determined minicircle segregation. For example, over 500 generations, when  $p=0.9$ , minicircles are lost linearly with a rate that results in loss of 70 classes on average. This high rate of loss is exaggerated by the population size used in this simulation. In this instance bottlenecks are proportional to the maximum population size of the simulation, so for a maximum population of 100,000 when cells are split 1/1000 only 100 cells are passing through the bottleneck, this vastly increases the stochasticity and thus increases the rate of minicircle loss. As expected rate of minicircle loss increases with lower segregation probabilities and in the extreme case where  $p=0.1$  nearly 260 minicircle classes are lost within 500 generations. Segregation probabilities below 0.5 are probably not biologically relevant as it is unlikely that the cells will actively segregate their kDNA genome asymmetrically, favouring one daughter over the other. That is unless they are kDNA segregation mutants of some description (i.e. have a defective TAC component).







*Figure 4.8: Small scale minicircle segregation simulations. For each segregation probability the simulation was run 20 times and the number of minicircles lost at each generation recorded. Each dot in the top panel represents the cumulative minicircle loss for a given simulation. Where replicate simulations produce the same cumulative minicircle loss at a given generation the dots are placed on top of each other and become more opaque. The bottom panel shows the simulation population size and bottlenecks with a maximum population size of  $10^5$ .*

Preliminary simulations were run with a population size peaking at  $10^7$  cells and were bottlenecked according to the same proportions used in the in the long term time-course. The simulation was seeded with 1000 cells and the cells were permitted to proliferate for up to a maximum population size of  $10^6$  split by a dilution factor of

1/100 the remaining cells are carried through to proliferate for to the maximum population size again; this is one cycle of bottlenecking and is continued for the remainder of the simulation. Every third cycle however the maximum population is set to  $10^7$  cells in a “bulking” cycle and the cells are split 1/1000 the generation after they reach this maximum population. The bottlenecks are close to experimental procedure however the maximum population size for the “bulking” cycle is magnitude lower than the maximum population in the long term time course due to current computational limitations. These simulations lose too few minicircle classes compared to expectations based on the time-course experiment even with low segregation probabilities (Figure 4.9). We see that with a segregation probability of  $p=0.1$ , only nine minicircles are lost from the population on average. Based on the time-course data we are expecting that after  $\sim 600$  generations of bottlenecked growth we should lose  $\sim 30$  classes of minicircle, this is the order of magnitude of minicircle loss we observe with low population simulations (Figure 3.2). However the large population simulations behave quantitatively differently. This is to be expected, the more cells in the population the lower the chances of losing a low copy number minicircle class as there are proportionally more cells with a single copy of a given class. For example if a minicircle represents 0.001% of the total minicircles, then in a population of 100,000 individuals 100 individuals would have one copy of this minicircle class. In a large population, say  $10^7$  cells,  $10^4$  cells would have a copy of this low copy number minicircle class. This means that in large populations low copy number minicircle classes persist for longer. The size of the bottleneck in the simulations is proportional to the maximum population size, as this is the case in the laboratory where cells are diluted proportionally (1/100, 1/1000 etc.) so for a large population size the bottlenecks are much larger than the small scale simulations and thus the stochastic effect of

bottlenecking the cells is reduced.

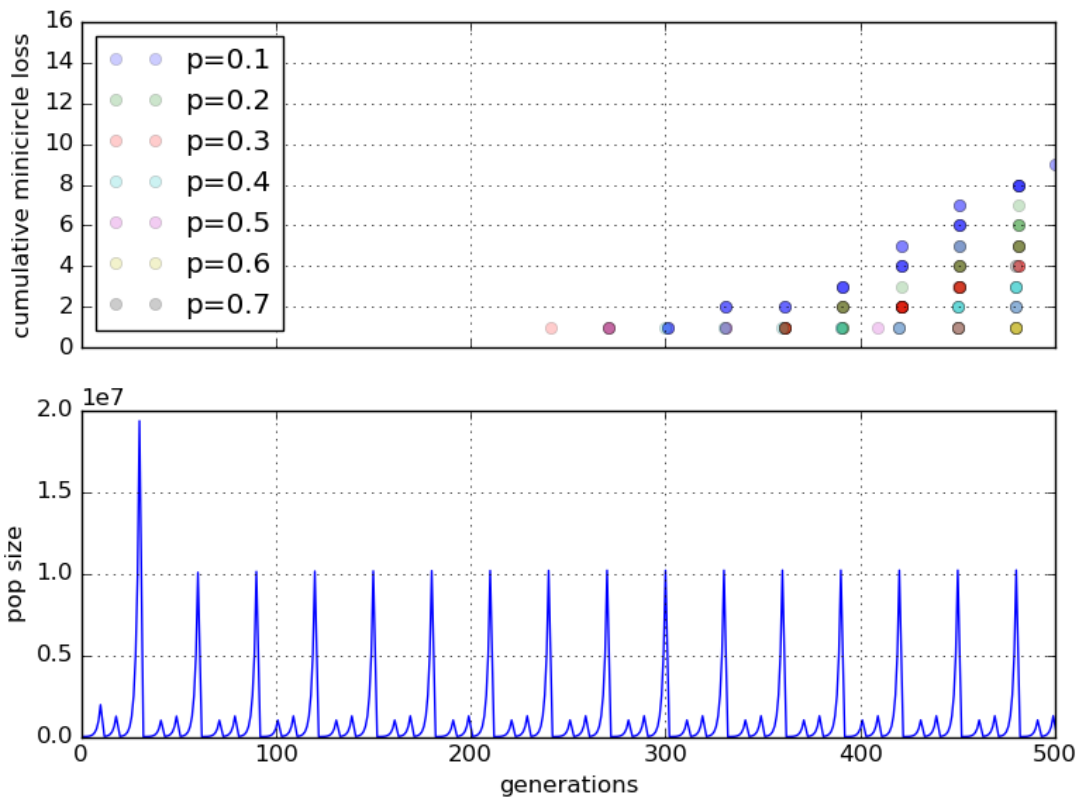


Figure 4.9: Large scale minicircle segregation simulations. Maximum population= $10^7$ , generations=500, number of replicate simulations=10.

This problem of the discrepancy between current model output and observed minicircle loss can be addressed in the future as follows. The primary approach for addressing this is to explore how to initialise the minicircle class copy numbers in each cell. The average minicircle class copy numbers from the reference sequences in Section 3 have been used to calculate the proportions that each minicircle class represents of the total network size of  $\sim 10,000$  minicircles. These proportions are used to initialise minicircle class copy numbers in each cell. We assume that the copy number of a particular class is binomially distributed with parameters  $p$ =average copy number of that minicircle

class and  $n=10,000$ . However copy numbers may be more dispersed than a binomial distribution, resulting in a greater proportion of cells with zero copies of a particular minicircle class. A second approach is to initialise the model for a several hundred generations with a constant population. The importance of stabilising the segregation model is discussed in Savill and Higgs (1999), however it was initially not deemed necessary as we now know the starting copy numbers of the minicircle classes thus we do not need to allow them to stabilise from a initially equal copy numbers as in the original model. The detection limit of low copy number minicircles in the time course experiments can be factored into the model. The definition for a lost minicircle in the simulations however is that a lost minicircle has copy number zero across the population. One way to obviate the requirement to supply the exact detection levels by experimentation would be to parametrise the segregation probability model based on rate of minicircle loss rather than absolute number of minicircles lost over 500 generations. This requires more exploration via simulation and higher resolution copy number information from sequencing data for ratification.

#### 4.4 Discussion

We now have a large number of samples with which to explore the dynamics of kDNA segregation and the initial analyses of end point samples from time courses has demonstrated that complexity on the kDNA genome is lost within just a few hundred generations. Previous observations, mainly based on observations in *Leishmania* (Thiemann et al., 1994; Gao et al., 2001; Simpson et al., 2015) and Dk/Ak subspecies of *T. brucei* (Schnauffer et al., 2002; Lai et al., 2008) had demonstrated that loss of complexity can occur and can affect the editing capacity of a population of cells. The observations made with *L. tarentolae* (Thiemann et al., 1994; Gao et al., 2001; Simpson

et al., 2015) showed that a 'fresh' isolate had approximately five times the kDNA complexity of the old lab strain. However these observations have been made in two separate cell lines, one of which has been cultured in the lab for an undefined period of time.

The observations made in the mutant cell line (L262P $\gamma$ ) (whilst mostly anecdotal at this stage) are interesting. We have observed in this particular kDNA-independent cell line that they lose kDNA wholesale within 6 weeks (~126 generations) of *in vitro* culture. It should be noted that other mutant kDNA-independent cell lines which have been generated are much more stable and do not become Ak so readily although it has been observed that homozygous L262P $\gamma$  cells appear to lose complexity more readily in order to circumvent expression of A6 (Dewer C. -in preparation). For comparison lab induced Ak cells were selected for using sub-lethal doses of acriflavin which selectively binds kDNA (Stuart, 1971). The data in this thesis highlight that the mechanism by which the mutant cell lines lose complexity is probably different from that of WT cells, where by it is likely a result of partially random kDNA segregation. There kDNA in the mutant cells in this particular experiment appears to be being actively selected against. In nature the question still remains, is loss of kDNA complexity and the subsequent generation of Ak cells a result of tsetse independent transmission or *visa versa*?

## 4.5 Outlook

Steps have been made towards a refined and accurate model of kDNA segregation which will initially provide information about the fidelity of kDNA inheritance. With this information predictive analyses can give indications of how long it takes for sufficient complexity to be lost to have an effect on editing capacity. Interesting

questions can be probed with this information in hand, for example, if the length of a chronic infection is directly related to the complexity of the kDNA genome how will transmission efficiency be affected? Chronic *T. b. gambiense* infections of humans have been recorded that can persist for several decades (Sudarshi et al., 2014) and trypanotolerant cattle with chronic *T. b. rhodesiense* infections serve as parasite reservoirs. The transmission dynamics of these long infections are unclear however, it is known that parasite levels are low during the chronic phase (Van Den Bossche et al., 2005), this is mediated by the host immune clearance and the cells differentiating to the non-dividing short stumpy morphology (MacGregor et al., 2011). Given this low parasitaemia (as low as 100 parasites per ml of blood (Franco et al., 2014)) and the small volumes of blood being taken up by the tsetse, in the order of 10s of milligrams of blood (Loder, 1997), the effect that even a small proportion of cells within the population not being able to survive in the insect vector has on transmissibility would presumably be significant. For example given a chronic infection with a parasitaemia of 100 parasites per mL of blood and a tsetse fly blood meal of 10  $\mu$ L only 1 parasite is taken up per blood meal. The effect of just one tenth of the cells in the the host having lost kDNA complexity to the degree that they are unable to survive in the insect vector would result in a reduction in transmission efficiency by 10%.

We can also begin to answer some long standing questions about why kDNA in *T. brucei* has evolved to be so complex, some of which are outlined by Simpson (1997). Is this related to the average length of time a BSF parasite is likely to be locked in a host? Do high levels of gRNA redundancy give a fitness advantage to parasitic trypanosomes? Or is it related to the editing machinery requiring many gRNAs for efficient editing cascades? Additionally it can provide information pertaining to the unique architecture of the kinetoplast, has the TAC evolved to increase the segregation

probability?



## 5. Conclusion

Taken together the data presented in this thesis provide the first essentially complete description of the total kDNA complexity in *T. brucei*, along with functional annotation based on small transcriptome data. Additionally we present the first quantitative data describing the dynamics of the kDNA genome.

The kDNA genome in *T. brucei* is more complex than previously thought (Steinert and Van Assel, 1980; Hong and Simpson, 2003) and, in the differentiation competent strain analysed here, made up of 365 classes of minicircles present in copy numbers ranging from less than one per cell on average to more than one hundred copies of a given class. On average we determined ~10,000 minicircles and ~30 maxicircles per cell. Of the ~1000 annotated and confirmed guide RNAs only half have a match in the known editing space (canonical gRNAs). The other half - non-canonical gRNAs - are highly similar to canonical ones in every way apart from the ratio of sense to anti-sense transcripts at steady state.

The large number of non-canonical gRNAs identified here is somewhat surprising. Transcriptome studies had previously eluded that a proportion of the gRNAs could be non-canonical however the true scale of the non-canonical gRNAs was not known (Hong and Simpson, 2003; Madej et al., 2008a; Aphasizheva et al., 2014; Suematsu et al., 2016b). The significance of non-canonical small RNAs identified in screens where the DNA template is not known is difficult to ascertain. For example in Suematsu et al., (2016b) the authors classified their non-canonical guide RNAs as any small RNA (isolated from mitochondrially enriched samples) which did not map to the nuclear genome and did not have a match in the known editing space. With no DNA template available the authors are forced to classify gRNA species using a cluster based

method, this resulted in identification of 73,361 non-canonical gRNAs and 64,441 gRNAs. This level of sequence heterogeneity is far beyond the number of minicircles we have defined or that have been previously estimated to be present (Steinert and Van Assel, 1980). Knowledge of the DNA template greatly helps in dealing with the question of how to classify guide RNAs. Analysis of canonical gRNA genes in minicircle sequences revealed a strong nucleotide bias upstream and downstream of these genes, which led to the development of a gRNA identification pipeline that was independent of sequence homology to edited mRNAs. This allowed more accurate and comprehensive quantification of the number of gRNA genes that are canonical versus non-canonical. Analyses of the canonical gRNAs compared to the non-canonical gRNAs revealed that they are the same in almost every respect, length, 3' U-tail length and abundance. Suematsu et al., (2016b) have reported that non-canonical gRNAs, as they define them, have shorter U-tails. Our study does not concur with this finding, and this difference is likely a consequence of the way in which they define non-canonical gRNAs; by taking all mitochondrially enriched small RNAs which do not have a match in the editing space or the nuclear genome to be non-canonical gRNAs.

The only difference between the canonical-gRNAs and the non-canonical gRNAs identified in the present study was the ratio of sense to anti-sense transcripts. This suggests that non-canonical gRNAs, or at least many of them, are indeed functionally different and not simply canonical gRNAs that had been missed due to alignment parameters being too stringent. This difference coupled with evidence from previous studies (Suematsu et al., 2016b) highlights the potential role of anti-sense transcripts in gRNA processing and/or function. Suematsu et al., (2016b) provided intriguing hypotheses related to the function of anti-sense transcripts. By definition, and as illustrated in the model of gRNA processing presented in that study, any

sense:anti-sense gRNA duplexes would be derived from opposing strands of the same minicircle. However, the lack of complementary minicircle data in that study made the reliable identification of anti-sense minicircle-derived transcripts extremely challenging. Given the model of gRNA processing they provide one would assume that any sense:anti-sense gRNA duplexes would be derived from opposing strands of the same minicircle. The anti-sense transcripts identified in the present study are shorter than their sense counterparts and have shorter U-tails, in contrast to the molecules identified by Suematsu et al., which resembled their sense counterparts in these aspects. Thus, the antisense RNAs identified in the present work would form quite different duplexes from those reported by Suematsu et al., (2016b). The difference in the ratio of sense to anti-sense transcripts for the canonical and non-canonical gRNAs suggests an interaction with the maxicircle derived mRNAs that has some bearing on the abundance of the anti-sense transcripts. Another possibility is that the canonical gRNAs are more efficiently transcribed on the anti-sense strand than non-canonical gRNAs. Ascertaining if gRNA:anti-gRNA duplexes indeed form *in vivo* is key to resolving these questions and as previously suggested could be resolved using RNA cross-linking methods (Helwak et al., 2013). Additionally the stability of the potential RNA duplexes identified in this study could be assessed as in Suematsu et al., (2016b) by calculating their free energy (Freier et al., 1986). Resolution of these issues may begin to give clues as to the function, if any, of non-canonical gRNAs.

The nucleotide bias observed in the regions flanking gRNA genes appears to be a key characteristic of a gRNA transcription unit and possibly explains why the previously identified inverted repeats (Pollard et al., 1990) are often described as “imperfect” (Jasmer and Stuart, 1986). Perhaps the transcription machinery does not need specific sequence for binding but rather requires an overall nucleotide

composition which tends towards, but does not fully depend on the imperfect repeats which have been previously reported [11]. If minicircles are indeed prone to fast evolution (Sanchez et al., 1984), flexibility in the recognition of transcription units could prove essential. *In vitro* transcription studies, such as have been carried out for nuclearly encoded genes (Park et al., 2012), using synthetic minicircle sequences with varying degrees of nucleotide bias could be used to test this suggestion experimentally.

We have completely assembled all, or at least 99% of minicircles for *T. brucei* strain AnTat90.13. In comparison to other sets of minicircles from *T. brucei* this represents a significant improvement. Previous attempts to examine the kDNA genome (Hong and Simpson, 2003; Ochsenreiter et al., 2007a) have compiled minicircle sequences from multiple strains and did not give a true representation of minicircle complexity for a single cell line and population. Despite the improvements made in this study it is likely that the minicircle complexity represented in AnTat90.13 is somewhat lower than what would be observed in a freshly isolated strain of *T. brucei*. The exact culturing history of strain AnTat90.13 is not known and given the evidence in section 4, and section 3.3.7 this strain may have lost kDNA complexity already. Sequencing kDNA from “fresh” *T. brucei* isolates may yield yet more complexity. The within population minicircle variation that is likely to be present in the kDNA genome could also be investigated using single cell sequencing methods (Junker and Van Oudenaarden, 2014).

Nonetheless, we have now realised accurate information for minicircle complexity required in a fully differentiation competent cell line. This information is key to confirming or adjusting previous hypotheses and answering many outstanding questions surrounding the kinetoplast. Some of these questions have gone unanswered for many decades; the tools are now in place to answer them. A hypothesis we have

started to address was posed over a decade ago (Savill and Higgs, 1999): that the fluctuations observed in kDNA complexity are due to the segregation of daughter minicircles being partially random. Minicircle dynamics data will prove to be an essential tool for confirming this hypothesis and in turn may be able to provide testable data relating to questions put forward by Borst, (1991), in which the author asks why the kinetoplast network exists at all. A properly parametrised kDNA segregation model will allow important insight into how quickly kDNA complexity can change during replicative life cycle stages of *T. brucei*. We also now have the tools to examine other factors that may contribute to kDNA complexity, for example genetic exchange in the tsetse (Gibson et al., 1997). The extent to which the minicircle repertoire may be replenished by recombination in the insect vector is presently unknown.

The kDNA dynamics data presented in section 4 also represent a useful data set for the validation of other models. Savill and Higgs, (2000) put forward a model whereby non-functional gRNAs are generated through drift and accumulate over time. The predicted result is minicircles which encode just one functional gRNA and an increase in minicircle complexity. The authors postulated that *T. brucei* is in the midst of this process. It is conceivable that the generation of non-functional gRNAs could be predicted by measuring the accumulation of variation within gRNA genes and may provide evidence to support his model. Relating the number of non-functional gRNA genes to trypanosomatid phylogeny may also support this hypothesis.

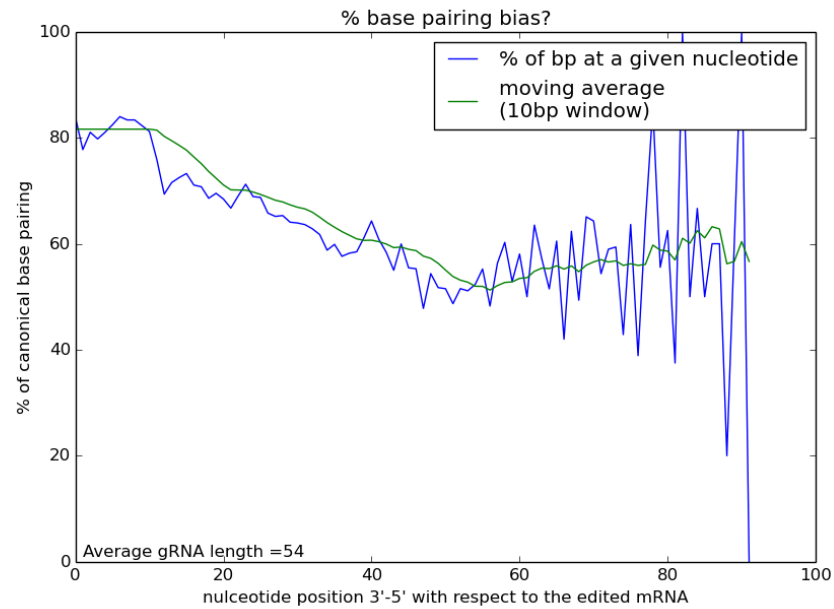
Likewise, a complete kDNA genome provides a tool to be able to address questions related to models of kDNA structure and topology (Michieletto et al., 2014; Diao et al., 2015). Genome topology analyses (Dekker, 2002) can give global insight into the degree to which minicircles and maxicircles are linked to each other. This approach coupled with methods for investigating cell cycle (Kabani et al., 2010) has

the potential to provide models of kDNA interactions during replication.

Sequencing and annotating the nuclear genome of *T. brucei*, as in other organisms, has dramatically changed research, giving a new understanding of the biology and allowing the development of new tools. The potential applications for sequencing the mitochondrial genome are not known. However, this is as it was before the nuclear genome was sequenced and true understanding of eukaryotic biology cannot be realised until we also have a complete understanding of the mitochondrial genome. Insights into how the nuclear and mitochondrial genomes interact will not be possible without detailed analyses of the “neglected” mitochondrial genome and its functional products (Pesole et al., 2012). There are many open questions and the scope for new applications for investigating kDNA structure, function and evolution are only helped by having a more comprehensive knowledge of its complexity and dynamics.

## 6. Appendix

### 6.1 Additional minicircle annotation statistics



*Figure 6.1 gRNA-mRNA duplexes have more canonical (Watson-Crick) base-pairing at their 5'-anchor end (3' with respect to the mRNA) than at their 3' end.*

## 6.2 PCR validation of assembled minicircles

A random subset of minicircles as described in section 3.3.2 and section 3.3.7.1 (if the appropriate genomic DNA was available) were selected, PCR amplified, visualised by gel electrophoresis and Sanger sequenced (Edinburgh Genomics). Gel with amplified minicircles for AnTat90.13 is shown in Figure 3.3. Lister 427 minicircles are shown in figure 6.2. EATRO164 minicircles were ligated into pGem T-easy (Promega) as per manufacturers instructions and transformed in E.coli. A fraction of the vector plus insert was then restriction digested with Taq (Promega) and visualised on an ethidium bromide gel to check that the insert was present (Figure 6.3). Inserts were Sanger sequenced (Edinburgh genomics).

Cell line	Minicircle ID	Forward Primer	Reverse Primer
AnTaT90.13	mO_@31	TCAGTATTCTTATCACCTCCATTAT	AAATCAGTAGGAAAAGTAAGGTGTA
	mO_@180	GGTTTTCTCACTTATTGGCTTTA	TGTAATTCTCTACCATATACTTAC
	mO_@335	GGTTTTCTGGAATTTTCAGCTTAT	CTTTAGAGTCAAACCTAATAACCCCT
	mO_@232	TCAACCTCTAACACTATTGATTCAA	ACTGAAGTAGTTAATTAGAATAATC
	mO_@347	CACACGGTTTTTTCACATTATTCA	AATTCCTAATAGCAGATTCTTCTTC
	mO_@7	TGCTTTTTCTGTTTATTCTGAGAT	TATCCACAGAAAATAACACTACTACT
	mO_@307	TATTTTATTCTTCCACCCCTGAATA	CTGATTTATAAGTTGGGAATAGAAG
	mO_@223	CGATTTTCTCCGATTTTACCTAAT	ATAAATCATGAGTTGTGGATCTTGT
	mO_@337	TTTTTCTCGAGATTTAGTGGGAAAT	ATCATCTGACTGGATTTTAACACT
	mO_@230	AAAATTCCCAAAATTCCTCCAAAAT	TGTAGATGGATGTGAATGGAATTTT
	mO_@138	TAAGAGGGACAGAGTAATATAAGTT	GATTAGCACTTAACTCTGCAATTA
	mO_@197	GGTTTGGAGTTTATCTAGGTTTTGA	TTTTCTTTCTTTATCTTGAAGCCCT
	mO_@45	ACTCTGCATAACTTCTACTAACAAA	TTTTGAGTTTGATAATTGGATGGGT
	mO_@170	ACGCCATAAGATTTAGGTTTTTTT	AACAAAATATCGACTTTTTGAGGCT
	mO_@248	GTGTTATTGAATGTCTCAAAGGTAA	TTTTATCTTCTAGACCTCGAAAA
EATRO 164	164_WT_minicircle_33	AATTATCCGAAAATCCCAGGAAAAA	TATATATACAATGGTTATTCACTGG
	164_WT_minicircle_17	AAACTGAAAAATCCCTTGAAAAAGC	CTTTAAAGAGTGTCCCTAGAAATAA
	164_WT_minicircle_89	ATATCCAATAAAAACTCCGGAAAA	AGATAAGGGTTTGCCTATTAGCTTA
	164_WT_minicircle_90	AGCAAACTCTCTGTAACATATAG	TGTTTCAGATATAGATTCTCTAGA
	164_WT_minicircle_19	AAACTCTGAAAATCATGGGAAAAAC	CAAGGGTTTTTTCAGAGATTTAGATT
	164_WT_minicircle_92	ATTTCTCCATCACAAAACCTACAAAA	GGTTAATCATCTGGAGTGTTTATA
	164_WT_minicircle_87	AAAAATTCAGATTTCTAAGCCTCC	GAGTTACAATGGATACTAATGTGAT
164_WT_minicircle_8	AACATCCGAAAATATCGAGAAAAAC	GTATATGAGGTTTTATATGGCAGTA	
Lister427	427_minicircle_10	AAGTCTCAAAAACCGTGTGA	AGGGTTTTTCTGAGAGTCAT
	427_minicircle_19	AAAATTCCCGAAAAACCCCT	GGAGTGGTTATTGGGTATTT
	427_minicircle_25	CTGAAAAAGGAGTGTGTGTA	GTTTTTGTGGTTTTGAGA
	427_minicircle_24	AAATCACCGAAAAACCGTGT	GAGTTAGAATGGGAGTTAGA
	427_minicircle_2	TCTATCTAGACGACTCGAAA	TGATAAATTAACAAACAATG
	427_minicircle_40	AAATCCCTGAAAAAGCAGTG	GGGTTTTGGGAGTTTTGATA
	427_minicircle_27	TAAACTCAATGAAAACCTGG	CACAATATTTTCAGATTTAA
	427_minicircle_21	TCGAAAACCCCTGGAAAAAAC	TATTTGTGGGTTTGTAGGA
	427_minicircle_5	TCTCGAAAACCTCAGAAAAAC	TGTTGAATTCAGGTTTATAG
	427_minicircle_51	TTACTCAAACCCCAAAAAC	ACAGGCTTGACAGATTAAT

Table 6.1 Primers used for validation of assembled minicircles.



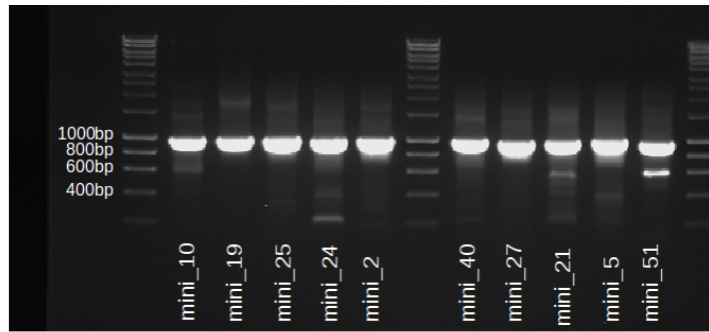


Figure 6.2 PCR validation of *Lister* 427 minicircles.

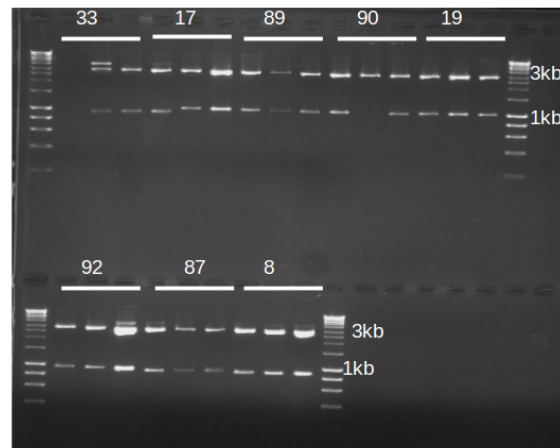
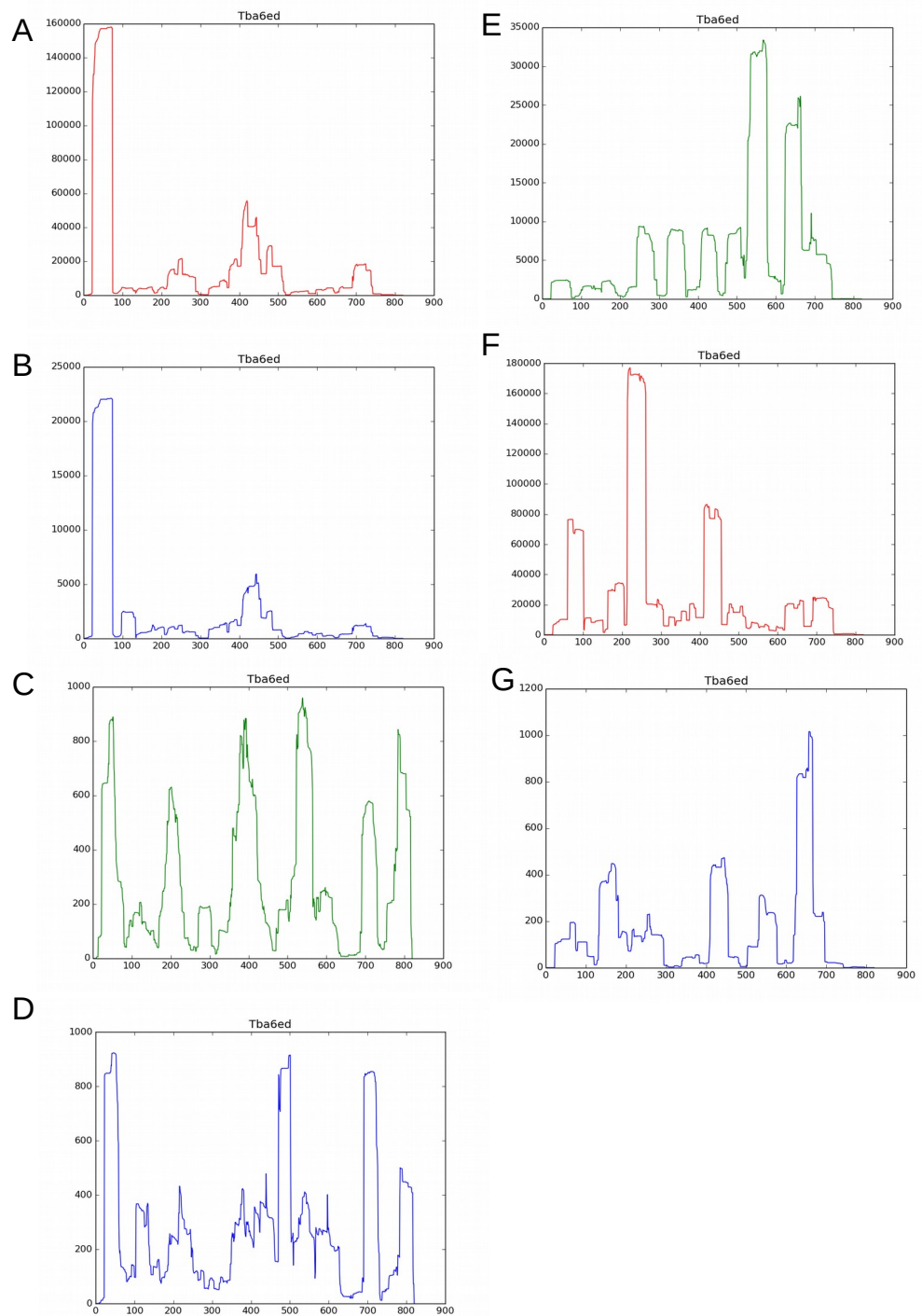


Figure 6.3 PCR validation of *EATRO* 164 minicircles. PCR products were ligated into *Pgem-T-easy* vector, transformed in *Ecoli* and subsequently digested using *Taq*. 3 kb band indicated vector backbone and 1 kb band indicates cloned minicircle.

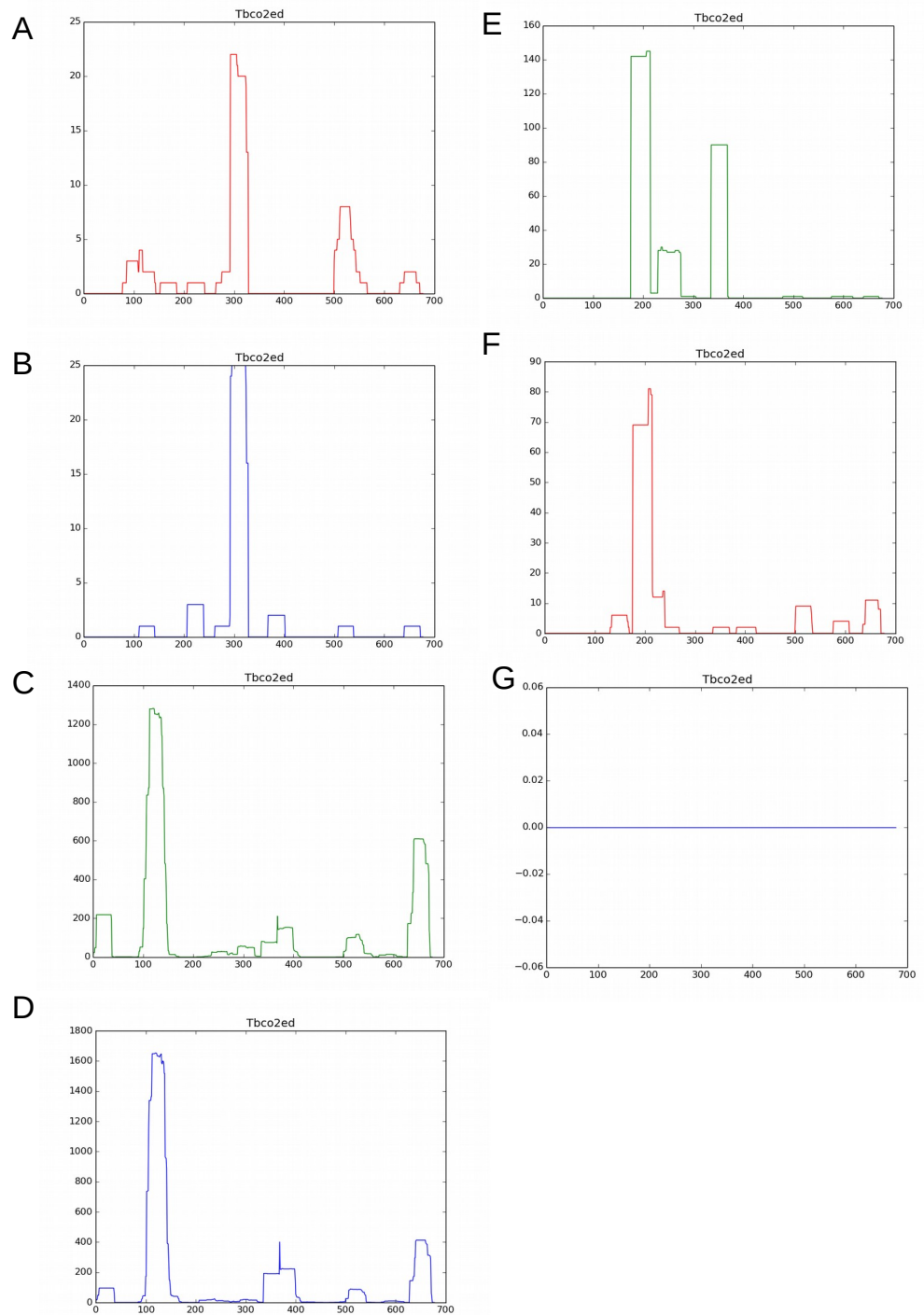
### 6.3 gRNA depth plots

The per nucleotide depth across the editing space for each of the edited genes is calculated and depth plots have been generated for each of the small RNA data sets analysed in this study.

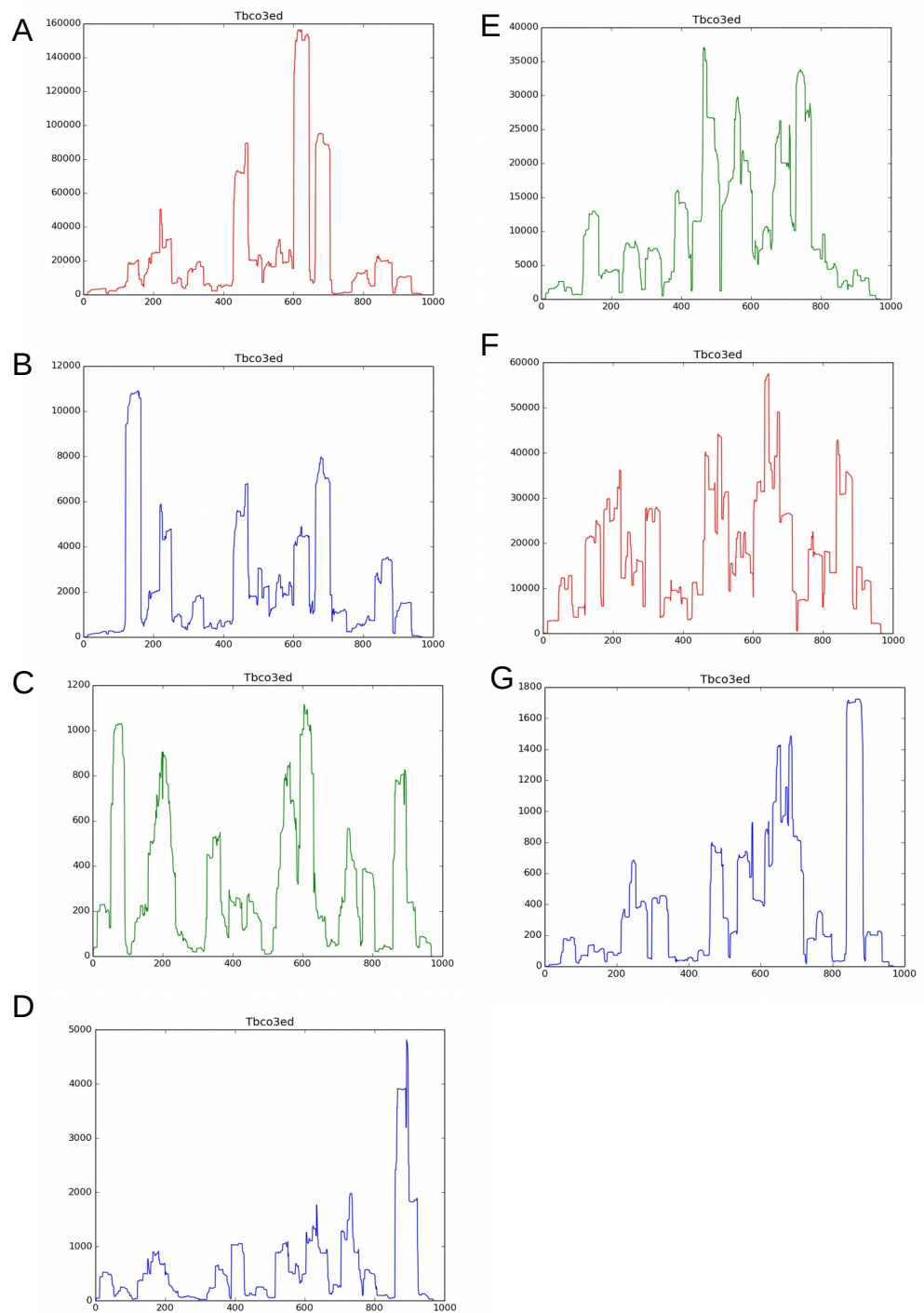
Figures 6.4-6.15: Depth plots for small RNA data sets which have been aligned to the edited maxicircle transcripts. **A:** AnTat90.13 BSF, **B:** AnTat90.13 PCF, **C:**TREU667 BSF, **D:** TREU667 PCF, **E:** EATRO 164 PCF (Koslowsky et al. 2014), **F:**L427\_29.13A (Suematsu et al 2016), **G:** L427\_29.13A (Madina et al. 2014).



**Figure 6.4 Edited A6 gRNA depth plots. A: AnTat90.13 BSF, B: AnTat90.13 PCF, C: TREU667 BSF, D: TREU667 PCF, E: EATRO 164 PCF (Koslowsky et al. 2014), F: L427\_29.13A (Suematsu et al 2016), G: L427\_29.13A (Madina et al. 2014).**



**Figure 6.5 Edited *co2* gRNA depth plots. A: AnTat90.13 BSF, B: AnTat90.13 PCF, C: TREU667 BSF, D: TREU667 PCF, E: EATRO 164 PCF (Koslowsky et al. 2014), F: L427\_29.13A (Suematsu et al 2016), G: L427\_29.13A (Madina et al. 2014).**



**Figure 6.6 Edited *co3* gRNA depth plots. A: AnTat90.13 BSF, B: AnTat90.13 PCF, C: TREU667 BSF, D: TREU667 PCF, E: EATRO 164 PCF (Koslowsky et al. 2014), F: L427\_29.13A (Suematsu et al 2016), G: L427\_29.13A (Madina et al. 2014).**

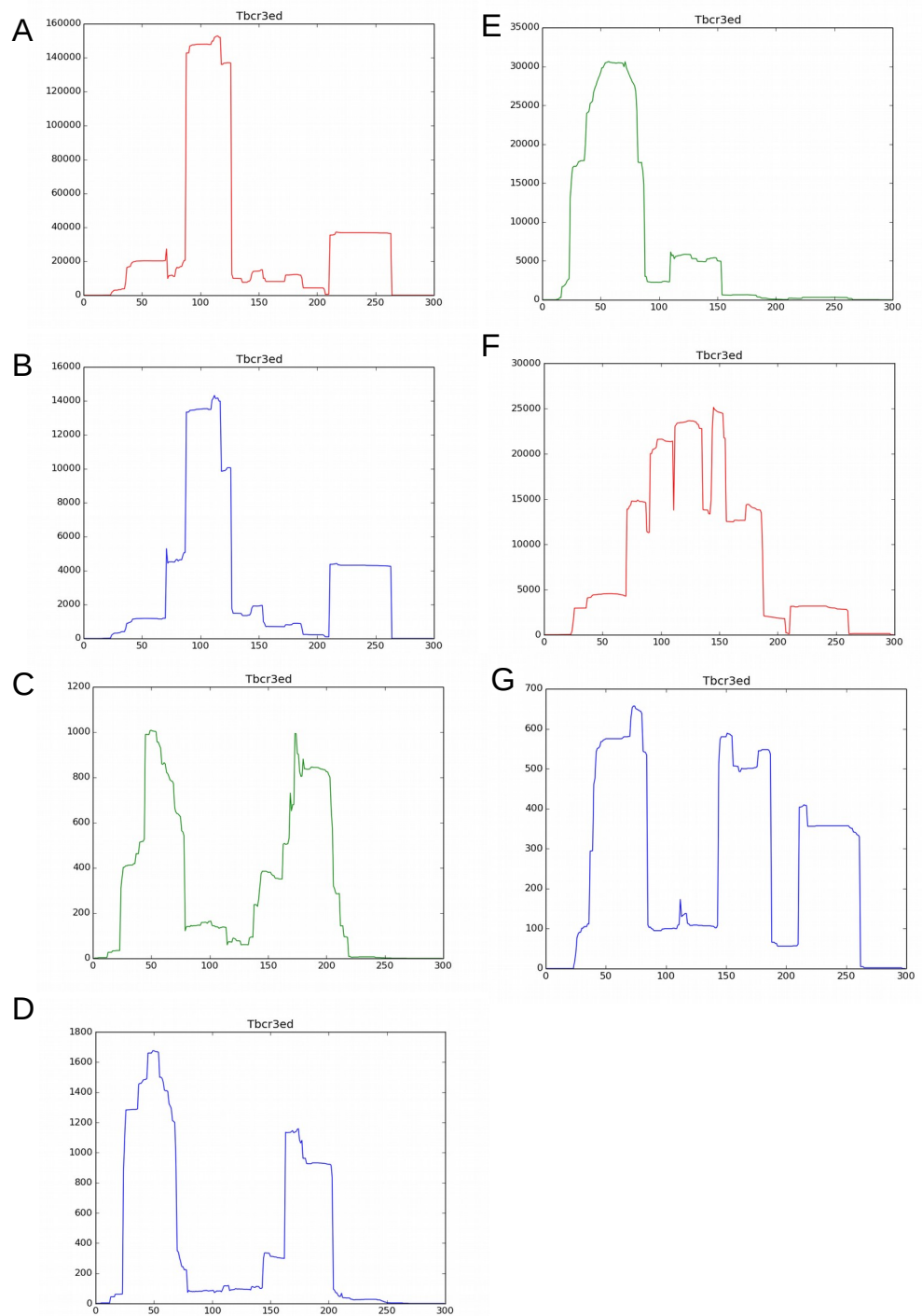
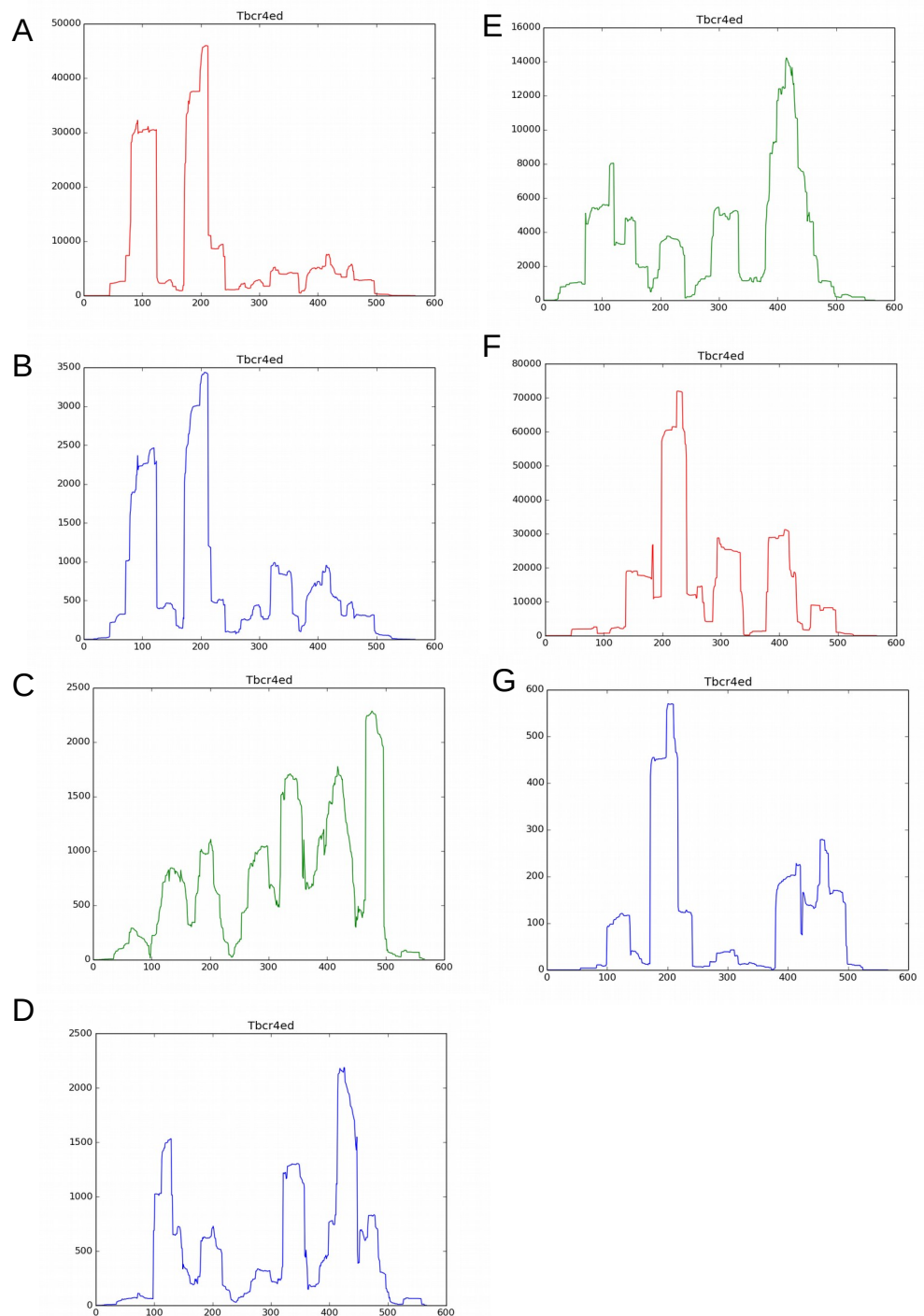
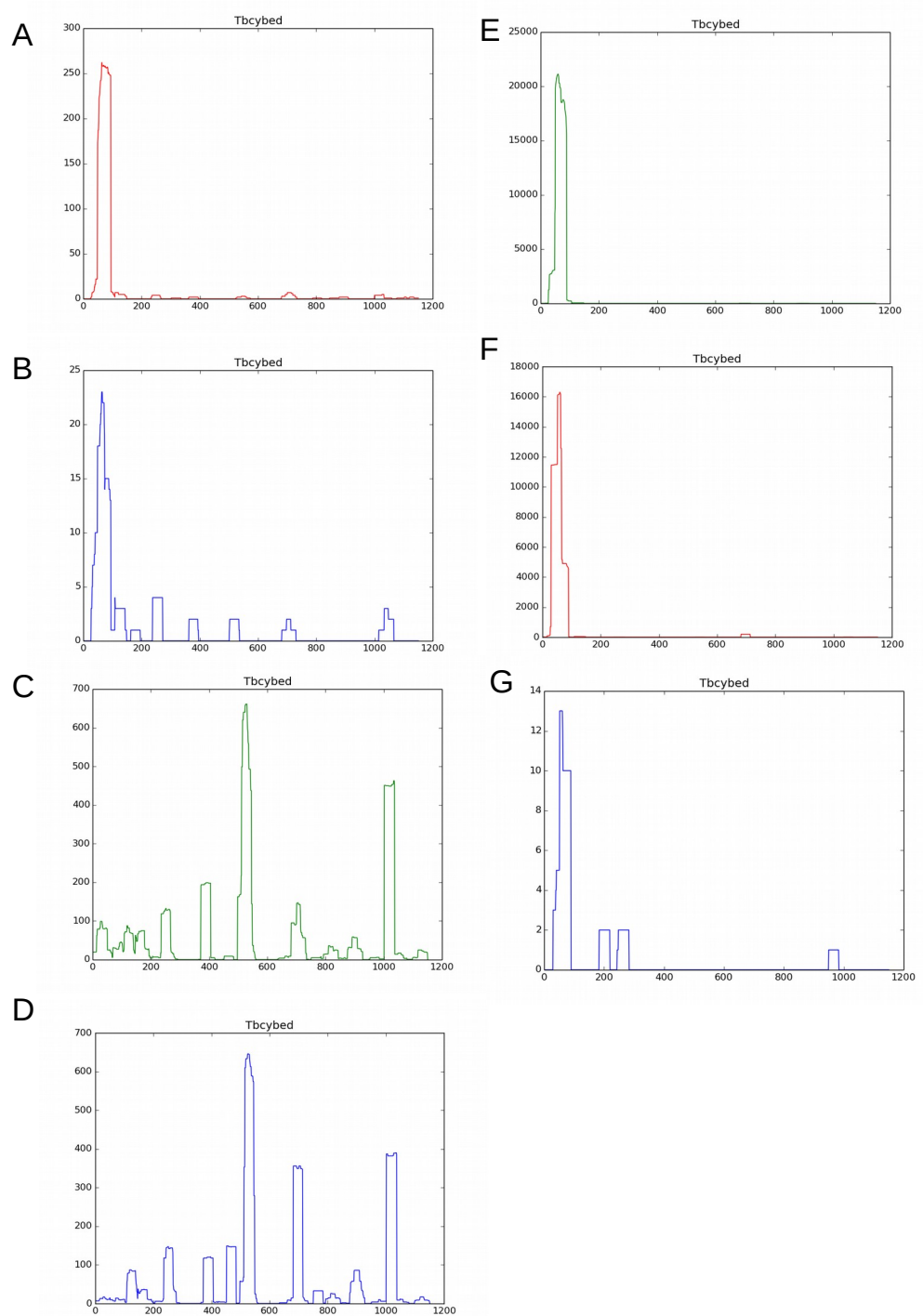


Figure 6.7 Edited cr3 gRNA depth plots. **A:** AnTat90.13 BSF, **B:** AnTat90.13 PCF, **C:** TREU667 BSF, **D:** TREU667 PCF, **E:** EATRO 164 PCF (Koslowsky et al. 2014), **F:** L427\_29.13A (Suematsu et al 2016), **G:** L427\_29.13A (Madina et al. 2014).



**Figure 6.8 Edited cr4 gRNA depth plots. A: AnTat90.13 BSF, B: AnTat90.13 PCF, C: TREU667 BSF, D: TREU667 PCF, E: EATRO 164 PCF (Koslowsky et al. 2014), F: L427\_29.13A (Suematsu et al 2016), G: L427\_29.13A (Madina et al. 2014).**



**Figure 6.9: Edited *cyb* gRNA depth plots. A: AnTat90.13 BSF, B: AnTat90.13 PCF, C: TREU667 BSF, D: TREU667 PCF, E: EATRO 164 PCF (Koslowsky et al. 2014), F: L427\_29.13A (Suematsu et al 2016), G: L427\_29.13A (Madina et al. 2014).**



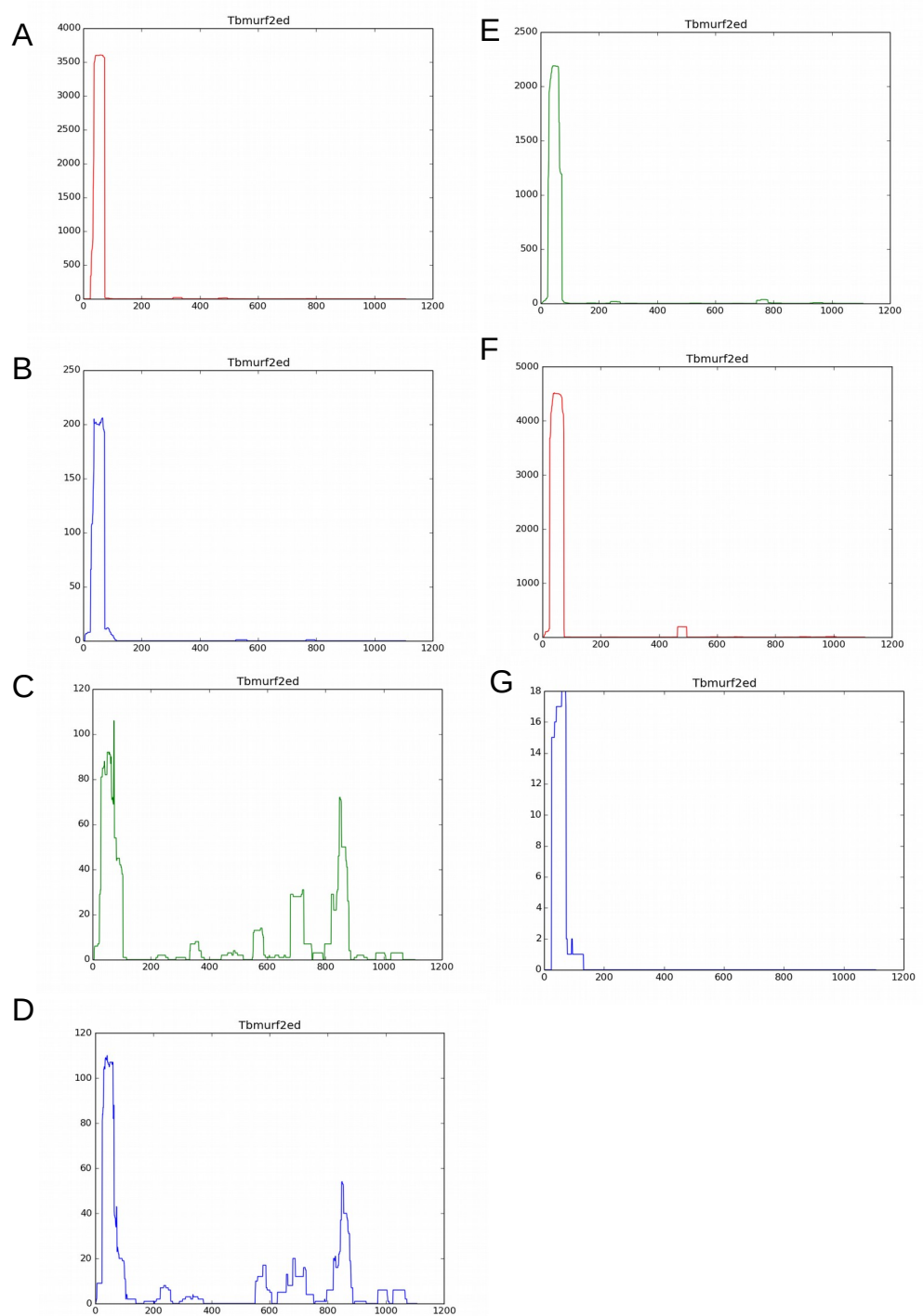


Figure 6.10: Edited murf2 gRNA depth plots. **A:** AnTat90.13 BSF, **B:** AnTat90.13 PCF, **C:** TREU667 BSF, **D:** TREU667 PCF, **E:** EATRO 164 PCF (Koslowsky et al. 2014), **F:** L427\_29.13A (Suematsu et al 2016), **G:** L427\_29.13A (Madina et al. 2014).

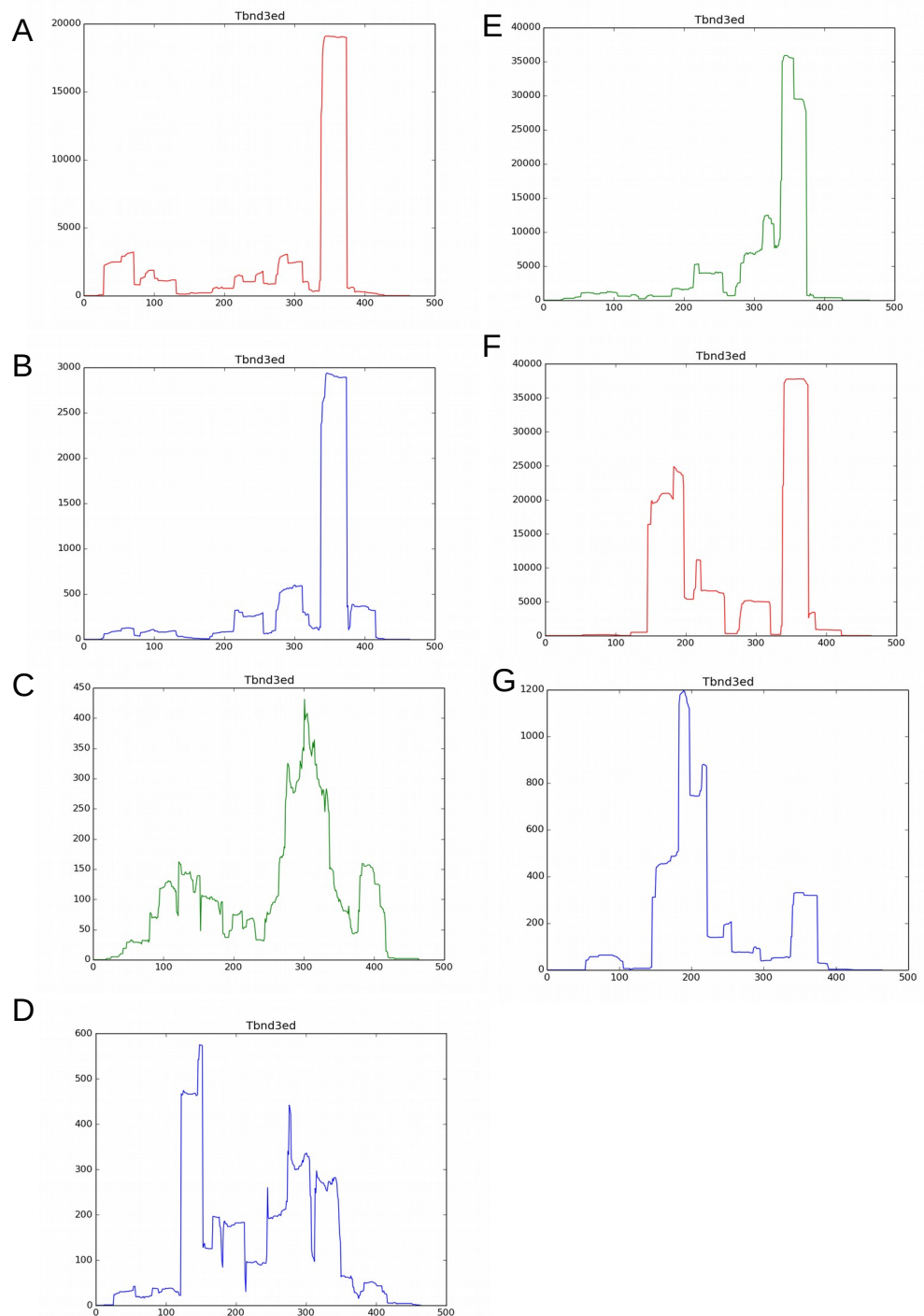
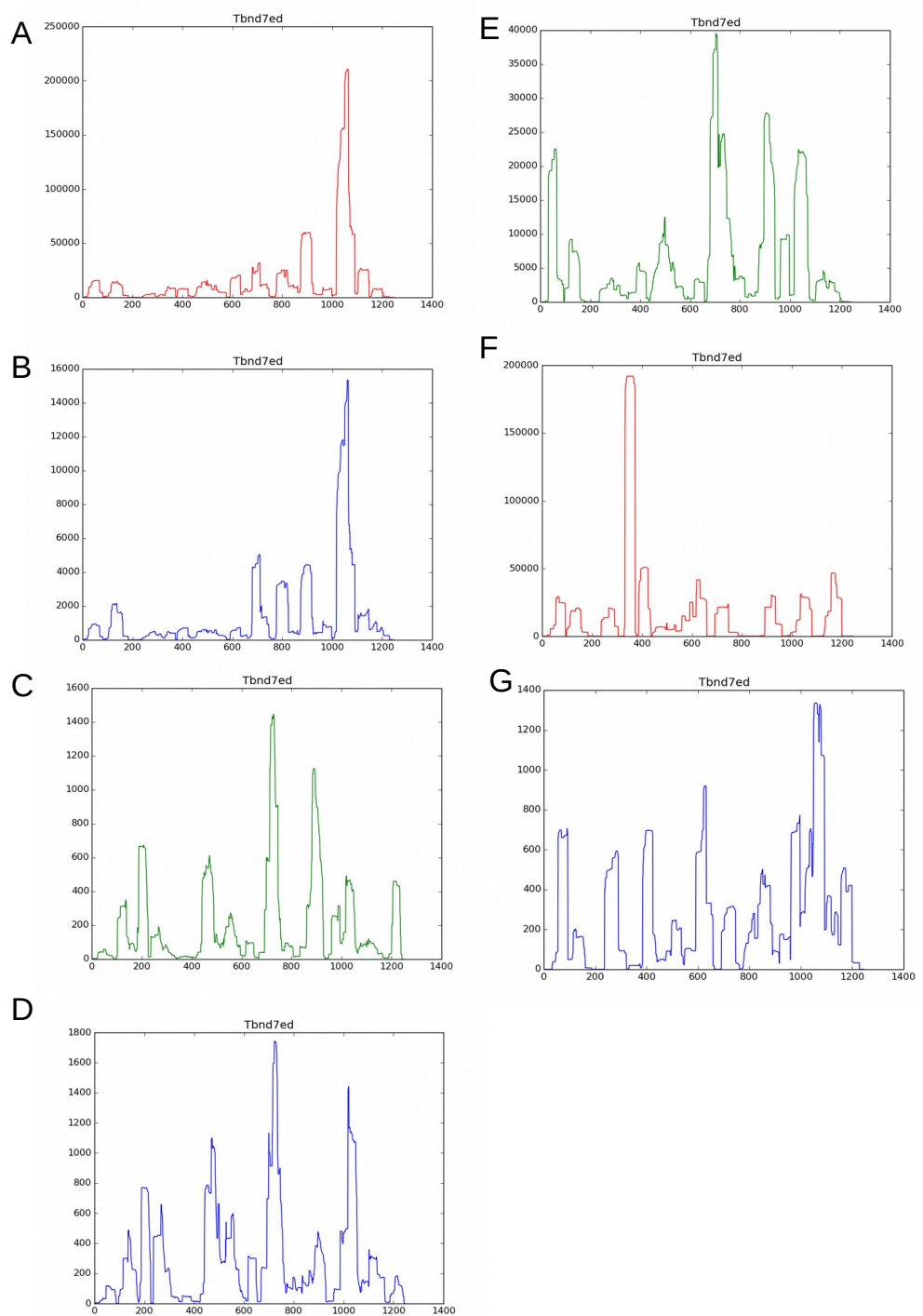
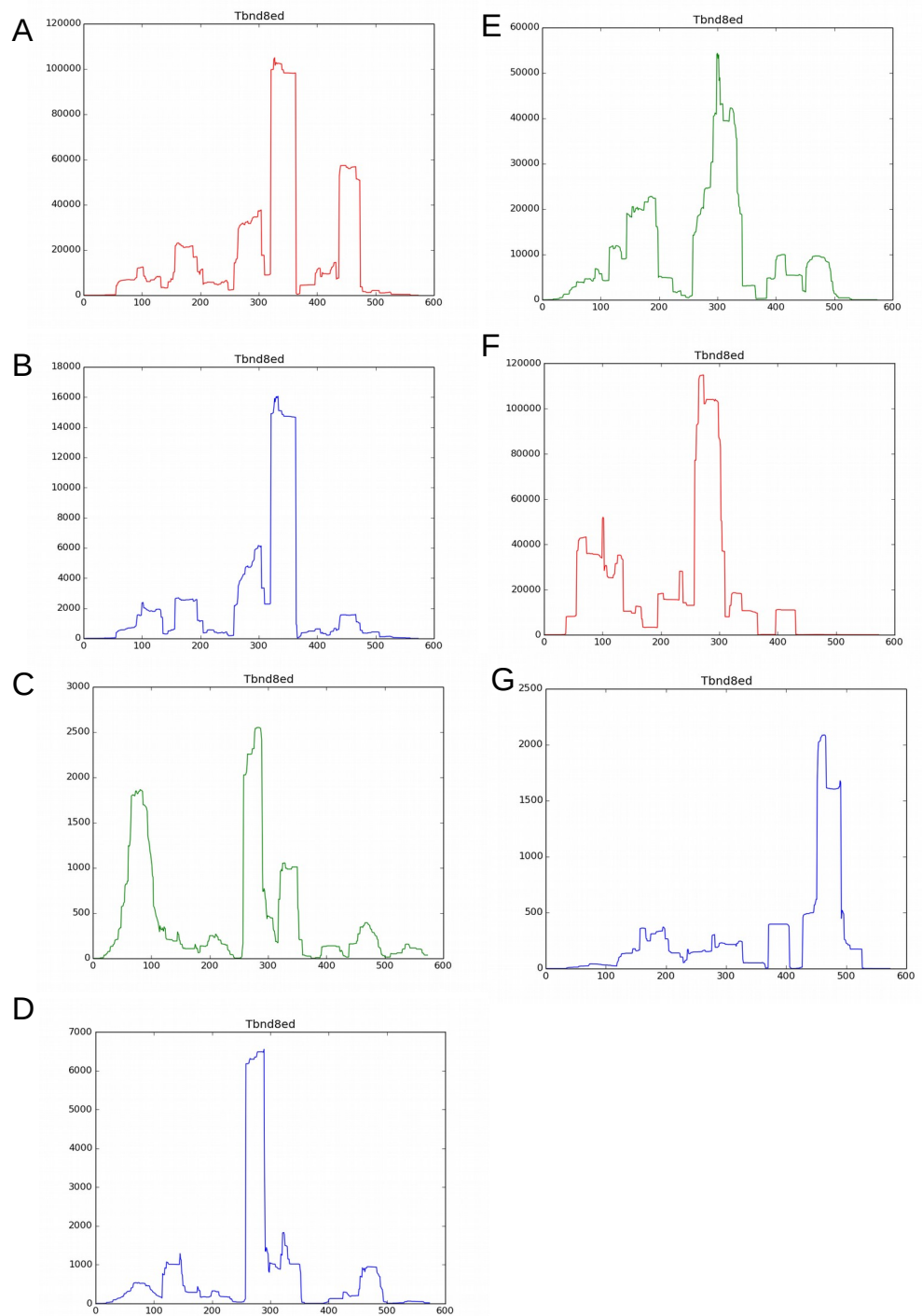


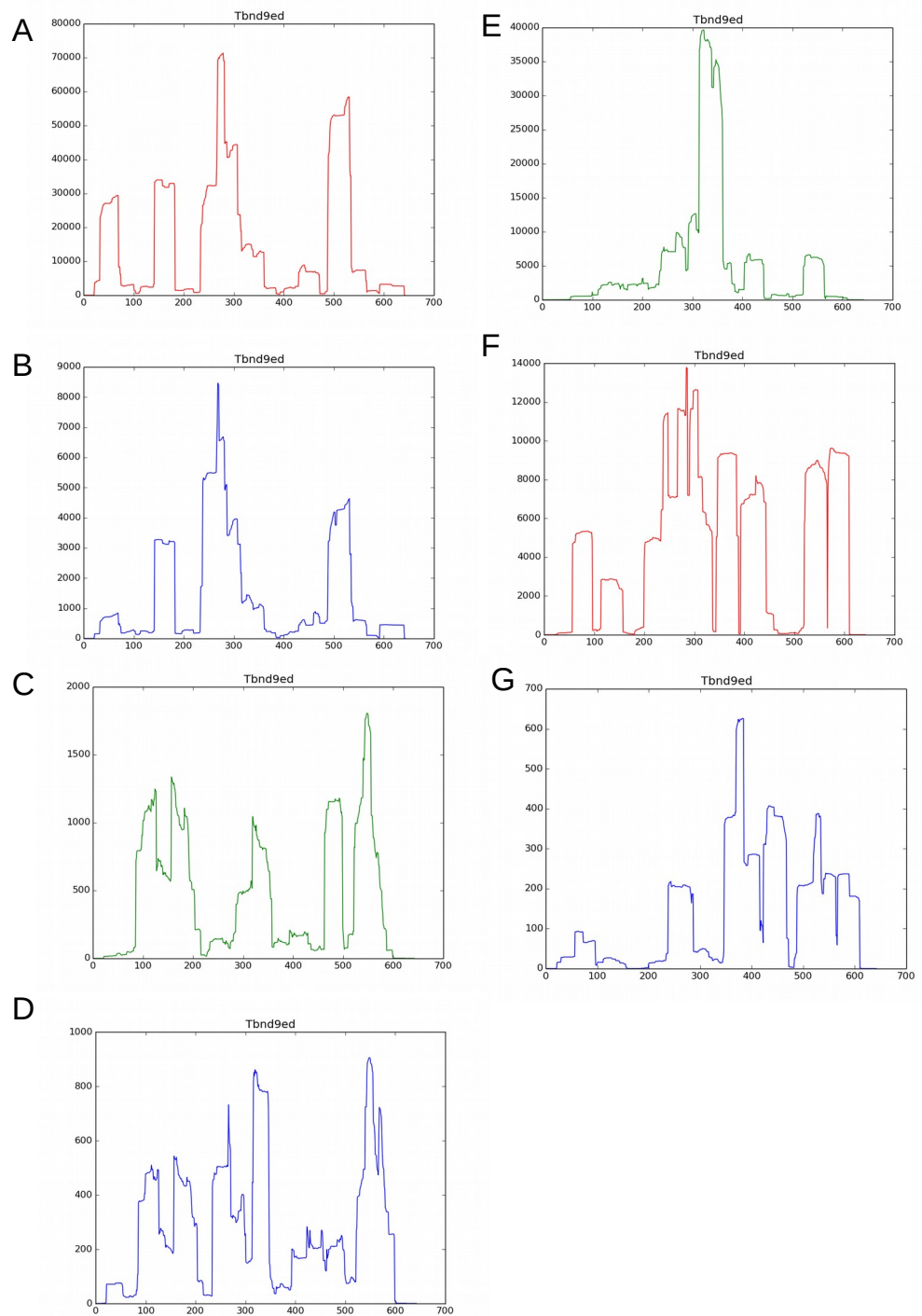
Figure 6.11: Edited *nd3* gRNA depth plots. **A:** AnTat90.13 BSF, **B:** AnTat90.13 PCF, **C:** TREU667 BSF, **D:** TREU667 PCF, **E:** EATRO 164 PCF (Koslowsky et al. 2014), **F:** L427\_29.13A (Suematsu et al 2016), **G:** L427\_29.13A (Madina et al. 2014).



**Figure 6.12: Edited *nd7* gRNA depth plots. A: AnTat90.13 BSF, B: AnTat90.13 PCF, C: TREU667 BSF, D: TREU667 PCF, E: EATRO 164 PCF (Koslowsky et al. 2014), F: L427\_29.13A (Suematsu et al 2016), G: L427\_29.13A (Madina et al. 2014).**



**Figure 6.13: Edited *nd8* gRNA depth plots. A: AnTat90.13 BSF, B: AnTat90.13 PCF, C: TREU667 BSF, D: TREU667 PCF, E: EATRO 164 PCF (Koslowsky et al. 2014), F: L427\_29.13A (Suematsu et al 2016), G: L427\_29.13A (Madina et al. 2014).**



**Figure 6.14: Edited *nd9* gRNA depth plots. A: AnTat90.13 BSF, B: AnTat90.13 PCF, C: TREU667 BSF, D: TREU667 PCF, E: EATRO 164 PCF (Koslowsky et al. 2014), F: L427\_29.13A (Suematsu et al 2016), G: L427\_29.13A (Madina et al. 2014).**

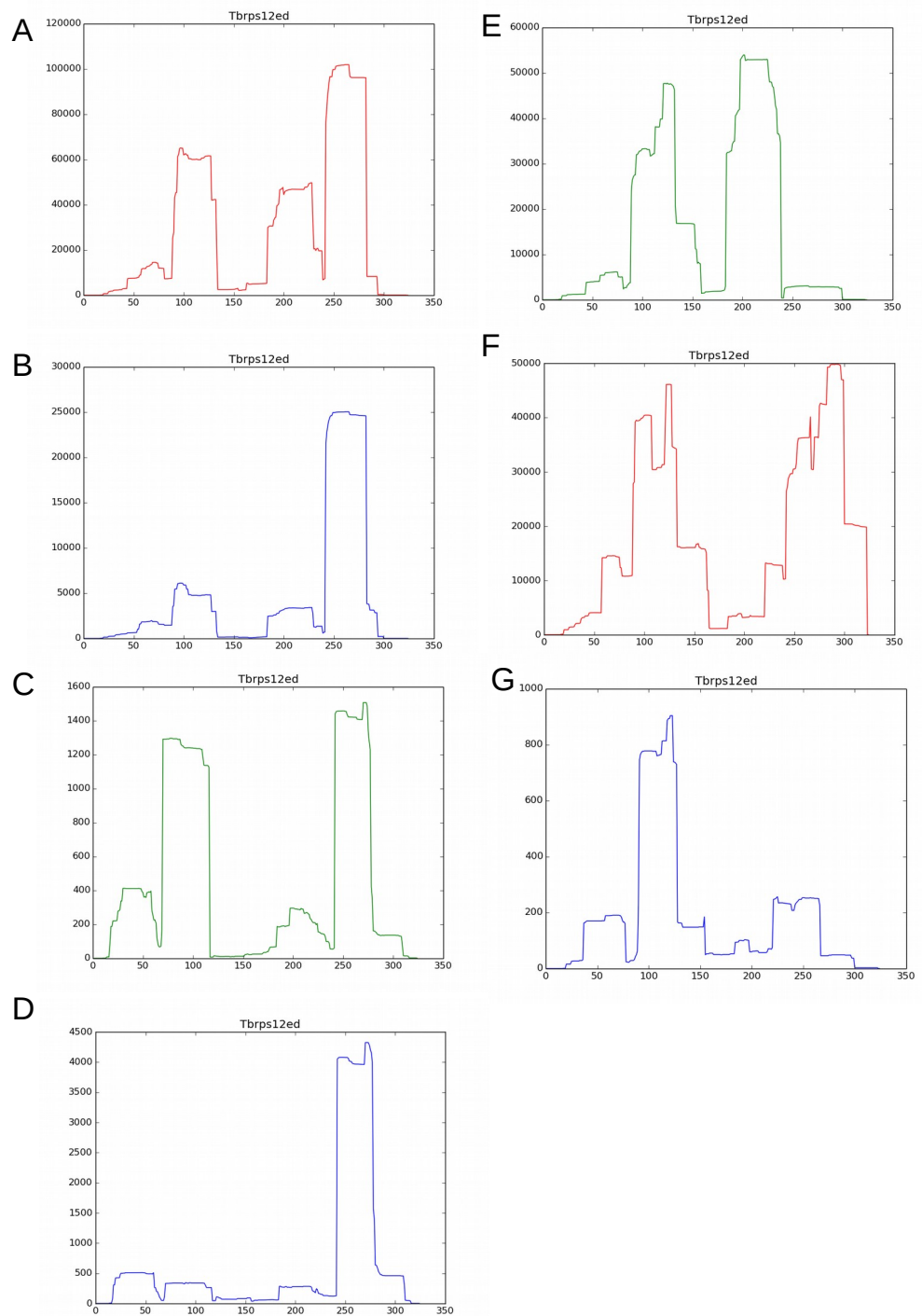


Figure 6.15 Edited *rps12* gRNA depth plots. **A:** AnTat90.13 BSF, **B:** AnTat90.13 PCF, **C:** TREU667 BSF, **D:** TREU667 PCF, **E:** EATRO 164 PCF (Koslowsky et al. 2014), **F:** L427\_29.13A (Suematsu et al 2016), **G:** L427\_29.13A (Madina et al. 2014).

## 6.4 A mutant cell line has minicircles enriched for A6 and RPS12 gRNAs

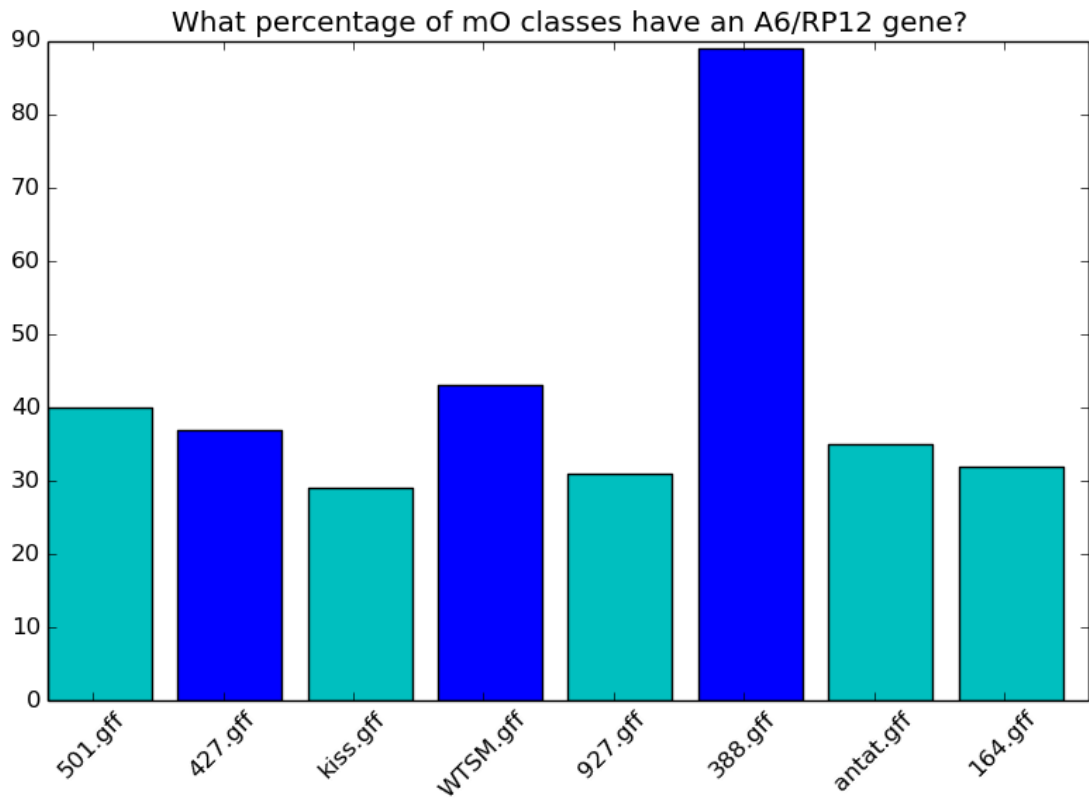


Figure 6.16: The minicircles from the mutant cell line 388.5 almost exclusively encode for A6/RPS12 gRNAs. The y axis represents % of minicircles encoding an A6 or RPS12 gRNA. Cyan bars are samples reported as being differentiation competent, blue bars are samples which are reported as being monomorphic.

## 6.5 *T. b. gambiense* minicircle complexity

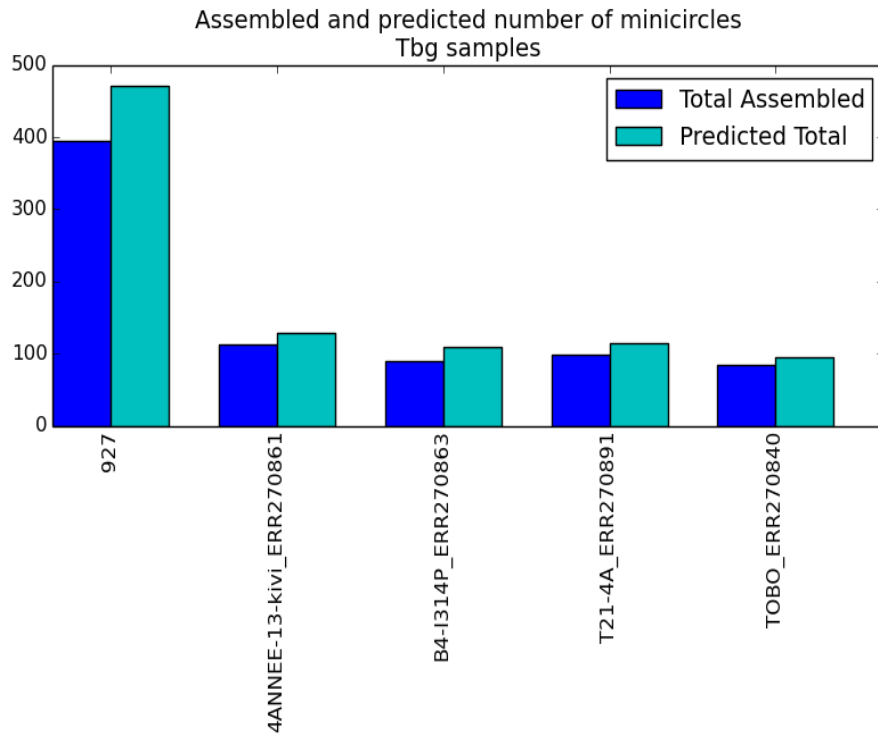


Figure 6.17: *T. b. gambiense* type I isolates are predicted to have a streamlined kDNA genome. The y-axis shows the number of minicircles, either assembled (blue bars) or predicted based on CSB3 mapping (Cyan). *T. b. b.* 927 is shown for comparison.

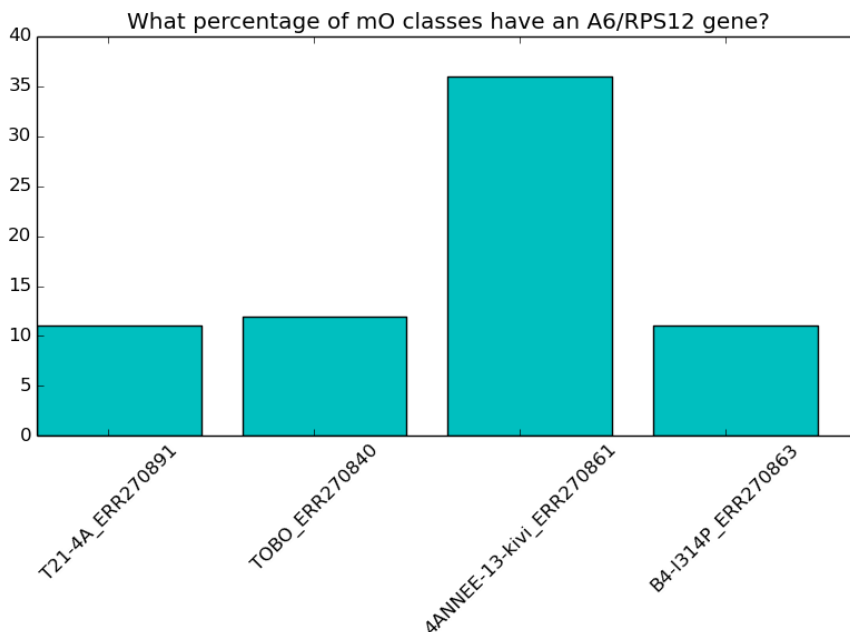


Figure 6.18: The proportion of assembled *T. b. g* type I minicircles with an A6/RPS12 gRNA. Y-axis shows percentage of minicircles with A6/RPS12 gRNA. Isolate 4ANNEE-13-kivi\_ERR270861 is enriched for A6/RPS12 gRNAs compared to others. Phylogenetic (12SSU) analysis shows that this isolate comes from a different clade (data not shown).



## 6.6 Long term time-course samples

Date	Sample ID	Date	Sample ID
05/06/15	<b>T0</b>	22/09/15	T16A
11/06/15	T1A	22/09/15	T16B
11/06/15	T1B	22/09/15	T16C
11/06/15	T1C	28/09/15	T17A
17/06/15	T2A	28/09/15	T17B
17/06/15	T2B	28/09/15	T17C
17/06/15	T2C	06/10/15	T18A
23/06/15	T3A	06/10/15	T18B
23/06/15	T3B	06/10/15	T18C
23/06/15	T3C	13/10/15	T19A
30/06/15	T4A	13/10/15	T19B
30/06/15	T4B	13/10/15	T19C
30/06/15	T4C	21/10/15	T20A
07/07/15	T5A	21/10/15	T20B
07/07/15	T5B	21/10/15	T20C
07/07/15	T5C	27/10/15	T21A
14/07/15	T6A	27/10/15	T21B
14/07/15	T6B	27/10/15	T21C
14/07/15	T6C	03/11/15	T22A
21/07/15	T7A	03/11/15	T22B
21/07/15	T7B	03/11/15	T22C
21/07/15	T7C	10/11/15	T23A
28/07/15	T8A	10/11/15	T23B
28/07/15	T8B	10/11/15	T23C
28/07/15	T8C	17/11/15	T24A
04/08/15	T9A	17/11/15	T24B
04/08/15	T9B	17/11/15	T24C
04/08/15	T9C	24/11/15	T25A
11/08/15	T10A	24/11/15	T25B
11/08/15	T10B	24/11/15	T25C
11/08/15	T10C	01/12/15	T26A
18/08/15	T11A	01/12/15	T26B
18/08/15	T11B	01/12/15	T26C
18/08/15	T11C	08/12/15	T27A
25/08/15	T12A	08/12/15	T27B
25/08/15	T12B	08/12/15	T27C
25/08/15	T12C	15/12/15	T28A
08/09/15	T14A	15/12/15	<b>T28B</b>
08/09/15	T14B	15/12/15	<b>T28C</b>
08/09/15	T14C	22/12/15	<b>T29A</b>
15/09/15	T15A	22/12/15	T29C
15/09/15	T15B		
15/09/15	T15C		

Table 6.2 *kDNA* timecourse dates and sample IDs.

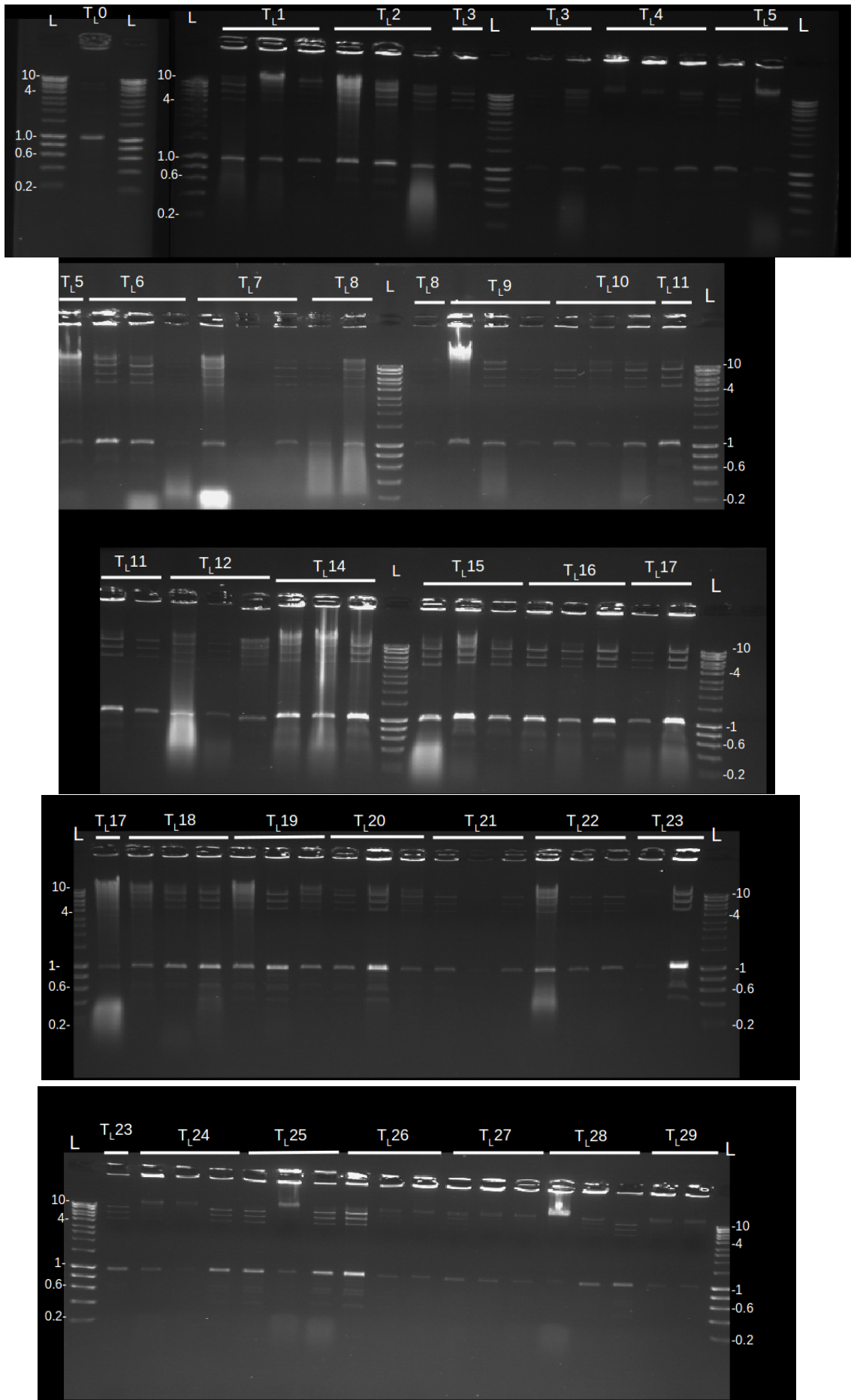


Figure 6.19: *EcoRI* digested time course samples. Samples from the long term ( $T_L$ ) time-course were digested with *EcoRI*. The replicates for each samples are in the order A,B,C.

## 6.7 Core chromosomal regions

For calculating copy numbers of minicircles and maxicircles the core chromosomal regions excluding repetitive sequences in telomeres were used as a standard for calculating sequencing coverage for total DNA prepared samples.

Chromosome	Core	
	Start	End
Tb427_01_v4	198051	521079
Tb427_01_v4	525253	776624
Tb427_01_v4	778767	989445
Tb427_02_v4	249668	354729
Tb427_02_v4	356197	393055
Tb427_02_v4	394391	959689
Tb427_02_v4	960866	1166450
Tb427_03_v4	123336	374531
Tb427_03_v4	397441	632831
Tb427_03_v4	638957	1585999
Tb427_04_v4	78295	228896
Tb427_04_v4	230349	319467
Tb427_04_v4	320464	848254
Tb427_04_v4	858326	1477029
Tb427_05_v4	65902	97652
Tb427_05_v4	99117	1236330
Tb427_05_v4	1237765	1374080
Tb427_06_v4	22580	115241
Tb427_06_v4	115467	851666
Tb427_06_v4	853350	992282
Tb427_06_v4	995020	1421548
Tb427_06_v4	1610566	1612057
Tb427_07_v4	24286	465662
Tb427_07_v4	512455	836328
Tb427_07_v4	839908	1768835
Tb427_07_v4	1782886	1929787
Tb427_07_v4	1931231	2205233
Tb427_08_v4	119912	125411
Tb427_08_v4	134250	1085475
Tb427_08_v4	1095115	2096001
Tb427_08_v4	2100983	2481190
Tb427_09_v4	166340	188039
Tb427_09_v4	319440	1133102
Tb427_09_v4	1170294	2396156
Tb427_09_v4	2464737	2484994
Tb427_09_v4	2802191	2810455
Tb427_09_v4	2938901	2943900
Tb427_10_v5	55118	3937721
Tb427_10_v5	3938229	3977607
Tb427_10_v5	3981373	4004720

*Table 6.3: Core chromosomal regions excluding any highly repetitive sequences for T. b. brucei Lister 427. Kindly provided by Bernardo Foth (Wellcome Trust Sanger Institute, Hinxton)*

## 7. Bibliography

- Abu-Elneel, K., Kapeller, I., and Shlomai, J. (1999). Universal minicircle sequence-binding protein, a sequence-specific DNA-binding protein that recognizes the two replication origins of the kinetoplast DNA minicircle. *J. Biol. Chem.* 274: 13419–26.
- Adler, B.K., Harris, M.E., Bertrand, K.I., and Hajduk, S.L. (1991). Modification of *Trypanosoma brucei* mitochondrial rRNA by posttranscriptional 3' polyuridine tail formation. *Mol. Cell. Biol.* 11: 5878–84.
- Alsford, S., Turner, D.J., Obado, S.O., Sanchez-Flores, A., Glover, L., Berriman, M., et al. (2011). High-throughput phenotyping using parallel sequencing of RNA interference targets in the African trypanosome. *Genome Res.* 21: 915–24.
- Ammerman, M.L., Downey, K.M., Hashimi, H., Fisk, J.C., Tomasello, D.L., Faktorová, D., et al. (2012). Architecture of the trypanosome RNA editing accessory complex, MRB1. *Nucleic Acids Res.* 40: 5637–5650.
- Aphasizhev, R., and Aphasizheva, I. (2014). Mitochondrial RNA editing in trypanosomes: small RNAs in control. *Biochimie* 100: 125–31.
- Aphasizhev, R., Aphasizheva, I., and Simpson, L. (2003). A tale of two TUTases. *Proc Natl Acad Sci U S A* 100: 10617–10622.
- Aphasizheva, I., and Aphasizhev, R. (2010). RET1-catalyzed uridylylation shapes the mitochondrial transcriptome in *Trypanosoma brucei*. *Mol. Cell. Biol.* 30: 1555–67.
- Aphasizheva, I., and Aphasizhev, R. (2016). U-Insertion/Deletion mRNA-Editing Holoenzyme: Definition in Sight. *Trends Parasitol.* 32: 144–56.
- Aphasizheva, I., Maslov, D.A., Qian, Y., Huang, L., Wang, Q., Costello, C.E., et al. (2016). Ribosome-associated pentatricopeptide repeat proteins function as translational activators in mitochondria of trypanosomes. *Mol. Microbiol.* 99: 1043–58.
- Aphasizheva, I., Maslov, D., Wang, X., Huang, L., and Aphasizhev, R. (2011). Pentatricopeptide repeat proteins stimulate mRNA adenylation/uridylation to activate mitochondrial translation in trypanosomes. *Mol. Cell* 42: 106–17.
- Aphasizheva, I., Zhang, L., Wang, X., Kaake, R.M., Huang, L., Monti, S., et al. (2014). RNA binding and core complexes constitute the U-insertion/deletion editosome. *Mol. Cell. Biol.* 34: 4329–42.
- Archibald, J.M. (2015). Endosymbiosis and eukaryotic cell evolution. *Curr. Biol.* 25: R911–R921.
- Arts, G.J., and Benne, R. (1996). Mechanism and evolution of RNA editing in kinetoplastida. *Biochim. Biophys. Acta - Gene Struct. Expr.* 1307: 39–54.
- Babokhov, P., Sanyaolu, A.O., Oyibo, W. a, Fagbenro-Beyioku, A.F., and Iriemenam,

- N.C. (2013). A current analysis of chemotherapy strategies for the treatment of human African trypanosomiasis. *Pathog. Glob. Health* 107: 242–52.
- Bailey, T.L. (2011). DREME: Motif discovery in transcription factor ChIP-seq data. *Bioinformatics* 27: 1653–1659.
- Bailey, T.L., Boden, M., Buske, F. a., Frith, M., Grant, C.E., Clementi, L., et al. (2009). MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Res.* 37: 202–208.
- Baker, N., Koning, H.P. de, Mäser, P., and Horn, D. (2013). Drug resistance in African trypanosomiasis: the melarsoprol and pentamidine story. *Trends Parasitol.* 29: 110–118.
- Bakker, B.M., Mensonides, F.I., Teusink, B., Hoek, P. van, Michels, P.A., and Westerhoff, H. V (2000). Compartmentation protects trypanosomes from the dangerous design of glycolysis. *Proc. Natl. Acad. Sci. U. S. A.* 97: 2087–92.
- Bakker, B.M., Michels, P.A.M., Opperdoes, F.R., and Westerhoff, H. V (1997). Glycolysis in Bloodstream Form *Trypanosoma brucei* Can Be Understood in Terms of the Kinetics of the Glycolytic Enzymes. *J. Biol. Chem.* 272: 3207–3215.
- Barrett, M.P., Vincent, I.M., Burchmore, R.J.S., Kazibwe, A.J.N., and Matovu, E. (2011). Drug resistance in human African trypanosomiasis. *Future Microbiol.* 6: 1037–47.
- Benne, R., Burg, J. Van den, Brakenhoff, J.P., Sloof, P., Boom, J.H. Van, and Tromp, M.C. (1986). Major transcript of the frameshifted *coxII* gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell* 46: 819–26.
- Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renauld, H., Bartholomeu, D.C., et al. (2005). The genome of the African trypanosome *Trypanosoma brucei*. *Science* 309: 416–22.
- Bienen, E.J., Maturi, R.K., Pollakis, G., and Clarkson, A.B. (1993). Non-cytochrome mediated mitochondrial ATP production in bloodstream form *Trypanosoma brucei brucei*. *Eur. J. Biochem.* 216: 75–80.
- Birkenmeyer, L., Sugisaki, H., and Ray, D.S. (1985). The majority of minicircle DNA in *Crithidia fasciculata* strain CF-C1 is of a single class with nearly homogeneous DNA sequence. *Nucleic Acids Res.* 13: 7107–18.
- Blom, D., Haan, A. de, Berg, M. van den, Sloof, P., Jirku, M., Lukes, J., et al. (1998). RNA editing in the free-living bodonid *Bodo saltans*. *Nucleic Acids Res.* 26: 1205–13.
- Blum, B., Bakalara, N., and Simpson, L. (1990). A model for RNA editing in kinetoplastid mitochondria: ‘guide’ RNA molecules transcribed from maxicircle DNA provide the edited information. *Cell* 60: 189–98.

- Blum, B., and Simpson, L. (1990). Guide RNAs in kinetoplastid mitochondria have a nonencoded 3' oligo(U) tail involved in recognition of the preedited region. *Cell* 62: 391–7.
- Blum, B., Sturm, N.R., Simpson, A.M., and Simpson, L. (1991). Chimeric gRNA-mRNA molecules with oligo(U) tails covalently linked at sites of RNA editing suggest that U addition occurs by transesterification. *Cell* 65: 543–50.
- Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114–20.
- Bonhivers, M., Landrein, N., Decossas, M., and Robinson, D.R. (2008). A monoclonal antibody marker for the exclusion-zone filaments of *Trypanosoma brucei*. *Parasit. Vectors* 1: 21.
- Borst, P. (1986). Discontinuous transcription and antigenic variation in trypanosomes. *Annu. Rev. Biochem.* 55: 701–32.
- Borst, P. (1991). Why kinetoplast DNA networks? *Trends Genet.* 7: 139–141.
- Borst, P., Fase-Fowler, F., and Gibson, W.C. (1987). Kinetoplast DNA of *Trypanosoma evansi*. *Mol. Biochem. Parasitol.* 23: 31–38.
- Borst, P., Fase-Fowler, F., Weijers, P.J., Barry, J.D., Tetley, L., and Vickerman, K. (1985). Kinetoplast DNA from *Trypanosoma vivax* and *T. congolense*. *Mol. Biochem. Parasitol.* 15: 129–142.
- Bossche, P. Van Den, Ky-Zerbo, A., Brandt, J., Marcotty, T., Geerts, S., and Deken, R. De (2005). Transmissibility of *Trypanosoma brucei* during its development in cattle. *Trop. Med. Int. Heal.* 10: 833–839.
- Bringaud, F., Barrett, M.P., and Zilberstein, D. (2012). Multiple roles of proline transport and metabolism in trypanosomatids. *Front. Biosci. (Landmark Ed.)* 17: 349–74.
- Brown, S. V, Hosking, P., Li, J., and Williams, N. (2006). ATP Synthase Is Responsible for Maintaining Mitochondrial Membrane Potential in Bloodstream Form *Trypanosoma brucei*. *Society* 5: 44–53.
- Bruhn, D.F., Sammartino, M.P., and Klingbeil, M.M. (2011). Three mitochondrial DNA polymerases are essential for kinetoplast DNA replication and survival of bloodstream form *Trypanosoma brucei*. *Eukaryot. Cell* 10: 734–43.
- Brun, R., Blum, J., Chappuis, F., and Burri, C. (2010). Human African trypanosomiasis. *Lancet (London, England)* 375: 148–59.
- Campbell, D.A., Thomas, S., and Sturm, N.R. (2003). Transcription in kinetoplastid protozoa: why be normal? *Microbes Infect.* 5: 1231–40.

- Capewell, P., Cooper, A., Clucas, C., Weir, W., and Macleod, A. (2015). A co-evolutionary arms race: trypanosomes shaping the human genome, humans shaping the trypanosome genome. *Parasitology* 142 Suppl: S108-19.
- Carnes, J., Anupama, A., Balmer, O., Jackson, A., Lewis, M., Brown, R., et al. (2015). Genome and phylogenetic analyses of *Trypanosoma evansi* reveal extensive similarity to *T. brucei* and multiple independent origins for dyskinetoplasty. *PLoS Negl. Trop. Dis.* 9: e3404.
- Carnes, J., Soares, C.Z., Wickham, C., and Stuart, K. (2011). Endonuclease associations with three distinct editosomes in *Trypanosoma brucei*. *J. Biol. Chem.* 286: 19320–30.
- Carnes, J., Trotter, J.R., Peltan, A., Fleck, M., and Stuart, K. (2008). RNA editing in *Trypanosoma brucei* requires three different editosomes. *Mol. Cell. Biol.* 28: 122–30.
- Carver, T., Harris, S.R., Otto, T.D., Berriman, M., Parkhill, J., and McQuillan, J.A. (2013). BamView: Visualizing and interpretation of next-generation sequencing read alignments. *Brief. Bioinform.* 14: 203–212.
- Chaudhuri, M., Ajayi, W., Temple, S., and Hill, G.C. (1995). Identification and partial purification of a stage-specific 33 kDa mitochondrial protein as the alternative oxidase of the *Trypanosoma brucei brucei* bloodstream trypomastigotes. *J. Eukaryot. Microbiol.* 42: 467–72.
- Chen, J., Englund, P.T., and Cozzarelli, N.R. (1995a). Changes in network topology during the replication of kinetoplast DNA. *EMBO J.* 14: 6339–47.
- Chen, J., Rauch, C. a, White, J.H., Englund, P.T., and Cozzarelli, N.R. (1995b). The topology of the kinetoplast DNA network. *Cell* 80: 61–9.
- Clayton, C.E. (2002). Life without transcriptional control? From fly to man and back again. *EMBO J.* 21: 1881–8.
- Connor, R.J. (1994). The impact of nagana. *Onderstepoort J. Vet. Res.* 61: 379–83.
- Corell, R. a, Feagin, J.E., Riley, G.R., Strickland, T., Guderian, J. a, Myler, P.J., et al. (1993). *Trypanosoma brucei* minicircles encode multiple guide RNAs which can direct editing of extensively overlapping sequences. *Nucleic Acids Res.* 21: 4313–20.
- Cortez, a P., Ventura, R.M., Rodrigues, a C., Batista, J.S., Paiva, F., Añez, N., et al. (2006). The taxonomic and phylogenetic relationships of *Trypanosoma vivax* from South America and Africa. *Parasitology* 135: 1317–1328.
- Cross, G. a M., Kim, H.-S., and Wickstead, B. (2014). Capturing the variant surface glycoprotein repertoire (the VSGnome) of *Trypanosoma brucei* Lister 427. *Mol. Biochem. Parasitol.* 195: 59–73.
- Dean, S., Gould, M.K., Dewar, C.E., and Schnauffer, A.C. (2013a). Single point

- mutations in ATP synthase compensate for mitochondrial genome loss in trypanosomes. *Proc. Natl. Acad. Sci. U. S. A.* *110*: 14741–6.
- Dean, S., Gould, M.K., and Schnauffer, A. (2013b). Single point mutations in ATP synthase compensate for mitochondrial genome loss in non- tsetse transmitted trypanosomes. *Press na*: na.
- Dekker, J. (2002). Capturing Chromosome Conformation. *Science* (80-. ). *295*: 1306–1311.
- Desquesnes, M., Dargantes, A., Lai, D.-H., Lun, Z.-R., Holzmüller, P., and Jittapalpong, S. (2013). *Trypanosoma evansi* and *surra*: a review and perspectives on transmission, epidemiology and control, impact, and zoonotic aspects. *Biomed Res. Int.* *2013*: 321237.
- Desquesnes, M., and Dia, M.L. (2003). *Trypanosoma vivax*: mechanical transmission in cattle by one of the most common African tabanids, *Atylotus agrestis*. *Exp. Parasitol.* *103*: 35–43.
- Diao, Y., Rodriguez, V., Klingbeil, M., and Arsuaga, J. (2015). Orientation of DNA Minicircles Balances Density and Topological Complexity in Kinetoplast DNA. *PLoS One* *10*: e0130998.
- Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* *26*: 2460–1.
- Embley, T.M., and Hirt, R.P. (1998). Early branching eukaryotes? *Curr. Opin. Genet. Dev.* *8*: 624–629.
- Englund, P.T. (1979). Free minicircles of kinetoplast DNA in *Crithidia fasciculata*. *J. Biol. Chem.* *254*: 4895–900.
- Engstler, M., and Boshart, M. (2004). Cold shock and regulation of surface protein trafficking convey sensitization to inducers of stage differentiation in *Trypanosoma brucei*. *Genes Dev.* *18*: 2798–2811.
- Engstler, M., Pfohl, T., Herminghaus, S., Boshart, M., Wiegertjes, G., Heddergott, N., et al. (2007). Hydrodynamic Flow-Mediated Protein Sorting on the Cell Surface of Trypanosomes. *Cell* *131*: 505–515.
- Estévez, a M., and Simpson, L. (1999). Uridine insertion/deletion RNA editing in trypanosome mitochondria--a review. *Gene* *240*: 247–60.
- Etheridge, R.D., Aphasizheva, I., Gershon, P.D., and Aphasizhev, R. (2008). 3' adenylation determines mRNA abundance and monitors completion of RNA editing in *T. brucei* mitochondria. *EMBO J.* *27*: 1596–608.
- Eyob, E., and Matios, L. (2013). Review on camel trypanosomosis (*surra*) due to *Trypanosoma evansi*: Epidemiology and host response. *J. Vet. Med. Anim. Heal.* *5*:



334–343.

Fairlamb, A.H., Weislogel, P.O., Hoeijmakers, J.H., and Borst, P. (1978). Isolation and characterization of kinetoplast DNA from bloodstream form of *Trypanosoma brucei*. *J. Cell Biol.* 76: 293–309.

Faktorová, D., Dobáková, E., Peña-Díaz, P., and Lukeš, J. (2016). From simple to supercomplex: mitochondrial genomes of euglenozoan protists. *F1000Research* 5: 392.

Feagin, J.E., Shaw, J.M., Simpson, L., and Stuart, K. (1988). Creation of AUG initiation codons by addition of uridines within cytochrome b transcripts of kinetoplastids. *Proc. Natl. Acad. Sci. U. S. A.* 85: 539–43.

Fèvre, E.M., Wissmann, B. v., Welburn, S.C., and Lutumba, P. (2008). The Burden of Human African Trypanosomiasis. *PLoS Negl. Trop. Dis.* 2: e333.

Fisk, J.C., Ammerman, M.L., Presnyak, V., and Read, L.K. (2008). TbRGG2, an essential RNA editing accessory factor in two *trypanosoma brucei* life cycle stages. *J. Biol. Chem.* 283: 23016–23025.

Forrester, S.J., and Hall, N. (2014). The revolution of whole genome sequencing to study parasites. *Mol. Biochem. Parasitol.* 195: 77–81.

Franco, J.R., Simarro, P.P., Diarra, A., and Jannin, J.G. (2014). Epidemiology of human African trypanosomiasis. *Clin. Epidemiol.* 6: 257–275.

Freier, S.M., Kierzek, R., Jaeger, J. a, Sugimoto, N., Caruthers, M.H., Neilson, T., et al. (1986). Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci. U. S. A.* 83: 9373–7.

Gao, G., Kapushoc, S.T., Simpson, A.M., Thiemann, O.H., and Simpson, L. (2001). Guide RNAs of the recently isolated LEM125 strain of *Leishmania tarentolae*: an unexpected complexity. *RNA* 7: 1335–47.

Gibson, W., Crow, M., and Kearns, J. (1997). Kinetoplast DNA minicircles are inherited from both parents in genetic crosses of *Trypanosoma brucei*. *Parasitol. Res.* 83: 483–8.

Giezen, M. van der (2011). Mitochondria and the Rise of Eukaryotes. *Bioscience* 61: 594–601.

Gott, J.M., and Emeson, R.B. (2000). Functions and mechanisms of RNA editing. *Annu. Rev. Genet.* 34: 499–531.

Grams, J., McManus, M.T., and Hajduk, S.L. (2000). Processing of polycistronic guide RNAs is associated with RNA editing complexes in *Trypanosoma brucei*. *EMBO J.* 19: 5525–32.

Grams, J., Morris, J.C., Drew, M.E., Wang, Z., Englund, P.T., and Hajduk, S.L. (2002). A trypanosome mitochondrial RNA polymerase is required for transcription and

replication. *J. Biol. Chem.* 277: 16952–9.

Gray, M.W. (2003). Diversity and evolution of mitochondrial RNA editing systems. *IUBMB Life* 55: 227–33.

Gray, M.W. (2014). The pre-endosymbiont hypothesis: a new perspective on the origin and evolution of mitochondria. *Cold Spring Harb. Perspect. Biol.* 6:.

Gray, M.W., Burger, G., and Lang, B.F. (1999). Mitochondrial evolution. *Science* 283: 1476–81.

Greca, F. La, and Magez, S. (2011). Vaccination against trypanosomiasis: can it be done or is the trypanosome truly the ultimate immune destroyer and escape artist? *Hum. Vaccin.* 7: 1225–33.

Greif, G., Rodriguez, M., Reyna-Bello, A., Robello, C., and Alvarez-Valin, F. (2015). Kinetoplast adaptations in American strains from *Trypanosoma vivax*. *Mutat. Res.* 773: 69–82.

Guilbride, D.L., and Englund, P.T. (1998). The replication mechanism of kinetoplast DNA networks in several trypanosomatid species. *J. Cell Sci.* 111 ( Pt 6): 675–9.

Gunasekera, K., Wüthrich, D., Braga-Lagache, S., Heller, M., and Ochsenreiter, T. (2012). Proteome remodelling during development from blood to insect-form *Trypanosoma brucei* quantified by SILAC and mass spectrometry. *BMC Genomics* 13: 556.

Haeseler, a von, Blum, B., Simpson, L., Sturm, N., and Waterman, M.S. (1992). Computer methods for locating kinetoplastid cryptogenes. *Nucleic Acids Res.* 20: 2717–24.

Hajduk, S., and Ochsenreiter, T. (2010). RNA editing in kinetoplastids. *RNA Biol.* 7: 229–236.

Hall, J.P.J., Wang, H., and Barry, J.D. (2013). Mosaic VSGs and the scale of *Trypanosoma brucei* antigenic variation. *PLoS Pathog.* 9: e1003502.

Harris, M.E., Moore, D.R., and Hajduk, S.L. (1990a). Addition of uridines to edited RNAs in trypanosome mitochondria occurs independently of transcription. *J. Biol. Chem.* 265: 11368–11376.

Harris, M.E., Moore, D.R., and Hajduk, S.L. (1990b). Addition of uridines to edited RNAs in trypanosome mitochondria occurs independently of transcription. *J. Biol. Chem.* 265: 11368–76.

Hashimi, H., Cicová, Z., Novotná, L., Wen, Y.-Z., and Lukes, J. (2009). Kinetoplastid guide RNA biogenesis is dependent on subunits of the mitochondrial RNA binding complex 1 and mitochondrial RNA polymerase. *RNA* 15: 588–99.

Hashimi, H., Zíková, A., Panigrahi, A.K., Stuart, K.D., and Lukes, J. (2008). TbRGG1,

an essential protein involved in kinetoplastid RNA metabolism that is associated with a novel multiprotein complex. *RNA* 14: 970–80.

Hashimi, H., Zimmer, S.L., Ammerman, M.L., Read, L.K., and Lukeš, J. (2013). Dual core processing: MRB1 is an emerging kinetoplast RNA editing complex. *Trends Parasitol.* 29: 91–9.

Hellemond, J.J. van, Opperdoes, F.R., and Tielens, a G.M. (2005). The extraordinary mitochondrion and unusual citric acid cycle in *Trypanosoma brucei*. *Biochem. Soc. Trans.* 33: 967–71.

Helwak, A., Kudla, G., Dudnakova, T., and Tollervy, D. (2013). Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* 153: 654–665.

Heo, I., Ha, M., Lim, J., Yoon, M.-J., Park, J.-E., Kwon, S.C., et al. (2012). Mono-uridylation of pre-microRNA as a key step in the biogenesis of group II let-7 microRNAs. *Cell* 151: 521–32.

Hermann, T., Schmid, B., Heumann, H., and Göringer, H.U. (1997). A three-dimensional working model for a guide RNA from *Trypanosoma brucei*. *Nucleic Acids Res.* 25: 2311–8.

Hernandez, A., Madina, B.R., Ro, K., Wohlschlegel, J.A., Willard, B., Kinter, M.T., et al. (2010). REH2 RNA helicase in kinetoplastid mitochondria: ribonucleoprotein complexes and essential motifs for unwinding and guide RNA (gRNA) binding. *J. Biol. Chem.* 285: 1220–8.

Hines, J.C., and Ray, D.S. (2010). A mitochondrial DNA primase is essential for cell growth and kinetoplast DNA replication in *Trypanosoma brucei*. *Mol. Cell. Biol.* 30: 1319–28.

Hines, J.C., and Ray, D.S. (2011). A second mitochondrial DNA primase is essential for cell growth and kinetoplast minicircle DNA replication in *Trypanosoma brucei*. *Eukaryot. Cell* 10: 445–54.

Hirumi, H., and Hirumi, K. (1989). Continuous cultivation of *Trypanosoma brucei* blood stream forms in a medium containing a low concentration of serum protein without feeder cell layers. *J. Parasitol.* 75: 985–9.

Hong, M., and Simpson, L. (2003). Genomic organization of *Trypanosoma brucei* kinetoplast DNA minicircles. *Protist* 154: 265–79.

Horstmann, D.M. (1974). Importance of disease surveillance. *Prev. Med. (Baltim).* 3: 436–42.

Hotez, P., Molyneux, D., and Fenwick, A. (2007). Control of neglected tropical diseases. *N Engl J Med* 357: 1018–27.

- Hotez, P.J., and Kamath, A. (2009). Neglected tropical diseases in sub-saharan Africa: review of their prevalence, distribution, and disease burden. *PLoS Negl. Trop. Dis.* 3: e412.
- Huang, X., and Madan, A. (1999). CAP3 : A DNA Sequence Assembly Program. 868–877.
- Ivens, A.C., Peacock, C.S., Worthey, E. a, Murphy, L., Aggarwal, G., Berriman, M., et al. (2005). The genome of the kinetoplastid parasite, *Leishmania major*. *Science* 309: 436–42.
- Jackson, A.P. (2014). Genome evolution in trypanosomatid parasites. *Parasitology* 1–17.
- Jackson, A.P., Goyard, S., Xia, D., Foth, B.J., Sanders, M., Wastling, J.M., et al. (2015). Global gene expression profiling through the complete life cycle of *Trypanosoma vivax*. *PLoS Negl. Trop. Dis.* 9: 1–29.
- Jensen, R.E., and Englund, P.T. (2012). Network News: The Replication of Kinetoplast DNA. *Annu. Rev. Microbiol.* 66: 473–491.
- Jensen, R.E., Simpson, L., and Englund, P.T. (2008). What happens when *Trypanosoma brucei* leaves Africa. *Trends Parasitol.* 24: 428–31.
- Jørgensen, T.S., Xu, Z., Hansen, M.A., Sørensen, S.J., and Hansen, L.H. (2014). Hundreds of circular novel plasmids and DNA elements identified in a rat cecum metatranscriptome. *PLoS One* 9: e87924.
- Junker, J.P., and Oudenaarden, A. Van (2014). Every cell is special: Genome-wide studies add a new dimension to single-cell biology. *Cell* 157: 8–11.
- Kabani, S., Waterfall, M., and Matthews, K.R. (2010). Cell-cycle synchronisation of bloodstream forms of *Trypanosoma brucei* using Vybrant DyeCycle Violet-based sorting. *Mol. Biochem. Parasitol.* 169: 59–62.
- Kable, M.L., Seiwert, S.D., Heidmann, S., and Stuart, K. (1996). A Mechanism for Editing : Uridylate mRNA into Precursor Insertion. 273: 1189–1195.
- Kao, C.-Y., and Read, L.K. (2005). Opposing effects of polyadenylation on the stability of edited and unedited mitochondrial RNAs in *Trypanosoma brucei*. *Mol. Cell. Biol.* 25: 1634–44.
- Kim, B., Ha, M., Loeff, L., Chang, H., Simanshu, D.K., Li, S., et al. (2015). TUT7 controls the fate of precursor microRNAs by using three different uridylation mechanisms. *EMBO J.* 34: 1801–15.
- Kirby, L.E., Sun, Y., Judah, D., Nowak, S., and Koslowsky, D. (2016). Analysis of the *Trypanosoma brucei* EATRO 164 Bloodstream Guide RNA Transcriptome. *PLoS Negl. Trop. Dis.* 10: e0004793.

- Kleisen, C.M., Weislogel, P.O., Fonck, K., and Borst, P. (1976). The structure of kinetoplast DNA. 2. Characterization of a novel component of high complexity present in the kinetoplast DNA network of *Crithidia luciliae*. *Eur. J. Biochem.* *64*: 153–60.
- Koslowsky, D., Sun, Y., Hindenach, J., Theisen, T., and Lucas, J. (2014). The insect-phase gRNA transcriptome in *Trypanosoma brucei*. *Nucleic Acids Res.* *42*: 1873–86.
- Koslowsky, D.J., Bhat, G.J., Read, L.K., and Stuart, K. (1991). Cycles of progressive realignment of gRNA with mRNA in RNA editing. *Cell* *67*: 537–46.
- Koslowsky, D.J., Riley, G.R., Feagin, J.E., and Stuart, K. (1992). Guide RNAs for transcripts with developmentally regulated RNA editing are present in both life cycle stages of *Trypanosoma brucei*. *Mol. Cell. Biol.* *12*: 2043–9.
- Koslowsky, D.J., and Yahampath, G. (1997). Mitochondrial mRNA 3' cleavage/polyadenylation and RNA editing in *Trypanosoma brucei* are independent events. *Mol. Biochem. Parasitol.* *90*: 81–94.
- Kumar, V., Madina, B.R., Gulati, S., Vashisht, A.A., Kanyumbu, C., Pieters, B., et al. (2016). REH2C Helicase and GRBC Subcomplexes May Base Pair through mRNA and Small Guide RNA in Kinetoplastid Editosomes. *J. Biol. Chem.* *291*: 5753–64.
- Lai, D.-H., Hashimi, H., Lun, Z.-R., Ayala, F.J., and Lukes, J. (2008). Adaptations of *Trypanosoma brucei* to gradual loss of kinetoplast DNA: *Trypanosoma equiperdum* and *Trypanosoma evansi* are petite mutants of *T. brucei*. *Proc. Natl. Acad. Sci. U. S. A.* *105*: 1999–2004.
- Lamour, N., Rivière, L., Coustou, V., Coombs, G.H., Barrett, M.P., and Bringaud, F. (2005). Proline metabolism in procyclic *Trypanosoma brucei* is down-regulated in the presence of glucose. *J. Biol. Chem.* *280*: 11902–11910.
- Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* *9*: 357–359.
- Lavrov, D. V., Adamski, M., Chevaldonné, P., and Adamska, M. (2016). Extensive Mitochondrial mRNA Editing and Unusual Mitochondrial Genome Organization in Calcarean Sponges. *Curr. Biol.* *26*: 86–92.
- Laxman, S., Riechers, A., Sadilek, M., Schwede, F., and Beavo, J. a (2006). Hydrolysis products of cAMP analogs cause transformation of *Trypanosoma brucei* from slender to stumpy-like forms. *Proc. Natl. Acad. Sci. U. S. A.* *103*: 19194–19199.
- Lee, M.G., and Ploeg, L.H. Van der (1997). Transcription of protein-coding genes in trypanosomes by RNA polymerase I. *Annu. Rev. Microbiol.* *51*: 463–489.
- Legros, D., Ollivier, G., Gastellu-Etchegorry, M., Paquet, C., Burri, C., Jannin, J., et al. (2002). Treatment of human African trypanosomiasis—present situation and needs for research and development. *Lancet Infect. Dis.* *2*: 437–440.

- Leung, S.S., and Koslowsky, D.J. (2001). Interactions of mRNAs and gRNAs involved in trypanosome mitochondrial RNA editing : structure probing of an mRNA bound to its cognate gRNA Interactions of mRNAs and gRNAs involved in trypanosome mitochondrial RNA editing : Structure probing of an mRNA bou.
- Liu, B., Liu, Y., Motyka, S. a, Agbo, E.E.C., and Englund, P.T. (2005). Fellowship of the rings: the replication of kinetoplast DNA. *Trends Parasitol.* 21: 363–9.
- Loder, P.M.J. (1997). Size of blood meals taken by tsetse flies ( *Glossina* spp.) (Diptera: Glossinidae) correlates with fat reserves. *Bull. Entomol. Res.* 87: 547.
- Lukes, J., Guilbride, D.L., Voty, J., Zíková, A., Benne, R., and Englund, P.T. (2002). MINIREVIEW Kinetoplast DNA Network : Evolution of an Improbable Structure. 1: 495–502.
- Lukes, J., Hashimi, H., and Zíková, A. (2005). Unexplained complexity of the mitochondrial genome and transcriptome in kinetoplastid flagellates. *Curr. Genet.* 48: 277–99.
- Lukes, J., Leander, B.S., and Keeling, P.J. (2009). Cascades of convergent evolution: the corresponding evolutionary histories of euglenozoans and dinoflagellates. *Proc. Natl. Acad. Sci. U. S. A.* 106 *Suppl*: 9963–9970.
- Lun, Z.R., Lai, D.H., Li, F.J., Lukeš, J., and Ayala, F.J. (2010). *Trypanosoma brucei*: Two steps to spread out from Africa. *Trends Parasitol.* 26: 424–427.
- Macgregor, P., Rojas, F., Dean, S., and Matthews, K.R. (2013). Stable transformation of pleomorphic bloodstream form *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* 190: 60–2.
- MacGregor, P., Savill, N.J., Hall, D., and Matthews, K.R. (2011). Transmission stages dominate trypanosome within-host dynamics during chronic infections. *Cell Host Microbe* 9: 310–318.
- MacGregor, P., Szöör, B., Savill, N.J., and Matthews, K.R. (2012). Trypanosomal immune evasion, chronicity and transmission: an elegant balancing act. *Nat. Rev. Microbiol.* 10: 431–8.
- Madej, M.J., Alfonzo, J.D., and Hüttenhofer, A. (2007). Small ncRNA transcriptome analysis from kinetoplast mitochondria of *Leishmania tarentolae*. *Nucleic Acids Res.* 35: 1544–54.
- Madej, M.J., Niemann, M., Hüttenhofer, A., and Göringer, H.U. (2008a). Identification of novel guide RNAs from the mitochondria of *Trypanosoma brucei*. *RNA Biol.* 5: 84–91.
- Madej, M.J., Niemann, M., Hüttenhofer, A., and Göringer, H.U. (2008b). Identification of novel guide RNAs from the mitochondria of *Trypanosoma brucei*. *RNA Biol.* 5: 84–91.

- Madina, B.R., Kumar, V., Metz, R., Mooers, B.H.M., Bundschuh, R., and Cruz-Reyes, J. (2014). Native mitochondrial RNA-binding complexes in kinetoplastid RNA editing differ in guide RNA composition. *RNA* 20: 1142–52.
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17: 10.
- Maslov, D.A., Thiemann, O., and Simpson, L. (1994). Editing and misediting of transcripts of the kinetoplast maxicircle G5 (ND3) cryptogene in an old laboratory strain of *Leishmania tarentolae*. *Mol. Biochem. Parasitol.* 68: 155–9.
- Maslov, D. a, Kolesnikov, A. a, and Zaitseva, G.N. (1984). Conservative and divergent base sequence regions in the maxicircle kinetoplast DNA of several trypanosomatid flagellates. *Mol. Biochem. Parasitol.* 12: 351–364.
- Maslov, D. a, and Simpson, L. (1992). The polarity of editing within a multiple gRNA-mediated domain is due to formation of anchors for upstream gRNAs by downstream editing. *Cell* 70: 459–67.
- Matthews, K.R. (1999). Developments in the differentiation of *Trypanosoma brucei*. *Parasitol. Today* 15: 76–80.
- Matthews, K.R. (2005). The developmental cell biology of *Trypanosoma brucei*. *J. Cell Sci.* 118: 283–290.
- Mazet, M., Morand, P., Biran, M., Bouyssou, G., Courtois, P., Daulouède, S., et al. (2013). Revisiting the Central Metabolism of the Bloodstream Forms of *Trypanosoma brucei*: Production of Acetate in the Mitochondrion Is Essential for Parasite Viability. *PLoS Negl. Trop. Dis.* 7: 1–14.
- McCulloch, R., and Field, M.C. (2015). Quantitative sequencing confirms VSG diversity as central to immune evasion by *Trypanosoma brucei*. *Trends Parasitol.* 31: 346–349.
- McManus, M.T., Adler, B.K., Pollard, V.W., and Hajduk, S.L. (2000). *Trypanosoma brucei* guide RNA poly(U) tail formation is stabilized by cognate mRNA. *Mol. Cell. Biol.* 20: 883–91.
- McNicoll, F., Müller, M., Cloutier, S., Boilard, N., Rochette, A., Dubé, M., et al. (2005). Distinct 3'-Untranslated Region Elements Regulate Stage-specific mRNA Accumulation and Translation in *Leishmania*. *J. Biol. Chem.* 280: 35238–35246.
- Michels, P.A.M., Bringaud, F., Herman, M., and Hannaert, V. (2006). Metabolic functions of glycosomes in trypanosomatids. *Biochim. Biophys. Acta - Mol. Cell Res.* 1763: 1463–1477.
- Michieletto, D., Marenduzzo, D., and Orlandini, E. (2014). Is the Kinetoplast DNA a Percolating Network of Linked Rings at its Critical Point ? *Nucleic Acids Res.* XX: 1–6.

- Militello, K.T., and Read, L.K. (2000). UTP-dependent and -independent Pathways of mRNA turnover in *Trypanosoma brucei* mitochondria. *Mol. Cell. Biol.* 20: 2308–2316.
- Mingler, M.K., Hingst, A.M., Clement, S.L., Yu, L.E., Reifur, L., and Koslowsky, D.J. (2006). Identification of pentatricopeptide repeat proteins in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* 150: 37–45.
- Mitashi, P., Hasker, E., Lejon, V., Kande, V., Muyembe, J.J., Lutumba, P., et al. (2012). Human African Trypanosomiasis Diagnosis in First-Line Health Services of Endemic Countries, a Systematic Review. *PLoS Negl. Trop. Dis.* 6:.
- Mony, B.M., MacGregor, P., Ivens, A., Rojas, F., Cowton, A., Young, J., et al. (2013). Genome-wide dissection of the quorum sensing signalling pathway in *Trypanosoma brucei*. *Nature* 505: 681–685.
- Morales, J., Hashimoto, M., Williams, T.A., Hirawake-mogi, H., Makiuchi, T., Tsubouchi, A., et al. (2016). Differential remodelling of peroxisome function underpins the environmental and metabolic adaptability of diplomonads and kinetoplastids.
- Moreira, S., Valach, M., Aoulad-Aissa, M., Otto, C., and Burger, G. (2016). Novel modes of RNA editing in mitochondria. *Nucleic Acids Res.* 44: 1–13.
- Morrison, L.J. (2011). Parasite-driven pathogenesis in *Trypanosoma brucei* infections. *Parasite Immunol.* 33: 448–455.
- Morrison, L.J., Marcello, L., and McCulloch, R. (2009). Antigenic variation in the African trypanosome: Molecular mechanisms and phenotypic complexity. *Cell. Microbiol.* 11: 1724–1734.
- Morrison, L.J., Tait, A., McCormack, G., Sweeney, L., Black, A., Truc, P., et al. (2008). *Trypanosoma brucei* gambiense Type 1 populations from human patients are clonal and display geographical genetic differentiation. *Infect. Genet. Evol.* 8: 847–854.
- Nass, S., and Nass, M.M.K. (1963). INTRAMITOCHONDRIAL FIBERS WITH DNA CHARACTERISTICS : II. Enzymatic and Other Hydrolytic Treatments. *J. Cell Biol.* 19: 613–629.
- Njiru, Z.K., Constantine, C.C., Masiga, D.K., Reid, S. a, Thompson, R.C. a, and Gibson, W.C. (2006). Characterization of *Trypanosoma evansi* type B. *Infect. Genet. Evol.* 6: 292–300.
- Ntambi, J.M., and Englund, P.T. (1985). A gap at a unique location in newly replicated kinetoplast DNA minicircles from *Trypanosoma equiperdum*. *J. Biol. Chem.* 260: 5574–9.
- Ochsenreiter, T., Anderson, S., Wood, Z. a, and Hajduk, S.L. (2008a). Alternative RNA editing produces a novel protein involved in mitochondrial DNA maintenance in trypanosomes. *Mol. Cell. Biol.* 28: 5595–604.



- Ochsenreiter, T., Cipriano, M., and Hajduk, S.L. (2007a). BIOINFORMATICS KISS : The kinetoplastid RNA editing sequence search tool. 1–4.
- Ochsenreiter, T., Cipriano, M., and Hajduk, S.L. (2007b). KISS: the kinetoplastid RNA editing sequence search tool. *RNA* 13: 1–4.
- Ochsenreiter, T., Cipriano, M., and Hajduk, S.L. (2008b). Alternative mRNA editing in trypanosomes is extensive and may contribute to mitochondrial protein diversity. *PLoS One* 3: e1566.
- Ochsenreiter, T., and Hajduk, S.L. (2006). Alternative editing of cytochrome c oxidase III mRNA in trypanosome mitochondria generates protein diversity. *EMBO Rep.* 7: 1128–33.
- Onn, I., Kapeller, I., Abu-Elneel, K., and Shlomai, J. (2006). Binding of the universal minicircle sequence binding protein at the kinetoplast DNA replication origin. *J. Biol. Chem.* 281: 37468–37476.
- Panigrahi, A.K., Zíková, A., Dalley, R. a, Acestor, N., Ogata, Y., Anupama, A., et al. (2008). Mitochondrial complexes in *Trypanosoma brucei*: a novel complex and a unique oxidoreductase complex. *Mol. Cell. Proteomics* 7: 534–545.
- Paris, Z., Hashimi, H., Lun, S., Alfonzo, J.D., and Lukeš, J. (2011). Futile import of tRNAs and proteins into the mitochondrion of *Trypanosoma brucei evansi*. *Mol. Biochem. Parasitol.* 176: 116–120.
- Park, S.H., Nguyen, T.N., and Günzl, A. (2012). Development of an efficient in vitro transcription system for bloodstream form *Trypanosoma brucei* reveals life cycle-independent functionality of class i transcription factor A. *Mol. Biochem. Parasitol.* 181: 29–36.
- Pays, E., Vanhollebeke, B., Uzureau, P., Lecordier, L., and Pérez-Morga, D. (2014). The molecular arms race between African trypanosomes and humans. *Nat. Rev. Microbiol.* 12: 575–584.
- Pérez-Morga, D., Amiguet-Vercher, A., Vermijlen, D., and Pays, E. (2001). Organization of telomeres during the cell and life cycles of *trypanosoma brucei*. *J. Eukaryot. Microbiol.* 48: 221–226.
- Pérez-Morga, D., and Englund, P.T. (1993). The structure of replicating kinetoplast DNA networks. *J. Cell Biol.* 123: 1069–79.
- perez-morga, D.L., and Englund, P.T. (1993). The attachment of minicircles to kinetoplast DNA networks during replication. *Cell* 74: 703–711.
- Pesole, G., Allen, J.F., Lane, N., Martin, W., Rand, D.M., Schatz, G., et al. (2012). The neglected genome. *EMBO Rep.* 13: 473–474.
- Pollard, V.W., and Hajduk, S.L. (1991). *Trypanosoma equiperdum* minicircles encode

- three distinct primary transcripts which exhibit guide RNA characteristics. *Mol. Cell. Biol.* 11: 1668–75.
- Pollard, V.W., Rohrer, S.P., Michelotti, E.F., Hancock, K., and Hajduk, S.L. (1990). Organization of minicircle genes for guide RNAs in *Trypanosoma brucei*. *Cell* 63: 783–90.
- Preußner, C., Jaé, N., and Bindereif, A. (2012). mRNA splicing in trypanosomes. *Int. J. Med. Microbiol.* 302: 221–224.
- Priest, J.W., and Hajduk, S.L. (1994). Developmental regulation of mitochondrial biogenesis in *Trypanosoma brucei*. *J. Bioenerg. Biomembr.* 26: 179–91.
- Pusnik, M., Small, I., Read, L.K., Fabbro, T., and Schneider, A. (2007). Pentatricopeptide repeat proteins in *Trypanosoma brucei* function in mitochondrial ribosomes. *Mol. Cell. Biol.* 27: 6876–88.
- Ray, D.S. (1989). Conserved sequence blocks in kinetoplast minicircles from diverse species of trypanosomes. *Mol. Cell. Biol.* 9: 1365–7.
- Read, L.K., Lukeš, J., and Hashimi, H. (2015). Trypanosome RNA editing: the complexity of getting U in and taking U out. *Wiley Interdiscip. Rev. RNA* 7: n/a-n/a.
- Read, L.K., Zimmer, S.L., and Ammerman, M.L. (2011). Marked for Translation: Long A/U Tails as an Interface between Completion of RNA Editing and Ribosome Recruitment. *Mol. Cell* 42: 6–8.
- Rice, P. (2000). The European Molecular Biology Open Software Suite EMBOSS : The European Molecular Biology Open Software Suite. 16: 2–3.
- Riley, G.R., Corell, R. a, and Stuart, K. (1994). Multiple guide RNAs for identical editing of *Trypanosoma brucei* apocytochrome b mRNA have an unusual minicircle location and are developmentally regulated. *J. Biol. Chem.* 269: 6101–8.
- Riou, G., and Delain, E. (1968). Electron microscopy of the circular kinetoplastic DNA from *Trypanosoma cruzi*: occurrence of catenated forms. *Proc. Natl. Acad. Sci. U. S. A.* 62: 210–7.
- Rohloff, P., Montalvetti, A., and Docampo, R. (2004). Acidocalcisomes and the contractile vacuole complex are involved in osmoregulation in *Trypanosoma cruzi*. *J. Biol. Chem.* 279: 52270–52281.
- Rohrer, S.P., Michelotti, E.F., Torri, a F., and Hajduk, S.L. (1987). Transcription of kinetoplast DNA minicircles. *Cell* 49: 625–32.
- Rosenthal, J.J.C. (2015). The emerging role of RNA editing in plasticity. *J. Exp. Biol.* 218: 1812–1821.
- Rotureau, B., and Abbeele, J. Van Den (2013). Through the dark continent: African trypanosome development in the tsetse fly. *Front. Cell. Infect. Microbiol.* 3: 53.

- Ryan, K. a, Shapiro, T. a, Rauch, C. a, and Englund, P.T. (1988). Replication of kinetoplast DNA in trypanosomes. *Annu. Rev. Microbiol.* 42: 339–58.
- Sanchez, D.O., Madrid, R., Engel, J.C., and Frasch, A.C.C. (1984). Rapid identification of *Trypanosoma cruzi* isolates by ‘dot-spot’ hybridization. *FEBS Lett.* 168: 139–142.
- Savill, N.J., and Higgs, P.G. (1999). A theoretical study of random segregation of minicircles in trypanosomatids. *Proc. Biol. Sci.* 266: 611–20.
- Savill, N.J., and Higgs, P.G. (2000). Redundant and non-functional guide RNA genes in *Trypanosoma brucei* are a consequence of multiple genes per minicircle. *Gene* 256: 245–52.
- Saxowsky, T.T., Choudhary, G., Klingbeil, M.M., and Englund, P.T. (2003). *Trypanosoma brucei* has two distinct mitochondrial DNA polymerase beta enzymes. *J. Biol. Chem.* 278: 49095–49101.
- Schnarwiler, F., Niemann, M., Doiron, N., Harsman, A., Käser, S., Mani, J., et al. (2014). Trypanosomal TAC40 constitutes a novel subclass of mitochondrial  $\beta$ -barrel proteins specialized in mitochondrial genome inheritance. *Proc. Natl. Acad. Sci. U. S. A.* 111: 7624–7629.
- Schnauffer, A. (2010a). Evolution of dyskinetoplastic trypanosomes: how, and how often? *Trends Parasitol.* 26: 557–558.
- Schnauffer, A. (2010b). Evolution of dyskinetoplastic trypanosomes: How, and how often? *Trends Parasitol.* 26: 557–558.
- Schnauffer, A., Clark-Walker, G.D., Steinberg, A.G., and Stuart, K. (2005). The F1-ATP synthase complex in bloodstream stage trypanosomes has an unusual and essential function. *EMBO J.* 24: 4029–40.
- Schnauffer, A., Domingo, G.J., and Stuart, K. (2002). Natural and induced dyskinetoplastic trypanosomatids: how to live without mitochondrial DNA. *Int. J. Parasitol.* 32: 1071–84.
- Schnauffer, A., Panigrahi, A.K., Panicucci, B., Igo, R.P., Wirtz, E., Salavati, R., et al. (2001). An RNA ligase essential for RNA editing and survival of the bloodstream form of *Trypanosoma brucei*. *Science* 291: 2159–2162.
- Schneider, a (2001). Unique aspects of mitochondrial biogenesis in trypanosomatids. *Int. J. Parasitol.* 31: 1403–15.
- Schroth, G.P., Siino, J.S., Cooney, C.A., Th’ng, J.P.H., Ho, P.S., and Bradbury, E.M. (1992). Intrinsically bent DNA flanks both sides of an RNA polymerase I transcription start site: Both regions display novel electrophoretic mobility. *J. Biol. Chem.* 267: 9958–9964.
- Schwede, A., and Carrington, M. (2010). Bloodstream form Trypanosome plasma

- membrane proteins: antigenic variation and invariant antigens. *Parasitology* 137: 2029–39.
- Seiwert, S.D., Heidmann, S., and Stuart, K. (1996). Direct visualization of uridylyate deletion in vitro suggests a mechanism for kinetoplastid RNA editing. *Cell* 84: 831–841.
- Shapiro, T.A., and Englund, P.T. (1995). THE STRUCTURE AND REPLICATION OF Kinetoplast DNA.
- Shapiro, T. a (1993). Kinetoplast DNA maxicircles: networks within networks. *Proc. Natl. Acad. Sci. U. S. A.* 90: 7809–13.
- Shapiro, T. a, Klein, V. a, and Englund, P.T. (1999). Isolation of kinetoplast DNA. *Methods Mol. Biol.* 94: 61–7.
- Shaw, J.M., Campbell, D., and Simpson, L. (1989). Internal Frameshifts within the Mitochondrial Genes for Cytochrome Oxidase Subunit II and Maxicircle Unidentified Reading Frame 3 of *Leishmania tarentolae* are Corrected by RNA Editing: Evidence for Translation of the Edited Cytochrome Oxidase Subunit II m. *Proc Natl Acad Sci U S A* 86: 6220–6224.
- Shaw, J.M., Feagin, J.E., Stuart, K., and Simpson, L. (1988). Editing of kinetoplastid mitochondrial mRNAs by uridine addition and deletion generates conserved amino acid sequences and AUG initiation codons. *Cell* 53: 401–11.
- Shlomai, J. (2004). The structure and replication of kinetoplast DNA. *Curr. Mol. Med.* 4: 623–47.
- Siegel, T.N., Gunasekera, K., Cross, G.A.M., and Ochsenreiter, T. (2011). Gene expression in *Trypanosoma brucei*: lessons from high-throughput RNA sequencing. *Trends Parasitol.* 27: 434–441.
- Simarro, P.P., Cecchi, G., Paone, M., Franco, J.R., Diarra, A., Ruiz, J.A., et al. (2010). The Atlas of human African trypanosomiasis: a contribution to global mapping of neglected tropical diseases. *Int. J. Health Geogr.* 9: 57.
- Simpson, L. (1968). Behavior of the kinetoplast of *Leishmania tarentolae* upon cell rupture. *J Protozool* 15: 132–136.
- Simpson, L. (1997). The genomic organization of guide RNA genes in kinetoplastid protozoa: Several conundrums and their solutions. *Mol. Biochem. Parasitol.* 86: 133–141.
- Simpson, L., Aphasizhev, R., Lukeš, J., and Cruz-Reyes, J. (2010). Guide to the Nomenclature of Kinetoplastid RNA Editing: A Proposal. *Protist* 161: 2–6.
- Simpson, L., Douglass, S.M., Lake, J. a., Pellegrini, M., and Li, F. (2015). Comparison of the Mitochondrial Genomes and Steady State Transcriptomes of Two Strains of the

- Trypanosomatid Parasite, *Leishmania tarentolae*. PLoS Negl. Trop. Dis. 9: e0003841.
- Simpson, L., and Emeson, R.B. (1996). RNA Editing. Annu. Rev. Neurosci. 19: 27–52.
- Simpson, L., Sbicego, S., and Aphasizhev, R. (2003). Uridine insertion / deletion RNA editing in trypanosome mitochondria : A complex business. 1: 265–276.
- Simpson, L., and Silva, A. da (1971). Isolation and characterization of kinetoplast DNA from *Leishmania tarentolae*. J. Mol. Biol. 56: 443–473.
- Simpson, L., Thiemann, O.H., Savill, N.J., Alfonzo, J.D., and Maslov, D. a (2000). Evolution of RNA editing in trypanosome mitochondria. Proc. Natl. Acad. Sci. U. S. A. 97: 6986–93.
- Simpson, R.M., Bruno, A.E., Bard, J.E., Buck, M.J., and Read, L.K. (2016). High-throughput sequencing of partially edited trypanosome mRNAs reveals barriers to editing progression and evidence for alternative editing. Rna rna.055160.115-.
- Sinha, K.M., Hines, J.C., Downey, N., and Ray, D.S. (2004). Mitochondrial DNA ligase in *Crithidia fasciculata*. Proc Natl Acad Sci U S A 101: 4361–4366.
- Sloof, P., Haan, A. de, Eier, W., Iersel, M. van, Boel, E., Steeg, H. van, et al. (1992). The nucleotide sequence of the variable region in *Trypanosoma brucei* completes the sequence analysis of the maxicircle component of mitochondrial kinetoplast DNA. Mol. Biochem. Parasitol. 56: 289–99.
- Smith, H.C., Gott, J.M., and Hanson, M.R. (1997). A guide to RNA editing. RNA 3: 1105–23.
- Speijer, D. (2006). Is kinetoplastid pan-editing the result of an evolutionary balancing act? IUBMB Life 58: 91–6.
- Speijer, D. (2010). Constructive neutral evolution cannot explain current kinetoplastid panediting patterns. Proc. Natl. Acad. Sci. U. S. A. 107: E25; author reply E26.
- Steinert, M., and Assel, S. Van (1975). Large circular mitochondrial DNA in *Crithidia luciliae*. Exp. Cell Res. 96: 406–409.
- Steinert, M., and Assel, S. Van (1980). Sequence heterogeneity in kinetoplast DNA: Reassociation kinetics. Plasmid 3: 7–17.
- Steinmann, P., Stone, C.M., Sutherland, C.S., Tanner, M., and Tediosi, F. (2015). Contemporary and emerging strategies for eliminating human African trypanosomiasis due to *Trypanosoma brucei gambiense*: review. Trop. Med. Int. Heal. 20: 707–718.
- Stoltzfus, A. (2012). Constructive neutral evolution: exploring evolutionary theory’s curious disconnect. Biol. Direct 7: 35.
- Stuart, K. (1979). Kinetoplast DNA of *Trypanosoma brucei*: Physical map of the maxicircle. Plasmid 2: 520–528.

- Stuart, K., and Gelvin, S.R. (1980). Kinetoplast DNA of normal and mutant *Trypanosoma brucei*. *Am. J. Trop. Med. Hyg.* 29: 1075–81.
- Stuart, K.D. (1971). Evidence for the retention of kinetoplast DNA in an acriflavine-induced dyskinetoplastic strain of *Trypanosoma brucei* which replicates the altered central element of the kinetoplast. *J. Cell Biol.* 49: 189–195.
- Stuart, K.D., Schnauffer, A., Ernst, N.L., and Panigrahi, A.K. (2005). Complex management: RNA editing in trypanosomes. *Trends Biochem. Sci.* 30: 97–105.
- Sturm, N.R., Maslov, D.A., Blum, B., and Simpson, L. (1992). Generation of unexpected editing patterns in *Leishmania tarentolae* mitochondrial mRNAs: misediting produced by misguiding. *Cell* 70: 469–76.
- Sturm, N.R., and Simpson, L. (1990). Kinetoplast DNA minicircles encode guide RNAs for editing of cytochrome oxidase subunit III mRNA. *Cell* 61: 879–84.
- Sudarshi, D., Lawrence, S., Pickrell, W.O., Eligar, V., Walters, R., Quaderi, S., et al. (2014). Human African Trypanosomiasis Presenting at Least 29 Years after Infection???What Can This Teach Us about the Pathogenesis and Control of This Neglected Tropical Disease? *PLoS Negl. Trop. Dis.* 8: 8–13.
- Suematsu, T., Zhang, L., Aphasizheva, I., Monti, S., Huang, L., Wang, Q., et al. (2016a). Antisense Transcripts Delimit Exonucleolytic Activity of the Mitochondrial 3' Processome to Generate Guide RNAs. *Mol. Cell* 1–15.
- Suematsu, T., Zhang, L., Aphasizheva, I., Monti, S., Huang, L., Wang, Q., et al. (2016b). Antisense Transcripts Delimit Exonucleolytic Activity of the Mitochondrial 3' Processome to Generate Guide RNAs. *Mol. Cell* 1–15.
- Sugisaki, H., and Ray, D.S. (1987). DNA sequence of *Crithidia fasciculata* kinetoplast minicircles. *Mol. Biochem. Parasitol.* 23: 253–263.
- Taylor, J.E., and Rudenko, G. (2006). Switching trypanosome coats: what's in the wardrobe? *Trends Genet.* 22: 614–620.
- Thiemann, O.H., Maslov, D. a, and Simpson, L. (1994). Disruption of RNA editing in *Leishmania tarentolae* by the loss of minicircle-encoded guide RNA genes. *EMBO J.* 13: 5689–700.
- Thomas, S., Martinez, L.L.I.T., Westenberger, S.J., and Sturm, N.R. (2007). A population study of the minicircles in *Trypanosoma cruzi*: predicting guide RNAs in the absence of empirical RNA editing. *BMC Genomics* 8: 133.
- Tielens, a G., and Hellemond, J.J. Van (1998). Differences in energy metabolism between trypanosomatidae. *Parasitol. Today* 14: 265–72.
- Timms, M.W., Deursen, F.J. Van, Hendriks, E.F., and Matthews, K.R. (2002). Mitochondrial Development during Life Cycle Differentiation of African

- Trypanosomes : Evidence for a Kinetoplast-dependent Differentiation Control Point. *13*: 3747–3759.
- Truc, P., Büscher, P., Cuny, G., Gonzatti, M.I., Jannin, J., Joshi, P., et al. (2013). Atypical human infections by animal trypanosomes. *PLoS Negl. Trop. Dis.* 7: e2256.
- Turner, C.M. (1999). Antigenic variation in *Trypanosoma brucei* infections: an holistic view. *J. Cell Sci.* 112 ( Pt 1): 3187–3192.
- Turner, C.M., Aslam, N., and Dye, C. (1995). Replication, differentiation, growth and the virulence of *Trypanosoma brucei* infections. *Parasitology* 111: 289–300.
- Uzureau, P., Uzureau, S., Lecordier, L., Fontaine, F., Tebabi, P., Homblé, F., et al. (2013). Mechanism of *Trypanosoma brucei* gambiense resistance to human serum. *Nature*.
- Vassella, E., Reuner, B., Yutzy, B., and Boshart, M. (1997). Differentiation of African trypanosomes is controlled by a density sensing mechanism which signals cell cycle arrest via the cAMP pathway. *J. Cell Sci.* 110: 2661–2671.
- Vickerman, K. (1965). Polymorphism and Mitochondrial Activity in Sleeping Sickness Trypanosomes. *Nature* 208: 762–766.
- Vickerman, K. (1985). Developmental cycles and biology of pathogenic trypanosomes. *Br. Med. Bull.* 41: 105–114.
- Vickerman, K., Tetley, L., Hendry, K. a, and Turner, C.M. (1988). Biology of African trypanosomes in the tsetse fly. *Biol. Cell* 64: 109–19.
- Visser, N., and Opperdoes, F.R. (1980). Glycolysis in *Trypanosoma brucei*. *Eur. J. Biochem.* 103: 623–32.
- VREYSEN M., M.J., Saleh, K.M., Ali, M.Y., Abdulla, a M., Zhu, Z.R., Juma, K.G., et al. (2000). *Glossina austeni* (Diptera: Glossinidae) eradicated on the island of Unguja, Zanzibar, using the sterile insect technique. *J. Econ. Entomol.* 93: 123–135.
- Wang, X., and Seed, B. (2003). Selection of oligonucleotide probes for protein coding sequences. *Bioinformatics* 19: 796–802.
- Wang, Z., and Englund, P.T. (2001). RNA interference of a trypanosome topoisomerase II causes progressive loss of mitochondrial DNA. *EMBO J.* 20: 4674–4683.
- Weelden, S.W.H. van, Fast, B., Vogt, A., Meer, P. van der, Saas, J., Hellemond, J.J. van, et al. (2003). Procyclic *Trypanosoma brucei* do not use Krebs cycle activity for energy generation. *J. Biol. Chem.* 278: 12854–12863.
- Weiner, A.M., and Maizels, N. (1990). RNA editing: Guided but not templated? *Cell* 61: 917–920.
- Weng, J., Aphasizheva, I., Etheridge, R.D., Huang, L., Wang, X., Falick, A.M., et al. (2008). Guide RNA-Binding Complex from Mitochondria of Trypanosomatids. *Mol.*

Cell 32: 198–209.

Westesson, O., Skinner, M., and Holmes, I. (2013). Visualizing next-generation sequencing data with JBrowse. *Brief. Bioinform.* 14: 172–7.

WILSON, S.G., MORRIS, K.R., LEWIS, I.J., and KROG, E. (1963). The effects of trypanosomiasis on rural economy with special reference to the Sudan, Bechuanaland and West Africa. *Bull. World Health Organ.* 28: 595–613.

Woodward, R., and Gull, K. (1990). Timing of nuclear and kinetoplast DNA replication and early morphological events in the cell cycle of *Trypanosoma brucei*. *J. Cell Sci.* 95 ( Pt 1): 49–57.

Yaro, M., Munyard, K.A., Stear, M.J., and Groth, D.M. (2016). Combatting African Animal Trypanosomiasis (AAT) in livestock: The potential role of trypanotolerance. *Vet. Parasitol.* 225: 43–52.

Yurchenko, V.Y., Merzlyak, E.M., Kolesnikov, a a, Martinkina, L.P., and Vengerov, Y.Y. (1999). Structure of *Leishmania* minicircle kinetoplast DNA classes. *J. Clin. Microbiol.* 37: 1656–7.

Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18: 821–9.

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* 30: 614–620.

Zhao, Z., Lindsay, M.E., Roy Chowdhury, A., Robinson, D.R., and Englund, P.T. (2008). p166, a link between the trypanosome mitochondrial DNA and flagellum, mediates genome segregation. *EMBO J.* 27: 143–154.