Hyperevolution of trypanosome *Variant Surface Glycoprotein* genes

Lindsey Jane Plenderleith

BA MSci MRes

Submitted in fulfilment of the requirements for the Degree of Doctor of Philosophy

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

August 2013

Abstract

The African sleeping sickness parasite *Trypanosoma brucei* evades the immune system of its mammalian host by periodically switching the variant surface glycoprotein (VSG) that forms its cell-surface coat. This process of antigenic variation utilises a large archive of *VSG* genes, which are primarily subtelomeric and appear to evolve rapidly. Subtelomeres are the location of multi-member, variable gene families in many organisms, and often seem to have an elevated rate of mutation. The *VSG* archive is a particularly striking example of an organism taking advantage of this environment to promote hyperevolution. The aim of this project was to investigate the changes that occur in *VSG* evolution. In collaboration with researchers at the Sanger Institute, genomes from two time-separated isolates of the same trypanosome strain were sequenced and assembled. The quality of the genome assemblies was assessed, and the genomes concluded to be of sufficient quality for further analysis. Chromosome core genes and *VSG* N-terminal domain (NTD) genes and pseudogenes were annotated in each genome, and mutations between the genomes in each gene were catalogued. *VSG* NTDs had a significantly higher mutation frequency than core genes, and the specific patterns of mutations differed significantly between the two genome regions. These results together implied that *VSG* are subject to different mutational processes from core genes. However, mutation frequency did not appear to differ between *VSG* NTDs and other subtelomeric sequence, indicating that it is the subtelomeres in general that are subject to elevated mutational activity. Further examination of the *VSG* NTDs within each new genome reinforced published observations in the reference genome strain *VSG* archive of extensive substructuring and abundance of pseudogenes. Finally, to address the question of which mechanisms may be involved in elevating the mutation rate in subtelomeres, an attempt was made to characterise two members of a gene family predicted to encode error-prone lesion bypass DNA polymerases, a class of enzymes that have been suggested to have a role in the systematic generation of mutations. Such results as were obtained suggested that the genes examined may not encode active polymerases, and the results did not provide any evidence for a role for these polymerases in *VSG* hyperevolution. Overall, however, the project has uncovered considerable detail of how hypermutation proceeds in this highly variable gene family.

Table of contents

List of tables

List of figures

List of accompanying material

1 CD containing Electronic Appendices (details in Appendix 2)

Acknowledgements

I would like to thank:

Dave, for inspiration, support and guidance. May your retirement be long and happy, especially now I've stopped badgering you about my thesis.

The Barry and McCulloch groups and other members of the WTCMP: it has been a pleasure to work with you. Particular thanks to Jamie for many ideas and fruitful chats, Barbara for pol kappa materials and guidance, Nick and Jon for bioinformatics tools and much advice, Alex and Ingrid for help with all the administrative things, the morning tea drinkers for the fascinating and far-fetched discussions, and everyone who has generously given me cake.

Dan for help with statistics and Erida for thoughts on modelling mutations; Thomas and Matt for welcoming me at Sanger and for your contribution to the project and my training; and everyone else who kindly gave me materials or tools.

Darren, Olwyn and Bill for running such an excellent PhD programme, the Wellcome Trust for funding it, and the past and present students of the programme for stimulating forums and retreats.

My family, for making an interest in biology inevitable, and for all your encouragement all along the way.

Stephen, for unwavering support and faith in me, and bonus maths consultancy skills. Thank you.

Author's declaration

The EATRO 3 and EATRO 2340 genome sequencing and initial assembly were carried out in collaboration with Dr Thomas Otto, Dr Matt Berriman (both Sanger Institute) and Prof Ed Louis (University of Nottingham), as detailed in section 3.2.1.

Two other, minor contributions from other researchers are detailed in the text (sections 3.4.4 and 5.3.2).

I declare that this thesis and the results presented within are entirely my own work except where otherwise stated. No part of this thesis has been previously submitted for a degree at any university.

Lindsey Plenderleith

List of abbreviations

ABACAS	Algorithm-Based Automatic Contiguation of Assembled Sequence
ACT	Artemis Comparison Tool
BAMview	Binary Alignment/Map viewer
BES	bloodstream expression site
BLAST	Basic Local Alignment Search Tool
bp	base pair
BSF	bloodstream form *Trypanosoma brucei*
BWA	Burrows-Wheeler Aligner
CI	confidence interval
contig	contiguous DNA sequence
CTD	VSG C-terminal domain
CV	column volumes
dH_2O	sterile deionised water
d_N	number of non-synonymous mutations per non-synonymous site
d_{N3}	number of mutations per site for which every possible mutation was non-synonymous
dNTP	deoxyribonucleotide triphosphate
d_S	number of synonymous mutations per synonymous site
d_{S3}	number of mutations per site for which every possible mutation was synonymous
dsDNA	double-stranded DNA
DTT	dithiothreitol
EATRO	East African Trypanosomiasis Research Organisation
EDTA	ethylenediaminetetraacetic acid
eGFP	enhanced green fluorescent protein
ES	expression site
ESAG	expression site-associated gene
ESB	expression site body
FISH	fluorescence *in-situ* hybridisation
GPI	glycosylphosphatidylinositol
HAT	human African trypanosomiasis

HMM	hidden Markov model
HSP	high-scoring segment pairs
IF	immunofluorescence
IMAGE	Iterative Mapping and Assembly for Gap Elimination
indel	insertion-deletion mutation
IPTG	isopropyl ß-D-1-thiogalactopyranoside
kb	thousand base pairs
kDNA	kinetoplast DNA
LOH	loss of heterozygosity
Mb	million base pairs
MBP	maltose binding protein
MES	metacyclic expression site
MRCA	most recent common ancestor
MUM	maximal unique match
NEB	New England Biolabs
NTD	VSG N-terminal domain
ORF	open reading frame
PAGE	polyacrylamide gel electrophoresis
(cf)PBS	(ice-cold, filter-sterilised) phosphate-buffered saline
PCR	polymerase chain reaction
PNA	peptide nucleic acid
Pol I	RNA polymerase I
Pol II	RNA polymerase II
pol κ	DNA polymerase κ
polκ001	*Tb11.12.0001* or its product
polκ10	*Tb11.01.0010* or its product
RATT	Rapid Annotation Transfer Tool
RHS	retrotransposon hot spot
SAMTools	Sequence Alignment/Map Tools
SDS	sodium dodecyl sulphate
SGA	String Graph Assembler
SNP	single nucleotide polymorphism
SRA	serum resistance-associated gene/protein
SSC	saline-sodium citrate buffer
SVM-VSG	Support Vector Machine for VSG

TAE	Tris-acetate-EDTA
TLF	trypanosome lytic factor(s)
VSG	variant surface glycoprotein
X-Gal	5-bromo-4-chloro-indolyl-ß-D-galactopyranoside

Chapter 1: Introduction

1 Introduction

1.1 Overview

1.1.1 Hyperevolution

> *Look abroad thro' Nature's range,*
> *Nature's mighty law is change.*
>
> Robert Burns

Change is a constant in biology, when considering DNA sequences. It appears to be either impossible or undesirable for organisms to maintain their genomes without some change, because an organism with a zero mutation rate has never been described. However, although everything changes, not everything changes in the same way. Rates of mutation vary between organisms (Lynch, 2010b), as might be expected from variation in DNA replication machinery, genome size, environment and selection pressures. Interestingly, mutation rates also vary between different genome sites within an organism (Hodgkinson & Eyre-Walker, 2011; Martincorena & Luscombe, 2013). Thus, hyperevolution, a significantly faster rate of sequence change than average, can occur on the scale of species, for example the rapid mutation rates of some bacteria compared with that of mammals; but can also be occurring within an organism, for example ectopic recombination in subtelomeres increasing the rate of mutation near chromosome ends.

1.1.2 Hyperevolution in trypanosome antigenic variation genes

The African sleeping sickness parasite *Trypanosoma brucei* possesses a large, highly variable family of *variant surface glycoprotein* (*VSG*) genes that it uses for antigenic variation. The gene family is located in the subtelomeres, which are genome regions that often have an elevated rate of mutation compared with chromosome cores and are often the location of members of variable multi-gene families (Mefford & Trask, 2002; Brown *et al*, 2010). The trypanosome *VSG* gene family is an excellent example of hyperevolution in subtelomeres, and is exciting to study because of the importance of the genes to parasite survival, and because the hyperevolution of *VSG*s is intertwined with the process of antigenic variation. In this Introduction, I will describe the importance of variation generated in subtelomeres, and why hyperevolution of *VSG*s is both intrinsically interesting to parasitologists but also of wider interest in the context of subtelomere evolution.

1.2 Subtelomere hyperevolution

1.2.1 Organisation of linear chromosome ends

Eukaryotes, and some prokaryotes (Bentley *et al*, 2002b; Kobryn & Chaconas, 2001), maintain their genetic material in the form of linear chromosomes. The maintenance of such chromosomes poses a problem, due to the existence of free ends of DNA at their extremities. As well as being exposed to nuclease activity, free ends of DNA usually trigger a DNA damage response (van Gent *et al*, 2001), because they are often the result of DNA damage causing a double-stranded break. The DNA damage response causes the cell to attempt repair and disrupts the cell cycle. The response can also involve recombination of DNA ends, which for chromosome ends could result in chromosome fusion (McEachern *et al*, 2000) or other inappropriate recombination. The linear structure of chromosomes also poses a problem for replication, because the standard DNA machinery requires a 3′ RNA primer, and hence cannot replicate the 3′ ends of linear DNA (de Lange *et al*, 1990). Telomeres, the ends of linear chromosomes, are therefore sequestered into specialised complexes that prevent the free DNA ends from interacting with other cellular components (Blackburn, 2001). Eukaryotic telomeres usually consist of tandem repeats of a short, G-rich sequence, which are usually replicated by the specialised enzyme telomerase (Blackburn, 2005). Telomeres are frequently organised in clusters at the nuclear periphery, a location that facilitates telomeric silencing and double-stranded break repair, and is also important in meiosis for accurate chromosome pairing and recombination (Gotta *et al*, 1996; Nimmo *et al*, 1998; Therizols *et al*, 2006).

Most of the chromosome is core sequence: in this region, homologous chromosome pairs share gene content, gene order and intergenic sequence, in an arrangement maintained by homologous recombination. However, frequently there is a transitionary region between the chromosome core and the telomere repeats, known as the subtelomere (Figure 1-1), which is more complex than the simple telomere repeats but is not particularly conserved between homologous chromosomes (Louis & Vershinin, 2005). The content of subtelomeric sequences varies widely between organisms, but in virtually all eukaryotes studied they are composed of a combination of various repeats (Pryde *et al*, 1997). These repeats are usually found in the subtelomeres of multiple chromosomes, and different blocks of repeats have different distributions with little regard to chromosome core homology (Mefford & Trask, 2002). The haploid nature of subtelomeres presents a challenge for sequencing efforts, compounded by the regions' repetitive nature and variability between strains within a species (Louis, 1995).

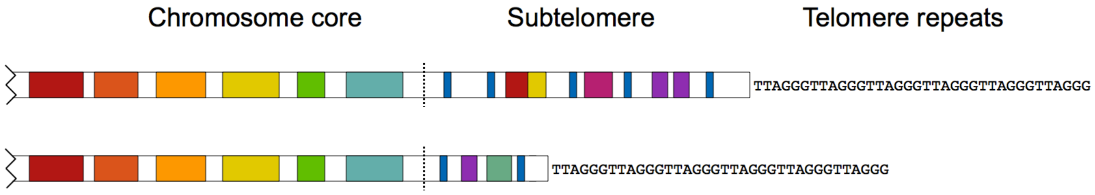

Chromosome core
Subtelomere
Telomere repeats
TTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
TTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG

Figure 1-1 – Schematic diagram of the ends of pair of homologous chromosomes showing chromosome cores, subtelomeres and telomere repeats.
Blocks of the same colour indicate homologous sequences, *i.e.* the cores of homologous chromosomes share syntenic, homologous genes; the subtelomeres share some homologous sequences but also contain sequences and parts of genes from the cores, from other parts of the genome, and some unique sequences.

One of the first organisms in which subtelomeres were characterised was *Saccharomyces cerevisiae*. Yeast subtelomeres are combinations of highly conserved repeat elements (X and Y′) (Chan & Tye, 1983); repeat elements found in variable numbers of chromosome ends, including open reading frames (ORFs) and probable former transposable elements (Louis, 1995); and, most chromosome-internally, sequence unique to each subtelomere (Louis & Haber, 1992). Human subtelomeres have also been well characterised and consist of several elements: subtelomeric repeats, which appear in numerous subtelomeric regions; segmental duplications of regions from within the chromosome core; and single-copy sequences unique to each subtelomere (Riethman *et al*, 2005). Parts of the subtelomeres are transcribed; transcribed regions are distributed throughout the subtelomeres and include repeat regions and segmental duplications as well as the subtelomere-unique sequences (Flint *et al*, 1997; Riethman *et al*, 2004).

Subtelomeres have been characterised in a number of other organisms, including *Plasmodium falciparum* (Gardner *et al*, 1998; Figueiredo *et al*, 2000; Scherf *et al*, 2001); *T. brucei* (discussed below), *Trypanosoma cruzi* (Moraes Barros *et al*, 2012), *Drosophila melanogaster* (Mason & Biessmann, 1995) and *Arabidopsis thaliana* (Kotani *et al*, 1999; Arabidopsis Genome Initiative, 2000). These studies reveal that repeated elements and segmental duplications are a common theme in subtelomere structure, but that the extent and composition of subtelomeres varies widely. Subtelomeres are highly polymorphic between chromosomes and between individuals, and between organisms there is very little conservation of their sequences or lengths (Mefford & Trask, 2002; Pryde *et al*, 1997).

1.2.2 Subtelomeres as sites of highly variable gene families

Most subtelomeric sequence is non-coding, but in many species the subtelomeres contain genes, encoding proteins with a wide variety of functions, which will be described in this section. Subtelomeric genes often belong to large, highly diverse gene families. For example, the gene families used by some eukaryotic parasites in antigenic variation (discussed further below), where diversity of the genes is critical to parasite survival in a mammalian host, are often subtelomeric (Barry *et al*, 2003). In the human malaria parasite *P. falciparum*, three subtelomeric families are involved in antigenic variation: the *var* genes, with approximately 60 members, most of which are subtelomeric; and the *rif* and *stevor* gene families, which are closely related to one another, belong to the same superfamily and likely have similar structure, and contain about 200 and 40 members respectively (Gardner *et al*, 1998; Scherf *et al*, 2008; Hernandez-Rivas *et al*, 1997; Duffy & Tham, 2007; Kyes *et al*, 1999; Cheng *et al*, 1998). *Plasmodium vivax*, another human malaria parasite, possesses a different subtelomeric gene set, the 600 to 1000-member *vir* superfamily, which may be involved in antigenic variation (Fernandez-Becerra *et al*, 2009). *T. brucei* also undergoes antigenic variation, for which it utilises an archive of *VSG* genes that are primarily subtelomeric and number between 1000 and 2000 in the reference genome strain TREU 927 (Berriman *et al*, 2005; Marcello & Barry, 2007b). Another subtelomeric antigenic variation system is used by *Pneumocystis carinii*, which uses up to 100 major surface glycoproteins in small subtelomeric clusters (Stringer & Keely, 2001). Other examples of variable subtelomeric gene families with relevance to the organism’s ecological niche include disaccharide utilisation genes in yeast (Brown *et al*, 2010) and various surface protein genes in *Trypanosoma cruzi* (Moraes Barros *et al*, 2012).

As an entirely different example, vertebrate olfactory receptor genes allow the detection and differentiation of thousands of millions of chemicals, and comprise the largest gene superfamily in the vertebrate genome (Glusman *et al*, 2001; Olender *et al*, 2008). Although most of the 900 olfactory receptor genes and pseudogenes in humans are located in chromosome cores, at least eight distinct genes are subtelomeric, and four of these are present in multiple copies that vary in copy number between individuals (Trask *et al*, 1998). At least one subtelomeric olfactory receptor gene is transcribed, so it is likely that these genes are not simply accidentally copied pseudogenes (Linardopoulou *et al*, 2001).

Although most characterised bacterial genomes are circular, there are a few species that possess linear chromosomes, and hence telomere equivalents. One example is the soil bacterium *Streptomyces coelicor*, which has a very large single linear chromosome. The core of this chromosome contains essential genes and is syntenic with the circular chromosomes of *Mycobacterium tuberculosis* and *Corynebacterium diptheriae*, with which it appears to share a common ancestor. However, nearly a million base pairs (Mb) of sequence at either end is *Streptomyces*-specific, and can tolerate extensive deletions and amplifications without compromising viability in laboratory conditions (Volff & Altenbuchner, 1998; Bentley *et al*, 2002b). These regions appear to be analogous to eukaryotic subtelomeres and contain ‘contingency’ genes, which are highly mutable loci encoding proteins that interact with the unpredictable aspects of the environment (Moxon *et al*, 1994). Such genes help the bacterium survive in the varied soil environment, for example encoding proteins to produce secondary metabolites (Bentley *et al*, 2002a). It is hypothesised that the linearisation of the *Streptomyces* genome occurred as a specific adaptation to accommodate the extra genes and promote their diversity (Barry *et al*, 2003).

1.2.3 Polymorphism and hyperevolution of subtelomeres

Because there are numerous variable subtelomeric gene families, often of crucial importance to their host organisms, subtelomeres are coming to be recognised as having a key role in variation in organisms (Riethman *et al*, 2005; Brown *et al*, 2010). The importance of a subtelomeric location for variable multi-gene families appears to lie with the rapid rate of change associated with these genomic regions.

Detailed examination of the human reference genome has revealed that human subtelomeres are complex, dynamic mosaics formed and shaped by the interaction of several processes. Segmental duplications arise throughout the subtelomeres, often as a result of aberrant non-homologous end-joining, and appear just as readily in regions containing genes as in non-coding sequence. Duplicated segments may then be modified by interchromosomal recombination, which means that blocks of sequence on separate subtelomeres do not evolve completely independently (Linardopoulou *et al*, 2005). Sister chromatid exchange occurs at rates several orders of magnitude higher in subtelomeres than in chromosome cores, indicating that rates of DNA breakage and repair are also elevated in subtelomeres (Rudd *et al*, 2007). The evolution of these patchworks of sequence is rapid: half of known human subtelomere sequence has arisen in human-specific events, and almost every subtelomeric block examined in multiple individuals shows some variation in copy number and location (Brown *et al*, 1990; Trask *et al*, 1998; Baird *et al*, 2000; Mefford & Trask, 2002; Linardopoulou *et al*, 2005). Gene copy number variation is therefore regarded as a key characteristic of human subtelomeres, being detectable at almost every subtelomere (Redon *et al*, 2006; Riethman, 2008a).

In general, what is known about subtelomeres in other organisms bears out a picture of a highly polymorphic and dynamic genome region. In *P. falciparum*, the size of chromosomes can vary by up to 300 thousand base pairs (kb) between strains, due to differences in subtelomeres caused by homologous recombination; and single nucleotide polymorphisms (SNPs) are also over-represented in these regions (Corcoran *et al*, 1988; de Bruin *et al*, 1994; Volkman *et al*, 2002). Further, the subtelomeric *var* genes have been shown to undergo frequent ectopic recombination (Freitas-Junior *et al*, 2000). Size polymorphisms of similar scale are seen in kinetoplastid parasites (Fu & Melville, 2002). Analysis of the *VSG* genes occupying the subtelomeres in the genome strain of *T. brucei* allowed inference of the occurrence of multiple types of mutation, including segmental duplication and recombination but also point mutations and short insertion-deletion mutations (indels) (Bernards *et al*, 1986; Marcello & Barry, 2007b).

Processes that duplicate genes are considered key in diversification and generation of novel function, because selection for the original function is somewhat relaxed for the new gene copy (Lynch & Conery, 2000; Moran *et al*, 2011). Relaxation of selection may allow random events to take a leading part in shaping evolution of the duplicate gene. Subtelomere hyperevolution may be an important process acting to provide such duplications (Riethman, 2008b). In yeast, comparative analysis of eight species revealed that subtelomeric gene families varied more in copy number and contained more members than families with no subtelomeric members, and the same study provided evidence for rapid functional divergence within selected subtelomeric families (Brown *et al*, 2010). The key role of subtelomeres in generating gene diversity is also illustrated by the evolution of the olfactory receptor genes. Although most of these genes are not subtelomeric, analysis of the evolutionary relationships between them suggested that a major event in the expansion of the olfactory receptor gene family was the duplication of a core chromosomal cluster into the subtelomere of chromosome 1, from where it was duplicated into numerous new locations (Glusman *et al*, 2001).

Even without considering the effects of selection on mutations and duplications, there are several possible factors that can be envisaged to contribute, to a greater or lesser degree, to the hyperevolution of subtelomeres compared with chromosome cores: a greater tendency to undergo the mutations observed, for example sequence that makes double-stranded breaks more likely, or the presence of retrotransposons; greater tolerance of mutations in the cell, for example by altered repair and damage-sensing pathways; and novel mechanisms actively generating mutations. Roles have been suggested for several of these factors in bringing about differences in mutation profile between specific regions (Sharp *et al*, 2005; Cooper *et al*, 2007; Marcello & Barry, 2007b), but little is understood about how subtelomeres in general are particularly prone to mutation.

All these data suggest that genes located in subtelomeres are subject to a higher mutation rate than those in chromosome cores, that this increased rate is important in organism and gene diversity, and that rather than subtelomeric genes being specifically targeted, hyperevolution is an inherent property of subtelomeres. It appears that subtelomeric hyperevolution occurs by a combination of multiple gene conversions and segmental duplications with smaller-scale mutations.

1.3 *Trypanosoma brucei* and its impact

1.3.1 African trypanosomes

The genus *Trypanosoma* is a monophyletic clade within the unicellular protist group Kinetoplastida (Simpson *et al*, 2006). Members of *Trypanosoma* are obligate, flagellated parasites; and the subgroup of African salivarian trypanosomes causes a set of diseases that impose severe health and economic burdens on the affected areas. Salivaria currently comprises at least six species: *Trypanosoma brucei*, *Trypanosoma evansi*, *Trypanosoma equiperdum*, *Trypanosoma congolense*, *Trypanosoma simiae* and *Trypanosoma vivax* (Gibson, 2007). In addition to sleeping sickness in humans, which is caused by two *T. brucei* subspecies, salivarian trypanosomes also cause several diseases in animals. Nagana is a disease of ungulant livestock (most often cattle), characterised by fever and wasting, and caused by *T. brucei*, *T. congolense* and *T. vivax*. These parasites are transmitted by tsetse flies (and thus confined to the tsetse belt of Africa), and also infect wild animals, causing diseases referred to generally as animal African trypanosomiases. Surra has similar symptoms to nagana, is caused by *T. evansi* and affects numerous large domestic and wild mammals including cattle, horses, camels and buffalo. Surra has a wider range than nagana, being found in Asia, Africa and South America, and is transmitted mechanically (*i.e.* there are no insect-specific life cycle stages) by a wide range of vectors, namely blood-sucking insects of the genera *Tabanus*, *Stomoxys*, *Atylotus* and *Lyerosia*. Dourine is a disease of equines caused by *T. equiperdum*, which has a cosmopolitan distribution and is transmitted between hosts during coitus. *T. simiae* causes an acute porcine trypanosomiasis in tropical Africa (Brun *et al*, 1998; Gibson, 2007; Lai *et al*, 2008). *T. brucei* subspecies are the most intensively studied, because of their importance as the only African trypanosome that causes human disease, and because they are the most readily cultured in the laboratory.

Tsetse flies (genus *Glossina*) are the insect vectors for African trypanosomes, although several species of the parasite have expanded beyond the range of these insects by evolving tsetse-independent transmission, as mentioned above. The range of the vector is the so-called ‘tsetse belt’, stretching across the humid and semi-humid zones between the south edge of the Sahara (14ºN) and the north edge of the Kalahari desert (29ºS) (Barrett *et al*, 2003), an area covering ten million square kilometres and including parts of 36 countries (Black *et al*, 2001).

1.3.2 The life cycle of *T. brucei*

The life cycle of *T. brucei* contains several morphologically distinct stages (Figure 1-2), comprising several extracellular mammal-specific stages followed by several insect-specific stages (Vickerman, 1985), and alternates proliferative stages with quiescent stages adapted for transmission between hosts or movement between parts of a single host (Barry & McCulloch, 2001). Development in the mammalian host begins when an infected tsetse bites, taking a blood meal and injecting metacyclic stage parasites from its salivary glands into dermal connective tissue. From the point of entry, trypanosomes move into the draining lymph vessels and then into the bloodstream, where they differentiate into the proliferative long slender trypomastigote. This stage inhabits the circulatory and lymphatic systems and, later in infection, the brain and cerebrospinal fluid. The long slender form can either divide to increase the population, with a doubling time of about 6 h, or differentiate to the non-proliferative stumpy form (Vickerman, 1985). The stumpy form remains in G0 phase and has a limited lifespan, and is adapted for uptake by tsetse during feeding (Macgregor & Matthews, 2010).

Parasites are exposed to the mammalian immune system throughout the mammal life cycle stages, and they protect themselves from both innate and adaptive mammalian immune effectors with a surface coat composed of approximately 10 million molecules of variant surface glycoprotein (VSG). This structure is present in all mammal-specific life cycle stages, and also in the metacyclic stage as a preadaptation for mammalian infectivity. An individual parasite can be tackled by an immune response raised against a particular VSG, but the adaptive response is evaded at the population level because some parasites spontaneously switch to express another VSG, in a process of antigenic variation (see below). The rapid proliferation of the long slender form, coupled with killing by the adaptive immune response, contributes to a fluctuating parasitaemia with characteristic waves (Vickerman, 1985; Lythgoe *et al*, 2007). The dynamics of parasitaemia are also affected by the density-dependent dynamics of differentiation into stumpy form, because when a long slender form parasite differentiates to stumpy form the cell can no longer contribute to increasing parasite numbers. This process is predicted to be important to parasitaemia dynamics when numerous different VSGs are expressed at one time in the population, and has been shown to be dominant in chronic infections (Gjini *et al*, 2010; Macgregor *et al*, 2011).

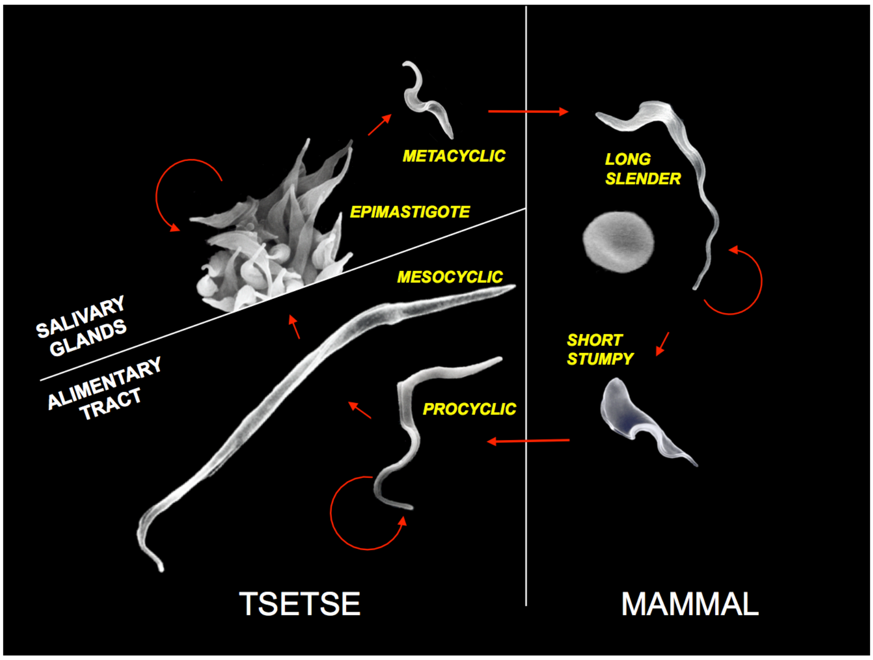


Figure 1-2 – Life cycle of *T. brucei*. Reproduced from Barry & McCulloch, 2001. Scanning electron micrographs of each life cycle stage are shown to scale, with an erythrocyte shown next to the long slender form. Curved red arrows indicate the potential for division of replication-capable life cycle stages, and straight arrows indicate differentiation into the next stage. White lines divide the different environments of mammalian tissue, the tsetse alimentary tract and the tsetse salivary glands.

When another tsetse bites the infected host, the insect ingests blood containing long slender and stumpy form parasites. Long slender parasites cannot proliferate in the tsetse, but stumpy form parasites differentiate into the proliferative procyclic form. Procyclic form parasites either divide to increase the population or differentiate to the mesocyclic form. Mesocyclic form parasites migrate to the salivary gland via the lumen of the gut, a step that appears to be a bottleneck in the life cycle (van den Abbeele *et al*, 1999; Oberle *et al*, 2010), and undergo an asymmetric cell division. The larger daughter cell is discarded and the other is the progenitor of the epimastigote form, a third proliferative stage that colonises the tsetse salivary gland. Epimastigotes can divide, or transform into metacyclic stage parasites that begin the life cycle again when the tsetse bites another host (Barry & McCulloch, 2001).

1.3.3 African trypanosomiases

Human African trypanosomiasis (HAT), or sleeping sickness, is caused by two *T. brucei* subspecies: *T. b. gambiense* and *T. b. rhodesiense*. The third subspecies, *T. b. brucei*, cannot infect humans. *T. b. gambiense* causes the chronic form of HAT, in west and central Africa, which accounts for around 90% of total HAT cases. *T. b. rhodesiense* causes the acute form of HAT in east and southern Africa. Infection with *T. b. rhodesiense* will eventually cause death if untreated; death is also a frequent outcome of untreated *T. b. gambiense* infection but tolerance and self-cure have also been reported (Jamonneau *et al*, 2012). Death will occur within a few weeks or months of infection with *T. b. rhodesiense*, whereas *T. b. gambiense* infections can last for several years before patients show major clinical signs (Cattand *et al*, 2001).

HAT has two clinically distinct stages: stage one (early), a generalised haemolymphatic infection, and stage two (late), when parasites have crossed the blood-brain barrier and invaded the central nervous system. Early stage disease has several symptoms, most of which are non-specific, including general malaise, rash, headache, weight loss, arthralgia and fatigue (Kennedy, 2008). One of the most common symptoms is an irregular fever that fluctuates with the waves of parasite proliferation (Stich *et al*, 2002). In stage two of infection, damage caused by the parasites and the host reaction results in a chronic encephalopathy, with severe headache and loss of concentration and co-ordination (Stich *et al*, 2002). Untreated, infection will eventually result in coma and death (Hide, 1999).

Early stage disease is treated with the drugs pentamidine and suramin, which are effective against *T. b. gambiense* and *T. b. rhodesiense* respectively. These drugs cannot cross the blood-brain barrier, so are ineffective if the disease has progressed to the late stage (Wilkinson & Kelly, 2009). Melarsoprol, an arsenic-based drug, is effective against late-stage infection by both parasites, but it is highly toxic, causing a reactive encephalopathy in around 10% of patients, resulting in death in 5% of all patients receiving this drug (Kennedy, 2008). Additionally, drug resistance and treatment failure are recognised problems (Legros *et al*, 1999). Alternative treatments are becoming available for late-stage *T. b. gambiense* disease: the drug eflornithine, which has few side effects, is being increasingly used, but is required in large doses that often make its use impractical (Barrett *et al*, 2003). Nifurtimox is a drug developed and licensed for use against South American trypanosomiasis, which is caused by a distinct trypanosome, but the drug has been used successfully against *T. b. gambiense* in combination with both melarsoprol and eflornithine; and an eflornithine-nifurtimox combination has been recommended by the WHO as a first-line treatment for the late-stage disease (Wilkinson & Kelly, 2009).

Animal trypanosomiases are similar to human sleeping sickness: wasting diseases that often will eventually result in death. A broader range of parasites cause disease in non-human animals than in humans, and the range includes *T. congolense*, *T. vivax*, *T. evansi*, *T. equiperdum* and all three subspecies of *T. brucei* (both human-infective subspecies plus *T. b. brucei*). *T. b. rhodesiense* and *T. b. brucei* have wide host ranges, including both domestic and wild animals. *T. b. brucei* is non-human-infective, and *T. b. rhodesiense* is considered primarily an animal parasite; that is, East African sleeping sickness is a zoonosis (Welburn *et al*, 2006). *T. b. gambiense* is considered to have its main reservoir in humans (Smith *et al*, 1998) but has also been identified in domestic pigs, and a number of wild animals including monkeys and duikers (Njiokou *et al*, 2006; Maudlin, 2006).

Different host species have varying degrees of trypanosome tolerance or resistance. Most cattle species and other domestic animals, if infected, will succumb to disease and usually die, as will a number of game animals including gazelle, jackal and monkey species (Black *et al*, 2001). Other species, for example bushbuck and hyena, and some African breeds of cattle such as N'dama, are susceptible to infection, but have a degree of trypanotolerance that allows parasitaemia to persist for some time without killing the animal, which may self-cure. Some species, for example baboons, are completely resistant to trypanosome infection (Mulla & Rickman, 1988). Humans are innately resistant to infection by most African trypanosome species due to the trypanosome lytic factors (TLF) of human serum: TLF1, which is a minor subclass of human high density lipoprotein (Hajduk *et al*, 1992); and TLF2, which is a high molecular weight protein complex (Raper *et al*, 1999). *T. b. rhodesiense* can infect humans by expressing the serum resistance associated protein (SRA), which prevents TLF-mediated lysis (Xong *et al*, 1998). The mechanism of action of SRA is still debated but the protein appears to be able to neutralise TLF1 once it has been taken up by the parasite (Vanhamme *et al*, 2003; Pérez-Morga *et al*, 2005; Oli *et al*, 2006; Stephens & Hajduk, 2011). *T. b. gambiense* resistance is SRA-independent, and the two subgroups of *T. b. gambiense* have different human serum resistance mechanisms: group 1 resistance arises partly from reduced expression of and mutations in the haptoglobin/haemoglobin receptor responsible for mediating TLF1 uptake, so TLF1 is not internalised (Kieft *et al*, 2010; Symula *et al*, 2012); while in group 2 TLF1 is internalised but prevented from bringing about lysis by an unknown mechanism (Capewell *et al*, 2011).

1.4 Antigenic variation in trypanosomes

1.4.1 Introduction to antigenic variation

In the process of antigenic variation, individual members of a clone or genotype of a pathogen periodically alter the molecules that they expose to the host immune system, such that an adaptive immune response raised against the first set is ineffective against the second set. This has two possible benefits for the pathogen: it allows infection to persist in the host, increasing the clone's chances of being transmitted to a new host (Barbour & Restrepo, 2000; Barry & McCulloch, 2001); and it allows re-infection of a previously infected host, overcoming herd immunity (Futse *et al*, 2008). There are three main requirements that allow a pathogen to use this strategy: access to a large number of genes encoding the antigens; the ability to express only one of these at once (allelic exclusion); and the ability to switch which gene is active (Borst & Genest, 2006). A large *permanent* repertoire is not essential, as there is an alternative strategy to provide numerous variants: the capacity to generate temporary genes from a small permanent repertoire. For example, short segments from different pseudogenes can be assembled in various combinations to form distinct functional genes, as occurs in *Anaplasma marginale* (Brayton *et al*, 2002), and also in *T. brucei,* where the strategy operates in conjunction with the use of a large permanent repertoire (see below). A high mutation rate for the antigen genes is advantageous, and is another alternative to a large permanent repertoire. Rapid mutation is important particularly in the case of viruses, which cannot directly encode many antigenic alternatives simultaneously due to the restricted size of their genome (Barry & McCulloch, 2001).

The strategy of antigenic variation is used by a number of pathogens other than *T. brucei* (Deitsch *et al*, 1997). For example, *P. falciparum* varies its erythrocyte membrane protein 1, which is displayed on the surface of the infected red blood cell, by expressing one of 60 *var* genes (Kraemer & Smith, 2006). *Giardia lamblia*, which causes an intestinal disease, has an immunogenic variable surface protein coat made by expressing one from about 190 possible genes (Prucca *et al*, 2008). Several *Borrelia* species use antigenic variation strategies, including the bacteria causing relapsing fever (*Borrelia hermsii*) and Lyme disease (*Borrelia burgdorferi*), which have surface coats of variable major protein and variable major protein-like protein respectively (Barbour & Restrepo, 2000). These pathogens, and numerous others including *Neisseria* species (bacterial species causing meningitis and gonorrhoea), the bacterium *A. marginale* (which infects cattle and goats) and the fungus *P. carinii*, all display antigenic variation as they persist in, or re-infect, their host (Deitsch *et al*, 1997).

1.4.2 Structure and function of VSG

Antigenic variation in *T. brucei* occurs in the VSGs that form the protective surface coat. Approximately five million VSG homodimers form a 12-15 nm layer on the outer surface of the plasma membrane of bloodstream form (BSF) and metacyclic form trypanosomes (Vickerman, 1985; Barry & McCulloch, 2001). By blocking antibody access to invariant potential antigens, such as glucose transporters and signalling proteins, the VSG coat enables the trypanosome to avoid the adaptive immune response (Overath *et al*, 1994; Barry & McCulloch, 2001). It also conceals invariant surface molecules and the cell membrane from the innate immune system, which they may activate (Ferrante & Allison, 1983; Schwede & Carrington, 2010).

The VSG coat itself is strongly immunogenic, and antibodies are raised by the host against the VSG that is exposed on the cell surface: this is why antigenic variation is required. Expression of *VSG* genes is tightly controlled, and only one gene at a time is expressed by each parasite. However, in addition to this single expressed gene, parasites possess a large number of silent *VSG*s, and long slender trypanosomes can switch which *VSG*s they express. A switch occurs stochastically in a fraction of the population, *i.e.* the antigen is varied. Switches are spontaneous, occurring even when parasites are grown in culture (Horn & Cross, 1997), and allow a fraction of the population to pre-emptively avoid an adaptive immune response. When antibodies appear against the original VSG, the parasites that have switched VSG survive to continue the infection, and proliferate to cause a new wave of parasitaemia (Barry & McCulloch, 2001). Parasites can probably express any one of thousands of VSGs capable of forming the surface coat, ensuring the success of trypanosome antigenic variation.

VSG probably has no function on the cell surface beyond forming the coat. VSG polypeptides form dimers with each subunit attached to the cell plasma membrane via a glycosylphosphatidylinositol (GPI) anchor (Ferguson, 1999). Most VSGs consist of 400 to 500 amino acid residues and contain two domains (Johnson & Cross, 1979; Carrington *et al*, 1991). The 350 to 400 residue N-terminal domain (NTD) is highly variable, with less than 20% sequence identity between different VSGs, and probably contains most of the epitopes against which antibodies are raised (Miller *et al*, 1984b; 1984a; Marcello & Barry, 2007b). The most conserved feature in the NTD primary sequence is the position of the cysteine residues that form disulphide bridges (Carrington *et al*, 1991). NTDs are divided into two types (A and B), based on the pattern of cysteines; originally a third group (C) was described, but this was later shown to be a subset of the A group (Carrington *et al*, 1991; Marcello & Barry, 2007b). The remainder of the VSG sequence, forming the C-terminal domain (CTD) is more conserved (around 40% sequence identity between genes) (Blum et al, 1993; Carrington et al, 1991). There are six CTD types (1 to 6), again defined by their patterns of cysteine residues (Carrington *et al*, 1991; Marcello & Barry, 2007b). Three CTD types appear to comprise two subdomains rather than a single CTD. From VSG studied so far, any combination of CTD and NTD types seems possible (Marcello & Barry, 2007b).

Despite the extensive variety in primary sequence in the NTD, higher-order structure appears to be conserved in the two VSGs for which the NTD structure has been solved (Freymann *et al*, 1990; Blum *et al*, 1993) (Figure 1-3). However, it should be pointed out that both these VSGs are type A, and no type B structure has been obtained experimentally. Two extended, antiparallel α helices interact to form a coiled coil (coloured blue in Figure 1-3), followed (in the primary sequence) by a surface-exposed region composed primarily of unstructured loops (coloured red in Figure 1-3). This ‘VSG fold’ appears to occur in a range of other trypanosome cell surface proteins with various functions, an arrangement which is thought to allow close packing of any VSG with the invariant proteins (Carrington & Boothroyd, 1996; Field & Boothroyd, 1996). The elongated NTD projects perpendicular to the plasma membrane, with the CTD adjacent to the membrane and covalently bound to the GPI anchor which attaches each monomer to the cell surface. The surface loops are unstructured, and so most amino acid changes here will not affect the protein folding fundamentally. Furthermore, the structure of the NTD is principally based on α helices, which do not depend on a specific primary sequence, but rather have requirements that can be provided by a number of residues and sequences. These structural features explain why the sequence of the NTD has great latitude for variation even though the higher-order structure is conserved.

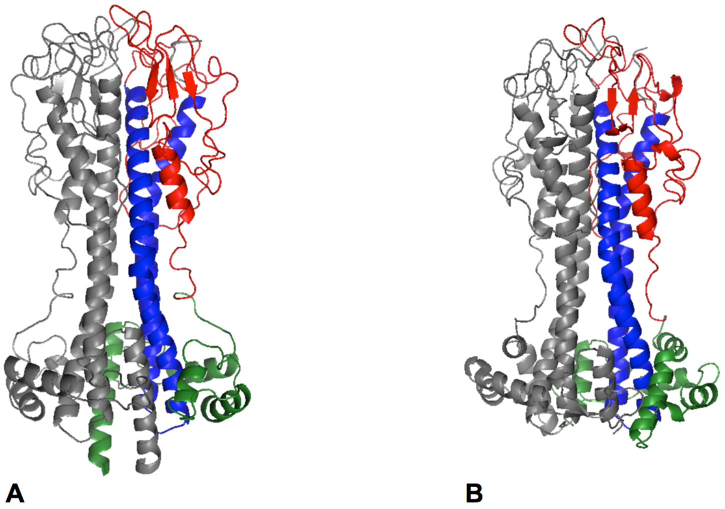


Figure 1-3 – Structures of VSG NTDs.
A) MITat 1.2, Freymann *et al*, 1990, PDB ID: 1VSG, B) ILTat 1.24, Blum *et al*, 1993, PDB ID: 2VSG. Solved structures from the indicated publications were downloaded from the RCSB Protein Data Bank (www.pdb.org) and visualised with PyMol. In each structure, one subunit is coloured grey; in the other, the coiled-coil helices are coloured blue, the surface loops are coloured red and the remaining, CTD-proximal region is coloured green.

1.5 Control of *VSG* expression

1.5.1 Control strategy

Given the critical role of VSG expression in antigenic variation, its regulation must achieve two things: only one VSG must be made at any one time (allelic exclusion), and the specific VSG made must periodically change. A further requirement is that VSG should only be produced at the appropriate life cycle stages. Although most regulation of protein production in the trypanosome occurs post-transcriptionally (Vanhamme & Pays, 1995), VSG expression appears to be an exception that is controlled at the level of transcription, and a *VSG* gene requires to be associated with an active promoter in order to be expressed. VSG production, therefore, is regulated by controlling which promoter is active and which gene is positioned so as to be transcribed from this promoter.

1.5.2 Allelic exclusion

VSG genes are transcribed from specialised telomeric environments, the VSG expression sites (ES) (Pays *et al*, 2001). Although there are several thousand *VSG* genes in the genome, transcription can occur only from these expression sites. There are two types of ES, metacyclic (MES) and bloodstream (BES), which are activated in the fly salivary glands and mammalian bloodstream respectively. MES and BES have somewhat different control mechanisms (Barry *et al*, 1998).

VSG ES are transcribed by RNA polymerase I (Pol I) (Lee & van der Ploeg, 1997; Günzl *et al*, 2003). In other organisms Pol I exclusively transcribes ribosomal DNA, while mRNA can only be synthesised by RNA polymerase II (Pol II). Trypanosomes are able to use Pol I to produce mRNA because they have separated the usual coupling of Pol II RNA synthesis with mRNA maturation processes such as 5′ capping (Günzl *et al*, 2003; Benz *et al*, 2005). However, most trypanosome protein-coding genes are still transcribed by Pol II (Vanhamme & Pays, 1995), so transcription of *VSG* by Pol I has implications for control of VSG expression.

Restriction of *VSG* transcription to ES prevents the transcription of most archive *VSG*s. However, a striking feature is that although theoretically there can be up to 23 BES (Young *et al*, 2008a), plus approximately 20 MES (Horn & Barry, 2005), only one ES at a time is used to produce protein (Barry & McCulloch, 2001). In probably all BES, polymerase is recruited and transcription is initiated, but transcriptional elongation is inhibited at all but the single active BES (Vanhamme *et al*, 2000). In the inactive MES, transcription initiation is inhibited (Graham & Barry, 1995). It is thought that this strict control of BES is achieved by specialised nuclear sublocalisation of the active ES (Navarro & Gull, 2001), coupled with effective repression of inactive ES.

The active ES does not localise within the nucleolus, as might have been expected from its transcription by Pol I (Chaves *et al*, 1998). Instead it is found in the expression site body (ESB), a subnuclear region where only the active ES appears to be transcribed (Navarro & Gull, 2001). This provides an explanation for the differential expression of BES: it is suggested that the ESB provides factors required to escape abortion of elongation, but can only accommodate one (the active) ES at a time, and all others are excluded and thus silenced (Navarro & Gull, 2001; Borst, 2002). Because DNA rearrangements are not required to switch which ES is active, several pathways of epigenetic control have been suggested to be involved in regulation. Depletion of proteins involved in chromatin remodelling (Hughes *et al*, 2007) and chromatin methylation (Figueiredo *et al*, 2008) has been shown to cause derepression of *VSG* in silent BES. Derepression also occurs on depletion of the trypanosome homologue of a yeast protein involved in the telomere position effect (Yang *et al*, 2009), a phenomenon in which genes located near telomeres are silenced (Gottschling *et al*, 1990; Glover & Horn, 2006), although inhibition or deletion of homologues of several other such genes appeared to have no effect on BES repression (Conway *et al*, 2002b; Alsford *et al*, 2007; Glover *et al*, 2007). Other proteins demonstrated to have a role in *VSG* silencing include ORC1, which is involved in DNA replication (Tiengwe *et al*, 2012b; Benmerzouga *et al*, 2012); histones, and histone chaperones (Povelones *et al*, 2012; Alsford & Horn, 2012). However, none of these studies described derepression of inactive BES to the levels of transcription seen from the active BES (Morrison *et al*, 2009), so it is clear that considerable detail remains to be elucidated.

1.5.3 *VSG* switching

The second key requirement for successful antigenic variation is that at least some parasites in the infecting population must switch the VSG that forms the surface coat. In non-laboratory-adapted strains, switching occurs at a rate of 10^{-2} to 10^{-3} switches/cell/generation (Turner, 1997). There are two main mechanisms whereby the VSG expressed can be changed: transcriptional (*in situ*) switching, and recombinational switching. Transcriptional switching involves silencing of the *VSG* previously being transcribed, accompanied by activation of the *VSG* (and associated genes) in another BES, without DNA rearrangement (Morrison *et al*, 2009). It is presumed that this mechanism of switching involves a change of the BES occupying the ESB (Borst, 2002), but details of how the change is initiated are not known, although there are indications that chromatin modification plays a role (Figueiredo *et al*, 2008), and that cell division is an important event for switching (Landeira *et al*, 2009).

Transcriptional switching, however, probably plays a minor role in generation of antigenic diversity in natural infections, not least because the pool of different *VSG*s available by this mechanism is small. The main mechanism driving antigenic variation is recombinational switching (Robinson *et al*, 1999). There are more than 1600 *VSG* genes in the genome strain of *T. brucei* (Marcello & Barry, 2007b), but most are located in subtelomeric arrays rather than ES (see section 1.6). Recombination reactions allow genes from this archive of silent, promoterless loci to be moved to the active ES, replacing the *VSG* there previously and thus changing which VSG is expressed.

There are three main types of recombinational switching, and it is not clear whether they share molecular mechanisms, nor whether each type of switch can be achieved by several different mechanisms (Morrison *et al*, 2009). The first type of switching is duplicative gene conversion (Robinson *et al*, 1999), in which a complete *VSG* gene is copied from a silent, promoterless site into the active BES (Barry, 1997). The second type of recombinational switching results in assembly of novel, mosaic *VSG* genes by combining different regions of pseudogenes (Thon *et al*, 1989; 1990). The third type, termed reciprocal recombination, exchanges chromosome ends between the active BES and an inactive, telomere-proximal *VSG*, so *VSG*s are swapped but the site of transcription initiation is not changed (Pays *et al*, 1985). However, this mechanism has been studied mainly in a laboratory-adapted strain (Rudenko *et al*, 1996; Aitcheson *et al*, 2005), and its relevance to antigenic variation in natural infections is unclear but probably minor (Verstrepen & Fink, 2009).

The pathways of homologous recombination are expected to be important in transferring silent *VSG*s to the active ES (Barry & McCulloch, 2001). A number of trypanosome homologues of proteins important in eukaryotic homologous recombination (San Filippo *et al*, 2008) have been tested for their effect on VSG switching. Several of these homologues were found to cause a decrease in the frequency of switching when mutated, including RAD51, the driver of homologous recombination (McCulloch & Barry, 1999); the RAD51-related protein RAD51-3, though other RAD51 paralogues appear to have no or a minor role (Proudfoot & McCulloch, 2005; Dobson *et al*, 2011); and BRCA2 (Hartley & McCulloch, 2008). It has been suggested that recombinational switching is initiated by double-stranded DNA breaks, which are usually repaired by homologous recombination (Barry, 1997; Horn, 2004). Twofold evidence for this hypothesis was provided by a study in which the induction of double-stranded breaks just upstream of the *VSG* gene in the active ES increased switching frequency, and spontaneous double-stranded breaks were detected in the same region (Boothroyd *et al*, 2009).

Switching has a semi-predictable order, such that the same VSGs are usually used to form a coat at broadly similar times in different infections with the same strain (Morrison *et al*, 2005). However, this order is not thought to be the result of the parasite favouring different switching mechanisms at different times, but rather, it emerges from the interaction of two features of antigenic variation: the existence of different types of switching mechanism, and the continuous pressure of immune selection. Since the various switching mechanisms involve, at least to some extent, different mechanisms and factors, they have different probabilities of occurring. For example, recombinational switching to activate *VSG*s resident in telomeric ES has a relatively high probability of occurring, because telomeres tend to interact with one another, promoting recombination (Barry *et al*, 2003), and because the inactive ES provides a long stretch of sequence homologous to the active ES (Hertz-Fowler *et al*, 2008), which also makes recombination more likely. By contrast, the assembly of an intact mosaic

VSG has a relatively low probability, because a single segmental conversion has a low probability of occurring due to the short regions of homology often involved, and assembly of a mosaic *VSG* requires a string of such events (Marcello *et al*, 2007b). Moreover, as there is no evidence so far of any mechanism to check mosaic genes for functionality, it can be inferred that even once a mosaic has been assembled, it may not be able to produce an effective coat.

The probability of each type of switch is thought to remain constant throughout infection. However, genes which have a high probability of activation will likely be first used to make surface coats, and thus seen by the immune system, quite early in infection. Therefore, although such genes will be frequently re-activated at later stages of infection, parasites that do so will be rapidly killed, because an immune response against their surface coat has already been raised (Barry & McCulloch, 2001). At this later stage of infection, although mosaic formation still has the same relatively low probability of occurring as it did at the early stages, when it does occur it will confer an antigenically novel surface coat upon the parasite, and thus allow escape from the immune response. These low-probability events will therefore make up a large proportion of successful switches at later stages of infection. In the discussion of different mechanisms I will refer to their *importance* at a particular stage of infection; this is not to be understood to mean that they do not occur at other stages, but that at stages where they are important, switches using these mechanisms constitute most of the successful switches that allows parasites to continue to proliferate.

1.6 Architecture of the *VSG* archive

1.6.1 *T. brucei* genome structure

The genome of *T. brucei* consists of 11 pairs of conventional diploid chromosomes (the ‘megabase chromosomes’), approximately 100 minichromosomes of 30 to 150 kb, and several intermediate chromosomes of 200 to 700 kb (Williams *et al*, 1982; Melville *et al*, 1998; 1999). There are estimated to be between 1600 and 2000 *VSG* genes in the genome of the reference strain TREU 927 (Berriman *et al*, 2005; Marcello & Barry, 2007b). *VSG* genes are found on all chromosome types, in four different locus types, which will now be described.

1.6.2 VSG expression sites

Two of the locus types are the VSG ES, which are adjacent to the telomeres of megabase and intermediate chromosomes. Genes in ES at the time of inoculation are important early in infection, as described above (Barry & McCulloch, 2001). There are estimated to be between five and 23 BES, with numbers varying between strains and subspecies due to the dynamic and unstable nature of telomeres (Becker *et al*, 2004; Hertz-Fowler *et al*, 2008; Young *et al*, 2008b). ES are usually between 40 and 80 kb in length, although ES as small as 24 kb have been observed (Zomerdijk *et al*, 1990; Berriman *et al*, 2002; Becker *et al*, 2004). A stylised representation of the structure of a BES is shown in Figure 1-4. The genes in the BES are all transcribed from a single, strong Pol I promoter (Zomerdijk *et al*, 1990); this type of polycistronic transcription is usual for trypanosome protein coding genes (Vanhamme & Pays, 1995). The *VSG* gene is always the most telomere-proximal gene, located adjacent to the telomere repeats (Vanhamme *et al*, 2001). Upstream of the *VSG* gene is an array of up to 10 kb of imperfect 70 base pair (bp) repeats. These AT-rich repeats are the 5′ limit of recombinational switching (Campbell *et al*, 1984; Liu *et al*, 1985; Shah *et al*, 1987). Such 70-bp repeats are also found associated with array *VSG* genes (see below), and are thought to provide the regions of homology necessary for homologous recombination of silent *VSG*s into the ES. In a low-switching, laboratory-adapted strain, deletion of 70-bp repeats did not abolish recombination of *VSG*s into the active BES (McCulloch *et al*, 1997), but the relevance of this observation is unclear because the strain used seems to be defective in recombinational switching, so the low level of switching that does occur probably differs in mechanism from recombinational switching in other strains (Barry, 1997). It is hypothesised that the 70-bp repeats are optimised to be prone to the double-stranded breaks that may initiate VSG switching (Alsford *et al*, 2009; Boothroyd *et al*, 2009), and indeed their sequence contains TAA motifs that have been shown to be physically unstable (Ohshima *et al*, 1996).

Between the 70-bp repeats and the promoter are several expression-site associated genes (*ESAG*s) (Cully *et al*, 1985), which encode various proteins, including likely surface receptors, transporters and signalling molecules (Pays *et al*, 2001). Some *ESAG*s are conserved between all characterised BES, while others are present only in some (Hertz-Fowler *et al*, 2008; McCulloch & Horn, 2009). Upstream of the promoter, and thus untranscribed, are several kb of 50-bp repeats (Zomerdijk *et al*, 1990), followed centromere-proximally by regions containing *RIME* and *INGI* retrotransposon and retrotransposon hot spot (*RHS*) sequences (Bringaud *et al*, 2002).

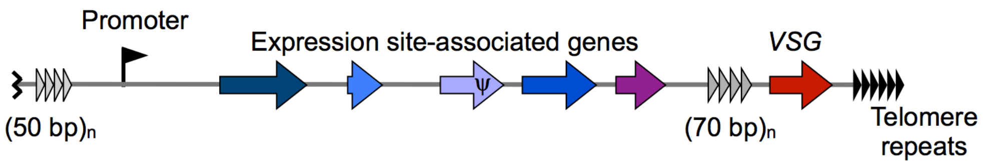


Figure 1-4 – Schematic diagram of a VSG bloodstream expression site.
Triangles represent short repeats (50-bp, 70-bp and telomere repeats); filled arrows represent genes, with Ψ denoting a pseudogene; the flag represents the promoter.

Expression from MES occurs in the metacyclic form and in BSF during the first few days of infection. The MES are unique among trypanosome protein-coding genes in being transcribed monocistronically, with a dedicated metacyclic-*VSG* promoter (Alarcon *et al*, 1994; Graham & Barry, 1995; Ginger *et al*, 2002). MES are used to express VSG in the fly salivary glands, at least partially as a preadaptation for mammalian infection, for which they may be essential. Their number is uncertain, but there may be up to 27 (Turner *et al*, 1988). Despite their monocistronic transcription, upstream of the promoter there is an untranscribed region containing *ESAG*s and *ESAG* pseudogenes, suggesting MES may have arisen from BES (Graham *et al*, 1999). As with the BES, in the MES the *VSG* gene is immediately telomere-proximal, but the promoter is much closer, located approximately 5 kb from the gene (Ginger *et al*, 2002). Between the gene and the promoter there are some of the 70-bp repeats seen in the BES, but here there are considerably fewer, usually only two or three. MES have also been described that do not contain any 70-bp repeats (Lenardo *et al*, 1986; Graham *et al*, 1999). Change in the MES *VSG* repertoire is slow, but turnover does occur (Barry *et al*, 1983).

1.6.3 Silent *VSG*

Although ES harbour a small pool of different *VSG*s, by far the majority of *VSG* genes lie elsewhere in the genome, where they are unexpressed, but can be copied into ES. Silent *VSG*s are found in two locus types: minichromosome subtelomeres, and megabase chromosome subtelomeric arrays.

Minichromosomes are linear molecules between 30 and 150 kb in size, consisting mainly of 177-bp repeats, and ending with standard telomere repeats (Weiden *et al*, 1991; Wickstead *et al*, 2004). Most, though not all, of the minichromosomes contain *VSG* genes, flanked upstream by a set of 70-bp repeats (Williams *et al*, 1982; van der Ploeg *et al*, 1984; Donelson, 2003). Each subtelomere can contain one *VSG*, located within 5 kb of the telomere repeats, and likely to be an intact gene (Alsford *et al*, 2001; Horn & Barry, 2005). There are approximately 100 minichromosomes in the *T. brucei* genome, making up around 10% of nuclear DNA, so there are potentially up to 200 minichromosomal *VSG*s, although few have been characterised (van der Ploeg *et al*, 1984; Barry *et al*, 2005). As the minichromosome sequence essentially consists only of the *VSG* gene and the tandem repeats, it has been suggested that minichromosomes have evolved to provide a pool of readily-accessible *VSG* genes (Wickstead *et al*, 2004; Barry & McCulloch, 2001). Intact minichromosomal *VSG*s tend to be important quite early in infections, presumably since only a single recombinational step is required to activate them, although they have shorter stretches of homology to the active ES than do *VSG*s in inactive ES (Robinson *et al*, 1999; Morrison *et al*, 2005).

The remainder of the *VSG* archive is contained in arrays located in the subtelomeres of the megabase chromosomes (Berriman *et al*, 2005) (Figure 1-5). Experimental infections examining a small number of VSGs have suggested that intact array genes usually become important later in infection than do minichromosomal genes, but still relatively early in infection (Morrison *et al*, 2005; Marcello & Barry, 2007b). Later in infection, however, mosaic VSGs assembled from pseudogenes appear to dominate, as explained above (Marcello & Barry, 2007b; Hall *et al*, 2013). The subtelomeric subarchive is arranged in tandem arrays, usually oriented away from the telomere, although some arrays contain strand switches in which the array changes orientation. In the genome strain trypanosome, an array can contain between three and around 250 genes and pseudogenes. Although *VSG* genes can be grouped into subfamilies by their N- and C-terminal domains, this organisation is not reflected in the arrangement of *VSG*s within and between arrays (Berriman *et al*, 2005; Marcello & Barry, 2007b). Array *VSG*s are arranged as cassettes similar to those of *VSG*s in other chromosomal loci. This cassette structure is thought to provide the silent *VSG*s with homology for recombination into BES. Within the cassette, upstream of the gene is at least one 70-bp repeat, and at the 3′ end there is a partially conserved region including coding and untranslated regions (Michels *et al*, 1983; Liu *et al*, 1983; Marcello & Barry, 2007b). The arrays are also interspersed with the retrotransposon *INGI*, particularly around strand switches (Berriman *et al*, 2005).

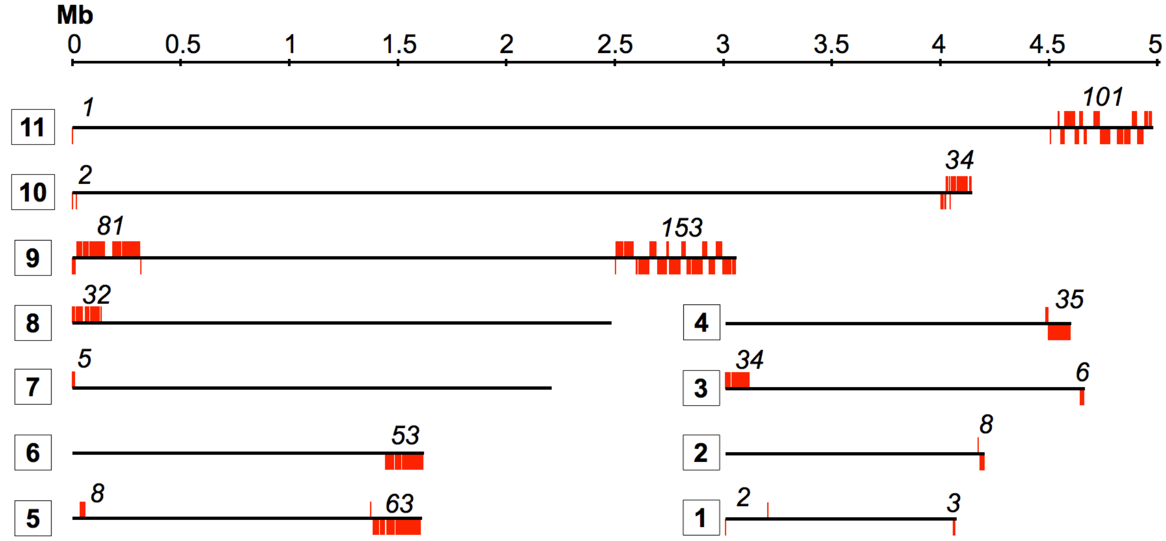


Figure 1-5 – Subtelomeric *VSG* arrays of TREU 927.
Subtelomeric *VSG* arrays in the TREU 927 genome assembly. Gene annotations were downloaded from TriTrypDB (Aslett *et al*, 2010) and filtered to include only those where the gene product description or primary name was 'VSG' or 'variant surface glycoprotein'. Chromosomes are represented by solid black lines; red lines projecting above the line of the chromosome represent *VSG* genes, pseudogenes or gene fragments on the forward strand; those projecting below represent the same on the reverse strand. Chromosome numbers are shown in boxes, numbers in italics above arrays indicate the number of genes at each subtelomere. For chromosome 11, the assembly consists of a main contiguation and two much smaller fragments: only the main contiguation is shown, although the fragments also contain *VSG*s. The scale bar indicates the lengths of chromosomes in Mb.

The arrays are effectively haploid, as they are not conserved between homologous chromosome pairs, and therefore the published trypanosome genome of eleven chromosomes does not cover the entire silent archive (Melville *et al*, 1999; Berriman *et al*, 2005; Hutchinson *et al*, 2007). Comparison between the assembled cores of megabase chromosomes and the physical size of these chromosomes, as estimated by pulsed-field gel electrophoresis (S. Melville, unpublished), has been used to estimate the size of the *VSG* arrays in the genome strain. This approach suggests that the arrays total approximately 8 Mb containing 1600 genes, half to two thirds of which have been analysed (Marcello & Barry, 2007b). However, the size of homologous chromosomes can vary widely between strains (Melville *et al*, 2000). A large amount of the difference between strains may lie in the *VSG* arrays, as has been shown for chromosome 1, one copy of which in one strain has over half of its length devoted to *VSG*s, amounting to around 3 Mb and potentially 600 *VSG* (Callejas *et al*, 2006; Marcello & Barry, 2007b).

Although the *VSG* arrays contain a large amount of DNA, analysis of the sequence has revealed that in TREU 927 only a small proportion (4.5%) of the annotated silent *VSG* cassettes contain fully functional genes (Figure 1-6). A further 9.5% are atypical genes, which may encode proteins with inconsistent VSG folding or post-translational modifications. Most (65%) are in fact pseudogenes, containing frameshifts or in-frame stop codons; and 21% are incomplete genes, encoding fragments of the protein, mostly CTDs (Berriman *et al*, 2005; Marcello & Barry, 2007b). This does not mean, however, that only 5% of the archive is useful for antigenic variation. Non-functional genes are still able to act as substrates for the recombinational process that generates novel genes that are mosaics or hybrids of genomic *VSG*s (Thon *et al*, 1990; Kamper & Barbet, 1992; Morrison *et al*, 2005; Marcello & Barry, 2007b). These mosaic genes are probably assembled by multiple steps of segmental gene conversion, as the recombination events are usually between sequences with only short stretches of homology (Conway *et al*, 2002b; Taylor & Rudenko, 2006). The predominance of pseudogenes in the archive means that the generation of mosaic VSGs might assume greater importance than was previously thought (Marcello & Barry, 2007b), and detailed analysis of the VSGs expressed during infection reinforced this importance (Hall *et al*, 2013).

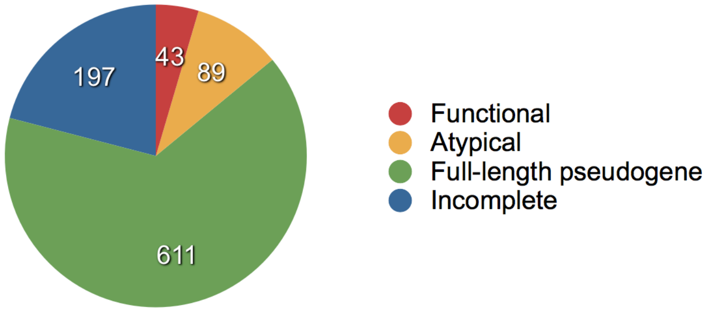


Figure 1-6 – Composition of the TREU 927 *VSG* archive.
Although the TREU 927 archive is estimated to contain around 1600 *VSG*, not all of the archive genes are present in the assembled genome. Around 900 genes have been annotated and analysed and it is these genes that are shown. Data are from Marcello & Barry, 2007b. An 'incomplete' gene is usually an isolated NTD or CTD. The number of genes in each category is indicated in the appropriate segment.

1.6.4 Role of archive composition and substructure in generation of antigenic variation

A *VSG* archive provides functional VSGs of sufficient antigenic diversity to achieve three things: the survival and persistence of a population of parasites in the mammalian host; the overcoming of herd immunity by allowing re-infection of previously infected hosts; and the possibility of competition between strains (Barry *et al*, 2005; Marcello & Barry, 2007a). The composition and substructure of the archive are well suited to providing such an antigenic range, in a number of ways. Simply the scale of the archive provides a considerable resource, even if only full-length, intact *VSG* genes are considered. If the *VSG*s analysed so far are representative of the entire archive, functional genes in the arrays number between 50 and 100 (Marcello & Barry, 2007b), and there are up to 200 on minichromosomes, which although not well characterised are probably mostly intact (Marcello *et al*, 2007). As described above, other pathogens successfully persist by antigenic variation using no more different variants than there are intact *VSG* genes: for example, *P. falciparum* uses only around 60 *var* genes.

Of course, the potential repertoire of VSGs is much greater than this, because of the parasite's ability to produce functional genes by combining pseudogenes, a process that is thought to be key to trypanosome survival (Kamper & Barbet, 1992; Barbet & Kamper, 1993). It has further been argued that the use of recombined pseudogenes rather than full-length genomic genes has been actively selected for as a means of overcoming herd immunity: if pseudogenes are used then mosaic *VSG*s will be assembled anew in every infection, meaning a somewhat different string of antigens will be produced every time (Marcello & Barry, 2007a; Barry *et al*, 2012). Analysis of the archive found around 40% of *VSG* sequences belong to small, high-identity subfamilies, usually with two or three members. This subfamily structure is suggested to be key to mosaic formation by providing regions of high identity between donor genes, allowing them to recombine readily (Marcello & Barry, 2007b). The other key feature of the *VSG* archive that promotes antigenic diversity is its fast rate of evolution, which will be discussed in the next section.

1.7 Evolution of the *VSG* archive

1.7.1 Hyperevolution of *VSG*s

The *VSG* archive appears to evolve rapidly. Strains that have minor differences in housekeeping genes can have large differences in their *VSG* repertoires, and repertoires have diverged to become strain-specific (Bernards *et al*, 1986; Hutchinson *et al*, 2007). An individual *VSG* gene will be invisible to selection unless it is expressed, which for most genes is probably only rarely. Furthermore, because there are so many genes, it seems likely that any single gene could be lost with very little effect on the overall capacity of a population for antigenic variation. It therefore seems unlikely that hyperevolution is the result of strong diversifying selection on individual *VSG* genes. More likely is the hypothesis that mechanisms have evolved, either *de novo* or through adaptation of existing processes, to generate mutations of various sorts in *VSG*s, leading to the observed hypermutation (Barry *et al*, 2012).

Analysis of the *VSG* arrays in the genome strain led to the proposal that there is a dynamic equilibrium between gene duplication and rapid diversification. The process results in the previously described subfamily structure that promotes mosaic formation, and produces novel genomic *VSG* genes. Inferred mutations from this analysis include duplication of whole genes and gene segments, point mutations and short indels, in addition to larger-scale contractions and expansions of the arrays mediated by the *INGI* retrotransposon (Marcello & Barry, 2007b).

Segregation of subtelomeric *VSG* genes from chromosome cores provides a means for them to be subject to different activities. For example, divergence of *VSG* repertoires between strains is hypothesised to be due to cessation of meiotic genetic exchange between telomeres and between subtelomeres. Meiotic recombination tends to homogenise gene sequences between homologous chromosomes, so its absence would allow different changes to accumulate in different strains (Hutchinson *et al*, 2007). Furthermore, because subtelomeres in general are subject to hyperevolution, it is proposed that the subtelomeric location of the *VSG* arrays is key to the evolutionary mechanisms that lead to diversification (Marcello & Barry, 2007b).

1.7.2 Possible mechanisms of evolution

Subtelomeres have been observed to undergo elevated levels of ectopic and homologous recombination (Freitas-Junior *et al*, 2000; Mefford *et al*, 2001; Rudd *et al*, 2007), and it is likely that these processes promote segmental duplication and the generation of new combinations of *VSG* sequence. Because the archive is very large, and the segments converted can be very small, it is possible that such recombination would generate sufficient antigenic diversity for the parasite to successfully persist in infections. However, it is by no means clear that recombination is able to generate enough diversity, and it may be necessary for new variation to be introduced into the archive. Further, inference of the events of *VSG* evolution by comparing genes within the TREU 927 archive suggests that smaller-scale sequence diversification such as point mutation and short indels are also important (Marcello & Barry, 2007b). No detailed study has been made of the processes causing these mutations in subtelomeres, and it is not clear whether they are genuinely different from those occurring in chromosome cores, or whether the difference is because relaxed selection on individual *VSG*s allows mutations to accumulate. However, there are several mechanisms that could plausibly play a role in elevating mutagenic processes in *VSG*, and these are discussed in the following sections.

1.7.2.1 Recombination and repair activities

Ongoing characterisation of recombination and repair pathways in *T. brucei* is revealing that many proteins involved in these pathways in other organisms have homologues in *T. brucei* that act in similar processes, although there are some large differences (reviewed in Machado *et al*, 2006). Trypanosome components of homologous recombination and DNA repair activities include RAD51, a key recombinase enzyme (McCulloch & Barry, 1999; San Filippo *et al*, 2008); five RAD-51 paralogues, which are thought to act as accessory factors to the enzyme (Proudfoot & McCulloch, 2005); BRCA2, a regulator of homologous recombination that interacts with RAD51, although the role of BRCA2 appears to be somewhat different in trypanosomes from other organisms (Hartley & McCulloch, 2008; Oyola *et al*, 2009; Trenaman *et al*, 2012); the multifunctional MRE11 (Tan *et al*, 2002; Robinson *et al*, 2002); and several mismatch repair proteins, MSH2, MSH3, MSH8, MLH1 and PMS1 (Bell *et al*, 2004). In addition to homologous recombination, there is a second, RAD51-independent pathway of microhomology-based recombination, for which components have not been characterised (Conway *et al*, 2002a; Barnes & McCulloch, 2007; Glover *et al*, 2008). For most of these proteins, the effects of deletion have been characterised, and two in particular had striking effects on the trypanosome's subtelomeres.

MRE11 is the central protein in a tripartite nuclease complex that in yeast and in mammals is important in maintaining genomic integrity, apparently through its involvement in checkpoint signalling and DNA replication, and is required for homologous recombination (D'Amours & Jackson, 2002). Deletion of both alleles of the *MRE11* gene in *T. brucei* resulted in gross chromosomal rearrangements, which involved only decreases in chromosome size (*i.e.* loss of sequence rather than duplicative translocations), and did not appear to affect telomeres or BES *VSG* but involved loss of array *VSG* genes in at least half of clones examined (Robinson *et al*, 2002). Chromosomal rearrangements were not detectable in the intermediate chromosomes and minichromosomes. The pattern of chromosomal rearrangements is in contrast to that in S. *cerevisiae* cells lacking MRE11, where 90% of rearrangements are either duplicative translocations or additions of new telomeres (Chen & Kolodner, 1999). Further analysis with chromosome core markers showed the loss of trypanosome sequence to be almost exclusively restricted to subtelomeres (A. Browitt, A. MacLeod, J.D. Barry, unpublished).

BRCA2 is also important in eukaryotic genome stability, and is involved in regulation of homologous recombination. In mammals and in *Leishmania infantum*, the protein appears to regulate RAD51 by binding it and facilitating its transport to the nucleus (Davies *et al*, 2001; San Filippo *et al*, 2008; Genois *et al*, 2012). In *T. brucei*, BRCA2 does interact with RAD51, but it does not appear to act in RAD51 transport to the nucleus (Trenaman *et al*, 2012). Deletion of the *T. brucei BRCA2* gene in BSF parasites, though not in the procyclic form, resulted in gross chromosomal rearrangements similar to those resulting from *MRE11* deletion: most chromosomes that changed appeared to become smaller; rearrangements were detectable in only the megabase chromosomes; and *VSG* loss occurred in at least some cases (Hartley & McCulloch, 2008; Trenaman *et al*, 2012). The findings for these two key DNA repair genes suggest that these activities may act differently, or with a different rate, in subtelomeres compared with chromosome cores. It can be envisaged that such differences in exposure to recombination and repair processes could result in differences in mutation profiles.

1.7.2.2 Error-prone polymerases

Cells have many strategies to allow survival following DNA damage, as its consequences may be catastrophic. One such strategy involves the use of lesion bypass polymerases, which can insert bases opposite to DNA lesions that act as blocks to replicative polymerases. Most eukaryotic lesion bypass polymerases belong to the Y family, and have specialised structures that include the classic DNA polymerase fold and catalytic residues despite having very little sequence identity to replicative polymerases (Waters *et al*, 2009). Their active sites are larger and more open than that of other DNA polymerases, a difference that allows accommodation of bulky adducts. Lesion bypass polymerases also make fewer contacts with the DNA substrate than do replicative polymerases (Prakash *et al*, 2005). These structural features, coupled with lack of 5′-3′ proofreading activity, give lesion bypass DNA polymerases low replicative fidelity, with error rates that range from two to seven orders of magnitude higher than that of replicative polymerases (Yang, 2005; Waters *et al*, 2009). A second consequence is reduced stability of the DNA-polymerase complex, a feature that, although it promotes translesion synthesis, also decreases enzyme processivity on undamaged DNA (Prakash *et al*, 2005).

In addition to their function in tolerance of DNA damage, the low fidelity of lesion bypass polymerases led to the suggestion that they have a role in the systematic generation of mutations (Radman, 1999). Such a mutagenic process is seen, for example, in the somatic hypermutation that occurs in immunoglobulin genes in B cells, promoting antibody diversity, and is also used by cells to generate adaptive mutations (Goodman, 2002). It has further been suggested that error-prone DNA polymerases have a role in generating mutations in trypanosome *VSG* genes, in a process involving DNA synthesis primed by homologous recombination (McKenzie & Rosenberg, 2001). Several human lesion bypass polymerases have mutation profiles that are similar to the indels and point mutations that are predicted to occur in *VSG* arrays (Marcello & Barry, 2007b). However, it is likely that there will be species differences in the specific types and frequencies of mutation each polymerase introduces, and perhaps more relevant than a particular mutation profile is the fact of a *T. brucei*-specific expansion in the DNA polymerase κ (pol κ) gene family.

Human pol κ has moderate processivity and low fidelity. Its most frequent error type is single-base substitutions (5.8×10^{-3}/base/replicative cycle) followed in frequency by single-base deletions (1.8×10^{-3}/base/replicative cycle) and then single-base insertions, and it can also cause larger or more complex deletions and additions (Johnson *et al*, 2000; Ohashi *et al*, 2000a; 2000b; Hile & Eckert, 2008). *Leishmania major* has three copies of the pol κ gene and *T. cruzi* has two, but *T. brucei* has ten tandemly arranged copies (Rajão *et al*, 2009). The *T. brucei* family of pol κ genes, which has diverged into two groups based on the sequence of the DNA binding domain, models well onto the solved structure of human pol κ, and has mutated residues lining the substrate-binding and active sites (Uljon *et al*, 2004; J.D. Barry, and B. Marchetti, unpublished). It is hypothesised that *T. brucei* has adapted pol κ to specifically mutate trypanosome subtelomeres.

1.7.2.3 The *VSG* archive as a model for evolution of subtelomeres

The evolution of subtelomeres is clearly important in the variation of organisms, but little is understood about the mechanisms that generate diversity in these chromosomal regions. I therefore suggest that the *VSG* archive of *T. brucei* is a suitable model to study specific mutations in subtelomere evolution and the mechanisms possibly involved. Several considerations make this system appropriate to use as a model. Firstly, the *VSG* archive as a whole is critical for parasite survival in the mammalian host, with a strong selective pressure acting on the phenotype — diversity of antigenic coat — and thus on the mechanisms that generate diversity. Secondly, as noted in the previous section, the archive evolves at a rapid rate, with divergence detectable in isolates separated by only a short evolutionary time. Thirdly, suitable resources exist to study this parasite: the genome of one strain has been sequenced and assembled to a high standard (Berriman *et al*, 2005) and several other strains have also been sequenced or efforts are under way (Jackson *et al*, 2010; Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/resources/downloads/protozoa/trypanosoma-brucei.html), 2013; this thesis); and a large collection of isolates from similar areas and known dates (from the 1960s onwards) exists, from which the archives can be compared. Finally, work on areas such as VSG switching has already provided some clues as to mechanisms that might play a part in hyperevolution of the subtelomeric archives.

1.8 Aims of the project

The broad purpose of the work described in this thesis was to investigate the processes occurring in hyperevolution of *VSG* genes in *T. brucei*. The first aim of the project was to test the hypothesis that the subtelomeric *VSG*s are subject to different and more rapid mutagenic processes than are genes in the chromosome cores. The strategy used was to identify the changes that had occurred in the *VSG* archive between two very closely-related isolates of the same trypanosome strain, and compare them with the changes that had occurred in the core genes. This aim required the sequencing and assembly of the genomes of the two isolates, the annotation and pairing of *VSG*s in and between the genomes, and the identification of mutations that had occurred. The second aim was to test the hypothesis that members of the expanded pol κ gene family may have a role in mutagenesis in the *VSG*s. The strategy used was to express and purify one gene from each of the two pol κ subfamilies, characterise their mutation profiles to determine whether either was consistent with the mutations observed in *VSG*s, and examine their subcellular localisation, in particular whether the proteins colocalised with subtelomeres.

Chapter 2: Methods and Materials

2 Methods and Materials

2.1 Bioinformatic methods

For all programs described, default settings and parameters were used unless otherwise stated. Full names of programs and appropriate references are given in Table 2-1, along with a short description of its function in this work.

2.1.1 Assembly

2.1.1.1 Assembly algorithms and settings

Assembly was carried out mainly at the Sanger Institute by Dr Thomas Otto. First, the EATRO 2340 454 libraries were assembled *de novo* using Celera assembler. The resulting contigs were ordered against *T. b. gambiense* strain DAL 972 (Jackson *et al*, 2010), using ABACAS, with minimum alignment length of 300 bp and identity cutoff of 97%. After ordering, the 6 kb and 8 kb 454 libraries were mapped to the new assembly and a custom script was used to find positions with fewer than three reads in proper pairs mapped, and to break the contigs at these positions. Finally, IMAGE was used to close gaps, with the following k-mer sizes, in order: 55 (4 iterations), 49 (7 iterations), 41 (6 iterations) and 37 (4 iterations).

The EATRO 3 Illumina reads were trimmed using SGA (version sga-0.9.9) using the following sequence of commands: ‘sga filter -t 10 -d 32’; ‘sga-0.9.9 index -t 10 --disk=50000’; ‘sga-0.9.9 correct -k 31 -x 3 -t 10 --discard’. The trimmed reads were assembled with Velvet Columbus, using the EATRO 2340 assembly as the reference. A custom script was used to change the base call to N (*i.e.* ambiguity) for any position not covered by at least two reads. ABACAS was used as before to order the resulting contigs to EATRO 2340. IMAGE was used for 13 iterations with a k-mer of 31.

2.1.1.2 Read mapping

Illumina reads were mapped to assembled genomes using the Burrows-Wheeler aligner. The mapping was performed for each set of reads (forward and reverse) by the algorithm bwa aln using a maximum insert size of 800, then forward and reverse read mappings were joined with bwa sampe. Further processing was performed with SAMTools 0.1.17, to convert .sam files to .bam files using samtools view, and to order and index .bam files using samtools sort and samtools index. Mapped reads were visualised with BamView in Artemis. To look at coverage, the samtools mpileup command was used, with the flag -g. From the mpileup output, plots for each contig of the number of reads and the number of good-quality reads (having a Phred-scaled base quality score of at least 23 at that position) mapped to each position were generated with the scripts pileup2plot.pl and pileup2plot_qual.pl, and coverage statistics calculated with the script coverage_from_plots.pl.

2.1.2 Genome annotation

2.1.2.1 RATT

Annotations of TREU 927 core genes were transferred automatically to the EATRO genome assemblies using RATT. The configuration file used was RATT.config_euk, provided with the program, and the transfer type used was Strain.

2.1.2.2 Annotation of NTDs in full-length *VSG* genes

ORFs in the subtelomeres were annotated with Artemis, using a minimum length of 430 amino acids. The resulting ORF sequences were processed using the script extract_genes.pl to trim them to the first in-frame ATG and then exclude genes that were not between 1290 and 1650 bp in length. The filtered genes were concatenated with a spacer sequence between each one consisting of 1 kb of sequence randomly drawn from TREU 927 chromosome 10. This concatenated sequence was screened for windows likely to contain VSG NTD or CTD sequence by using SVM-VSG, a support vector machine-based program provided by Dr Jon Wilkes, using the script arrayFinder.pl with the settings arrayGuide.txt followed by the script array2vsg.pl. ORFs that were predicted to contain both NTD and CTD sequence, *i.e.* were likely to be full-length *VSG* genes, were extracted from the concatenated file by the script extract_vsg.pl.

To separate the NTD and CTD, the ORFs were translated in frame 1 and queried with hidden Markov models (HMM) for NTD type A, NTD type B and CTD (provided by Dr Bill Wickstead, University of Nottingham) using hmmsearch from the HMMer 3.0 suite of programs. If an NTD was found, the end of the NTD was used to define the domain boundary, and the NTD type of the gene was defined according to which HMM had a hit. If no NTD was found the ORF was queried with the CTD HMM, and the conserved CTD cysteines in the DNA sequence of the HMM hit were located by eye. The domain boundary was defined as 50 residues upstream of the first conserved cysteine, and the NTD type was defined as 'unknown'. The 'VSG translation' (see section 4.5.4) of these NTDs was simply the frame 1 translation.

2.1.2.3 Annotation of NTDs in *VSG* pseudogenes

Windows containing likely NTD sequence were located in contigs longer than 10 kb using SVM-VSG, as for the concatenated ORFs; and windows containing assembly gaps were excluded. Translations were produced for the window sequences in all three frames, then the translations were queried with the NTD A-type and B-type HMMs using hmmsearch. The results were parsed using the script hmmer_3frames.pl (parameters: eval=0.001, length=40, length2=300, met_range=20, dis_limit=50, max_overlap=200). Windows in which an NTD could not be annotated using HMMer were used in BLASTX to query a protein BLAST database generated with makeblastdb using NTDs from the appropriate genome that had been annotated with HMMer and from full-length genes. The results were parsed using the script blastx_hits.pl (parameters: id_cut=40, len_cut=40, len2=900, dist_limit=100, met_range=20, max_overlap=200). Both hmmer_3frames.pl and blastx_hits.pl output a predicted 'VSG translation' for each NTD based on the region of each translation frame used to annotate the NTD.

2.1.3 Sequence comparisons and analyses

2.1.3.1 Phylogenetic analysis

The chromosomes were concatenated for each strain and all ten possible pairings of the full genomes were aligned using MUMmer (version 3.23) to find maximal unique matches (MUMs), using the script make_coords_tbrucei.sh, which was based on a script provided by Dr Nick Dickens. Using the script mums2distance_tbrucei.pl, also based on a script provided by Dr Nick Dickens, a distance matrix was constructed for each genome pair A and B using the formula –log(((fraction of genome A covered by MUMs to genome B) + (fraction of genome B covered by MUMs to genome A))/(length of A + length of B)). A neighbour-joining tree was generated using BIONJ and visualised in Dendroscope.

2.1.3.2 Global SNP annotation

Putative SNPs were called from the mapped .bam files using SAMTools. The command samtools mpileup was used to parse the .bam file to extract information on base and mapping quality and compute SNP likelihoods, with filters for base quality and downgrading of reads containing excessive mismatches (command line: 'samtools mpileup -g -Q 23 -C 10 -f <.fasta file> <.bam file>), which generated a raw .bcf file. This was followed by calling SNPs and filtering using bcftools view and vcfutils varFilter from the SAMTools package (commands: 'bcftools view -bcgv <raw.bcf>'; 'bcftools view <unfiltered.bcf> | vcfutils.pl varFilter -D 300). The resulting .vcf file was filtered further for minimum and maximum depth and to remove SNPs with multiple alternative alleles using the script vcf_filter.pl (parameters: dp4_min=10, dp4_max=200). Plots of SNPs for each contig for each set of mapped reads were generated with the script snps_plot.pl. Finally, the script compare_snps.pl was used to find positions in the contig where a SNP had been called using one set of mapped reads, but not the other, to check that the position had sufficient coverage with both sets for a SNP to be called, and to generate a plot of the resulting EATRO 3-EATRO 2340 SNPs.

2.1.3.3 Identification of gene pairs

Initial pairings were made by querying a nucleotide BLAST database of the contigs from one genome (Genome A), with the NTDs annotated from the other genome (Genome B), using BLASTX. The best hit or set of hits was identified based on score, length and pairwise identity, using the script find_homologues.pl (parameters: dis_limit=400, max_overlap=200). The sequence of the best hit was extracted from Genome A using hits_for_reblast.pl, with any gaps represented by a single 'N', and was used to query a database Genome B. The script reciprocal_blast.pl (reciprocal_blast_anygene.pl for core genes) was used to check whether the best hit for these sequences was the gene in Genome B that hit them in the first step; if not, the gene in Genome B was excluded from the analysis.

Genes with a hit with at least 95% identity and 95% coverage were extracted from the results. For the core genes, because only EATRO 2340 queries were used, the gene partner was simply taken as the best hit in EATRO 3 (if any met the identity and coverage thresholds), translations were those of frame 1, and no further processing was carried out. For the *VSG* NTDs, instances where the high-identity hit corresponded to a previously annotated NTD were identified for each set of NTD queries using the script find_shared_orfs.pl, and the reciprocity of gene pairings was checked with check_pairings.pl, and any that were not reciprocal pairs were manually checked. The scripts orf_covered_part.pl and full_orf_hit.pl were used to find if there was any sequence at the end of the query gene that was not matched to any sequence in the hit, and then to extend the co-ordinates of the ‘hits’ to compensate for this; for example, if a hit covered positions 5-1000 of an NTD that was 1005 bp long, the hit was extended by 4 bp at the start and 5 bp at the end. For ‘hits’ that comprised several BLAST high-scoring segment pairs (HSP), this process was done manually. Finally, to annotate new genes from the BLAST hits, the sequences of hits that did not correspond to previously annotated NTDs were extracted and their co-ordinates noted. Hits containing internal gaps were excluded and hits beginning or terminating in gaps were cropped to remove the gap, and the co-ordinates adjusted accordingly using the scripts seqs_without_ns.pl, trim_off_ns.pl and fix_ns_coords.pl. To predict ‘VSG translations’, the new genes were used in BLASTX to query a BLASTP database constructed from all the NTDs annotated previous to this step. The BLASTX hits were filtered so that each new NTD only had hits to the gene that had been used to annotate it originally, then the script blastx_hits_rehits.pl was used to predict the VSG translations based on which frames the BLASTX hits were in.

2.1.3.4 SNP cataloguing and processing

For ease of data handling and analysis, a set of Perl object classes was created. The packages containing methods used by these classes are contained in Appendix E1 as the files <class name>.pm. The class ORFPair handled paired-up sequences, containing information like the co-ordinates and sequence of each partner and their pairwise identity (and used the class ORFs, which contained simple details of the origin and co-ordinates of each NTD). ORFPair objects were generated from paired NTDs by the script seq_pair_object.pl. The script align_orfpairs.pl then worked on these objects to produce a pairwise alignment with ClustalW2 (non-default parameter: -pwgapopen=4). Alignments of gene pairs where the hit had been pieced together from multiple adjacent hits were manually inspected, and if necessary corrected using Jalview. For core genes, the alignments were not manually checked, and the script cds_alns.pl was used to generate ORFPair objects from paired sequences.

For cataloguing SNPs, the classes Singlediffs, Plotdiffs and Changes were used, for raw and confirmed SNPs respectively. For Method 1 (data not used for final analysis), SNPs were catalogued with the scripts orfs_alns.pl (to find differences in the alignments, to make Singlediffs objects), check_snps.pl (to check whether either position already had a SNP called in the SNP plot), and class_snps.pl (to make Changes objects), using SNP plots generated as described in section 2.1.3.2. For Method 2, snps_from_plots.pl found SNPs in the gene in either genome using the SNP plots; singlediffs_from_plots.pl discarded SNPs where there was a SNP at the same position in the alignment in both genomes, and generated Plotdiffs objects; and class_snps_fromplots.pl made Changes objects.

The objects and classes described above were used for the SNP analysis. Substitution type and strand bias were examined using the scripts class_subs.pl, and composition.pl. The effect on predicted protein sequence was examined with the script synon_subs_sitetypes.pl and the numbers of each type of site (synonymous/non-synonymous) were counted with synon_sites.pl and synon_sites_sepgenomes.pl. The distribution of SNPs was examined with sub_pos.pl and sub_relative2cys.pl. Clustering was examined with Permute III, provided by Prof Dan Haydon, which uses random permutation to find windows with significantly higher or lower numbers of events than expected, with window sizes of 10 and then 30-600 bp at 30 bp intervals, using 1000 permutations with a confidence level of 0.01. Secondary structure prediction was carried out with the Jpred 3 server, and the structures extracted from the output with jpred_parse.pl. Statistical analyses were carried out in R using the MASS library.

2.1.3.5 Archive analyses

The class NTDs was created to handle and access attributes of the NTDs, including their sequences, locations, pseudogene status, method of annotation and which regions were used to make the translation. NTDs objects were created with the scripts ntds_objects.pl and ntds_objects2.pl. The functionality of each NTD was predicted with the script functional_genes.pl, with CTD positions used from the SVM predictions. NTDs predicted to be intact were tested for signal sequences using SignalP with the settings ‘eukaryote’ and ‘do not include TM regions’.

For analysis of subfamilies, a BLAST protein database was generated from predicted VSG translations and queried with the same sequences using default settings. Pairs with higher identity than 50% were identified, aligned and grouped into subfamilies with the script families.pl.

2.1.4 Other programs

Searches with BLAST were carried out with programs from the command line suite BLAST 2.2.25+. Common subroutines for custom Perl scripts are given in the file SequenceHandling.pm in Appendix E1. The BioPerl version used was BioPerl-1.6.901. The R version was R 2.15.1 GUI; models were fitted with glm (for Poisson distributions) and glm.nb from the library MASS (for negative binomial distributions).

2.1.5 Glossary of software

Program	Full name	Function (in this work)	Reference
ABACAS	Algorithm-Based Automatic Contiguation of Assembled Sequences	Aligns *de novo* assembled contigs to reference genome	Assefa *et al*, 2009
ACT	Artemis Comparison Tool	Tool for viewing pairwise comparisons of DNA sequences	Carver *et al*, 2005
Artemis	-	Sequence and annotation viewing tool	Rutherford *et al*, 2000
BAMview	Binary Alignment/Map viewer	Viewing of BAM files in Artemis	Carver *et al*, 2010
BIONJ	-	Generates neighbour-joining phylogenetic trees	Gascuel, 1997
BioPerl	-	Perl code modules for biology	Stajich *et al*, 2002
BLAST	Basic Local Alignment Search Tool	Aligning sequences and finding homologues	Altschul *et al*, 1990; Camacho *et al*, 2009
BWA	Burrows-Wheeler Aligner	Mapping of Illumina reads to assembled genome	Li & Durbin, 2009
Celera Assembler	-	*De novo* assembler of short sequencing reads	http://sourceforge.net/apps/mediawiki/wgs-assembler/index.php?title=Main_Page
Clustalw1.8	-	Aligning sequences (version built into BioPerl)	Thompson *et al*, 1994
clustalw2	-	Aligning sequence; this version used unless version 1.8 is specified	Larkin *et al*, 2007
Dendroscope	-	Tree viewer	Huson & Scornavacca, 2012
HMMer	-	Searches sequences for homologues using hidden Markov models	Eddy, 2011
IMAGE	Iterative Mapping and Assembly for Gap Elimination	Targeted reassembly of sequencing reads to fill gaps in assembly	Tsai *et al*, 2010
Jalview	-	Alignment viewing and editing	Waterhouse *et al*, 2009
Jpred 3	-	Prediction of secondary structure from amino acid sequences	Cole *et al*, 2008
MUMmer	Maximal Unique Matcher	Sequence aligner	Kurtz *et al*, 2004
Newbler	-	*De novo* assembler of 454 sequencing reads	http://www.454.com/products/analysis-software/index.asp
RATT	Rapid Annotation Transfer Tool	Automatic transferral of annotations from reference genome	Otto *et al*, 2011

Table 2-1 – Glossary of software used in the project.
Continued on next page.

Program	Full name	Function (in this work)	Reference
SAMTools	Sequence Alignment/Map Tools	Processing of BWA output, SNP calling and filtering	Li *et al*, 2009
SGA	String Graph Assembler	Trimming of poor-quality bases from ends of Illumina reads	Simpson & Durbin, 2012
SignalP	-	Predicts signal peptide (cleavage sites) in amino acid sequences	Petersen *et al*, 2011
Velvet Columbus	-	Short sequencing read assembly using reference genome as a guide, followed by *de novo* assembly of reads that don't assemble to reference	Zerbino & Birney, 2008; http://www.ebi.ac.uk/~zerbino/velvet/Columbus_manual.pdf
SVM-VSG	Support Vector Machine for VSG	Identification of sequence likely to contain VSG, by support vector machine learning	Jon Wilkes, unpublished

Table 2-1 continued.

2.2 Molecular biology methods

All procedures were carried out at room temperature (approximately 25°C) unless otherwise specified.

2.2.1 High-throughput sequencing

Purified DNA was prepared for sequencing by standard library methods, or by a PCR-free approach (Kozarewa *et al*, 2009), as indicated in Table 2-2. Sequence sets 5-7, 13 and 14 (see Table 3-1) were generated by Illumina sequencing, and sequence sets 15 and 16 were generated by 454 Titanium sequencing, both at the Sanger Institute (collaboration, Dr Matt Berriman). Sequence sets 1-3, and 8-11 were generated by 454 Titanium sequencing at the University of Liverpool; and sequence sets 4 and 12 were generated by 454 Titanium sequencing at the University of Nottingham (collaboration, Prof Ed Louis).

2.2.2 Growth and manipulation of trypanosomes

2.2.2.1 Trypanosome strains

For *in vitro* analysis of pol κ, BSF Lister 427 90:13 was used (Wirtz *et al*, 1999). EATRO 3 was grown in mice from a stabilate of the original ETat 1.2 clone, derived from the original EATRO 3 isolate by 32 passages, mainly through mice (McNeillage *et al*, 1969; Lumsden & Herbert, 1975; in both of these papers ETat 1.2 is referred to as ETat 2). EATRO 2340 was grown from a stabilate of a clonal, fly-passaged line derived from the original EATRO 2340 isolate, stock F described by Barry and colleagues (Barry *et al*, 1983); the stabilate was first inoculated into an immunosuppressed mouse, then passaged into a second mouse, then extracted into procyclic form culture.

2.2.2.2 *In vitro* trypanosome culture

BSF Lister 427 90:13 parasites and their derivatives were grown in HMI-9 growth medium (Hirumi & Hirumi, 1989) supplemented with 20% (v/v) heat-inactivated foetal calf serum (FCS, Sigma) and 1% (v/v) penicillin/streptomycin mix (Sigma), in a humidified 5% CO_2 incubator at 37°C. Cell lines were maintained in 2 ml cultures in 24-well tissue culture plates, and were passaged three times a week by transferring 20 µl culture into fresh media. Drugs used were phleomycin (2 µg/ml), neomycin (2.5 µg/ml), hygromycin (5 µg/ml) and tetracycline (2 µg/ml). Lister 427 90:13 required neomycin and hygromycin for maintenance, phleomycin was used as the selective drug for parasites transfected with overexpression constructs, and tetracycline was used to induce overexpression.

2.2.2.3 Parasite density and growth curves

Parasite density in cultures was estimated using an Improved Neubauer haemocytometer (Hawksley) by taking a 10 µl sample, counting the number of cells in several 1 mm^2 fields (0.1 µl volume) culture and calculating the mean number of cells/ml in the fields. If the number of cells was low, the mean of four fields was used, otherwise the mean of two was used. For growth curves, parasites were counted then diluted to approximately 2×10^5 cells/ml in 4 ml cultures in 6-well tissue culture plates. If necessary, appropriate volumes of tetracycline (5 mg/ml in 50% ethanol) and/or hydrogen peroxide (3% solution in sterile, deionised water (dH_2O)) were added.

2.2.2.4 Growth of trypanosomes in mice

This work was carried out by Alana Hamilton. ICR mice (Charles River) were immunosuppressed by intraperitoneal injection of 200 µl of 31.25 mg/ml cyclophosphamide solution, 24 h before inoculation with parasites. Parasites were prepared for injection by placing a blood straw into 250 µl CBSS (0.023 M HEPES, pH 7.4, 0.12 M NaCl, 5.41 mM KCl, 0.4 mM $MgSO_4$, 5.6 mM Na_2PO_4, 0.035 M glucose, 0.05 mM phenol red) containing 10000 U/l heparin, and the whole volume of CBSS/blood mixture was injected into one mouse. After 3 days, the mice were euthanised and infected blood from each mouse was harvested by cardiac puncture into 200 µl CBSS with 10000 U/l heparin.

2.2.2.5 Transfection of trypanosomes

Parasite transfections were carried out using the Amaxa Nucleofector apparatus (Lonza). Approximately 3×10^7 BSF parasites were harvested by centrifugation at 460×*g* for 10 min and resuspended in 100 ml Amaxa T-cell solution (Lonza) at 4°C. Linearised plasmid DNA dissolved in dH_2O (10 µg DNA in a maximum volume of 15 µl) was added to a cuvette, followed by the resuspended cells, and the parasites were transfected using program X-001. The transfected cells were immediately transferred to 30 ml HMI-9 containing drugs appropriate to growth of the parental strain (but not the selective drug) and mixed. Three ml of this resuspension was transferred to 27 ml HMI-9 containing parental strain drugs. The contents of each tube was aliquoted into a 24-well tissue culture plate, 1 ml per well. The plates were incubated at 37°C for 18 h. Media was prepared by supplementing HMI-9 with the parental strain drugs, and the selective drug at twice its required concentration. One ml of this was added to each well of transfected cells. Plates were inspected for growth of transfectants after 5 days. Transfectants were transferred to culture plates and cultured as normal.

2.2.2.6 Isolation of genomic DNA from trypanosomes

For small amounts of genomic DNA, the DNeasy Blood and Tissue kit (Qiagen) was used. For BSF parasites grown in culture, approximately 5×10^6 cells were harvested by centrifugation at 460×*g* for 10 min and washed twice with 1 ml ice-cold, filter-sterilised phosphate-buffered saline (cfPBS). The cells were resuspended in 200 µl PBS and processed according to the manufacturer’s protocol for animal blood or cells. To extract DNA from parasites grown in mice, aliquots of blood from infected mice extracted into anticoagulant were centrifuged at 1620×*g* for 10 min at 25°C, which separated the blood into several layers, with the trypanosomes present in the ‘buffy coat’ layer on top of lysed blood cells. The buffy coat was transferred to a new Eppendorf tube and resuspended in 200 µl PBS per 200 µl original sample, then processed according to the manufacturer's protocol for animal blood or cells using one column for each 200 µl resuspended cells.

2.2.3 Basic laboratory procedures

2.2.3.1 Polymerase chain reaction (PCR)

For routine PCRs, *Taq* DNA polymerase (New England Biolabs, NEB) was used. Reactions contained template, 0.1 mM each deoxyribonucleotide triphosphate (dNTP), 0.5 µM each primer, 1X Thermopol buffer (supplied with enzyme) and 0.5 U *Taq* enzyme in a total volume of 20 µl, made up with dH_2O. Template was either approximately 10 ng genomic DNA, 1 ng plasmid DNA, or an *Escherichia coli* colony picked from a plate and resuspended in 20 µl dH_2O, in which case 5 µl of the resuspension was used. Standard cycling conditions were 5 min at 95°C, followed by 30 cycles of 30 s at 95°C, 30 s at the annealing temperature (Table 2-2) and 1 min/kb of product at 72°C, followed by a final extension step of 5 min at 72°C.

For PCRs where maximum accuracy was required (in production of vectors that would be expressed in *E. coli* or in *T. brucei*, and for amplification of *VSG* genes for sequencing), a proof-reading polymerase was used, either Herculase II Fusion DNA polymerase (Agilent Technologies) or Phusion High-Fidelity DNA polymerase (NEB) if amplification with Herculase was unsuccessful. Herculase reactions contained 10 ng plasmid DNA or 160 ng genomic DNA, 1 mM each dNTP, 0.25 µM each primer, 1X Herculase reaction buffer (supplied with enzyme) and 1 µl Herculase enzyme (from kit, catalogue #600675, concentration not available) in a total volume of 50 µl, made up with dH_2O. Standard cycling conditions for Herculase were 5 min at 95°C, followed by 30 cycles of 15 s at 95°C, 30 s at the annealing temperature (Table 2-2) and 30 s/kb of product at 72°C, followed by a final extension step of 5 min at 72°C. Phusion reactions contained 10 ng plasmid DNA, 1 mM each dNTP, 0.5 µM each primer, 1X Phusion HF reaction buffer (supplied with enzyme) and 1 U Phusion enzyme in a total volume of 50 µl, made up with dH_2O. Standard cycling conditions for Phusion were 30 s at 98°C, followed by 30 cycles of 10 s at 98°C, 30 s at the annealing temperature (Table 2-2) and 30 s/kb of product at 72°C, followed by a final extension step of 10 min at 72°C.

All oligonucleotides were synthesised by Eurofins MWG Operon, and sequences are given in Table 2-2.

2.2.3.2 Agarose gel electrophoresis and gel extraction

DNA separation was usually performed by electrophoresis in gels consisting of 1% (w/v) agarose in 1X Tris-acetate-EDTA (TAE) buffer (40 mM Tris pH 8.0, 19 mM acetic acid, 1 mM ethylenediaminetetraacetic acid (EDTA)) with 1X SYBR-Safe DNA gel stain (Invitrogen). Loading buffer (0.25% bromophenol blue, 40% sucrose in dH_2O) was added to DNA samples (1 volume buffer to 9 volumes sample) before loading on to the gel. Gels were run at 100 V in 1X TAE buffer, typically for 40 min. DNA size standards (1 kb Plus DNA ladder, Invitrogen) were resolved alongside sample DNA. Stained DNA in the gels was visualised in a GelDoc, and images captured with Quantity One software (BioRad). Apparatus was supplied by BioRad.

Where required, DNA bands were excised from the gel using a sterile scalpel blade. DNA was extracted from the excised agarose using a Qiaquick Gel Extraction kit (Qiagen), according to the manufacturer's instructions. DNA was quantified using a Nanodrop 1000 spectrophotometer (Nanodrop).

2.2.3.3 Restriction digests

Enzymes and buffers for restriction digests were supplied by NEB. Reactions were carried out with approximately 10 units of enzyme per 1 µg DNA, and incubated at 37°C for 1 h (diagnostic digests), 3 h (preparation of DNA for ligation) or overnight (linearisation of DNA for transfection into parasites).

2.2.3.4 Transformation and growth of *E. coli*

Where home-made competent cells were used (see section 2.2.3.5), approximately 10 ng of purified plasmid or 10 µl ligation (see section 2.2.3.6) were added to 100 µl competent cells. The cells were incubated on ice for 5-10 min then heat-shocked at 42°C for 45 s and returned to ice to 10-15 min. Next, 250 µl SOC (20 g/l tryptone, 5 g/l yeast extract, 0.5 g/l NaCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$ and 20 mM glucose) was added and the transformations were incubated for 1 h at 37°C with shaking. The transformations were then spread on

LB agar plates containing the appropriate antibiotics and incubated at 37°C overnight. For transformations with ligations, typically 200 µl of transformation per plate was used; for transformations with purified plasmid, typically 20 µl per plate was used. For transformation using commercial competent cells, transformation was carried out according to the manufacturer's instructions. Colonies that grew were picked from plates, if necessary checked by colony PCR, and then used to inoculate 5 ml of LB with appropriate antibiotics, which was then incubated overnight at 37°C with shaking.

Stabilates were optionally prepared from overnight cultures by adding 500 µl culture to 1 ml 2% tryptone, 40% glycerol in dH_2O. Plasmids were purified from overnight cultures using the Qiaprep Spin Miniprep kit (Qiagen). All sequencing was done using plasmids prepared from transformed *E. coli* by overnight culture and miniprep, by Eurofins MWG Operon.

2.2.3.5 Preparation of competent cells

A colony was picked from a fresh overnight plate of the relevant strain and used to inoculate 100 ml LB in a 500 ml flask, then incubated at 37°C with shaking until the culture reached an OD_{600} of 0.35. Next, 50 ml of this culture was cooled on ice then centrifuged at 2100×*g* for 20 min at 4°C. The supernatant was discarded and the cells drained, then resuspended in 30 ml ice-cold filtered 80 mM $MgCl_2$, 20 mM $CaCl_2$ in dH_2O, centrifuged as before, resuspended in 2 ml ice-cold filtered 0.1 M $CaCl_2$ in dH_2O and used directly for transformation.

2.2.3.6 Cloning

Gel-purified PCR product (insert) and miniprep of plasmid DNA (vector) were treated with the appropriate restriction enzymes. Enzyme was removed from digested insert using Qiaquick PCR purification kit (Qiagen), and the appropriate part of the digested vector was isolated and purified by gel extraction. A standard ligation reaction used approximately 150 ng of insert and a 5:1 insert:vector molar ratio with 400 U T4 DNA ligase (NEB) in the supplied reaction buffer, in a volume of 20 µl, for 3-4 h at 25°C.

PCR products for sequencing were cloned into pCR4-TOPO using the TOPO TA Cloning Kit for sequencing (Invitrogen). PCR products of the correct size were purified by gel extraction and A-tailed: 8.7 µl of gel-extracted PCR product (approximately 200 ng DNA) was incubated in 1X Thermopol buffer (NEB), 0.5 U *Taq* enzyme and 0.2 mM each dNTP in a total volume of 10 µl for 20 min at 72°C. 4 µl of the A-tailing reaction was used directly for the ligation reaction and then this was transformed into One Shot TOP10 competent cells, both according to the manufacturers' instructions.

2.2.3.7 Protein electrophoresis

Protein separation was performed by polyacrylamide gel electrophoresis (PAGE) in NuPAGE Novex 10% Bis-Tris polyacrylamide gels (Invitrogen). Protein samples were prepared by adding NuPAGE LDS Sample Buffer (Invitrogen) to 1X concentration, and heating at 100°C, 10 min. Gels were run at 200 V in 1X MOPS SDS running buffer (Invitrogen). Apparatus was supplied by Invitrogen. To visualise protein, gels were stained with Coomassie stain solution (2.5 g/l Coomassie brilliant blue R (Sigma) in 45% methanol, 45% dH_2O, 10% glacial acetic acid, (v/v)) for 1.5 to 3 h, then washed overnight in destaining solution (10% glacial acetic acid, 40% methanol, 50% dH_2O). To produce figures, gels were dried in a gel dryer (BioRad) and scanned.

2.2.3.8 Western blotting

Proteins were transferred from polyacrylamide gels to Hybond ECL nitrocellulose membrane (Amersham) using Xcell blotting apparatus (Invitrogen). Gel, membrane, filter paper and sponges were equilibrated in 1X NuPAGE transfer buffer (Invitrogen) before being assembled in the apparatus according to the manufacturer's instructions. Transfer was carried out at 30 V for 1 h.

2.2.3.9 Probing of western blots

Nitrocellulose membranes to which protein had been transferred were incubated in blocking buffer (5% milk powder (w/v, Marvel), 0.1% Tween 20 in PBS) overnight at 4°C. Membranes were rinsed in PBS-T (0.1% Tween 20 in PBS), then incubated for 1 h in blocking buffer containing the primary antibody. The primary antibodies used were rabbit polyclonal anti-GFP (Abcam), used at a dilution of 1:1000; and mouse monoclonal anti-MYC (Sigma), used at a dilution of 1:7000. The membranes were washed in PBS-T for 15 minutes followed by two five-minute washes, then incubated for 1 h in blocking buffer containing goat anti-rabbit or anti-mouse antibody conjugated to horseradish peroxidase (Molecular Probes) at a dilution of 1:5000. Membranes were washed three times in PBS-T as before. SuperSignal West Pico Chemiluminescent Substrate (Pierce) was applied to the membrane (2 ml per blot) and incubated for five minutes and placing the membrane in the dark. Signal from the reaction was detected by exposing the membrane to X-ray film (Kodak) for 5 s to 4 h. X-ray films were developed in a Kodak M35M Xomat processor.

2.2.3.10 Site-directed mutagenesis

Site-directed mutagenesis of pol κ was carried out with the QuikChange II XL mutagenesis kit (Agilent Technologies). The reactions were carried out according to the manufacturer's instructions, using 10 ng plasmid DNA as a template. The mutant strand synthesis reaction conditions were 1 min at 95°C followed by 18 cycles of 50 s at 95°C, 50 s at 60°C and 8.5 min at 68°C, then a final extension step of 7 min at 68°C. Mutants were screened by sequencing with P74, P78, T7 and T7term.

2.2.4 Protein production and purification

2.2.4.1 Production of recombinant protein in *E. coli*

A glycerol stock of Rosetta2 *E. coli* cells that had been transformed with pMAL-K10 or pMAL-K001 was used to inoculate 10 ml LB containing 34 µg/ml chloramphenicol and 100 µg/ml ampicillin, and grown at 37°C overnight. This overnight culture was used to inoculate 50-200 ml LB supplemented with 2 g/l glucose and containing 34 µg/ml chloramphenicol and 100 µg/ml ampicillin, using 1 ml overnight culture per 100 ml LB. The new culture was incubated at 37°C until the culture gave an absorbance at 600 nm of approximately 0.5 U, usually around 3 h. Expression of the fusion protein was induced by adding isopropyl β-D-1-thiogalactopyranoside (IPTG) to a final concentration of 0.3 mM. The culture was grown at 37°C for 2 h, or at 15°C overnight, or at 30°C for 4 h, then cells were harvested by centrifugation at 2100×*g* for 20 min at 4°C.

2.2.4.2 Amylose affinity chromatography

Harvested cells were resuspended in 40 ml column buffer (20 mM Tris, pH 7.5, 200 mM NaCl, 1 mM EDTA, 1 mM dithiothreitol (DTT)) per 1 l original culture and frozen at least overnight at -20ºC. The cell suspension was then thawed on ice and sonicated on ice with 6 cycles of 16 s on followed by 10 s off. Insoluble material was removed by centrifugation at 3275×*g* for 30 min. The soluble material was retained and diluted 1:5 in column buffer.

Purification columns were prepared by placing a filter into a 2.5×10 cm plastic column (Pierce), pouring 10 ml amylose resin (NEB) per 1 l original cell culture (usually 4 ml) and placing a second filter on top of the resin. The resin was equilibrated by washing with 8 column volumes (CV) of column buffer. Cell lysate was loaded onto the column, then washed with at least 15 CV column buffer plus 0.1 mM maltose. Fusion protein was eluted from the column with column buffer plus 10 mM maltose. If only one round of purification was carried out, eluted protein was collected as 10×1 ml fractions and one 5 ml fraction. If a second round was to be carried out, the protein was eluted in 4 CV and the maltose was removed as far as possible by reducing the eluate volume to approximately 500 µl in a Vivaspin 20 centrifugal concentrator with a 30 kDa molecular weight cut-off (Sartorius), then making up to 20 ml with column buffer (no maltose). The volume reduction was repeated twice more, making the volume up to 50 ml after the final reduction. The amylose column was regenerated according to the manufacturer’s protocol, with 3 CV dH_2O, 3 CV 0.1% sodium dodecyl sulphate (SDS), 2 CV dH_2O and 3 CV column buffer.

For a second round of purification, the regenerated column was equilibrated with 20 ml column buffer and the first-round eluate loaded on to it, washed as before, eluted with column buffer plus 10 mM maltose and collected as 10×1 ml fractions and one 5 ml fraction. The concentration of protein was determined with a Nanodrop 1000 spectrophotometer (Nanodrop). Fractions containing protein were pooled and concentrated as far as possible with a Vivaspin 20 centrifugal concentrator with a 30 kDa molecular weight cut-off.

2.2.4.3 Inclusion body purification

MBP-polκ protein was produced in a 200 ml culture by induction at 37°C and harvested by centrifugation as described. The supernatant was removed and the mass of the cell pellet determined, and the pellet was resuspended in 3 ml cell lysis buffer I (50 mM Tris-Cl, pH 8.0, 1 mM EDTA, 100 mM NaCl) per 1 g of pellet. Protease inhibitor cocktail (20X stock made by dissolving one tablet of Complete EDTA protease inhibitor (Roche) in 2 ml dH_2O) was added to 1X concentration, followed by 80 µl of 10 mg/ml lysozyme (Sigma) per 1 g of pellet, and the suspension was stirred for 20 min at 4°C. The lysate was then transferred to 37°C and stirred occasionally until it became viscous. To digest DNA, approximately 2.5 U DNase I (amplification grade, Invitrogen) was added per 1 g pellet, and the lysate was stirred at 25°C until it was no longer viscous.

The lysate was centrifuged at 21000×*g* for 15 min at 4°C to pellet insoluble material. The supernatant was removed and the pellet was resuspended in 9 volumes of cell lysis buffer II (50 mM Tris-Cl pH 8.0, 10 mM EDTA, 100 mM NaCl, 0.5% Triton X-100, 5 mM DTT), and incubated at 25°C for 5 min. The suspension was centrifuged, resuspended and centrifuged as before, and the supernatant was removed. The pellet was resuspended in 9 volumes 50 mM Tris, pH 8.0, 10 mM EDTA and transferred to a tube of known weight, then centrifuged as before. The supernatant was removed and the pellet dried as completely as possible, then the weight of the inclusion body was determined and it was stored at -70°C.

2.2.4.4 Inclusion body solubilisation and refolding

Inclusion bodies were solubilised by resuspending in a small volume of 50 mM Tris pH 8.0, then adding appropriate volumes of the other components of the solubilisation buffer (7 M guanidine hydrochloride (Pierce), 50 mM Tris pH 8.0, 5 mM DTT), using a total of 1 ml solubilisation buffer per 40 mg inclusion body, then incubating overnight at 4°C. The solubilised protein was then centrifuged at 21000×*g* for 15 min at 4°C to pellet any remaining insoluble matter, and the supernatant retained. This yielded a protein solution of approximately 20 mg/ml. For low protein concentration refolding experiments, a 1 mg/ml solution was prepared by dilution in solubilisation buffer.

For each refolding condition, 950 µl of a 1.1X refolding buffer was prepared at 4°C. Solubilised protein was added as 5×10 µl aliquots, with mixing by pipetting between each addition. When all 50 µl had been added, the solutions were vortexed then incubated at 4°C for 18-24 h. After this time, the solutions were centrifuged at 21000×*g* for 15 min at 4°C to pellet any aggregated protein. To remove chaotropic agents that might be artificially keeping unfolded protein in solution, and would interfere with downstream analyses, 100 µl of each refolding reaction was dialysed into column buffer using Slide-a-Lyzer mini dialysis cups (Pierce), for 2 h at 4°C in 1 l buffer per 8 cups. Dialysed refolded protein was stored at 4°C.

2.2.5 Assays

2.2.5.1 Growth of M13 bacteriophage

To produce M13mp18 double-stranded DNA (dsDNA), 0.02 ng M13mp18 RF1 DNA (NEB) or the product of the gap-filling assay was used to transform MAX Efficiency DH5αF′IQ competent cells (Invitrogen), according to the manufacturer's instructions, except that the volumes of competent cells and SOC media used were reduced to one-quarter of that specified. After transformation, 200 µl transformed cells were added to 3 ml aliquots of 2X YT (16 g/l tryptone, 10 g/l yeast extract, 5 g/l NaCl) top agar (7 g/l agar) at 45°C containing 0.03% 5-bromo-4-chloro-indolyl-β-D-galactopyranoside (X-Gal) and 3.3mM IPTG, followed by 100 µl DH5αF'IQ lawn cells (Invitrogen). The top agar was poured on top of 2X YT agar plates (15 g/l agar) and the plate incubated at 37°C overnight.

To grow larger amounts of M13mp18 dsDNA, DH5αF'IQ lawn cells were grown overnight in 2X YT broth, and 5 ml of this culture was used to inoculate 500 ml 2X YT broth. A blue plaque was picked from a plate containing cells transformed with M13mp18 and added to this broth, and the culture was grown at 37°C overnight. The cells were harvested by centrifugation at 2100×*g* for 20 min at 4°C, and M13mp18 dsDNA was extracted using a Genelute HP maxiprep kit (Sigma), according to the manufacturer's instructions, using two columns. Smaller amounts were produced by inoculating a 6 h culture of lawn cells grown in LB with cells picked from a white plaque, incubating the culture at 37°C overnight and extracting M13mp18 dsDNA using a Qiaprep Spin Miniprep kit (Qiagen).

2.2.5.2 Construction of gapped M13 DNA substrate

To construct the template for the gap-filling assay (Bebenek & Kunkel, 1995), M13mp18 double-stranded DNA was digested to completion with *Kas*I and *Pvu*I, giving two restriction fragments of 407 bp and 6848 bp. The 6848 bp fragment was purified by agarose gel electrophoresis and gel extraction, followed by dilution of the DNA 1:10 in dH_2O to decrease the salt concentration as far as possible. The DNA was heated at 70°C for 5 min to denature strands. This low temperature, low salt concentration denaturation temperature was developed and validated in the original paper (Bebenek & Kunkel, 1995), and was used to avoid heat-induced DNA damage to the substrate. Next, sufficient M13mp18 single-stranded circular DNA (NEB) was added to produce a 1:1 molar ratio of minus strand, 6848-bp fragment to circular DNA. The hybridisation reaction was incubated on ice for 5 min, then 20X saline-sodium citrate buffer (SSC, 3 M NaCl, 300 mM sodium citrate) was added to a final concentration of 2X SSC. The reaction was heated at 60°C for 5 min then placed on ice. The hybridised DNA was cleaned up using a QiaQuick PCR purification kit (Qiagen), and analysed by agarose gel electrophoresis to check the size of the product.

2.2.5.3 Gap-filling assay

The reaction conditions for the gap-filling reaction were 50 mM NaCl, 10 mM Tris-HCl, pH 7.9, 10 mM $MgCl_2$, 1 mM DTT, 200 µg/ml BSA, approximately 300 ng of polymerase and approximately 40 ng of gapped substrate, in a total volume of 25 µl. Reactions were incubated at 37°C for 2 to 3 h, then stopped by the addition of EDTA to a final concentration of 15 mM. Agarose gel electrophoresis was used to analyse 20 µl of the reaction mixture. A 1:50 dilution of reaction mixture in dH_2O was used to transform MAX Efficiency DH5αF′IQ competent cells (Invitrogen), and grown as for DNA production (section 2.2.5.1).

2.2.5.4 Primer extension assays

Substrates and assays were based on the assay of Rajão and colleagues (Rajão *et al*, 2009). The bipartite primer-template consisted of fluorescein-labelled M13 primer and unlabelled DloopF (template) oligonucleotides. The tripartite substrate for the strand displacement assay consisted of the same primer and template plus the unlabelled oligonucleotide PL3. The substrates were produced by annealing the oligonucleotides in 50 mM Tris (pH 7.5) and 100 mM NaCl in a volume of 10-50 µl; 0.5 µM primer and 0.8 µM template, with 0.8 µM PL3 for the tripartite substrate, were heated to 80ºC for 5 min, then cooled slowly to 25ºC.

For both assays, DNA polymerase (approximately 1 µg) was incubated with 50 nM substrate for 2 h at 37ºC, in a 10 µl reaction. Reaction conditions were 10 mM Tris, pH 7.9, 10 mM $MgCl_2$, 50 mM NaCl, 5 mM DTT, 100 µg/ml BSA and 200 µM each dNTP. Reactions were stopped by addition of Novex TBE-urea sample buffer (Invitrogen) to 1X concentration and incubation at 70ºC for 3 min. Products were analysed by electrophoresis on 15% Novex TBE-Urea polyacrylamide gels (Invitrogen), and fluorescein signal detected using a Typhoon 8610 imager.

2.2.5.5 Reversion assay

This assay was based on an assay described by Osheroff and colleagues (Osheroff *et al*, 1999). M13Stop was incubated with Nb.*Bbv*CI for 2 h at 37ºC to generate the nicked template, checked for cutting by agarose gel electrophoresis, then the enzyme and buffer were removed using a Qiaquick PCR purification kit (Qiagen). The reaction conditions for the DNA synthesis reaction were 50 mM NaCl, 10 mM Tris-HCl, pH 7.9, 10 mM $MgCl_2$, 1 mM DTT, 200 µg/ml BSA, approximately 400 ng of polymerase and 250 ng of nicked template, in a total volume of 20 µl. Reactions were incubated for 3 h at 37ºC, then stopped by the addition of EDTA to a final concentration of 15 mM. The reaction mix was diluted 1:20 in dH_2O and 1 µl of this was used to transform MAX Efficiency DH5αF′IQ as in the gap filling assay. Plates were incubated at 37ºC overnight then at 4ºC for 2-4 h to allow the blue colour of the plaques to develop, then the number of blue and white plaques on each plate was counted.

2.2.5.6 Tet repressor escape assay

The MG1655 *cI*-Tet cell line (full genotype: MG1655 attλ::*cI* (Ind$^-$) λpR *tetA* Δ*ara*::FRT Δ*metRE*::FRT; (Bjedov *et al*, 2007)) was provided by Dr Ivan Matic (Pasteur Institute). Competent cells prepared from the cell line were transformed with pMAL-K10 plasmid. A glycerol stock of transformed cells was used to inoculate 100 ml LB with 100 µg/ml ampicillin, and this was incubated at 37°C with shaking for 4 h. The density of the culture was then estimated using a spectrophotometer and the volume corresponding (very approximately) to 100 cells was used to inoculate 10 ml LB containing 100 µg/ml ampicillin and various concentrations of IPTG. The new cultures were incubated overnight at 37°C with shaking. An aliquot of 500 µl was removed for analysis by SDS-PAGE, centrifuged at 17900×*g*, resuspended in 100 µl loading buffer, and approximately 1×10^8 cells were analysed by SDS-PAGE. The overnight cultures were serially diluted in LB, and 200 µl aliquots of 1:10, 1:100, 1:1000, 1:10^4, 1:10^5 and 1:10^6 dilutions and 20 µl of 1:10^6 dilution were spread on LB plates containing either 100 µg/ml ampicillin and 12.5 µg/ml tetracycline or only ampicillin, to allow estimation of the total number of cells plated. Plates were incubated for 24 h at 37°C then the resulting colonies counted for each plate.

2.2.6 Microscopy

2.2.6.1 Preparation of slides and fixation

BSF parasites were grown to a density of between 5×10^5 and 1×10^6 cells/ml and were harvested by centrifugation at 460×*g* for 10 min. Cell pellets were washed twice with 1 ml cfPBS then resuspended in a small volume of cfPBS to an approximate density of 2×10^7 cells/ml. For live cell microscopy with parasites expressing enhanced green fluorescent protein (eGFP)-labelled protein,10 µl resuspended cells were pipetted onto a clean glass slide, then a cover slip was placed on top and the cells used immediately for microscopy.

For methanol fixation, one spot of 5 µl resuspended cells was spread onto a clean glass slide, then the density of parasites in the spot was checked with a low-power light microscope (Leitz) to ensure the cells were neither too dense nor too sparse. If either was the case, the volume of resuspended cells was adjusted by 1 µl down or up respectively for a further one or two spots. The slides were allowed to air dry in a laminar flow hood for at least 15 min. Dried slides were immersed in methanol for 20 min, then washed in cfPBS twice for 5 min.

For formaldehyde fixation, an equal volume of 4% formaldehyde in cfPBS was added to the resuspended parasites. The parasites were incubated for 10 min, then washed twice with PBS and resuspended to a density of approximately 1×10^7 cells/ml. Next, 100 µl of resuspended cells were spotted on to a silanised glass slide and allowed to settle by gravity for 15 min. Slides were then washed four times in cfPBS for 5 min. Fixed parasites on slides were permeabilised by placing 1 ml 0.2% NP40 in cfPBS on to the slide and incubating for 5 min at 25ºC, then rinsing the slide three times in cfPBS. Slides were air-dried and then proceeded to antibody incubation. For refixation after incubation with antibody (see below), 100 µl 3% formaldehyde in cfPBS was pipetted onto each cell spot and incubated for 10 min. Slides were rinsed in cfPBS then washed three times for 5 min in cfPBS and air-dried.

2.2.6.2 Immunofluorescence (IF): blocking and antibody hybridisation

Slides with fixed parasites were incubated in blocking buffer (10% FCS, cfPBS) for 1 h at 4ºC. A primary antibody diluted in blocking buffer was spotted onto the slides and these were incubated for 1 h. Slides were washed twice for 5 min in blocking buffer, then incubated with secondary antibody diluted in blocking buffer for 1 h in the dark. For MYC-tagged protein, the primary antibody was mouse monoclonal anti-MYC antibody (Sigma), diluted 1:7000 and the secondary antibody was Alexa594-conjugated anti-mouse antibody (Invitrogen), diluted 1:5000; for eGFP-tagged protein, the primary antibody was rabbit polyclonal anti-GFP (Abcam), used at a dilution of 1:1000, and the secondary antibody was Alexa488-conjugated anti-rabbit antibody (Invitrogen), diluted 1:5000. For IF alone, after incubation with secondary antibody, slides were washed for 5 min in blocking buffer, blotted dry, then a drop of Vectashield mounting medium with DAPI (Vector Laboratories), and a coverslip was placed on top and sealed with nail varnish. For IF with fluorescence *in-situ* hybridisation (FISH), slides were either used directly for FISH (see below) or re-fixed before proceeding to FISH.

2.2.6.3 FISH

For FISH, 25 µl plastic frames (Geneframe, Thermo scientific) were applied to slides with fixed parasites, each one surrounding a spot of cells. A volume of 15 µl peptide nucleic acid (PNA) probe complementary to telomere repeats and conjugated to Cy5 (Cambridge Research Biochemicals), at a concentration of 800 ng/µl in hybridisation buffer (10 mM $NaHPO_4$, 10 mM NaCl, 20 mM Tris, pH 7.5, 70% formamide) was applied to each parasite spot and the frames sealed with coverslips, after which the slides were kept in the dark. The slides with probe were incubated on an *in-situ* block (Eppendorf) at 85°C for 5 min (probe and cell DNA denaturation) followed by slow cooling to 25°C for 2 h (hybridisation). The coverslips and frames were then removed and the slides were washed in 0.1% Tween-20 in filtered PBS for 20 min at 50°C in a hybridisation oven, followed by a wash of 1 min in 2X SSC at room temperature on a rolling shaker. The slides were then air-dried, mounted with Vectashield mounting medium with DAPI, covered with a coverslip and sealed with nail varnish.

2.2.6.4 Protocols and visualisation

Several combinations of the procedures described below were used. Fluorescence from eGFP-tagged parasites was visualised directly by live cell microscopy using an Axioskop 2 fluorescence microscope (Zeiss) or from methanol-fixed parasites, using either an Axioskop 2 or DeltaVision (Applied Precision) fluorescence microscope. In the latter case, the images were deconvoluted using Softworx software. Both eGFP and 12MYC tags were visualised by IF, with or without FISH, and imaged using an Axioskop 2.

2.2.7 Details of oligonucleotides

Name	Sequence (5′ to 3′)	Annealing temp (˚C)	To amplify
DLoopF	CTTCTCATCTCTCGGTCGTGACTGGGAAAACAAGTGGTCAGTGGT	NA	NA
LEWSEQ1	CAGTGATAGAGATCCCTGAG	50	pB2X-GFP
LEWSEQ2	CACCCTTAATGAGCGCTTTCG	50	pB2X-GFP
M13 primer	GTTTTCCCAGTCACGAC	NA	NA
M13 rev (-29)	CAGGAAACAGCTATGACC	seq	various
M13 uni (-21)	TGTAAAACGACGGCCAGT	seq	various
P73	TGTCCCATCATCGACAGAGC	50	K10
P74	AGCCCCCTCACAATTCTCGG	50	K001
P76	GCTCTGTCGATGATGGGACA	50	K10
P77	CCGAGAATTGTGAGGGGGCT	50	K001
P78	GGTCGTCAGACTGTCGATGAAGCC	seq	pMAL constructs
P79	ACGTTGTAAAACGACGGCCAGTG	52	pMAL constructs
P86	TGCCAAGCTTACTAGTAATTATGGTAAC	50	pMAL-K10 and -K001
P95	TCAGAATTCAAGCTTATGTGTGGGGGAG	50	pMAL-K10 and -K001
PL1	GATTACGAATTCGAGCTGAGGATCGGTACCCGGGGATCC	52	M13mp18 LacZα
PL2	TTTGAGAGATCTACAAAGGCTATC	52	M13mp18 LacZα
PL3	CGAGAGATGAGAAG	NA	NA
PL4	GATCTTATGCGAGAAACTAGTGAACA	65	12MYC
PL5	CAGTGACTCGAGGTCCGCGTGGATCCTCAC	65	12MYC
PL8	AAGCTTGATCTTATGCGAGAATCTAGAGAACA	65	12MYC
PL9	GATATCCTCGAGGTCCGCGTGGATCCTCAC	65	12MYC
PL12	CAGTGAAAGCTTGATATCATGTGTGGGGGAGAAGAG	50	K10 and K001
PL13	CAGTGATCTAGAAATTATGGTAACTTCCCTC	50	K10 and K001
PL15	CATCAGCTTTGGTTTGGCCGAACTGACCCTTGAGG	60	K10 mutagenesis
PL16	CCTCAAGGGTCAGTTCGGCCAAACCAAAGCTGATG	60	K10 mutagenesis
PL17	ATTACATAGTTGTTGGTTTGGCCGACTTTACTCTCGAAGTAAGTG	60	K001 mutagenesis
PL18	CACTTACTTCGAGAGTAAAGTCGGCCAAACCAACAACTATGTAAT	60	K001 mutagenesis
PL19	CATCAGCTTTGGTTTGGCC	52	mutant K10
PL21	CATAGTTGTTGGTTTGGCC	52	mutant K001
PL33	CTTTGTCTTTGCCTCTAAAAG	50	E2340_ORF312
PL34	ATGGCAAACTACAACACTCTTC	50	E2340_ORF312
PL35	GAACCAAGCAAGCCCCTCTTGC	50	E2340_ORF312

Table 2-2 – Details of oligonucleotides.
NA = not applicable (primer was for polymerase primer extension assay); seq = annealing temperature determined by sequencing provider; * = no product was obtained. Continued on next page.

Name	Sequence	Annealing temp (˚C)	To amplify
PL38	GAGCAGAAAACAGGCTGCGATG	55	E3_ORF250
PL39	TACTTTGGGTTGGTGCCGGCTG	55	E3_ORF250
PL40	GCGCAGAGGCCAGACCTCGATG	55	E3_ORF250
PL41	TATTAAAGACGCGGAATGATG	52	E3_NTD0250
PL42	ATGAGCTTCTATATCCTTGATG	52	E3_NTD0250
PL43	ATGAAAGCGCTTAGACAAGG	50	E3_NTD0866
PL44	CCTTTCTCGTTGTTTCGATTG	50	E3_NTD0866
PL45	CCTTTCTCATTGCTTCGATTG	50	E3_NTD0866
PL46	ATAAAAATGAAGAATAAACTAG	50	E3_ORF328
PL47	TTCTTGTTCCTGTTTCTGGAC	50	E3_ORF328
PL48	ATGGAACACAAAATTCTGC	52	E3_ORF189
PL50	CATCAGTTCCCTCTTTGTTGC	52	E3_ORF189
PL51	ACAAGAGGGGTGGAGGGC	50	E2340_NTD1110
PL52	GTTCAAAAGGCTTTGAATTATC	50	E2340_NTD1110
PL53	ATGAAAAGAGGTTTCGCCGT	50	E2340_NTD1104
PL54	GTTGGCTGCTGTAGGCGATG	50	E2340_NTD1104
PL55	ATGGCGGTAACTTTTTTTATAGC	50*	E2340_NTD1648
PL56	CCGGTGCCGACGCTATAAAAC	50*	E2340_NTD1648
PL57	ATGATTACCGGTAACGTCATA	50	E2340_NTD1861
PL58	GGTTGAGGCCTTGGAGTTCTG	50	E2340_NTD1861
PL62	CAGATGAGGGTCAGTTTATTC	55	E2340_NTD1171
PL63	TGGCGTTGTTGTTACTTGTG	55	E2340_NTD1171
PL64	ATGATAAAGAAACCGCGGTTTTG	55	E2340_NTD0947
PL65	GGCGGTCTTTTTGCGATCGTAG	55	E2340_NTD0947
PL66	AGCATGACGCAGCAAGCGGTC	50	E3_NTD0529
PL67	TATTTCTGCTTTGTAGAGCTC	50	E3_NTD0529
PL68	AGCAGGTCAAAGAACACGACC	50	E3_NTD0529
PL69	ACAATGATCACAAGAAATAGC	52	E2340_NTD0451
PL70	TTGGCCTGTAGTTGGTGCTGG	52	E2340_NTD0451
PL71	GTCTGCAATACGCTATTGGC	50*	E3_NTD0257
PL72	CAGCTACCGTTCACGATATT	50*	E3_NTD0257
PL73	ATGCACAGGACGCTTCTTAAG	50	E2340_NTD0909
PL74	ATCTCCAATCCAAAGAAAAGCC	50	E2340_NTD0909
T7	TAATACGACTCACTATAGGG	seq	various
T7term	CTAGTTATTGCTCAGCGGT	seq	various

Table 2-2 continued.

Chapter 3: Assembly and global comparison of EATRO 3 and EATRO 2340 genomes

3 Assembly and global comparison of EATRO 3 and EATRO 2340 genomes

3.1 Introduction

Previous studies of the evolution of subtelomeres and subtelomeric genes have been based on a single, haploid, assembled genome and have studied the differences between related genes and chromosome segments within that genome to deduce changes that have occurred (Marcello & Barry, 2007b; Linardopoulou *et al*, 2005). This type of approach can provide valuable information. However, it is limited by several factors: it relies on duplications within the genome; it is very difficult to estimate the divergence time of duplicated genes; and the divergence time between genes studied may be sufficiently great that separating two overlaid mutational events is very challenging. This project, therefore, aimed to study changes in the subtelomeres by sequencing two time-separated isolates of the same trypanosome strain, effectively sequencing the same genome twice, but with sufficient time between sampling for changes to occur. After assembly, the goal was to find novel and previously identified subtelomeric genes, and to identify the same gene in both genomes, so that the changes in these genes could be analysed. Identifying the changes would then allow patterns of mutation at the levels of individual genes and of the whole genome to be considered. As described in the Introduction, the subtelomeric antigen gene family of *VSG*s is enormous, highly variable and crucial to the survival of the parasite in its mammalian host, and I therefore focused on *VSG* genes in this study.

The aims of the work described in this chapter were to generate draft-quality genome assemblies for the two *T. b. rhodesiense* isolates, to assess the quality of these assemblies, and to determine the overall mutation rates in the chromosome cores and in the subtelomeres. The overall goals were to provide sequence data from which *VSG* genes could be annotated and compared between isolates; and to ensure that this analysis would be likely to provide insights into the mutational processes in subtelomeres by checking that they had, as hypothesised, a higher rate of mutation than the chromosome cores.

3.1.1 Isolates used in study

The two isolates used in the study were from a *T. b. rhodesiense* sleeping sickness focus in south-east Uganda, close to Lake Victoria, which appeared in 1940 (Robertson & Baker, 1958). Samples were taken from this disease focus every few years, and in this longitudinal set of samples we can follow the evolution of the causative strain. EATRO 3 (East African Trypanosomiasis Research Organisation isolate 3) was isolated in 1960 from flies or a fly in Busoga (Lumsden & Herbert, 1975), and EATRO 2340 was isolated in 1977 from a human in Samia (stock F in Barry *et al*, 1983; J.D. Barry, pers. comm.). The two isolates have been shown serologically to share a VSG set (Barry *et al*, 1983), and they belong to a set of isolates that have been indicated by population genetic analysis to have propagated clonally in the field (Tait *et al*, 1985; Hide *et al*, 1991). The clonal relationship has been confirmed by microsatellite and minisatellite genotyping with Genescan (L. Morrison and A. MacLeod, unpublished): out of six markers examined, five (JS2, 18, 5L5, 5 and m12) were identical between the two isolates, and for the other marker (PLC), one allele was different. Four other isolates from the same focus were examined in the same set of experiments, and all of these isolates had an identical genotype to EATRO 3, indicating that all six isolates are from the same clonal lineage. Details of the serology and marker genotypes of the isolates are given in Appendix 1. The serological and genotypic similarity of the EATRO 3 and EATRO 2340 isolates makes them ideal for the purpose of following *VSG* evolution, as we are likely to be able to identify the same *VSG* genes in both genomes.

Our knowledge of the relationship between the strains means that the approximate number of generations separating them (in the sense of cell divisions rather than complete life cycles) can be estimated (Figure 3-1). These estimates were obtained by assuming a doubling time of approximately 6 h through most of the life cycle; although this assumption ignores the time spent in non-proliferative life-cycle stages, it should provide a reasonable figure on which to base a conservative estimate of mutation rate (Turner *et al*, 1995; van den Abbeele *et al*, 1999; Salmon *et al*, 2005; Koffi *et al*, 2009).

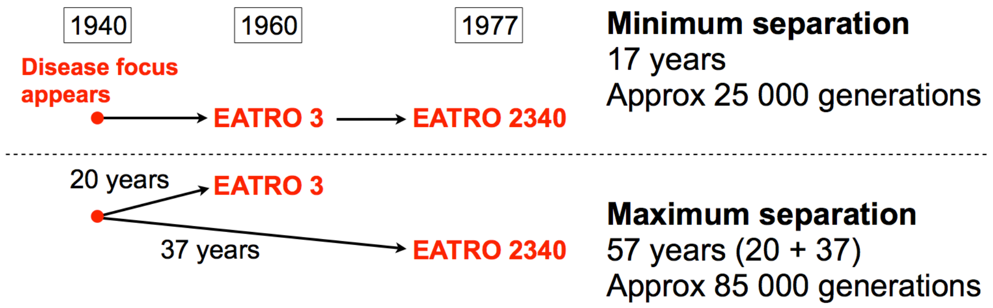


Figure 3-1 – Range of possible relationships between isolates EATRO 3 and EATRO 2340. In both cases, the origin of the disease focus is shown in 1940. In the upper panel, the relationship is shown where EATRO 2340 is the direct clonal descendant of EATRO 3, and hence the separation between the two is the time interval from 1960 to 1977 (17 years). In the lower panel, the maximum expected separation is shown, where the EATRO 3 and EATRO 2340 lineages diverged at the beginning of the disease outbreak, and hence differences have accumulated in the 20 years from 1940 to 1960 and the 37 years from 1940 to 1977.

The sizes of chromosomes 1 to 6 of EATRO 2340 have been analysed by pulsed-field gel electrophoresis, as have those from EATRO 795, a 1964 isolate from the same disease focus and isolate sequence, and from TREU 927, the strain used for genome sequencing (Melville *et al*, 1998; and S. Melville, unpublished). The sizes estimated for the TREU 927 chromosomes in the pulsed-field gel experiment corresponded well with the sizes found when its genome was sequenced (Berriman *et al*, 2005). As mentioned above, TREU 927 has relatively short subtelomeres, and both EATRO 2340 and EATRO 795 were shown to have several Mb more sequence in their subtelomeres, despite only the six smallest chromosome pairs being considered. In the chromosomes examined, these isolates had a total of 11.12 Mb and 8.96 Mb respectively of subtelomeric sequence, compared with 4.83 Mb in the same subtelomeres of TREU 927. The difference in subtelomere size even between EATRO 2340 and EATRO 795 demonstrates the dynamism of these genome regions, which evidently can change size dramatically even as the repertoire of expressed *VSG*s remains substantially the same, as judged from serological analyses (Barry *et al*, 1983).

3.1.1.1 Note on subspecies classification

There has been some discussion over the classification of EATRO 2340 as *T. b. rhodesiense*, due to an apparent lack of the *rhodesiense*-defining gene *SRA* in a derivative of the isolate used to study BESs (Young *et al*, 2008b). However, since the original isolate was from a human, we consider that EATRO 2340 should not be considered as *T. b. brucei*; and the classification of the disease focus as caused by *T. b. rhodesiense* is well-established. Further, microsatellite analysis has confirmed that the parasite line used in the Young *et al* study, which was considerably more passages removed from the original isolate than the EATRO 2340 cell line used in this study, differs from other EATRO 2340 lines, and thus it is most likely that the Young *et al* strain is not EATRO 2340 (A. MacLeod, M. Turner, G. Rudenko and K. Matthews, pers. comm.). The EATRO 2340 used in this work was the original line cloned from the EATRO 2340 stabilate in our laboratory (Barry *et al*, 1983) and, having been derived from a human, is *T. b. rhodesiense*. The EATRO 3 clone used here is infective to humans (Herbert *et al*, 1980) and is the strain from which *SRA* was initially characterised (De Greef & Hamers, 1994), so it certainly is *T. b. rhodesiense*. The presence of *SRA* in the genome is examined in section 3.2.4.4.

3.2 Assembly of genomes

3.2.1 Assembly strategy

The work in this section was performed in collaboration with Dr Thomas Otto (Sanger Institute), who carried out or directed the majority of the assembly steps; Dr Matt Berriman (Sanger Institute), who provided Illumina and 454 sequencing data; and Prof Ed Louis (University of Nottingham), who provided 454 sequencing data. I began to work on the project after the sequence data had been generated, and worked on the assembly steps under the direction of Dr Otto. After the steps detailed in this section I took over full control of the data and analysis. A glossary of the software used in assembly is given in Table 2-1 (which also includes details of other programs used in the project). DNA was prepared from BSF EATRO 3 and a culture of procyclic form EATRO 2340 parasites. Sequence data were obtained from DNA of both isolates by Illumina sequencing and 454 sequencing (Table 3-1 , read sets 1-14). Separate *de novo* assemblies were attempted with the Newbler assembler from each isolate’s 454 data sets. The EATRO 2340 assembly was the more successful, and to improve this assembly a further two 454 data sets were generated (Table 3-1, read sets 15 and 16). A new set of *de novo* contiguous DNA sequences (contigs) was generated from the five EATRO 2340 454 data sets using Celera Assembler, a newer program. To produce assembled chromosomes, the EATRO 2340 *de novo* contigs were aligned to a reference genome, *T. b. gambiense* strain DAL 972 (Jackson *et al*, 2010), using ABACAS. This reference genome was chosen because the assembled chromosomes contain very little subtelomeric sequence. Because subtelomeric sequence changes rapidly at both small and larger scales, these regions vary considerably between parasite strains (Melville *et al*, 2000), so it was not appropriate to use the subtelomeres of another parasite strain as a template for contig ordering. After ordering, the three 8 kb 454 libraries were mapped to the new assembly, and a plot was generated of the coverage of mate pairs that were mapped at the correct distance apart. Contigs were broken at points where the paired read coverage was below three, *i.e.* at likely misassembled points. Finally, gaps in the assembly were filled, as far as possible, by targeted reassembly with IMAGE in contigs longer than 15 kb, which closed 1210 gaps.

Set	Isolate	Technology	Type	Insert/ fragment size	No. reads	Total read length (bp)	Mean read length (bp)
1*	EATRO 2340	454	PEL	2 kb	347728	63397799	182
2*	EATRO 2340	454	PEL	2 kb	896497	168031785	187
3*	EATRO 2340	454	PEL	3 kb	983398	201707497	205
4*	EATRO 2340	454	PEL	8 kb	1146306	189307841	165
5*	EATRO 2340	Illumina	PEP	300 bp	48192764	3662650064	76
6	EATRO 2340	Illumina	PEL	3 kb	11028950	838200200	76
7	EATRO 2340	Illumina	PEL	3 kb	18420270	1399940520	76
8	EATRO 3	454	PEL	2.5 kb	729630	172417460	236
9	EATRO 3	454	PEL	3 kb	589856	154717106	262
10	EATRO 3	454	R	-	515398	164008964	318
11	EATRO 3	454	R	-	512420	178841313	349
12	EATRO 3	454	PEL	3 kb	746428	261645469	351
13*	EATRO 3	Illumina	PEP	300 bp	37446496	2845933696	76
14*	EATRO 3	Illumina	PEL	3 kb	23514048	1787067648	76
15*	EATRO 2340	454	PEL	8 kb	1138453	241502445	212
16*	EATRO 2340	454	PEL	8 kb	1223578	225366130	184

Table 3-1 – Sequencing data sets generated for the project.
*** denotes sequence sets that were used in the final draft assemblies. PEL = paired-end library; PEP = paired-end PCR free sequencing; R = random sequencing.**

As the two genomes were expected to be very similar, the EATRO 2340 assembly was used as a reference to produce an EATRO 3 genome assembly. Illumina reads generated from PCR-free and 3 kb libraries of EATRO 3 DNA were trimmed to remove poor-quality bases at the ends of reads, using SGA. The trimmed reads were then used for assembly with Velvet Columbus. This program initially used the EATRO 2340 assembly (chromosomes and unmapped contigs) as a guide for assembly, then did *de novo* assembly to generate contigs from reads that could not be mapped to the reference. In the part of the EATRO 3 assembly that had been assembled to EATRO 2340, any position that was not covered by at least two reads was considered to be poorly supported and the base called was changed to an N to indicate ambiguity. The resulting contigs were ordered using ABACAS, to the EATRO 2340 chromosomes[1]. Finally, IMAGE was used to fill, as far as possible, gaps in contigs longer than 15 kb, resulting in the closure of 5751 gaps. The assembly strategy is summarised in Figure 3-2.

[1] In the actual ABACAS run, the EATRO 2340 full assembly was used as a reference; however, mappings to EATRO 2340 bin contigs were then ignored.

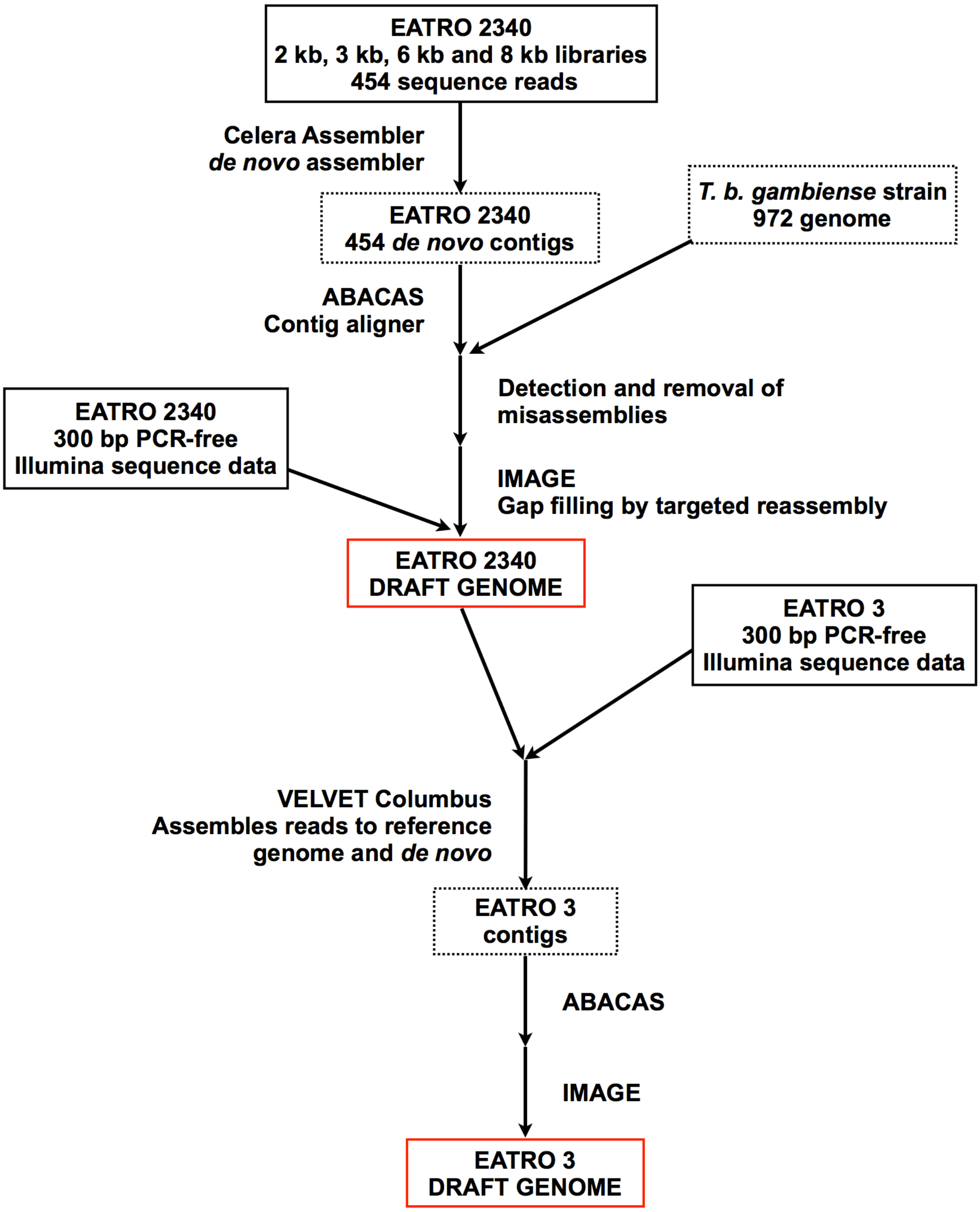


Figure 3-2 – Strategy used to generate EATRO 3 and EATRO 2340 genome assemblies. Software used at each step is described beside the arrows representing the appropriate step(s), see also Table 2-1. Red boxes show draft genomes used in the further analyses described below.

3.2.2 Summary statistics of assembled genomes

The assembly steps produced an EATRO 2340 assembly consisting of 11 chromosomes and 831 unmapped ('bin') contigs, and an EATRO 3 assembly of 11 chromosomes and 5122 bin contigs (Table 3-2). These assemblies both contained a large number of very short contigs. To simplify downstream analyses, I excluded contigs smaller than 1000 bases (excluding gaps), producing the full draft genome versions for EATRO 2340 and EATRO 3.

Assembly	No. contigs	Total length (bp)	N50 (bp)	N90 (bp)	No. gaps
EATRO 2340 all contigs and chromosomes	842	38008219	1247284	30262	4178
EATRO 2340 chromosomes	11	22490340	2065168	1131951	1130
EATRO 2340 contigs ≥ 1 kb	550	15342690	71827	19729	3045
EATRO 2340 final draft assembly (chromosomes plus contigs ≥ 1 kb)	561	37833030	1247284	31106	4175
EATRO 3 all contigs and chromosomes	5133	36968657	1171069	3636	13944
EATRO 3 chromosomes	11	20510521	1917240	1171069	3359
EATRO 3 contigs ≥ 1 kb	1702	14774087	31683	2706	9438
EATRO 3 final draft assembly (chromosomes plus contigs ≥ 1 kb)	1713	35284608	1278464	9274	12797

Table 3-2 – Summary statistics of the assemblies generated for EATRO 3 and EATRO 2340. 90% of sequence is in contigs of the N90 length or greater; similarly 50% of sequence is in contigs of the N50 length or greater.

3.2.3 Annotation of genomes

3.2.3.1 Annotation of boundaries of chromosome cores

Although the *T. brucei* reference genome (strain TREU 927) has been extensively annotated (Berriman *et al*, 2005; Marcello & Barry, 2007b; Kolev *et al*, 2010), there is no readily available description of the boundaries of the chromosome cores. Therefore, in order to define the chromosome cores and subtelomeres in EATRO 3 and EATRO 2340, I first annotated the core boundaries in TREU 927. I used TriTrypDB (Aslett *et al*, 2010) to compare each chromosome from TREU 927 and *T. b. gambiense* strain DAL 972 (Jackson *et al*, 2010), and define the core boundaries by the first and the last genes that 1) were syntenic between the two genomes and 2) were neither typical subtelomeric genes such as *VSG* genes, expression-site associated genes and retrotransposon hot spot protein genes, nor unlikely hypothetical genes (Table 3-3). This annotation used the most up-to-date versions of the TREU 927 chromosomes available on TriTrypDB at the time: version five of chromosome 10, and version four of all other chromosomes.

To find the core boundaries in EATRO 3 and EATRO 2340, I aligned each chromosome to its homologue in TREU 927 using BLAST, and viewed alignments of at least 90% identity using ACT. Using these alignments, I defined the chromosome cores by identifying the EATRO 3 or EATRO 2340 copies of the genes defining the chromosome cores in TREU 927. If the gene could not be found, or was poorly assembled, the core boundary was taken to be the point where the EATRO chromosome started to have high (at least 90%) identity to TREU 927. The lengths of chromosome cores and subtelomeres defined by this annotation process are shown in Table 3-3.

3.2.3.2 Automated transfer of core annotations

To annotate individual genes in the chromosome cores, I used RATT, which uses the MUMmer alignment algorithm to identify regions of synteny and automatically transfer annotations between a reference and a related genome. The reference genome was the TREU 927 chromosome cores only. Of 8508 coding sequences in TREU 927 cores, 6253 were transferred to both EATRO 3 and EATRO 2340; a further 1408 were transferred to EATRO 2340 but not to EATRO 3, and 337 were transferred to EATRO 3 but not to EATRO 2340.

3.2.4 Assessment of quality of genomes

3.2.4.1 Length of sequence and coverage of reference genome

The assemblies contained 19.11 Mb of EATRO 3 and 20.44 Mb of EATRO 2340 core sequence, which is 88% and 94% respectively of the TREU 927 core sequence. To check how similar the EATRO assemblies were to TREU 927, I used NUCmer from the MUMmer suite of programs to align the chromosomes of each isolate with the TREU 927 core sequences. When a 90% identity cut-off was used, 82% of the TREU 927 core sequence was contained in alignments to the EATRO 3 chromosomes, and 92% to the EATRO 2340 chromosomes. These figures imply that most of the core sequence that could be assembled in the EATRO isolates had high identity to the reference genome.

Chromosome	Subtelomere before core			Core			Subtelomere after core			Total subtelomere length			Total chromosome length		
	TREU 927	EATRO 3	EATRO 2340	TREU 927	EATRO 3	EATRO 2340	TREU 927	EATRO 3	EATRO 2340	TREU 927	EATRO 3	EATRO 2340	TREU 927	EATRO 3	EATRO 2340
1	203195	3094	6112	780753	648501	689842	80724	38451	65041	283919	41545	71153	1064672	690046	760995
2	306670	4985	8793	854238	707850	739447	33040	78140	40380	339710	83125	49173	1193948	790975	788620
3	147063	421840	475047	1434810	1195895	1480428	71352	421840	12926	218415	843680	487973	1653225	2039575	1968401
4	80880	0	9941	1384257	1387424	1296285	125295	63819	109725	206175	63819	119666	1590432	1451243	1415951
5	162089	79636	91706	1168567	1064568	1115794	277542	26865	39784	439631	106501	131490	1608198	1171069	1247284
6	175684	2177	5923	1237849	1268560	1110193	205382	7727	15835	381066	9904	21758	1618915	1278464	1131951
7	27071	38082	125312	2115307	1843705	1897202	62855	35453	42654	89926	73535	167966	2205233	1917240	2065168
8	135692	9414	38411	2345496	1993350	1934450	2	6078	15943	135694	15492	54354	2481190	2008842	1988804
9	326350	243500	479727	2155682	1518685	1799856	575515	146950	220695	901865	390450	700422	3057547	1909135	2500278
10	68602	34084	39131	3905553	3516335	3805074	170220	15215	10545	238822	49299	49676	4144375	3565634	3854750
11	31836	45203	32007	4450307	3965959	4574538	494939	86189	161593	526775	131392	193600	4977082	4097351	4768138
All	**1665132**	**882015**	**1312110**	**21832819**	**19110832**	**20443109**	**2096866**	**926727**	**735121**	**3761998**	**1808742**	**2047231**	**25594817**	**20919574**	**22490340**

Table 3-3 – Core and subtelomere lengths in bp for the three genomes annotated.

The assembled chromosomes contained 1.81 Mb of subtelomeric sequence in EATRO 3 and 2.05 Mb in EATRO 2340, compared with 3.76 Mb of subtelomeres in TREU 927. However, due to the contigs being aligned to strain DAL 972, which has minimal subtelomeric sequence, this low coverage was expected. Much of the subtelomeric sequence was likely to be on contigs that were not mapped on to reference chromosomes. Such sequence totalled 14.77 Mb in EATRO 3 and 15.34 Mb in EATRO 2340. Some of these ‘bin’ contigs probably corresponded to the homologous partners of sequence already assembled into chromosomes, where the two homologues were different enough to be seen as two separate sequences by the assembly algorithm, but this possibility was not examined in detail.

3.2.4.2 Read coverage

In order to assess the depth of coverage in the assemblies, I mapped each set of paired-end Illumina reads to the corresponding genome using BWA, using as a reference all the chromosomes and bin contigs available for each genome (Table 3-2, rows 1 and 5). I then calculated the mean coverage with good-quality sequence (Phred-scaled base quality score of at least 23) for all bases in the chromosomes and the final set of bin contigs over 1 kb (Table 3-2, rows 4 and 8). This self-mapping gave a mean coverage of 25-fold for EATRO 3 and 54-fold for EATRO 2340, with 92.4% and 97.3% of bases respectively having a coverage with good-quality sequence in at least eight reads, demonstrating that the genomes were well-supported. To call SNPs and heterozygosities (see below) I used more stringent coverage thresholds, requiring good-quality coverage between 10 and 200 inclusive. Using these thresholds, 89.1% of EATRO 3 and 96.9% of EATRO 2340 positions had coverage suitable for calling SNPs, so most of both the genomes could be examined for differences between the isolates. The difference in coverage between the two strains was partly due to the difference in the amount and quality of sequence obtained from each experiment: 24 million pairs were generated from the EATRO 2340 PCR-free sequencing experiment compared with 18 million pairs of reads from EATRO 3. Further, while 88% of forward reads and 71% of reverse reads from EATRO 2340 contained more than 80% good-quality bases, these figures were only 80% and 60% respectively for EATRO 3.

When I performed mappings of the EATRO 2340 reads to the EATRO 3 genome, and the EATRO 3 reads to the EATRO 2340 genome, the mean coverages were 46-fold and 28-fold respectively. These figures are somewhat counterintuitive when compared with the results for the self-mapping, but are explicable. The EATRO 2340 reads gave higher coverage of the EATRO 3 genome than did the EATRO 3 reads: this was likely due to the previously discussed better quality and number of EATRO 2340 reads compared with the EATRO 3 reads. The EATRO 3 reads mapped with higher coverage to the EATRO 2340 genome than they did to the EATRO 3 genome: this was probably due to the read coverage tending to drop off at the ends of contigs and towards gaps, because reads where only part of the sequence is present will usually not be aligned. The EATRO 3 genome was more fragmented than the EATRO 2340 genome, *i.e.* more sequence was in the bin contigs than for EATRO 2340; the bin sequence was spread over 1702 contigs with an N50 of 32 kb, compared with 550 contigs with an N50 of 72 kb; and there were nearly 13000 gaps compared with 4000 in EATRO 2340.

3.2.4.3 Coverage of previously sequenced telomeric and subtelomeric genes

The related strain EATRO 795 has been used extensively in work characterising various aspects of the biology of VSGs, as result of which the sequences of a number of telomere-proximal *VSG* genes from the EATRO lineage are available (Miller & Turner, 1981; Robinson *et al*, 1999). In order to assess the coverage of the *VSG* repertoire by the genome sequences, I used BLASTN to query the EATRO 3 and EATRO 2340 genomes with 16 ILTat genes. Of these 16 ILTat genes, good matches for nine were found in EATRO 3 and for eight in EATRO 2340 (Table 3-4). This represents a good level of coverage, given that in EATRO 795 some of these genes had only telomere-proximal copies (Robinson *et al*, 1999), which means that they were in one of the most difficult regions of the genome to assemble. It would be expected that EATRO 3 (isolated four years before EATRO 795) would share more of these genes with EATRO 795 than would EATRO 2340 (isolated 13 years later), and it is encouraging to see this reflected somewhat in the assembled genomes despite less EATRO 3 than EATRO 2340 sequence having been assembled.

Gene	Result in EATRO 3					Result in EATRO 2340					Location and copy no. in EATRO 795		
	Hit % ID	% query covered	Location	Gap?	Consider same gene?	Hit % ID	% query covered	Location	Gap?	Consider same gene?	T	I	MC
ILTat 1.1	96.9	66.5	bin contig	yes	-	90.2	99.1	chr 4	-	yes	nd		
ILTat 1.2	99.4	47.9	bin contig	yes	-	82.8	15.7	bin contig	-	-	nd		
ILTat 1.21	100	36	bin contig	yes	-	100	2.4	bin contig	-	-	1	0	1
ILTat 1.22	99.7	96.5	bin contig	-	yes	99.9	98.6	bin contig	-	yes	1	0	0
ILTat 1.23	95.9	99.5	bin contig	-	yes	95.8	99.5	bin contig	-	yes	2	0	1
ILTat 1.24	96.8	90	bin contig	-	yes	96.8	90	bin contig	-	yes	nd		
ILTat 1.25	92	83.4	chr 2	-	-	92.9	83.1	chr 2	-	-	2	0	1
ILTat 1.3	98.4	82.8	bin contig	-	-	94.7	83	bin contig	-	-	nd		
ILTat 1.4	96.5	100.3	bin contig	-	yes	96.6	100.1	bin contig	-	yes	nd		
ILTat 1.61	99.9	100	chr 11	yes	yes	99.8	35.5	chr 11	yes	-	nd		
ILTat 1.63	No hit	-	-	-	-	No hit	-	-	-	-	nd		
ILTat 1.64	99.8	100	chr 3	-	yes	99.8	100	chr 3	-	yes	nd		
ILTat 1.67	99.7	99	bin contig	-	yes	99.7	99	bin contig	-	yes	1	0	0
ILTat 1.68	99.8	92.3	bin contig	yes	yes	99.5	73.8	bin contig	-	-	1	1	1
ILTat 1.69	99.8	90.8	bin contig	yes	yes	79.8	81.6	bin contig	-	-	2	0	1
ILTat 1.71	No hit	-	-	-	-	100	100	bin contig	-	yes	1	2	1

Table 3-4 – Quality of matches to *VSG* genes previously characterised in the parasite lineage.

Table 3-4 continued. In the ‘Gap?’ column, a ‘yes’ indicates that the hit was interrupted or truncated by an assembly gap. Data on EATRO 795 are from Robinson *et al*, 1999. MC = minichromosomal, T = telomeric, I = internal (most likely in subtelomere arrays), nd = not done

3.2.4.4 *SRA*

As described above, the history of both EATRO 3 and EATRO 2340 predicts that they contain the *T. b. rhodesiense*-defining gene *SRA*. I therefore used BLASTN to query both genomes with four published sequences of *SRA*: Z37159, the original *SRA* sequence, from EATRO 3 (De Greef & Hamers, 1994); AF097331 from a Ugandan isolate (Milner & Hajduk, 1999); AJ345057 from a Kenyan isolate and AJ345058 from a Zambian isolate (Gibson *et al*, 2002). Despite their widely separated geographic origins, the minimum pairwise identity between these published *SRA* sequences is 97.8% (once the sequences have been truncated so that they all start and end at the same point as the shortest sequences, AJ345057 and AJ345058), reflecting the high level of conservation of *SRA*. In EATRO 2340, all four gave good hits to a sequence close to the end of a short (15 kb) bin contig. In the new genomes, finding *SRA* would likely present a challenge because the gene is in a BES in all trypanosomes examined, and hence in one of the most difficult regions to assemble. The EATRO 2340 hit was truncated by 97 bp at the 3′ end compared with the shortest published sequence, but had a minimum pairwise identity of 97.8% in the region that was present in all five sequences. EATRO 3 did not have such good hits, even to Z37159. I therefore checked whether the *SRA* homologue in EATRO 2340 was well-covered by the BWA-mapped EATRO 3 PCR-free sequence reads described in section 3.2.4.2. This proved to be the case, indicating that the EATRO 3 DNA sequence did indeed contain the *SRA* gene, as expected (Figure 3-3). There was only one likely EATRO 3-EATRO 2340 SNP in the *SRA* sequence (see section 3.4.1), indicating that the EATRO 3 *SRA* gene was also very highly conserved. Because no EATRO 3 *SRA* gene was assembled from this analysis, a comparison with the previously reported *SRA* from EATRO 3 (Z37159) was not possible, but the comparison with EATRO 2340 indicated that the two versions of the gene from EATRO 3 shared at least the 99% level of identity that was found between the assembled EATRO 2340 *SRA* and Z37159.

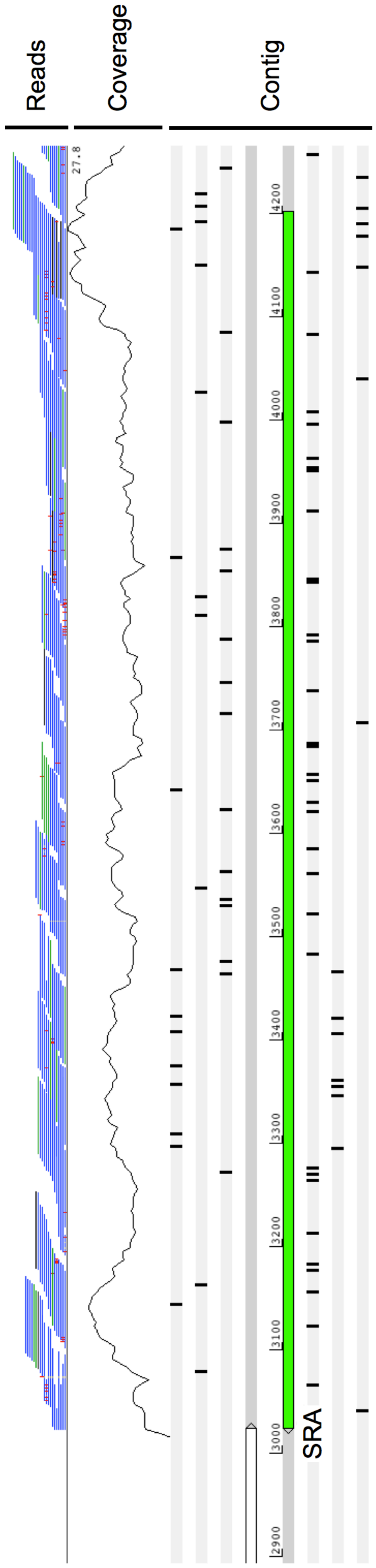


Figure 3-3 – EATRO 3 PCR-free sequence reads mapped with BWA to EATRO 2340 *SRA* homologue confirm the presence of *SRA* in the EATRO 3 genome sequenced.
BLAST queries with *SRA* genes showed a homologue in EATRO 2340 contig249, which contig is shown in part (bottom), visualised with Artemis, with the likely *SRA* gene shown by a green box. The scale is bp from the start of contig249. PCR-free reads from EATRO 3 mapped to this region are shown (top plot) as blue (forward orientation) or green (reverse) horizontal bars. The reads were mapped with good coverage (middle plot, reads per base in non-overlapping 8 bp windows, scale is from 0 to 27.8), and few SNPs (red strokes in top plot), indicating that *SRA* was present in the EATRO 3 genome sequenced.

3.3 Comparison of EATRO genomes and reference strains

3.3.1 Pairwise identities of coding sequences in EATRO 3, EATRO 2340 and *T. b. brucei* TREU 927

As described above, I used automatic annotation transfer to annotate coding sequences in EATRO 3 and EATRO 2340 chromosome cores. This approach yielded 6253 genes that had been annotated in both EATRO 3 and EATRO 2340. For each of these, I performed three pairwise alignments: EATRO 3 *vs* EATRO 2340, EATRO 3 *vs* TREU 927 and EATRO 2340 *vs* TREU 927 (Figure 3-4). In the EATRO 3 *vs* EATRO 2340 comparisons, around half of the coding genes examined (2982) were identical, and 96.7% had at least 99% identity. However, between TREU 927 and each EATRO genome, only around 5% were identical, and 86% (EATRO 3) and 87% (EATRO 2340) had at least 99% identity. This is consistent with EATRO 3 and EATRO 2340 being more closely related to each other than either is to TREU 927.

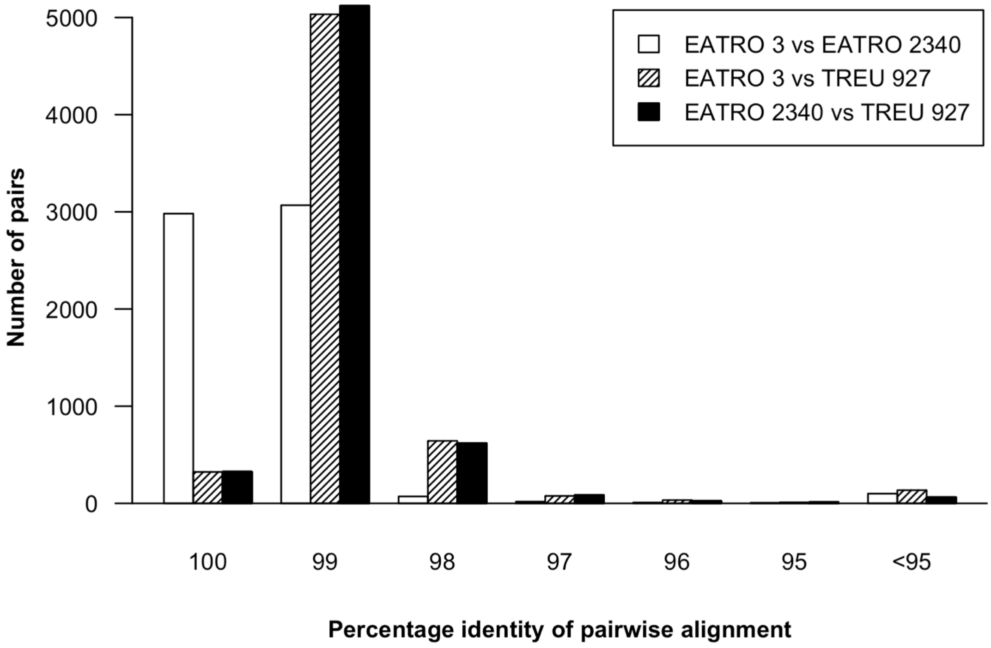


Figure 3-4 – Pairwise comparisons of all coding sequences annotated in EATRO 3, EATRO 2340 and TREU 927.
RATT was used to transfer annotations automatically from TREU 927 to EATRO 3 and EATRO 2340. Genes containing gaps and those less than 50% of the TREU 927 length were excluded, as were genes that had been transferred successfully to only one genome. Pairwise alignments were done for each gene for all three possible pairings, as shown in the key, using Clustalw1.8.

3.3.2 Comparison with other reference genomes

I decided to use the EATRO sequences to perform inference of the relationships between these isolates and other sequenced *T. brucei* strains, to test whether EATRO 3 and EATRO 2340 grouped closely, as expected. For this analysis, I used the publicly-available genomes of *T. b. brucei* strains TREU 927 (Berriman *et al*, 2005) and Lister 427 (downloaded from TriTrypDB), and *T. b. gambiense* strain DAL 972 (Jackson *et al*, 2010). My first strategy was to attempt to create alignments of genes that had been used previously to study phylogeny within *Trypanosoma* (Alvarez *et al*, 1996; Stevens & Gibson, 1999; Hamilton *et al*, 2004), but this approach was not successful. In some cases, for example phosphoglycerate kinase and various rRNA subunits, the genes were unusable because they were not properly assembled in one or both EATRO genomes, most likely because the genes are found in tandem arrays, which makes them difficult to assemble. In other cases, for example trypanothione reductase and heat-shock protein 70, there were only one or two differences within the five-genome alignment, which meant the alignments were unlikely to be reliable bases for trees.

Instead, I used a whole-genome comparison approach to generate a tree of the relationships between strains. This approach involved finding maximal unique matches (MUMs) between the concatenated genomes of each pair of strains using MUMmer, and estimating phylogenetic distances between two strains based on the fraction of each sequence that was covered by MUMs, scaled logarithmically. The analysis used 25.6 Mb of sequence in TREU 927, 25.6 Mb in Lister 427, 22.1 Mb in DAL 972, 21.5 Mb in EATRO 3 and 23.6 Mb in EATRO 2340. This approach generated a well-separated tree that was consistent with the expected relationships between the strains (Figure 3-5).

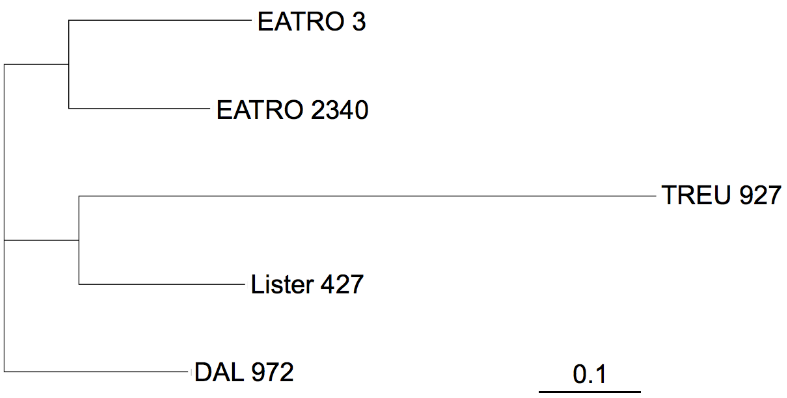


Figure 3-5 – Unrooted tree showing relationships between five *T. brucei* whole genomes using new EATRO 3 and EATRO 2340 genome assemblies.
The chromosomes were concatenated for each strain and MUMs for all ten possible pairings of the full genomes were found using MUMmer. A distance matrix was constructed by calculating, for each pair of sequences A and B, –log((fraction of genome A covered by MUMs to genome B) + (fraction of genome B covered by MUMs to genome A)/(length of A + length of B)). A neighbour-joining tree was generated using BIONJ and visualised in Dendroscope. The scale bar represents 0.1 distance units.

The strategy of using MUMs has several limitations, but is sufficient to give a general impression of the relationships. One limitation is caused by the fact that the genomes vary in length and in assembled subtelomere content, which means that the longest genome may have contained unmatched sequence that had simply not been assembled in the others; and that the genome with the highest assembled subtelomere content will have appeared to contain a higher proportion of sequence that is not found in the others. Both of these complications would have artificially inflated the distance of the genome in question from the others, and are probably at least partially responsible for the long branch length separating the best-assembled genome, TREU 927, from the others. A second caveat is that it is possible that the distance between EATRO 3 and EATRO 2340 was under-estimated somewhat, because EATRO 3 used EATRO 2340 as a reference for assembly. However, to produce any topology other than what is seen here, this bias would have to be considerably more than seems likely.

3.4 Comparison of EATRO 3 and EATRO 2340 genomes

3.4.1 Calling of mutations, and quality control

To look at mutations across the genome, I used EATRO 2340 as the reference sequence because this genome was better assembled than EATRO 3. I used BWA to map PCR-free, paired-end Illumina sequencing reads from both EATRO 2340 and EATRO 3 on to the EATRO 2340 genome assembly (including bin contigs), which gave mean coverages with good-quality reads of 54-fold and 28-fold respectively. I used the program SAMTools to locate single-nucleotide differences (either SNPs or indels) between the reference genome and the aligned reads, and to filter these differences to exclude any where the read coverage was unusually high (above 200), or too low to be confident of the call (below 10). For the EATRO 2340 to EATRO 2340 mapping, 96.9% of positions had appropriate coverage for SNP calling; this figure was 92.5% for the EATRO 3 to EATRO 2340 mapping. SAMTools identified 176334 'good quality' differences using the EATRO 3 reads and 184161 using the EATRO 2340 reads, of which 34982 and 41497 respectively were indels. I also used the information from SAMTools to classify substitutions into homozygous differences, where all (or almost all) reads differed from the reference, and likely heterozygosities, where the numbers of reads identical to the reference and reads that differed were consistent with two alleles being present. This analysis found that the majority of differences called as substitutions were likely to be due to heterozygosities: 137964 out of 141352 (97.6%) with EATRO 3 reads and 139234 out of 142664 (97.6%) for EATRO 2340.

I then developed a Perl script to identify differences that were present either when comparing the EATRO 2340 genome and the EATRO 3 reads, or when comparing the EATRO 2340 genome and the EATRO 2340 reads, but not in both. This step was to discard differences that were called either due to errors in the EATRO 2340 genome assembly, or due to heterozygosities that were present in both genomes. As a further check, if there was a SNP called in only one genome, the script also tested whether coverage in the other genome was sufficient for a SNP to have been called; if not, the SNP was discarded, as I could not be confident that there was not a SNP in both genomes.

Of the putative differences between EATRO 3 and EATRO 2340, 146433 were discarded due to there being a SNP in both isolates, and 11769 were discarded because coverage was insufficient in one or other genome for a SNP to be called. This analysis yielded 55860 likely EATRO 3-EATRO 2340 SNPs, of which 9596 were indels, 45608 were heterozygous substitutions (*i.e.* the genome with a new allele at the SNP position probably also contained the old allele) and 656 were homozygous substitutions.

I used this approach to generate a plot of differences for every EATRO 2340 chromosome and contig. The distributions of differences for each chromosome are shown in Figure 3-6, plotted on an ACT view showing hits of at least 90% identity in BLAST alignments of the EATRO 3 and EATRO 2340 chromosomes. From these plots it can be seen that, in general, there were many more differences between isolates in the subtelomeres than in the chromosome cores.

There were some regions within the cores that showed a greater difference between the isolates than would be expected. One of the most obvious of these is on chromosome 4, where there were many differences along the chromosome, approximately from the homologue of *Tb927.4.3760* to the end of the chromosome core after *Tb927.4.5380*. This region corresponds quite well with a region of the chromosome that is a segmental duplication shared with chromosome 8, between *Tb927.4.3860* and *Tb927.4.5390* (Jackson, 2007a). Although the corresponding region on chromosome 8 was assembled, it did not show a similarly high number of SNPs; however, it seems likely that the apparently anomalously high rate of mutation on chromosome four is due to misassemblies or misalignments caused by this duplication. This possibility is not of particular concern to the main analysis, which will concentrate on the genes in the subtelomeres.

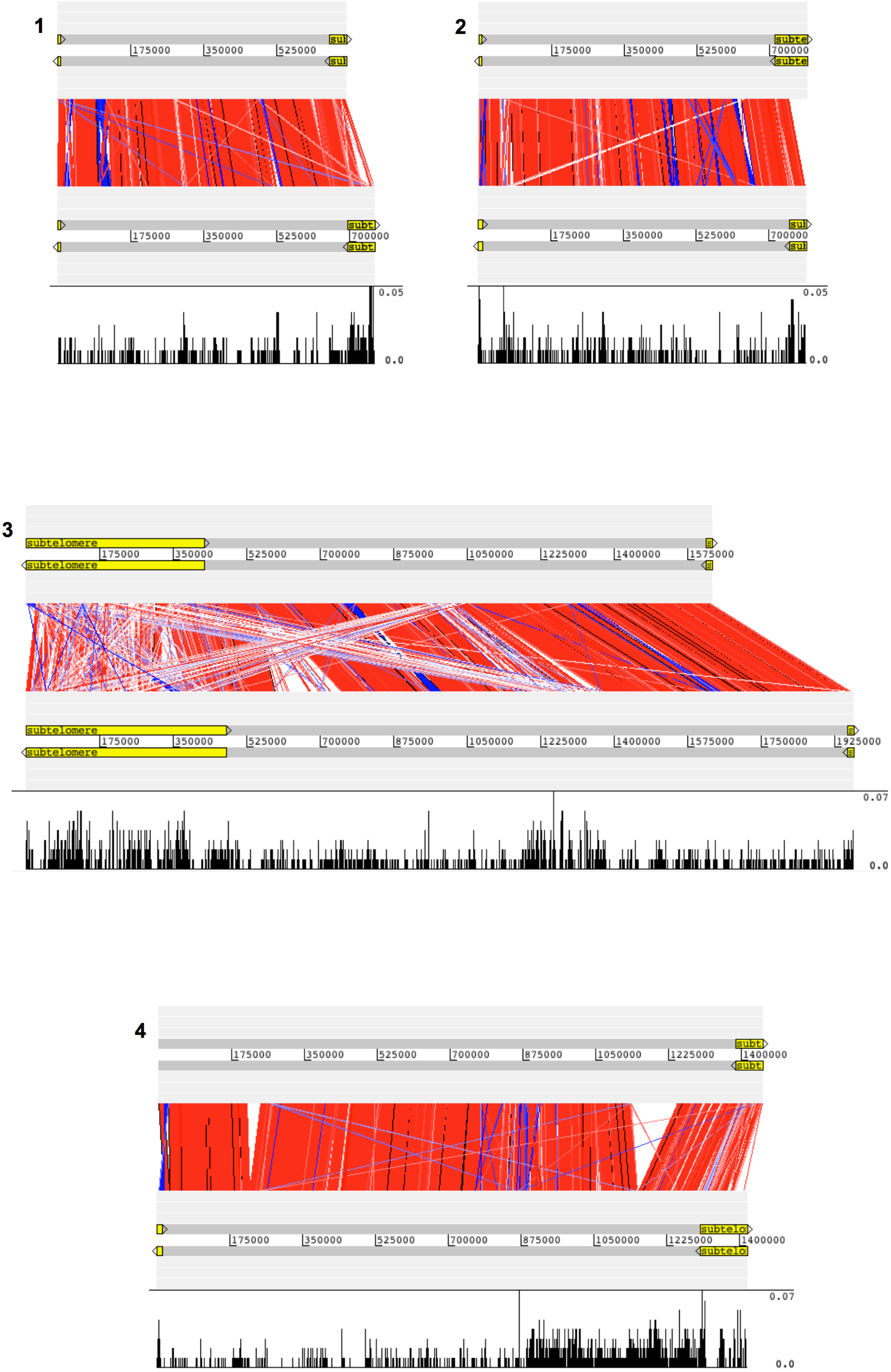


Figure 3-6 – Distribution of differences between EATRO 3 and EATRO 2340. Continued on next pages.

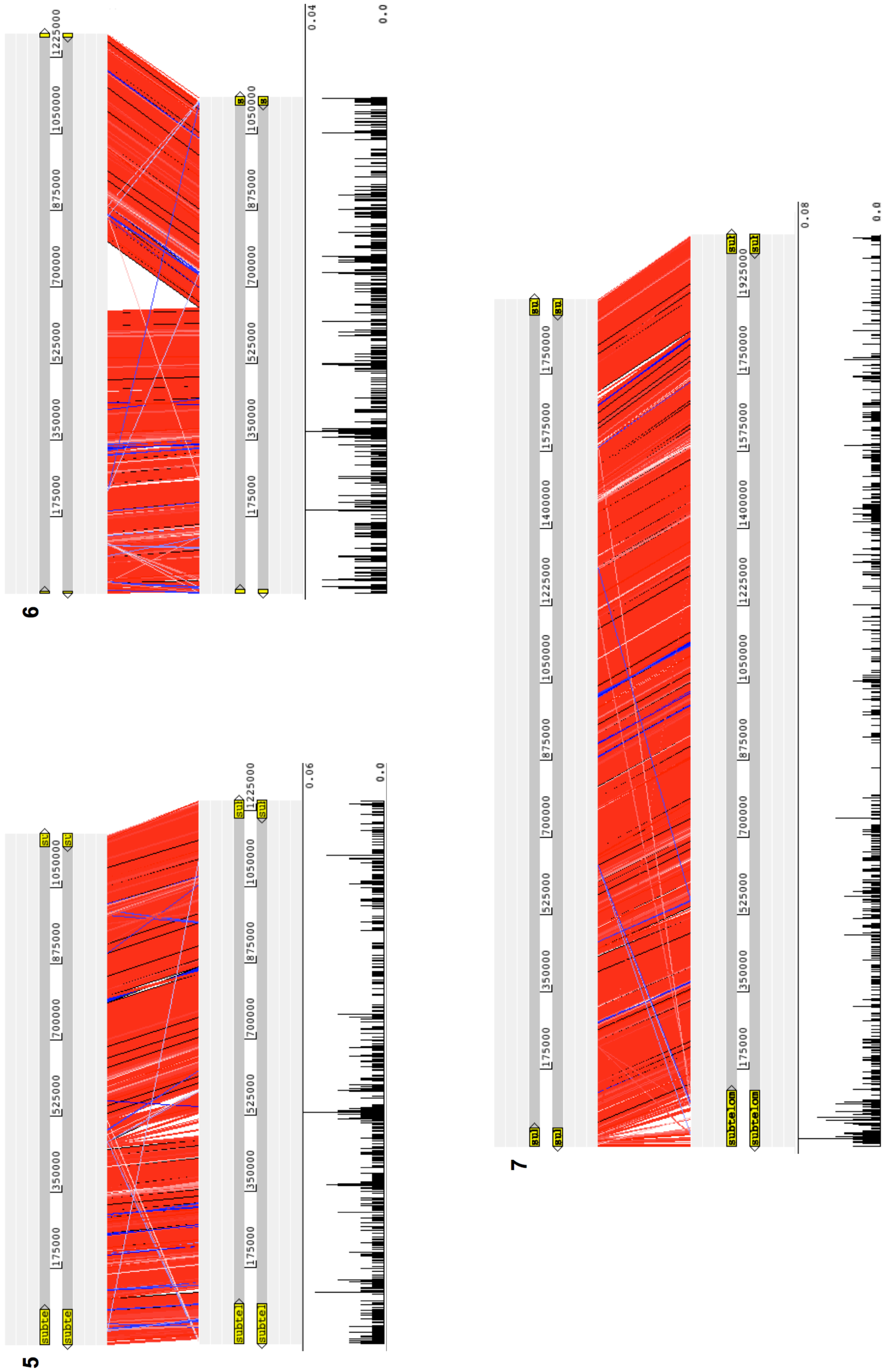


Figure 3-6 continued.

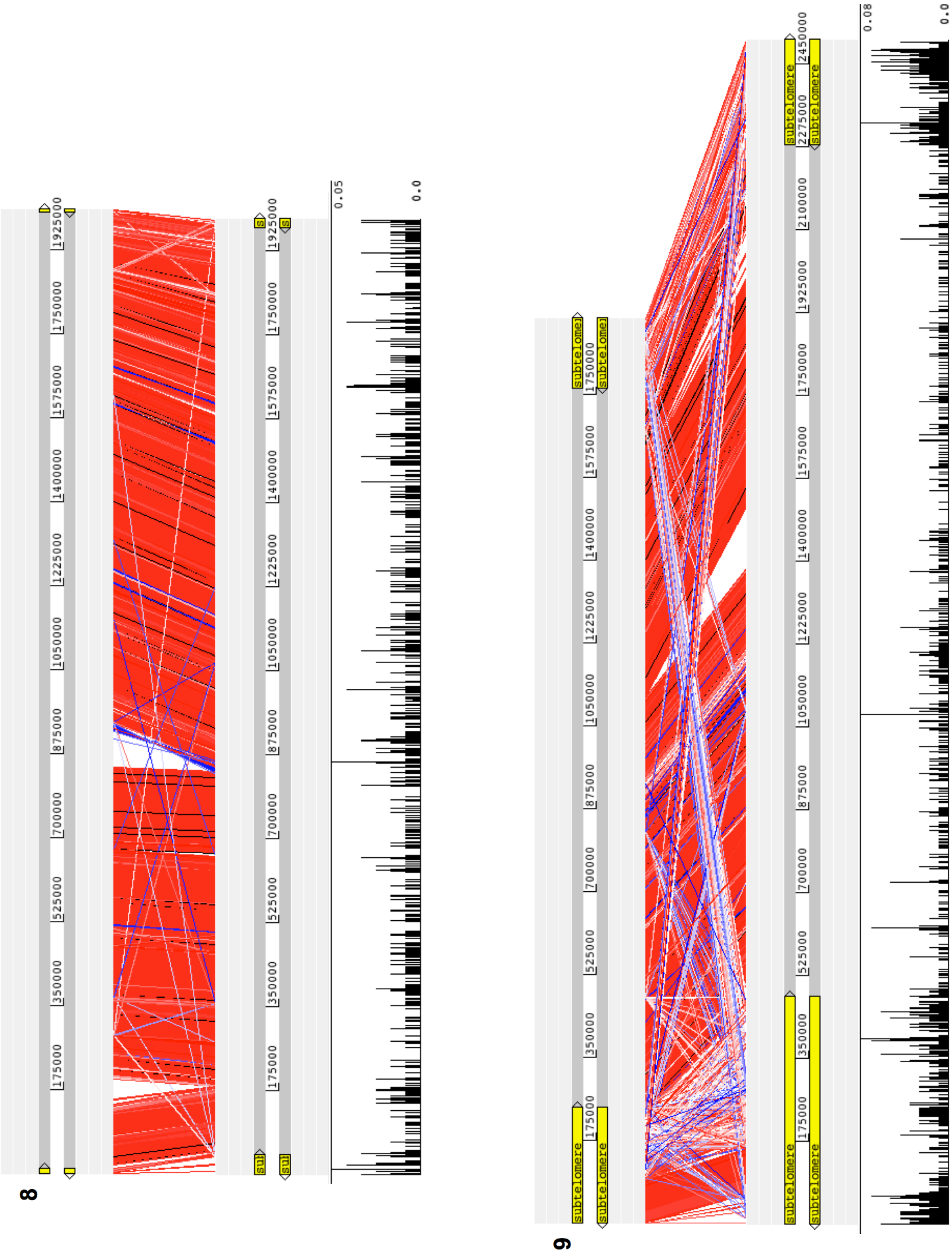


Figure 3-6 continued.

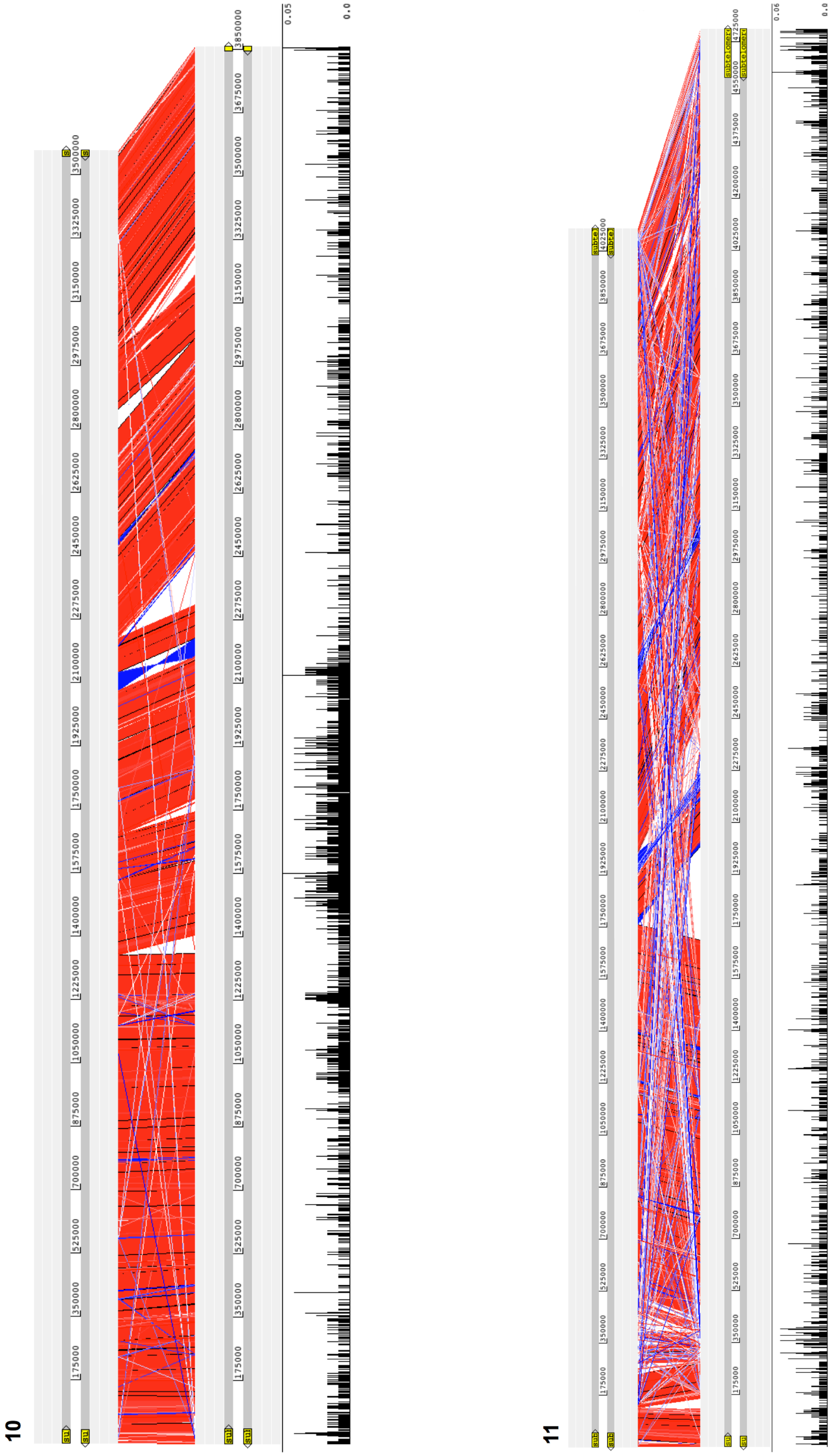


Figure 3-6 continued.

Figure 3-6 (previous four pages). Chromosomes were aligned by BLASTN, and hits of at least 90% identity were visualised using ACT, showing hits between the EATRO 3 chromosome (top set of grey bars) and the EATRO 2340 chromosome (bottom set) connected by red (hit in the same orientation) and blue (reverse orientation) bars. The scale bars in each chromosome indicate the distance along it in bp. Chromosomes are shown at the same scale, except for chromosomes 10 and 11, which are shown at smaller scales to enable each to fit on to one page. Subtelomeres are shown as yellow boxes on both the forward and reverse strand for each isolate and chromosome. Validated differences between EATRO 2340 and EATRO 3, annotated using PCR-free reads as described in the text, are plotted on the bottom graph, relative to the EATRO 2340 chromosome. These plots show the number of differences per base in non-overlapping 100 bp windows along the chromosome, with minimum 0 and maximum as shown on the right-hand end of each plot.

3.4.2 Comparison of mutation rates in cores and subtelomeres

From these SNP plots I calculated the overall mutation frequency in chromosome cores, subtelomeres and bin contigs (Table 3-5). Because the approximate relationship between the two isolates is known (see Figure 3-1), it was also possible to estimate upper and lower bounds for per-generation mutation rates. The mutation frequency in subtelomeres (2.4×10^{-3} SNPs/base) was 2.9-fold higher than in the chromosome cores, with a similar difference (2.7-fold) between the bin contigs and chromosome cores.

Genome region	Total length (bp)	Total SNPs	SNPs/base			SNPs/base/generation	
			Mean	99% CI lower	99% CI upper	Min	Max
Core	20443109	16817	0.00082	0.00081	0.00084	9.7×10^{-9}	3.3×10^{-8}
ST	2047231	4935	0.00241	0.00233	0.00249	28.4×10^{-9}	9.6×10^{-8}
Bin	15342690	34108	0.00222	0.00220	0.00225	26.1×10^{-9}	8.9×10^{-8}

Table 3-5 – Overall mutation frequencies and likely range of mutation rates in cores, subtelomeres and bin contigs.
CI = confidence interval: 99% confidence intervals were calculated assuming a binomial distribution. ST = subtelomeres

To further examine the difference, I calculated the mutation frequency in 10 kb windows across each chromosome core and each subtelomere (Figure 3-7). To simplify analysis I only considered full-length windows, *e.g.* if a contig was 65000 bp long, I examined only the first six windows. This meant that some sequence was ignored from the ends of each region: 63109 bp from the cores, 117231 bp from assembled subtelomeres and 1942690 bp from bin contigs. I used R to model the mutation frequency using a negative binomial distribution, and to test whether genome region made a significant difference to mutation frequency in this model. For this analysis I considered assembled subtelomeres and bin contigs as a single subtelomere region. For the model, the residual deviance was

4044.0 on 3569 degrees of freedom, which indicates that the model described the data reasonably well. The analysis found that genome region (subtelomere or core) was a significant determinant of mutation frequency in a window, and that subtelomeres had a significantly higher mutation rate than cores ($p < 0.0001$). The variation in the mutation frequency was considerably higher in subtelomeres and bin contigs than in the cores, with an interquartile range of 2.2×10^{-3} SNPs/base compared with 8×10^{-4} SNPs/base for cores.

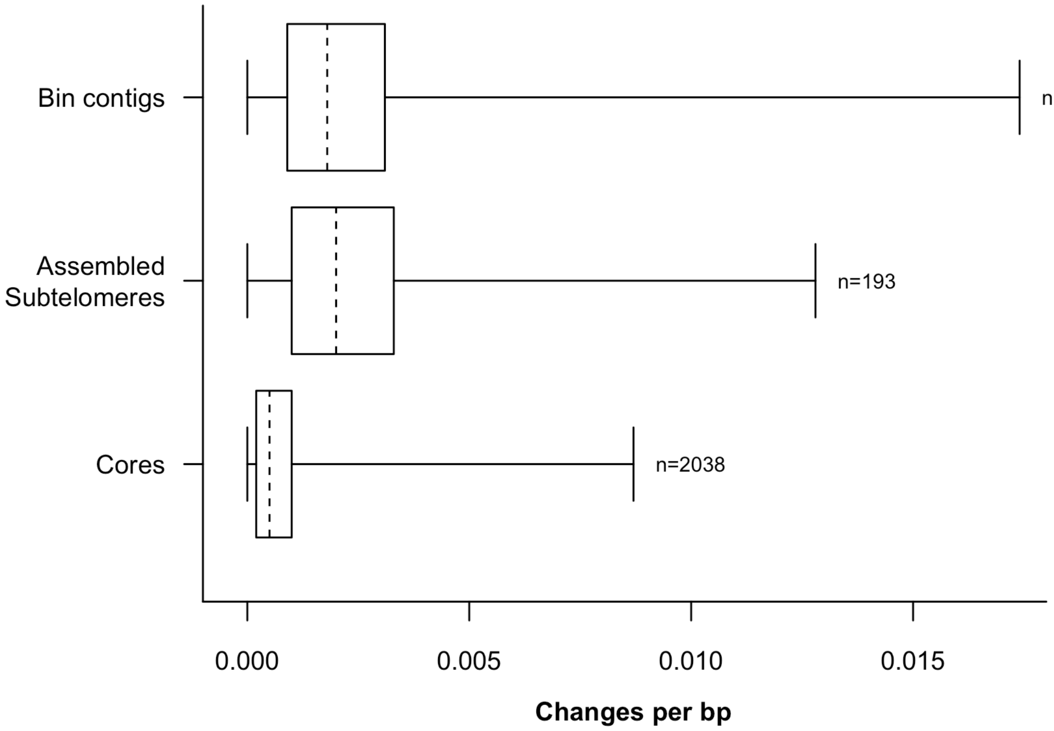


Figure 3-7 – Box plot of mutation frequencies in 10 kb windows across cores, assembled subtelomeres and bin contigs.
The x-axis shows the number of mutations per bp in 10 kb windows across the indicated regions. Boxes enclose values between Q_1 and Q_3; dotted lines indicate the median value; the whiskers extend to the maximum and minimum values. Subtelomere (combined assembled subtelomeres and bin contigs) mutation frequency was significantly higher than core mutation frequency ($p < 0.0001$).

3.4.3 Distribution of new mutations

As described above, new mutations were annotated by identifying SNP positions where one isolate had an allele that was not present in the other. Most of these positions were homozygous in one isolate, and in the second were heterozygous, with one new allele and one allele the same as in the first isolate. Hence, the lineage in which the new mutation arose can be deduced with high probability. The distribution of such new mutations between EATRO 3 and EATRO 2340 can be used to inform the estimate of the relationship between the two isolates. For example, if EATRO 2340 is indeed a direct clonal descendant of EATRO 3, then EATRO 2340 would already contain all alleles that had been present in EATRO 3, and so any new alleles should only have been found in EATRO 2340. In any case, the ratio of the number of new mutations in EATRO 2340 to the number in EATRO 3 should be approximately proportional to the ratio of the time from each to their most recent common ancestor (MRCA). I therefore determined the distribution between the two isolates of annotated new mutations (Table 3-6). This analysis was performed on only substitution mutations, because the calling of indel mutations was likely to be less reliable.

Chromosome	Core		Subtelomere	
	EATRO 3	EATRO 2340	EATRO 3	EATRO 2340
1	71	70	61	100
2	125	175	64	70
3	306	538	354	525
4	1224	153	254	42
* 4 minus anomalous regions	54	112	19	13
5	201	273	46	70
6	187	365	19	21
7	140	286	136	143
8	331	446	45	55
9	188	352	853	959
10	2787	527	58	58
* 10 minus anomalous regions	186	503	58	58
11	628	1265	197	231
Bin contigs	NA	NA	14807	15921
Initial total	6188	4450	16894	18195
Initial EATRO 2340:EATRO 3 ratio	0.72		1.08	
* Total minus anomalous regions	2417	4385	16659	18166
* Revised EATRO 2340:EATRO 3 ratio	1.81		1.09	

Table 3-6 – Distribution of new mutations between EATRO 3 and EATRO 2340 lineages. Lines marked with asterisks are those referred to in the text.

The distribution was initially somewhat surprising, because in the chromosome cores there were considerably more new mutations annotated in EATRO 3 than in EATRO 2340 (6188 compared with 4485). On inspection of the distribution on each chromosome, it became apparent that this unexpected result was dominated by the values for two chromosomes, 4 and 10 (lines five and 12, Table 3-6), where the number of mutations in EATRO 3 far exceeded the number in EATRO 2340. To examine these anomalous results further, I plotted the distribution of new EATRO 3 and new EATRO 2340 alleles along these chromosomes (Figure 3-8). This analysis revealed that on each of these two chromosomes there was one large region where the number of EATRO 3 mutations suddenly increased, and there were very few EATRO 2340 mutations.

In chromosome 10, this observation was reminiscent of a phenomenon observed in other *T. brucei* strains: loss of heterozygosity (LOH) in large regions of chromosome 10 (A. Cooper, PhD thesis, 2009). LOH was observed in multiple cell lines from several strains that were being maintained in cell culture (seven lines from three strains and a fourth strain that was a cross of two of the independent strains), and was associated with an increased growth rate compared with wild-type. It was hypothesised that this loss had occurred on multiple independent occasions because one of the chromosome 10 haplotypes contains a dispersed combination of alleles that together confer a growth advantage, and that converting both chromosomes to this haplotype would give an even greater growth advantage. Therefore, LOH in one direction would be selected for (A. Cooper, PhD thesis, 2009). If such a LOH event had occurred in EATRO 2340, heterozygosities that had been present in the ancestor of both isolates would be removed, and the allele that had been lost from EATRO 2340 would appear to be a new mutation in EATRO 3. Furthermore, EATRO 2340 would be expected to be homozygous throughout this region, or at least nearly so, since it possible that some new mutations could have arisen in one or other homologue since the loss of heterozygosity event. I therefore examined the number of heterozygous positions in the core region of EATRO 2340 chromosome 10 (*i.e.* the number of positions where a heterozygous SNP had been called in the EATRO 2340 reads mapped to the EATRO 2340 genome). This analysis revealed that there were indeed far fewer heterozygous SNPs called across the putative LOH region, with more than 17-fold higher density in the core sequence outside the anomalous region than within it: 0.00019 heterozygous SNPs/base inside the anomalous region compared with 0.00326 heterozygous SNPs/base outside it.

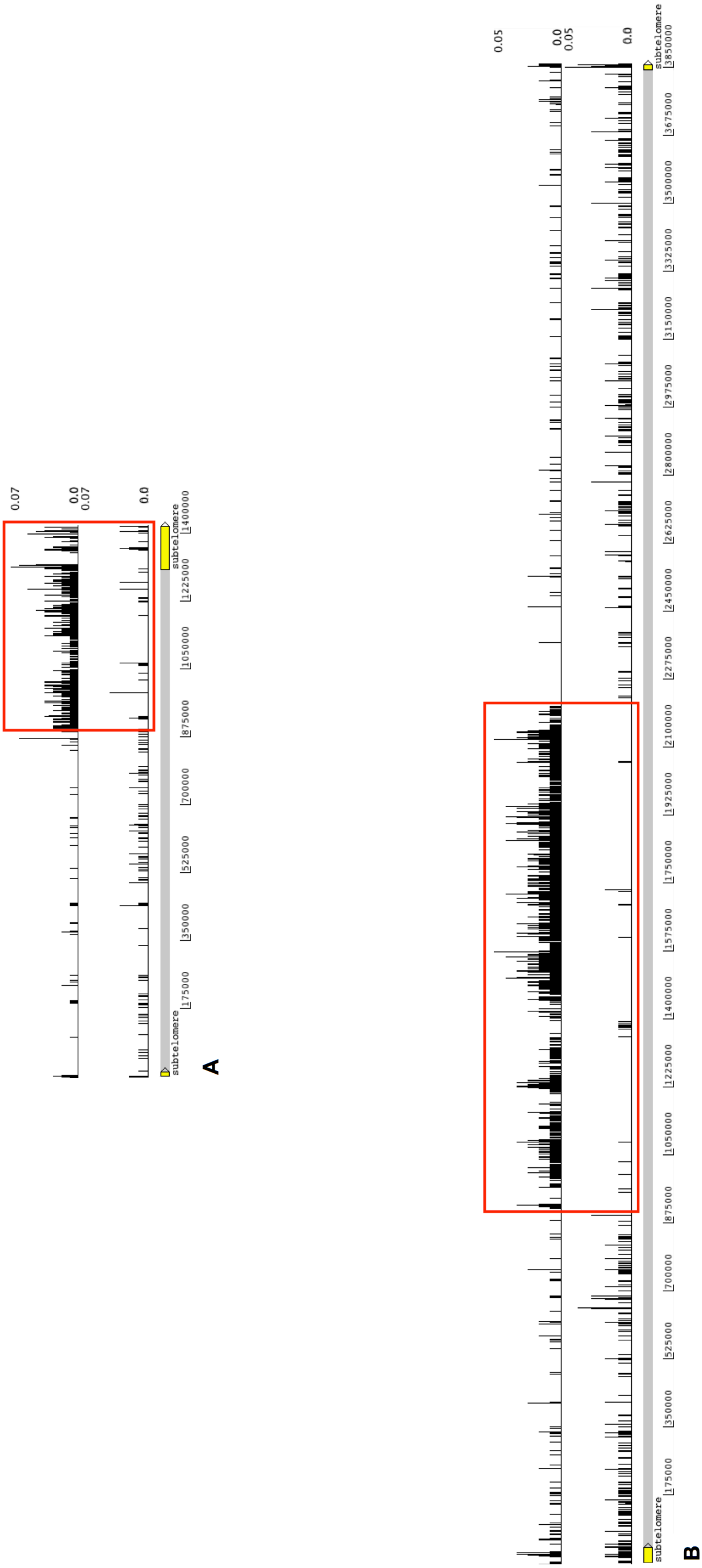


Figure 3-8 – Anomalous regions of new mutation distribution in A) chromosome 4 and B) chromosome 10.

Figure 3-8 continued. Plots are of the number of new mutations per base in 100 bp windows in EATRO 3 (top plots) and EATRO 2340 (bottom plots). Red boxes approximately indicate the anomalous regions described in the text. Yellow boxes below the scale bar indicate subtelomeres, as in Figure 3-6.

The anomalous region of EATRO 2340 chromosome 10 extended approximately from positions 915900 to 2204700, including the genes from *Tb927.10.3680* to *Tb927.10.9100* inclusive. The region between these genes in TREU 927 chromosome 10 (version 5) is positions 958508 to 2258748. The LOH region previously observed has been characterised in four cell lines, and was found to extend from at least the first genetic marker on the chromosome (at position 91140 in the current TREU 927 assembly), to a varying end position between 50% and 70% of the way along the chromosome. The end position was determined at high resolution for only one strain, and the estimated boundaries of the losses of heterozygosity ranged from approximately positions 2250000 to 3195925 in the current assembly (A. Cooper, PhD thesis, 2009). The potential LOH region in EATRO 2340, therefore, was consistent in its end point with previously characterised occurrences. The start point was considerably further from the beginning of the chromosome than in previously described LOH events, where it is likely LOH extended to the start of the chromosome. It was suggested that the LOH occurred by break-induced replication (A. Cooper, PhD thesis, 2009). Meiotic exchange was not considered since in several of the cell lines the event could only have occurred in culture, and trypanosome mating takes place in the tsetse fly (MacLeod *et al*, 2005). Break-induced replication would usually be expected to proceed from the break event to the end of the chromosome (Paques & Haber, 1999), which is consistent with what has been seen previously, but not with the apparently internal conversion in EATRO 2340 chromosome 10. However, it is possible that template switching from the homologous partner to the sister chromatid, or disengagement of the repair machinery, could have ended the LOH (Smith *et al*, 2007); that some other mechanism operated, for example two mitotic crossover events; or that heterozygosity was restored in some way and subsequently selected for. It is not possible to distinguish between these possibilities with the data available.

In general, a LOH event in part of chromosome 10 would explain the observations of the distributions of new mutations between isolates on this chromosome, and the plausibility of this explanation is supported by previous observations in other strains. This finding implies that apparent new mutations in this region should perhaps not be considered in further analyses of mutation rate and profile.

For chromosome 4 the anomalous region extended from approximately position 896500 (just upstream of *Tb927.4.3770*) until the end of the chromosome (including the subtelomere). As has already been noted, in TREU 927 there is a segmental duplication on chromosome 4 shared with chromosome 8, which begins around *Tb927.4.3860* and extends into the subtelomere, and there was an apparent increase in the number of EATRO 3–EATRO 2340 mutations across this region compared with the rest of the chromosome. It therefore seems likely that the anomaly in 'new mutation' distribution is connected with this duplication, and is perhaps due to misassemblies or misalignments caused by it. An alternative is that a LOH event similar to that hypothesised in chromosome 10 has occurred in chromosome 4, in a region that, by its segmental duplication, has already shown itself to be prone to rearrangement. (In fact, the approximate start of the anomalous region is in a region of repeats.) The excess of apparent new mutations in EATRO 3 in this region, therefore, is not of undue concern to the main analysis, although again it is implied that apparent new mutations in this region should perhaps not be considered in further analyses.

The observations in both chromosomes 4 and 10 could be explained by the physical loss of part of one chromosome from each homologous pair in EATRO 2340. However, there was little difference in read coverage of EATRO 2340 between the anomalous and normal regions of both chromosomes, which makes this explanation unlikely. In both cases, it remains possible that meiotic recombination is the explanation for the anomalous distributions of new mutations. However, it seems highly unlikely that mating with another strain has occurred, since genetic analysis has strongly suggested that expansion has been clonal, and inter-strain mating in one of the lineages would have produced a far higher number of new alleles than was observed. A second possibility is that mating between two individuals from the same clonal lineage could occur, *i.e.* inbreeding. Given that in general there is a reasonable level of heterozygosity in both isolates, it is unlikely that inbreeding mating has been a common occurrence. However, meiotic exchange followed by mating would result in partial LOH in all offspring (unless the relevant chromosome had had an exchange event at exactly the same place in both parents), and could produce a phenotype such as is seen in chromosomes 4 and 10. However, examination of the plots of distribution of new mutations in the other chromosomes did not reveal any other substantial regions of this nature, which are unlikely to be absent if an inbreeding mating event had occurred. Furthermore, it is striking that the only anomalous regions should be those where previously reported phenomena could provide an alternative explanation to meiotic exchange.

If these anomalous regions on chromosomes 4 and 10 are removed, the distribution of new mutations between EATRO 3 and EATRO 2340 is more in line with what might be expected (Table 3-6, asterisked lines), with an EATRO 2340 to EATRO 3 ratio of 1.8:1 in the chromosome cores. This core ratio implies that the time from the MRCA to EATRO 2340 is 1.8-fold that of the time for EATRO 3, *i.e.* the time since the MRCA is approximately 21 years for EATRO 3 and 38 years for EATRO 2340 with a divergence date of 1939. The minimum expected ratio, given the range of genealogies outlined in Figure 3-1, would be 1.9:1 (37 years:20 years). This result therefore implies that the separation between the two isolates is around the maximum expected from their known history. Using a divergence date of 1939 implies that the mutations observed have accumulated over 59 years, which corresponds to approximately 86000 generations and yields mutation rate estimates of 4.22×10^{-9} mutations/base/generation for the chromosome cores and 2.34×10^{-8} mutations/base/generation for the subtelomeres and bin contigs combined.

The ratio from the assembled subtelomeres and bin contigs combined (1:1.1) is somewhat lower than both the core ratio and minimum expected ratio[2]. However, because of the haploid nature of these regions, and probable high level of ectopic recombination, it is not surprising that the apparent distribution is not straightforward. Moreover, this global analysis was intended only to give an initial approximation of the mutations occurring in EATRO 3 and EATRO 2340. In particular, it is possible that there were more false positive new mutations called in EATRO 3 than in EATRO 2340, especially in the subtelomeres, since the

[2] The subtelomeres of chromosome 4 initally had a large deviation from this average, with a ratio of 6.0:1. However when the anomalous regions of chromosome 4 were removed as described above, many subtelomeric EATRO 3 mutations were removed, giving a less unusual ratio of 1.5:1 (19 EATRO 3 to 13 EATRO 2340, Table 3-6).

quality of reads was somewhat lower for EATRO 3. This caveat also applies to the estimates of mutation rates calculated above from the core ratio. A more careful and rigorous mutation rate determination and comparison will be described in Chapter 5.

3.4.4 Kinetoplast DNA

The initial BLAST searches described in this section were carried out by Prof Dave Barry. BLASTN was used to query both genomes with the sequence of the coding region of the kinetoplast maxicircle of *T. brucei* Lister 427 (GenBank reference M94286). This search identified one contig of 31 kb in the EATRO 3 genome that contained the coding region of the maxicircle, with 99% identity to the Lister 427 sequence, and some sequence matching the repetitive region between the end of the coding region and the replication origins. The contig appears to be composed of the 3′ two-thirds of the coding region, followed by a poorly assembled version of the repetitive region, followed by a second copy of most of the coding region (Figure 3-9). This assembly is likely due to the fact that kinetoplast DNA (kDNA) exists in circular form, but the assembly algorithm is designed for linear DNA, and has assembled one-and-a-half rounds of the circle.

No contig corresponding to the maxicircle was identifiable in the EATRO 2340, although one EATRO 2340 contig of 1033 bp was 99.32% identical to part of the EATRO 3 contig, containing part of *ND1* and part of *MURF1*. I therefore examined the data generated from mapping the PCR-free sequencing reads to the EATRO 3 contig. The contig had high mean good-quality read coverage with both EATRO 3 (206-fold) and EATRO 2340 reads (456-fold) , indicating, unsurprisingly, that the maxicircle was present in EATRO 2340. No SNPs had been called in the EATRO 3 PCR-free reads mapped to the contig. One SNP had been called in the EATRO 2340 PCR-free reads, but the SNP was caused by a length difference in a homopolymeric adenosine tract following a gap, and so was unlikely to be genuine. There may be other SNPs that were not called, but in general, this analysis indicated that there was very little difference between the kDNA of the two isolates.

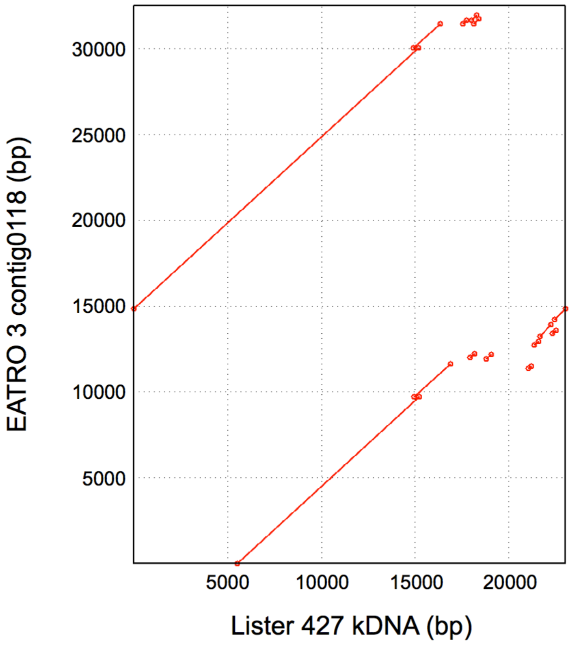


Figure 3-9 – EATRO 3 contig containing kDNA maxicircle.
The EATRO 3 contig was identified by BLASTN with the kDNA maxicircle of Lister 427, then aligned with the maxicircle sequence using NUCmer and the alignment visualised with MUMmerplot. The Lister 427 kDNA is on the x-axis and the EATRO 3 contig is on the y-axis, with the scale indicating the distance in bp along each. Where the sequence could be aligned, a red line is drawn, with dots marking the start and end of the alignment.

3.5 Summary

In this chapter, I have described the assembly, annotation and comparison of two new *T. brucei* genomes, from very closely related isolates of *T. b. rhodesiense*. These genomes were of draft quality with many gaps, and a large amount of the assembled sequence could not be mapped onto chromosomes. However, there was a large amount of sequence present; the genomes had high mean coverage depths; and most annotated TREU 927 coding genes could be transferred to both genomes. These metrics demonstrated that the genomes were of sufficiently high quality for further study. The assemblies did not contain every gene that would be expected in the isolates' genomes. However, the coverage of known genes was sufficiently high that it seemed reasonable to expect to find a substantial proportion of the isolates' *VSG* genes in the assemblies.

Comparison of the conserved chromosome cores with other trypanosome genomes confirmed that EATRO 3 and EATRO 2340 are extremely closely related. This analysis supported the presumed relationship between the two isolates, which was derived from the provenance of the original parasites and from serology. Despite this close relationship, however, comparison of the two genomes uncovered numerous differences between the isolates. Previous observations led us to predict that trypanosome subtelomeres are subject to an elevated rate of mutation, and indeed, in these genomes the subtelomeres had a significantly higher mutation rate than the cores.

Confirmation both of the close relationship and the presence of difference between the two genomes gave me confidence that a comparison of EATRO 3 and EATRO 2340 would yield fruitful insights into *VSG* evolution: the isolates appeared to be similar enough that it would be possible to identify the same *VSG* gene in both genomes, but different enough that there would be changes apparent in many genes. Further, at a first approximation, the way that mutation proceeds was quantifiably different between cores and subtelomeres, which suggested that closer analysis would be able to uncover any more detailed quantitative and qualitative differences. I therefore proceeded with analysis of the *VSG* archives within and between the two isolates.

Chapter 4: The *VSG* archive of EATRO 3 and EATRO 2340

4 The *VSG* archive of EATRO 3 and EATRO 2340

4.1 Introduction

Chapter 3 described the assembly of two new trypanosome genomes, those of *T. b. rhodesiense* isolates EATRO 3 and EATRO 2340. These assemblies substantially increased the number of sequence data available for a range of genomic studies in *T. brucei*. However, the primary purpose of the genome sequencing was to provide new subtelomeric sequence in which a new set of *VSG* genes could be annotated. The first use of these data was to compare the *VSG* genes between the two isolates in order to gain insight into the detailed processes of subtelomere evolution. A second outcome was the availability of a new data set that could be compared with previously described archives to study the *VSG* archive more generally.

In this chapter I will first describe the annotation of sequences encoding *VSG* NTDs in both genomes, a non-trivial process due to the considerable variability inherent in *VSG*s; the identification of sequences that corresponded to the same gene in the two genomes; and the cataloguing of the substitution mutations that had occurred in the *VSG* NTDs and in the core genes. I will then present a detailed comparison of mutation in the subtelomeric *VSG* NTDs with that occurring in chromosome core genes, and with that occurring across subtelomeres as a whole. The aims of the analysis were to test in detail the hypothesis (already supported tentatively by the analyses described in Chapter 3) that mutation in the subtelomeric genes is a different, more rapid process from what occurs in core genes; to examine the effect of mutations on the subtelomeric antigen genes; and, if this first hypothesis was supported, to test the hypothesis that the elevated mutation rate of *VSG* NTDs applies across the subtelomeres rather than being gene-specific. The project focused on evolution of *VSG* genes but the chapter also contains a brief analysis of substructuring and pseudogenicity in the newly annotated EATRO 3-EATRO 2340 *VSG* archive.

4.2 Annotation of *VSG* NTDs

4.2.1 Support vector machine-based *VSG* prediction

Because the sequences of VSG and *VSG* genes are so diverse, and because genes often include frameshifts, annotation approaches based solely on sequence similarity will give an incomplete picture. Therefore, to aid my annotation of *VSG*s I used a support vector machine-based prediction program (SVM-VSG), developed by Dr Jon Wilkes. This program examines DNA sequence in overlapping 1000 bp windows, and decides for each window whether it is likely to contain *VSG* domain sequence, with separate analyses for NTDs and CTDs. It can therefore be used to query whether a particular sequence is likely to be a *VSG* gene, and *de novo* to find windows of sequence that likely contain *VSG* sequence.

4.2.2 Full-length, potentially intact genes

I began annotation by looking for *VSG*s that were encoded by a complete ORF. Examination of previously annotated full-length *VSG* genes in VSGdb (Marcello *et al*, 2007) gave a range of 1317 to 1623 bp for this type of gene; I therefore used a range of 1290 to 1650 bp for my search. I used Artemis to find all ORFs above 1290 bp in every subtelomere and bin contig for both genomes. I then extracted the sequences for the ORFs and used a Perl script to trim them to the first in-frame ATG, remove any that contained assembly gaps, and filter the final sequences for maximum and minimum length. These sequences were analysed with the SVM-VSG program, and a sequence was retained as a likely full-length *VSG* if both an NTD and a CTD were predicted within the ORF.

Within the ORFs, the NTD-CTD boundaries were annotated using HMMs. HMMs for the A- and B-type NTDs and for the CTD were provided by Dr Bill Wickstead (University of Nottingham). The ORFs were translated, and then queried with the NTD HMMs using hmmsearch from the HMMer suite of programs. If either type of NTD was found, the domain boundary was defined after the last residue of this domain, and the NTD type was defined according to which HMM had a hit. If no NTD was found, the ORF was queried with the CTD HMM, and the DNA sequence of the CTD hit was manually examined to locate the conserved pattern of cysteines. The domain boundary was defined as 50 residues upstream of the first conserved cysteine, and the NTD type was defined as ‘unknown’. This approach identified 280 *VSG*s in EATRO 3 and 348 in EATRO 2340.

4.2.3 Pseudogenes

To extend the analysis, I used the SVM-VSG program to locate windows of sequence containing likely NTDs in the chromosomes and in the contigs longer than 10 kb (the VSG-SVM program could not analyse shorter contigs). After excluding windows containing assembly gaps, this yielded 883 NTD windows in EATRO 3 and 2179 in EATRO 2340. In order to define the *VSG* NTDs within these sequences, I used a combination of HMMer and BLASTX. I produced translations of all three frames of the NTD windows, and queried these with the NTD HMMs. When the HMM had a near full-length hit, this single hit was used to define the boundary of the domain. For those that did not have a near full-length hit but did hit adjacent fragments of NTDs in different frames, I developed a script to piece hits together where possible to give a near full-length domain. For all HMM-defined NTDs, I developed another script to check the start of the domain: if the sequence did not begin with a methionine (ATG) codon, the sequence immediately 5′ was scanned for a suitable start codon. If a start codon was found, the sequence was extended to include it; if not, the original start position was retained. This approach delineated 421 NTDs in EATRO 3 and 1077 in EATRO 2340.

There remained a large number of NTD windows where an NTD could not be found using HMMer. I generated protein BLAST databases from all the NTDs I had earlier annotated in both genomes, and used BLASTX to query the EATRO 3 database with the remaining unannotated EATRO 3 NTD windows, and the EATRO 2340 database with the remaining EATRO 2340 windows. Again, any near full-length BLASTX hits were used, where possible, to define the boundaries of the domains; then for NTD windows without near full-length hits, adjacent fragments found in different frames were used, where possible, to delineate domain boundaries. I then used the script developed for use with HMMer results to check the start positions of these domains. NTD windows for which it was not possible to annotate domain boundaries by either the HMMer or the BLASTX approach were not considered further (393 SVM windows in EATRO 3, 860 for EATRO 2340).

Finally, I checked all the genes annotated in each genome to ensure that they were all non-overlapping, so that no part of the genome was considered twice. This resulted in five genes being discarded from EATRO 3 and nine from EATRO 2340. The final data from the annotation process in these two sections are summarised in Table 4-1. The larger number of genes in EATRO 2340 was to be expected, given that the EATRO 2340 genome assembly was both larger and of better quality than the EATRO 3 assembly, as discussed in Chapter 4. Given that the length of assembled subtelomeric sequence was less than the subtelomere lengths of these parasite strains, determined from molecular karyotypes, neither annotated gene set was likely to represent the full *VSG* NTD archive existing in the genome.

Genome		**EATRO 3**	**EATRO 2340**
Full-length ORFs		279	348
SVM NTD windows		883	2179
HMMer	**From single hit**	286	657
	From several hits	132	411
BLASTX	**From single hit**	43	125
	From several hits	29	126
Initial annotation total		769	1667
BLASTN from other genome		341	130
Total VSG *NTD*s		**1110**	**1797**

Table 4-1 – *VSG* NTD coding sequences annotated in EATRO 3 and EATRO 2340 genomes.

Although many of the NTDs contained probable frameshift mutations, it was possible to predict the most likely protein sequence, *i.e.* the sequence that looked most like it could form part of a functional VSG, for almost all the genes. For NTDs that were annotated as full-length ORFs, I simply took the translation of the NTD part of the ORF. For NTDs that were annotated by searches with HMMer or BLASTX, I used the positions of the HMM or protein sequence hits to annotate where the reading frame shifted, and used the translation of several frames. If stop codons remained in the translated sequence, they were assumed to be the result of substitution mutations, and there was no way of determining the original sequence, so they were allowed to remain in the sequence.

4.3 Gene pairs: identifying and comparing the same NTD gene in the two genomes

4.3.1 Identifying gene pairs

The analysis described above generated a set of 769 *VSG* NTDs in EATRO 3 and a set of 1667 NTDs in EATRO 2340. For each NTD in EATRO 3, I attempted to identify the same gene in EATRO 2340, and vice versa. I used BLASTN to query a database of the EATRO 2340 genome with the EATRO 3 NTDs. I used the highest-scoring hit for each NTD to define the region of the EATRO 2340 most likely to correspond to the same gene. In some cases, several adjacent hits on the same EATRO 2340 contig could be concatenated together to give sequence that was a better match for the NTD than any single hit, for example if the EATRO 2340 sequence was interrupted by a gap. In these cases I took all the sequence between the start of the first hit and the end of the last hit. I then repeated this process with the EATRO 2340 NTDs and an EATRO 3 genome sequence database. Gene pairs were classified by quality, according to what percentage of the original NTD was covered by its hit, and to the degree of identity of the aligned sequence.

In pairing genes by this method, it is possible that gene pairs that appeared to be high quality were not in fact true pairs, for example if more than one NTD was paired with the same sequence in the opposite genome. In order to reduce the likelihood of such false pairings, I carried out reciprocal BLASTN searches. I extracted from the EATRO 2340 genome the sequences that had been defined as the partners of EATRO 3 NTDs, and used these EATRO 2340 sequences to query the EATRO 3 database. When the best hit for the EATRO 2340 sequence corresponded to the EATRO 3 NTD with which it was paired, I accepted the pairing. If not, the pairing was discarded and the EATRO 3 NTD was classed as having ‘Poor or no hit’. I checked the EATRO 2340 NTD pairings in a similar way (Figure 4-1A and B). This approach led to the rejection of 28 pairings from the EATRO 3 NTDs and 431 from the EATRO 2340 NTDs. Most of the pairs excluded had an identity below 90% (334 out of 459 genes excluded). I repeated this process to identify how many EATRO 3 and EATRO 2340 NTDs had identifiable counterparts in TREU 927, which is from a distinct lineage: far fewer had good hits, or indeed any hits at all (Figure 4-1C and D).

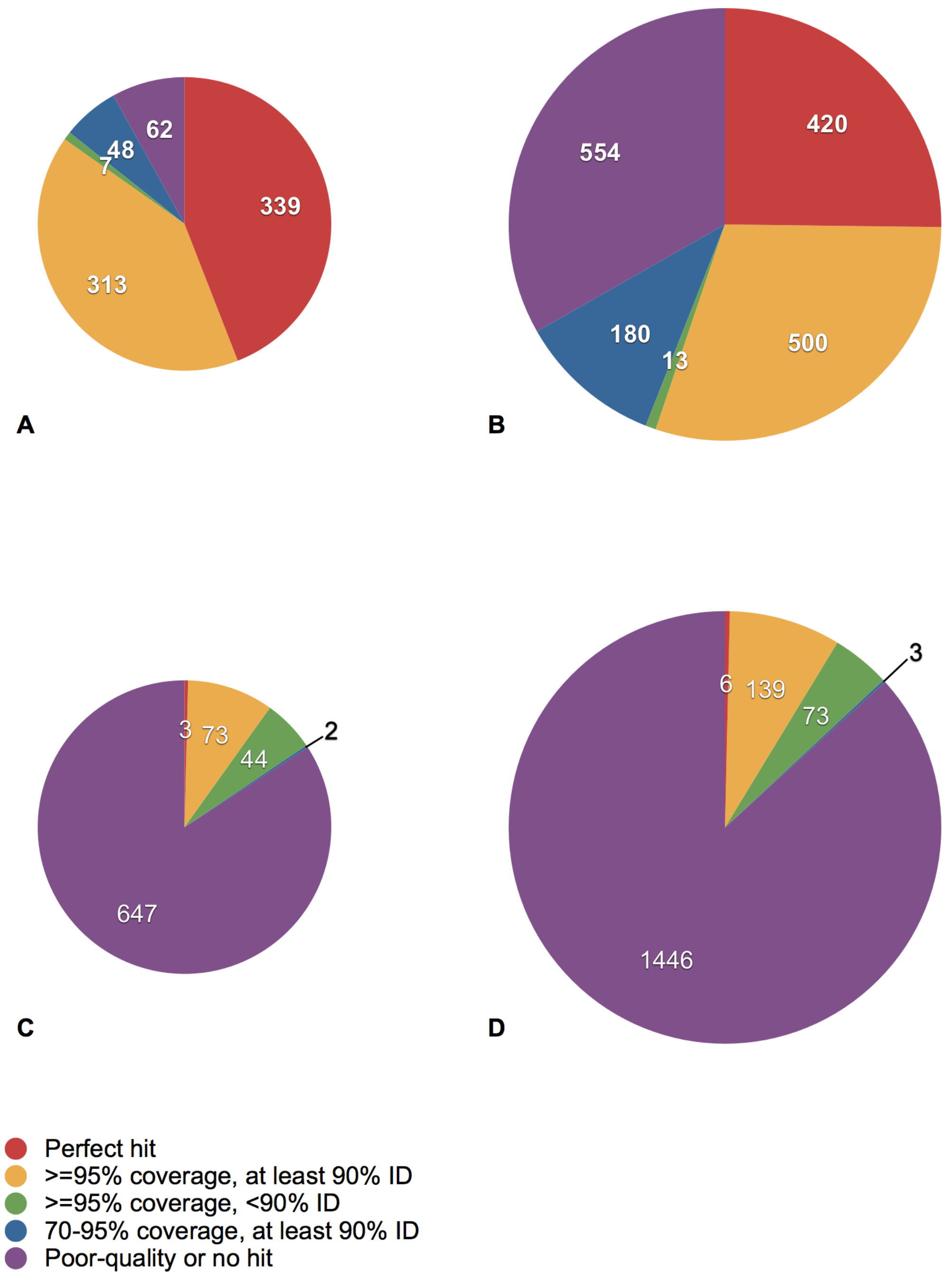


Figure 4-1 – Results of pairing by BLASTN searches with *VSG* NTDs, after checking pairings by reciprocal BLASTN.
Results for all queries are shown, classified by percentage identity of the hit (% ID) and by the percentage of the query covered by the hit (% coverage). A) EATRO 3 NTDs queried against EATRO 2340 genome. B) EATRO 2340 NTDs queried against EATRO 3 genome. C) EATRO 3 NTDs queried against TREU 927 genome. D) EATRO 2340 NTDs queried against TREU 927 genome. The areas of the pies are approximately proportional to the relative number of query sequences used in each comparison. The number of NTDs in each category is indicated in the appropriate segment. 'Poor-quality' hits had either less than 70% coverage, less than 90% ID, or both.

4.3.2 Pairwise identities of gene pairs

Where 95% or more of the sequence of an NTD was contained in its assigned partner (described hereafter as at least '95% coverage'), and reciprocal BLASTN confirmed the pairing, it seemed plausible that the two sequences genuinely represented the same gene in the two genomes. When considering the percentage sequence identity of such high-coverage sequence pairs (Figure 4-2), it became apparent that the vast majority of pairs had high identity: in EATRO 3, 96% of high-coverage pairs had identity of at least 95%, and in EATRO 2340 this figure was 95%. This observation suggests that the mutation processes acting on these genes are mainly small-scale, for example base substitution or short indels. (However, it should be noted that it is possible that larger-scale mutations had also taken place, but that genes in which these occurred had had their sequence changed to such an extent that they were excluded from this analysis by the thresholds used.)

In order to be most confident about the gene pairings, I took sequence pairs with at least 95% coverage and 95% identity for further analysis as ‘gene pairs’. It seemed likely that there would be a large degree of overlap between the annotated EATRO 3 and EATRO 2340 NTDs. I therefore checked the gene pairs to see whether the hit of an EATRO 3 NTD in EATRO 2340 had already been annotated as an NTD, and vice versa. If this was the case, I excluded from further analysis whichever NTD was shorter; or, if the two were the same length, I excluded the EATRO 2340 NTD. This step removed 489 duplicates from the analysis, 90 from EATRO 3 and 399 from EATRO 2340. There were also four cases where the hit from one gene of the pair fitted the criteria for a ‘high-quality match’, but the other did not, and so only one gene from the pair had been included in the analysis anyway. In all four cases the asymmetry was the result of the first gene being shorter than the second, leading to the percentage coverage differing although the length of sequence matched was approximately the same. In these cases, the pairing was retained and only the shorter gene was included in the analysis. These filtering steps left 1030 pairs that were unique in the analysis, with 493 of these matched with an already-annotated gene.

Next, for each gene pair that had less than 100% coverage, I extended the start and/or end of the hit so that, as far as possible, each position in the query NTD had a corresponding base in the hit. For example, if a hit covered positions 5-1000 of an NTD that was 1005 bp long, I extended the hit by 4 bp at the start and 5 bp at the end. This step was so that mutations at the very ends of genes, which would have caused the BLAST hits to be truncated, would not be missed.

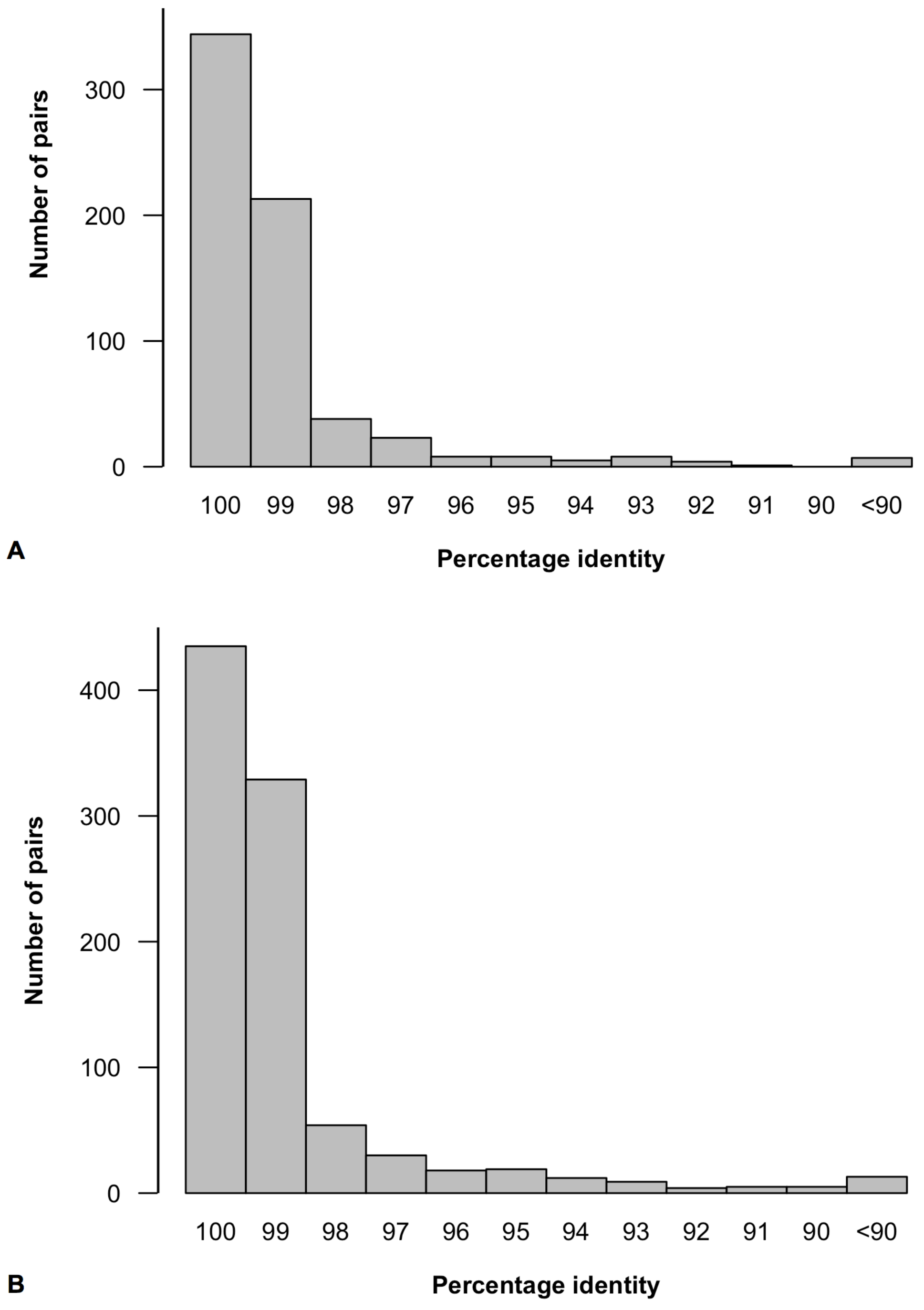


Figure 4-2 – Percentage identities of query-hit pairs for which the best BLAST hit covered at least 95% of the query NTD sequence.
A) EATRO 3 NTDs queried against the EATRO 2340 genome. B) EATRO 2340 NTDs queried against the EATRO 3 genome. The percentage identity is shown in bins of 1%.

In cases where there was a high-identity match, but the gene had been annotated in only one genome using the HMMer and BLASTX annotation approach described in section 4.2, I used the hit information to annotate new NTDs. If a gene had a region with an internal gap, this region was not annotated as a new gene; if one end of the hit was truncated by an assembly gap then the region was annotated as a gene, but its possible incompleteness was noted for downstream analyses. I then double-checked that the ‘new’ genes did not overlap with each other or with any previously annotated genes. This annotation gave 341 new genes in EATRO 3 and 130 in EATRO 2340, bringing the total number of annotated genes to 1110 in EATRO 3 and 1797 in EATRO 2340, with 964 paired up between the genomes. There were a further 66 genes that had a good BLAST hit in the other genome (13 from EATRO 3 and 53 from EATRO 2340), but the hit sequence was interrupted by an assembly gap. These pairs were considered in the SNP analysis, but the gapped hit sequences were not annotated as new genes.

I classified all NTDs into A-type and B-type according to the type of sequence used to annotate them. In EATRO 3, there were 519 A-type NTDs, 560 B-type, and 31 that could not be classified. In EATRO 2340, there were 843 A-type NTDs, 921 B-type, and 33 that could not be classified.

4.4 Detecting changes in *VSG* NTDs

4.4.1 Core gene pairs

As described in Chapter 3, I used an automatic annotation program to transfer core gene annotations to both genomes, and many genes were annotated in both. However, to produce a set of control gene pairs where the method of pairing of sequences was more comparable with that of the *VSG* NTD gene pairs, I used a similar BLASTN analysis to produce a set of pairs. This set of pairings would allow comparison of the mutations in *VSG* genes with those in core genes. Using the 7605 coding sequences automatically annotated in the EATRO 2340 chromosome cores, I repeated the steps of BLASTN, finding of the best hit, and checking by reciprocal BLASTN to generate a set of gene pairs. A far higher proportion of the original genes could be paired with high confidence than was the case for *VSG* NTDs (Figure 4-3A). I applied the same filters to these gene pairs (at least 95% identity and at least 95% coverage) as to the *VSG* NTD gene pairs, to give a comparable set of genes. Some 97% of the gene pairs with at least 95% coverage had 99% pairwise identity or above (Figure 4-3B), compared with 85% for *VSG* NTDs originally annotated in EATRO 3 and 82% for NTDs from EATRO 2340.

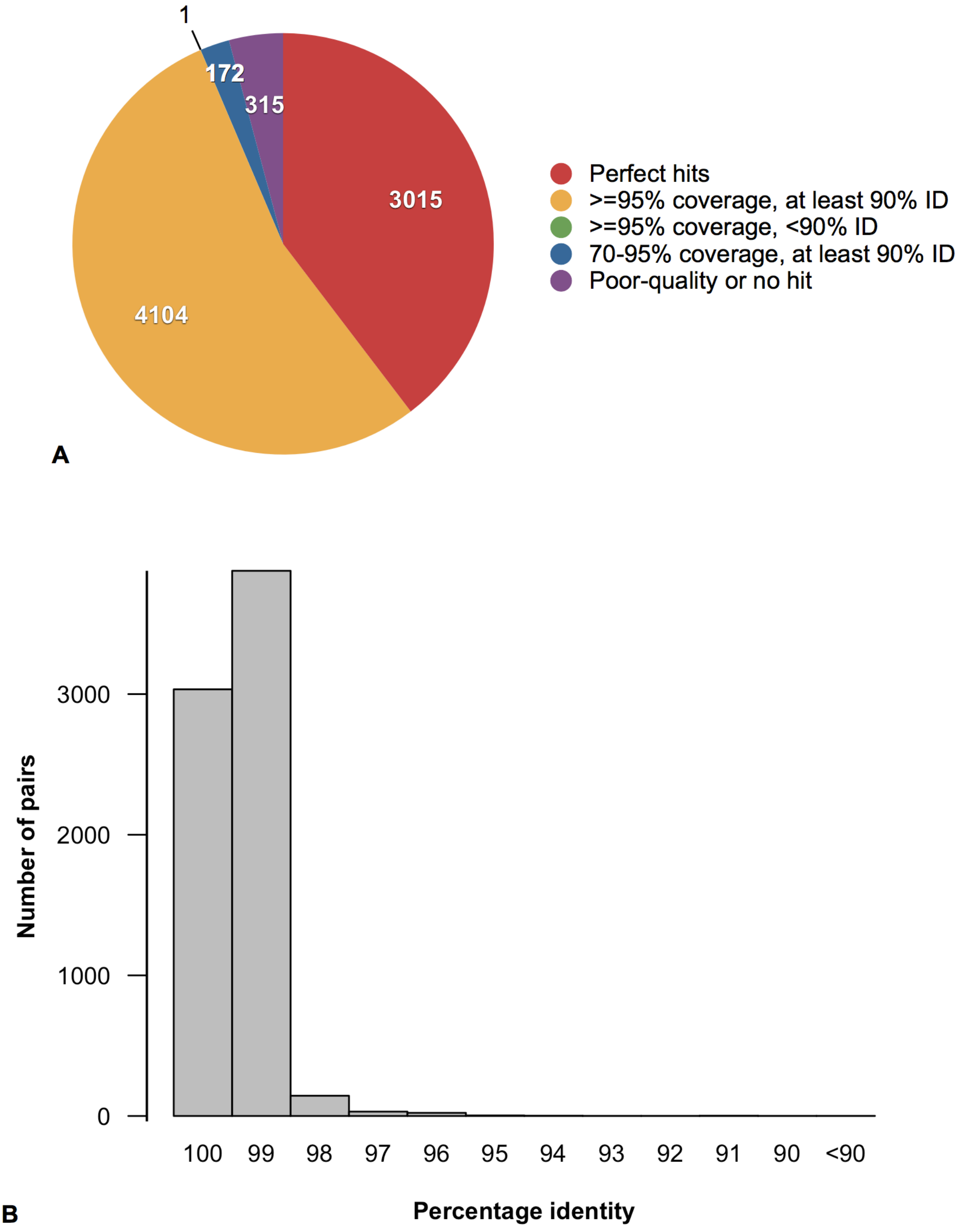


Figure 4-3 – Gene pairing with chromosome core genes.
A) Results of BLASTN search of EATRO 3 genome with coding sequences annotated in the EATRO 2340 chromosome cores. B) Percentage identities of gene-hit pairs with at least 95% coverage of the query by the best BLAST hit. Analyses were carried out as in Figure 4-1 (part A) or Figure 4-2 (part B). The pairwise percentage identity in B is shown in bins of 1%.

4.4.2 Cataloguing of changes

Initially, I catalogued EATRO 3-EATRO 2340 SNPs by aligning and comparing the sequences of *VSG* NTDs paired up by BLAST (Method 1). I attempted to minimise the number of false SNPs by checking each position that differed to see if the apparent difference could be attributed to an error or ambiguity in the assembly in both sequences. For each potential change, I therefore checked to see if a ‘self-mapping SNP’ had been called at this position when PCR-free sequencing reads were mapped back to the genome from which they came (*e.g.* EATRO 2340 reads to the EATRO 2340 genome), and discarded the EATRO 3-EATRO 2340 SNP if a self-mapping SNP had been called in both genomes.

In order to check the reliability of the SNP calling and genome assembly, I selected 16 pairs of genes, exhibiting a range of different types of change, for resequencing. For templates, I purified genomic DNA from EATRO 3 parasites grown in mice (by Alana Hamilton), and obtained genomic DNA from cultured procyclic form EATRO 2340 parasites (cells grown by Dr Barbara Marchetti and DNA extracted by Jamie Hall by phenol/chloroform extraction). I designed primers to each gene (sequences given in Table 2-2), and used them for *de novo* PCR with EATRO 3 or EATRO 2340 genomic DNA as a template, cloned gel-purified PCR products into the sequencing vector pCR4-TOPO and had the inserts sequenced by capillary Sanger sequencing with the primers M13 uni and M13 rev. Two pairs of genes failed to amplify in PCR, but in both cases there was no product from either genomic DNA template, implying that the failure was due to the PCR primers or conditions rather than the gene being absent from the genome.

Three clones from each successful PCR were sequenced. However, many of the sequences obtained differed greatly from the expected sequence, in one instance having only 33% identity. In these cases, it seems likely that due to the relatively low annealing temperature used in PCR (50-55°C), the products were from mis-primed PCRs. Additionally, the primers were designed to bind to sequence at or just beyond the ends of the NTDs, so they might contain sequence complementary to conserved flanks or conserved sequence in adjacent CTDs, making mis-priming more likely. Unfortunately, time did not permit the repetition of these experiments with more stringent reaction conditions.

Of the 28 genes examined, only six had PCR products that were identical to the expected sequence. Two of these genes also yielded PCR products with over 100 differences from the expected sequence. A further eight each gave products with over 100 differences from the expected sequence, with no products similar to what was expected. In the remaining genes, there were between two and 41 differences between the recovered sequence and the expected sequence. The differences between the recovered and the expected sequence often coincided with positions annotated as EATRO 3-EATRO 2340 SNPs. However, if a SNP coincided with a difference it does not necessarily mean that the SNP was invalid: for example, consider a gene that has two copies in each genome, only one of which had been assembled in each. If position 10 in copy two had changed from A to G in EATRO 2340, then position 10 in the composite gene may be A or G in the EATRO 2340 assembly, but only A in the EATRO 3 assembly. If the position had been called as G in EATRO 2340, it would have been annotated as an EATRO 3-EATRO 2340 SNP. The described experiment could recover a sequence with A instead of G at this position in EATRO 2340, but it does not follow that the A to G change did not happen. In resequenced genes with reasonably good matches to the reference, such a scenario could have occurred in approximately half of the SNPs predicted; in the other half, both isolates contained both possible alleles. Therefore, the resequencing experiments were not as informative as had been expected, but where the results could be used to check SNP calls, it appeared that in many cases the apparent SNPs were not in fact true differences between isolate genomes.

The results of the PCR resequencing cast some doubt on the reliability of the SNPs annotated by comparing the two sequences, and implied that the archive is more complex than was suggested by the assembled genome. That is not to say that there was any substantial problem with the *VSG* genes that were assembled: it was possible to recover some of the expected sequences from an experiment with low-stringency PCR conditions; and, more convincingly, out of around 1800 NTDs annotated in EATRO 2340, around 670 had matches with at least 99.9% identity in the EATRO 3 genome — neither of which is likely to have happened if the genes had been primarily assembly artefacts.

I examined in more detail the SNPs called in the whole-genome comparison, where the PCR-free reads mapped to the EATRO 2340 genome had a SNP called in one genome but not the other (see previous chapter), and found that the vast majority of these (139234 out of 142664, 97.6%) had been called as heterozygous rather than homozygous SNPs. This heterozygosity implied that in most cases where there was a SNP, the sequence in which it was found was present in two copies, that only one copy had been assembled for EATRO 2340, and that one copy differed at the SNP position between EATRO 3 and EATRO 2340. Heterozygosity would be expected from the core sequences, which are diploid, but would not necessarily be expected for the subtelomeric *VSG* genes: as indicated by the variation in subtelomere size between homologous chromosomes, subtelomeres can be considered as effectively haploid, and previous studies have shown haploidy (Melville *et al*, 1999; Callejas *et al*, 2006; Hutchinson *et al*, 2007). However, the preponderance of heterozygous SNPs in the subtelomeres indicated that most genes (or at least most genes that change) had a partner somewhere in the archive of sufficient similarity that the two genes could not be separated from each other by the assembly algorithms. In this case, it was quite likely that if there were more than one difference between the partners, and the differences were widely separated (such that they would not be connected by the fragment size used in the assembly), the gene in the assembly would have been a composite of both partners, not absolutely identical to either.

The possibility of composite assembled genes represents a complication that must be considered in analysis of the *VSG* archive. However, if the sequences are sufficiently similar to be conflated by the assembler, then using the composite gene in the analysis will give a reasonable general idea of the archive, though we must be careful when focusing on specifics. Additionally, for almost every SNP called there were only two variants, and the alleles were predicted to be in approximately a 1:1 ratio[3]. These data suggest that almost all composite assembled genes represented only two genes, which makes comparison with the diploid core valid when looking at SNPs called from reads, as it removes the possibility that more SNPs were called in a same-sized window in the subtelomeres than the cores solely because the subtelomere window actually sampled more genes.

[3] SNPs that had more than two variants, or where the alleles were predicted to be present in a different ratio noticeably different from 1:1, were excluded early in the SNP processing procedure, so there are no data for how many of SNPs with these characteristics would have passed all the filters, and none were present in the final EATRO 3-EATRO 2340 SNP data set. However, before those SNPs that were present in both genomes were excluded, there were 184358 SNPs with appropriate coverage called from mapping EATRO 2340 reads to the EATRO 2340 genome, and 176579 SNPs called from mapping EATRO 3 reads to the EATRO 2340 genome; 77 and 103 respectively had more than two alleles called; 120 and 142 respectively had only two alleles called but the frequency of the alternative allele was not predicted to be between 0.4 and 0.6.

The results of the PCR resequencing experiments suggested that there was more variation in both genomes than had been captured in the genome assemblies, but that I could use the data from mapping the PCR-free reads to each genome to try to find a more reliable method of annotating SNPs in the *VSG*. I therefore developed a second strategy for SNP annotation, Method 2. The EATRO 2340 PCR-free Illumina sequencing reads were used only in the gap-filling step of the EATRO 2340 assembly, although the EATRO 3 PCR-free reads were used to assemble the EATRO 3 genome. Both sets of PCR-free sequencing reads had already been mapped to the EATRO 2340 genome, and SNPs called. For the EATRO 2340 half of each high-identity pair of *VSG* genes, I compared the SNPs called from mapping the EATRO 3 reads and from mapping the EATRO 2340 reads. For positions where a SNP had been called in only one set of reads, I checked that the coverage of these positions with the other set was sufficient that, if a SNP were present, it would likely have been called. If this were the case, I annotated the position as a putative SNP. I then carried out the reciprocal process of mapping both sets to the EATRO 3 genome assembly and calling putative SNPs for the EATRO 3 half of each pair. I then compared the putative SNPs called in both genes of the pair, and annotated a SNP only if an identical SNP had been called at the equivalent position in both. This last step gave a further advantage to Method 2 over Method 1 because the step would reduce the number of SNPs called due to incorrect pairing of EATRO 3 and EATRO 2340 genes: the method required a SNP to be called at the same place in both genes of the pair, and this is unlikely to have happened if the genes being compared were not truly the same gene.

This analysis yielded 446 substitutions in 136 pairs, compared with 2213 substitutions in 203 pairs from the first analysis, with Method 1. This second, more stringent analysis was intended to decrease the chance of SNPs being called in regions that were different due to misassembly (and hence the quality of mapping not being sufficient to call SNPs), and generally to provide an extra layer of quality control. The new analysis predicted the genotype of each isolate at the SNP position, usually with one being a heterozygote and the other homozygous for a reference allele, and this genotype prediction could be tested by looking at the sequences recovered. Within the 14 resequenced gene pairs, this analysis called 18 SNPs, allowing 28 genotype predictions to be tested (eight predictions were in a gene that had no resemblance to the expected sequence). Ten of these predicted genotypes were homozygous: in nine cases only the predicted allele was found, but in one a second allele was found that matched the 'SNP' predicted at that position in the other genome. The remaining 18 were heterozygous: in five cases, both predicted alleles were found; in eight cases (all in the same gene), only one of the predicted alleles was found, but it was the one not already represented in the reference gene; in five cases, only the allele in the reference gene was found. The results from Method 2, therefore, gave SNP predictions that had a reasonable degree of support from the data, although these predictions were not 100% correct.

The Method 2 results were not directly comparable with the results of the first attempt at cataloguing SNPs by directly comparing the assembled sequences of the genes in each genome (Method 1), as the outputs were different. Additionally, the resequencing experiments, designed to test the results from Method 1, did not provide many data to test Method 2. However, comparison of the results from Method 2 with the available resequencing data indicated that looking for SNPs in paired genes was a more reliable method than simply looking at the assembled sequence as in Method 1, although neither method was without false positives. The strategy in Method 2 has certain limitations: the program used to annotate SNPs from the read data is unreliable at calling indels, so Method 2 can only be used to study substitutions; and Method 2 will also be poor at detecting segmental conversions donated by less closely-related genes, because the mapping algorithm will discard a read that contains more than a certain number of differences from the reference, and so no SNPs will be called in the converted sequence. However, this second point could potentially be addressed to some extent by looking for regions that show considerably decreased coverage compared with the rest of the sequence. In general, Method 2 will probably generate more false negatives than the assembled sequence-based approach in Method 1, because of the described problems calling indels and segmental conversions, and because the more stringent filters will likely exclude more true SNPs than Method 1.

In summary, there appeared to be considerably more diversity within each genome than initially appeared from the assembly, which made detection of changes from EATRO 3 to EATRO 2340 more difficult. Without substantial further resequencing, which was not feasible in this project, it was not possible to fully evaluate different strategies for analysing SNPs, or to find a completely reliable method. However, I developed a strategy (Method 2, Figure 4-4) involving finding SNPs in the annotated genes using PCR-free reads, which appeared to have reasonable reliability. I therefore used this strategy to annotate SNPs to provide data for more detailed analysis of the substitutions that occur in *VSG* genes, with the caveat that there may be some incorrectly annotated SNPs.

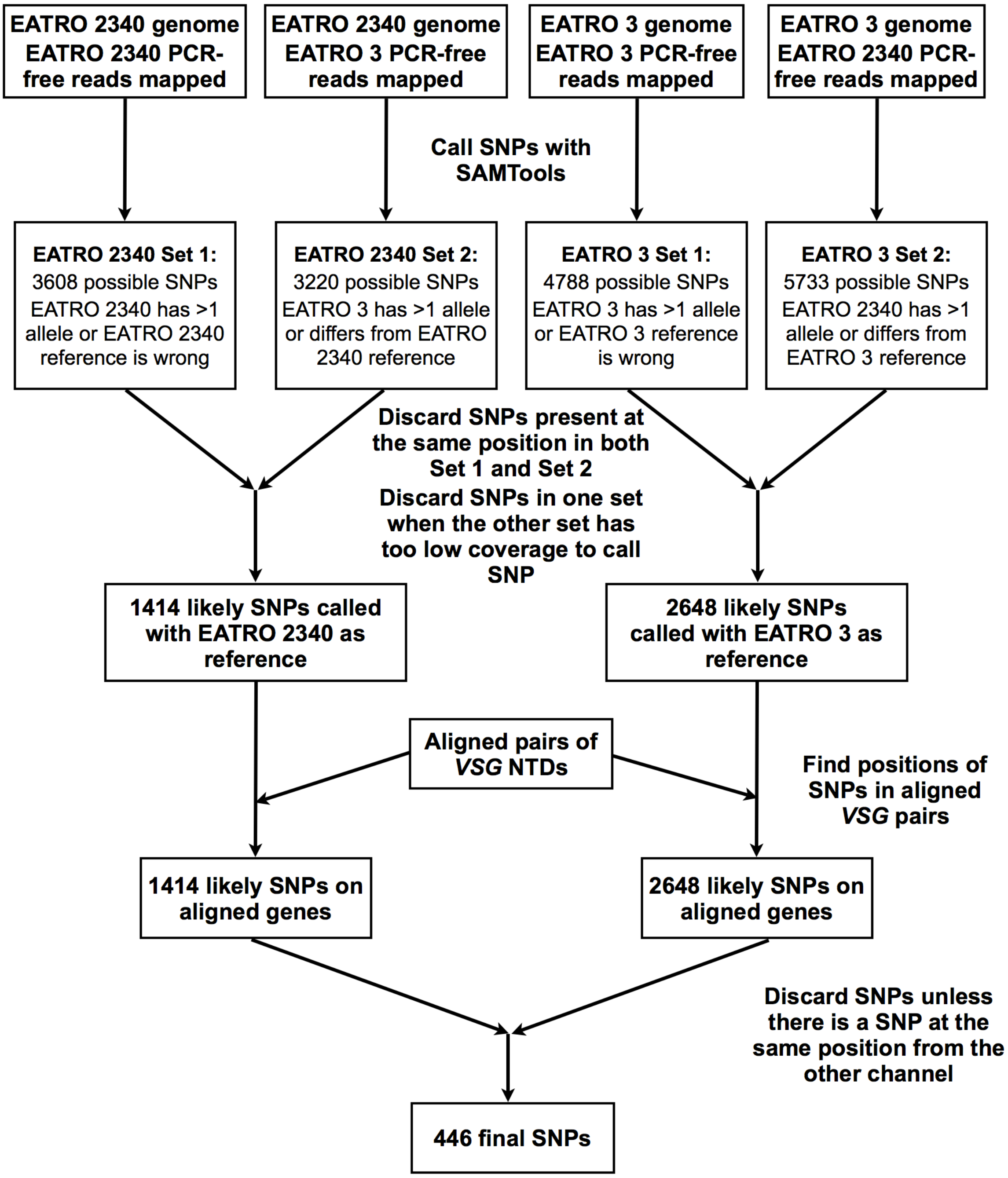

EATRO 2340 genome EATRO 2340 PCR-free reads mapped
EATRO 2340 genome EATRO 3 PCR-free reads mapped
EATRO 3 genome EATRO 3 PCR-free reads mapped
EATRO 3 genome EATRO 2340 PCR-free reads mapped
Call SNPs with SAMTools
EATRO 2340 Set 1: 3608 possible SNPs EATRO 2340 has >1 allele or EATRO 2340 reference is wrong
EATRO 2340 Set 2: 3220 possible SNPs EATRO 3 has >1 allele or differs from EATRO 2340 reference
EATRO 3 Set 1: 4788 possible SNPs EATRO 3 has >1 allele or EATRO 3 reference is wrong
EATRO 3 Set 2: 5733 possible SNPs EATRO 2340 has >1 allele or differs from EATRO 3 reference
Discard SNPs present at the same position in both Set 1 and Set 2
Discard SNPs in one set when the other set has too low coverage to call SNP
1414 likely SNPs called with EATRO 2340 as reference
2648 likely SNPs called with EATRO 3 as reference
Aligned pairs of VSG NTDs
Find positions of SNPs in aligned VSG pairs
1414 likely SNPs on aligned genes
2648 likely SNPs on aligned genes
Discard SNPs unless there is a SNP at the same position from the other channel
446 final SNPs

Figure 4-4 – Outline of steps in SNP annotation Method 2, showing the order of steps, the input data, and the number of SNPs remaining at each step.

4.4.3 Mutation analysis pipeline

In order to process a large number of gene pairs, I developed an object-oriented Perl pipeline. This set of programs extracted the sequences of each gene in the pair and aligned them for use with the SNP-finding strategies described in the previous section. A final step grouped together single-base changes if they seemed likely to represent a single mutation event (*e.g.* substitution of multiple adjacent bases simultaneously), to allow later estimation of the frequency of mutational events as well as the frequency of mutated bases.

4.5 Substitution mutations in *VSG* evolution

4.5.1 Mutation frequency

I used the strategy described above to search for substitution mutations in the *VSG* NTDs and in the control core genes, and detected 446 substituted bases in the *VSG* NTDs and 1781 in the core genes. Some substituted bases were immediately adjacent to each other and so were grouped together as being likely to represent a single mutation event (four pairs in the *VSG* and 12 pairs in the core genes), giving a total of 431 mutations in the NTDs and 1769 in the core genes. These data yield a mutation frequency of 3.97×10^{-4} mutations/base in the *VSG* NTDs and 1.67×10^{-4} mutations/base in the core genes, which indicates a 2.4-fold higher mutation rate in *VSG* NTDs than in core genes; and the difference was significant ($p < 0.001$), assuming a straightforward binomial distribution. These numbers were calculated by dividing the number of SNPs by the total length of the genes in gene pairs; however, not every base was suitable for calling SNPs, for example due to low coverage. I therefore determined the number of bases in total that were suitable for calling SNPs, *i.e.* the coverage with all four genome-read pairings was at least 10 and not greater than 200 reads with good-quality bases. When the frequencies were calculated based on the number of SNPs per base suitable for SNP calling, the frequencies were 4.79×10^{-4} mutations/base in the *VSG* and 1.69×10^{-4} mutations/base in the core genes, a ratio of 2.8 (Table 4-2).

Region	Total positions changed	Total no. changes	Total gene length (bp)	Substitutions per base	Substitutions/base/generation	
					Min	Max
Core	1781	1769	10581389	0.000167	2.0×10^{-9}	6.7×10^{-9}
***VSG* NTD**	446	442	1112229	0.000397	4.7×10^{-9}	1.6×10^{-8}

Table 4-2 – Mutation frequencies and rates in *VSG* NTDs and in core genes.

When I examined the number of substituted bases per gene in *VSG* NTDs, it seemed possible that there were more genes with no substitutions than would be predicted from the mutation frequency if the occurrence of every substitution were random (*i.e.* the data were overdispersed) (Figure 4-5).

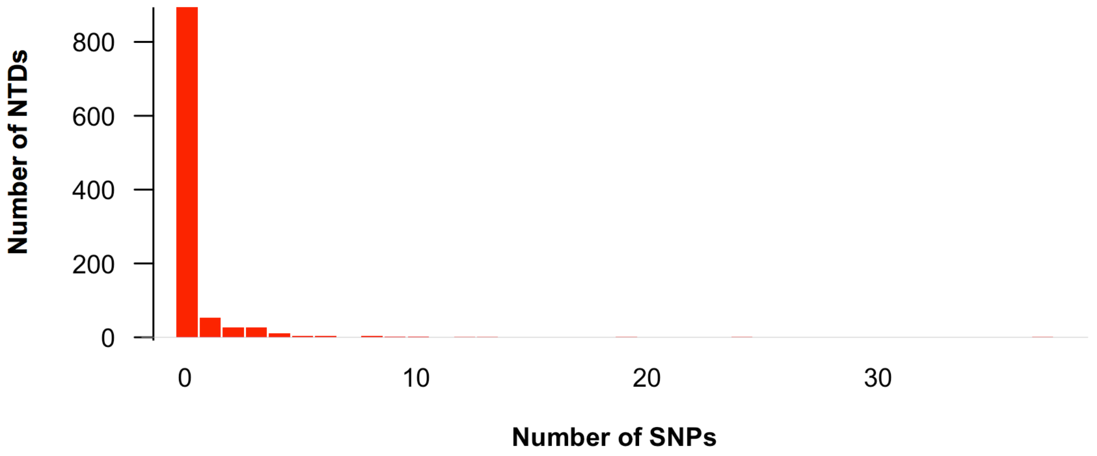


Figure 4-5 – Number of SNPs per gene in *VSG* NTDs.
The data are overdispersed, *i.e.* there were more NTDs with no new mutations than would have been expected if the occurrence of mutations was random.

I therefore used R to model the data of number of SNPs per gene to two distributions, a Poisson and a negative binomial distribution. The models also incorporated length, because the number of mutations in a gene would be expected to depend on the number of nucleotides in a gene as well as the mutation rate (*i.e.* SNPs depends on number of bases as well as SNPs/base). The log likelihood of the Poisson model was -1287, whereas the log likelihood of the negative binomial model was -665, indicating that the negative binomial model described the data much better than did the Poisson model, and that the assumptions of the Poisson distribution likely did not hold in the process producing the data. The most likely reason for these data not fitting the Poisson distribution is that the mutations did not occur entirely independently of one another. The negative binomial model predicted that the expected number of SNPs for a gene of length 1000 bp was 0.56. This length was chosen because it was a typical length for the NTDs examined, and hence the region where the model was most likely to be a good reflection of the data. This expected value corresponds to 5.6×10^{-4} mutations/base, which is reasonably close to the value of the mean number of 3.97×10^{-4} mutations/base calculated directly from the data.

Since the negative binomial model described the *VSG* data reasonably well, I used a similar model to test whether the mutation rates in the *VSG* NTDs and core genes differed significantly. I used R to produce a single model of the number of SNPs per gene in both regions (subtelomere and core) using a negative binomial distribution, again incorporating the length of gene. This analysis indicated that the region that a gene came from was a highly significant determinant of the number of SNPs it contained ($p < 0.0001$), after gene length had been taken into account. For a gene of length 1000, the expected number of SNPs was 0.407 for *VSG* and 0.180 for core genes, indicating that the mutation rate in *VSG* NTDs was 2.41-fold higher than in cores, which was very close to the value of 2.37-fold calculated directly from the data.

4.5.2 Distribution of mutations between lineages

The mutations annotated by this approach have been more stringently checked than those in the larger set discussed in Chapter 4. The new mutation annotations can therefore can be used to obtain a more reliable estimate of the relationship between the two isolates, by looking again at the distribution of new mutations. For the reasons discussed in section 3.4.3, genes from the anomalous regions of chromosomes 4 and 10 were excluded from the analysis. Excluding genes from these regions gives mutation frequencies of 3.97×10^{-4} substitutions/base in the *VSG* NTDs and 8.81×10^{-5} substitutions/base in the core genes. The EATRO 2340:EATRO 3 ratio was 3.5 for core genes and 2.8 for *VSG* NTDs (Table 4-3). These values imply approximate times to the MRCA of seven years for EATRO 3 and 24 years for EATRO 2340 (using the core genes ratio), or nine years for EATRO 3 and 26 years for EATRO 2340 (using the *VSG* NTDs ratio), with divergence dates of 1953 and 1951 respectively. The values for the two genome regions therefore seem acceptably similar. Using the divergence date estimated for the core genes, the mutations observed have accumulated over 31 years, which corresponds to approximately 45000 generations and gives mutation rates of 8.82×10^{-9} substitutions/base/generation for *VSG* NTDs and 1.96×10^{-9} substitutions/base/generation for core genes. This core mutation rate is a credible value since it is reasonably consistent with a previously estimated minimum mutation rate of approximately 10^{-9} substitutions/base/cell/generation (Valdes *et al*, 1996).

Region	EATRO 3	EATRO 2340	EATRO 2340: EATRO 3 ratio	Total subs	Mutation frequency (subs/base)	Mutation rate (subs/base/ generation)
Core	188	665	3.54	853	0.000088	1.96×10^{-9}
***VSG* NTDs**	111	309	2.78	435	0.000397	8.82×10^{-9}

Table 4-3 – Distribution of new mutations in core genes and *VSG* NTDs between EATRO 3 and EATRO 2340 lineages.
Genes in anomalous regions of chromosomes 4 and 10 have been excluded from the analysis. Note that the total number of mutations includes those for which both isolates were homozygous and so it could not be determined which lineage the mutation occurred in. subs = substitutions.

It should be noted that excluding genes from the anomalous regions of chromosomes 4 and 10 substantially decreased the observed mutation frequency of core genes, but had very little effect on *VSG* NTDs (since only 11 of the latter were excluded from the analysis). The estimated *VSG* NTD substitution frequency thus became 4.5-fold higher than the core substitution frequency rather than 2.4-fold higher. However, this change does not substantially affect the conclusions of the work: the *VSG* NTD frequency is still higher than the core frequency, and the difference between the two regions is still the same order of magnitude. Moreover, presumably the excluded mutations from EATRO 3 core genes arose largely by the same mutational process which gave rise to all the other core gene mutations, albeit over a longer time period than expected, and not in the time since the MRCA of the two isolates. Hence, when the mutation *profiles* are compared (see sections below), the comparison is still between a sample of mutations in the core genes and a sample of mutations in the *VSG* NTDs; the possibility that each sample was accumulated over a different time period is essentially irrelevant.

4.5.3 Specific base changes

I counted the number of each type of substitution mutation in the *VSG* NTDs and core genes, taking the base in EATRO 3 as the starting base. Note that although I examined the bases in the DNA strand containing the coding sequence, the mutation could have occurred in either strand, *e.g.* a mutation that appears to be from A to G in the *VSG* coding strand could in fact have come about by a mutation from T to C in the complementary strand. For this reason, mutations were considered in their complementary pairs. The proportions of each type of substitution appeared quite different in the *VSG* NTD genes from the core genes (Figure 4-6). I analysed these results in R using Pearson's x^2 test of independence, which indicated that the distributions were significantly different (p < 0.0001).

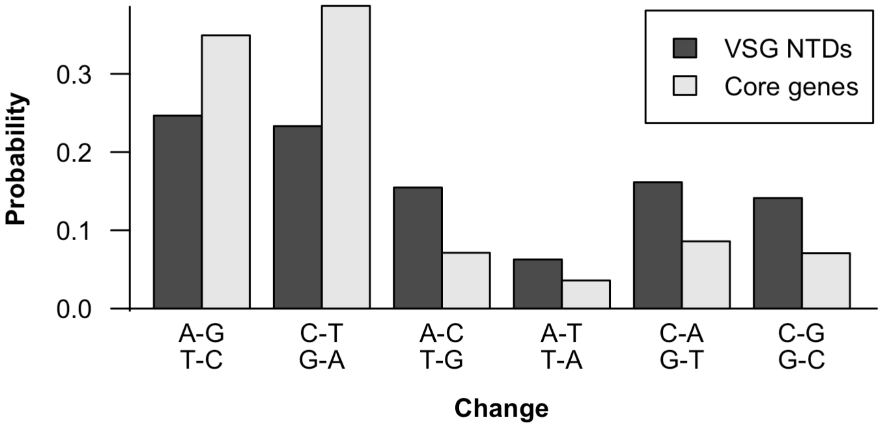


Figure 4-6 – Proportion of each type of substitution mutation in *VSG* NTDs and core genes. The x-axis shows the type of change: Label 'A-G' indicates A in EATRO 3 and G in EATRO 2340. Complementary mutations (*e.g.* A-G and T-C) have been grouped together as it was not possible to distinguish between them in the analysis used. *VSG* NTDs are represented in dark grey and core genes in dark grey, as shown in the key.

Analysis of the raw counts of the number of each type of substitution gave a first approximation comparison between the two genome regions, but such an analysis may suffer from bias if the regions differ in composition. For example, if the coding sequences of one region contained a higher proportion of As than the other, then genes in that region would be expected to exhibit a higher proportion of A substitutions even if the probability a given A will mutate was the same in both regions. The comparison described above also assumed that there was no difference between mutations in the coding and the non-coding strands. I therefore determined the base composition of the EATRO 3 sequence in the coding strand of high-identity pairs of *VSG* NTDs and core genes (excluding pairs where there was a sequencing gap in either genome) (Table 4-4). EATRO 3 was used to be consistent with the description of each type of substitution as 'base in EATRO 3' to 'base in EATRO 2340'. The compositions of the coding strands in the two genome regions were revealed to be strikingly different. It would be expected that bases which are A or T would be distributed equally between A and T (*i.e.* P(A|(A or T)) = 0.5), and similarly for C and G. In fact, the A/T and C/G distributions differed significantly from this expectation (p < 0.001) for both core genes and *VSG* NTDs. However, for all but the VSG A/T distribution, the magnitude of the difference from 0.5 was small (<0.05). Importantly, there was a significant difference (p < 0.001) between *VSG* NTDs and core genes for both A/T and C/G distributions. The substantial excess of A over T bases is consistent with observations in TREU 927 *VSG* (L. Marcello, PhD thesis, 2007).

Base	*VSG*		Core	
	Count	%	Count	%
A	381402	34.3	2627855	24.7
C	289790	26.1	2499189	23.4
G	270204	24.3	2947698	27.7
T	169936	15.3	2583071	24.2

Table 4-4 – Composition of EATRO 3 genes in coding strands of paired *VSG* NTDs and core gene sequence.
All probabilities were significantly different from 50% (p < 0.001), assuming a binomial distribution between A and T for all A/T positions, and C and G for all C/G positions.

As mentioned above, it was impossible to tell whether (for example) an apparent A to G mutation in the coding strand was the result of an actual A to G mutation in the coding strand, or the result of a T to C mutation in the non-coding strand. However, the primary aim of comparing the mutation profiles between *VSG* NTDs and core genes was to determine whether the two profiles were different. If the two mutation profiles were identical, and the compositions of the two genome regions were the same, then the proportion of mutations that were A to G in the coding strand would be expected to be the same in both regions, as would the proportion of mutations that are T to C in the non-coding strand. Therefore, if the mutations are assumed for simplicity always to have taken place in the coding strand, the proportion of apparent A to G mutations should be approximately the same between the two genome regions. If they are not, it can be concluded that the mutation profiles differ in some way, although it will not be possible to tell whether they differ in the A to G substitution rate in the coding strand, or the T to C substitution rate in the non-coding strand, or both.

A similar idea can be applied to allow for the possibility that the assumption that all mutations are in the EATRO 3 to EATRO 2340 direction was an oversimplification. If the MRCA of EATRO 3 and EATRO 2340 was not EATRO 3 then the reported mutation profile may not be correct, because, for example, some of the mutations annotated as A to G may actually be G to A mutations that have occurred in the EATRO 3 lineage since the MRCA. However, if the core and *VSG* mutation profiles were identical, the number of each (A to G in the MRCA-EATRO 2340 lineage and G to A in the MRCA-EATRO 3 lineage), and hence the apparent A to G totals would be expected to be similar in the core genes and the *VSG* NTDs. If the totals are significantly different, again it can be concluded that the mutation profiles are not identical, although the specific difference will not be clear.

The concept that allows simplification of the analysis to consider only one strand can be extended to allow comparison between the *VSG* and core substitutions with correction for the different compositions of the two regions. The reasoning supporting this statement can be expressed as an example with the A to G in the coding and T to C in the non-coding strand:

expected n A to G mutations = AG = P(AG) . nA, where P(AG) is the probability any given A will mutate to G, and nA is the number of As in the sequence
Let $_c$ denote a parameter for the coding strand and $_{nc}$ a parameter for the non-coding strand.
apparent AG_c = $AG_c + TC_{nc}$
= $P(AG)_c \cdot nA_c + P(TC)_{nc} \cdot nT_{nc}$
but $nT_{nc} = nA_c$ as the two strands are complementary
so apparent AG_c = $nA_c\ (\ P(AG)_c + P(TC)_{nc}\ )$
$P(AG)_c + P(TC)_{nc}$ = (apparent AG_c) / nA_c
Hence if (apparent AG_c) / nA_c for the *VSG* does not equal (apparent AG_c) / nA_c for the core genes, the *VSG* and core mutation profiles are different.

A similar approach was used to generate 'corrected' numbers of each type of substitution, *i.e.* the number of each type that would be expected in the *VSG* NTD genes if they had the same base composition as the core genes:

Using the equation
$P(AG)_c + P(TC)_{nc}$ = (apparent AG_c) / nA_c
we can calculate $P(AG)_c + P(TC)_{nc}$ from the observed data, and use it in the equation
apparent AG_{coding} = $nA_c\ (\ P(AG)_c + P(TC)_{nc})$
to calculate the expected value of apparent AG_c if nA_c changes.
If the *VSG* sequences had the same base composition as the core sequences
$nA_{c,\ VSG}$ / $total_{VSG}$ = $nA_{c,\ core}$ / $total_{core}$
so in this case, for the *VSG*

$$\text{corrected } AG_c = (P(AG)_c + P(TC)_{nc}) \,.\, nA_{c,\,core} / total_{core} \,.\, total_{VSG}$$
$$= (\text{apparent } AG_{c,\,VSG}) / nA_{c,\,VSG} \,.\, nA_{c,\,core} / total_{core} \,.\, total_{VSG}$$

This process allowed the ‘corrected’ numbers to be compared with the observed numbers of each type in the core genes using a x^2 test, to determine whether there was a significant difference between the mutation profiles of core and *VSG* NTDs. Again, this comparison will not have revealed precisely what the differences are. The ‘corrected’ data are given in Table 4-5 and plotted in Figure 4-7. I re-analysed these new data in R using Pearson’s x^2 test of independence, which again indicated that the distributions were significantly different ($p < 0.0001$). This result implies that mutation is qualitatively as well as quantitatively different in *VSG* NTDs and core genes.

Base change (EATRO 3 to EATRO 2340)	Raw number	Corrected number (to core composition)	Corrected number (to equal base frequency)
A-C	45	32	33
A-G	80	57	58
A-T	11	8	8
C-A	56	50	54
C-G	37	33	35
C-T	31	28	30
G-A	73	83	75
G-C	26	30	27
G-T	16	18	16
T-A	17	27	28
T-C	30	48	49
T-G	24	38	39

Table 4-5 – Counts of each mutation type with *VSG* data adjusted for A-richness.

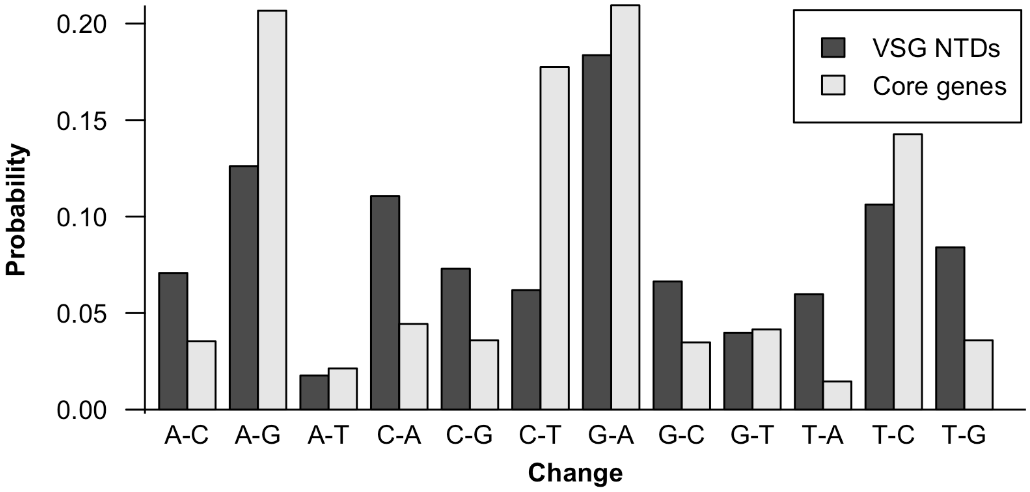


Figure 4-7 – Proportion of each type of substitution mutation in *VSG* NTDs and core genes with *VSG* data adjusted for A-richness.
Label 'A-G' indicates A in EATRO 3 and G in EATRO 2340 on the coding strand. *VSG* NTDs are represented in dark grey and core genes in dark grey, as shown in the key.

4.5.3.1 Strand bias

The asymmetry of composition of *VSG* genes, as observed previously (L. Marcello, PhD thesis, 2007), suggested that there may be a difference between the mutation processes that act on the coding and on the non-coding strands. Therefore, to examine the *VSG* NTD mutations for strand bias I generated a second set of corrected mutation frequencies, this time assuming that all bases occurred with equal frequency (Table 4-5). I then examined the frequency of complementary mutations, for example A-G and T-C. If there were no strand bias, *i.e.* both strands had the same probability of each mutation, then complementary mutations should be approximately equal: the apparent A-G on the coding strand is the sum of the actual A-G on coding strand plus the actual T-C on the non-coding strand, and the apparent T-C on the coding strand is the sum of the actual T-C on the coding strand plus the actual A-G on the coding strand. This was not the case in the mutations found in the *VSG* genes: there was considerable asymmetry in several cases, with a particularly noticeable bias towards mutations generating As in the coding strand (Table 4-6). Further, the number of mutations generating As was 1.6-fold higher than the number of As mutating to other bases, and the of number Ts mutating to other bases was 2.1-fold higher than the number of mutations generating Ts, an observation that was consistent with the A-richness of the coding strand of the *VSG* NTDs. Some asymmetry was seen in the core gene mutations (which were not corrected, because the core gene base composition was approximately even), but not to anything like the extent of the *VSG*s.

	Change	Number	Complement	Number	Ratio
VSG (frequencies corrected)	**AC**	33	**TG**	39	0.84
	AG	58	**TC**	49	1.19
	AT	8	**TA**	28	0.29
	CA	54	**GT**	16	3.26
	CG	35	**GC**	27	1.33
	CT	30	**GA**	75	0.4
Core genes	**AC**	63	**TG**	64	0.98
	AG	368	**TC**	254	1.45
	AT	38	**TA**	26	1.46
	CA	79	**GT**	74	1.07
	CG	64	**GC**	62	1.03
	CT	316	**GA**	373	0.85

Table 4-6 – Frequency of substitutions displayed in pairs corresponding to an identical mutation occurring in the coding strand or the non-coding strand.

4.5.4 Effects on predicted protein sequence

Substitution mutations are presumably only of functional significance to the parasite’s antigenic variation if they result in a change to what is presented to the immune system, *i.e.* a different amino acid in an expressed VSG. I therefore examined whether the substitution mutations detected would change the protein sequence. *VSG* genes often include frameshifts, an observation which was borne out in my data by the fact that it was often necessary to use HMMer or BLASTX hits in several frames to annotate a full-length *VSG* NTD (see section 4.2). The frequency of frameshifts meant that it was not straightforward to predict the ‘translation’ of a gene: we were interested in the amino acid sequence that the gene would contribute to a functional VSG, rather than necessarily a full translation of the first frame of the gene. A functional VSG can be achieved by assembly of a mosaic gene from silent donor *VSG* genes or pseudogenes, and the final polypeptide will not necessarily be the same as the amino acid sequence obtained by translating the domain in a single frame from start to finish. However, I was able to use the information from HMMer and

BLASTX hits during annotation to predict a 'VSG translation' for almost every NTD, which I then used to examine the effects of substitutions.

For each gene pair with a substitution mutation, I examined the predicted protein sequence of the gene to find the codon affected by the mutation, and compared the amino acids encoded by the two alternative codons at this position to determine whether the substitution was synonymous or non-synonymous; and determined the total number of synonymous and non-synonymous sites in paired *VSG* NTDs. (The results given were calculated with the assumption that EATRO 3 contained the ancestral base, but redoing the analysis with the assumption that EATRO 2340 contained the ancestral base made very little difference to the results.) I also carried out a similar analysis for mutations in the core genes. The results of these analyses (Table 4-7) revealed that there were a considerable number of non-synonymous mutations in both *VSG* NTDs and core genes. For both genome regions, I estimated confidence intervals (CI) for the number of mutations per site for which every possible mutation was non-synonymous (d_{N3}, *i.e.* 3 mutations are non-synonymous), and the number of mutations per site for which every possible mutation was synonymous (d_{S3}), because the mutation frequency at both these types of site is essentially a binomial distribution. Importantly, d_{S3} for the *VSG* (7.9×10^{-4}/site ± 2.0×10^{-4}, 99.9% CI) was significantly higher ($p < 0.001$) than d_{S3} for the core genes (3.0×10^{-4}/site ± 4.0×10^{-5}). The mutation rate at fourfold degenerate sites (considered here as d_{N3}) is frequently taken as a measure of the background rate of mutation. This result therefore strengthened the conclusion that the rate of substitution mutation was genuinely higher in *VSG* NTDs than in core genes, and was not just an artefact of relaxed selection on *VSG*s. The *VSG* d_{N3} value was also significantly higher than that of the core genes (VSG: 3.0×10^{-4}/site ± 6.4×10^{-5}, 99.9% CI; core: 1.1×10^{-4}/site ± 1.2×10^{-5}), and in fact the difference between the two was greater than for the d_{S3} values, but the core value may have been decreased because of selection acting to remove fatal or strongly deleterious mutations, so the significance of these values in terms of differences between the mutation processes in core genes and *VSG* cannot really be evaluated.

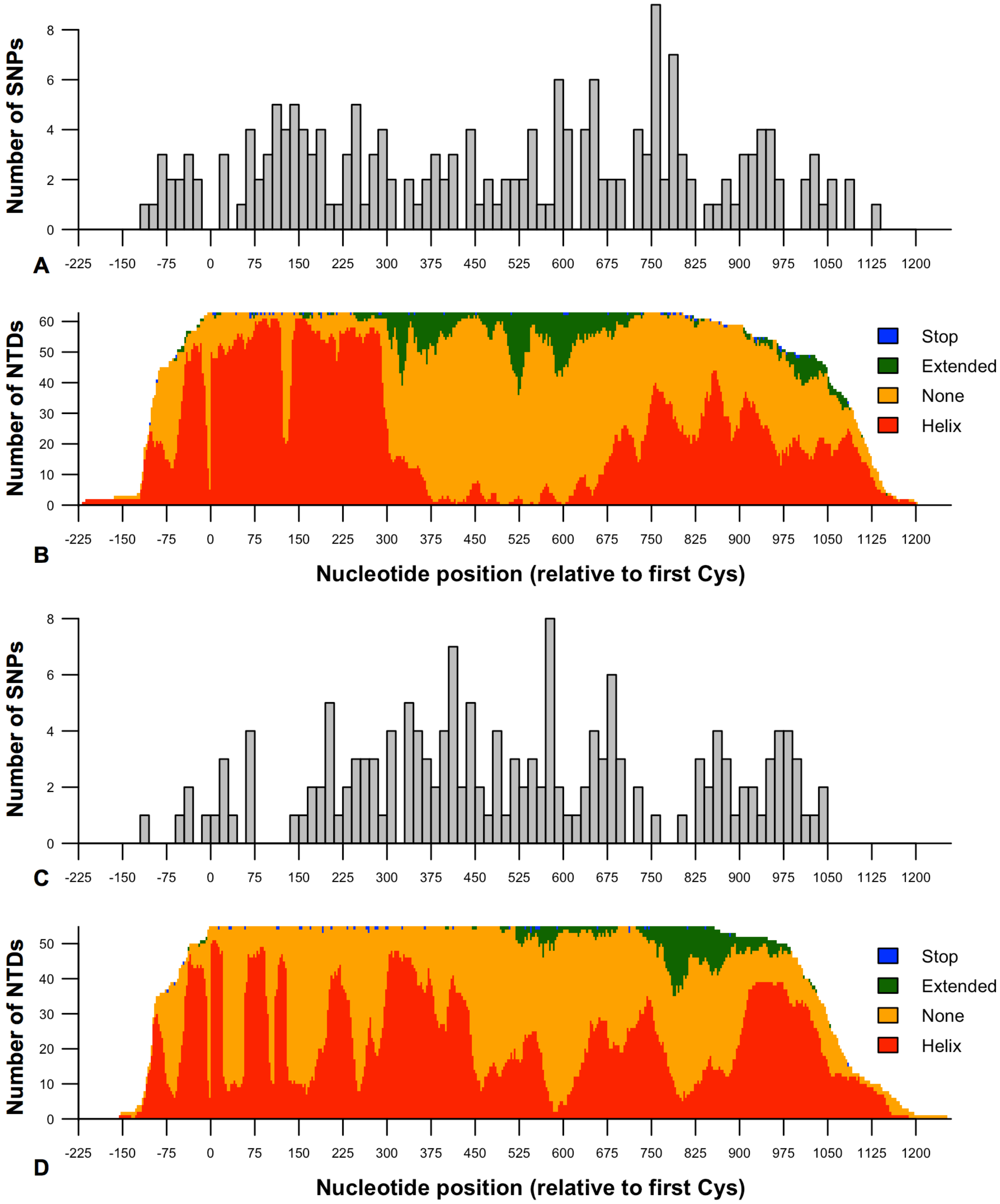
Region	***VSG***	**Core**
Non-synonymous changes	254	852
Non-synonymous sites	840942	8169245
Synonymous changes	189	917
Synonymous sites	256479	2487988
Total changes	442	1769
d_N	$3.02{\times}10^{-4}$	$1.04{\times}10^{-4}$
d_S	$7.37{\times}10^{-4}$	$3.69{\times}10^{-4}$
d_N/d_S	0.41	0.28

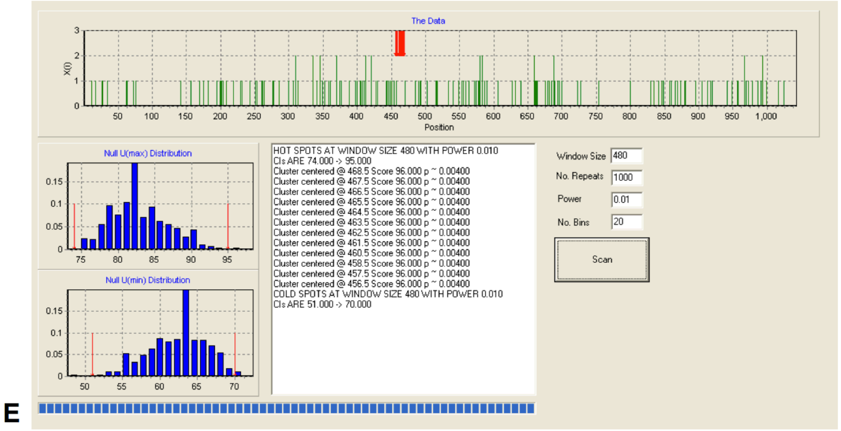
Table 4-7 – Synonymous and non-synonymous changes in VSG NTD and core protein coding genes.
Only gene pairs where neither partner contained an assembly gap were considered. Note that, as is standard practice, sites where some possible changes are synonymous and some are non-synonymous have been counted according to the proportion of each, *e.g.* if one of the possible changes is synonymous and two are non-synonymous, the site will add 0.33 to the synonymous site total and 0.67 to the non-synonymous site total.

In studying sequence evolution, it is often useful to consider the ratio between d_N, the number of non-synonymous mutations per non-synonymous site, and d_S, the number of synonymous mutations per synonymous site. In the case of the *VSG* NTDs, the d_N/d_S ratio (based entirely on counting sites and mutations) was 0.4 (Table 4-7) (and the d_{S3} was significantly higher than the d_{N3}, $p < 0.05$). Usually such a d_N/d_S value would be taken to mean that on average the *VSG* genes were under weak purifying selection. Given that the function of VSG essentially is to be diverse, this conclusion seems highly unlikely. There are several factors that can be considered to explain this anomalous result. Firstly, the time-scale under consideration, only a few decades, is very short in evolutionary time. Originally, studies of d_N/d_S were used to detect selection on particular nucleotide sites by looking at several species, with the implicit assumption that the polymorphisms under consideration had gone to fixation in their respective species. Various studies have concluded that d_N/d_S is not a reliable indicator of selection in very closely-related lineages, and has a complex and unclear dependence on divergence time as well as direction and strength of selection (Rocha *et al*, 2006; Kryazhimskiy & Plotkin, 2008; Peterson & Masel, 2009). Secondly, the *VSG* genes are not considered to be under direct selection (Barry *et al*, 2012), and the effects of the true selective pressure, to have a diverse VSG repertoire, would likely be expressed in more subtle ways than a high d_N/d_S ratio. Therefore, the d_N/d_S ratio is not a particularly relevant statistic to *VSG* evolution, and for this reason, it seems unnecessary to consider more sophisticated approaches to estimate the d_N/d_S ratio, for example by using models that take into account codon bias and transition/transversion bias.

4.5.5 Distribution

To examine the distribution of SNPs across the *VSG* NTDs, I found the position of each *VSG* SNP relative to the first cysteine codon of the gene containing it. Figure 4-8A and C show the distribution of SNPs in A-type and B-type NTDs. To test for clustering within the gene, I used the program Permute III, provided by Prof Dan Haydon, which uses permutation to find windows with significantly higher (hot spots) or lower (cold spots) numbers of events than expected. Because different genes had different sequence lengths before the first cysteine, for the Permute analysis I considered only SNPs after this position. I tested for clustering in 10 bp windows and then in windows from 30-600 bp at 30 bp intervals, using 1000 permutations with a confidence level of 0.01. No hot spots or cold spots were found in A-type NTDs using this analysis. In the B-type NTDs, there was evidence of clustering: window sizes from 360-570 bp found hot-spots centred around positions from 415-520 bp after the cysteine codon, with larger window sizes having centres more 5′ in the gene (example in Figure 4-8E).





```
HOT SPOTS AT WINDOW SIZE 480 WITH POWER 0.010
CIs ARE 74.000 -> 95.000
Cluster centered @ 468.5 Score 96.000 p ~ 0.00400
Cluster centered @ 467.5 Score 96.000 p ~ 0.00400
Cluster centered @ 466.5 Score 96.000 p ~ 0.00400
Cluster centered @ 465.5 Score 96.000 p ~ 0.00400
Cluster centered @ 464.5 Score 96.000 p ~ 0.00400
Cluster centered @ 463.5 Score 96.000 p ~ 0.00400
Cluster centered @ 462.5 Score 96.000 p ~ 0.00400
Cluster centered @ 461.5 Score 96.000 p ~ 0.00400
Cluster centered @ 460.5 Score 96.000 p ~ 0.00400
Cluster centered @ 458.5 Score 96.000 p ~ 0.00400
Cluster centered @ 457.5 Score 96.000 p ~ 0.00400
Cluster centered @ 456.5 Score 96.000 p ~ 0.00400
COLD SPOTS AT WINDOW SIZE 480 WITH POWER 0.010
CIs ARE 51.000 -> 70.000
```

Window Size: 480
No. Repeats: 1000
Power: 0.01
No. Bins: 20

Figure 4-8 – Distribution of SNPs and predicted secondary structure across *VSG* NTDs.

Figure 4-8 continued. A) SNP positions in A-type NTDs (193 genes, 179 SNPs), showing number of SNPs at each relative position, split into 15-bp bins. B) Jpred 3 predicted secondary structure in the predicted VSG translations of A-type NTDs, showing number of NTDs with each type of structure prediction at each position. As shown by the key, blue indicates a predicted stop codon, green indicates an extended (β-sheet) structure, yellow indicates no predicted secondary structure, and red indicates α-helical structure. Codon positions have been converted to nucleotide positions, so each line is 3 bp wide. C) SNP positions in B-type NTDs (162 genes, 157 SNPs), showing number of SNPs at each relative position, split into 15-bp bins . D) Jpred 3 predicted secondary structure in the predicted VSG translations of B-type NTDs, showing number of NTDs with each type of structure prediction at each position, colours as in part B. E) Example screenshot from a run of Permute III using a window size of 480 bp, showing clusters of SNPs in B-type NTDs, and scaled to approximately match parts A-D. Centres of hot-spot regions are indicated by red arrows. In all figures, position 0 corresponds to the first base of the first cysteine codon in the translation. Ten genes containing SNPs did not have a cysteine residue with 75 aa of the start of the sequence, and so were not considered, and a further 60 genes were of unknown type. Note that the plot in the screenshot (E) is of the number of SNPs at each position, whereas A and C use the frequency in 15-bp non-overlapping windows, and B and D use the frequency in 3-bp overlapping windows (since the structure predictions are converted from codon positions).

To investigate whether the observed region of clustering corresponded to a particular structural feature, I used the Jpred 3 server to make secondary structure predictions for every A-type and B-type predicted VSG NTD sequence that contained a SNP (Figure 4-8B and D). Most regions were either predicted to form alpha helix (red in Figure 4-8) or to have no structure (orange in Figure 4-8). In the A-type NTDs, the location of the predicted helices and loops corresponded reasonably well with what would be expected from published VSG sequences and structures (Freymann *et al*, 1990; Blum *et al*, 1993; Carrington & Boothroyd, 1996), with helices predicted at either end of the domain and unstructured loops in the middle region. The pattern in B-type NTDs was less clear, although this is likely to have been simply because the structures lined up less well than the A-type structures, perhaps due to more length variation. In any case, there did not appear to be a strong correspondence between higher frequency of SNPs (Figure 4-8C) and either secondary structure type. It is notable that the A-type NTDs, where regions corresponding to the loops and helices were discernible, had no significant clustering of SNPs. It therefore seems unlikely that the clustering of SNPs in B-type NTDs, although statistically robust, has any particular biological significance.

4.5.6 Comparison of mutation rate in *VSG* NTDs with mutation rate across the subtelomeres

To examine whether *VSG* NTD genes specifically had higher mutation frequencies than core genes, or whether an elevated mutation frequency was a general feature of subtelomeres, I compared the mutation frequency in *VSG* NTDs with that across all sequence in the subtelomeres. As described above, for a large number of *VSG* NTDs, the same gene could be identified in both genome assemblies. However, it was not possible to repeat that approach to pair up all sequences across the subtelomeres. The strategy of double-checking mutations in *VSG* genes by examining SNP calls to the same gene in both genomes could therefore not be used for verification of mutations across the subtelomeres. This strategy (Method 2 described in section 4.4.2) was used to ensure that the SNPs studied in detail were as accurate as possible, and false positives were as few as possible. However, to compare the mutation frequency in *VSG* NTD genes with the mutation frequency across the subtelomeres, a less stringent method was acceptable, as there was no reason to think that false positive SNPs would be a *VSG*-specific phenomenon.

I therefore examined the substitution mutation frequencies in the EATRO 2340 genome assembly using the strategy of considering all good-quality SNPs called from mapping reads, as described in section 3.4.1. To calculate the mutation frequency, I counted the number of SNPs in the *VSG* NTD genes, and then in the assembled subtelomeres in chromosomes plus all the bin contigs on which *VSG* domains had been annotated. The mutation frequencies in *VSG* NTD genes and in all subtelomere sequence (including *VSG*) are shown in Table 4-8. The frequencies for the two regions were very similar, with the frequency in the subtelomeres as a whole being 1.16-fold the *VSG* frequency. To further examine the difference in frequency between *VSG* NTDs and total subtelomere sequence in EATRO 2340, I counted the number of SNPs in 1 kb windows across the subtelomeres (*i.e.* windows approximately the length of a typical NTD), and used R to produce a negative binomial model of the number of SNPs in 1 kb windows and *VSG* NTD genes. Although the subtelomere 1 kb windows would have included *VSG* sequence, there were approximately 10 times as many ‘subtelomere’ windows as annotated *VSG* NTD genes (17156 compared with 1858), so considering all the subtelomere sequence windows would give a reasonable approximation of the picture in non-*VSG* subtelomere sequence. Testing with this model found that region (VSG or subtelomere) was not a significant determinant of mutation SNP count (p > 0.1). These data indicate that the probability of mutation in *VSG* NTD was no different from anywhere else in the subtelomeres.

Region	No. SNPs	Length (bp)	SNPs/base	Ratio to VSG
EATRO 2340 *VSG* NTD	3866	2056405	0.00187998	1
EATRO 2340 all ST	34373	15724393	0.00218596	1.163
EATRO 2340 ST minus *VSG*	30507	13667988	0.00223200	1.187

Table 4-8 – Mutation frequencies in *VSG* NTDs and in all likely subtelomere sequences. ST = subtelomere.

4.6 Pseudogenes and substructuring of the EATRO 3 and EATRO 2340 *VSG* archive

4.6.1 Functionality of NTD coding sequences

I screened each *VSG* NTD sequence for which a translation could be predicted (including those identified only by BLASTN hits) for features likely to render it a pseudogene. The pseudogene criteria were considered sequentially in a pipeline such that if a gene fitted one criterion this gene was given a pseudogene classification and excluded from further checks. The features considered to make an NTD likely to be pseudogene were (in the order checked): lack of an initial ATG codon; an in-frame stop codon; and a likely frameshift (*i.e.* any genes that required several BLASTX or HMMer hits in different frames in order to annotate them). Additionally, NTDs with no CTD predicted by SVM in the near 3′ proximity were classed as ‘fragments’; and NTDs with fewer than four cysteines in the predicted translation, or for which the NTD type could not be determined, were classed as ‘atypical’, following standard practice for *VSG* (Berriman *et al*, 2005; Marcello & Barry, 2007b). I then tested the predicted translations of otherwise intact NTDs for the presence of a signal sequence using SignalP, and classed as ‘atypical’ any that did not have a signal sequence predicted. Because the study did not consider CTDs in detail, and CTD domain boundaries and translations were not annotated, the presence of GPI anchor signal sequences was not examined, even if there was a CTD predicted by SVM in close proximity to the NTD under consideration. The results of the analysis are shown in Figure 4-9, alongside a similar breakdown of TREU 927 *VSG* NTDs, using classifications from VSGdb (Marcello *et al*, 2007).

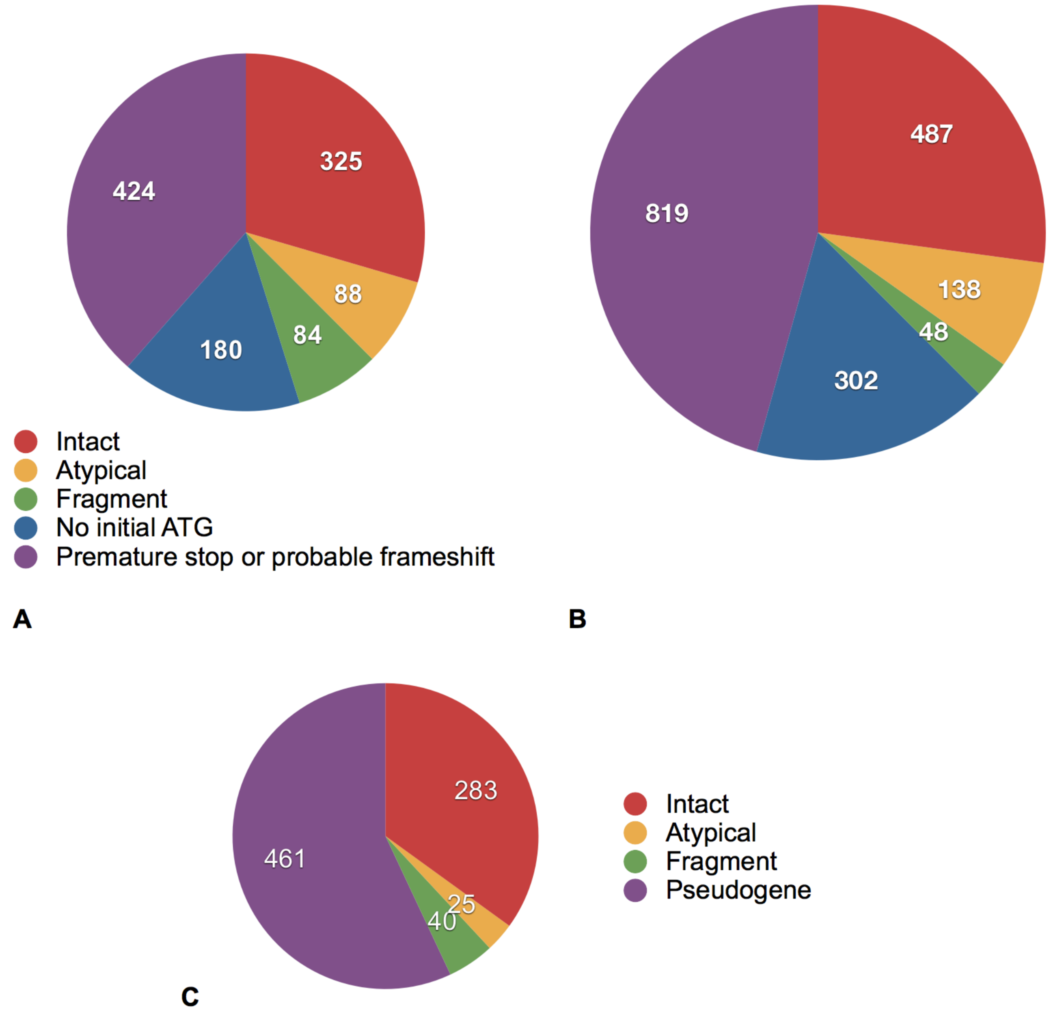


Figure 4-9 – Predicted functionality of *VSG* NTD genes.
A) EATRO 3. B) EATRO 2340. C) TREU 927 (sequences and classifications from VSGdb). Red represents intact NTDs, yellow represents atypical but otherwise intact NTDs, green shows fragments, *i.e.* NTDs for which there is no nearby CTD. In parts A and B, blue represents NTDs with no initial ATG, and purple represents NTDs that are pseudogenes due to a premature in-frame stop codon or a probable frameshift. In part C, purple represents NTDs that are pseudogenes for any reason, including no initial ATG, a premature stop codon or a frameshift. Classifications are detailed further in the text. The number of NTDs in each category is indicated in the appropriate segment.

I then checked the high-identity NTD pairs to see whether any had changed between being classed as intact and being classed as not intact (including atypical NTDs) (Table 4-9). This analysis was intended to give only an idea of the dynamics of functionality: given the problems verifying SNPs and the true sequence of genes, discussed in section 4.4.2, any specific changes in the classification should be treated with caution. I counted separately those gene pairs that had changed because one partner was classed as intact but the other had no close predicted CTD (‘fragments’ in Figure 4-9). Such changes were quite likely to be due to the somewhat more fragmented nature of the EATRO 3 genome assembly, rather than to genuine mutations; 7.6% of NTDs in EATRO 3 were classed as fragments, which was considerably higher than the corresponding figure for EATRO 2340, 2.7%.

Classification in EATRO 3	Classification in EATRO 2340	No. gene pairs
Intact	Intact	192
Pseudogene	Pseudogene	604
Intact	Pseudogene	70
Pseudogene	Intact	27
Intact	Fragment	15
Fragment	Intact	48

Table 4-9 – Functionality classification of genes with a high-identity partner in the other genome.

For most of the gene pairs, either both were intact or both were pseudogenes, but there were some pairs where one gene was intact but the other was a pseudogene. Some changes from intact to pseudogenes were to be expected, given the fact that the parasite can make use of pseudogenes and hence there is not strong selection on individual genes to remain intact. However, a change from pseudogene to an intact gene was much less likely to have occurred by chance, and therefore was of particular interest in this analysis. The gene pairs most likely to represent this scenario were those that were annotated as pseudogenes in EATRO 3 but were predicted to be intact in EATRO 2340; although of course it is possible that in such cases the gene in the common ancestor was intact, and had become pseudogenised in EATRO 3 but not EATRO 2340.

Of the 27 gene pairs that changed from pseudogenes in EATRO 3 to intact in EATRO 2340, 12 had a premature stop codon, 11 had no start codon and four were atypical in EATRO 3. Of these 27 pairs, only four contained validated SNPs. In three of these four cases the call as a pseudogene was not due a change caused by the SNP, but in fact due to slight differences in the annotated gene boundaries (meaning the EATRO 2340 gene contained the ATG, but the EATRO 3 gene did not, although the same sequence was present in both genomes), or to an assembly gap truncating the EATRO 3 gene. In the fourth, changing the allele of the SNP would not have changed the pseudogene call. Therefore, it seems that there was no particular evidence that genes which started out as pseudogenes in the archive had changed so as to regain function.

If it is assumed, instead, that all changes are from intact to pseudogene rather than *vice versa*, then the distribution of such changes between the two lineages should reflect the time of each since the MRCA of the two isolated (as with mutations, as discussed in sections 3.4.3 and 4.5.2). The ratio of the number of genes that were intact in EATRO 3 but pseudogenes in EATRO 2340, to the number of genes that were intact in EATRO 2340 but pseudogenes in EATRO 3 is 2.6:1, a value that is similar to the 2.8:1 calculated for *VSG* NTD substitutions in the more stringent gene-based analysis of distribution (Table 4-3).

4.6.2 *VSG* gene subfamilies

A striking finding of the analysis of the TREU 927 *VSG* archive was that around 40% of *VSG* NTDs were in subfamilies, falling into outlier groups comprising two or more partners with approximately 50% or greater amino acid sequence identity (Marcello & Barry, 2007b). To examine whether such substructuring occurred in the EATRO *VSG* NTDs that I had annotated, I carried out individual pairwise amino acid alignments of predicted VSG sequence for all possible combinations of NTDs within each archive. The resulting distribution of identities indicated that although most pairwise comparisons had low identity, there was an outlier group with pairwise peptide identities above approximately 50% (Figure 4-10). It is also worth noting that the modal pairwise identity for both archives was between 10% and 20%, which is well into the ‘twilight zone’ of protein sequence alignment (less than 20-25% pairwise identity), below which proteins are usually expected not to be homologous (Rost, 1999).

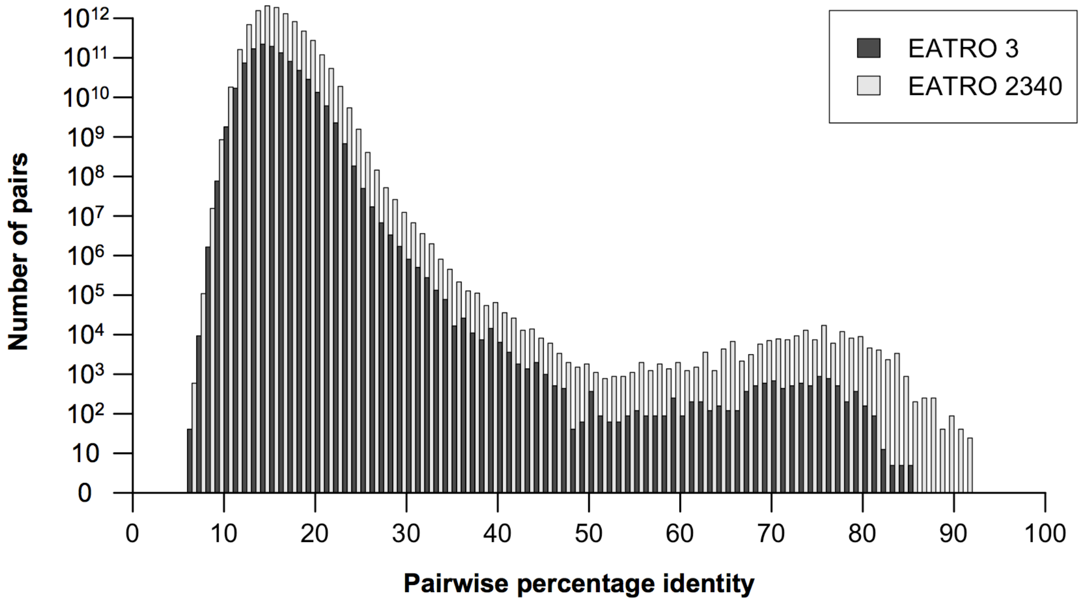


Figure 4-10 – Percentage identities in all pairwise comparisons of predicted NTD translations within each *VSG* archive.
Dark grey represents comparisons within the EATRO 3 NTD archive, light grey represents comparison within the EATRO 2340 NTD archive. Pairwise alignments were carried out with ClustalW. The pairwise percentage identity is shown in bins of 1%. The number of comparisons for each identity bin is shown on a log scale to allow observation of the low-frequency outlier group (percentage identities greater than 50).

To identify subfamilies, I used a combined BLASTP and ClustalW approach. For each genome, I first identified candidate pairs by using the full set of predicted gene translations of *VSG* NTDs to query a protein database made from the same set of sequences. After removing the trivial matches of genes to themselves, for pairs of genes that had a highest-scoring pair with 50% sequence identity I then carried out a ClustalW alignment of the full amino acid sequences, and retained the two genes as a high-identity pair if this second alignment demonstrated at least 50% peptide identity between the two, *i.e.* if the pair fell in the outlier group in Figure 4-10. Pairs of genes were joined together to make larger subfamilies when every member of the subfamily had at least 50% identity with at least one other member (but not necessarily every other member). The revealed subfamilies in each archive are summarised in Table 4-10.

Genome	Genes with no close relatives	No. genes in families with … members							TOTAL
		2	3	4	5	6	7	8	
EATRO 3	622	300	117	44	20	0	0	0	1103
EATRO 3 %	56.4	27.2	10.6	4	1.8	0	0	0	-
EATRO 2340	563	502	333	180	110	72	28	8	1796
EATRO 2340 %	31.3	28	18.5	10	6.1	4	1.6	0.4	-

Table 4-10 – NTD subfamilies in the EATRO 2340 and EATRO 3 *VSG* archives.
Note that the total number of genes considered in this analysis was slightly smaller than the total number of genes annotated, because translations could not be predicted for some genes.

Both genomes contained a large number of genes in subfamilies, although they differed to some extent in the proportion of genes in subfamilies, and the number of genes in each subfamily. Given that we expect EATRO 3 and EATRO 2340 to possess essentially the same *VSG* archive, it was perhaps surprising that they appeared to have somewhat different substructures. However, the subfamilies were numerous and usually had only a few members, and we know that both sets of annotated genes were each likely to be only a subset of the full archive. Therefore, we would expect that whichever gene set samples the true archive more thoroughly would appear to have a larger proportion of its genes in subfamilies, as well as a larger absolute number. This is indeed what was seen in the data: only 31% of the EATRO 2340 NTDs had no high-identity partner, compared with 56% of EATRO 3 NTDs. By the same token, we should be cautious when drawing conclusions about the detailed substructuring of the EATRO *VSG* archive from the annotated EATRO 2340 NTDs, because they too represented only a sample of the full data. A further caveat is that some of the NTDs in the gene sets were annotated based on similarity to the predicted protein sequences of other annotated genes, which would lead us to expect the proportion of genes in subfamilies in the samples to be an overestimate of the real value for the archive. However, the 69% of annotated EATRO 2340 genes that had high-identity partners indicated that this archive has a considerable amount of substructuring. In this respect the EATRO 3-EATRO 2340 archive was similar to the TREU 927 archive. The high mean pairwise identity of the subfamilies (72% in EATRO 2340 subfamilies and 69% in EATRO 3 subfamilies) is also informative, because it implies that the mechanism producing closely related genes, which is presumably gene duplication followed by divergence, is ongoing. However, the main point emerging from this work is the large degree of substructuring in the archive, with its implications for the ease of segmental conversion and mosaic formation as a fundamental mechanism of generating new antigenic variants during infection, as discussed by Marcello and Barry (Marcello & Barry, 2007b).

4.7 Discussion

In this section I will give an overview and brief discussion of the principal findings of the chapter; in Chapter 6, the results will be further evaluated and discussed in the wider context of the project and current understanding of antigenic variation and subtelomere evolution.

4.7.1 Gene annotation

The first outcome of the work described in this chapter was the annotation of thousands of *VSG* NTDs in the subtelomeres of the EATRO 3 and EATRO 2340 genome assemblies. The annotated gene sequences constituted the raw data required for the analyses described in the remainder of the chapter. The genes also provide a valuable resource for future studies (some possibilities discussed in Chapter 6) of the evolution and properties of the *VSG* archive in this strain and in *T. brucei* more generally, which were not possible within the scope of this project. The conclusions that could be drawn from analysing these data are discussed below, but there are two main points that emerged from the initial results. Firstly, the success of the annotation demonstrates that it is possible to extract a large number of data from the difficult subtelomeric regions even if they cannot be assembled into chromosomes, and that sequence that is often discarded in assembly efforts can be extremely informative.

Secondly, when compared by BLASTN, the genes from EATRO 3 had good hits in the subtelomeres of EATRO 2340, and vice versa, but neither archive matched well to the TREU 927 genome. This result illustrates how similar the two archives remain. Further, the lack of similarity with TREU 927 validated the annotation approach: because the TREU 927 and the EATRO 3-EATRO 2340 archives differed considerably, an annotation approach based solely on sequence similarity to previously annotated genes would have missed many *VSG*s in the EATRO genomes. The approach used was also far less time-consuming than annotating

*VSG*s by manual inspection for conserved cysteine patterns, as was used for some of the TREU 927 annotation (L. Marcello, PhD thesis, 2007).

4.7.2 Gene pairing and unpaired genes

Next, I described the identification of sequences that corresponded to the same gene in both genomes, a key step required to be able to follow changes in genes between the genomes. The pairing approach used was able to identify 1034 genes that were likely present in both genome assemblies. In conjunction with SNP annotation from PCR-free sequencing reads, this pairing of genes revealed the changes that had occurred in each gene between EATRO 3 in 1960 and EATRO 2340 in 1977. However, there remained 130 EATRO 3 and 778 EATRO 2340 *VSG* NTDs that could not be matched with high confidence to a sequence in the other genome. Considering the incomplete nature of both genomes, it is likely that many genes were missing partners simply because the partner was not assembled. In these cases, it may be speculated that one partner but not the other was assembled because the two were in somewhat different genomic contexts, but there was not sufficient information to investigate this possibility. It is striking that far more EATRO 2340 than EATRO 3 genes were unmatched, even when considered as a proportion of the total rather than an absolute count (43% of EATRO 2340 compared with 12% of EATRO 3 genes). However, this observation can be explained to some extent by the larger amount of sequence and smaller number of gaps in EATRO 2340: EATRO 2340 had 2.04 Mb of subtelomeric sequence in the assembled chromosomes, and 13.7 Mb of sequence on bin contigs containing annotated *VSG* NTDs; EATRO 3 had 1.81 Mb and 9.7 Mb respectively.

In an important step in pairing genes, a match was excluded because the sequence annotated as the partner of a particular gene did not hit that gene when it was used as the query in the reciprocal BLAST step. Often, the lack of a reciprocal hit was because there were two genes in the query genome that had both hit the same, or nearly the same, sequence in the other genome. In this case, whichever gene had the higher-identity match was excluded, and the other was classed as a ‘poor hit’. However, in 17 cases the gene that had the lower-identity match was nonetheless a good hit, and would have been accepted as a high-confidence match in the absence of a second hit. In two further instances there were two genes in the same genome that were very similar to one another, and had hit nearly the same sequence in the other genome, but there was a small difference between the precise limits of the hit that had prevented reciprocal BLAST from identifying either pair as invalid. Both of these scenarios could have been the result of one of the genes' partner not being assembled. However, the close similarity of all three genes (two in the query genome and one in the other genome) means a plausible alternative explanation is that there was originally just one gene in each genome, and it was duplicated in one genome but not the other. These results therefore provide some evidence that gene duplication had occurred in the course of *VSG* archive evolution in the time-scale studied.

4.7.3 Evolution of *VSG* NTDs

The next and principal part of the study described in this chapter was the comparison of genes and detection of changes that had occurred in the *VSG* NTDs between EATRO 3 and EATRO 2340, and analysis of what these changes could tell us about hyperevolution of *VSG* genes. One important technical point to emerge from the annotation of changes in the genes was that the assembled genomes represented only part of the true genomes, such that highly similar genes tended to have been conflated into a single gene. Hence, to compare the sequence of the assembled genes was over-simplistic. Therefore, such approaches should be used with caution, and sequences should be verified as far as possible by targeted resequencing. However, this study demonstrated that it was possible to gain data with reasonable reliability; albeit with some caveats, particularly that of not focusing too closely on specific SNPs.

An important conclusion obtained from the analysis of gene changes was that *VSG* NTDs accumulated substitution mutations significantly faster than core genes, with 2.4-fold higher substitution frequency. This conclusion was strengthened by the finding that the frequency of mutations at positions where any change was synonymous, a measure that removes the effects of selection, was significantly higher in *VSG* NTDs than in core genes. Substitution mutations account for only one of several mutation processes that change the sequence of *VSG* (Marcello & Barry, 2007b), so the substitution rate obtained for *VSG* NTDs (4.7×10^{-9} to 1.6×10^{-8} substitutions/base/generation) represents a minimum rate of change of *VSG* genes, and the overall rate of evolution is likely to be substantially higher. Further, it is likely that the figure obtained underestimated the difference in substitution frequency: as described above, the method used for SNP annotation was likely to have a non-negligible false negative rate. It seems probable that there will be more false negatives in the *VSG*s than in the core genes, for reasons that include more uncertainty in the assembly and more missing sequence, and that can be summed up as consequences of the *VSG*s being in a genome region that is more complex to assemble. The complexity and uncertainty of the subtelomeres' assembly is particularly relevant because successful annotation of SNPs relied on correct alignment of paired genes. This reasoning was borne out to some extent by observations of SNP numbers using different cataloguing methods: as the filters became more stringent, the number of SNPs in *VSG*s decreased more rapidly than the number of SNPs in core genes. Although most of the removed SNPs were likely to have been false positives, it does not seem implausible that some true positives were also excluded. This work therefore presents a picture of *VSG* genes evolving rapidly, at several times the rate of core genes.

The analysis also revealed that the substitution mutation profile of *VSG* NTDs was significantly different from that of core genes, even when corrected for the A-richness of the *VSG*s. Further, a preliminary comparison of strand bias in the two gene types suggested that the mutations occurring in *VSG* NTDs had a more pronounced strand bias than those in core genes. Taken together with the faster rate of change in *VSG* NTDs, these results supported the hypothesis that substitution mutation in *VSG*s is a process that is distinct from substitution mutation in chromosome cores. Such a difference could be produced either by different mechanisms acting in each genome region, or by the same basic mechanisms combining in different ways.

Having concluded that substitution in *VSG* NTDs was different from and faster than substitution in core genes, I explored the *VSG* NTD data further to attempt to gain insight into the processes occurring in *VSG* genes. First, the distribution of the number of SNPs per gene was not well described by a Poisson distribution; that is, one or more of the assumptions of the Poisson distribution was not true. Because of the observed overdispersion of the data, a probable explanation for violation of the Poisson distribution assumptions was that the occurrences of

SNPs were not independent. In particular, the probability a gene with at least one SNP would have more than one SNP (0.61) was considerably greater than the probability that any gene would have at least one SNP (0.13, *i.e.* P(n>1| n>0) > P(n>0), where n = number of SNPs in a gene), whereas if SNPs occurred completely independently we would expect the two probabilities to be equal. One possible explanation for this lack of independence is that despite the stringent filters used, in some cases 'SNPs' had been annotated which were actually due to the existence of two highly similar genes in both genomes that had a difference at that position in both genomes.

However, there are also plausible biological explanations that do not invoke technical artefacts. Firstly, multiple SNPs could arise from a single event. For example, segmental conversion between two highly similar genes could produce the outcomes observed: if a recent duplication event produced two copies of a particular gene, point mutations would accumulate independently in both daughter copies, and then a single segmental conversion event could copy several mutations from one gene to the other. This explanation is probably the most likely, because gene duplications, point mutations and segmental conversions have been predicted to occur in *VSG*s (Marcello & Barry, 2007b). A related possibility is that such duplication and point mutation do occur, but rather than sharing mutations between the two daughter copies, segmental conversion has overwritten changes in one copy by converting it to the ancestral sequence still present in the other copy. This mechanism would elevate the number of genes with no mutations beyond the number that would be expected from random mutations. A more speculative explanation is that point mutation occurs at random and independently, but only in a subset of genes. Such a subset could be created by targeting of mutagenic activity; however, no research has been done that could confirm or reject the idea of such targeting of mutations within subtelomeres. The observed distribution of the number of SNPs per gene therefore most probably resulted from the effects of segmental conversions but may well include some effects of sequencing and assembly artefacts.

When the SNPs were examined in the context of predicted protein sequence, a substantial number were found to result in an amino acid change, demonstrating that the observed mutations had potential to contribute to antigen diversity.

The frequency of non-synonymous mutations at non-synonymous sites (d_N) was higher than the frequency of synonymous mutations at synonymous sites (d_S), with a d_N/d_S ratio of 0.4, but as discussed above, this was unlikely to indicate purifying selection on the *VSG* genes.

On examination of the distribution of SNPs within genes, it appeared that distribution was random in type A NTDs but that there was some clustering of SNPs in type B NTDs. However, there seemed to be little relationship between predicted secondary structure and the region of SNP clustering, so the biological significance of the clustering is unclear. It is surprising that the two NTD types should differ in their SNP distribution, because there has been no previous indication that they evolve by distinct processes, and the genes' classification is based on their lineage rather than any separation of function (Jackson et al, 2012). Additionally, NTD genes of both types are intermingled in the genome, so it is not the case that B-type genes could experience a completely different genomic environment from A-type (Berriman *et al*, 2005; Marcello & Barry, 2007b). It is possible that aspects of the SNP annotation method biased the position of recovered SNPs: specifically, that the central region, which is more variable between genes, was more likely to have sequence that was unique to each gene and hence to have Illumina reads mapped with high confidence, which was necessary for a SNP to be called. Overall, there may be a bias in SNP occurrence towards certain parts of the gene, but the bias observed was not strong, and the difference between domain types was probably not relevant. Importantly, despite a possible bias in B-type VSGs towards the centre of the NTD, it was possible for a SNP to occur in any part of the gene.

Although SNPs may not be distributed at random within *VSG* sequence, this study found no evidence that a SNP was more likely to occur in *VSG* NTD than in subtelomeric sequence that was not annotated as *VSG* NTD. There was no difference in the frequency of substitution mutations in *VSG* NTDs and in subtelomeres generally. Further, any bias introduced by the method of SNP calling would be expected to bias the overall subtelomere substitution frequency downwards, because SNPs would be harder to call in the lower complexity sequence that tends to occur outside the *VSG*s, because fewer reads would map uniquely in these regions. Therefore it seems highly unlikely that mutation proceeds more rapidly in *VSG*s than anywhere else in the subtelomeres. This conclusion supports the hypothesis that hypermutation in *VSG*s is a consequence of their location in the hypermutational environment of the subtelomeres, although more evidence, such as a more detailed study of the types of mutation occurring, would be required to fully accept the hypothesis.

Examination of the composition of the *VSG* NTD archive annotated in both genomes tended to support the conclusions of *VSG* evolution drawn from study of the TREU 927 archive (Berriman *et al*, 2005; Marcello & Barry, 2007b). The majority of NTDs examined were pseudogenes, from a variety of causes including frameshift mutations, premature stop codons and lack of a CTD. There was some indication that a few pseudogenes in EATRO 3 had regained functionality in EATRO 2340, but examination of the genes involved did not exclude the possibility that the observed changes were artefactual. Both archives exhibited a large degree of substructuring, greater than that seen in TREU 927, implying that an even greater fraction of diversity in the archive was the result of duplication followed by divergence. Further evidence for an important role of such a process of duplication was supplied by the suggestion of a highly similar copy of many genes given by the large number of genes containing ‘heterozygous SNPs’.

The principal conclusion that can be drawn from the analysis of changes in *VSG* NTD genes is that substitution mutation proceeds differently in the *VSG* genes from in genes in the chromosome cores. The study went some way towards dissecting the substitution process in *VSG*s in more detail, but the detailed picture is unclear and few firm conclusions could be reached. There was no evidence obtained that the mutation processes were specific to *VSG* genes within the subtelomeres, although there was some tentative evidence that substitutions did not occur completely at random within the genes. Examination of genes within the archive supported the conclusion from TREU 927 that gene duplication is important in the evolution of the archive.

Chapter 5: Characterisation of DNA polymerase κ

5 Characterisation of DNA polymerase κ

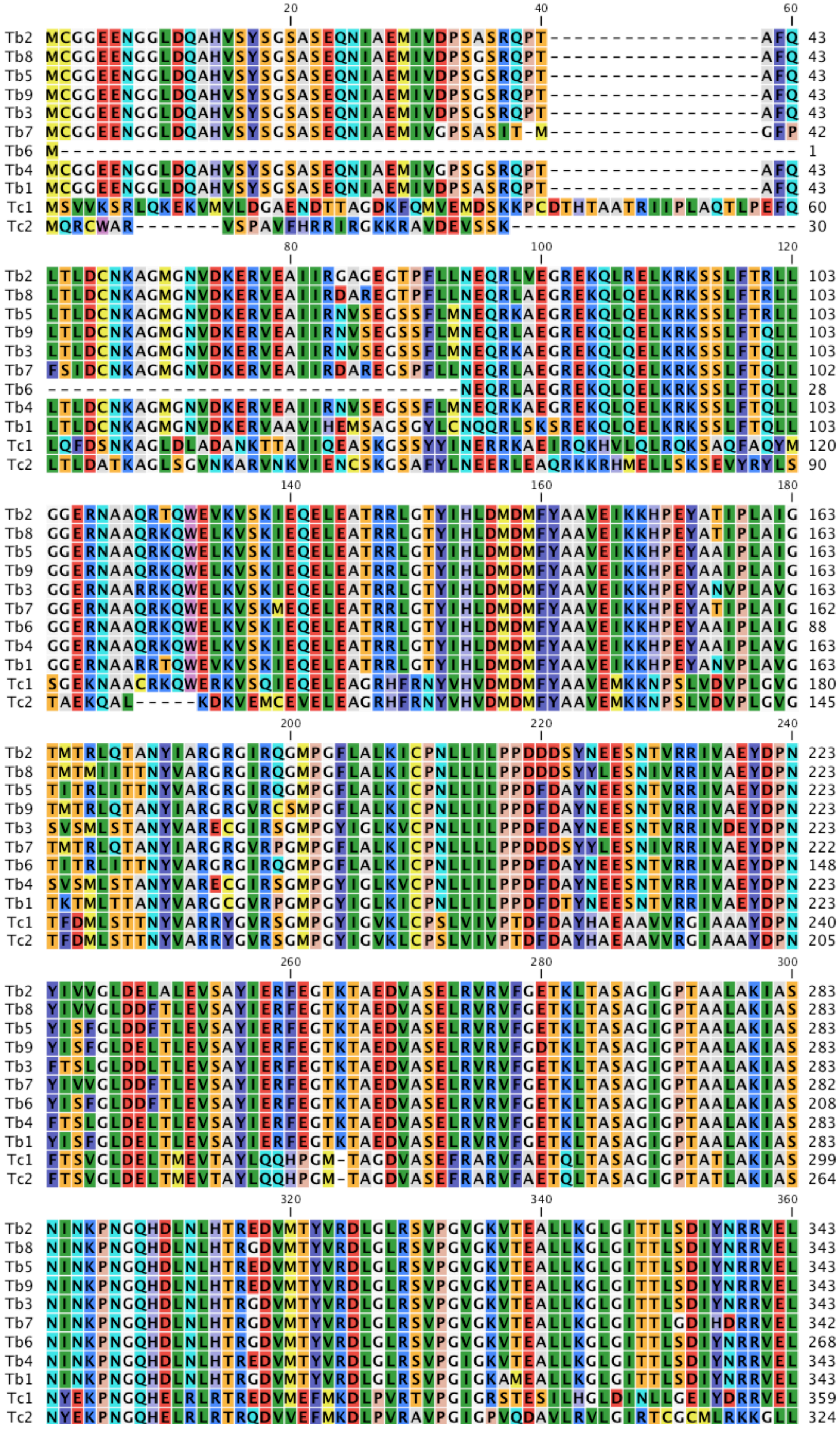
5.1 Introduction

The first aim of the work in this thesis was to characterise the mutations that occur during *VSG* evolution, as described in the previous two chapters. The second aim, which will be covered in this chapter, was to address the problem of which mechanisms are involved in generating (whether specifically or otherwise) these mutations. This is clearly an immense question, and is unlikely to have a single answer, and my work focused on examining the potential for involvement of a single mechanism hypothesised to play a role.

Lesion bypass DNA polymerases are enzymes that can insert bases opposite to DNA lesions that would otherwise block DNA synthesis (Prakash *et al*, 2005). The enzymes are error-prone compared with replicative DNA polymerases, primarily because of two structural features: their role requires a larger active site than replicative DNA polymerases, to accommodate DNA lesions; and they do not have 5′-3′ proofreading activity (Yang, 2005; Waters *et al*, 2009). This low replication fidelity has led to the suggestion that lesion bypass polymerases have a role in the systematic generation of mutations, including in trypanosome *VSG* genes (Goodman, 2002; McKenzie & Rosenberg, 2001).

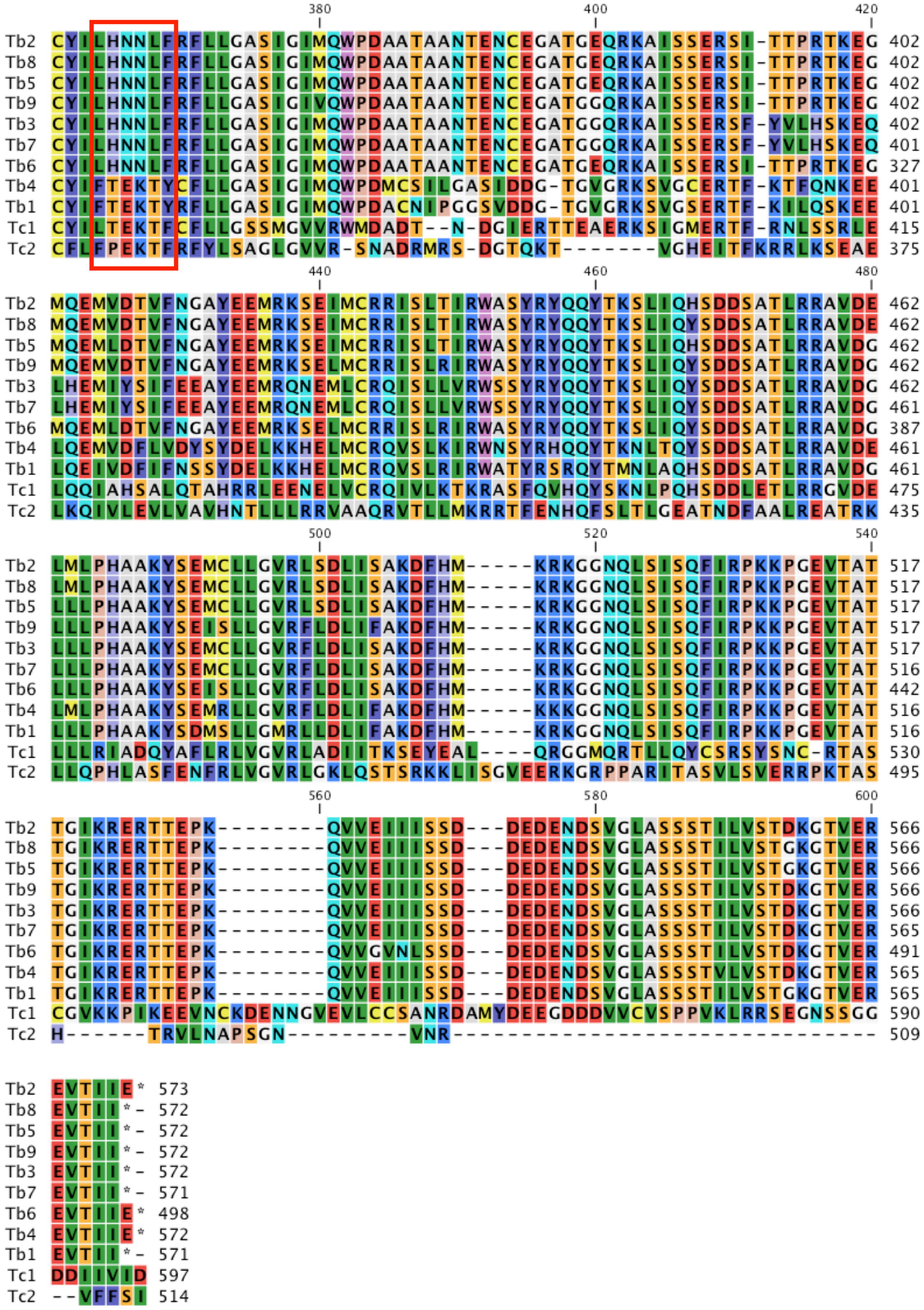
A lesion bypass polymerase of particular interest in the context of *VSG* evolution is pol κ, a member of the Y-family of DNA polymerases. Pol κ has undergone a *T. brucei*-specific expansion in gene number. *Leishmania major* has three copies of the pol κ gene, *T. cruzi* has two, but *T. brucei* has ten tandemly arranged copies, although one is truncated (Rajão *et al*, 2009). The characterised mutation profile of human pol κ (along with that of other human lesion bypass polymerases) is similar to the spectrum of mutations that are predicted to occur in *VSG* arrays (Johnson *et al*, 2000; Ohashi *et al*, 2000b; 2000a; Hile & Eckert, 2008; Marcello & Barry, 2007b). The *T. brucei* family of pol κ genes has diverged into two groups based on whether the sequence of the DNA-binding domain predicts a Y-family typical domain or an atypical domain. The predicted proteins model well onto the solved structure of human pol κ. However, the genes encode mutated residues lining the substrate-binding and active sites, which may render the enzyme even more permissive in its substrate, and hence more error-prone (Uljon *et al*, 2004; J.D. Barry and B. Marchetti, unpublished). The mean pairwise percentage amino acid identities between the predicted translations of the nine full-length pol κ genes in *T. brucei* TREU 927 and each of the two *T. cruzi* pol κ proteins are 43% (*Tc00.1047053503755.10*) and 36% (*Tc00.1047053503755.30*); and between *T. brucei* pol κs and human pol κ is 21%. The mean pairwise percentage amino acid identity between all full-length *T. brucei* pol κs is 86%; if the genes with canonical DNA-binding domains (n=2) and with atypical DNA-binding domains (n=7) are considered separately, the mean percentage identities are 88% and 89% respectively. Figure 5-1 shows an alignment of the predicted amino acid sequences of the full-length *T. brucei* pol κ genes and the *T. cruzi* pol κ protein sequences.

The expansion of the pol κ gene family and its likely error profile therefore made the gene products plausible candidates for DNA polymerases that introduce mutations into *VSG*. This chapter describes work carried out to attempt to test the hypothesis that *T. brucei* has adapted members of the pol κ gene family to specifically mutate trypanosome subtelomeres. This question was approached in several ways. Firstly, I attempted to characterise the *in vitro* mutation profile of two members of the pol κ family, one from each subfamily. The aim of this work was to examine how closely the actual mutation profile of the enzymes matched the mutations observed to occur in the subtelomeres. Secondly, I studied the two pol κ proteins *in vivo*, examining their subcellular localisation and the effect of their overexpression, to test whether these features were consistent with a role for these lesion bypass polymerases at subtelomeres.



```
                                   20                  40                  60
Tb2  MCGGEENGGLDQAHVSYSGSASEQNIAEMIVDPSASRQPT-----------------AFQ  43
Tb8  MCGGEENGGLDQAHVSYSGSASEQNIAEMIVDPSGSRQPT-----------------AFQ  43
Tb5  MCGGEENGGLDQAHVSYSGSASEQNIAEMIVDPSGSRQPT-----------------AFQ  43
Tb9  MCGGEENGGLDQAHVSYSGSASEQNIAEMIVDPSGSRQPT-----------------AFQ  43
Tb3  MCGGEENGGLDQAHVSYSGSASEQNIAEMIVDPSGSRQPT-----------------AFQ  43
Tb7  MCGGEENGGLDQAHVSYSGSASEQNIAEMIVGPSASIT-M-----------------GFP  42
Tb6  M-----------------------------------------------------------   1
Tb4  MCGGEENGGLDQAHVSYSGSASEQNIAEMIVGPSGSRQPT-----------------AFQ  43
Tb1  MCGGEENGGLDQAHVSYSGSASEQNIAEMIVDPSASRQPT-----------------AFQ  43
Tc1  MSVVKSRLQKEKVMVLDGAENDTTAGDKFQMVEMDSKKPCDTHTAATRIIPLAQTLPEFQ  60
Tc2  MQRCWAR-------VSPAVFHRRIRGKKRAVDEVSSK-----------------------  30
                                   80                 100                 120
Tb2  LTLDCNKAGMGNVDKERVEAIIRGAGEGTPFLLNEQRLVEGREKQLRELKRKSSLFTRLL 103
Tb8  LTLDCNKAGMGNVDKERVEAIIRDAREGTPFLLNEQRLAEGREKQLQELKRKSSLFTRLL 103
Tb5  LTLDCNKAGMGNVDKERVEAIIRNVSEGSSFLMNEQRKAEGREKQLQELKRKSSLFTRLL 103
Tb9  LTLDCNKAGMGNVDKERVEAIIRNVSEGSSFLMNEQRLAEGREKQLQELKRKSSLFTQLL 103
Tb3  LTLDCNKAGMGNVDKERVEAIIRNVSEGSSFLMNEQRKAEGREKQLQELKRKSSLFTQLL 103
Tb7  FSIDCNKAGMGNVDKERVEAIIRDAREGSPFLLNEQRLAEGREKQLQELKRKSSLFTQLL 102
Tb6  ---------------------------------NEQRLAEGREKQLQELKRKSSLFTQLL  28
Tb4  LTLDCNKAGMGNVDKERVEAIIRNVSEGSSFLMNEQRKAEGREKQLQELKRKSSLFTQLL 103
Tb1  LTLDCNKAGMGNVDKERVAAVIHEMSAGSGYLCNQQRLSKSREKQLQELKRKSSLFTQLL 103
Tc1  LQFDSNKAGLDLADANKTTAIIQEASKGSSYYINERRKAEIRQKHVLQLRQKSAQFAQYM 120
Tc2  LTLDATKAGLSGVNKARVNKVIENCSKGSAFYLNEERLEAQRKKRHMELLSKSEVYRYLS  90
                                  140                 160                 180
Tb2  GGERNAAQRTQWEVKVSKIEQELEATRRLGTYIHLDMDMFYAAVEIKKHPEYATIPLAIG 163
Tb8  GGERNAAQRKQWELKVSKIEQELEATRRLGTYIHLDMDMFYAAVEIKKHPEYATIPLAIG 163
Tb5  GGERNAAQRKQWELKVSKIEQELEATRRLGTYIHLDMDMFYAAVEIKKHPEYAAIPLAIG 163
Tb9  GGERNAAQRKQWELKVSKIEQELEATRRLGTYIHLDMDMFYAAVEIKKHPEYAAIPLAIG 163
Tb3  GGERNAARRKQWELKVSKIEQELEATRRLGTYIHLDMDMFYAAVEIKKHPEYANVPLAVG 163
Tb7  GGERNAAQRKQWELKVSKMEQELEATRRLGTYIHLDMDMFYAAVEIKKHPEYATIPLAIG 162
Tb6  GGERNAAQRKQWELKVSKIEQELEATRRLGTYIHLDMDMFYAAVEIKKHPEYAAIPLAIG  88
Tb4  GGERNAAQRKQWELKVSKIEQELEATRRLGTYIHLDMDMFYAAVEIKKHPEYAAIPLAVG 163
Tb1  GGERNAARRTQWEVKVSKIEQELEATRRLGTYIHLDMDMFYAAVEIKKHPEYANVPLAVG 163
Tc1  SGEKNAACRKQWERKVSQIEQELEAGRHFRNYVHVDMDMFYAAVEMKKNPSLVDVPLGVG 180
Tc2  TAEKQAL-----KDKVEMCEVELEAGRHFRNYVHVDMDMFYAAVEMKKNPSLVDVPLGVG 145
                                  200                 220                 240
Tb2  TMTRLQTANYIARGRGIRQGMPGFLALKICPNLLILPPDDDSYNEESNTVRRIVAEYDPN 223
Tb8  TMTMIITTNYVARGRGIRQGMPGFLALKICPNLLLLPPDDDSYYLESNIVRRIVAEYDPN 223
Tb5  TITRLITTNYVARGRGIRQGMPGFLALKICPNLLILPPDFDAYNEESNTVRRIVAEYDPN 223
Tb9  TMTRLQTANYIARGRGVRCSMPGFLALKICPNLLILPPDFDAYNEESNTVRRIVAEYDPN 223
Tb3  SVSMLSTANYVARECGIRSGMPGYIGLKVCPNLLILPPDFDAYNEESNTVRRIVDEYDPN 223
Tb7  TMTRLQTANYIARGRGVRPGMPGFLALKICPNLLLLPPDDDSYYLESNIVRRIVAEYDPN 222
Tb6  TITRLITTNYVARGRGIRQGMPGFLALKICPNLLILPPDFDAYNEESNTVRRIVAEYDPN 148
Tb4  SVSMLSTANYVARECGIRSGMPGYIGLKVCPNLLILPPDFDAYNEESNTVRRIVAEYDPN 223
Tb1  TKTMLTTANYVARGCGVRPGMPGYIGLKICPNLLILPPDFDTYNEESNTVRRIVAEYDPN 223
Tc1  TFDMLSTTNYVARRYGVRSGMPGYIGVKLCPSLVIVPTDFDAYHAEAAVVRGIAAAYDPN 240
Tc2  TFDMLSTTNYVARRYGVRSGMPGYIGVKLCPSLVIVPTDFDAYHAEAAVVRGIAAAYDPN 205
                                  260                 280                 300
Tb2  YIVVGLDELALEVSAYIERFEGTKTAEDVASELRVRVFGETKLTASAGIGPTAALAKIAS 283
Tb8  YIVVGLDDFTLEVSAYIERFEGTKTAEDVASELRVRVFGETKLTASAGIGPTAALAKIAS 283
Tb5  YISFGLDDFTLEVSAYIERFEGTKTAEDVASELRVRVFGETKLTASAGIGPTAALAKIAS 283
Tb9  YISFGLDELTLEVSAYIERFEGTKTAEDVASELRVRVFGDTKLTASAGIGPTAALAKIAS 283
Tb3  FTSLGLDDLTLEVSAYIERFEGTKTAEDVASELRVRVFGETKLTASAGIGPTAALAKIAS 283
Tb7  YIVVGLDDFTLEVSAYIERFEGTKTAEDVASELRVRVFGETKLTASAGIGPTAALAKIAS 282
Tb6  YISFGLDDFTLEVSAYIERFEGTKTAEDVASELRVRVFGETKLTASAGIGPTAALAKIAS 208
Tb4  FTSLGLDELTLEVSAYIERFEGTKTAEDVASELRVRVFGETKLTASAGIGPTAALAKIAS 283
Tb1  YISFGLDELTLEVSAYIERFEGTKTAEDVASELRVRVFGETKLTASAGIGPTAALAKIAS 283
Tc1  FTSVGLDELTMEVTAYLQQHPGM-TAGDVASEFRARVFAETQLTASAGIGPTATLAKIAS 299
Tc2  FTSVGLDELTMEVTAYLQQHPGM-TAGDVASEFRARVFAETQLTASAGIGPTATLAKIAS 264
                                  320                 340                 360
Tb2  NINKPNGQHDLNLHTREDVMTYVRDLGLRSVPGVGKVTEALLKGLGITTLSDIYNRRVEL 343
Tb8  NINKPNGQHDLNLHTRGDVMTYVRDLGLRSVPGVGKVTEALLKGLGITTLSDIYNRRVEL 343
Tb5  NINKPNGQHDLNLHTREDVMTYVRDLGLRSVPGVGKVTEALLKGLGITTLSDIYNRRVEL 343
Tb9  NINKPNGQHDLNLHTREDVMTYVRDLGLRSVPGVGKVTEALLKGLGITTLSDIYNRRVEL 343
Tb3  NINKPNGQHDLNLHTRGDVMTYVRDLGLRSVPGVGKVTEALLKGLGITTLSDIYNRRVEL 343
Tb7  NINKPNGQHDLNLHTRGDVMTYVRDLGLRSVPGVGKVTEALLKGLGITTLGDIHDRRVEL 342
Tb6  NINKPNGQHDLNLHTRGDVMTYVRDLGLRSVPGVGKVTEALLKGLGITTLSDIYNRRVEL 268
Tb4  NINKPNGQHDLNLHTREDVMTYVRDLGLRSVPGIGKVTEALLKGLGITTLSDIYNRRVEL 343
Tb1  NINKPNGQHDLNLHTRGDVMTYVRDLGLRSVPGIGKAMEALLKGLGITTLSDIYNRRVEL 343
Tc1  NYEKPNGQHELRLRTREDVMEFMKDLPVRTVPGIGRSTESILHGLDINLLGEIYDRRVEL 359
Tc2  NYEKPNGQHELRLRTRQDVVEFMKDLPVRAVPGIGPVQDAVLRVLGIRTCGCMLRKKGLL 324
```

Figure 5-1 – Aligned *T. brucei* and *T. cruzi* pol κ protein sequences (continued next page).



```
                                   380                 400                 420
                                     |                   |                   |
Tb2  CYILHNNLFRFLLGASIGIMQWPDAATAANTENCEGATGEQRKAISSERSI-TTPRTKEG 402
Tb8  CYILHNNLFRFLLGASIGIMQWPDAATAANTENCEGATGEQRKAISSERSI-TTPRTKEG 402
Tb5  CYILHNNLFRFLLGASIGIMQWPDAATAANTENCEGATGEQRKAISSERSI-TTPRTKEG 402
Tb9  CYILHNNLFRFLLGASIGIVQWPDAATAANTENCEGATGGQRKAISSERSI-TTPRTKEG 402
Tb3  CYILHNNLFRFLLGASIGIMQWPDAATAANTENCEGATGGQRKAISSERSF-YVLHSKEQ 402
Tb7  CYILHNNLFRFLLGASIGIMQWPDAATAANTENCEGATGGQRKAISSERSF-YVLHSKEQ 401
Tb6  CYILHNNLFRFLLGASIGIMQWPDAATAANTENCEGATGEQRKAISSERSI-TTPRTKEG 327
Tb4  CYIFTEKTYCFLLGASIGIMQWPDMCSILGASIDDG-TGVGRKSVGCERTF-KTFQNKEE 401
Tb1  CYIFTEKTYRFLLGASIGIMQWPDACNIPGGSVDDG-TGVGRKSVGSERTF-KILQSKEE 401
Tc1  CYILTEKTFCFLLGSSMGVVRWMDADT--N-DGIERTTEAERKSIGMERTF-RNLSSRLE 415
Tc2  CFLFPEKTFRFYLSAGLGVVR-SNADRMRS-DGTQKT-------VGHEITFKRRLKSEAE 375
                                   440                 460                 480
                                     |                   |                   |
Tb2  MQEMVDTVFNGAYEEMRKSEIMCRRISLTIRWASYRYQQYTKSLIQHSDDSATLRRAVDE 462
Tb8  MQEMVDTVFNGAYEEMRKSEIMCRRISLTIRWASYRYQQYTKSLIQYSDDSATLRRAVDE 462
Tb5  MQEMLDTVFNGAYEEMRKSEIMCRRISLTIRWASYRYQQYTKSLIQHSDDSATLRRAVDG 462
Tb9  MQEMVDTVFNGAYEEMRKSELMCRRISLRIRWASYRYQQYTKSLIQYSDDSATLRRAVDG 462
Tb3  LHEMIYSIFEEAYEEMRQNEMLCRQISLLVRWSSYRYQQYTKSLIQYSDDSATLRRAVDG 462
Tb7  LHEMIYSIFEEAYEEMRQNEMLCRQISLLVRWSSYRYQQYTKSLIQYSDDSATLRRAVDG 461
Tb6  MQEMLDTVFNGAYEEMRKSELMCRRISLRIRWASYRYQQYTKSLIQYSDDSATLRRAVDG 387
Tb4  LQEMVDFLVDYSYDELKKHELMCRQVSLKIRWNSYRHQQYTKNLTQYSDDSATLRRAVDE 461
Tb1  LQEIVDFIFNSSYDELKKHELMCRQVSLRIRWATYRSRQYTMNLAQHSDDSATLRRAVDG 461
Tc1  LQQIAHSALQTAHRRLEENELVCRQIVLKTKRASFQVHQYSKNLPQHSDDLETLRRGVDE 475
Tc2  LKQIVLEVLVAVHNTLLLRRVAAQRVTLLMKRRTFENHQFSLTLGEATNDFAALREATRK 435
                                   500                 520                 540
                                     |                   |                   |
Tb2  LMLPHAAKYSEMCLLGVRLSDLISAKDFHM-----KRKGGNQLSISQFIRPKKPGEVTAT 517
Tb8  LMLPHAAKYSEMCLLGVRLSDLISAKDFHM-----KRKGGNQLSISQFIRPKKPGEVTAT 517
Tb5  LLLPHAAKYSEMCLLGVRLSDLISAKDFHM-----KRKGGNQLSISQFIRPKKPGEVTAT 517
Tb9  LLLPHAAKYSEISLLGVRFLDLIFAKDFHM-----KRKGGNQLSISQFIRPKKPGEVTAT 517
Tb3  LLLPHAAKYSEMCLLGVRFLDLISAKDFHM-----KRKGGNQLSISQFIRPKKPGEVTAT 517
Tb7  LLLPHAAKYSEMCLLGVRFLDLISAKDFHM-----KRKGGNQLSISQFIRPKKPGEVTAT 516
Tb6  LLLPHAAKYSEISLLGVRFLDLISAKDFHM-----KRKGGNQLSISQFIRPKKPGEVTAT 442
Tb4  LMLPHAAKYSEMRLLGVRFLDLIFAKDFHM-----KKKGGNQLSISQFIRPKKPGEVTAT 516
Tb1  LLLPHAAKYSDMSLLGMRLLDLIFAKDFHM-----KRKGGNQLSISQFIRPKKPGEVTAT 516
Tc1  LLLRIADQYAFLRLVGVRLADIITKSEYEAL----QRGGMQRTLLQYCSRSYSNC-RTAS 530
Tc2  LLQPHLASFENFRLVGVRLGKLQSTSRKKLISGVEERKGRPPARITASVLSVERRPKTAS 495
                                   560                 580                 600
                                     |                   |                   |
Tb2  TGIKRERTTEPK--------QVVEIIISSD---DEDENDSVGLASSSTILVSTDKGTVER 566
Tb8  TGIKRERTTEPK--------QVVEIIISSD---DEDENDSVGLASSSTILVSTGKGTVER 566
Tb5  TGIKRERTTEPK--------QVVEIIISSD---DEDENDSVGLASSSTILVSTGKGTVER 566
Tb9  TGIKRERTTEPK--------QVVEIIISSD---DEDENDSVGLASSSTILVSTDKGTVER 566
Tb3  TGIKRERTTEPK--------QVVEIIISSD---DEDENDSVGLASSSTILVSTDKGTVER 566
Tb7  TGIKRERTTEPK--------QVVEIIISSD---DEDENDSVGLASSSTILVSTDKGTVER 565
Tb6  TGIKRERTTEPK--------QVVGVNLSSD---DEDENDSVGLASSSTILVSTDKGTVER 491
Tb4  TGIKRERTTEPK--------QVVEIIISSD---DEDENDSVGLASSSTVLVSTDKGTVER 565
Tb1  TGIKRERTTEPK--------QVVEIIISSD---DEDENDSVGLASSSTILVSTGKGTVER 565
Tc1  CGVKKPIKEEVNCKDENNGVEVLCCSANRDAMYDEEGDDDVVCVSPPVKLRRSEGNSSGG 590
Tc2  H------TRVLNAPSGN---------VNR------------------------------- 509

Tb2  EVTIIE* 573
Tb8  EVTII*- 572
Tb5  EVTII*- 572
Tb9  EVTII*- 572
Tb3  EVTII*- 572
Tb7  EVTII*- 571
Tb6  EVTIIE* 498
Tb4  EVTIIE* 572
Tb1  EVTII*- 571
Tc1  DDIIVID 597
Tc2  --VFFSI 514
```

Figure 5-1 continued. The truncated (487 bp) gene *Tb11.01.0060* has not been included. The DNA-binding domain is indicated by a red box on the top line of the second page. Sequences were aligned and visualised with CLC Genomics Workbench 6. Positions in the alignment are shown above each row. Tc1 = *Tc00.1047053503755.10*; Tc2 = *Tc00.1047053503755.30*; Tb1 = *Tb11.01.0010*; Tb2 = *Tb11.01.0020*; Tb3 = *Tb11.01.0030*; Tb4 = *Tb11.01.0040*; Tb5 = *Tb11.01.0050*; Tb6 = *Tb11.01.0080*; Tb7 = *Tb11.12.0001*; Tb8 = *Tb11.12.0002*; Tb9 = *Tb11.12.0003*.

5.2 Biochemical analysis

5.2.1 Generation of MBP-tagged pol κ constructs and active site mutants

One pol κ gene was studied from each of the two branches of the family: *Tb11.12.0001* (referred to as polκ001, atypical DNA binding domain) and *Tb11.01.0010* (referred to as polκ10, typical DNA binding domain). Expression vectors were the gift of Dr Barbara Marchetti and were constructed by amplifying the genes from *T. brucei* genomic DNA and cloning them into the vector pMAL-c4x (NEB). The vectors (Figure 5-2) allowed expression in *E. coli* of the trypanosome genes fused to maltose binding protein (MBP) at their N-termini, when induced with IPTG.

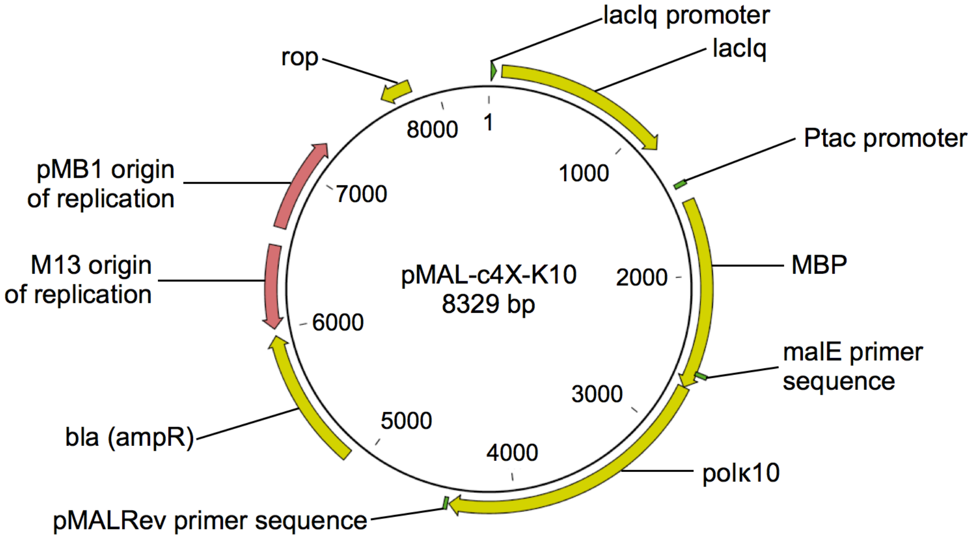


Figure 5-2 – MBP-tagged polκ10 expression construct.
The central circular scale is in bp from an arbitrary position (the beginning of the laclq promoter). Notable features are the tag (labelled MBP); the pol κ gene (labelled polκ10); and the ampicillin resistance selectable marker for bacterial transformation and maintenance (labelled bla (ampR)). The MBP-polκ001 construct was identical except for the pol κ gene.

I generated a predicted null mutant of each of the two MBP-polκ genes, using site-directed mutagenesis with primers specific to each gene, using either pMAL-K10 (primers PL15 and PL16) or pMAL-K001 (primers PL17 and PL18) as the template. Mutants were screened using colony PCR with PL19 (pMAL-K10) or PL21 (pMAL-K001) and P79 primers, then verified by sequencing using the primers P74, P78, T7 and T7term. These mutants were generated to provide a negative control to check that any polymerase activity of the MBP-polκ preparations was due to *T. brucei* pol κ, rather than due to contamination with polymerase activity from *E. coli* proteins. The mutation was of a critical aspartate in the catalytic triad conserved in Y-family polymerases, which in the related pol η completely abolished polymerase activity (Kondratick *et al*, 2001). Both wild-type constructs and both mutant constructs were used to transform Rosetta2 *E. coli* competent cells (Novagen, gift of Anna Trenaman), a strain optimised for protein expression.

5.2.2 Purification of MBP-polκ fusion proteins by amylose affinity chromatography

Initially, I grew Rosetta2 cells expressing MBP-polκ fusion proteins at 20ºC and purified MBP-polκ fusion proteins by amylose affinity chromatography. PAGE analysis demonstrated the final preparations of soluble protein each contained a protein band of the appropriate apparent size, although much of the fusion protein produced appeared to be insoluble. Some extra protein bands were visible on the gel after purification, indicating potential contamination or degradation. I made an attempt to remove contaminating proteins with relatively low maltose affinity by including small concentrations of maltose in the wash buffer and by doing a second round of affinity chromatography. However, this change did not noticeably improve the protein purity, and the second round of purification considerably reduced the protein yield (Figure 5-3).

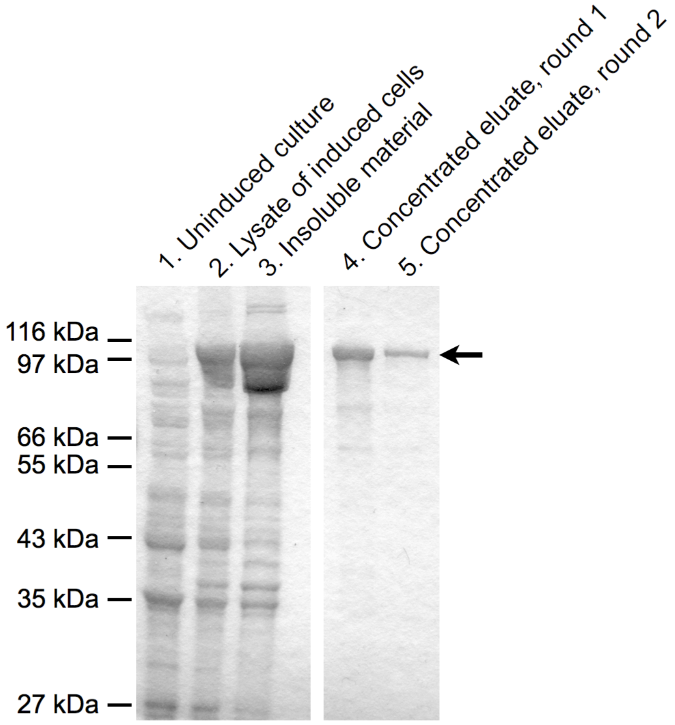

1. Uninduced culture
2. Lysate of induced cells
3. Insoluble material
4. Concentrated eluate, round 1
5. Concentrated eluate, round 2
116 kDa
97 kDa
66 kDa
55 kDa
43 kDa
35 kDa
27 kDa

Figure 5-3 – Representative PAGE showing stages in maltose affinity purification of pol κ-MBP proteins.
E. coli **cells transformed with pMAL-K10 (Lane 1) were induced with IPTG, grown overnight at 15°C, harvested and lysed (Lane 2). Insoluble material in the lysate (Lane 3) was removed by centrifugation, and appears to contain a large amount of the fusion protein, which ran below its expected molecular weight, probably due to its high concentration. Soluble MBP-polκ10 was purified by amylose affinity chromatography, and concentrated (Lane 4). A second round of amylose affinity purification and concentration was carried out with this concentrated preparation (Lane 5). The black arrow indicates approximately the expected size of MBP-polκ10 (106 kDa). Lanes have been cropped from the gel for clarity. The size markers to the left of the gel image indicate the positions of protein size standards (Protein Marker, Broad Range (2-212 kDa) (NEB)).**

5.2.3 Development of assays to characterise pol κ activity

I developed and optimised several assays to characterise the activity of the pol κ DNA polymerase, testing the efficacy of the assays with tagged wild-type pol κ proteins. This work was carried out in parallel with generation and preparation of the tagged mutant pol κ proteins. Results in this section therefore do not contain the appropriate pol κ negative controls, and are shown for the purpose of illustrating the development processes. For all assays described I used reaction conditions that had been successful for MBP-tagged *T. cruzi* pol κ (Rajão *et al*, 2009). The conditions included magnesium as the divalent cation, which has been shown to be an effective cofactor for human pol κ (Gerlach *et al*, 2001).

5.2.3.1 Gap-filling assay for characterisation of mutational profile

I first attempted to adapt for the pol κ proteins a forward mutation assay intended to determine the mutagenic profile of a polymerase on undamaged DNA (Bebenek & Kunkel, 1995). The template, which I prepared by restriction digestion and hybridisation, was M13mp18 dsDNA containing a 407-bp single-stranded gap in the *lacZα* gene. In the assay, the gapped template is incubated with a polymerase that synthesises DNA across the gap and the products are transformed into *E. coli* and plated on a lawn of *E. coli* α-complementation cells on plates containing X-Gal. If synthesis has been error free, the *lacZα* gene product complements the defective ß-galactosidase activity of the lawn cells, allowing hydrolysis of X-Gal and the production of blue plaques. White plaques arise from products of reactions in which an error occurred and these can be counted and the mutants sequenced.

I used MBP-polκ fusion proteins as the polymerase in this gap-filling assay. When I analysed reaction products by agarose gel electrophoresis, no difference in band mobility was observable between the MBP-polκ reactions and the negative control (Figure 5-4, bands running at approximately 7 kb in lanes 1 and 2 (MBP-polκ) and 4 (no-enzyme negative control)). An observable difference would be expected between gapped substrate and the product of the gap-filling reaction (Bebenek & Kunkel, 1995), and there was an observable difference from the negative control in the product of a gap-filling reaction when *E. coli* Klenow fragment DNA polymerase was used (Figure 5-4, lane 3). When I used the reaction products to transform *E. coli*, there was no consistent difference, either quantitatively or qualitatively (plaque colour), between the plaques on the MBP-polκ and negative control plates.

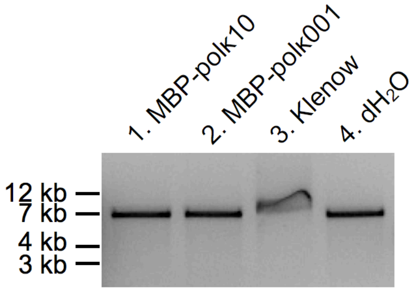


Figure 5-4 – Agarose gel analysis of the products of a typical gap-filling reaction. Approximately 300 ng of the polymerase indicated in the lane label was incubated with approximately 40 ng of gapped DNA substrate for 3 h at 37ºC, then the reactions were stopped by the addition of EDTA, and 20 µl of each 30 µl reaction was analysed by agarose gel electrophoresis. Klenow = *E. coli* Klenow fragment DNA polymerase. The size markers to the left of the gel image indicate the positions of DNA size standards (1 kb Plus DNA ladder, Invitrogen).

5.2.3.2 Primer extension assay for detection of polymerase activity

One hypothesis for this lack of activity in the gap-filling assay was that the purified polymerases were inactive. Therefore, to test in a simplified system whether the MBP-polκ preparations had any polymerase activity, I used a primer extension assay based on an experiment used for *T. cruzi* pol κ (Rajão *et al*, 2009). A 17-bp 5′ fluorescein-labelled primer (M13 primer) was hybridised to a 31-bp template (DLoopF, Figure 5-5A). The substrate was incubated with polymerase to allow extension of the primer. Polymerase activity was detected by a mobility shift of the extended primer in denaturing PAGE. A novel second, similar assay used a template in which, during the hybridisation step, I added a third type of DNA molecule: an unlabelled oligonucleotide (PL3) complementary to the part of the template that was not complementary to the fluorescein-labelled primer. This hybridisation generated a tripartite substrate that could be used to check the activity of polymerases in strand displacement DNA synthesis (Figure 5-5B).The assay demonstrated that wild-type MBP-polκ preparations did indeed have polymerase activity, and were able to synthesise DNA without and with strand displacement (Figure 5-5C).

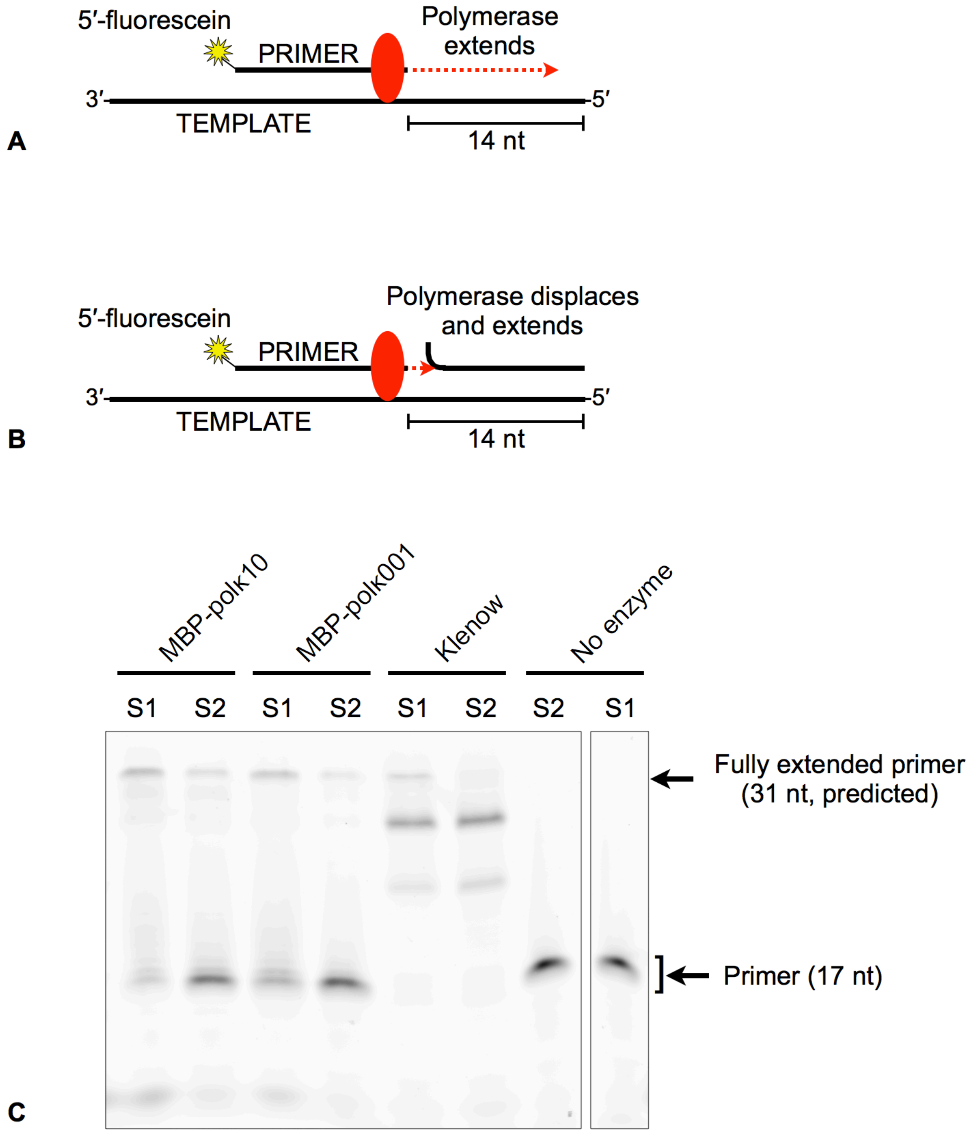


Figure 5-5 – Primer extension assays
A) Schematic diagram of primer extension assay (substrate 1, S1). B) Schematic diagram of strand displacement primer extension assay (substrate 2, S2). C) Pilot experiment showing both assays. The indicated polymerases (1 µg) or no enzyme were incubated with 200 ng of the indicated substrate for 2 h at 37ºC. The products were analysed on polyacrylamide gels and imaged using a Typhoon imager. Black arrows indicate the position of the starting primer and the fully extended primer.

5.2.3.3 Reversion assay for estimation of mutation rate

Given that the purified polymerases appeared to be active, another possible explanation for lack of activity in the gap-filling assay was that this assay reaction was so far from the polymerases’ physiological role that the enzymes were unable to synthesise DNA. I therefore developed a mutation assay that had conditions closer to the enzymes’ likely *in vitro* activity. This assay was loosely based on a reversion assay described by Osheroff and colleagues (Osheroff *et al*, 1999). I amplified part of the *lacZα* gene in M13mp18, using primers PL1 and

PL2, which added a recognition site for Nb.*Bbv*CI, which includes a stop codon, in the multiple cloning site near the 5′ end of the coding region of the gene. I then used existing *Bam*HI and *Bgl*II sites in the gene to remove the original sequence between these sites and replace it with the modified PCR product. This generated a version of M13mp18, referred to as M13Stop, with a nicking enzyme restriction site and a stop codon within the *lacZa* gene, close to the start of the ORF. The construct was checked by sequencing with primer PL2. The insertion of the stop codon resulted in a colourless plaque phenotype when the double-stranded bacteriophage was transformed into *E. coli*.

I treated the modified bacteriophage template with the enzyme Nb.*Bbv*CI, providing a substrate for strand displacement DNA synthesis by the DNA polymerases, on otherwise undamaged template. If a single error was made by the polymerase in synthesis of the stop codon, for eight out of nine possible changes this would result in a functional *lacZa* gene, and the bacteriophage would be able to produce blue plaques again (Figure 5-6).

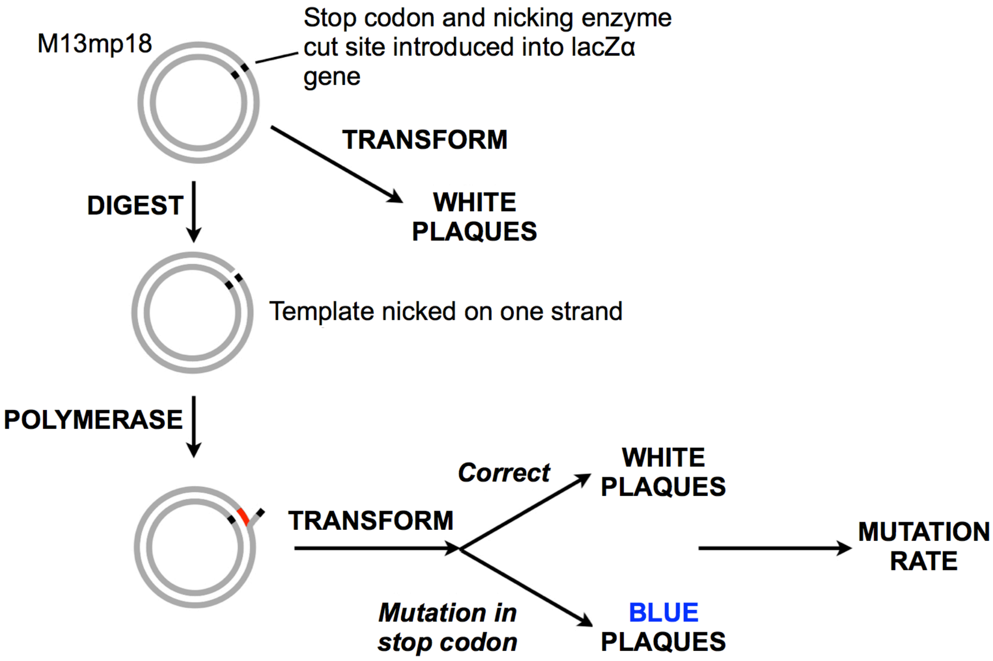


Figure 5-6 – Schematic diagram of steps in the M13-Stop reversion assay. Double-stranded M13-Stop was treated with Nb.*Bbv*CI to generate a nicked template, which was incubated with polymerase and the products transformed in *E. coli*. If an error was made in synthesising the stop codon, blue plaques would be produced; otherwise, white plaques would be produced.

Pilot experiments with wild-type MBP-polκ preparations resulted in the generation of some blue colonies, indicating that this approach could be used to estimate the mutation rates of pol κ proteins. Given the mutation rate of human pol κ on undamaged DNA (Ohashi *et al*, 2000a), we would expect approximately one revertant per 65 plaques in this assay if human pol κ were used. However, the preliminary results for *T. brucei* pol κ indicated a considerably lower revertant rate: seven blue plaques out of 5500 scored for MBP-polκ10, and one blue plaque out of 900 scored for MBP-polκ001.

5.2.3.4 Quantification of escape from tet repressor for estimation of mutation rate

As an alternative to *in vitro* characterisation of mutation rate with purified protein, I attempted to adapt a forward mutation approach that has been used to characterise spontaneous mutation rates and profiles in *E. coli* under various conditions (Bjedov *et al*, 2007), using a cell line that was the gift of Dr Ivan Matic. In this line, referred to as MG1655 *cI*-Tet (see Chapter 2 for full genotype), cells contain a *tetA* (tetracycline resistance) gene whose promoter has been replaced by the λ pR promoter; and the λ *cI* (Ind$^-$) gene, which produces a protein, cI, that represses expression from λ pR. If a mutation occurs that inactivates or prevents expression of cI, the cells will gain tetracycline resistance, allowing the mutation rate to be estimated, and the mutation(s) in λ *cI* (Ind$^-$) to be characterised.

In *E. coli*, overexpression of the pol κ homologue *dinB* increases the spontaneous mutation rate by factors of up to several hundred (Kim *et al*, 1997). If *T. brucei* polκ10 is a typical Y-family DNA polymerase that can interact with the *E. coli* DNA synthesis and repair system, its overexpression should give a similar phenotype of increased mutation rate. Therefore, to test whether this system could be used with *T. brucei* pol κ, I transformed pMAL-K10 into competent MG1655 *cI*-Tet cells. PAGE of lysates of cells to which IPTG had been added indicated that induction of overexpression was successful (Figure 5-7). I grew both induced and uninduced *E. coli* cells from a very small starting count, spread the culture on tetracycline plates, incubated overnight, and counted the resulting tetracycline-resistant colonies. Initial experiments showed very little difference between the number of resistant colonies from induced and uninduced cells (data not shown). It is possible that further experiments might have allowed a significant difference to be observed, but it seemed unlikely that this approach would be useful. One possible explanation for the lack of a positive result was failure of polκ10 to engage with the *E. coli* replication machinery, so we did not consider that this outcome showed that polκ10 was not an active polymerase.

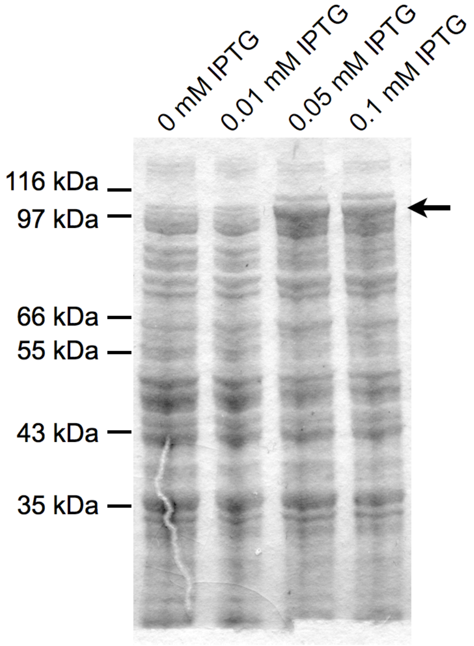


Figure 5-7 – Expression of MBP- polκ10 in MG1655 *cI*-Tet cells.
MG1655 *cI*-Tet cells transformed with pMAL-K10 were induced with the indicated concentration of IPTG, and their approximate density determined by spectrophotometry. Sample of approximately the same number of cells for each culture were lysed with protein running buffer and analysed by PAGE. The expected size of MBP-polκ10 is indicated by the black arrow. The size markers to the left of the gel image indicate the positions of protein size standards (Protein Marker, Broad Range (2-212 kDa) (NEB)).

5.2.4 Activity of affinity-purified MBP-polκ fusion proteins

As described above, primer extension assays demonstrated that wild-type pol κ preparations contained DNA polymerase activity. However, when the simple primer extension experiment was repeated with null mutant MBP-polκ10, a similar level of polymerase activity was observed, indicating that most of the activity did not come from the pol κ component of the preparations, but was likely to be due to *E. coli* contaminants (Figure 5-8).

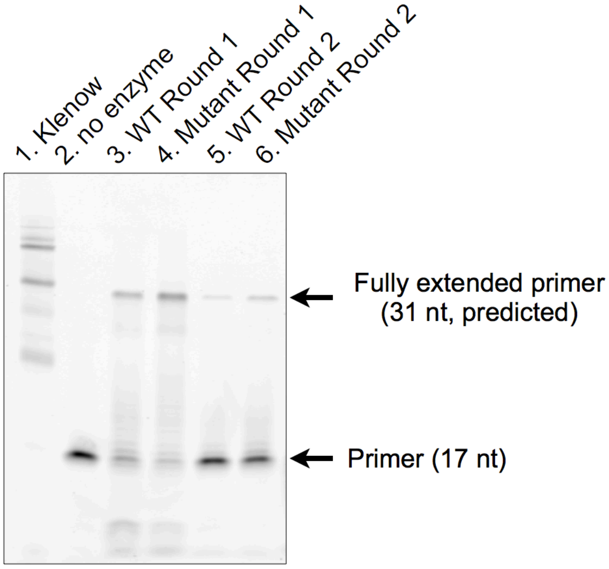


Figure 5-8 – Primer extension assay with affinity-purified wild-type (WT) and mutant MBP-polκ10.
Primer extension assays were carried out as in Figure 5-5 with substrate 1. Black arrows indicate the positions of the starting primer and the fully extended primer. 'Round 1' and 'Round 2' refer to whether the protein preparation was the product of one or of two rounds of affinity chromatography.

5.2.5 Purification of MBP-polκ fusion proteins by inclusion body isolation, solubilisation and refolding

Because recovery of pol κ was already low due to insolubility of much the fusion protein, it seemed more practical to attempt to increase the purity of the pol κ preparations by an alternative purification method rather than by gel filtration of the affinity chromatography-purified protein. The contaminating polymerase activity was presumably due to a soluble *E. coli* protein, perhaps binding to the amylose beads, whereas a large proportion of the MBP-polκ protein produced appeared to be insoluble (Figure 5-3). The approach used, therefore, was to isolate the insoluble MBP-polκ, which would likely be free of soluble contaminant, as native *E. coli* enzymes should reliably be soluble in *E. coli* cultures. I induced expression of MBP-polκ10, and extracted, washed, and solubilised the inclusion body from *E. coli* lysate. I then set up a screen to attempt refolding of MBP-polκ10 (Table 5-1). After refolding, I dialysed solutions to remove chaotropic buffers. After this step, any remaining unfolded protein should have precipitated. I determined the concentration of soluble protein, which indicated good yields of soluble protein (Table 5-1), and PAGE of selected conditions suggested good purification (Figure 5-9A).

Condition	EDTA (mM)	$MgCl_2$ $+CaCl_2$ (mM each)	DTT (mM)	GSH (mM added)	GSSG (mM added)	L-arginine (mM)	guanidine (mM)	Protein (mg/ml approx.)	Yield (mg/ml)			
									MBP-κ10	MBP-κ10 mutant	MBP-κ001	MBP-κ001 mutant
1	1	0	5.25	0	0	0	350	1	0.45	NA*	0.57	NA*
2	0	2	0.25	2	0.2	440	350	1	0.30	0.26	0.60	NA*
3	0	4	0.25	2	0.4	880	350	1	0.38	0.28	0.66	0.40
4	0	2	0.25	2	0.4	0	900	1	0.30	NA*	0.44	NA*
5	0	4	5	0	0	440	900	1	0.36	NA*	0.49	NA*
6	1	0	0.25	2	0.2	880	900	1	0.33	NA*	0.40	0.20
7	0	4	0.25	2	0.2	0	1450	1	0.34	NA*	0.45	NA*
8	1	0	0.25	2	0.4	440	1450	1	0.29	NA*	0.41	NA*
9	0	2	5	0	0	880	1450	1	NA	NA*	0.48	NA*
10	0	0	0.0125	0	0	0	0.35	1	NA	NA*	0.33	NA*
11	1	0	5.0125	0	0	0	350	0.05	<0.01	NA*	<0.01	NA*
12	0	2	0.0125	2	0.2	440	350	0.05	<0.01	NA*	<0.01	<0.01
13	0	4	0.0125	2	0.4	880	350	0.05	<0.01	<0.01	<0.01	NA*
14	0	2	0.0125	2	0.4	0	900	0.05	<0.01	NA*	<0.01	NA*
15	0	4	5.0125	0	0	440	900	0.05	<0.01	NA*	<0.01	<0.01
16	1	0	0.0125	2	0.2	880	900	0.05	<0.01	<0.01	<0.01	NA*
17	0	4	0.0125	2	0.2	0	1450	0.05	<0.01	NA*	<0.01	NA*
18	1	0	0.0125	2	0.4	440	1450	0.05	<0.01	NA*	<0.01	NA*
19	0	2	5.0125	0	0	880	1450	0.05	NA	NA*	<0.01	NA*
20	0	0	0.0125	0	0	0	0.35	0.05	NA	NA*	<0.01	NA*

Table 5-1 – Conditions used for refolding screen and protein yields from each condition. All refolding reactions also included 21 mM NaCl, 0.88 mM KCl and 55 mM Tris, pH 8.2. NA = dialysis was not carried out, NA* = refolding reaction was not carried out.

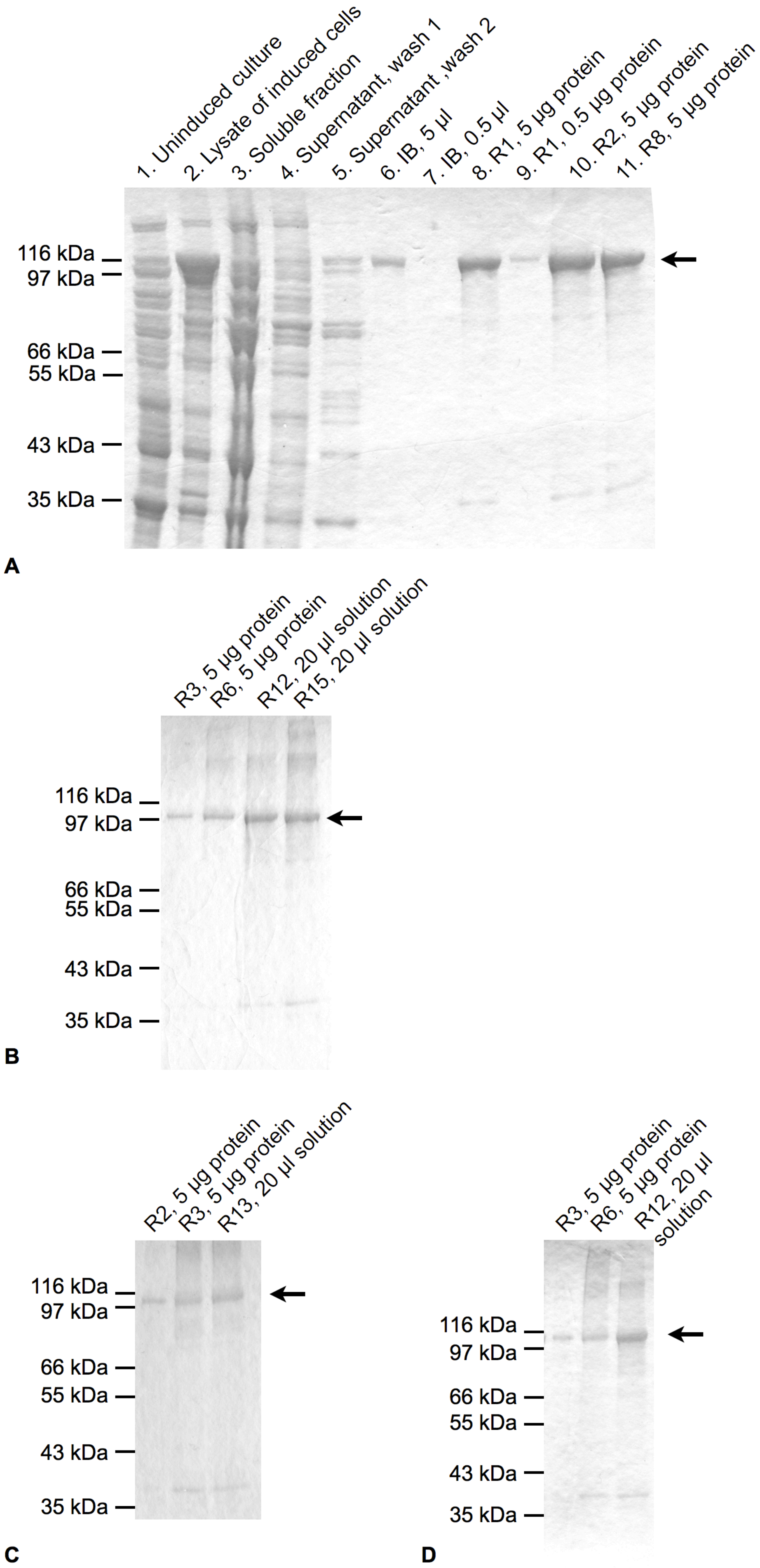


Figure 5-9 – PAGE of stages of inclusion body purification and refolding.

Figure 5-9 continued. A) polκ10-MBP. *E. coli* cells transformed with pMAL-K10 (Lane 1) were induced with IPTG, grown for 2 h at 37ºC, harvested and lysed (Lane 2). The lysate was centrifuged to pellet insoluble material, and the soluble fraction (Lane 3) was removed. Insoluble matter, expected to contain the majority of the fusion protein, was washed twice (Lanes 4 and 5 contain the wash supernatant). The washed inclusion body (Lane 6 and 7) was solubilised in guanidine hydrochloride solution, and refolding was attempted (Lanes 8, 9 and 10). B) polκ001-MBP, products of four refolding experiments, carried out as for polκ10-MBP. C) polκ10-D230A, products of three refolding experiments. D) polκ001-D229A, products of three refolding experiments. Where a volume rather than a mass is given, the protein solution was too dilute for the concentration to be determined, so a fixed volume of was used. IB = solubilised inclusion body; R1 = refolding with condition 1; similarly R2 = refolding with condition 2, and so on. For details of refolding conditions, see Table 5-1. The size markers to the left of the gel images indicate the positions of protein size standards (Protein Marker, Broad Range (2-212 kDa) (NEB)).

I tested the activity of the refolded protein preparations in the primer extension assay as before (Figure 5-10), which demonstrated that several preparations had some polymerase activity. I repeated the expression, purification, solubilisation and the refolding screen using MBP-polκ001, with similar results (Table 5-1, Figure 5-9B). I produced solubilised protein for both mutant pol κs, and refolded each using the conditions that gave the best activity for the corresponding wild-type pol κ. There was no evidence for any contaminants co-purifying with the mutant proteins that were not in the wild-type preparations (Figure 5-9C and D). I compared the wild-type proteins with the mutant proteins using the primer extension assay, which showed similar activity in each, indicating that the contamination problem had not been solved (Figure 5-11).

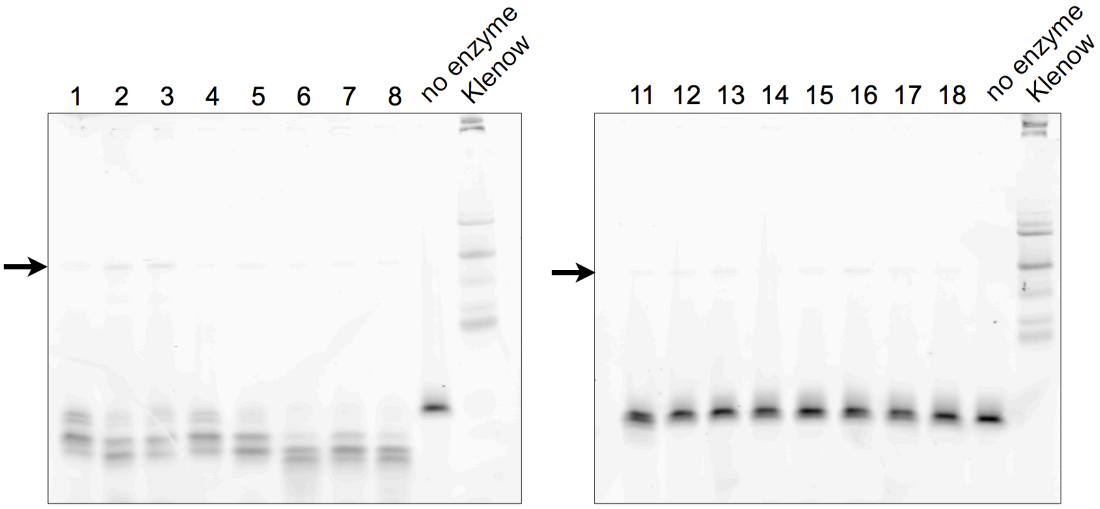


Figure 5-10 – Primer extension assay with products of refolding screen of solubilised polκ-MBP inclusion bodies.
Primer extension assays were carried out as for Figure 5-5, with substrate 1. The substrate was incubated with 1 µg of protein product from each set of refolding conditions, indicated above each lane and numbered as in Table 5-1. Black arrows indicate the location of faint bands corresponding to extended primer.

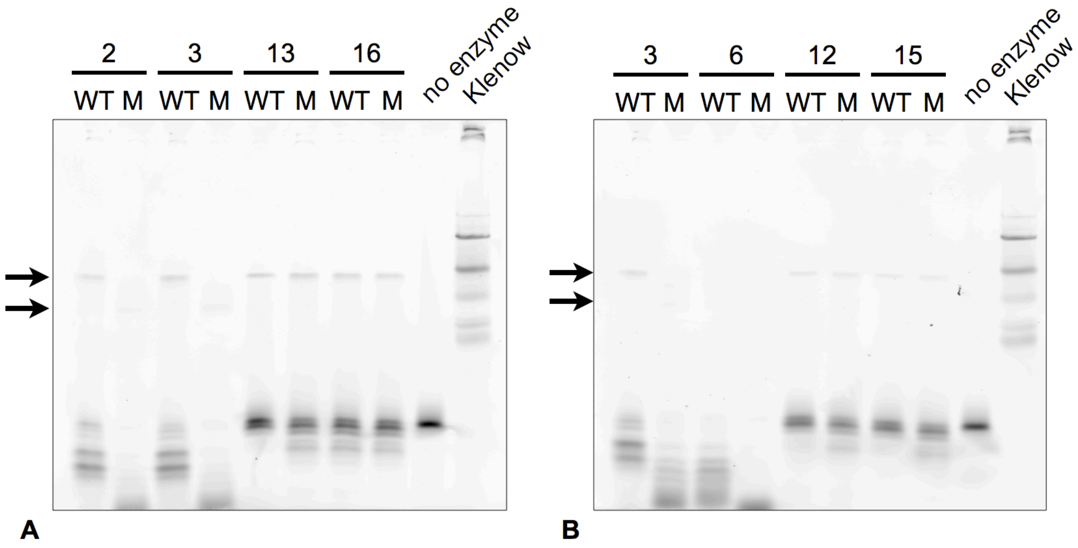


Figure 5-11 – Primer extension assays with refolded proteins showed no difference in polymerase activity between wild-type and mutant MBP-polκ preparations.
A) MBP-polκ10. B) MBP-polκ001. Primer extension assays were carried out as for Figure 5-5, with substrate 1. Black arrows indicate the location of faint bands corresponding to extended primer. WT = wild-type; M = active site mutant. Refolding conditions are indicated above each pair of lanes and are numbered as in Table 5-1.

5.3 Cell biological analysis

5.3.1 Generation and fluorescence microscopy of eGFP-tagged pol κ parasite lines

The 10 pol κ genes exist as a tandem array of highly similar sequences on chromosome 11. This structure meant that it was impractical to tag or to knock out the endogenous genes, previous attempts having failed (J.D. Barry, pers. comm.). Therefore, to study the *in vivo* localisation of the products of the pol κ genes used in the biochemical assays, I generated *T. brucei* cell lines containing tagged ectopic copies of the genes. To achieve this, I cloned pol κ genes from the pMAL vectors into the plasmid pB2X. This vector was the gift of Dr Michal Swiderski, and was generated from pLEW100v5B1D, a variant of pLEW100 (Wirtz *et al*, 1999), by replacing the luciferase gene with a eGFP gene. I carried out PCR with primers P86 and P95 to amplify both pol κ genes from the pMAL vectors, restore the initial ATG codon, remove the stop codon and add flanking *Spe*I and *Hin*dIII restriction sites, then I used *Spe*I and *Hin*dIII to digest the PCR products and the vector, and then ligated the PCR products into the vector. This cloning procedure created the plasmids pB2X-K10-GFP and pB2X-K001-GFP, the sequences of which were confirmed by sequencing with the primers LEWSEQ1 and either P73 and P76 (for pB2X-K10-GFP) or P74 and P77 (for pB2X-K001-GFP).

Each of these plasmids contained a gene encoding pol κ fused to eGFP at its C-terminus, under the control of an RNA Pol I promoter with a Tet repressor binding site; a phleomycin resistance gene under the control of the T7 polymerase promoter; and the sequence of the trypanosome ribosomal RNA spacer, to allow insertion of the construct into the genome (Figure 5-12). I transfected these constructs into the BSF RNAi cell line Lister 427 90:13 (Wirtz *et al*, 1999). This cell line constitutively expresses T7 polymerase, which allowed expression of the phleomycin resistance gene from the pB2X plasmids, and the Tet repressor protein, which suppressed expression of the pol κ fusion protein until tetracycline was added.

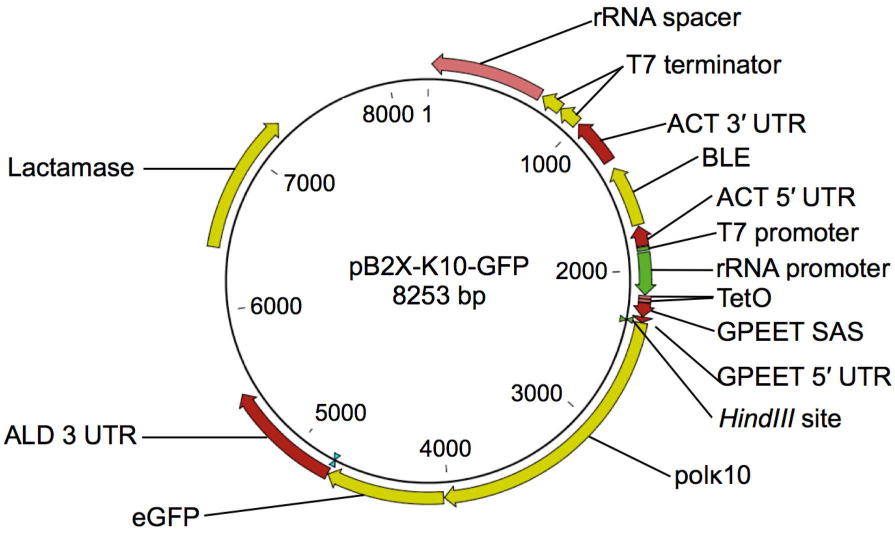


Figure 5-12 – pB2X-K10-GFP, the construct used for polκ10-GFP overexpression. The central circular scale is in bp from an arbitrary position (the 3′ end of the rRNA spacer). Notable features are the rRNA spacer, in which the plasmid is cut for transformation, and which allows the construct's integration into the genomic rRNA array; the phleomycin resistance selectable marker for parasite transfection (labelled BLE); the tag (labelled eGFP); and the pol κ gene (labelled polκ10). The plasmid pB2X-K001-GFP was identical except for the pol κ gene.

Western blot analysis with an anti-GFP polyclonal antibody suggested that an eGFP-tagged protein of approximately the expected size was present in transfectants when expression of the fusion protein was induced by tetracycline, but not in uninduced parasites, for both examined clones from the tagged polκ10 line (Figure 5-13). The same analysis with tagged polκ001 lines did not give a positive result, so results are only shown for polκ10-GFP parasites.

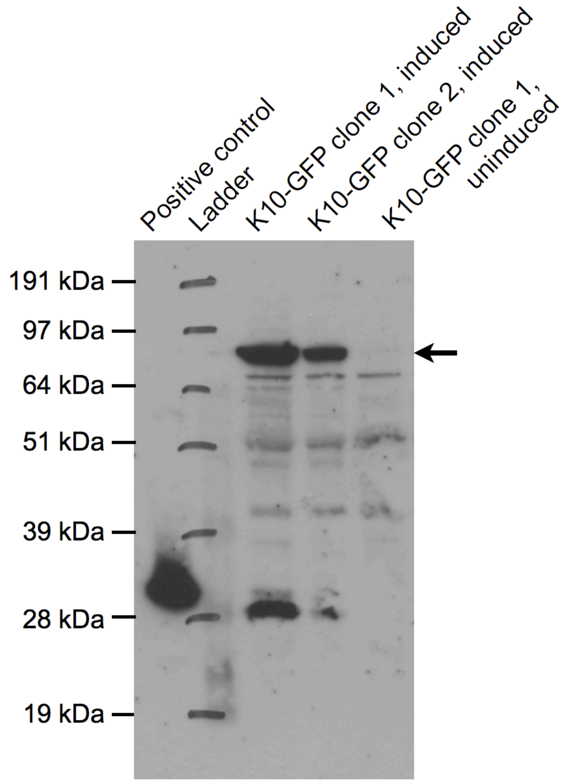


Figure 5-13 – Western blot analysis of parasites transformed with polκ10 overexpression construct.
Parasites that had been stably transfected with pB2X-K10-GFP were grown either with (induced) or without (negative control, uninduced) tetracycline to induce overexpression of eGFP-tagged polκ10. The parasites were harvested and lysed, then the lysate was separated by PAGE and transferred to a nitrocellulose membrane. The membrane was probed with an anti-GFP primary antibody, followed by a horseradish peroxidase-conjugated secondary antibody, which was detected by adding a chemiluminescent substrate and exposing to X-ray film. The positive control was lysate of *E. coli* expressing eGFP from a plasmid. Ladder = SeeBlue pre-stained standard (Invitrogen); the sizes to the left of the image indicate the sizes of these protein standards marked on the image.

I used parasites from two pol κ10 clones in fluorescence microscopy to determine the localisation of eGFP fluorescence. I attempted live cell fluorescence imaging, but did not observe any eGFP fluorescence in induced parasites that was not also visible in uninduced cells (data not shown). I therefore prepared slides for microscopy using methanol fixation. Fluorescence appeared to be localised to the nucleus in induced parasites, and nuclear localisation was confirmed by high-resolution microscopy (Figure 5-14).

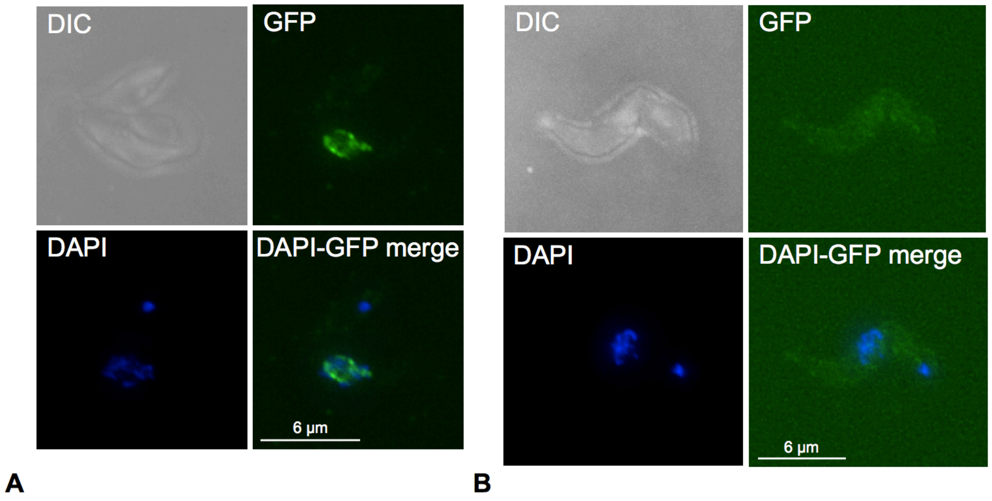


Figure 5-14 – Fluorescence microscopy of polκ-GFP overexpressing parasites. Parasites that had been stably transfected with pB2X-K10-GFP were grown either with (induced) or without (negative control, uninduced) tetracycline to induce overexpression of eGFP-tagged pol κ. The parasites were harvested, fixed with methanol and mounted with Vectashield mounting medium with DAPI (Vector Laboratories). A) Polκ10-GFP parasites with expression induced. B) Control: polκ10-GFP parasites without expression induced. Fluorescence was imaged using a DeltaVision fluorescence microscope, and the images deconvoluted using Softworx software. Projection uses maximum intensity layer. Images shown are representative of the cultures indicated, and of approximately 100 parasites checked in each case, almost all showed the same patterns of localisation. The scale bar indicates 6 µm for all four panels in one part of the figure. DIC = phase contrast image; GFP = fluorescence from eGFP (FITC filter); DAPI = fluorescence from DAPI (DAPI filter); DAPI-GFP merge = images from eGFP and DAPI superimposed by the software.

5.3.2 Fluorescence *in-situ* hybridisation with eGFP-tagged pol κ cell lines

In order to test whether polκ10-GFP co-localises with telomeres within the nucleus, I decided to carry out fluorescence *in-situ* hybridisation (FISH). The work in this section was carried out in collaboration with Dr Barbara Marchetti. We used a PNA probe labelled with Cy5 to label telomere repeats in methanol-fixed cells. However, no eGFP fluorescence was visible in parasites that had been prepared using the FISH protocol (data not shown). It seems likely that this failure was due to the harsh hybridisation conditions denaturing the eGFP sufficiently to abolish fluorescence.

We therefore attempted to use immunofluorescence (IF) to detect eGFP in FISH parasites. However, when IF was followed by FISH, we were unable to detect any IF signal. Possible explanations for this failure were that the FISH protocol disrupted the antibody-GFP interaction, or that it disrupted the nuclear structure itself, dispersing the eGFP-labelled protein throughout the cell. We altered the protocol to try and address this issue: we replaced methanol with the more powerful fixative formaldehyde, to better preserve cell structure; and we re-fixed parasites after incubation with antibody, in order to strengthen the antibody-epitope interaction by cross-linking. However, the IF signal was not improved by these changes.

As can be seen from Figure 5-13, the polyclonal anti-GFP antibody showed a considerable amount of non-specific binding. We decided that in order to optimise the FISH-IF protocol, it would be better to use a tag that can be detected more cleanly than eGFP. Additionally, the large size of the eGFP tag was potentially disruptive to the function of the tagged protein. The main benefit of using eGFP-tagged protein is direct visualisation of the fusion protein through eGFP fluorescence; a benefit that is lost if an antibody is required to detect eGFP.

5.3.3 Generation and fluorescence microscopy of MYC-tagged pol κ cell lines

To replace the polκ-GFP cell lines, I generated polκ-12MYC cell lines. As there was no MYC version of the overexpression vector I used the 12MYC tag from an existing tagging construct (gift of Anna Trenaman). First, I attempted to use PCR with PL4 and PL5 followed by restriction digestion with *Spe*I and *Xho*I and ligation to directly substitute this multiple tag for eGFP in the pB2X-polκ-GFP vectors, but this strategy was unsuccessful, probably due to the small size and repetitive nature of the 12MYC PCR product. I therefore carried out PCR with primers PL8 and PL9 to amplify the 12MYC PCR product and add flanking 5′ *Hin*dIII and *Xba*I and 3′ *Xho*I and *Eco*RV recognition sites, and with primers PL12 and PL13 to amplify the pol κ genes from the pMAL vectors and add flanking 5′ *Hin*dIII and *Eco*RV and 3′ *Xba*I recognition sites. I ligated the 12MYC PCR product into a TA cloning vector, then digested the 12MYC vector and the pol κ PCR product with *Hin*dIII and *Xba*I and ligated the pol κ PCR product into the 12MYC vector, creating polκ-12MYC fusion genes. I used *Hin*dIII and *Xho*I to digest pB2X-GFP and the polκ-12MYC vectors, gel purified the pB2X backbone and the polκ-12MYC fusion genes, and ligated the fusion genes into pB2X to generate the plasmids pB2X-K10-12MYC and pB2X-K001-12MYC. The plasmid sequences were confirmed by sequencing with the primer LEWSEQ2 and the same primers as were used for pB2X-K10-GFP and pB2X-K001-GFP. I transfected pB2X-K10-12MYC and pB2X-K001-12MYC into BSF 427 90:13. Western analysis with an anti-MYC monoclonal antibody confirmed that the cell lines were expressing MYC-tagged proteins of the appropriate size (Figure 5-15).

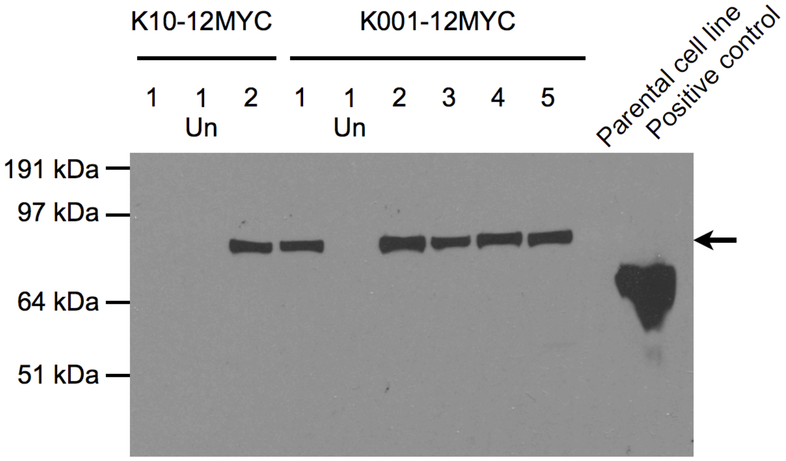


Figure 5-15 – Western analysis of parasites transfected with pB2X-polκ-12MYC constructs. Parasites that had been stably transfected with pB2X-12MYC constructs were grown and the lysate transferred to a nitrocellulose membrane as for Figure 5-13. The membrane was probed with an anti-MYC primary antibody, followed by a horseradish peroxidase-conjugated secondary antibody, which was detected by adding a chemiluminescent substrate and exposing to X-ray film. The positive control was a lysate of *T. brucei* expressing MYC-tagged BRCA2 (gift of Anna Trenaman). Numbers above lanes indicate clone numbers. Un = uninduced. The size markers to the left of the image indicate the positions of protein size standards that were run on the same gel (SeeBlue pre-stained standard (Invitrogen)).

I used the polκ-12MYC cell lines for IF analysis with an anti-MYC polyclonal antibody and a fluorescence-tagged secondary antibody. Initial experiments suggested that both polκ001-12MYC and polκ10-12MYC were located in the nucleus (Figure 5-16). However, the localisation work was not taken further because at this point the conclusions emerging from my biochemical studies of pol κ suggested that my time would better be spent on other aspects of my project.

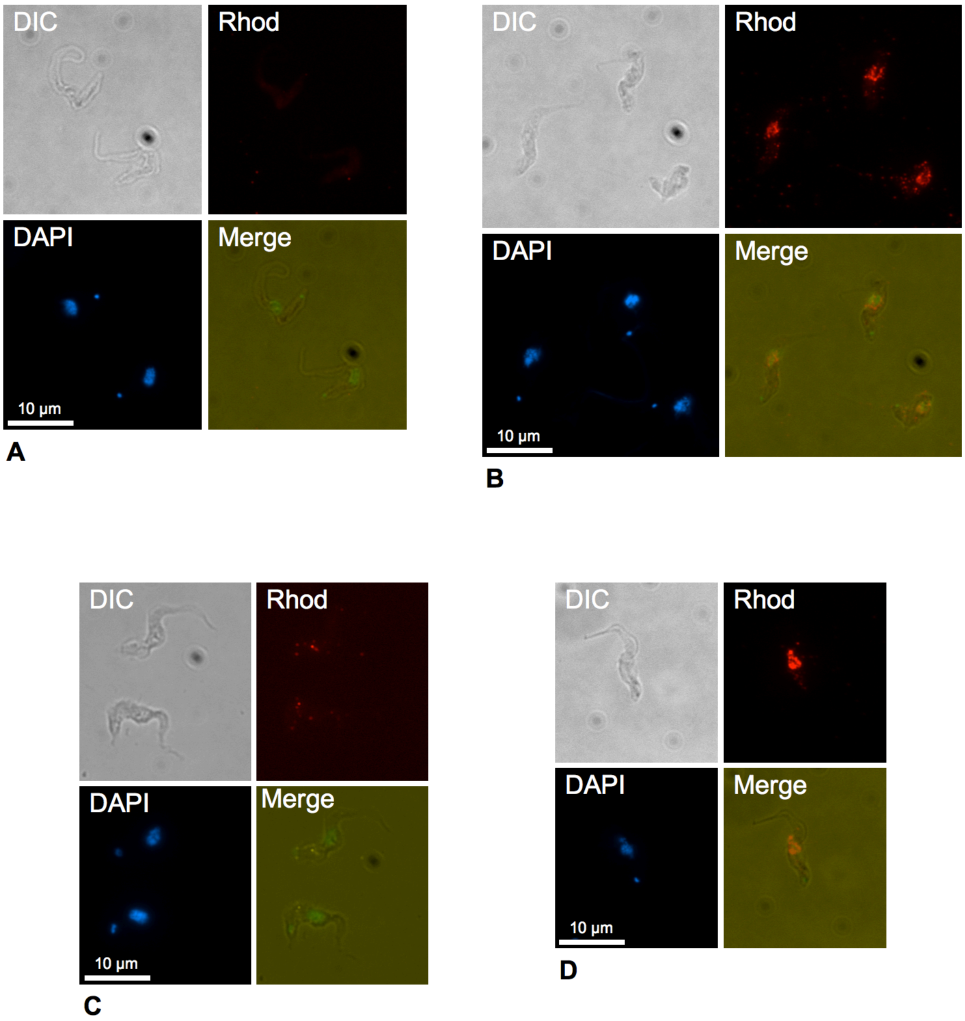


Figure 5-16 – Immunofluorescence microscopy of parasites transfected with 12MYC-tagged pol κ constructs.
Parasites stably transfected with pB2X-K10-12MYC or pB2X-K001-12MYC were grown and fixed with methanol as for Figure 5-14, then used for immunofluorescence with an anti-MYC primary antibody and Alexa594-conjugated secondary antibody. Imaging was performed using an Axioskop 2 fluorescence microscope. A) polκ10-12MYC, uninduced. B) polκ10-12MYC, induced. C) polκ001-12MYC, uninduced. D) polκ001-12MYC, induced. The scale bar indicates 10 μm for all four panels in each section of the figure. DIC = phase contrast image; Rhod = fluorescence from secondary antibody against anti-MYC antibody, indicating location of MYC-tagged pol κ (Rhod filter); DAPI = fluorescence from DAPI (DAPI filter); Merge = images from Rhod and DAPI merged using the Axioskop included software.

5.3.4 Overexpression phenotype of pol κ proteins

In other eukaryotes, including *T. cruzi*, overexpression of pol κ increases resistance to the DNA-damaging agent hydrogen peroxide (Rajão *et al*, 2009). I therefore checked whether overexpression of either MYC-tagged *T. brucei* pol κ gave a similar phenotype, indicating a similar role. I grew overexpression-induced and -uninduced parasites in the presence of 200 μM hydrogen peroxide, and took cell counts after 24 h and 48 h. There was no observable difference between induced and uninduced cultures exposed to hydrogen peroxide, nor between induced and uninduced cultures in standard conditions (Figure 5-17).

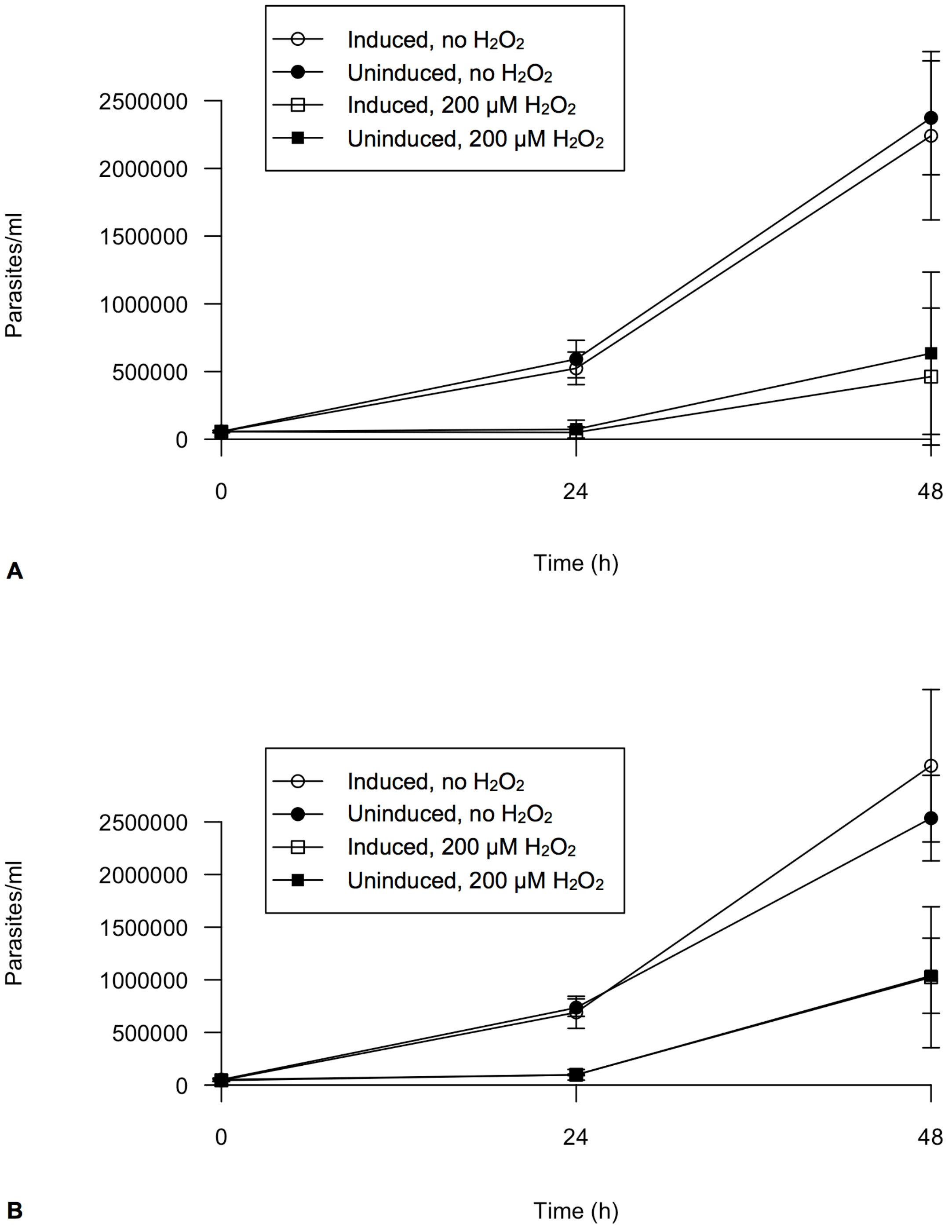


Figure 5-17 – Overexpression of polκ-12MYC proteins has no effect on hydrogen peroxide tolerance in BSF parasites.
Parasites stably transfected with pB2X-K10-12MYC (for polκ10-12MYC) or pB2X-K001-12MYC (for polκ001-12MYC) were grown from a low density in the indicated concentration of hydrogen peroxide, either with (induced) or without (negative control, uninduced) tetracycline to induce overexpression of tagged pol κ. Cell density was determined every 24 h and is shown on the y-axis. Error bars represent one standard deviation. A) polκ10-12MYC. B) polκ001-12MYC.

5.4 Discussion

The work in this chapter aimed to characterise the proteins encoded by two members of the *T. brucei* pol κ gene family, with the aim of determining whether these error-prone polymerases had the potential to be involved in *VSG* hyperevolution. However, biochemical characterisation of the mutation profiles of the enzymes could not be carried out due to persistent contamination of MBP-polκ preparations with *E. coli* DNA polymerase activity. The contamination was detected because primer extension assays with wild-type and with active site mutant pol κ proteins produced approximately equal amounts of product. Contamination was present in preparations of both proteins tested. Further, the contaminating activity was present in preparations using two purification methods, one of which involved purifying insoluble protein from *E. coli* cultures; and a similar contamination has been seen in amylose affinity preparations of MBP-tagged pol κ from *T. cruzi* (C. Machado, pers. comm.). The persistence of the contaminating activity suggests that its presence in the preparations was due to specific association with the MBP-polκ protein, rather than association with the amylose resin used in the first purification attempt. Its presence in multiple experiments with both proteins also makes it unlikely that the apparent polymerase activity of the presumed active site mutants was actually due to a mix-up between wild-type and mutant plasmids or *E. coli* cell lines.

One alternative explanation for the polymerase activity of the mutant preparations is that the mutation did not in fact abolish the pol κ activity, and the polymerase activity actually came from the mutant protein rather than from *E. coli* protein. Activity of the presumed inactive mutant could be due to mis-annotation of the active site, or due to the mutated active site residue not actually being critical for activity. However, due to the very high conservation of the active site across Y-family polymerases (Prakash *et al*, 2005), both possibilities seem unlikely[4]. Contamination with *E. coli* protein is a far more likely explanation for the activity of the mutant protein.

[4] If the active site does not work as it does in other Y-family polymerases, then these *T. brucei* enzymes are essentially not Y-family polymerases. In this case, it would be beyond the scope of this project to characterise what the *T. brucei* enzymes actually are, especially as there is no reason to assume that they have the characteristics of the Y-family that led us to investigate them in the first place.

As we conclude that the polymerase activity in mutant pol κ preparations was due to *E. coli* protein, it seems reasonable to assume that the same *E. coli* activity contributed a similar amount of activity in the wild-type pol κ preparations. Although the amount of product was not quantified in any of the primer extension experiments, visual inspection of Figure 5-8 and Figure 5-11 suggests that usually the mutant produced at least as much product as the wild-type. The amount of product was not determined simply by the amount of substrate added, as in most cases there was still unextended primer visible. These results indicate that most, if not all, of the polymerase activity in the wild-type preparations also came from *E. coli* proteins, which in turn implies that the trypanosome pol κ proteins themselves contributed little or no polymerase activity. This conclusion is supported to some extent by the data from the reversion assay (section 5.2.3.3): although the results were only from preliminary experiments, the number of revertants indicated that the polymerase preparations contain an enzyme with a considerably lower mutation rate than human pol κ, a result that is more consistent with a processive *E. coli* replicative polymerase than with a member of the pol κ family.

The fact that MBP-polκ overexpressed in *E. coli* and enriched *in vitro* had little activity does not necessarily mean that the *T. brucei* proteins are not active polymerases *in vivo*. Importantly, it was not possible to tell from the experiments described whether the enzymes have no activity at all, or whether it was just that the activity was so low as to be undetectable against the background of *E. coli* polymerase activity. Further, low activity in the experiment does not necessarily mean the enzymes have a low activity *in vivo*. Firstly, the activity of the enzyme may have been lost because of degradation during the expression and purification steps. This possibility cannot be excluded, but seems less likely when we consider that multiple preparations of two enzymes using two strategies all gave a similar result, and that a similar purification protocol yielded active *T. cruzi* enzyme. A second possibility is that the apparent low activities of the MBP-polκ fusion proteins were an inherent feature of the experimental setup: for example, the MBP tag may have impeded activity in some way; or the reaction conditions may not have been appropriate; or perhaps the inherent unprocessive nature of lesion bypass enzymes resulted in only a very small amount of fully-extended primer being made in the conditions used. However, as mentioned above, the reaction conditions used (including the tag) have previously been successful for obtaining activity with *T. cruzi* pol κ (Rajão *et al*, 2009); included a divalent cation that has been shown to be an effective cofactor for human pol κ (Gerlach *et al*, 2001); and were able to support activity of *E. coli* Klenow fragment DNA polymerase (*e.g.* Figure 5-5). In an experiment almost identical to that described in section 5.2.3.2, *T. cruzi* MBP-polκ produced an amount of fully extended primer from which the fluorescein signal was easily detectable upon scanning the gel (Rajão *et al*, 2009). It is therefore unlikely that the low activity of *T. brucei* fusion proteins was due to such an inherent problem with experimental design, although of course this explanation is still possible.

A related potential explanation for the lack of observed activity is that the purified enzymes would have been capable of synthesising DNA, but were inactive in the polymerase assays because some factor required for activity *in vivo* was not provided. One possibility for such a factor is another protein that might be required to form a hetero-oligomeric complex with the pol κ used, for example multiple different members of the pol κ family might need to associate together for any of them to be active. A second possibility is that the enzymes were never offered their true substrate, for example if one or both of the enzymes act *in vivo* on a specific DNA adduct. One candidate to be such a substrate is base J. This hypermodified base replaces T, and is found in repetitive sequences, primarily in telomeres, including inactive ES and 177-bp and 70-bp repeats, and also in regions flanking Pol II transcription units (Gommers-Ampt *et al*, 1993; van Leeuwen *et al*, 1997; Cliffe *et al*, 2010). Of the two gene products studied, it would be more likely that polκ001, which has the atypical DNA-binding domain, would be able to accommodate this glucosylated base. No polymerases were identified in a study of proteins binding specifically to base J (Cross *et al*, 1999), but this absence is not necessarily inconsistent with a polκ-base J interaction, because any association of polymerase with DNA would likely be very transient and thus difficult to detect. However, these suggestions are entirely speculative and considerable further experimentation would be required to provide evidence for them.

These alternative explanations cannot be discounted, but after considering them it still remains a likely possibility that the *in vitro* experiments with MBP-polκ described in this chapter indicated that the two genes examined did not encode fully active polymerases. The plausibility of this explanation is strengthened by the results of cell biological studies. In general, overexpression of pol κ increases tolerance to hydrogen peroxide, but in BSF *T. brucei* neither of the pol κ genes examined affected tolerance when overexpressed. Localisation experiments indicated that both tagged, overexpressed proteins were present in the nucleus, so the lack of an effect on tolerance was probably not due to a failure of expression or of nuclear localisation. An obvious conclusion is therefore that the pol κ gene products were unable to carry out lesion bypass DNA synthesis and thus their overexpression did not increase repair activity in the parasite. Of course, this result alone does not demonstrate that the gene products were not active lesion bypass polymerases. A rigorous investigation would include further controls, including quantification of the amount of extra (tagged) pol κ present in transgenic trypanosomes, repetition of the experiment at several concentrations of hydrogen peroxide, and an experiment where overexpression is induced before hydrogen peroxide is added. An additional positive control would be to include *T. cruzi* pol κ in the experiments, to confirm that the problem was with the *T. brucei* enzymes rather than with some aspect of my handling of them. However, when this work is taken in conjunction with the *in vitro* studies of the protein, the results are very suggestive that the pol κ gene products were inactive.

One caveat to this conclusion is that if, as speculated above, pol κ requires an extra factor that was not provided in the *in vitro* experiments, such a requirement could prevent the overexpression of pol κ from having a detectable effect on hydrogen peroxide tolerance. If several proteins were required for activity but only pol κ was overexpressed, the limited supply of the others would mean no increase in overall pol κ DNA synthesis could occur, and hence the parasite would not be protected. Similarly, if pol κ can only act on DNA with a specific adduct, overexpression of pol κ would not increase the rate of repair in the regions of the genome that do not contain the modification. If the adduct were base J, this category would cover the majority of housekeeping genes, and so again the parasite's essential genes would gain no extra protection from damage.

Although there are many caveats, the most likely conclusion from the work described in this chapter is that the two pol κ gene family members examined did not encode active polymerases. The work certainly provided no evidence to support the hypothesis that the pol κ gene family encodes polymerases that are involved in the hypermutation of subtelomeres.

Chapter 6: Discussion

6 Discussion

6.1 Introduction

The purpose of this thesis project was to investigate the processes occurring in hyperevolution of *VSG* genes in *T. brucei*. The specific aims were to test the hypothesis that the subtelomeric *VSG*s are subject to different and more rapid mutagenic processes than are genes in the chromosome cores; and to test the hypothesis that members of the expanded pol κ gene family may have a role in mutagenesis in the *VSG*. The key outcomes and findings of the project were:

1) Good draft quality genomes of EATRO 3 and EATRO 2340 were assembled, with a high level of coverage and a substantial amount of subtelomere sequence.
2) A large number of *VSG* NTD genes were annotated in and paired between the EATRO 3 and EATRO 2340 genome assemblies; the archives were similar to the published TREU 927 *VSG* archive in terms of the degree of pseudogenicity and substructuring.
3) Substitution mutations in *VSG* NTD genes appeared to occur with a higher frequency than and different pattern from those in core genes, indicating that differences in the evolution of the two genome regions were not simply due to relaxation of direct selection on *VSG*, but rather that different substitution mutation processes were operating in each region.
4) Frequency of mutations in the subtelomeres appeared not to differ between *VSG* and likely non-*VSG* sequence, implying that a similar mutation process operates throughout the subtelomeres
5) Exhaustive attempts to demonstrate polymerase activity in two members of the pol κ family failed, indicating that pol κ proteins are unlikely to play a role in promoting mutation in subtelomeres. However, it is also possible that the experimental conditions omitted a crucial *in vivo* factor.

6.2 New trypanosome genomes and *VSG* archive

6.2.1 Genome assembly and annotation

The work described in Chapter 3 provided a preliminary indication that the mutation rate in subtelomeres was higher than that in chromosome cores, and confirmed the close relationship of EATRO 3 and EATRO 2340. The main outcome of the work in that chapter, however, was to produce raw data (*i.e.* genome assemblies) for extensive further analysis. New genomes for two isolates of a *T. b. rhodesiense* strain were assembled using paired-end short sequence reads, taking advantage of new assembly algorithms that combined assembly to a reference genome with *de novo* contig assembly. Although part of the assembly process relied on the previously assembled genome of TREU 927 (Berriman *et al*, 2005), the use of *de novo* assembly allowed a considerable amount of new sequence to be obtained in regions where the two EATRO genomes differed substantially from TREU 927, *i.e.* the subtelomeres. The assembled genomes therefore demonstrate the power of the relatively high-throughput approaches used, particularly when several sequencing technologies are combined, since even such difficult regions as subtelomeres could be assembled from the short reads. The assemblies are of draft quality with numerous gaps, but the analyses in Chapter 3 indicated that the genome assemblies, including likely subtelomere sequence, were of sufficiently high quality and coverage for *VSG* archive analysis. Further, the EATRO 2340 genome assembly likely contained proportionately more (as well as absolutely more) subtelomere sequence than TREU 927: the reference genome contains only one chromosome from each homologous pair, *i.e.* only half the subtelomeres, whereas the EATRO 2340 assembly has retained many of the missing subtelomere sequence in the contigs that were not mapped to chromosomes. The assembled EATRO genomes can be described as 'good enough': they are not as high quality as a painstakingly finished reference genome, but provide plenty of information for a meaningful analysis of *VSG*s and were far less laborious and time-consuming to generate.

A similar evaluation applies to the other automated or semi-automated analyses used throughout the project, particularly gene annotation transfer with RATT, *VSG* NTD gene annotation with VSG-SVM and BLAST, and SNP annotation with SAMtools and project-specific Perl scripts. It is unlikely that every datum considered was absolutely correct. However, due to the large number of data being considered and the stringency of the analyses, it is likely that the broad conclusions drawn from analysing the data will be reliable.

6.2.2 *VSG* archive

The work described in Chapter 4 led to the annotation of 1110 *VSG* NTDs in EATRO 3 and 1797 in EATRO 2340, with 964 from each set identified as being the same gene in both genomes. This EATRO 3-EATRO 2340 archive is probably incomplete. In TREU 927, the observed *VSG* density is approximately one gene (NTD, CTD or both) per 5 kb (Marcello & Barry, 2007b). If only the assembled subtelomere sequence is considered, in EATRO 2340 the density was one NTD per 8.8 kb, which is not tremendously different. However, pulsed-field gel electrophoresis estimated the total size of EATRO 2340 subtelomeres in the diploid genome as approximately 31 Mb. Only 15.7 Mb of subtelomeric sequence was present in my assemblies, and it seems unlikely that the missing sequence contained no *VSG*s. Despite this incomplete coverage, however, there were still far more NTDs annotated in EATRO 2340 than there are in TREU 927 (1797 compared with 809 NTDs in VSGdb), thus this project considerably increased the amount of *VSG* sequence available for analysis even though the EATRO 2340 genome assembly contained many more gaps than TREU 927.

The EATRO 3-EATRO 2340 *VSG* archive contained a high proportion of non-functional NTDs. The NTD annotation strategy made use of sequence translations, so genes that contained no frameshifts would have been easier to annotate, likely meaning that a disproportionate number of the genes recovered would have been intact. The high proportion of pseudogenes in the annotated archive is therefore particularly striking. It is possible that the number of pseudogenes was inflated by sequencing errors, because an error is more likely to render an intact gene a pseudogene than restore a pseudogene to functionality. This possibility is particularly likely when pseudogenes contained a single frameshift, in particular if the frameshift was due to an indel in a homopolymeric tract. There were 352 gene pairs annotated as intact in one or other genome, and 192 of these were intact in both genomes. Of those that were described as intact in one genome and as pseudogenes due to frameshifts or premature stop codons in the other, 25 were pseudogenes in EATRO 3 and intact in EATRO 2340, and 54 were intact in EATRO 3 and pseudogenes in EATRO 2340. The EATRO 3 genome assembly probably contained more errors than the EATRO 2340 genome assembly, so the fact that more of the changed pairs were pseudogenes in EATRO 2340 implies that most changes were the result of genuine changes, and thus sequencing error did not play a major role in generating apparent pseudogenes.

The abundance of pseudogenes in the EATRO 3-EATRO 2340 archive underlined the conclusion drawn from TREU 927 about the importance and necessity to the parasite of making use of sequence from pseudogenes in antigenic variation. In fact, an even smaller proportion of the EATRO 3-EATRO 2340 archive was intact than in TREU 927. Because it is likely that neither archive was complete, however, it is difficult to draw any conclusions from this difference. Furthermore, the number of intact genes in the EATRO 2340 genome was higher than the number in TREU 927 (527 compared with 283), so it is not possible to draw any further conclusions about whether there is a minimum number of intact genes required for successful antigenic variation. Indeed, any such number would be likely to be different between strains of different genotype, given there are differences in growth rates and other relevant parameters.

Examination of closely-related genes within the archive revealed that there was considerable substructuring in both EATRO 3 and EATRO 2340, with 69% of EATRO 2340 NTDs in families of at least 50% sequence identity, compared with 40% of TREU 927 NTDs. The families also tended to be larger in EATRO 3-EATRO 2340 than in TREU 927, where almost all groups were pairs or triplets (Marcello & Barry, 2007b): 10% of EATRO 2340 genes were in families with four or more members. These EATRO 2340-TREU 927 differences are likely at least partly due to the larger number of recovered EATRO 2340 NTDs, but they do suggest even more extensive duplication than was inferred from the TREU 927 archive. The role of duplications in *VSG* evolution is discussed in more detail below. Marcello and Barry (Marcello & Barry, 2007b) found that pseudogenes that were combined to construct expressed mosaic *VSG* genes had high identity to one another, in at least parts of the gene, and concluded that the substructuring of the TREU 927 archive was key in providing multiple donors that were of sufficient identity to assemble mosaic *VSG* genes. This conclusion was reinforced by a more detailed study of the *VSG* mosaics expressed during infection, in which all mosaics examined were found to have been generated from donors from the same subfamily (Hall *et al*, 2013). The extensive substructuring of another *VSG* archive provides further support for this hypothesis, and for the idea that maintaining sets of high-identity genes in the genome is crucial.

The general conclusion to be drawn is that the analyses performed on the EATRO 3-EATRO 2340 *VSG* archive supported the initially surprising conclusions from analysis of the TREU 927 archive (Berriman *et al*, 2005; Marcello & Barry, 2007b), but suggested that the processes inferred from the TREU 927 genome are even more important than previously thought. However, only a few aspects of the *VSG* archive were addressed in this thesis, and the EATRO 3-EATRO 2340 archive provides data for the exploration of more questions, for example the dispersal of gene subfamilies throughout the genome, indications of past segmental conversions and recombination, and detailed examination of the changes that cause pseudogenisation. Further, the project included very little examination of CTDs, but running VSG-SVM on the genome assemblies indicated there were numerous CTDs present, so annotating and analysing these fully would provide a fuller picture of the EATRO 3-EATRO 2340 *VSG* archive.

6.3 *VSG* evolution

6.3.1 Possible caveats

The first potential problem with the work described is the spectre of sequencing and assembly errors that is present in all genome-scale studies. Resequencing experiments did suggest that there were indeed some errors (false positives) among the SNPs considered. However, considerable care was taken throughout the *VSG* analysis to prevent such errors from contributing too much to the SNP analyses. In particular, SNPs were called only when there was good coverage of the base with high-quality reads; and when it was possible to call a SNP at the same position in the gene-pair alignment from sequencing read alignments to both genomes. Further, although the EATRO 3 genome assembly was partially based on the EATRO 2340 genome assembly, most of the EATRO 3 contigs that were not mapped to chromosomes were assembled *de novo*; the fact that many annotated *VSG* genes could be matched up within such *de novo* contigs was an encouraging indication that such difficult genome regions had reasonably good assembly. In general, there are likely to be some errors in the results reported, but as long as the potential for errors is kept in mind, and analysis is focused on the general picture rather than on specific data, the study will be informative. The work also underlines the importance of at least some verification of these type of data by targeted resequencing.

A second point to consider is that for the purposes of *VSG* analysis EATRO 3 was assumed to be the direct ancestor of EATRO 2340, when in fact the relationship may not be quite so direct. As discussed in Chapter 3, EATRO 2340 may well be descended from EATRO 3, but another possibility is that is descended from an ancestor of EATRO 3 dating from between 1940 and 1960 (the MRCA). However, for comparisons of mutation frequency and substitution profile, the precise relationship between the isolates is not critical: whatever the true relationship and time separation of the two isolates, both are identical for the chromosome cores and for the *VSG*s, since the strain to which they both belong appears to have expanded clonally. If the same processes are occurring, we would expect the same results for both the chromosome cores and the subtelomeres; the fact that we do not implies that the regions are not equivalent.[5]

If the relationship is not what was assumed, the substitution profile obtained for *VSG*s may not be the true one. Nevertheless, even with the maximum likely separation (MRCA in 1940), EATRO 2340 has had nearly twice as long to accumulate mutations than EATRO 3 has had (37 years compared with 20 years). If we consider a rough classification of mutations into frequent or infrequent, such classifications are likely to be broadly correct. Further, if mutations are considered only as transitions or transversions the profile would be correct because this classification is independent of the direction of change. For the analysis of synonymous and non-synonymous mutations, the analysis was done first assuming that EATRO 3 contained the ancestral base, and then assuming EATRO 2340 did. The results obtained from both analyses were very similar, so it seems reasonable to conclude that the precise relationship made little difference to the conclusions drawn.

[5] This statement only holds true, of course, if we assume that in a given genome region, the mutational processes are exactly the same in EATRO 3 as in EATRO 2340. This second assumption is much more critical to the work than the first, because if it is not true then there is little point in comparing the subtelomeres at all, but it does not seem an unreasonable one to make.

6.3.2 Distinct substitution processes in subtelomeric genes and core genes

The work provided convincing evidence that the substitution mutation processes occurring in *VSG* NTDs are different from, and faster than, those in core genes, rather than simply that more substitutions persist in *VSG*s because deleterious mutations are not removed by direct selection. Point mutations have previously been shown to be important in *VSG* archive evolution (Marcello & Barry, 2007b), but this is the first work to compare directly the mutations in *VSG* genes and cores, and demonstrate that such mutations arise more rapidly in *VSG*s. This finding of different mutation processes occurring in *VSG* NTDs is consistent with the idea that selection for *VSG* diversity has resulted in the evolution of mutagenic mechanisms that promote diversity, rather than individual *VSG* being selected for (King & Kashi, 2007; Barry *et al*, 2012). A high rate of substitution mutation would increase the turnover of *VSG* sequence, providing material for the expression and/or mosaic assembly of antigenically new variants.

A previous study estimated the mutation rate in *T. brucei* chromosome cores to be approximately 10^{-9} mutations/base/cell/generation (Valdés *et al*, 1996). In the work described in this thesis, the rate of mutation at synonymous sites in core genes was estimated to be $3.5×10^{-9}$ to $1.2×10^{-8}$ substitutions/base/generation, while the mutation rate of *VSG* NTDs was estimated at $9.4×10^{-9}$ to $3.2×10^{-8}$ substitutions/base/generation[6]. The core mutation rate was therefore reasonably consistent with the previous estimate. These data would tend to suggest that the lower estimate of mutation rate is more likely to be correct, *i.e.* that the divergence time of the parasites is at the higher end of the range allowed. However, Valdes and colleagues considered their mutation rate to be a minimum, so any divergence time within the 17 to 57 year range is not necessarily inconsistent with a higher value for the mutation rate.

[6] Note that the ranges given for core and *VSG* mutation rates are not confidence intervals, but were calculated from point estimates of the mutation frequency for each divided by the maximum and minimum number of years likely separating the two isolates, which will be constant for both genome regions, *i.e.* if the *VSG* mutation rate is $9.4×10^{-9}$ then the corresponding core rate is $3.5×10^{-9}$ substitutions/base/generation. Thus the overlapping of the quoted core and *VSG* mutation rate ranges should not be taken to mean that there is no significant difference between the two.

The overall rate of point mutation in *VSG* NTDs was 2.4-fold higher than that in cores. This is not a dramatic difference, for example it is not on the scale of the million-fold increase in the mutation rate in somatic hypermutation in B-cell activation (Odegard & Schatz, 2006). As discussed in Chapter 4, there were likely to have been more false negatives in the *VSG*s than in the core genes, so this analysis probably underestimated the difference between the two point mutation rates. Further, substitution mutations account for only one of several mutation processes that act on *VSG*s (Marcello & Barry, 2007b). The other processes include segmental conversions, indels and recombination, all of which are potentially more damaging to core genes than are point mutations. They will therefore likely be tolerated much less than substitutions in the chromosome cores by the DNA maintenance machinery, and so the differences in the rates of these processes between cores and *VSG*s are likely to be higher than for substitutions. Considering both of these points, it seems likely that the overall difference in mutation rate between core genes and *VSG*s is considerably higher than the given estimate from substitution frequencies. Even a small increase over the background mutation rate would be useful to generate new variation, and such an increase is provided in substitution frequency. However, given the reasonably small rate difference in substitutions between core and *VSG*s it is likely that there is a major contribution from other mutagenic processes that will further elevate the mutation rate in *VSG*.

The patterns of substitutions, *i.e.* the relative contribution of each specific substitution to the total, differed between *VSG* NTDs and core genes, even when the difference in base composition between the two regions was taken into account. Because of uncertainty in the direction of mutation, it is probably unwise to draw firm conclusions from the precise pattern. However, one notable feature is robust to changes in assumption about the direction or strand in which the mutation occurred: the mutations in the core were primarily transitions (interchanges between the pyrimidines C and T or between the purines A and G), whereas mutations in *VSG* had a substantial contribution from transversions (figures Figure 4-6 and Figure 4-7). Across prokaryotes and eukaryotes there is a general bias in mutations in favour of C:G pairs mutating to T:A pairs, and this bias is likely to be at least partially the result of deamination of methylated C to produce T (Lynch, 2010a; Hershberg & Petrov, 2010). Transitions between the chemically similar pyrimidines or purines therefore tend to dominate mutation spectra, and such a profile is what we observe in the trypanosome chromosome cores. The *VSG* mutation profile is therefore somewhat atypical. The increased frequency of transversions hints at a mechanism that actively promotes mutations of several different types, rather than relying on the accumulation of mutations that occur as a by-product of replication. A related possibility is that the general bias towards transitions arises because transversions may be easier to repair, perhaps because the greater differences between chemical structures interchanged in a transversion mutation make a substitution easier to detect. In this case, the *VSG* profile could reflect a relaxation of repair activities rather than the induction of new mutations. The observed *VSG* profile may therefore be due to active mutagenesis, relaxed error-correction, or a combination of both these processes.

An observation related to the mutation profile is that mutations in *VSG* NTDs had a much more noticeable strand bias than did those in core genes. That is, the pattern of substitutions in *VSG* NTDs suggested that the mutations that occurred on the non-coding strand were different from those on the coding strand. The existence of this phenomenon is generally inferred from the compositional strand bias it generates, and it has been extensively studied in bacteria (Frank & Lobry, 1999; Francino & Ochman, 2001), and has also been shown to occur in eukaryotes (Niu *et al*, 2003; Touchon *et al*, 2005). Broadly, asymmetry proceeds from the two strands being in different environments, such as occurs between the leading and lagging strands in DNA replication, and between the template (non-coding) and non-template (coding) strands in transcription, the two main causes of strand bias so far identified (Francino & Ochman, 2001; Rocha, 2004; Mugal *et al*, 2009). Given that the observed difference in *T. brucei* NTD mutations was between the coding and non-coding strands, an obvious conclusion would be that the mutation bias was the result of transcription. However, involvement of transcription seems unlikely because the array *VSG*s have no promoters and so are probably never transcribed *in situ* (although see section 6.3.5 for a discussion of one scenario in which transcription might be able to contribute to *VSG* mutation strand bias). Because the *VSG* genes are arranged in directional arrays (usually oriented away from the telomeres), however, a link is possible between leading or lagging strand and coding or non-coding strand: in a directional array, if the first gene's coding sequence is on the leading strand, so will be the coding sequence of all other genes in the array. Such a correlation seems a more plausible explanation of the observed strand bias than the existence of an entirely novel and undescribed mechanism generating strand bias, although a novel mechanism cannot be ruled out. A further consideration is that one important cause of strand mutation bias in replication in bacteria is the higher rate of deamination of C to T in the lagging strand, probably due to the template strand being single-stranded for longer than in leading strand replication (Rocha, 2004). Such a process occurring in the non-coding strand could be a contributing factor to the A-richness observed in the coding strand, as T mutations accumulated in its complement. It should be noted that the A+T content of the *VSG* genes was not atypical, and the genes examined had a G+C content of 50.4%, compared with 51.1% in core genes and 44.3% in the genome overall; it was the distribution of As and Ts between the coding and non-coding strands that was biased.

The arrangement of directional arrays of genes also applies in the chromosome cores, so a difference in the observable strand bias of coding and non-coding strands would imply some difference in the replication mechanism of the cores and *VSG* arrays. This idea receives some support from work carried out to locate replication origins in *T. brucei*, which identified no evidence for replication origin activity in the *VSG* arrays, despite a high density of binding of the ORC1/CDC6 protein that appears to constitute the trypanosome origin recognition complex (Tiengwe *et al*, 2012a; 2012b). In summary, the observed difference in the strand bias of cores and *VSG* NTDs underlines the different mechanisms operating in each region, and the bias in *VSG* NTDs may occur due to the mechanism of replication, and perhaps involves deamination.

A large number of mutations in the *VSG* NTDs (252 out of 442 total) were predicted to change the sequence of expressed VSG, and thus had potential to contribute to antigen diversity. It was found that there was a higher number of synonymous mutations per synonymous site than non-synonymous mutations per non-synonymous site, with a d_N/d_S ratio of approximately 0.4. However, as discussed previously, this value is unlikely to indicate that the *VSG* genes are under purifying selection, because of the short time separating the two genomes and the nature of *VSG* expression. The principal point to note from the analysis of the synonymity of mutations is that there was a high frequency of non-synonymous changes, and in fact the number of non-synonymous changes was higher than the number of synonymous changes. If we consider selection to be acting to produce a mutagenic mechanism that elevates the diversification of the archive, the particular facet of such a mechanism that would be operated on by selection would be the amount of epitope diversity it produced in the expressed protein. That is, a high rate of non-synonymous mutation would produce more diversity, and hence would be favoured. The effects of an elevated synonymous mutation rate, however, would be much smaller. There might be some cost to the parasite if a synonymous mutation resulted in change away from the optimal codon, but although there is codon bias in trypanosomes, it appears that the bias is less important in *VSG*s than other genes, probably because mechanisms stronger than optimal codon usage promote high VSG expression (Horn, 2008). There would be no particular selection pressure for an elevated synonymous mutation rate, but neither would there be strong selection against it. In order to increase non-synonymous mutations, therefore, the parasite could evolve a mechanism that promoted mutations at both kinds of site. Indeed, a mechanism that increases mutations generally is easier to envisage than a mechanism that specifically acts to produce non-synonymous mutations, since a particular base change can be either synonymous or non-synonymous depending on the sequence context. The particular ratio of d_N to d_S would then result from the interaction of the mutation mechanism and the specific base and codon composition of the *VSG* archive; such interactions are accounted for in more sophisticated approaches to estimating and comparing d_N and d_S, (*e.g.* Zhang *et al*, 2006), but were not taken into account here because of a lack of specific information about them.

The study found no evidence that substitution mutation frequency in *VSG* NTDs was higher than in the subtelomeres in general. The work therefore supported the hypothesis that *VSG* NTDs had a higher frequency of mutation than did chromosome cores because the subtelomeres as a whole experience this elevated frequency, and hence hypermutation of *VSG*s is a consequence of their subtelomeric location. More detailed analysis would be required to allow the hypothesis to be accepted, but this result provides a solid foundation to be built on by further work such as comparison of the mutation profiles in subtelomeres within and outwith *VSG*s. The location of *VSG*s in the subtelomeres is therefore likely to be an adaptation to promote diversity by exposing the genes to this region's hypermutational environment, and can therefore be viewed as part of the consequences of the second-order selection envisaged to operate to produce mechanisms promoting antigenic variation (Barry *et al*, 2012). The adaptive nature of *VSG* genomic location has long been assumed, because of the generally elevated rate of ectopic recombination in subtelomeres, but this work provides some evidence that the same principle applies to point mutations.

6.3.3 Further experiments: other mutagenic processes

For practical reasons, the initial analysis of *VSG* hyperevolution described in this thesis focused on point mutations. However, it seems unlikely that point mutations are the only important process in *VSG* evolution. In this section I will discuss other processes that may be involved, and how the data presented in this thesis could be used to examine them. The data generated in the project hold much more information about *VSG* evolution than was extracted in the analyses discussed, but detailed examination of other components of hyperevolution would require the development of more software and analytical tools, and was not possible in the timescale of the project.

6.3.3.1 Segmental conversions

The substitution mutations identified in *VSG* NTDs were screened for possible segmental conversions using a simple method of searching for clusters of SNPs flanked by sequence that was identical between the isolates. None were found by this method, but this negative result does not demonstrate that segmental conversions do not occur in archive evolution. Firstly, as discussed in Section 4.4.2, the method used for identifying SNPs is likely to have missed base substitutions that are due to conversions, because it relied on mapping reads to the genome assembly, and reads with many mismatches would have been discarded or mapped elsewhere. Secondly, many of the genes in the archive had highly similar partners within the same genome, and conversion between such highly similar sequences might not generate enough differences to be detectable by the method used. Interestingly, a gene with at least one SNP was more likely to have a second SNP than would be expected if mutations occurred independently, although none of the genes with multiple SNPs were annotated as having segmental conversions. One process that could bring about these observations would be segmental conversions between sequences that were very similar, so that the converted tract did not show up as a cluster of SNPs, but multiple SNPs were introduced at once. (A second, more speculative, possible explanation for the observations is discussed in section 6.3.5.)

Finally, the method used was somewhat simplistic, and more reliable results could be achieved by the method used by Gjini and colleagues (Gjini *et al*, 2012), who searched for segmental conversions using a much more sophisticated HMM approach to separate regions with differing probabilities of mutations, corresponding to segmental conversions and regions where SNPs are due to point mutations. In summary, the analyses performed here do not provide evidence that segmental conversions do occur, but given that other studies have shown that conversions are likely to occur and to be important (Marcello *et al*, 2007b; Hall *et al*, 2013), this work does not provide enough evidence to conclude that they do not.

Segmental conversions could be searched for more closely in the data available using the same fact that may have prevented their identification in the current study, *i.e.* that if a segmental conversion has occurred in one genome, reads from the converted tract from that genome will have many mismatches to the same region in the other genome, and will probably be poorly mapped. The mapped reads could therefore be screened for regions in one genome where the number of reads from the second genome mapped to a *VSG* drops over a short stretch of the gene. Strategies based on a similar principle are used to look for copy number variation in related genomes, with the difference that such strategies look for regions of higher rather than lower than expected coverage (*e.g.* Simpson *et al*, 2010). In the first SNP annotation method attempted, when assembled genes were compared and the SNPs were not double-checked using sequencing reads, numerous potential segmental conversions were found; this method could be used in combination with screening for low-coverage regions to identify the most likely locations of segmental conversions. As has been demonstrated by the work of this project, the likely locations would need to be confirmed by targeted resequencing, but such a combinatorial approach would identify which regions should be targeted in this way.

6.3.3.2 Indels

Insertion-deletion mutations are more difficult to study than substitutions, because correct identification of mutations relies on correct alignment of genes, which is made less likely by the presence of an indel. This problem is exacerbated in large data sets where it is unfeasible to manually check individual alignments, and also applies to aligning sequencing reads, which is the reason that SAMTools is not reliable when calling indels, and hence why indels were not considered in this study. Furthermore, the types of sequence where indels are most likely to occur naturally, such as repetitive sequence and homopolymeric tracts, are also the types of sequence most prone to sequencing errors artificially introducing indels. Analysis of TREU 927 suggested that indels do occur in *VSG* genes with reasonably high frequency, because there are a considerable number of *VSG*s that are pseudogenes due to frameshifts (Berriman *et al*, 2005; Marcello & Barry, 2007b). At least some of the changes that appeared to cause intact genes in EATRO 3 to become pseudogenes in EATRO 2340 were frameshifts, and so presumably due to indel mutations, although it was not ruled out that these sequence changes could have been caused by sequencing errors. However beyond this it is difficult to infer from the project's data any information about how indels occur in *VSG* evolution.

6.3.3.3 Large-scale rearrangements

In this project, analysis has been focused on small-scale changes. The effects of rearrangement of large chromosome segments have not been considered, mainly because I was interested in mutations that would affect the sequence of individual *VSG* genes, and such rearrangements would tend to translocate entire *VSG* cassettes through exchange delimited by the recombination-prone 70-bp repeats, and/or the 3′ end of the coding sequence, as occurs in recombinational switching, or via different copies of the retroelement *ingi*, which are dispersed throughout the *VSG* arrays (Marcello & Barry, 2007b). Ectopic recombination of chromosome segments is a common feature in the subtelomeres of other organisms (Louis & Vershinin, 2005; Freitas-Junior *et al*, 2000), and there are several possible effects on antigenic variation. Although crossing-over points would be most likely to occur between cassettes, recombination could result in the fusion of the N-terminal portion of one *VSG* gene to the C-terminal portion of another. Recombination would also change the genomic environment of a translocated gene, which may affect which other genes it could interact with, and may also affect the activation probability of the gene, although the factors determining activation probability are not well understood (Morrison *et al*, 2005; Gjini *et al*, 2013). The extent of large-scale rearrangements between EATRO 3 and EATRO 2340 can be examined by taking advantage of the paired-end reads generated for the project: if one read of the pair is mapped far away from the other in one genome but not the other, this would indicate a rearrangement. Such approaches have been used previously to characterise structural variation in genomes (*e.g.* Campbell *et al*, 2008).

6.3.3.4 Gene duplications

Tentative evidence of gene duplication in the time-scale of the study was obtained from the reported *VSG* analysis, provided by several instances where there were two NTDs in one genome with high pairwise identity, but a partner for only one could be identified in the other genome. Indications of the broader importance of duplication in the evolution of the *VSG* archive were given by the extensive substructuring of the archive; and by the presence of numerous SNPs called as heterozygosities at the same positions in the same EATRO 3 and EATRO 2340 genes, which implied the existence of very similar, unassembled homologues of genes that were in the assembly. The importance of gene duplication in the assembly of mosaics has been discussed above. A complementary role for gene duplication in archive evolution is also possible, because duplication of intact genes or genes with intact domains would provide another starting point in which mutations could accumulate and generate diversity.

6.3.4 Molecular mechanisms of hyperevolution

One aim of the project was to test the hypothesis that *T. brucei* pol κ proteins were involved in *VSG* hyperevolution. However, no evidence was found to support the hypothesis. In fact, one interpretation of the results obtained was that the genes tested did not encode fully functional polymerases. However, homologues of pol κ are found from bacteria to vertebrates (Gerlach *et al*, 1999), suggesting an important role in the cell, and it seems peculiar that

T. brucei should have tandemly duplicated a non-functional gene family. It therefore seems unlikely that the entire pol κ gene family is non-functional. In fact, the apparent lack of functionality of the gene members examined may be connected with the gene family expansion: if a gene produces a product with low functionality, then up-regulating its expression is one strategy to compensate for the low functionality, and expression of multiple copies of genes appears to achieve this up-regulation in *T. brucei* (Jackson, 2007b). Such a duplication (or series of duplications) is considerably more likely to occur than mutations repairing the damaged gene. A possible alternative explanation for the apparent lack of functionality, as discussed in section 5.4, is that the experiments failed to demonstrate polymerase activity in the gene products because the reaction conditions omitted a necessary *in vivo* factor, for example a modified DNA substrate, or a second protein that was required to associate with pol κ to form an active polymerase. *In vitro* experiments with purified pol κ from other species, including another trypanosome species, have been successful (Ohashi *et al*, 2000a; Rajão *et al*, 2009), but it does not necessarily follow that *T. brucei* pol κ proteins must be able to act in isolation. The suggestion that *T. brucei* pol κ acts more-or-less exclusively on a modified DNA substrate could be seen as an extreme version of the tendency of lesion bypass polymerases to be able to bypass certain DNA lesions much more effectively and with higher fidelity than for other lesions (Ohashi *et al*, 2000b; 2000a; Bjedov *et al*, 2007; McCulloch & Kunkel, 2008). A specific case of this general explanation is that the *T. brucei*-expanded pol κ gene family could act specifically on sequences containing base J, which is an intriguing possibility, particularly because of the enrichment of base J in repetitive regions and trypanosome telomeres. However, the experiments described here did not provide any information that could inform such a hypothesis, so the possibility remains speculative.

The work on pol κ described in this thesis therefore did not support the hypothesis that the enzyme is involved in *VSG* hyperevolution, but neither did they conclusively rule out pol κ involvement. Comparisons of the EATRO 3 and EATRO 2340 genomes indicated that point mutations occurred at a higher frequency in *VSG* NTDs than in core genes, and DNA polymerases remain a likely candidate for introducing such mutations. The observed mutation strand bias in *VSG* NTDs also provides a clue to the mechanism of mutagenesis because, as discussed above, it implies that the process of DNA replication contributes to the difference between *VSG* and core genes.

It would be feasible to compare the mutations observed with the reported error profiles of different polymerases, but this approach is unlikely to be helpful on its own for two reasons. Firstly, no trypanosome polymerases have had their error profiles established, so we have no idea how conserved the profiles might be, given the genes and proteins themselves are substantially different in trypanosomes from human or yeast polymerases. Secondly, and more fundamentally, it is likely that the observed substitutions were the result of the interaction of multiple processes, so comparing them with the profile of an individual polymerase can only be useful in terms of which mutations the polymerase is able to cause, rather than considering any of the relative proportions of mutations[7]. There are several other approaches that could shed light on other candidate mechanisms. The continuing characterisation of the mechanisms of DNA replication in *T. brucei* may provide insight into mutation in subtelomeres. A technology such as PICh (Proteomics of Isolated Chromatin segments, Déjardin & Kingston, 2009), which uses a specific nucleic acid probe to pull down a particular DNA sequence and the proteins associated with it, could be used to identify proteins that associate with the *VSG* arrays, although the likely transient nature of the interactions might limit the usefulness of this approach.

Finally, whatever factors are involved in hypermutation, the mechanistic basis of the difference in mutational processes between cores and subtelomeres remains to be examined. One broad possibility is that a difference between cores and subtelomeres could be achieved by the physical interactions of each with different structures, or their localisation in different parts of the nucleus. Such differential localisation of chromosome regions can already be seen in the tethering of telomeres at the nuclear periphery, which could bring subtelomeres into a somewhat different environment from chromosome cores (Dubois *et al*,

[7] If it had been possible to determine the error profile of pol κ, the argument would still have applied that merely having a profile consistent with what was seen in *VSG*s would not have been sufficient to demonstrate involvement in hyperevolution. However, pol κ was worth investigating in this context because there were particular reasons to think that it could have a role in hyperevolution, namely the gene family expansion, and the likely more permissive substrate binding site, and because the localisation of the protein was also investigated and could have provided another strand of evidence.

2012), although such an arrangement would probably produce a gradient of activity rather than a sharp delineation. A second possibility is linked to the finding that sequence in the subtelomeres, specifically the repetitive, AT-rich 70 bp repeats, is particularly prone to DNA breaks, and that the repair of these breaks by recombination may be a trigger for VSG switching (Boothroyd *et al*, 2009; Alsford *et al*, 2009). Although the original study was focused on switching, the idea that DNA sequence can be adapted to be easy to damage, for the sake of the consequences of repair activity, is also applicable to subtelomere mutation. Such a mechanism might rely, as switching is proposed to do, on the properties of the demonstrably destabilising TAA:TTA motif in the 70 bp repeats (Boothroyd *et al*, 2009; Ohshima *et al*, 1996). Alternatively, a potential mechanism might involve some other sequence feature prone to causing blockages that require repair, perhaps by error-prone lesion bypass polymerases during replication — a possibility to which the observation of strand bias in *VSG* mutation is relevant, because of its possible links to replication, as discussed above. However, any sort of DNA repair has the potential to introduce errors, so a mutagenic mechanism that involved inducing repair need not necessarily involve polymerases. A further means to target subtelomeres, which would be relevant if polymerases are involved, has been hinted at above: perhaps the target genome regions contain some DNA modification, such as base J, that the polymerases require to function. Whatever the process, the evidence provided in this thesis that some such mechanism exists to allow mutational mechanisms to differentiate between subtelomeres and cores provides a fascinating basis for future work on the subject.

6.3.5 Potential for a role for the expression site in archive evolution

Array *VSG*s are activated when an entire gene or a part of it is copied into the active ES by a mechanism involving homologous recombination (Barry & McCulloch, 2001). This process is asymmetrical: the newly activated *VSG* now has two copies in the genome, but the previous *VSG* is lost from the ES. If the previous *VSG* had been copied from an intact array *VSG*, then the gene will persist in the genome despite being lost from the ES. However, if the previous *VSG* was a mosaic, then this switching mechanism probably means that the replaced gene would be lost altogether. It is assumed that this conversion process only occurs in the array to ES direction, but the possibility of duplicative transposition from the ES *VSG* into the arrays has not really been examined, although it has been hypothesised that switching by telomere exchange between the ES and minichromosomes could act to preserve ES mosaics (Taylor & Rudenko, 2006). This hypothesis is consistent with the observation that many minichromosomal *VSG*s are important early in infection, and so are presumably intact (Liu *et al*, 1985; Robinson *et al*, 1999).

If it were possible for *VSG*s to be copied from the ES into the arrays, events in the ES would have the potential to contribute to archive evolution. Such a process might confer the ability to retain mosaic genes as new, intact array *VSG*, which could contribute to continuing regeneration of the archive. Additionally, there is some evidence that, although point mutations are unlikely to contribute to increasing the scope of antigenic variation within an infection, the process of being copied into the ES and transcribed could in itself be somewhat mutagenic: single base mutations in the ES copy compared with the genome copy have been observed, albeit infrequently (Graham & Barry, 1996; Hall *et al*, 2013); and, more generally, there appears to be a link between transcription and an elevated mutation rate, which has been inferred from genomic compositional strand bias (Francino & Ochman, 2001; Mugal *et al*, 2009). The results reported in this thesis provide some hints that such retention of ES *VSG*s might be a possibility. Firstly, one hypothesis to explain the overdispersion in the number of SNPs per gene could be that those NTDs with an unexpectedly large number of mutations represented genes that had been copied back from the ES. Secondly, the mutations observed in *VSG* NTDs did appear to have a strand bias. Finally, there was some indication that genes that were pseudogenes in EATRO 3 had become intact genes in EATRO 2340. However, intriguing as such a possibility would be, there are alternative explanations for all these observations: genes with many SNPs could be the result of segmental conversion within the archive; the strand bias could be associated with replication rather than transcription, as discussed above; and it was not possible to rule out technical artefacts as the reason for pseudogenes that appearing to become intact genes. Therefore, the hypothesis of a contribution to *VSG* evolution from ES will probably need to remain as speculation.

6.4 Wider implications of the work

6.4.1 Antigenic variation

The finding of this work that is perhaps most relevant in considering the broader picture of antigenic variation is the rate of evolution of *VSG*s. Leaving aside the observation of non-random distribution of mutations, the estimated mutation rate of 9.4×10^{-9} to 3.2×10^{-8} mutations/base/generation implies that a typical *VSG* NTD of 1000 bp would be expected to have a substitution mutation less than once per hundred thousand generations. Even for a repertoire of several thousand genes, this rate implies a given parasite could wait tens of generations for any of its *VSG*s to change. It therefore seems reasonable to conclude that point mutation in *VSG* NTDs, rapid though it is compared with the process in chromosome cores, is unlikely to be able to generate new sequence quickly enough to make a substantial contribution to antigenic variation within a single infection of a single host. This finding complements previous work suggesting that single point mutations are unlikely to change VSG sufficiently to allow evasion of the polyclonal antibody response (Graham & Barry, 1996; Hall *et al*, 2013). It can certainly be envisaged that an antigenically novel VSG could be encoded by a mosaic created by combining parts of various genes each containing point mutations, but again the rate of point mutation suggests that VSGs from this scenario are unlikely to contribute significantly in the relatively short term of a single infection.

The role of *VSG* hyperevolution, then, is more likely to be in the promotion of VSG diversity in the longer term, and on a larger scale. The accumulation of point mutations over a longer timescale would introduce new diversity to the archive, which could be combined in new ways by recombination. Such a process could give an advantage at the scale of the parasite population, if it were to provide enough novel diversity to overcome immunity in a host that had previously been exposed to a similar strain, *i.e.* the development of strain-specific repertoires could facilitate reinfection. An effect at this level is consistent with the idea that *VSG* hyperevolution has evolved under second-order selection for VSG diversity: such a selection process presumably occurs at the level of the population rather than the individual (Barry *et al*, 2012).

6.4.2 Subtelomere hyperevolution

Do the reported results provide simply an interesting story about the evolution of a family of trypanosome genes, or can wider conclusions be drawn? Most studies of subtelomere evolution have focused on the effects of recombination, so it is an important finding of this study that another mutagenic process in trypanosome subtelomeres, point mutation, appears to be not only elevated but also mechanistically distinct from what occurs in chromosome cores. Substitution processes are difficult to study in the subtelomeres of other organisms because these regions tend to be more repetitive than in trypanosomes, but the finding has particular relevance for other organisms that have subtelomeric antigenic variation gene families. A second broadly relevant conclusion is that the *VSG*s hypermutate simply because of their location in subtelomeres. This study of the evolution of the trypanosome *VSG* arrays therefore underlines the importance of subtelomeres in providing an environment for hyperevolution, and complements previous findings that movement of a gene family into subtelomeres can trigger the expansion and accelerated evolution of that family (Glusman *et al*, 2001; Brown *et al*, 2010).

6.5 Concluding remarks

The significance of subtelomeres to variation only relatively recently became apparent, when the availability of genome sequences meant it became possible to explore their contents in detail, but the power of subtelomeres to drive evolution is now widely appreciated. Subtelomere hyperevolution has been studied in a range of organisms, but trypanosomes stand out due to the scale of their principal subtelomeric gene family and its critical role in their survival strategy. The work presented here has given some hints about how hyperevolution and mutation proceeds in subtelomeres in this system, suggesting an environment that is substantially different from that experienced by genes in the chromosome core; and the results have also given some insight into the process of antigenic variation. Hopefully, in the future we can look forward to more details being revealed by other studies tracking the events of hyperevolution. The project has underlined that the importance of studying trypanosomes arises not only from their niche as a devastating human and veterinary pathogen, but also because their biology is inherently fascinating and can be effectively used as a model to investigate and illuminate strategies that are widely used by eukaryotes.

Appendices

Appendix 1: Population genetics and serology of EATRO 3, EATRO 2340 and related strains

EATRO 3 and EATRO 2340 were two of a series of samples from a sleeping sickness focus in SE Uganda (see section 3.1.1). EATRO 3 and EATRO 2340 were shown serologically to share a VSG set with one another and with many other samples from the focus; samples that were found to belong to this variable antigen type are shown in red in Table A-1 (Barry *et al*, 1983; J.D. Barry, pers. comm.). Isolates that are shown in black belong to other variable antigen types. Some of these stocks were analysed by microsatellite and minisatellite genotyping, and where available these data are also shown in Table A-1 (L. Morrison and A. MacLeod, unpublished).

Source	area	year	EATRO ID	Marker name												
				JS2		PLC		18		5L5		5		m12		
bushbuck	Sakwa	1958	204/222													
man		1959	174	99.73	109.1	156.7	165.9	160.2	160.2	121.7	121.7	152.5	173	107.1	107.1	genescan
				4	5	8	7	1	1	2	2	1	3	2	2	allele
fly	Lugala,Busoga	1960	3	99.73	109	156.6	165.8	159.4	159.4	121.6	121.6	152.5	172.9	107.2	107.2	genescan
				4	5	8	7	1	1	2	2	1	3	2	2	allele
fly	C.Nyanza	1961	7													
	C.Nyanza	1961	18													
man	Alego	1961	94	99.72	108.1	156.7	165.8	160.2	160.2	121.7	121.7	152.6	172.9	107.1	107.1	genescan
				4	5	8	7	1	1	2	2	1	3	2	2	allele
man	Alego	1961	95													
man	Alego	1961	96													
man	Alego	1961	97													
man	Alego	1961	98													
man	Sakwa	1961	103													
man	Sakwa	1961	148													
man	Uyoma	1961	149													
man	Yimbo	1961	156													
cow	Uhembo,Alego	1964	795	99.8	109.1	156.8	165.9	160.3	160.3	121.7	121.7	152.5	172.9	107.2	107.2	genescan
				4	5	8	7	1	1	2	2	1	3	2	2	allele
cow	Uhembo,Alego	1964	811													
cow	Uhembo,Alego	1964	812													
man	Sakwa	1964	846													
cow	Yimbo	1966	1051													
man	Sakwa	1966	1095													
cow	Alego	1967	1155													
cow	Malengo	1968	1216													
cow	Malengo	1968	1217													
man		1976	2274													
man	Samia	1977	2340	100.6	109	143.6	155.7	160.2	160.2	121.7	121.7	152.5	179.1	107.2	107.2	genescan
				4	5	1	8	1	1	2	2	1	3	2	2	allele
man	Samia	1977	2344													
man	Samia	1977	2350													
dog	Luuka,Busoga	1979	2503													
man	Luuka,Busoga	1979	2509													
		1981	UTRO3	99.72	109	156.6	165.8	160.3	160.3	121.7	121.7	152.5	172.8	107.2	107.2	genescan
				4	5	8	7	1	1	2	2	1	3	2	2	allele

Table A-1 – Serology and genetic markers in samples from the SE Uganda sleeping sickness focus.
See text for details.

Appendix 2: Details of electronic appendices

Appendix E1: Scripts and modules developed for or used in the project.

The purpose and usage of the scripts are given in Chapter 3. Unless otherwise indicated, the scripts were written as part of the project.

Appendix E2: Genome assemblies and *VSG* NTDs of EATRO 3 and EATRO 2340.

Each genome is given as a single GenBank file with multiple entries. *VSG* NTDs are annotated within the GenBank files.

Appendix E3: *VSG* NTD matches.

The appendix comprises a table showing *VSG* NTDs from EATRO 3 and EATRO 2340 genomes for which the corresponding gene was considered to have been found in the other genome. The *VSG* NTD used as the query is given in the first column, followed by details of the hit quality, and the corresponding gene in the other genome, if applicable. Genes not in this table were not considered in the SNP analysis.

Appendix E4: Changes in *VSG* NTDs.

The appendix comprises a table showing changes annotated between the EATRO 3 and EATRO 2340 *VSG* archives. Changes were annotated using Method 2 (see section 4.4.2). The first column describes the gene as named in Appendix E3. GT = genotype. As discussed in section 4.4.2, most SNPs were found to be heterozygous, so two alleles are given for the gene in each genome.

List of references

van den Abbeele J, Claes Y, van Bockstaele D, Le Ray D & Coosemans M (1999) *Trypanosoma brucei* spp. development in the tsetse fly: characterization of the post-mesocyclic stages in the foregut and proboscis. *Parasitology* **118:** 469–478

Aitcheson N, Talbot S, Shapiro J, Hughes K, Adkin C, Butt T, Sheader K & Rudenko G (2005) VSG switching in *Trypanosoma brucei*: antigenic variation analysed using RNAi in the absence of immune selection. *Mol Microbiol* **57:** 1608–1622

Alarcon CM, Son HJ, Hall T & Donelson JE (1994) A monocistronic transcript for a trypanosome variant surface glycoprotein. *Mol Cell Biol* **14:** 5579–5591

Alsford S & Horn D (2012) Cell-cycle-regulated control of VSG expression site silencing by histones and histone chaperones ASF1A and CAF-1b in *Trypanosoma brucei. Nucleic Acids Res* **40:** 10150–10160

Alsford S, Horn D & Glover L (2009) DNA breaks as triggers for antigenic variation in African trypanosomes. *Genome Biol* **10:** 223

Alsford S, Kawahara T, Isamah C & Horn D (2007) A sirtuin in the African trypanosome is involved in both DNA repair and telomeric gene silencing but is not required for antigenic variation. *Mol Microbiol* **63:** 724–736

Alsford S, Wickstead B, Ersfeld K & Gull K (2001) Diversity and dynamics of the minichromosomal karyotype in *Trypanosoma bruce*i. *Mol Biochem Parasitol* **113:** 79–88

Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215:** 403–410

Alvarez F, Cortinas MN & Musto H (1996) The analysis of protein coding genes suggests monophyly of *Trypanosoma. Mol. Phylogenet. Evol.* **5:** 333–343

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana. Nature* **408:** 796–815

Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, Carrington M, Depledge DP, Fischer S, Gajria B, Gao X, Gardner MJ, Gingle A, Grant G, Harb OS, Heiges M, Hertz-Fowler C, Houston R, Innamorato F, Iodice J, Kissinger JC, *et al* (2010) TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Res* **38:** D457–62

Assefa S, Keane TM, Otto TD, Newbold C & Berriman M (2009) ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* **25:** 1968–1969

Baird DM, Coleman J, Rosser ZH & Royle NJ (2000) High levels of sequence polymorphism and linkage disequilibrium at the telomere of 12q: implications for telomere biology and human evolution. *Am J Hum Genet* **66:** 235–250

Barbet AF & Kamper SM (1993) The importance of mosaic genes to trypanosome survival. *Parasitol Today (Regul Ed)* **9:** 63–66

Barbour AG & Restrepo BI (2000) Antigenic variation in vector-borne pathogens. *Emerging Infect Dis* **6:** 449–457

Barnes RL & McCulloch R (2007) *Trypanosoma brucei* homologous recombination is dependent on substrate length and homology, though displays a differential dependence on mismatch repair as substrate length decreases. *Nucleic Acids Res* **35:** 3478–3493

Barrett MP, Burchmore RJS, Stich A, Lazzari JO, Frasch AC, Cazzulo JJ & Krishna S (2003) The trypanosomiases. *Lancet* **362:** 1469–1480

Barry JD (1997) The relative significance of mechanisms of antigenic variation in African trypanosomes. *Parasitol Today (Regul Ed)* **13:** 212–218

Barry JD & McCulloch R (2001) Antigenic variation in trypanosomes: enhanced phenotypic variation in a eukaryotic parasite. *Adv Parasitol* **49:** 1–70

Barry JD, Crowe JS & Vickerman K (1983) Instability of the *Trypanosoma brucei* rhodesiense metacyclic variable antigen repertoire. *Nature* **306:** 699–701

Barry JD, Ginger ML, Burton P & McCulloch R (2003) Why are parasite contingency genes often associated with telomeres? *Int J Parasitol* **33:** 29–45

Barry JD, Graham SV, Fotheringham M, Graham VS, Kobryn K & Wymer B (1998) VSG gene control and infectivity strategy of metacyclic stage *Trypanosoma brucei*. *Mol Biochem Parasitol* **91:** 93–105

Barry JD, Hall JPJ & Plenderleith L (2012) Genome hyperevolution and the success of a parasite. *Ann N Y Acad Sci* **1267:** 11–17

Barry JD, Marcello L, Morrison LJ, Read AF, Lythgoe K, Jones N, Carrington M, Blandin G, Böhme U, Caler E, Hertz-Fowler C, Renauld H, El-Sayed N & Berriman M (2005) What the genome sequence is revealing about trypanosome antigenic variation. *Biochem Soc Trans* **33:** 986–989

Bebenek K & Kunkel TA (1995) Analyzing fidelity of DNA polymerases. *Meth Enzymol* **262:** 217–232

Becker M, Aitcheson N, Byles E, Wickstead B, Louis E & Rudenko G (2004) Isolation of the repertoire of VSG expression site containing telomeres of *Trypanosoma brucei* 427 using transformation-associated recombination in yeast. *Genome Res* **14:** 2319–2329

Bell JS, Harvey TI, Sims A-M & McCulloch R (2004) Characterization of components of the mismatch repair machinery in *Trypanosoma brucei*. *Mol Microbiol* **51:** 159–173

Benmerzouga I, Concepción-Acevedo J, Kim H-S, Vandoros AV, Cross GAM, Klingbeil MM & Li B (2012) *Trypanosoma brucei* Orc1 is essential for nuclear DNA replication and affects both VSG silencing and VSG switching. *Mol Microbiol* **87:** 196–210

Bentley S, Holden M, Sebaihia M, Cerdeno-Tarraga A & Parkhill J (2002a) Genome giants. *Trends Microbiol* **10:** 309–310

Bentley SD, Chater KF, Cerdeño-Tárraga A-M, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H, Harper D, Bateman A, Brown S, Chandra G, Chen CW, Collins M, Cronin A, Fraser A, Goble A, Hidalgo J, Hornsby T, *et al* (2002b) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417:** 141–147

Benz C, Nilsson D, Andersson B, Clayton C & Guilbride DL (2005) Messenger RNA processing sites in *Trypanosoma brucei*. *Mol Biochem Parasitol* **143:** 125–134

Bernards A, van der Ploeg LH, Gibson WC, Leegwater P, Eijgenraam F, de Lange T, Weijers P, Calafat J & Borst P (1986) Rapid change of the repertoire of *variant surface glycoprotein* genes in trypanosomes by gene duplication and deletion. *J Mol Biol* **190:** 1–10

Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, Lennard NJ, Caler E, Hamlin NE, Haas B, Böhme U, Hannick L, Aslett MA, Shallom J, Marcello L, Hou L, Wickstead B, Alsmark UCM, Arrowsmith C, Atkin RJ, *et al* (2005) The genome of the African trypanosome *Trypanosoma brucei*. *Science* **309:** 416–422

Berriman M, Hall N, Sheader K, Bringaud F, Tiwari B, Isobe T, Bowman S, Corton C, Clark L, Cross GAM, Hoek M, Zanders T, Berberof M, Borst P & Rudenko G (2002) The architecture of variant surface glycoprotein gene expression sites in *Trypanosoma brucei*. *Mol Biochem Parasitol* **122:** 131–140

Bjedov I, Dasgupta CN, Slade D, Le Blastier S, Selva M & Matic I (2007) Involvement of *Escherichia coli* DNA Polymerase IV in tolerance of cytotoxic alkylating DNA lesions *in vivo*. *Genetics* **176:** 1431–1440

Black SJ, Seed JR & Murphy NB (2001) Innate and acquired resistance to African trypanosomiasis. *J Parasitol* **87:** 1–9

Blackburn EH (2001) Switching and signaling at the telomere. *Cell* **106:** 661–673

Blackburn EH (2005) Telomeres and telomerase: their mechanisms of action and the effects of altering their functions. *FEBS Letters* **579:** 859–862

Blum ML, Down JA, Gurnett AM, Carrington M, Turner MJ & Wiley DC (1993) A structural motif in the variant surface glycoproteins of *Trypanosoma brucei*. *Nature* **362:** 603–609

Boothroyd CE, Dreesen O, Leonova T, Ly KI, Figueiredo LM, Cross GAM & Papavasiliou FN (2009) A yeast-endonuclease-generated DNA break induces antigenic switching in *Trypanosoma brucei*. *Nature* **459:** 278–281

Borst P (2002) Antigenic variation and allelic exclusion. *Cell* **109:** 5-8

Borst P & Genest P-A (2006) Parasitology: switching like for like. *Nature* **439:** 926–927

Brayton KA, Palmer GH, Lundgren A, Yi J & Barbet AF (2002) Antigenic variation of *Anaplasma marginale* msp2 occurs by combinatorial gene conversion. *Mol Microbiol* **43:** 1151-1159

Bringaud F, Biteau N, Melville SE, Hez S, El-Sayed NM, Leech V, Berriman M, Hall N, Donelson JE & Baltz T (2002) A new, expressed multigene family containing a hot spot for insertion of retroelements is associated with polymorphic subtelomeric regions of *Trypanosoma brucei*. *Eukaryot Cell* **1:** 137–151

Brown CA, Murray AW & Verstrepen KJ (2010) Rapid expansion and functional divergence of subtelomeric gene families in yeasts. *Curr Biol* **20:** 895–903

Brown WR, MacKinnon PJ, Villasanté A, Spurr N, Buckle VJ & Dobson MJ (1990) Structure and polymorphism of human telomere-associated DNA. *Cell* **63:** 119-132

Brun R, Hecker H & Lun ZR (1998) *Trypanosoma evansi* and *T. equiperdum*: distribution, biology, treatment and phylogenetic relationship (a review). *Vet Parasitol* **79:** 95–107

Callejas S, Leech V, Reitter C & Melville S (2006) Hemizygous subtelomeres of an African trypanosome chromosome may account for over 75% of chromosome length. *Genome Res* **16:** 1109–1118

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K & Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinformatics* **10:** 421

Campbell DA, van Bree MP & Boothroyd JC (1984) The 5'-limit of transposition and upstream barren region of a trypanosome *VSG* gene: tandem 76 base-pair repeats flanking $(TAA)_{90}$. *Nucleic Acids Res* **12:** 2759–2774

Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, Stebbings LA, Leroy C, Edkins S, Hardy C, Teague JW, Menzies A, Goodhead I, Turner DJ, Clee CM, Quail MA, Cox A, Brown C, Durbin R, Hurles ME, *et al* (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* **40:** 722-729

Capewell P, Veitch NJ, Turner CMR, Raper J, Berriman M, Hajduk SL & MacLeod A (2011) Differences between *Trypanosoma brucei gambiense* groups 1 and 2 in their resistance to killing by trypanolytic factor 1. *PLoS Negl Trop Dis* **5:** e1287

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

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

Carver T, Böhme U, Otto TD, Parkhill J & Berriman M (2010) BamView: viewing mapped read alignment data in the context of the reference sequence. *Bioinformatics* **26:** 676–677

Carver TJ, Rutherford KM, Berriman M, Rajandream M-A, Barrell BG & Parkhill J (2005) ACT: the Artemis Comparison Tool. *Bioinformatics* **21:** 3422-3423

Cattand P, Jannin J & Lucas P (2001) Sleeping sickness surveillance: an essential step towards elimination. *Trop Med Int Health* **6:** 348–361

Chan CS & Tye BK (1983) Organization of DNA sequences and replication origins at yeast telomeres. *Cell* **33:** 563–573

Chaves I, Zomerdijk J, Dirks-Mulder A, Dirks RW, Raap AK & Borst P (1998) Subnuclear localization of the active *variant surface glycoprotein* gene expression site in *Trypanosoma brucei*. *Proc Natl Acad Sci USA* **95:** 12328-12333

Chen C & Kolodner RD (1999) Gross chromosomal rearrangements in *Saccharomyces cerevisiae* replication and recombination defective mutants. *Nat Genet* **23:** 81–85

Cheng Q, Cloonan N, Fischer K, Thompson J, Waine G, Lanzer M & Saul A (1998) *stevor* and *rif* are *Plasmodium falciparum* multicopy gene families which potentially encode variant antigens. *Mol Biochem Parasitol* **97:** 161–176

Cliffe LJ, Siegel TN, Marshall M, Cross GAM & Sabatini R (2010) Two thymidine hydroxylases differentially regulate the formation of glucosylated DNA at regions flanking polymerase II polycistronic transcription units throughout the genome of *Trypanosoma brucei*. *Nucleic Acids Res* **38:** 3923-3935

Cole C, Barber JD & Barton GJ (2008) The Jpred 3 secondary structure prediction server. *Nucleic Acids Res* **36:** W197–201

Conway C, McCulloch R, Ginger ML, Robinson NP, Browitt A & Barry JD (2002a) Ku is important for telomere maintenance, but not for differential expression of telomeric *VSG* genes, in African trypanosomes. *J Biol Chem* **277:** 21269-21277

Conway C, Proudfoot C, Burton P, Barry JD & McCulloch R (2002b) Two pathways of homologous recombination in *Trypanosoma brucei*. *Mol Microbiol* **45:** 1687-1700

Cooper A (2009) PhD thesis, University of Glasgow

Cooper GM, Nickerson DA & Eichler EE (2007) Mutational and selective effects on copy-number variants in the human genome. *Nat Genet* **39:** S22–9

Corcoran LM, Thompson JK, Walliker D & Kemp DJ (1988) Homologous recombination within subtelomeric repeat sequences generates chromosome size polymorphisms in *P. falciparum*. *Cell* **53:** 807–813

Cross M, Kieft R, Sabatini R, Wilm M, de Kort M, van der Marel GA, van Boom JH, van Leeuwen F & Borst P (1999) The modified base J is the target for a novel DNA-binding protein in kinetoplastid protozoans. *EMBO J* **18:** 6573–6581

Cully DF, Ip HS & Cross GA (1985) Coordinate transcription of *variant surface glycoprotein* genes and an expression site associated gene family in *Trypanosoma brucei*. *Cell* **42:** 173-182

D'Amours D & Jackson SP (2002) The Mre11 complex: at the crossroads of dna repair and checkpoint signalling. *Nat Rev Mol Cell Biol* **3:** 317–327

Davies AA, Masson JY, McIlwraith MJ, Stasiak AZ, Stasiak A, Venkitaraman AR & West SC (2001) Role of BRCA2 in control of the RAD51 recombination and DNA repair protein. *Molecular Cell* **7:** 273–282

de Bruin D, Lanzer M & Ravetch JV (1994) The polymorphic subtelomeric regions of *Plasmodium falciparum* chromosomes contain arrays of repetitive sequence elements. *Proc Natl Acad Sci USA* **91:** 619–623

De Greef C & Hamers R (1994) The serum resistance-associated (SRA) gene of Trypanosoma brucei rhodesiense encodes a variant surface glycoprotein-like protein. *Mol Biochem Parasitol* **68:** 277–284

de Lange T, Shiue L, Myers RM, Cox DR, Naylor SL, Killery AM & Varmus HE (1990) Structure and variability of human chromosome ends. *Mol Cell Biol* **10:** 518–527

Deitsch KW, Moxon ER & Wellems TE (1997) Shared themes of antigenic variation and virulence in bacterial, protozoal, and fungal infections. *Microbiol Mol Biol Rev* **61:** 281–293

Déjardin J & Kingston RE (2009) Purification of proteins associated with specific genomic Loci. *Cell* **136:** 175–186

Dobson R, Stockdale C, Lapsley C, Wilkes J & McCulloch R (2011) Interactions among *Trypanosoma brucei* RAD51 paralogues in DNA repair and antigenic variation. *Mol Microbiol* **81:** 434–456

Donelson JE (2003) Antigenic variation and the African trypanosome genome. *Acta Trop* **85:** 391–404

Dubois KN, Alsford S, Holden JM, Buisson J, Swiderski M, Bart J-M, Ratushny AV, Wan Y, Bastin P, Barry JD, Navarro M, Horn D, Aitchison JD, Rout MP & Field MC (2012) NUP-1 Is a large coiled-coil nucleoskeletal protein in trypanosomes with lamin-like functions. *PLoS Biol* **10:** e1001287

Duffy MF & Tham W-H (2007) Transcription and coregulation of multigene families in *Plasmodium falciparum*. *Trends Parasitol* **23:** 183–6

Eddy SR (2011) Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7:** e1002195

Ferguson MA (1999) The structure, biosynthesis and functions of glycosylphosphatidylinositol anchors, and the contributions of trypanosome research. *J Cell Sci* **112:** 2799–2809

Fernandez-Becerra C, Yamamoto MM, Vêncio RZN, Lacerda M, Rosanas-Urgell A & del Portillo HA (2009) Plasmodium vivax and the importance of the subtelomeric multigene vir superfamily. *Trends Parasitol* **25:** 44–51

Ferrante A & Allison AC (1983) Alternative pathway activation of complement by African trypanosomes lacking a glycoprotein coat. *Parasite Immunol* **5:** 491–498

Field MC & Boothroyd JC (1996) Sequence divergence in a family of variant surface glycoprotein genes from trypanosomes: coding region hypervariability and downstream recombinogenic repeats. *J Mol Evol* **42:** 500–511

Figueiredo LM, Janzen CJ & Cross GAM (2008) A histone methyltransferase modulates antigenic variation in African trypanosomes. *PLoS Biol* **6:** e161

Figueiredo LM, Pirrit LA, Scherf A & Pirritt LA (2000) Genomic organisation and chromatin structure of Plasmodium falciparum chromosome ends. *Mol Biochem Parasitol* **106:** 169–174

Flint J, Bates GP, Clark K, Dorman A, Willingham D, Roe BA, Micklem G, Higgs DR & Louis EJ (1997) Sequence comparison of human and yeast telomeres identifies structurally distinct subtelomeric domains. *Hum Mol Genet* **6:** 1305–1313

Francino MP & Ochman H (2001) Deamination as the basis of strand-asymmetric evolution in transcribed *Escherichia coli* sequences. *Mol Biol Evol* **18:** 1147-1150

Frank AC & Lobry JR (1999) Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene* **238:** 65–77

Freitas-Junior LH, Bottius E, Pirrit LA, Deitsch KW, Scheidig C, Guinet F, Nehrbass U, Wellems TE & Scherf A (2000) Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of *P. falciparum*. *Nature* **407:** 1018–1022

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

Fu G & Melville SE (2002) Polymorphism in the subtelomeric regions of chromosomes of Kinetoplastida. *Trans R Soc Trop Med Hyg* **96 Suppl 1:** S31–40

Futse JE, Brayton KA, Dark MJ, Knowles DP & Palmer GH (2008) Superinfection as a driver of genomic diversification in antigenically variant pathogens. *Proc Natl Acad Sci USA* **105:** 2123–2127

Gardner MJ, Tettelin H, Carucci DJ, Cummings LM, Aravind L, Koonin EV, Shallom S, Mason T, Yu K, Fujii C, Pederson J, Shen K, Jing J, Aston C, Lai Z, Schwartz DC, Pertea M, Salzberg S, Zhou L, Sutton GG, *et al* (1998) Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282:** 1126–1132

Gascuel O (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol* **14:** 685–695

Genois M-M, Mukherjee A, Ubeda J-M, Buisson R, Paquet E, Roy G, Plourde M, Coulombe Y, Ouellette M & Masson J-Y (2012) Interactions between BRCA2 and RAD51 for promoting homologous recombination in *Leishmania infantum*. *Nucleic Acids Res* **40:** 6570–6584 van Gent DC, Hoeijmakers JH & Kanaar R (2001) Chromosomal stability and the DNA double-stranded break connection. *Nat Rev Genet* **2:** 196–206

Gerlach VL, Aravind L, Gotway G, Schultz RA, Koonin EV & Friedberg EC (1999) Human and mouse homologs of *Escherichia coli* DinB (DNA polymerase IV), members of the UmuC/DinB superfamily. *Proc Natl Acad Sci USA* **96:** 11922–11927

Gerlach VL, Feaver WJ, Fischhaber PL & Friedberg EC (2001) Purification and characterization of pol kappa, a DNA polymerase encoded by the human *DINB1* gene. *J Biol Chem* **276:** 92–98

Gibson W (2007) Resolution of the species problem in African trypanosomes. *Int J Parasitol* **37:** 829–838

Gibson W, Backhouse T & Griffiths A (2002) The human serum resistance associated gene is ubiquitous and conserved in Trypanosoma brucei rhodesiense throughout East Africa. *Infect Genet Evol* **1:** 207–214

Ginger ML, Blundell PA, Lewis AM, Browitt A, Günzl A & Barry JD (2002) Ex vivo and in vitro identification of a consensus promoter for VSG genes expressed by metacyclic-stage trypanosomes in the tsetse fly. *Eukaryot Cell* **1:** 1000–1009

Gjini E, Haydon DT, Barry JD & Cobbold CA (2010) Critical interplay between parasite differentiation, host immunity, and antigenic variation in trypanosome infections. *Am Nat* **176:** 424–439

Gjini E, Haydon DT, Barry JD & Cobbold CA (2012) The impact of mutation and gene conversion on the local diversification of antigen genes in African trypanosomes. *Mol Biol Evol* **29:** 3321–3331

Gjini E, Haydon DT, Barry JD & Cobbold CA (2013) Linking the antigen archive structure to pathogen fitness in African trypanosomes. *Proc R Soc B: Biol Sci* **280:** 20122129

Glover L & Horn D (2006) Repression of polymerase I-mediated gene expression at *Trypanosoma brucei* telomeres. *EMBO Rep* **7:** 93–99

Glover L, Alsford S, Beattie C & Horn D (2007) Deletion of a trypanosome telomere leads to loss of silencing and progressive loss of terminal DNA in the absence of cell cycle arrest. *Nucleic Acids Res* **35:** 872–880

Glover L, McCulloch R & Horn D (2008) Sequence homology and microhomology dominate chromosomal double-strand break repair in African trypanosomes. *Nucleic Acids Res* **36:** 2608–2618

Glusman G, Yanai I, Rubin I & Lancet D (2001) The complete human olfactory subgenome. *Genome Res* **11:** 685–702

Gommers-Ampt JH, van Leeuwen F, de Beer AL, Vliegenthart JF, Dizdaroglu M, Kowalak JA, Crain PF & Borst P (1993) beta-D-glucosyl-hydroxymethyluracil: a novel modified base present in the DNA of the parasitic protozoan *T. brucei*. *Cell* **75:** 1129–1136

Goodman MF (2002) Error-prone repair DNA polymerases in prokaryotes and eukaryotes. *Annu Rev Biochem* **71:** 17–50

Gotta M, Laroche T, Formenton A, Maillet L, Scherthan H & Gasser SM (1996) The clustering of telomeres and colocalization with Rap1, Sir3, and Sir4 proteins in wild-type *Saccharomyces cerevisiae*. *J Cell Biol* **134:** 1349–1363

Gottschling DE, Aparicio OM, Billington BL & Zakian VA (1990) Position effect at *S. cerevisiae* telomeres: reversible repression of Pol II transcription. *Cell* **63:** 751–762

Graham SV & Barry JD (1995) Transcriptional regulation of metacyclic *variant surface glycoprotein* gene expression during the life cycle of *Trypanosoma brucei*. *Mol Cell Biol* **15:** 5945–5956

Graham SV, Terry S & Barry JD (1999) A structural and transcription pattern for *variant surface glycoprotein* gene expression sites used in metacyclic stage *Trypanosoma brucei*. *Mol Biochem Parasitol* **103:** 141–154

Graham VS & Barry JD (1996) Is point mutagenesis a mechanism for antigenic variation in *Trypanosoma brucei*? *Mol Biochem Parasitol* **79:** 35–45

Günzl A, Bruderer T, Laufer G, Schimanski B, Tu L-C, Chung H-M, Lee P-T & Lee MG-S (2003) RNA polymerase I transcribes procyclin genes and variant surface glycoprotein gene expression sites in *Trypanosoma brucei*. *Eukaryot Cell* **2:** 542-551

Hajduk SL, Hager K & Esko JD (1992) High-density lipoprotein-mediated lysis of trypanosomes. *Parasitology Today* **8:** 95–98

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

Hamilton PB, Stevens JR, Gaunt MW, Gidley J & Gibson WC (2004) Trypanosomes are monophyletic: evidence from genes for glyceraldehyde phosphate dehydrogenase and small subunit ribosomal RNA. *Int J Parasitol* **34:** 1393–1404

Hartley CL & McCulloch R (2008) *Trypanosoma brucei* BRCA2 acts in antigenic variation and has undergone a recent expansion in BRC repeat number that is important during homologous recombination. *Mol Microbiol* **68:** 1237–1251

Herbert WJ, Parratt D, van Meirvenne N & Lennox B (1980) An accidental laboratory infection with trypanosomes of a defined stock. II. Studies on the serological response of the patient and the identity of the infecting organism. *J. Infect.* **2:** 113–124

Hernandez-Rivas R, Mattei D, Sterkers Y, Peterson DS, Wellems TE & Scherf A (1997) Expressed *var* genes are found in *Plasmodium falciparum* subtelomeric regions. *Mol Cell Biol* **17:** 604–611

Hershberg R & Petrov DA (2010) Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet* **6:**

Hertz-Fowler C, Figueiredo LM, Quail MA, Becker M, Jackson A, Bason N, Brooks K, Churcher C, Fahkro S, Goodhead I, Heath P, Kartvelishvili M, Mungall K, Harris D, Hauser H, Sanders M, Saunders D, Seeger K, Sharp S, Taylor JE, *et al* (2008) Telomeric expression sites are highly conserved in *Trypanosoma brucei*. *PLoS ONE* **3:** e3527

Hide G (1999) History of sleeping sickness in East Africa. *Clin Microbiol Rev* **12:** 112–125

Hide G, Buchanan N, Welburn S, Maudlin I, Barry JD & Tait A (1991) *Trypanosoma brucei rhodesiense*: characterisation of stocks from Zambia, Kenya, and Uganda using repetitive DNA probes. *Exp Parasitol* **72:** 430–439

Hile SE & Eckert KA (2008) DNA polymerase kappa produces interrupted mutations and displays polar pausing within mononucleotide microsatellite sequences. *Nucleic Acids Res* **36:** 688–696

Hirumi H & Hirumi K (1989) Continuous cultivation of *Trypanosoma brucei* blood stream forms in a medium containing a low concentration of serum protein without feeder cell layers. *J Parasitol* **75:** 985–989

Hodgkinson A & Eyre-Walker A (2011) Variation in the mutation rate across mammalian genomes. *Nat Rev Genet* **12:** 756–766

Horn D (2004) The molecular control of antigenic variation in *Trypanosoma brucei*. *Curr. Mol. Med.* **4:** 563–576

Horn D (2008) Codon usage suggests that translational selection has a major impact on protein expression in trypanosomatids. *BMC Genomics* **9:** 2

Horn D & Barry JD (2005) The central roles of telomeres and subtelomeres in antigenic variation in African trypanosomes. *Chromosome Res* **13:** 525–533

Horn D & Cross GA (1997) Analysis of *Trypanosoma brucei* VSG expression site switching in vitro. *Mol Biochem Parasitol* **84:** 189–201

Hughes K, Wand M, Foulston L, Young R, Harley K, Terry S, Ersfeld K & Rudenko G (2007) A novel ISWI is involved in VSG expression site downregulation in African trypanosomes. *EMBO J* **26:** 2400–2410

Huson DH & Scornavacca C (2012) Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. *Syst Biol* **6:** 1061-1067

Hutchinson OC, Picozzi K, Jones NG, Mott H, Sharma R, Welburn SC & Carrington M (2007) *Variant surface glycoprotein* gene repertoires in *Trypanosoma brucei* have diverged to become strain-specific. *BMC Genomics* **8:** 234

Jackson AP (2007a) Evolutionary consequences of a large duplication event in *Trypanosoma brucei*: chromosomes 4 and 8 are partial duplicons. *BMC Genomics* **8:** 432

Jackson AP (2007b) Tandem gene arrays in *Trypanosoma brucei*: Comparative phylogenomic analysis of duplicate sequence variation. *BMC Evol Biol* **7:** 54

Jackson AP, Sanders M, Berry A, McQuillan J, Aslett MA, Quail MA, Chukualim B, Capewell P, MacLeod A, Melville SE, Gibson W, Barry JD, Berriman M & Hertz-Fowler C (2010) The genome sequence of *Trypanosoma brucei gambiense*, causative agent of chronic human African trypanosomiasis. *PLoS Negl Trop Dis* **4:** e658

Jamonneau V, Ilboudo H, Kaboré J, Kaba D, Koffi M, Solano P, Garcia A, Courtin D, Laveissière C, Lingue K, Büscher P & Bucheton B (2012) Untreated human Infections by *Trypanosoma brucei gambiense* are not 100% fatal. *PLoS Negl Trop Dis* **6:** e1691

Johnson JG & Cross GA (1979) Selective cleavage of variant surface glycoproteins from *Trypanosoma brucei*. *Biochem J* **178:** 689–697

Johnson RE, Prakash S & Prakash L (2000) The human *DINB1* gene encodes the DNA polymerase Pol theta. *Proc Natl Acad Sci USA* **97:** 3838–3843

Kamper SM & Barbet AF (1992) Surface epitope variation via mosaic gene formation is potential key to long-term survival of *Trypanosoma brucei*. *Mol Biochem Parasitol* **53:** 33–44

Kennedy PGE (2008) The continuing problem of human African trypanosomiasis (sleeping sickness). *Ann Neurol* **64:** 116–126

Kieft R, Capewell P, Turner CMR, Veitch NJ, MacLeod A & Hajduk S (2010) Mechanism of *Trypanosoma brucei gambiense* (group 1) resistance to human trypanosome lytic factor. *Proc Natl Acad Sci USA* **107:** 16137–16141

Kim SR, Maenhaut-Michel G, Yamada M, Yamamoto Y, Matsui K, Sofuni T, Nohmi T & Ohmori H (1997) Multiple pathways for SOS-induced mutagenesis in *Escherichia coli*: an overexpression of dinB/dinP results in strongly enhancing mutagenesis in the absence of any exogenous treatment to damage DNA. *Proc Natl Acad Sci USA* **94:** 13792–13797

King DG & Kashi Y (2007) Indirect selection for mutability. *Heredity (Edinb)* **99:** 123–124

Kobryn K & Chaconas G (2001) The circle is broken: telomere resolution in linear replicons. *Curr Opin Microbiol* **4:** 558–564

Koffi M, De Meeûs T, Bucheton B, Solano P, Camara M, Kaba D, Cuny G, Ayala FJ & Jamonneau V (2009) Population genetics of *Trypanosoma brucei gambiense*, the agent of sleeping sickness in western Africa. *Proc Natl Acad Sci USA* **106:** 209–214

Kolev NG, Franklin JB, Carmi S, Shi H, Michaeli S & Tschudi C (2010) The transcriptome of the human pathogen *Trypanosoma brucei* at single-nucleotide resolution. *PLoS Pathog* **6:** e1001090

Kondratick CM, Washington MT, Prakash S & Prakash L (2001) Acidic residues critical for the activity and biological function of yeast DNA polymerase eta. *Mol Cell Biol* **21:** 2018–2025

Kotani H, Hosouchi T & Tsuruoka H (1999) Structural analysis and complete physical map of *Arabidopsis thaliana* chromosome 5 including centromeric and telomeric regions. *DNA Res* **6:** 381–386

Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M & Turner DJ (2009) Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods* **6:** 291–295

Kraemer SM & Smith JD (2006) A family affair: *var* genes, PfEMP1 binding, and malaria disease. *Curr Opin Microbiol* **9:** 374–380

Kryazhimskiy S & Plotkin JB (2008) The population genetics of dN/dS. *PLoS Genet* **4:** e1000304

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C & Salzberg SL (2004) Versatile and open software for comparing large genomes. *Genome Biol* **5:** R12

Kyes SA, Rowe JA, Kriek N & Newbold CI (1999) Rifins: a second family of clonally variant proteins expressed on the surface of red cells infected with *Plasmodium falciparum*. *Proc Natl Acad Sci USA* **96:** 9333–9338

Lai D-H, Hashimi H, Lun Z-R, Ayala FJ & 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 USA* **105:** 1999–2004

Landeira D, Bart J-M, van Tyne D & Navarro M (2009) Cohesin regulates VSG monoallelic expression in trypanosomes. *J Cell Biol* **186:** 243–254

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ & Higgins DG (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23:** 2947–2948

Lee MG & van der Ploeg LH (1997) Transcription of protein-coding genes in trypanosomes by RNA polymerase I. *Annu. Rev. Microbiol.* **51:** 463–489

Legros D, Evans S, Maiso F, Enyaru JC & Mbulamberi D (1999) Risk factors for treatment failure after melarsoprol for *Trypanosoma brucei gambiense* trypanosomiasis in Uganda. *Trans R Soc Trop Med Hyg* **93:** 439–442

Lenardo MJ, Esser KM, Moon AM, van der Ploeg LH & Donelson JE (1986) Metacyclic *variant surface glycoprotein* genes of *Trypanosoma brucei* subsp. *rhodesiense* are activated in situ, and their expression is transcriptionally regulated. *Mol Cell Biol* **6:** 1991–1997

van Leeuwen F, Wijsman ER, Kieft R, van der Marel GA, van Boom JH & Borst P (1997) Localization of the modified base J in telomeric *VSG* gene expression sites of *Trypanosoma brucei*. *Genes Dev* **11:** 3232–3241

Li H & Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25:** 1754–1760

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R & 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079

Linardopoulou E, Mefford HC, Nguyen O, Friedman C, van den Engh G, Farwell DG, Coltrera M & Trask BJ (2001) Transcriptional activity of multiple copies of a subtelomerically located olfactory receptor gene that is polymorphic in number and location. *Hum Mol Genet* **10:** 2373–2383

Linardopoulou EV, Williams EM, Fan Y, Friedman C, Young JM & Trask BJ (2005) Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature* **437:** 94–100

Liu AY, Michels PA, Bernards A & Borst P (1985) Trypanosome *variant surface glycoprotein* genes expressed early in infection. *J Mol Biol* **182:** 383–396

Liu AY, van der Ploeg LH, Rijsewijk FA & Borst P (1983) The transposition unit of variant surface glycoprotein gene 118 of *Trypanosoma brucei*. Presence of repeated elements at its border and absence of promoter-associated sequences. *J Mol Biol* **167:** 57–75

Louis EJ (1995) The chromosome ends of *Saccharomyces cerevisiae*. *Yeast* **11:** 1553–1573

Louis EJ & Haber JE (1992) The structure and evolution of subtelomeric Y' repeats in *Saccharomyces cerevisiae*. *Genetics* **131:** 559–574

Louis EJ & Vershinin AV (2005) Chromosome ends: different sequences may provide conserved functions. *Bioessays* **27:** 685–697

Lumsden WH & Herbert WJ (1975) Pedigrees of the Edinburgh Trypanosoma (Trypanozoon) antigenic types (ETat). *Trans R Soc Trop Med Hyg* **69:** 205–208

Lynch M (2010a) Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci USA* **107:** 961–968

Lynch M (2010b) Evolution of the mutation rate. *Trends Genet* **26:** 345–352

Lynch M & Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* **290:** 1151–1155

Lythgoe KA, Morrison LJ, Read AF & Barry JD (2007) Parasite-intrinsic factors can explain ordered progression of trypanosome antigenic variation. *Proc Natl Acad Sci USA* **104:** 8095–8100

Macgregor P & Matthews KR (2010) New discoveries in the transmission biology of sleeping sickness parasites: applying the basics. *J. Mol. Med.* **88:** 865–871

Macgregor P, Savill NJ, Hall D & Matthews KR (2011) Transmission stages dominate trypanosome within-host dynamics during chronic infections. *Cell Host Microbe* **9:** 310–318

Machado CR, Augusto-Pinto L, McCulloch R & Teixeira SMR (2006) DNA metabolism and genetic diversity in Trypanosomes. *Mutat Res* **612:** 40–57

MacLeod A, Tweedie A, McLellan S, Hope M, Taylor S, Cooper A, Sweeney L, Turner CMR & Tait A (2005) Allelic segregation and independent assortment in *T. brucei* crosses: proof that the genetic system is Mendelian and involves meiosis. *Mol Biochem Parasitol* **143:** 12–19

Marcello L & Barry JD (2007a) From silent genes to noisy populations-dialogue between the genotype and phenotypes of antigenic variation. *J Eukaryot Microbiol* **54:** 14–17

Marcello L & Barry JD (2007b) Analysis of the *VSG* gene silent archive in *Trypanosoma brucei* reveals that mosaic gene expression is prominent in antigenic variation and is favored by archive substructure. *Genome Res* **17:** 1344–1352

Marcello L, Menon S, Ward P, Wilkes JM, Jones NG, Carrington M & Barry JD (2007) VSGdb: a database for trypanosome variant surface glycoproteins, a large and diverse family of coiled coil proteins. *BMC Bioinformatics* **8:** 143

Marcello L (2007) PhD thesis, University of Glasgow

Martincorena I & Luscombe NM (2013) Non-random mutation: the evolution of targeted hypermutation and hypomutation. *Bioessays* **35:** 123–130

Mason JM & Biessmann H (1995) The unusual telomeres of *Drosophila*. *Trends Genet* **11:** 58–62

Maudlin I (2006) African trypanosomiasis. *Ann Trop Med Parasitol* **100:** 679–701

McCulloch R & Barry JD (1999) A role for RAD51 and homologous recombination in *Trypanosoma brucei* antigenic variation. *Genes Devt* **13:** 2875–2888

McCulloch R & Horn D (2009) What has DNA sequencing revealed about the VSG expression sites of African trypanosomes? *Trends Parasitol* **25:** 359–363

McCulloch R, Rudenko G & Borst P (1997) Gene conversions mediating antigenic variation in Trypanosoma brucei can occur in variant surface glycoprotein expression sites lacking 70-base-pair repeat sequences. *Mol Cell Biol* **17:** 833–843

McCulloch SD & Kunkel TA (2008) The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell Res.* **18:** 148–161

McEachern MJ, Iyer S, Fulton TB & Blackburn EH (2000) Telomere fusions caused by mutating the terminal region of telomeric DNA. *Proc Natl Acad Sci USA* **97:** 11409–11414

McKenzie GJ & Rosenberg SM (2001) Adaptive mutations, mutator DNA polymerases and genetic change strategies of pathogens. *Curr Opin Microbiol* **4:** 586–594

McNeillage GJ, Herbert WJ & Lumsden WH (1969) Antigenic type of first relapse variants arising from a strain of *Trypanosoma* (Trypanozoon) *brucei*. *Exp Parasitol* **25:** 1–7

Mefford HC & Trask BJ (2002) The complex structure and dynamic evolution of human subtelomeres. *Nat Rev Genet* **3:** 91–102

Mefford HC, Linardopoulou E, Coil D, van den Engh G & Trask BJ (2001) Comparative sequencing of a multicopy subtelomeric region containing olfactory receptor genes reveals multiple interactions between non-homologous chromosomes. *Hum Mol Genet* **10:** 2363–2372

Melville SE, Gerrard CS & Blackwell JM (1999) Multiple causes of size variation in the diploid megabase chromosomes of African tyrpanosomes. *Chromosome Res* **7:** 191–203

Melville SE, Leech V, Gerrard CS, Tait A & Blackwell JM (1998) The molecular karyotype of the megabase chromosomes of *Trypanosoma brucei* and the assignment of chromosome markers. *Mol Biochem Parasitol* **94:** 155–173

Melville SE, Leech V, Navarro M & Cross GA (2000) The molecular karyotype of the megabase chromosomes of *Trypanosoma brucei* stock 427. *Mol Biochem Parasitol* **111:** 261–273

Michels PA, Liu AY, Bernards A, Sloof P, van der Bijl MM, Schinkel AH, Menke HH, Borst P, Veeneman GH, Tromp MC & van Boom JH (1983) Activation of the genes for variant surface glycoproteins 117 and 118 in *Trypanosoma brucei*. *J Mol Biol* **166:** 537–556

Miller EN & Turner MJ (1981) Analysis of antigenic types appearing in first relapse populations of clones of *Trypanosoma brucei*. *Parasitology* **82:** 63–80

Miller EN, Allan LM & Turner MJ (1984a) Topological analysis of antigenic determinants on a variant surface glycoprotein of *Trypanosoma brucei*. *Mol Biochem Parasitol* **13:** 67–81

Miller EN, Allan LM & Turner MJ (1984b) Mapping of antigenic determinants within peptides of a variant surface glycoprotein of *Trypanosoma brucei*. *Mol Biochem Parasitol* **13:** 309–322

Milner JD & Hajduk SL (1999) Expression and localization of serum resistance associated protein in *Trypanosoma brucei rhodesiense*. *Mol Biochem Parasitol* **104:** 271–283

Moraes Barros RR, Marini MM, Antônio CR, Cortez DR, Miyake AM, Lima FM, Ruiz JC, Bartholomeu DC, Chiurillo MA, Ramirez JL & da Silveira JF (2012) Anatomy and evolution of telomeric and subtelomeric regions in the human protozoan parasite *Trypanosoma cruzi*. *BMC Genomics* **13:** 229

Moran GP, Coleman DC & Sullivan DJ (2011) Comparative genomics and the evolution of pathogenicity in human pathogenic fungi. *Eukaryot Cell* **10:** 34–42

Morrison LJ, Majiwa P, Read AF & Barry JD (2005) Probabilistic order in antigenic variation of *Trypanosoma brucei*. *Int J Parasitol* **35:** 961–972

Morrison LJ, Marcello L & McCulloch R (2009) Antigenic variation in the African trypanosome: molecular mechanisms and phenotypic complexity. *Cell Microbiol* **11:** 1724–1734

Moxon ER, Rainey PB, Nowak MA & Lenski RE (1994) Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr Biol* **4:** 24–33

Mugal CF, Grünberg von H-H & Peifer M (2009) Transcription-induced mutational strand bias and its effect on substitution rates in human genes. *Mol Biol Evol* **26:** 131–142

Mulla AF & Rickman LR (1988) How do African game animals control trypanosome infections? *Parasitol Today (Regul Ed)* **4:** 352–354

Navarro M & Gull K (2001) A pol I transcriptional body associated with VSG mono-allelic expression in *Trypanosoma brucei*. *Nature* **414:** 759–763

Nimmo ER, Pidoux AL, Perry PE & Allshire RC (1998) Defective meiosis in telomere-silencing mutants of *Schizosaccharomyces pombe*. *Nature* **392:** 825–828

Niu DK, Lin K & Zhang D-Y (2003) Strand compositional asymmetries of nuclear DNA in eukaryotes. *J Mol Evol* **57:** 325–334

Njiokou F, Laveissière C, Simo G, Nkinin S, Grébaut P, Cuny G & Herder S (2006) Wild fauna as a probable animal reservoir for *Trypanosoma brucei gambiense* in Cameroon. *Infect Genet Evol* **6:** 147–153

Oberle M, Balmer O, Brun R & Roditi I (2010) Bottlenecks and the maintenance of minor genotypes during the life cycle of *Trypanosoma brucei*. *PLoS Pathog* **6:** e1001023

Odegard VH & Schatz DG (2006) Targeting of somatic hypermutation. *Nat Rev Immunol* **6:** 573–583

Ohashi E, Bebenek K, Matsuda T, Feaver WJ, Gerlach VL, Friedberg EC, Ohmori H & Kunkel TA (2000a) Fidelity and processivity of DNA synthesis by DNA polymerase kappa, the product of the human *DINB1* gene. *J Biol Chem* **275:** 39678–39684

Ohashi E, Ogi T, Kusumoto R, Iwai S, Masutani C, Hanaoka F & Ohmori H (2000b) Error-prone bypass of certain DNA lesions by the human DNA polymerase kappa. *Genes Dev* **14:** 1589–1594

Ohshima K, Kang S, Larson JE & Wells RD (1996) TTA.TAA triplet repeats in plasmids form a non-H bonded structure. *J Biol Chem* **271:** 16784–16791

Olender T, Lancet D & Nebert DW (2008) Update on the olfactory receptor (OR) gene superfamily. *Hum Genomics* **3:** 87–97

Oli MW, Cotlin LF, Shiflett AM & Hajduk SL (2006) Serum resistance-associated protein blocks lysosomal targeting of trypanosome lytic factor in *Trypanosoma brucei*. *Eukaryot Cell* **5:** 132–139

Osheroff WP, Jung HK, Beard WA, Wilson SH & Kunkel TA (1999) The fidelity of DNA polymerase beta during distributive and processive DNA synthesis. *J Biol Chem* **274:** 3642–3650

Otto TD, Dillon GP, Degrave WS & Berriman M (2011) RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Res* **39:** e57

Overath P, Chaudhri M, Steverding D & Ziegelbauer K (1994) Invariant surface proteins in bloodstream forms of *Trypanosoma brucei*. *Parasitol Today (Regul Ed)* **10:** 53–58

Oyola SO, Bringaud F & Melville SE (2009) A kinetoplastid BRCA2 interacts with DNA replication protein CDC45. *Int J Parasitol* **39:** 59–69

Pâques F & Haber JE (1999) Multiple pathways of recombination induced by double-strand breaks in *Saccharomyces cerevisiae*. *Microbiol Mol Biol Rev* **63:** 349–404

Pays E, Guyaux M, Aerts D, van Meirvenne N & Steinert M (1985) Telomeric reciprocal recombination as a possible mechanism for antigenic variation in trypanosomes. *Nature* **316:** 562–564

Pays E, Lips S, Nolan D, Vanhamme L & Pérez-Morga D (2001) The VSG expression sites of *Trypanosoma brucei*: multipurpose tools for the adaptation of the parasite to mammalian hosts. *Mol Biochem Parasitol* **114:** 1–16

Petersen TN, Brunak S, Heijne von G & Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* **8:** 785–786

Peterson GI & Masel J (2009) Quantitative prediction of molecular clock and ka/ks at short timescales. *Mol Biol Evol* **26:** 2595–2603

Pérez-Morga D, Vanhollebeke B, Paturiaux-Hanocq F, Nolan DP, Lins L, Homblé F, Vanhamme L, Tebabi P, Pays A, Poelvoorde P, Jacquet A, Brasseur R & Pays E (2005) Apolipoprotein L-I promotes trypanosome lysis by forming pores in lysosomal membranes. *Science* **309:** 469–472

van der Ploeg LH, Schwartz DC, Cantor CR & Borst P (1984) Antigenic variation in *Trypanosoma brucei* analyzed by electrophoretic separation of chromosome-sized DNA molecules. *Cell* **37:** 77–84

Povelones ML, Gluenz E, Dembek M, Gull K & Rudenko G (2012) Histone H1 plays a role in heterochromatin formation and VSG expression site silencing in *Trypanosoma brucei*. *PLoS Pathog* **8:** e1003010

Prakash S, Johnson RE & Prakash L (2005) Eukaryotic translesion synthesis DNA polymerases: specificity of structure and function. *Annu Rev Biochem* **74:** 317–353

Proudfoot C & McCulloch R (2005) Distinct roles for two RAD51-related genes in *Trypanosoma brucei* antigenic variation. *Nucleic Acids Res* **33:** 6906–6919

Prucca CG, Slavin I, Quiroga R, Elías EV, Rivero FD, Saura A, Carranza PG & Luján HD (2008) Antigenic variation in *Giardia lamblia* is regulated by RNA interference. *Nature* **456:** 750–754

Pryde FE, Gorham HC & Louis EJ (1997) Chromosome ends: all the same under their caps. *Curr Opin Genet Dev* **7:** 822–828

Radman M (1999) Enzymes of evolutionary change. *Nature* **401:** 866–7, 869

Rajão MA, Passos-Silva DG, DaRocha WD, Franco GR, Macedo AM, Pena SDJ, Teixeira SM & Machado CR (2009) DNA polymerase kappa from *Trypanosoma cruzi* localizes to the mitochondria, bypasses 8-oxoguanine lesions and performs DNA synthesis in a recombination intermediate. *Mol Microbiol* **71:** 185–197

Raper J, Fung R, Ghiso J, Nussenzweig V & Tomlinson S (1999) Characterization of a novel trypanosome lytic factor from human serum. *Infect Immun* **67:** 1910–1916

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, González JR, Gratacòs M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, *et al* (2006) Global variation in copy number in the human genome. *Nature* **444:** 444–454

Riethman H (2008a) Human subtelomeric copy number variations. *Cytogenet Genome Res* **123:** 244–252

Riethman H (2008b) Human telomere structure and biology. *Annu Rev Genomics Hum Genet* **9:** 1–19

Riethman H, Ambrosini A & Paul S (2005) Human subtelomere structure and variation. *Chromosome Res* **13:** 505–515

Riethman H, Ambrosini A, Castaneda C, Finklestein J, Hu X-L, Mudunuri U, Paul S & Wei J (2004) Mapping and initial analysis of human subtelomeric sequence assemblies. *Genome Res* **14:** 18–28

Robertson DH & Baker JR (1958) Human trypanosomiasis in south-east Uganda. 1. A study of the epidemiology and present virulence of the disease. *Trans R Soc Trop Med Hyg* **52:** 337–348

Robinson NP, Burman N, Melville SE & Barry JD (1999) Predominance of duplicative *VSG* gene conversion in antigenic variation in African trypanosomes. *Mol Cell Biol* **19:** 5839–5846

Robinson NP, McCulloch R, Conway C, Browitt A & Barry JD (2002) Inactivation of Mre11 does not affect *VSG* gene duplication mediated by homologous recombination in *Trypanosoma brucei*. *J Biol Chem* **277:** 26185–26193

Rocha EPC (2004) The replication-related organization of bacterial genomes. *Microbiology* **150:** 1609–1627

Rocha EPC, Smith JM, Hurst LD, Holden MTG, Cooper JE, Smith NH & Feil EJ (2006) Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J. Theor. Biol.* **239:** 226–235

Rost B (1999) Twilight zone of protein sequence alignments. *Protein Eng.* **12:** 85-94

Rudd MK, Friedman C, Parghi SS, Linardopoulou EV, Hsu L & Trask BJ (2007) Elevated rates of sister chromatid exchange at chromosome ends. *PLoS Genet* **3:** e32

Rudenko G, McCulloch R, Dirks-Mulder A & Borst P (1996) Telomere exchange can be an important mechanism of *variant surface glycoprotein* gene switching in *Trypanosoma brucei*. *Mol Biochem Parasitol* **80:** 65–75

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA & Barrell B (2000) Artemis: sequence visualization and annotation. *Bioinformatics* **16:** 944–945

Salmon D, Paturiaux-Hanocq F, Poelvoorde P, Vanhamme L & Pays E (2005) *Trypanosoma brucei*: growth differences in different mammalian sera are not due to the species-specificity of transferrin. *Experimental Parasitology* **109:** 188–194

San Filippo J, Sung P & Klein H (2008) Mechanism of eukaryotic homologous recombination. *Annu Rev Biochem* **77:** 229–257

Scherf A, Figueiredo LM & Freitas-Junior LH (2001) Plasmodium telomeres: a pathogen's perspective. *Curr Opin Microbiol* **4:** 409–414

Scherf A, Lopez-Rubio JJ & Riviere L (2008) Antigenic variation in *Plasmodium falciparum*. *Annu. Rev. Microbiol.* **62:** 445–470

Schwede A & Carrington M (2010) Bloodstream form trypanosome plasma membrane proteins: antigenic variation and invariant antigens. *Parasitology*: 1–11

Shah JS, Young JR, Kimmel BE, Iams KP & Williams RO (1987) The 5' flanking sequence of a *Trypanosoma brucei variable surface glycoprotein* gene. *Mol Biochem Parasitol* **24:** 163–174

Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, Oseroff VV, Albertson DG, Pinkel D & Eichler EE (2005) Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet* **77:** 78–88

Simpson AGB, Stevens JR & Lukes J (2006) The evolution and diversity of kinetoplastid flagellates. *Trends Parasitol* **22:** 168–174

Simpson JT & Durbin R (2012) Efficient *de novo* assembly of large genomes using compressed data structures. *Genome Res* **22:** 549–556

Simpson JT, McIntyre RE, Adams DJ & Durbin R (2010) Copy number variant detection in inbred strains from short read sequence data. *Bioinformatics* **26:** 565–567

Smith CE, Llorente B & Symington LS (2007) Template switching during break-induced replication. *Nature* **447:** 102–105

Smith DH, Pepin J & Stich AH (1998) Human African trypanosomiasis: an emerging public health crisis. *Br Med Bull* **54:** 341–355

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H, Lehväslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, *et al*

(2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* **12:** 1611–1618

Stephens NA & Hajduk SL (2011) Endosomal localization of the serum resistance-associated protein in African trypanosomes confers human infectivity. *Eukaryot Cell* **10:** 1023–1033

Stevens JR & Gibson W (1999) The molecular evolution of trypanosomes. *Parasitol Today (Regul Ed)* **15:** 432–437

Stich A, Abel PM & Krishna S (2002) Human African trypanosomiasis. *BMJ* **325:** 203–206

Stringer JR & Keely SP (2001) Genetics of surface antigen expression in *Pneumocystis carinii*. *Infect Immun* **69:** 627–639

Symula RE, Beadell JS, Sistrom M, Agbebakun K, Balmer O, Gibson W, Aksoy S & Caccone A (2012) *Trypanosoma brucei gambiense* Group 1 Is Distinguished by a Unique Amino Acid Substitution in the HpHb Receptor Implicated in Human Serum Resistance. *PLoS Negl Trop Dis* **6:** e1728

Tait A, Barry JD, Wink R, Sanderson A & Crowe JS (1985) Enzyme variation in *T. brucei* ssp. II. Evidence for *T. b. rhodesiense* being a set of variants of *T. b. brucei*. *Parasitology* **90:** 89–100

Tan KSW, Leal STG & Cross GAM (2002) *Trypanosoma brucei* MRE11 is non-essential but influences growth, homologous recombination and DNA double-strand break repair. *Mol Biochem Parasitol* **125:** 11–21

Taylor JE & Rudenko G (2006) Switching trypanosome coats: what's in the wardrobe? *Trends Genet* **22:** 614–620

Therizols P, Fairhead C, Cabal GG, Genovesio A, Olivo-Marin J-C, Dujon B & Fabre E (2006) Telomere tethering at the nuclear periphery is essential for efficient DNA double strand break repair in subtelomeric region. *J Cell Biol* **172:** 189–199

Thompson JD, Higgins DG & Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22:** 4673–4680

Thon G, Baltz T & Eisen H (1989) Antigenic diversity by the recombination of pseudogenes. *Genes Dev* **3:** 1247–1254

Thon G, Baltz T, Giroud C & Eisen H (1990) Trypanosome variable surface glycoproteins: composite genes and order of expression. *Genes Dev* **4:** 1374–1383

Tiengwe C, Marcello L, Farr H, Dickens N, Kelly S, Swiderski M, Vaughan D, Gull K, Barry JD, Bell SD & McCulloch R (2012a) Genome-wide analysis reveals extensive functional interaction between DNA replication initiation and transcription in the genome of *Trypanosoma brucei*. *Cell Rep* **2:** 185–197

Tiengwe C, Marcello L, Farr H, Gadelha C, Burchmore R, Barry JD, Bell SD & McCulloch R (2012b) Identification of ORC1/CDC6-Interacting Factors in *Trypanosoma brucei* reveals critical features of Origin Recognition Complex architecture. *PLoS ONE* **7:** e32674

Touchon M, Nicolay S, Audit B, Brodie of Brodie E-B, d'Aubenton-Carafa Y, Arneodo A & Thermes C (2005) Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins. *Proc Natl Acad Sci USA* **102:** 9836–9841

Trask BJ, Massa H, Brand-Arpon V, Chan K, Friedman C, Nguyen OT, Eichler E, van den Engh G, Rouquier S, Shizuya H & Giorgi D (1998) Large multi-chromosomal duplications encompass many members of the olfactory receptor gene family in the human genome. *Hum Mol Genet* **7:** 2007–2020

Trenaman A, Hartley C, Prorocic M, Passos-Silva DG, Hoek MVD, Nechyporuk-Zloy V, Machado CR & McCulloch R (2012) *Trypanosoma brucei* BRCA2 acts in a life cycle-specific genome stability process and dictates BRC repeat number-dependent RAD51 subnuclear dynamics. *Nucleic Acids Res*

Tsai IJ, Otto TD & Berriman M (2010) Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol* **11:** R41

Turner CM (1997) The rate of antigenic variation in fly-transmitted and syringe-passaged infections of *Trypanosoma brucei*. *FEMS Microbiol Lett* **153:** 227–231

Turner CM, Aslam N & Dye C (1995) Replication, differentiation, growth and the virulence of *Trypanosoma brucei* infections. *Parasitology* **111:** 289–300

Turner CM, Barry JD, Maudlin I & Vickerman K (1988) An estimate of the size of the metacyclic variable antigen repertoire of *Trypanosoma brucei rhodesiense*. *Parasitology* **97:** 269–276

Uljon SN, Johnson RE, Edwards TA, Prakash S, Prakash L & Aggarwal AK (2004) Crystal structure of the catalytic core of human DNA polymerase kappa. *Structure* **12:** 1395–1404

Valdés J, Taylor MC, Cross MA, Ligtenberg MJ, Rudenko G & Borst P (1996) The viral thymidine kinase gene as a tool for the study of mutagenesis in *Trypanosoma brucei*. *Nucleic Acids Res* **24:** 1809–1815

Vanhamme L & Pays E (1995) Control of gene expression in trypanosomes. *Microbiol Rev* **59:** 223–240

Vanhamme L, Lecordier L & Pays E (2001) Control and function of the bloodstream variant surface glycoprotein expression sites in *Trypanosoma brucei*. *Int J Parasitol* **31:** 523–531

Vanhamme L, Paturiaux-Hanocq F, Poelvoorde P, Nolan DP, Lins L, van den Abbeele J, Pays A, Tebabi P, van Xong H, Jacquet A, Moguilevsky N, Dieu M, Kane JP, De Baetselier P, Brasseur R & Pays E (2003) Apolipoprotein L-I is the trypanosome lytic factor of human serum. *Nature* **422:** 83–87

Vanhamme L, Poelvoorde P, Pays A, Tebabi P, van Xong H & Pays E (2000) Differential RNA elongation controls the variant surface glycoprotein gene expression sites of *Trypanosoma brucei*. *Mol Microbiol* **36:** 328–340

Verstrepen KJ & Fink GR (2009) Genetic and epigenetic mechanisms underlying cell-surface variability in protozoa and fungi. *Annu Rev Genet* **43:** 1–24

Vickerman K (1985) Developmental cycles and biology of pathogenic trypanosomes. *Br Med Bull* **41:** 105–114

Volff JN & Altenbuchner J (1998) Genetic instability of the *Streptomyces* chromosome. *Mol Microbiol* **27:** 239–246

Volkman SK, Hartl DL, Wirth DF, Nielsen KM, Choi M, Batalov S, Zhou Y, Plouffe D, Le Roch KG, Abagyan R & Winzeler EA (2002) Excess polymorphisms in genes for membrane proteins in *Plasmodium falciparum*. *Science* **298:** 216–218

Waterhouse AM, Procter JB, Martin DMA, Clamp M & Barton GJ (2009) Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25:** 1189–1191

Waters LS, Minesinger BK, Wiltrout ME, D'Souza S, Woodruff RV & Walker GC (2009) Eukaryotic translesion polymerases and their roles and regulation in DNA damage tolerance. *Microbiol Mol Biol Rev* **73:** 134–154

Weiden M, Osheim YN, Beyer AL & van der Ploeg LH (1991) Chromosome structure: DNA nucleotide sequence elements of a subset of the minichromosomes of the protozoan *Trypanosoma brucei*. *Mol Cell Biol* **11:** 3823–3834

Welburn SC, Coleman PG, Maudlin I, Fèvre EM, Odiit M & Eisler MC (2006) Crisis, what crisis? Control of Rhodesian sleeping sickness. *Trends Parasitol* **22:** 123-128

Wickstead B, Ersfeld K & Gull K (2004) The small chromosomes of *Trypanosoma brucei* involved in antigenic variation are constructed around repetitive palindromes. *Genome Res* **14:** 1014–1024

Wilkinson SR & Kelly JM (2009) Trypanocidal drugs: mechanisms, resistance and new targets. *Expert Rev Mol Med* **11:** e31

Williams RO, Young JR & Majiwa PA (1982) Genomic environment of *T. brucei VSG* genes: presence of a minichromosome. *Nature* **299:** 417–421

Wirtz E, Leal S, Ochatt C & Cross GA (1999) A tightly regulated inducible expression system for conditional gene knock-outs and dominant-negative genetics in *Trypanosoma brucei*. *Mol Biochem Parasitol* **99:** 89–101

Xong HV, Vanhamme L, Chamekh M, Chimfwembe CE, van den Abbeele J, Pays A, van Meirvenne N, Hamers R, De Baetselier P & Pays E (1998) A VSG expression site-associated gene confers resistance to human serum in *Trypanosoma rhodesiense*. *Cell* **95:** 839–846

Yang W (2005) Portraits of a Y-family DNA polymerase. *FEBS Letters* **579:** 868–872

Yang X, Figueiredo LM, Espinal A, Okubo E & Li B (2009) RAP1 is essential for silencing telomeric variant surface glycoprotein genes in *Trypanosoma brucei*. *Cell* **137:** 99–109

Young JM, Endicott RM, Parghi SS, Walker M, Kidd JM & Trask BJ (2008a) Extensive copy-number variation of the human olfactory receptor gene family. *Am J Hum Genet* **83:** 228–242

Young R, Taylor JE, Kurioka A, Becker M, Louis EJ & Rudenko G (2008b) Isolation and analysis of the genetic diversity of repertoires of VSG expression site containing telomeres from *Trypanosoma brucei gambiense*, *T. b. brucei* and *T. equiperdum*. *BMC Genomics* **9:** 385

Zerbino DR & Birney E (2008) Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res* **18:** 821–829

Zhang Z, Li J, Zhao X-Q, Wang J, Wong GK-S & Yu J (2006) KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics* **4:** 259–263

Zomerdijk JC, Ouellette M, Asbroekten AL, Kieft R, Bommer AM, Clayton CE & Borst P (1990) The promoter for a *variant surface glycoprotein* gene expression site in *Trypanosoma brucei*. *EMBO J* **9:** 2791–2801