Author:
Cook, Kate F

Title:
The minichromosomes of livestock-infecting African Trypanosomes

Author:
Cook, Kate F

Title:
The minichromosomes of livestock-infecting African Trypanosomes

The minichromosomes of livestock-infecting African Trypanosomes.

By Kate Cook

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Master of Science by Research in the Faculty of Life Sciences. School of Biological Sciences - 11/2020

Word count (main text): 17,552

Abstract

Animal-infecting trypanosomes amongst subgenus *Nannomonas* pose severe threats to food security across Sub-Saharan Africa, yet the species associated with these infections have been largely left behind in the modern era of trypanosome research, with only a handful of incomplete genomic sequences available. The objective of this investigation was to characterise a similarly ignored portion of the trypanosome genome – the minichromosomes (MCs) - for six taxa within subgenus *Nannomonas,* and to compare them to the better known *T.b. brucei* MCs*.* MCs are highly repetitive, linear chromosomes involved in the process of antigenic variation, acting as VSG gene libraries. The traditional karyotype visualisation technique Pulsed Field Gel Electrophoresis (PFGE) was refined for the comparative analysis of MCs belonging to the *Nannomonas* species and findings of species-specific MC size classes were used to guide the recovery of minichromosomes from whole genome sequence data. A bioinformatic pipeline for the extraction, assembly and annotation of MCs from PacBio genome data was generated. This was more successful for some species than others, which in itself reflected species-specific differences in MC structure. Statistical comparisons characterised these differences in detail and annotation of assembled MC contigs showed that they also displayed species-specific differences in gene content. These findings challenge our previous understanding of MC structure and genesis, highlighting the importance of avoiding inter-specific generalisations based on *T. brucei* and providing direction for further investigation.

Acknowledgements

Thank you to my supervisors, Dr Tom Williams and Professor Wendy Gibson, who presented this collaborative project opportunity and both provided guidance that enhanced my success with it. A special thanks to Wendy who has welcomed me into her lab over the years since planting a seed of parasite enthusiasm in me from the start of my undergraduate degree.

I would also like to thank other members of the Trypanosome Research Group at Bristol, especially Dr Christopher Kay, with whom I've had many enthusiastic and intellectually stimulating discussions about the mysterious 'dark matter' (his words) of trypanosome genomes. I am grateful for his patience with my endless questions about bioinformatics and for providing much of the code used in these analyses. Thank you to Lori Peacock who prepared several of the PFGE samples and to Rachel Hutchinson and Clare Collett for the general advice and kindness!

It must be acknowledged that my completion of this thesis would not have been possible without the backbone of support provided by loved ones. Thank you to my parents, my Nan (Nutty Norma) and everyone else who has consistently encouraged my academic determination, never questioning my capabilities even during times of personal challenge.

Finally, I am indebted to the fascinating organism that is the trypanosome, which throughout this project has demonstrated to me that it is worth investigating even the smallest parts of what makes us who we are. The resources required to adapt and thrive may be contained in the most unexpected of places.

Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ..... [redacted] ..... DATE: 27/11/2020

Table of Contents

List of figures

List of tables

Chapter 1. Introduction

1.1 African Animal Trypanosomiasis and its impact

Nagana or African Animal Trypanosomiasis (AAT) is a fatal wasting disease of livestock, transmitted cyclically from tsetse fly (*Glossina spp.*) to mammalian host. The causative agents *Trypanosoma spp.* are unicellular haemoparasites, which also infect humans with Human African Trypanosomiasis (HAT) or Sleeping Sickness across the tsetse belt of sub-Saharan Africa.

Trypanosomes have complex life cycles, characterised by the presence of stage-specific developmental forms. The morphology of these forms varies between species but the sequence of events remains relatively conserved between the tsetse-transmitted species: as a tsetse vector bites a host, infective metacyclic trypanosomes enter the host's skin tissue, where they multiply by binary fission, before differentiating into bloodstream form trypanosomes. Bloodstream forms are carried throughout the body via the bloodstream and localise in several possible regions, including myocardial blood vessels (Batista et al., 2019) and the central nervous system (Zweygarth et al., 1987), causing varying levels of disease. Whilst some infected animals may remain asymptomatic due to possessing a trypanotolerance trait (Murray et al., 1990), others are severely affected, becoming anaemic and immunosuppressed (Taylor and Mertens, 1999). When a tsetse fly takes a bloodmeal from the infected host, the bloodstream form trypanosome is ingested, and it differentiates into a procyclic form in the tsetse midgut, before migrating and attaching higher up in the digestive tract as an epimastigote where it goes on to develop into the infective metacyclic form (Matthews, 1999; Coustou et al., 2010). Hybrid and population genomics studies have also proved that some species of trypanosome undergo sexual reproduction when inside the tsetse fly as epimastigotes and procyclics (Gibson and Bailey, 1994; Tihon et al., 2017).

Research on trypanosomiasis has primarily been directed towards human-infecting species, allowing *T.brucei* to become the top model in our understanding of trypanosome biology. However, a 95% decline in the number of reported cases of HAT between 2000 and 2018 (World Health Organisation, 2018) means that the burden of trypanosome infection on human health is near eradication.

Conversely, the effects of AAT remain pressing. AAT is one of the most dangerous disease threats to livestock (Auty et al., 2015) limiting livestock production in the 10 million km² tsetse-endemic zone (Bossche and Delespaux, 2011) of Africa, where camels, horses, sheep, goats, pigs and >46 million cattle are at risk of contracting the disease (Kristjansen et al., 1999). Of all animals used in African agriculture, cattle are the most heavily relied on - producing both milk and meat for human consumption, which means that the disease poses an immediate threat to human food security (Haile-Meskel, 2016). Although it is difficult to know precise prevalence figures for AAT over the continent due to the sheer mass of sampling that would be required, case studies have demonstrated a prevalence in cattle of 15% in the Tororo district of Uganda (Muhanguzi et al., 2014) and 27.08% in Cameroon (Nimpaye et al., 2011), representing East and central parts of the tsetse belt. Although already alarming, these data are likely to be under representative of the prevalence of AAT in all domestic animals across Africa due to the limited number in both host and trypanosome species sampled.

Diseases affecting livestock decreased agricultural productivity by up to 30% in developing countries (Food and Agriculture Organisation, 1990). The direct effects of AAT on cattle include reduced growth of adult animals, reduced grazing levels (Wacher et al., 1994), reduced milk yield, lower calving rates and an increase in calf mortality. This results in at least a 50% reduction in the yield of meat and milk available for human consumption. Indirect effects of AAT on agricultural productivity are mostly concerned with a reduction in number of cattle available for animal powered mechanisation, which reduces the efficiency of crop cultivation (Swallow, 2000). There are also constraints on human migration and colonisation of areas where AAT is prevalent (Reid et al., 1999). The threat of AAT means that administering trypanocidal drugs is common practice amongst African farmers, costing the agricultural industry up to $4.5 billion per year (Haile-Meskel, 2016). The cost of treatment combined with costs of vector control and cattle loss is estimated to be up to $2.8 billion in East Africa (Shaw et al., 2014), affecting the welfare of both producers and consumers across the continent. Implementing sufficient control measures is crucial in alleviating the burdens of economic loss and famine caused by AAT.

Although *T. brucei* does infect livestock, other trypanosome species have more devastating effects, such as the widespread bovine-infecting *T. congolense,* which is more prevalent in cattle than *T. brucei* (Desta et al., 2013). As the differences between human-infecting and livestock-infecting trypanosome species have been elucidated, it has become clear that findings gained from studying *T.brucei* may not be generalisable to other species. *T. brucei* belongs to subgenus *Trypanozoon* unlike *T. congolense* and other livestock-infecting trypanosomes such as *T. simiae* and *T. godfreyi*, which all belong to subgenus *Nannomonas.* A range of biological differences between the subgenera have been noted. In terms of the parasite lifecycle, *T. congolense* epimastigotes and metacyclics develop in the proboscis of the tsetse fly instead of the salivary glands as in *T. brucei* (Peacock et al., 2012). Several surface protein differences have been noted between the two species (Gibson et al., 2017) in addition to observed differences in metabolism (Ryley, 1956; Silvester et al., 2018). In more recent years, next generation sequencing technologies have revealed that *T. congolense* and *T. brucei* genomes display differences in the classes of variable surface glycoprotein (VSG) genes involved in antigenic variation as well as their locations and expression sites (Jackson et al., 2013). These differences are unsurprising considering the fact that different trypanosome species infect different hosts and therefore must be specialised to do so. However, what is surprising is that existing literature on the subject reflects a somewhat limited understanding of this host specificity. This is perhaps because much of what we know about AAT as a disease is built upon generalised findings from human-infecting or one or two animal-infecting species, with few attempts having been made to investigate the species-specific differences between these parasites. Considering a broader range of the trypanosome species that contribute towards the AAT burden may provide a more complete picture of the disease, particularly when it comes down to understanding pathogenicity mechanisms.

1.2 Trypanosome species classifications and distributions

1.2.1 Taxonomy and phylogenetic relationships

Vertebrate-infecting trypanosomes are uniflagellate protozoa belonging to the class Kinetoplastea and the genus *Trypanosoma*, which is monophyletic (Simpson et al., 2002; Hamilton et al., 2004). This diverse genus contains both intracellular and extracellular parasites, with widely varied geographical distributions and host ranges. Trypanosomes that infect mammals are often classified according to their mode of transmission - Stercorarians are species transmitted via the hindgut of blood sucking insect vectors, whilst Salivarians are species transmitted via the mouthparts of biting insects, developing in the anterior part of the digestive tract (Haag et al., 1998). The species that cause African trypanosomiasis in both humans and animals lie within the extracellular Salivarian clade, which consists of four subgenera. These are *Trypanozoon*, *Nannomonas*, *Pycnomonas* and *Duttonella* (Haag et al., 1998, Gibson et al., 2007). The advancement of molecular species identification techniques has enabled new species to be discovered and taxonomic classifications previously based on morphological data to become more refined (Gibson et al., 2001; Adams et al., 2010; Enyaru et al., 2010). Members of subgenus *Trypanozoon* include the three sub-species of *T. brucei: T. b. gambiense,* which causes chronic HAT in Western and Central Africa; *T. b. rhodesiense*, which causes acute HAT in Eastern Africa and *T. b. brucei,* which causes AAT (Gibson, 2007). Also belonging to subgenus *Trypanozoon* are *T. evansi,* which infects several wild and domestic animal species but not humans (Desquesnes et al., 2013) and *T. equiperdum,* which infects horses. The subgenus *Nannomonas* contains animal-infecting trypanosomes only, although there have been reported cases of *T. congolense* infection in humans (Truc et al., 2013). *Nannomonas* encompasses *T. congolense, T. godfreyi, T. simiae* and *T. simiae tsavo. T. congolense* infects cattle and small ruminants as well as wild animals such as rodents, ungulates and primates, which all act as wild reservoirs for the parasite (Njiokou et al., 2004). *T. godfreyi , T. simiae* and *T. simiae tsavo* are predominantly suid-specific species, which infect wild suids (Claxton et al., 1992; Kaare et al., 2007) and cause fatal disease in domestic pigs (Janssen and Wijers, 1974; McNamara et al., 1994; Zweygarth et al., 1994). *Pycnomonas* contains the

recently rediscovered suid-specific *T. suis* (Hutchinson and Gibson, 2015) and *Duttonella* contains the widespread bovine-infecting *T. vivax* (Gardiner, 1989).

1.2.2 A closer look at subgenus *Nannomonas*

Although *Trypanozoon* species are the most extensively researched of the *Trypanosoma* genus, *Nannomonas* is of particular interest when considering AAT. This heterogenous subgenus (Table 1) contains species that cause disease in a wide range of hosts as well as species that appear to be highly specialised to only infect or cause fatal parasitaemia in a limited number of host species. There is also variation in the levels of virulence between species and sub-types that share host ranges and appear morphologically identical (Bengaly et al., 2002).

Following the discovery of several behaviourally distinct species belonging to subgenus *Nannomonas* in the early 1900s, Hoare (1972) concluded that some of these species were synonymous to *T. congolense* and that *T. congolense* and *T. simiae* were the only two species that could be formally defined. Comparisons of their enzyme profiles using isoenzyme electrophoresis confirmed their independent status from each other (Gashumba et al., 1986a). *T. simiae* was initially thought to only infect pigs, with no effect on bovids (Stephen, 1966; Gashumba, 1988) but this has been challenged by observation of a *T. simiae* prevalence rate of 22% in cattle in Sudan (Salim et al., 2011). This study however did not report whether the cattle were symptomatic and to what level, so it is possible that the fatal effects of *T. simiae* infection are only restricted to pigs. Less virulent strains of *T. simiae* have also been identified (van Dijk et al., 1973). It seems that although some level of host-specificity has been observed, the host-parasite dance is perhaps not a simple species-specific matter. It is also unknown whether the variation in pathogenicity levels between *T. simiae* strains is caused by genetic variation within the genome of the parasite or that of the host species, or an interaction between the two.

T. congolense was split into the two subtypes Savannah and Forest due to distinct variation between the enzyme profiles of East and West African *T. congolense* isolates (Young and Godfrey,

1982), suggesting that variation between the isolates may be accounted for by their geographical origin. Comparative DNA typing studies confirmed these findings, providing evidence that the genes encoding proteins involved in antigenic variation differed between isolates from different regions (Majiwa et al., 1986). Further analysis of zymodeme variations between a greater number of stocks revealed that the Western and Eastern groupings were still too simplistic to account for the diversity observed within *T. congolense* (Gashumba et al., 1986b). Since then, an accumulation of data has corroborated that taxonomic distinctions within the species correlate more closely with their specific ecological origin (Majiwa et al., 1992), with *T. congolense* isolates each being classified as one of the three subtypes: *T. congolense* Kilifi, *T. congolense* Forest or *T. congolense* Savannah. Epidemiological surveys of *T. congolense* infection in tsetse flies across Africa complicated this ecological biome-based classification system by identifying the presence of subtypes in biomes different to those in which they were discovered, as well as the presence of multiple subtypes in the same area (Majiwa and Otiento, 1990; Solano et al., 2001). Comparative molecular studies on *T. congolense* uncovered significant differences between the three subtypes - *T. congolense* Kilifi has notably larger minichromosomes than do *T. congolense* Forest and *T. congolense* Savannah (Garside et al*.,* 1994) and all three subtypes have different characteristic satellite DNA sequences (Gibson et al., 1988), which have been used to identify them in the wild. *T. congolense* Savannah and *T. congolense* Forest have been identified in multiple wild animal species with *T. congolense* Forest appearing to have a wider host range (Njiokou et al., 2004). In domestic animals, both species infect sheep, goats and pigs (Gashumba et al., 1988) and all three *T. congolense* subtypes are prevalent in cattle (Gillingwater et al., 2010). When experimentally infected with one of the three subtypes, cattle infected with *T. congolense* Savannah suffered more than those infected with either of the other subtypes, which recovered after three months (Bengaly et al., 2002). Phylogenetic studies based on SSU rRNA and GAPDH genes consistently placed the subtypes into separate clades (Stevens and Gibson, 1999; Hamilton et al., 2004). Despite the differences in behaviour, genetics, epidemiology and pathogenicity between the *T. congolense* subtypes, whether the level of divergence between them is suitable to designate them as separate species remains unclear.

Table 1 Classifications and characteristics of species and subtypes within subgenus *Nannomonas*. The three *T. congolense* subtypes have similar host ranges yet some unique characteristics are present. Similarly, *T. godfreyi* and the two *T. simiae* taxa all infect pigs but display notable genomic differences.

Species	Subtype	Unique characteristics	Domestic animal hosts
T. congolense	Forest	Wide host range and distribution	Cattle, sheep, goats, pigs
	Savannah	Most virulent of the *T. congolense* subtypes to cattle	Cattle, sheep, goats
	Kilifi	Larger minichromosomes than the Forest and Savannah T. congolense subtypes	Cattle, sheep, goats
T. godfreyi		Many large minichromosomes	Pigs
T. simiae		Often causes fatal disease in pigs but some strains are less virulent	Pigs, cattle
T. simiae Tsavo		Originally classed as another *T. congolense* subtype, re-classified due to having SSU rRNA more similar to that of *T. simiae* than *T. congolense*	Pigs

In 1995, restriction fragment length polymorphism (RFLP) analysis of a newly identified *T. congolense* stock characterised it as a unique genotype that did not fit into any of the existing *T. congolense* subtypes. It was designated as a new type and named after the region of Kenya it was discovered in – *T. congolense* Tsavo (Majiwa et al., 1993). Subsequent phylogenetic analysis of SSU rRNA sequences re-designated this new discovery as *T. simiae Tsavo* due it having to a greater sequence similarity to *T. simiae* than to *T. congolense* (Gibson et al., 2001). This species has since been identified in regions other than that in which it was discovered, including Uganda (Magona et al., 2003), Tanzania (Malele et al., 2003) and Zambia (Dennis et al., 2014), proving it to be more widespread than first thought.

As the existence of a growing number of *Nannomonas* species was elucidated, it became apparent that the development of species identification tools would be crucial for gaining an accurate understanding of the species that exist and their relatedness. Morphological identification could not be relied on considering that *Nannomonas* species are difficult to distinguish from each other. Identifying trypanosome species based on where in the tsetse flies they are located (Lloyd and Johnson., 1924) also has limitations, as it cannot distinguish between the *T. congolense* subtypes and other *Nannomonas* species that inhabit the same sites in the digestive tract of the tsetse fly. Species-specific DNA probes developed in the 1980's have proved an invaluable resource, used for decades following their advent to diagnose trypanosome infections in livestock (Enyaru et al., 2010). These probes are

short (120-600 bp) repetitive DNA sequences with a high copy number obtained from nuclear genetic material that had been enzymatically digested and separated on a density gradient gel (Kukla et al., 1987; Gibson et al., 1988). The application of Polymerase Chain Reaction (PCR) enabled the development of highly sensitive species identification assays using the satellite DNA probes (Masiga et al., 1992).

The final subgenus *Nannomonas* species to be discovered was isolated from tsetse flies during epidemiological screening in The Gambia for all species for which probes had been developed. The new isolate did not hybridize with any of the existing probes (McNamara et al., 1989; McNamara & Snow, 1991) and a new specific DNA probe was developed (McNamara et al., 1994). RFLP and karyotype analysis revealed that the new isolate was easily distinguishable from other *Nannomonas* species due to possessing a previously unseen set of enzyme banding patterns and large, abundant minichromosomes (McNamara et al., 1994). *T. godfreyi* is widespread and has been found in Eastern and Western Africa (Lehane et al., 2000; Malele et al., 2003), causing acute infection in pigs.

Although useful as tools for species identification due to their species-specific variation in sequence, the satellite DNA sequences used as probes are not of much use in determining the relatedness of the species within subgenus *Nannomonas.* The satellite sequences are short, repetitive and non-coding, so they do not generate much phylogenetic signal. Additional species-specific DNA probes were developed for this purpose, based on spliced leader gene repeat and glutamine and alanine rich protein (GARP) sequences. Spliced leader gene repeats are small (35 nucleotides in length) sequences located at the 5' ends of messenger RNAs in the cytoplasm (Walder et al., 1986). Their genomic transcript is derived from approximately 200 tandem DNA repeats – also known as the mini-exon gene repeats (De Lange et al., 1983). GARP is a surface glycoprotein expressed only by *T. congolense* epimastigotes (Garside and Gibson, 1995; Butikofer et al., 2002). Comparisons of the mini-exon gene sequences between all the *Nannomonas* species provided evidence that most closely related groups were *T. congolense* Savannah and Forest, with *T. congolense* Kilifi being as far distantly related to them as it is from the other *Nannomonas* species (Figure 1) (Garside and Gibson., 1995). This study also identified the presence of a conserved GARP gene across all three *T. congolense* subtypes, which

did not exist in any other *Nannomonas* species, providing the only evidence that *T. congolense* Kilifi may be related to the other *T. congolense* subtypes rather than being a separate species, as the previously discussed evidence suggested (Gibson et al., 1999; Hamilton et al., 2004).

Attempting taxonomic classification at the sub genus level has proven a useful line of inquiry throughout the history of trypanosome research in that it enabled the discovery of new species. That being said, conflicting conclusions arising from different methodological approaches to classify *Nannomonas* species have not been completely resolved. As is true of any phylogenetic investigation, it is important to highlight that these analyses serve as a framework for pointing researchers towards compelling biological questions rather than definitive conclusions other than the evolutionary relationships between taxa. In themselves, these relationships fail to provide a full understanding of the complexity of the underlying biology such as why there is a level of host-specificity, and the processes of genome evolution involved. When reviewing *T. congolense* variation, Majiwa (1992) stated that understanding the genetic population structure of the species and its subtypes in the future may be achieved "by the global comparative analysis of the genomes of the parasites", which may also be true at the subgenus level. One method of achieving such comparisons would be to use a greater range of strains to recreate previous trypanosome phylogenetic reconstructions based on conserved

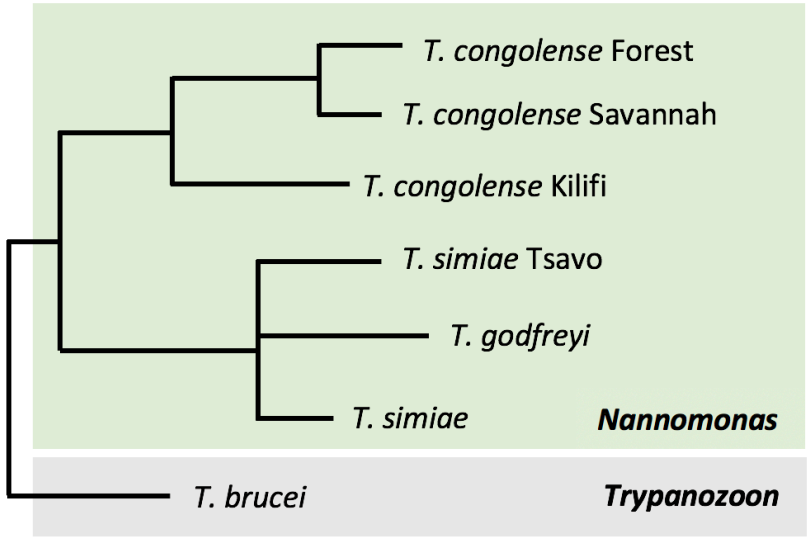

Figure 1 Phylogenetic relationships between the main taxa within subgenus *Nannomonas* (green), redrawn with permission from Gibson et al. (2007). This representation has been drawn to represent species relevant to the present investigation, based on alignments of 18S rRNA sequences previously published by Malele et al. (2003).

genes such as SSU rRNA and GAPDH (Haag et al., 1998; Hamilton et al., 2004). However, such an investigation would not provide any answers regarding the molecular basis of traits such as virulence, host range and specificity, which appear to vary greatly between the *Nannomonas* species. To obtain such answers, it seems pivotal that we have a clear understanding of the mechanisms involved in the host-parasite battle and their genomic basis. Such knowledge is likely to provide insight into the direction towards which comparative genomics studies should be focused to understand these mechanisms from an evolutionary perspective.

1.3 The host-parasite battle

The extracellular nature of animal-infecting trypanosomes renders them vulnerable to the host's immune response throughout the infective cycle in the blood, lymphatic fluids and later in the cerebrospinal fluid. To combat this, African trypanosomes evolved several mechanisms for evading the mammalian humoral immune response. The mechanism that has been the most well characterised is that of antigenic variation in *T. brucei*. Once the infective metacyclic forms develop in the tsetse fly, each parasite expresses many copies of a single variable surface glycoprotein (VSG) on its surface membrane, with the population in the vector expressing a total of approximately 15 different VSGs (Lenardo et al., 1984). After entering the host's bloodstream, these metacyclic VSGs continue to be expressed for up to seven days before they are switched to bloodstream form VSGs (Esser et al., 1982). The parasite then continually changes the molecular structure of its cell surface by sequentially switching the VSG that it exposes to the host. These surface antigens are attached to the outside of the trypanosome membrane by glycosylphosphatidylinositol (GPI) anchors (Pays et al., 1994). Approximately 10^7 (Cross, 1975) rod-like (Blum et al., 1993) VSG molecules form a densely packed protective coat, preventing host antibodies from accessing the non-variant antigens underneath. Even if the infected host initiates an effective antibody response against the VSG that is expressed, periodic switching of the specific expressed VSG causes a change in the primary structure of the surface coat, enabling

the trypanosome population to remain a step ahead and circulate throughout the body of the host establishing a chronic infection.

Much of the early research on this mechanism was focused on trypanosomes belonging to subgenus *Trypanozoon.* Bloodstream form *T. b. rhodesiense* changes the VSG it expresses at a rate of about 10^{-3} changes per cell, per division (Turner and Barry, 1989) and the genome of *T.b brucei* TREU927 contains at least 806 VSG genes (Berriman et al., 2005). 2563 complete and partial VSG genes have been identified in strain Lister 427 (Cross et al., 2014). Metacyclic form trypanosomes prepare to enter the mammalian bloodstream whilst inside the tsetse fly – a single VSG gene is switched on and the VSG is expressed. Metacyclic form expression sites (MESs) (Figure 2) are positioned adjacent to telomeres on the major chromosomes and they consist of a functional VSG gene with a promoter located upstream from it (Borst and Ulbert, 2001). Throughout the infective cycle of bloodstream form trypanosomes in the host, only one VSG is transcriptionally active at a time whilst the rest remain silenced. This monoallelic form of gene expression is not uncommon in medically relevant protozoan parasites that undergo antigenic variation – the malaria parasite *Plasmodium falciparum* also expresses one variable surface protein at a time (Guizetti et al., 2013), as does the intestinal parasite *Giardia lamblia* (Prucca and Lujan, 2009) as a means of host immune evasion. The epigenetic mechanisms involved in silencing expressed variable antigen genes are well understood for all these parasites and vary greatly between them. They include telomere position effect in *T. brucei* (Yang et al., 2009), methylation in *P. falciparum* (Jiang et al., 2013) and RNA interference in *G. lamblia* (Prucca et al., 2008). In *T. brucei rhodesiense,* active VSGs are positioned adjacent to the telomeres at one of the 20 bloodstream form expression sites (BESs) (Figure 2) on the major and intermediate chromosomes (Marcello and Barry, 2007). The active BES is 40-65 kbp in length (Pays and Nolan, 1998) and is transcribed as a polycistronic unit by RNA Polymerase I (Günzl et al., 2003), which also transcribes metacyclic VSGs in the tsetse fly as monocistronic units (Kolev et al., 2017). Important structural components of the *T. brucei* BES transcription unit are the telomeric repeats, the RNA polymerase I promoter, the VSG, a 70 bp repeat sequence and expression site associated gene (ESAG)

arrays (Borst and Ulbert, 2001), which includes genes encoding surface proteins such as transferrin receptors (Young et al., 2008).

Trypanosomes have been experimentally observed to use various methods for switching the VSG gene that is transcribed in the active BES, including transcriptional switching (Borst, 1986) and more commonly - recombination-based methods (Borst and Ulbert, 2001). There are several mechanisms that use recombination to enable VSG switching. In VSG gene conversion, a silent VSG gene located elsewhere in the genome is duplicated and inserted into the active BES, coinciding with the deletion of the previously active VSG. This method involves homologous recombination between regions upstream and downstream of the VSG genes being duplicated and switched, such as the telomeric repeats and the 70 bp repeat sequences that flank more than 90% of the VSGs in *T. brucei* (Marcello and Barry, 2007). These 70 bp repeat sequences are not present in *T. congolense* (Hovel-Miner et al., 2016). Reciprocal VSG recombination involves the crossover of chromosome ends that both have VSGs positioned next to the telomeres, resulting in a silent VSG gene becoming activated in the BES and the previously active VSG gene moving to a chromosome end that does not possess the active BES (Borst et al., 1996). In segmental gene conversion, mosaic VSG genes are formed from VSG gene segments or pseudogenes originating from various genomic locations being spliced together, which generates numerous possibilities in terms of antigenic variability (Barbet and Kamper,

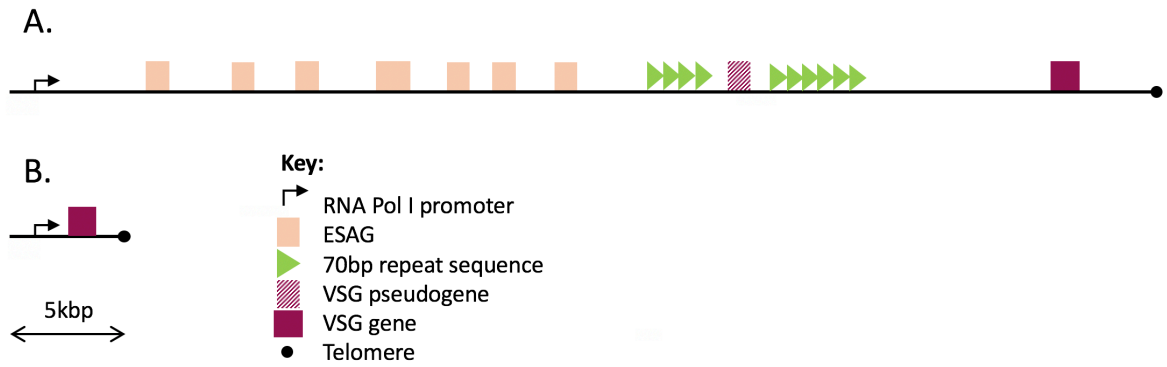

Figure 2 Typical *T. brucei* bloodstream form variable surface glycoprotein (VSG) expression site (BES) (A) and metacyclic form expression site (MES) (B), redrawn from Bangs (2018) and Borst and Ulbert (2001). BES's contain several expression site associated genes (ESAGs) and a tandemly repeated 70 bp repeat motif. Diagrams drawn approximately to scale – BES's range from 45-65 kbp in length and MES's range from 3-6 kbp (Kolev et al., 2017).

1993). This mechanism appears to be the preferred mechanism of *T. brucei* as it enters a chronic phase of infection within the host (Marcello and Barry, 2007). Not all VSG genes and pseudogenes are located at BES's and the subtelomeric regions of the major and intermediate chromosomes of Salivarian trypanosomes. Species belonging to subgenera *Trypanozoon* and *Nannomonas,* also possess small chromosomes or 'minichromosomes' (Weiden et al., 1991), which contain a "silent archive" of VSG genes without promoters (Ersfeld et al., 1999) at each of the telomeric ends (Williams et al., 1982).

In understanding the molecular mechanisms involved in the interactions of trypanosomes with their hosts, particularly those involved in evading the host's immune response, it becomes apparent just how well adapted these parasites are to their multi-host life cycle. *T. brucei* models provide insights into how these mechanisms change over the course of the infective cycle, allowing the parasites to survive in two completely different environments – the digestive tract of the insect and the bloodstream of the mammal. It must be acknowledged however that the relevance of these models to species outside of subgenus *Trypanozoon* is limited*,* which is of particular concern on our pursuit to understand host-parasite interactions in livestock-infecting species such as *T. congolense* that has a wider host range and other *Nannomonas* species that are able to cause varying levels of virulence in different host animals.

Fortunately, researchers have utilised emerging technologies in the post-genomics era to begin to identify where gaps may exist in our understanding of trypanosome immune evasion. Whole genome sequencing has enabled the revelation of striking contrasts between *Trypanozoon* and *Nannomonas* genomes in terms of the VSG gene families present. The *T. brucei* genome contains two types of VSG families – a-VSGs and b-VSGs, defined by the protein's N-terminal domain types (Marcello and Barry, 2007), whilst *T. congolense* has two b-VSG subfamilies and no a-VSGs at all (Jackson et al., 2012). This comparative genomics study by Jackson et al (2012) also characterised VSG phylogenies. The findings revealed that *T. brucei* displayed a greater frequency of recombination within the gene families than did *T. congolense,* suggesting that the mechanisms for generating antigenic diversity may differ between *Trypanozoon* and *Nannomonas* (Jackson et al., 2012). It is not surprising that

between-sub-genus or perhaps even between-species differences in the genomic basis of immune evasion exist, considering that different trypanosome species come into contact with slightly different immune responses depending on their host species. All mammalian immune systems are made up of the same components but comparisons of the genomic basis of these components reveals that they often have distinct evolutionary histories in different species, such as the genes that encode natural killer cell receptors in cattle and pigs (Bailey et al., 2013). When reviewing the necessity of increasing research focus on AAT-causing trypanosome species, Morrison et al. (2016) proposed that bovid ultralong CDR3 domain antibodies may be involved in their immune response to trypanosome burden and investigating whether they interact with VSGs in any particular way may reveal an interesting story of host-parasite evolution. It is this level of enquiry that may prove essential in generating more accurate *Nannomonas* phylogenies or in proving existing ones to be correct whilst adding contextual meaning to their frameworks. Detailed understandings of the evolutionary history of *Nannomonas* trypanosomes may lead to more specific host-parasite approaches to developing AAT-combatting tools against each trypanosome species. Nevertheless, it is important not to jump ahead in pursuing these approaches when trypanosome genomes have still not been fully characterised, especially those of *Nannomonas* species.

1.4 The trypanosome genome

The body of knowledge about trypanosome DNA has been developed over many decades, with increasingly detailed findings occurring as technology for investigating genetic material advanced. Initial investigations were not focused largely on nuclear DNA but on the mitochondrial genome of the parasite - DNA that resides in the kinetoplast, or kDNA (Riou and Pautrizel., 1969). The kinetoplast is an organelle located inside the mitochondria possessed by all trypanosomes and other members of the class Kinetoplastea. Each mitochondrion contains a single kinetoplast, which consists of a network of circular DNA molecules. There are two types of kDNA molecule: ~23 kbp maxicircles, which encode ribosomal RNAs (rRNAs) and proteins involved in mitochondrial respiration (Simpson,

1987; Aphasizheva et al., 2020) and ~1 kbp minicircles, which encode guide RNAs involved in the post-transcriptional editing of maxicircle transcript products (Hong and Simpson, 2003).

In addition to their mitochondrial genome, trypanosomes also have a complex nuclear genome. Early knowledge of this was based on studies investigating specific, functional genomic structures such as the BES (Pays et al., 1989). Microarray studies enabled the identification of genes expressed at different trypanosome life stages and major chromosomes have been sequenced and mapped (El-Sayed et al., 2000; Berriman et al., 2005). Despite such extensive efforts at characterising the genomes of different trypanosome species conducted by researchers employing a wide range of technologies over the years, there remains gaps in our knowledge, specifically in terms of the *Nannomonas* genomes. There is only one annotated reference genome sequence available, representing a single *T. congolense* subtype: *T. congolense* Savannah IL3000 (Jackson et al., 2012). Illumina data from a number of other *T. congolense* Savannah and *T. congolense* Forest isolates has been mapped to this reference (Jackson et al., 2012; Tihon et al., 2017).

1.4.1 Chromosomal karyotype

Trypanozoon and *Nannomonas* nuclear genomes consist of linear chromosomes of three size classes. Prior to the development of the chromosome-sized molecule fractioning method known as pulsed field gradient electrophoresis (PFGE), it was impossible to gain an accurate understanding of the number and sizes of trypanosome chromosomes as they do not condense during mitosis and therefore cannot be visualised individually under the microscope (Vickerman and Preston, 1970). Restriction endonuclease mapping of VSGs and other known trypanosome genes provided some clues about chromosome structure, including the fact that some VSGs are located on chromosome ends (Williams et al., 1982), but it was not until PFGE was invented and applied to trypanosomes (Schwartz and Cantor, 1984; Van Der Ploeg et al., 1984) that the chromosomes could be visualised, and their size classes were defined. Chromosomal size classes were quantified based on the respective migration distances of DNA molecules within a sample through an agarose gel subject to electrical fields of

alternating polarities. Smaller molecules migrate further down the gel, whilst longer ones remain near the top as it takes more time for them to adjust to field direction switches (Schwarz and Cantor., 1984). The first application of PFGE to the *T. brucei* nuclear genome revealed that it consisted of over a hundred chromosomes that fitted into one of three size classes: the megabase or major chromosomes (1-6 Mb), the intermediate chromosomes (200-900 kbp) and the minichromosomes (50-150 kbp) (Van der Ploeg et al., 1984). PFGE also provided opportunities for experiments involving blotting and hybridization with known gene probes that generated further conclusions on the nature of the trypanosome genome, including: the fact that housekeeping genes and active BESs are located only on the major chromosomes (Gibson and Borst, 1986; Melville et al., 1998; Borst et al., 1998); VSG gene location (Melville et al., 2000) and the revelation that there are eleven megabase chromosomes in *T. brucei* (Melville et al., 1998) that are all diploid, with many homologs that are different sizes to each other (Gottesdiener et al., 1990). These 11 chromosomes are numbered from one to eleven, from smallest to largest size, which are ~1 Mb and >6 Mb respectively in *T. brucei* (Melville et al., 1998; Turner et al., 1997). There are however considerable size differences between the chromosomes of different stocks and species (Melville et al., 1999).

PFGE was also applied to various *Nannomonas* species, revealing that *T. simiae, T. godfreyi* and all *T. congolense* subtypes displayed a similar pattern of chromosomal size classes as those discovered in *T. brucei,* with variation in the sizes of their minichromosomes and in predicted numbers of intermediate chromosomes (Garside et al., 1994; Gibson and Borst, 1986). *T. simiae* appeared to have the smallest minichromosomes with some that were less than 50 kbp in length, *T. congolense* Savannah and Forest minichromosomes were both between 50 – 100 kbp and those of *T. congolense* Kilifi and *T. godfreyi* were the largest at between 100 – 250 kbp (Garside et al., 1994). The 177 bp satellite repeat species identification probe hybridized to *T. brucei* 427 minichromosomes but not those of species belonging to subgenus *Nannomonas*, proving that this probe is specific not only to the smallest size class of trypanosome chromosomes (Sloof et al., 1983), but also displays specificity on at least the subgenus level (Gibson and Borst, 1986).

1.4.2 Genome sequence

Whole genome sequencing (WGS) and genome assembly technologies have enabled researchers to obtain knowledge about the origin, transmission and host adaptation of many zoonotic disease agents. Evolutionary genomics studies on severe acute respiratory system (SARS)-causing coronaviruses have enabled researchers to understand the origin of the outbreaks, ultimately contributing towards successful control of them (Zhao, 2007; Benvenuto et al., 2020). Viruses are fast-evolving and have markedly different genomes compared to protozoan pathogens, yet they provide a useful model for understanding the utility of WGS in controlling disease outbreaks. Functional genomics studies on *P. falciparum* have also enabled researchers to gain a clearer understanding of pathogenicity mechanisms and aided the discovery of potential malaria drug targets (Mu et al., 2007).

The *T. brucei* genome, which was published in 2005, was the first trypanosome genome to be sequenced using both shotgun sequencing and bacterial artificial chromosome walking methods (Berriman et al., 2005). Sequencing of the 11 major chromosomes resulted in a number of novel findings relating to the biology of the parasite and the structure of its genome. These include the realisations that the 26 Mb genome consists of 20% subtelomeric genes and most VSG genes in are in fact pseudogenes rather than functional VSG genes (Berriman et al., 2005).

It is still early days in terms of discovering host-specific pathogenicity mechanisms and their origins in the trypanosome genome. A promising beginning to this process occurred when the *T. congolense* IL3000 genome was sequenced using the same methods as those used to sequence *T. brucei.* The VSG gene repertoires of the two species were compared, uncovering differences between the families of VSG genes present and tracing the distinct evolutionary origins of their diversification (Jackson et al., 2012). Use of long-read sequencing later enabled the characterisation of *T. congolense* VSG gene expression sites, which are structurally similar to the telomeric BES's previously modelled in *T. brucei* (Abbas et al., 2018). Interestingly, the majority of VSG genes present in the *T. congolense* genome were predicted to encode functional proteins, unlike those in the *T. brucei* genome that appeared to be mostly non-functional. Additionally, the researchers proposed that the high conservation levels of non-coding elements located at these regions in both the major chromosomes and

the minichromosomes may indicate that they contain promoters and therefore are all able to actively transcribe VSGs (Abbas et al., 2018). It is yet to be confirmed whether *T. congolense* minichromosomes can express VSG genes rather than simply act as a reservoir of them that must undergo recombination with the active BES as in *T. brucei*. The majority of the telomeric sequences described in *T. congolense* IL3000 were found on minichromosomes, which Abbas et al. (2018) identified in single reads from the sequence data. This research has shone a light onto the minichromosomes – a component of the trypanosome genome that has been mostly a side note in previous genomic studies. The remarkable differences in minichromosome size and structure between different trypanosome species has been acknowledged but not yet analysed in great detail. Such analysis is likely to be informative considering the involvement of minichromosomes in antigenic variation.

1.4.3 Minichromosomes

The first minichromosome (hereafter abbreviated as 'MC') to be described was identified in an early VSG restriction endonuclease mapping study, in which there were two copies of a VSG gene at each end of a single DNA molecule in several *T. brucei* clones, located next to double strand breaks likely to represent chromosome ends (Williams et al., 1982). Karyotyping studies revealed that these small linear chromosomes were present in many *Trypanozoon* and *Nannomonas* species (Garside et al., 1994; Gibson and Borst, 1986). The use of Southern blotting hybridisation techniques proved that the MCs contained species-specific satellite DNA sequences, which were utilised to generate probes for species-identification (Gibson et al., 1987; McNamara et al., 1989; Masiga et al., 1992).

The role of MCs as reservoirs of VSG genes (Van der Ploeg et al., 1984) appointed them as an essential component of the trypanosome genome. The probing of PFGE gels with known VSG gene sequences enabled researchers to track VSG gene-related genomic rearrangements, revealing that there is an active mechanism of VSG gene exchange between the subtelomeric regions of minichromosomes and active telomeric BESs on major chromosomes (Robinson et al., 1999). Investigations of the mitotic mechanism of minichromosome segregation showed that they segregate with high levels of fidelity (Wickstead et al., 2003) due to associations with the mitotic spindle (Ersfeld and Gull, 1997),

which explained how individual trypanosomes are able to maintain high levels of VSG gene diversity within their minichromosome populations.

When considering MC structure, the satellite DNA sequences may also be referred to as 'core repeat units' as they are tandemly repeated in the centre or 'core repeat region' of the MC (Figure 3), constituting approximately 55% of total MC DNA and possibly providing stability to MC structure during mitosis (Wickstead et al., 2004). In *T. brucei,* the MC core repeat region has been described as a 'palindromic repeat' as the 177 bp tandem repeat sequence was observed to be inverted at the centre of the core region, which may be a possible origin of replication in MC genesis (Wickstead et al., 2004). All species within subgenus *Trypanozoon* have MCs with a 177 bp repeat sequence (Gibson and Borst., 1986), whereas each species within subgenus *Nannomonas* has a unique core repeat sequence. In *T. congolense* Savannah, the core repeat region consists of 369 bp repeats (Kukla et al., 1987; Gibson et al., 1988) of up to 25 kbp in length (Abbas et al., 2018). Each taxonomic group within *T. congolense* has a unique core repeat unit, with that of *T. congolense* Forest being ~350 bp in length and that of *T. congolense* Kilifi being ~400 bp in length (Gibson et al., 1988). *T. godfreyi* has a 373 bp core repeat unit (Masiga et al., 1996) and *T. simiae* has one that is 500-600 bp in length (Majiwa et al., 1987; Masiga et al., 1992; Majiwa et al., 1993). Subgenus-level differences also appear to exist in structure of the entire core repeat region - unlike *T.b brucei,* which has mostly inverted core repeat regions (Wickstead et al., 2004), *T. congolense* Savannah minichromosome core regions are rarely inverted (Abbas et al., 2018). The structure of core repeat regions in *T. godfreyi* and *T. simiae* MCs is unknown.

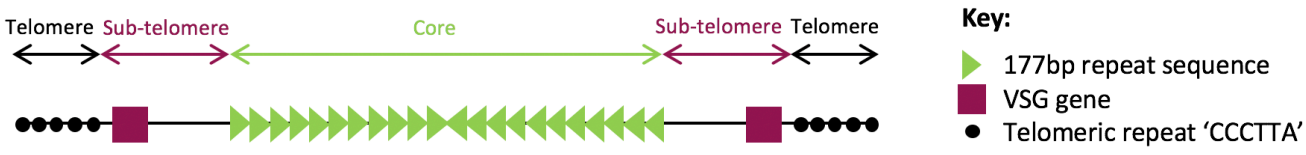

Figure 3 A typical *T. brucei* minichromosome. There is a central inversion point in the core region - arrow direction indicates the orientation of the repeat sequence - flanked by VSG-gene containing subtelomeric regions and a telomere at each end. Redrawn from Wickstead et al (2004).

MCs have the telomeric repeat sequence 'CCCTAA' at each end of the linear molecule (Gull et al., 1998), which is the same sequence that constitutes the telomeres of the major chromosomes (Blackburn and Challoner, 1984) and indeed the chromosomes of most kinetoplastids and eukaryotes (Van der Ploeg et al., 1984).

Cultured *T. brucei* populations display great heterogeneity in the lengths of their MCs. This is less pronounced when analysing the minichromosomal karyotype of a single clone, which remains stable over ~360 generations (Alsford et al., 2000). The mechanisms responsible for MC length variations both within and between species are not clearly understood, however several theories have been proposed. Within-species length differences may be simply due to the shortening and lengthening of telomeric sequences over time (Alsford et al., 2000). These differences may additionally or alternatively be caused by the frequent rearrangements of MCs undergoing recombination with other chromosome ends (Gibson and Borst, 1986), which may initiate changes in the total lengths of their core repeat regions. The core repeat region length appears to be the main determinant of MC length in *T. congolense,* with subtelomeric regions always measuring ~5 kbp in length (Abbas et al., 2018).

Whilst MCs are generally considered as repertoires of silent VSG genes, Abbas et al. (2018) speculated that in *T. congolense,* they may possess active promoters and expression sites due to the presence of conserved non-coding elements adjacent to the core repeat regions amongst populations of MCs. These sequences are not present in *T.b. brucei* MCs and although there is no experimental evidence that they have any regulatory function, the fact that they are conserved across MCs may indicate some functional importance (Abbas et al., 2018). Whether this is related to VSG gene transcription, the maintenance of MC structure or other unknown factors remains unclear. It is also unknown whether *T. godfreyi* and *T. simiae* MCs possess similar conserved noncoding elements that may have regulatory function.

1.5 Aims and objectives

The primary aim in conducting this project is to characterise the content and structure of minichromosomes from populations of representative trypanosome isolates within subgenus *Nannomonas* and to draw comparisons between the species. This will be conducted on two levels of analysis:

1.5.1 PFGE analysis

The objectives in using PFGE are to refine the conditions used in previous karyotyping studies for development of a method for the specific analysis of MCs, and to use this method for comparison of the minichromosomal karyotypes of representative taxa within subgenera *Nannomonas* and *Trypanozoon*.

1.5.2 Bioinformatic analysis

Bioinformatic tools will be tested on currently unpublished sequence data from the genomes of *T.b. brucei* and all species within subgenus *Nannomonas,* with the objective of developing a protocol for the assembly and annotation of minichromosomes. Detailed comparisons of all minichromosomal elements will be drawn between the species, providing a clear picture of areas worth investigating further in future study.

Chapter 2. Pulsed field gel electrophoresis analysis

2.1 Introduction

Early comparisons of the karyotypes of species within subgenus *Nannomonas* provided vague indications of minichromosomal size distributions, with MCs being compressed at the end of the PFGE gels (Gibson and Borst, 1986; Garside et al, 1994). Although PFGE was subsequently utilised to produce more detailed separations of the minichromosomal portion of the karyotype, these studies have mostly been focused on *T. brucei* (Wickstead et al., 2004) or have been conducted for purposes other than comparative karyotyping, such as in minichromosome population analysis (Alsford et al., 2000).

In the present investigation, the sets of PFGE conditions presented for optimal analysis of MCs were arrived at by modifying the conditions used in these previous studies in accordance with the knowledge that shorter switch times combined with shorter overall run times allows a greater level of separation of the smaller chromosomes in relation to that of the intermediate and megabase chromosomes (Bio-Rad Laboratories Inc, 2011). Whilst the ultimate aim in producing PFGE gels is to characterise the minichromosomal content of species amongst subgenus *Nannomonas,* images of all chromosomal size classes will be valuable for gaining a perspective of minichromosomes in their wider genomic context.

2.2 Methods

2.2.1 Sample preparation

A total of fourteen samples were prepared for PFGE analysis using at least two representative isolates for each species within subgenus *Nannomonas* and one representative sample of *T.b. brucei* from subgenus *Trypanozoon* (Table 2). Lysis and deproteinization of DNA occurred in-situ in agarose blocks, which were prepared as described previously by Van der Ploeg et al (1984) at a final cell concentration of 5 x 10^7 trypanosomes per block.

Table 2 All isolates used in this analysis and origins of the samples. At least two representatives were used for each taxon studied using PFGE and some of these were also the isolates for which PacBio genome data was generated and assembled.

Classification				Origin of isolation			
Subgenus	Species	Subtype	Isolate	Host	Location	Year	Reference
Trypanozoon	*T.b. brucei*		J10[a]	Hyena	Zambia	1973	Gibson & Borst, 1984
			Lister 427[b]	Sheep	Uganda	1960	Peacock et al., 2008
Nannomonas	*T. congolense*	Forest	ANR3[c]	Tsetse	The Gambia	1988	Garside et al., 1994
			TSW 103[a]	Pig	Liberia	1975	Garside et al., 1994
		Savannah	IL3000[a]	Cow	Kenya	1966	Gibson, 2012
			GAM2[c]	Cow	The Gambia	1977	Gashumba et al., 1988
			1/148[a]	Cow	Nigeria	1960	Gashumba et al., 1988
		Kilifi	WG84[c]	Goat	Kenya	1981	Garside et al., 1994
			ERA D1[a]	Tsetse	Tanzania	2006	Adams et al., 2008
	T. godfreyi		KEN7[c]	Tsetse	The Gambia	1988	Garside et al., 1994
			ERA F1[a]	Tsetse	Tanzania	2006	Adams et al., 2008
	T. simiae		TV008[c]	Tsetse	The Gambia	1985	Dukes et al., 1989
			ERA C2[c]	Tsetse	Tanzania	2006	Adams et al., 2008
	T. simiae Tsavo		114[c]	Tsetse	Tanzania	2000	Hamilton et al., 2004
			KETRI 3436[a]	Tsetse	Kenya	1970	Gibson et al., 2001

[a] Analysed using PFGE
[b] Analysed bioinformatically
[c] Analysed bioinformatically and using PFGE

2.2.2 PFGE parameters

PFGE was conducted using a contour clamped homogenous electric field apparatus (Biorad CHEFDR-III). Three different programs were used for the production of three gels displaying different levels of analysis. All gels consisted of 1% agarose and were run in 0.5 x Tris-borate-EDTA buffer at 12°C.

For visualisation of the distribution of minichromosomal, intermediate and megabase chromosome size classes in the karyotype, chromosomal DNA from *Hansenula wingei* was used as a marker for fragment size and a three-phase programme was used:

- Block 1:
    - Electrode switching time: 1800 s
    - Induced angle: 106°

    - Voltage: 2 V/cm
    - Time: 15 h
- Block 2:
    - Electrode switching time: ramped linearly from 300 to 900 s
    - Induced angle: 106°
    - Voltage: 3 V/cm
    - Time: 32 h
- Block 3:
    - Electrode switching time: ramped linearly from 60 s to 180 s
    - Induced angle: 120°
    - Voltage: 4 V/cm
    - Time: 8 h

For closer comparison of minichromosomal and intermediate chromosome karyotypes, chromosomal DNA from *Saccharomyces cerevisiae* was used as a size marker and a single-phase programme was used:

- Block 1:
    - Electrode switching time: ramped linearly from 20 s to 60 s
    - Induced angle: 120°
    - Voltage: 4.6 V/cm
    - Time: 32 h

For detailed analysis of the minichromosomal karyotype and species-specific size distributions, a Midrange PFG marker (New England Biolabs) was used as a size marker and a single-phase programme was used:

- Block 1:
    - Electrode switching time: ramped linearly from 1 s to 10 s
    - Induced angle: 120°
    - Voltage: 4.6 V/cm

- Time: 32 h

Upon completion of PFGE, gels were removed from the tank and stained by submersion in 500ml 0.5 x Tris-borate EDTA buffer containing 500µl ethidium bromide stock solution (1mg/L). The container was placed on a shaker and left overnight.

2.3 Results and discussion

The first set of conditions produced a satisfactory separation of all chromosomal size classes (Figure 4A), revealing that the genomes of all species had small chromosomes (<1.05 Mb), although the *T. b. brucei* sample displayed a weak signal. Species within subgenus *Nannomonas* were uniformly stained and well-resolved: the brightest DNA bands for all species, indicating the most abundant chromosomal classes, were amongst the megabase chromosome region of 1.66 – 3.13 Mb and what appeared to be the minichromosome region, although the size marker was not sufficient to define these smaller DNA fragments. The brightest band on the gel existed within this region of the *T. godfreyi* ERA F1 sample, indicating that this isolate may have a greater number of MCs in comparison to the other samples tested. There was a notable absence of abundant DNA bands in the intermediate chromosome region (200 – 900 kbp) for all species.

The general lack of intermediate chromosomes was confirmed by the gel produced under the second set of conditions (Figure 4B), which sufficiently restricted the migration of the megabase chromosomes seen as bright bands of DNA near the top of the gel, whilst allowing separation of the intermediate chromosome region. This portion of the karyotype appeared sparse for most species within subgenus *Nannomonas. T. godfreyi, T. simiae* and *T. simiae* Tsavo had no intermediate chromosomes, *T. congolense* Forest had one band at ~555 kbp, *T. congolense* Savannah had 1-5 bands ranging from 300 – 610 kbp and one of the *T. congolense* Kilifi isolates had several bands ranging from 250 to 400 kbp, whilst the other representative isolate from this subtype had none. All species had MCs < 225 kbp in length.

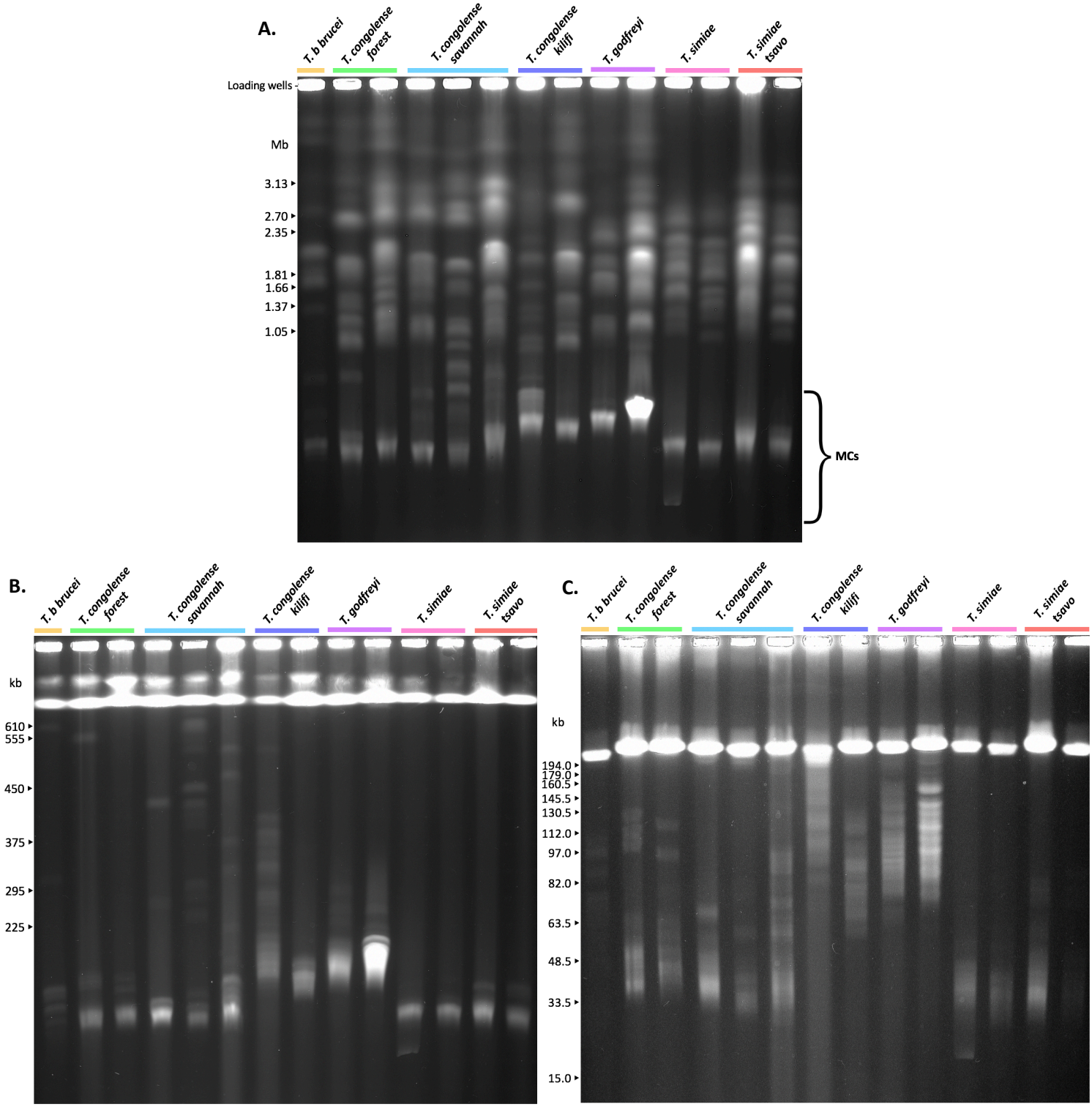

Figure 4 Comparisons of the molecular karyotypes of African Animal Trypanosomiasis (AAT)-causing trypanosomes. These agarose gels stained with ethidium bromide display the size distributions of: A = minichromosomes (<250 kb) in relation to the intermediate (~200 – 900 kb) and some of the megabase (1 – 6 Mb) chromosomes; B = minichromosomes and some intermediate chromosomes and C = the most abundant minichromosomal bands separated at high resolution. Fragment size markers were chromosomal DNA from *Hansenula wingei* (A), chromosomal DNA from *Saccharomyces cerevisiae* (B) and Midrange PFG marker (New England Biolabs) (C). Compression of the megabase chromosomes (B) and the megabase and intermediate chromosomes (C) can be seen as bright bands across all samples near the top of the gels. The minichromosomes were compressed in gels produced using greater switch times (A and B) but have been well separated as a result of the shorter pulse frequencies generated by reducing switch times (C). Samples from left to right: *T. b brucei*: J10, *T. congolense* Forest: ANR3, TSW 103, *T. congolense* Savannah: IL3000, GAM2, 1/148, *T. congolense* Kilifi: WG84, ERA D1, *T. godfreyi*: KEN7, ERA F1, *T. simiae*: TV008, ERA C2, *T. simiae Tsavo*: 114, KETRI 3436. MCs = minichromosomes

Table 3 Fragment length ranges from minichromosomal karyotypes of *T.b brucei* and all species from subgenus *Nannomonas.*

Species	Size range (kbp)
T.b. brucei	70 - 100
T. congolense Forest	35 - 130
T. congolense Savannah	33 - 100
T. congolense Kilifi	60 - 200
T. godfreyi	70 - 200
T. simiae	25 - 50
T. simiae Tsavo	25 - 85

The third set of PFGE conditions, which differed from the previous run only in that switch times were reduced, produced greater separation of the MCs for all species (Figure 4c) and the size marker used enabled precise definition of minichromosomal size distributions (Table 3). *T.b. brucei* MCs were 70 – 100 kbp. *T. congolense* Forest MCs (35 – 130 kbp) and *T. congolense* Savannah MCs (33– 100 kbp) existed within a similar size range, with the most abundant minichromosomal bands being at the lower end of the range (33-50 kbp). The largest MCs were present in *T. congolense* Kilifi (60 – 200 kbp) and *T. godfreyi* (70 – 200 kbp). *T. simiae* had the smallest MCs (25 – 50 kbp), with *T. simiae* TV008 also having a notably distinct smaller band of ~20 kbp. *T. simiae* Tsavo MCs were 25-85 kbp, with the brightest bands appearing at 30 – 40 kbp. *T. simiae* Tsavo KETRI 3436 displayed a weak signal in this gel, likely due to insufficient sample density, however the bands that are faintly visible appear to be congruent with those of *T. simiae* Tsavo 114.

The MCs of *T. congolense* Kilifi, *T. godfreyi, T. simiae* and *T. simiae* Tsavo were relatively evenly distributed over their respective size ranges in comparison to *T. congolense* Forest and *T. congolense* Savannah, which appeared to have more distinct minichromosomal size classes. *T. congolense* Forest had two subsets of MCs – one of which ranged from 35 – 65 kbp in an even distribution and another that contained ~4 distinct bands within the 95 – 130 kbp range.

Within subgenus *Nannomonas,* the minichromosomal karyotype mostly displayed species-level specificity, with the exception of *T. congolense,* which displayed subtype-level specificity. Within this species, *T. congolense* Forest and *T. congolense* Savannah had the most similar patterns in MC

size distribution, with the MCs of *T. congolense* Kilifi being significantly larger and more evenly distributed within their size range. The *T. congolense* Savannah isolates displayed a greater degree of within-subtype heterogeneity than the *T. congolense* Forest isolates, with *T. congolense* Savannah 1/148 containing clear minichromosomal bands ranging from 85 – 97 kbp. These larger MCs were not present in *T. congolense* Savannah IL3000 or *T. congolense* Savannah GAM2. *T. congolense* Kilifi also displayed within-species heterogeneity in minichromosomal size, with *T. congolense* Kilifi ERA D1 possessing a smaller range of MCs (60 – 130 kbp) than *T. congolense* Kilifi WG84 (80 - >200 kbp).

Considering that previous studies reported MCs of up to 250 kbp in *T. congolense* Kilifi and *T. godfreyi* (Garside et al., 1994), it is possible that this gel did not quite capture the full range of MCs for both of these species. The gel produced from the second set of conditions (Figure 4b) displayed chromosomal bands between 200 – 300 kbp in *T. congolense* Kilifi WG84 and both *T. godfreyi* isolates, however these bands were faint in comparison to those displayed on the final gel that was used for size characterisation, suggesting that the majority of minichromosomes were in fact <200 kbp.

The gels produced in this analysis have uncovered novel findings concerning the diverse nature of MCs within subgenus *Nannomonas,* providing refined separations of various regions of the karyotype and the most detailed descriptions of the comparative size distributions of MCs to date. Previous estimates of the species-specific MC size distributions were revised due to the PFGE conditions in this study enabling more specific visualisation of them. Some of these measurements were expanded upon, with 50-100 kbp in previous karyotyping studies of *T. congolense* Savannah (Garside et al., 1994) being measured as 35 – 130 kbp in the present study. Some previous estimates were refined, with *T. simiae* MCs ranging from 25 – 50 kbp here as compared to previous estimates of <50 – 100 kbp (Garside et al., 1994). The measurements described in this investigation are likely to be more accurate than karyotyping studies that did not provide such a 'zoomed in' view of the MCs or did not use a size marker allowing the accurate measurement of fragments <250 kbp. However, it must be considered that the size distribution findings reported here may not be entirely generalisable on the species- and subtype- levels for those in which isolates displayed within-species karyotype heterogeneity. The within-species variation in the minichromosomal karyotypes of *T. congolense*

Savannah and *T. congolense* Kilifi is also a novel finding within this subgenus, although the same pattern has been observed in *T.b brucei* MCs (Gibson and Borst, 1986). VSG gene exchange and the chromosomal rearrangement that occurs as a result of it has been proven to cause variation in *T. brucei* intermediate chromosome lengths, which may explain why different isolates from the same species can display such vast differences in intermediate chromosome size class distributions (Van der Ploeg et al., 1984). Considering that MCs are subject to the same recombination processes, it is likely that the same explanation accounts for the dynamic nature of MC size distributions. The minichromosomal banding patterns in *T. congolense* Forest, *T. godfreyi, T. simiae* and *T. simiae* Tsavo were relatively constant between isolates within each species in the present study, but it may be necessary to analyse a wider range of isolates from each species to determine if this is truly the case.

This analysis provided some useful baseline information for any comparative genomics studies that may be pursued in relation to trypanosomes from subgenus *Nannomonas.* It has enabled the identification of MCs and their size distributions within populations of all species from the subgenus. Future investigations may involve probing the southern blots of the PFGE gels produced here for known minichromosomal elements such as the respective core repeat sequences of each species. It would be interesting to discover whether these known structural components of MCs are entirely species/subtype specific or whether they also vary between MCs of different isolates or between MCs of different sizes within the same isolate.

Chapter 3. Assembly and annotation of minichromosomal genomes

3.1 Introduction

The currently published trypanosome genome assemblies contain only megabase chromosome data, with no MCs or intermediate chromosomes having been assembled and annotated. The *T. congolense* reference genome (IL3000) consists of assembled short read sequence data that was mapped to the 11 megabase chromosomes of the *T.b. brucei* 927 reference genome, with remaining genomic fragments placed in a 'bin' chromosome (https://tritrypdb.org/tritrypdb/app/record/dataset/DS_26f6be1159#GenomeHistory). The use of short-read sequencing combined with the nature of the intermediate and MCs are likely to have caused this exclusion of MCs from the assemblies. It is well known that short reads do not allow highly repetitive and low-complexity sequences to be resolved (Pollard et al., 2018). This is primarily due to the fact that short reads tend to be shorter than 300 bp in length, which means individual repeat units longer than 300 bp are not resolved and repeat regions > 300 bp have no genomic context in the read pool. In contrast, long-read sequencing allows these long repeat units to be captured in their entirety. This has resulted in the development of higher quality, more repeat-inclusive *de novo* genome assemblies in the past decade (Huddleston et al., 2014), including those of the intracellular animal-infecting trypanosomes *Leishmania infantum* and *L. braziliensis* (González-de la Fuente et al., 2017, 2019). More broadly, the ability of long reads to resolve repetitive regions has uncovered an abundance of information regarding different types of repetitive elements. These findings include the chromosomal locations of tandemly repeated genes in *L. infantum* (González-de la Fuente et al., 2017), the reconstruction of transposable elements within *Drosophila* genomes (McCoy et al., 2014) and the resolution of long stretches of short tandem repeat sequences involved in human genomic disease (Schmidt and Pearson, 2016).

The 177 bp repeat unit that forms the MC core region in *T.b. brucei* is estimated to constitute 10-20% of the whole genome and the repetitive INGI/RIME transposable element contributes a

further 5% of genomic material (Bhattacharya et al., 2002). Considering these repeat elements alone, it is apparent that at least 25% of the *T.b. brucei* genome sequence is missing from the published reference genome assembly for the species (Berriman et al., 2005). A similar void is likely to exist in the *T. congolense* reference genome that was mapped onto that of *T.b. brucei.* Given the importance of repetitive sequences in driving the genome evolution of parasitic protozoa (Wickstead et al, 2003), it seems imperative that long-read sequencing and repeat-friendly assembly methods ought to be used in future trypanosome genome projects. This is especially relevant to the intermediate and MCs, which should be fully characterised at the genomic level considering their pivotal role in antigenic variation.

The application of long-read sequencing has already provided insights into MC structure, as they fall within the 10-80 kbp length range of reads produced by long read technologies (van Dijk et al., 2018**)**. Recent research has identified MCs within single reads from PacBio long-read sequencing data for *T. congolense* IL3000 and *T. congolense* 1/148 (Abbas et al., 2018), yet this has likely only uncovered a small proportion of MCs and may not provide an accurate picture of MC sequences given the 11-15% error rate of PacBio long reads (Rhoads et al., 2015). Assembling MCs is necessary for resolving such errors, yet previous attempts at doing so proved unsuccessful (Cross et al., 2014).

Genome assembly poses its own challenges in terms of fully resolving long, repetitive regions. Tandem repeats are prone to mis-assembly, whereby the repeated sequence is collapsed into a single copy (Treangen and Salzberg, 2012). Although there has been little effort at resolving the sequences and structures of repetitive regions in trypanosome reference genomes, it may be possible to physically map them using a combination of labour-intensive steps such as those conducted by Wickstead et al (2004) to construct the first high-resolution *T.b. brucei* MC maps. However, this would not provide the same level of information that bioinformatic approaches offer and long-read sequencing may provide a more efficient solution if there are reads long enough to span entire repetitive regions. Assembling long reads instead of short reads should also allow greater overlap of partial repetitive regions captured within the single reads.

The PFGE analysis in Chapter 2 provided knowledge of MC size for various taxa in subgenus *Nannomonas* (Table 3). All species have abundant MC but MC size varies between species. *T. congolense* Forest and Savannah groups have MC ~30-130 kbp, whilst *T. congolense* Kilifi and *T. godfreyi* MCs are larger at 60-200 kbp in length. *T. simiae* and *T. simiae* Tsavo MCs both have a smaller size range of 25-50 kbp and 25-80 kbp, respectively. The objectives of the analyses presented in this chapter are (1) to recover a representative set of MCs from each taxon by mining PacBio long-read data, (2) to determine the structure of MC and the causes of length variation and (3) to identify the protein-coding genes present on MC and compare between taxa.

3.2 Methods

3.2.1 Minichromosome recovery, assembly and annotation

The long-read genomic sequence data for the seven taxa from subgenus *Nannomonas* (Table 2) used in this investigation was produced using SMRT sequencing technology (Pacific BioSciences, USA) as part of ongoing trypanosome genome projects by the Trypanosome Research Group at University of Bristol. Genomic data for *T. b. brucei* Lister 427 produced using the same methods was obtained from the European nucleotide archive via ENA accession number PRJEB18945 (Müller et al., 2018).

For each species, MCs were obtained either from raw single genomic reads or from the assembly of filtered read pools following a series of methodological steps optimised specifically for the sub-assembly of partial MC, using the highly repetitive telomeric and core repeat sequences (Figure 5). Due to the complex and MC-specific nature of this sub-assembly protocol, preliminary investigations of various parts of the workflow were necessary to test which tools would yield the most useful (if any) results for MC sub-assembly. The assembly programs Flye (Lin et al., 2016) and Canu (Koren et al., 2017) were compared to discover which would assemble the greatest number of MCs for each species. Canu was chosen for MC sub-assembly as it was able to assemble complete MCs for all species amongst subgenus *Nannomonas,* whilst Flye did not assemble any MCs for *T. godfreyi* KEN7, *T. simiae* TV008 and *T. simiae* Tsavo 114. Canu was at least twice as effective at MC sub-assembly than Flye for all species, as determined by the number of complete MCs assembled from the read pool (Appendix

A, Table A1). Whilst these exploratory tests were conducted by myself, the final, automated sub-assembly pipeline (Figure 5) was written by Dr Christopher Kay and included iterations of the assembly and filtering process for the optimal extraction of MC sequence-containing reads.

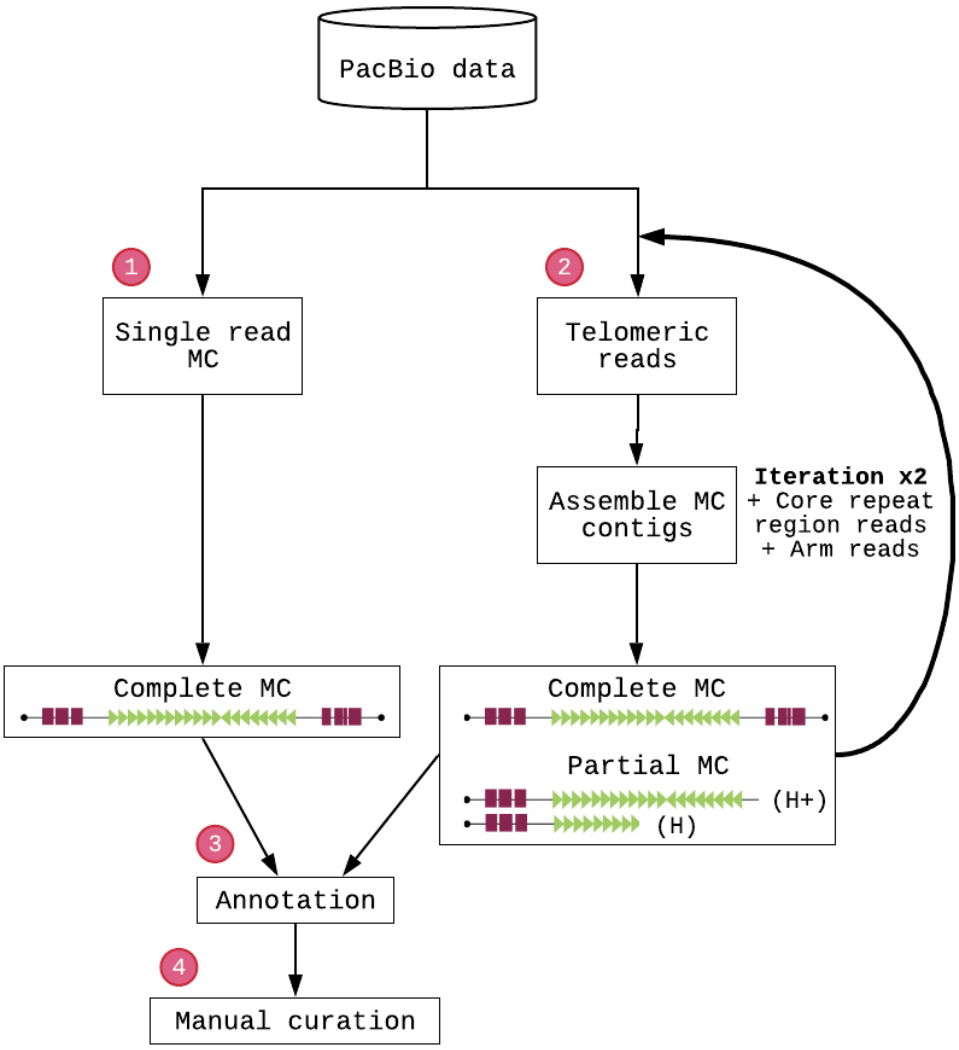

Figure 5 Simplified bioinformatic workflow for the assembly and annotation of Trypanosome minichromosomes from PacBio long read genome data. See text section titled '3.2.1 Minichromosome recovery, sub-assembly and annotation' for detailed explanation of each step and the programs used. Green arrowheads = MC repeat unit, maroon rectangles = VSG and other genes, black circle = telomere.

The numbered stages of the final pipeline represented in Figure 5 are as follows:

1) Recovery of MCs captured within single reads

MCs fall within the length range of reads produced using SMRT sequencing (10,000-80,000 bp), which enabled 'single read MC' sequences to be filtered out from the genomic read pools. Custom Bash scripting was used to search for all reads that contained the telomeric sequences 'TAACCC'

or ‘GGGTTA’ (Van der Ploeg et al., 1984) within the first and last 500 bp of each read. Reads that contained telomeres at both ends were considered as ‘Complete MC’ before being manually inspected (step 4 below).

2) Assembly of telomeric, core repeat region and arm reads

An iterative process was used to produce partial and complete MCs from the assembly of reads containing telomeric, core repeat region and arm (the subtelomeric region running from the core repeat region to the telomere) sequences as follows:

a) A pool of ‘telomeric reads’ was extracted by searching the PacBio data for reads that contained the telomere sequence within 500 bp at either the beginning or end of the read. These telomeric reads were sub-assembled to produce the first pool of assembled complete and partial MC contigs. Canu was used for all three rounds of sub-assembly using reads >500 bp (minReadLength=500) with a predicted genome size of 2000 kbp (genomeSize=2000k). The mhap overlapper was used (overlapper=mhap) with the following settings: utgReAlign=true, enableOEA=true correctedErrorRate=0.15.

b) Core repeat region unit sequences were determined de novo by applying Tandem Repeat Finder (TRF) (Benson, 1999) to the smallest MCs assembled in (2a) using the following parameters: 2 (match), 7 (mismatch), 7 (indels), 80 (PM), 10 (PI), 100 (Minscore), 1000 (MaxPeriod) -f -d -m -h. The repeat sequences obtained were aligned and re-orientated using MAFFT v7.427 (Katoh and Standley, 2013) and then de-duplicated by clustering using CD-HIT (Li and Godzik, 2006) with a sequence identity threshold of 0.95. BLASTn v2.2.31+ (Altschul et al., 1990) was used to filter all ‘core repeat reads’ out from the PacBio data using the derived repeat sequence with an E-value of 10^{-7}.

c) The FASTA file of core repeat reads was concatenated to that of the telomeric reads and the combined reads were de-duplicated using CD-HIT (as above) and assembled using Canu (as above), producing a second pool of assembled complete and partial MC contigs.

d) Partial MC contigs that contained either a core repeat region plus part of an MC arm or a telomere plus part of an MC arm were masked using TRF (Benson, 1999) so the repetitive

sequences were removed and only the unique 'arm' sequences remained. BLASTn v2.2.31+ (Altschul et al., 1990) was used to filter all 'arm reads' out from the genomic read pool using the derived arm sequences with an E-value of 10^{-7}. The arm reads were concatenated with the core repeat reads and the telomeric reads and the combined read pool was then de-duplicated and assembled using the same methods as above, producing a third pool of assembled MC contigs.

e) The FASTA files of assembled contigs from the three rounds of sub-assembly were concatenated and deduplicated as above, producing the final assembled pool of complete and partial MC contigs.

3) MC annotation

Whole and partial (H+ contigs; Figure 5) MCs were annotated by the Companion web server tool (Steinbiss et al., 2016) using *Trypanosoma brucei TREU 927* as a reference species for extracting matching annotations using Rapid Annotation Transfer Tool (Otto et al., 2011). De novo gene prediction by AUGUSTUS was enabled using a score inclusion threshold of 0.6. The list of proteins generated by Companion was filtered using tBLASTn (Altschul et al., 1990) with an E-value of 10^{-7} against a database of annotated *T. congolense* IL3000 proteins (Jackson et al., 2012) which were acquired from TriTrypDB (Aslett et al., 2010). Clustering of filtered protein sequences was conducted using C-HIT (Li and Godzik, 2006) with default parameters and any hypothetical proteins that had >90% sequence match to VSGs were declared as VSGs, as was the case for other proteins.

4) Manual curation of single read and assembled minichromosomes

MCs generated by the assembly pipeline were inspected by visual analysis of self-dot plots generated with FlexiDot using the default calculation parameters (Seibt et al., 2018). The purpose of this analysis was to contribute to a final contig filtering step, whereby the final populations of MC contigs presented here were manually curated for each species. Contigs were classed as genuine MCs and were included in statistical outputs if they contained only two telomeres (one for H and H+ contigs) and at least two complete copies of the core repeat unit. Contigs that contained multiple housekeeping genes and appeared to be assembled subtelomeric ends of major

chromosomes based on comparison of the annotated sequences to those on the *T. congolense* IL3000 major chromosomes (Jackson et al., 2012) using BLASTp (Altschul et al., 1990) were discarded. Dot plots were also utilised to manually characterise some general structural features such as gene locations and the inverted structure of some core repeat regions. The repeat sequences detected in MC contigs in step 2b. were compared to the existing satellite repeat species identification probes using BLASTn (Altschul et al., 1990).

3.2.2 Statistical analysis

Some statistical measurements were generated as part of the sub-assembly pipeline by Dr Christopher Kay (unpublished). Other assembly metrics, such as assembled MC contig N50 and GC content of core repeat sequences, were generated using custom Python scripts. MC populations were assessed statistically in terms of three variables: MC length, core repeat region length and arm length. For each of these variables, univariate ANOVA tests and post hoc Tukey HSD tests were conducted using R project (R Core Team, 2013) to determine whether significant differences existed between species. Pearson correlation coefficient, r^2 and *p* values were generated using the Python package Pingouin (Vallat, 2018) to investigate whether the variables were related at the within-species level. Final statistical data visualisation plots were generated using the MatPlotLib Python library (Hunter, 2007).

3.3 Results and discussion

3.3.1 Single read minichromosomes

MCs that fell within the range of the length of the PacBio reads could be captured within single reads (Figure 6). These "single-read MCs" were filtered out from the genomic read pools and were confirmed to contain telomeric repeats within the first and last 500 bp of each read (Figure 5). Single read MCs were identified in raw reads at varying numbers per species (Table 4). The number of MCs filtered out of the read pool per species somewhat appeared to reflect the raw read pool size, except in *T. simiae* ERA C2, which had a disproportionately greater number of single-read MCs compared to the other species in relation to read pool size (Table 4). 217 MCs were filtered out from the 22.3 gbp read pool for *T. simiae* ERA C2, whereas only 53 were identified for *T. congolense* GAM2, which had only a

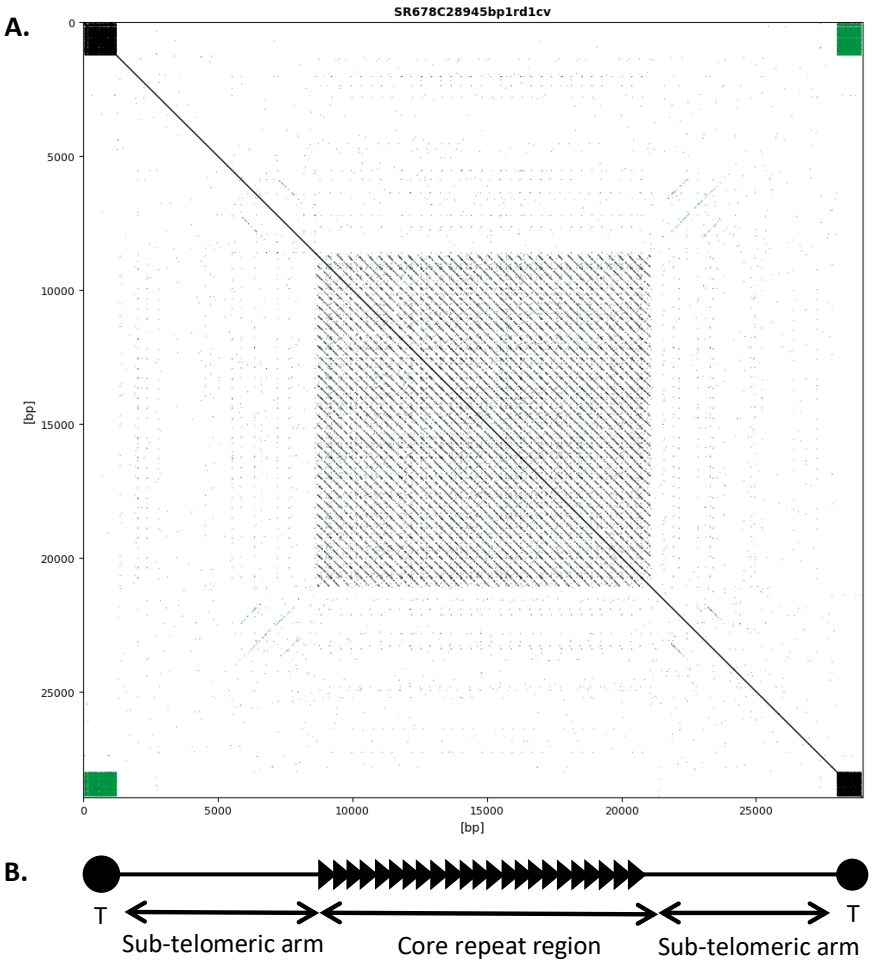

Figure 6 Dotplot of a *T. congolense* Savannah single read MC (A) and representative drawing detailing MC structure (B). This single read MC is ~28 kbp in length with a central ~13 kbp core repeat region consisting of the 369 bp satellite repeat sequence (non-inverted in this case) flanked by unique 7-8 kb subtelomeric arms and terminating with telomeric repeats 'GGGTTA' at each end. T = telomere.

Table 4 Minichromosomes identified in and assembled from genomic data for eight trypanosome isolates. Species abbreviations are: Tbb = *T. brucei brucei;* Tcf = *T. congolense* Forest; Tcs = *T. congolense* Savannah; Tck = *T. congolense* Kilifi; Tgo = *T. godfreyi;* Tsi = *T. simiae;* Tst = *T. simiae* Tsavo. ? = data not available.

Species	PFGE size range (kbp)	PacBio reads (gbp)	Pacbio read N50 (kbp)	No. single read MCs	Assembled MCs			Total no. whole MCs
					No. whole	Whole N50 (kbp)	No. partial	
Tbb 427	70 - 100	?	?	10	8	42.1	62	18
Tcf ANR3	35 - 130	2.1	7.9	2	12	31.5	129	14
Tcs GAM2	33 - 100	18.1	21.4	53	207	29.2	118	260
Tck WG84	60 - 200	9.8	21.6	4	18	77.9	263	22
Tgo KEN7	70 - 200	6.0	23.6	1	15	72.0	248	16
Tsi TV008	25 - 50	2.5	8.3	14	1	-	30	15
Tsi ERAC2	25 - 50	22.3	25.9	217	3	76.0	25	220
Tst 114	25 - 85	10.2	22.4	47	4	38.1	70	51

slightly smaller read pool at 18.1 gbp. This may reflect the fact that *T. simiae* MCs are smaller than those of other species (Table 4) and so they are more likely to be captured within single reads. At the other end of the spectrum, only one MC was filtered out from the *T. godfreyi* KEN7 read pool, which may reflect the fact that this species has some of the largest MCs of all species tested (Table 4), so most would have been too large to have been captured within a single read. Thus, the viability of single-read analysis of MCs is species-dependent, providing useful results for species that have MCs < 50 kbp but producing a limited yield of MCs for species in which they are mostly larger than this size, such as *T. godfreyi* and *T. congolense* Kilifi (Table 4, Figure 8).

The process of filtering reads of a specific size range out of a read pool, which is inherent to MC analysis, depends largely on the quality and length of reads present in the sequence data. In this part of the analysis, the N50 statistic conventionally used to assess genome assembly quality acts as a measurement of the usefulness of mining the read pool for completely intact MCs. In the cases of *T. congolense* Forest ANR3 and *T. simiae* TV008, raw read pools were small (2-2.5 gbp) and N50 values were less than 50% of those for the other species, which all had N50's greater than 20 kbp. The low N50 values for these two species indicate that there were fewer MC-length reads present in the read pools compared to those available for all other species. Despite similar N50 values (8.3, 7.9), filtering the *T. simiae* TV008 read pool resulted in the identification of 7x the number of MCs than the same

process did for *T. congolense* ANR3 (14 compared to 2; Table 4), which may be due to the smaller size range of MCs in *T. simiae* compared to *T. congolense* Forest. It is important to note that single reads suffer a greater rate of sequencing error compared to assembled contigs generated by overlapping and correcting reads. This does not pose a significant limitation in terms of using single read MC to gauge the numbers and sizes of MC present in different species but it may hinder other levels of analysis such as open reading frame (ORF) prediction.

A common feature to all species tested except *T.b. brucei* was that a number of chromosomes 5-12 kbp in length existed in the read data. These very small chromosomes were not included in further analysis as they did not contain core repeat regions and therefore could not be defined as true minichromosomes based on the accepted model of what constitutes an MC (Figure 3). In *T. simiae* and *T. simiae* Tsavo, several of these were 8 – 10 kbp in length and contained one or two copies of the core repeat unit; these were included in the analysis when they adhered to the criteria of what constitutes an MC as defined in step 4 of the Methods section (3.2.1) above.

3.3.2 Assembly of whole and partial MC

Additional whole and partial MCs were assembled via the pipeline shown in Fig 5 as explained in the Methods section. Numbers of whole and partial (H+) MCs assembled for each species are shown in Table 4. The largest dataset was that of *T. congolense* Savannah GAM2, for which 207 whole and 118 partial MCs were assembled. The assembly pipeline did not work as effectively for other taxa, particularly *T. simiae* and *T. simiae* Tsavo with only 1-4 complete MCs assembled for each, regardless of read pool size and quality.

Some assembled *T. congolense* Savannah, *T. simiae* and *T. simiae* Tsavo MC adhered to the existing *T.b. brucei* model of core repeat region structure (Wickstead et al., 2004), with the exception that the inversion point of the core repeat region was slightly off centre rather than completely central (Figure 7).

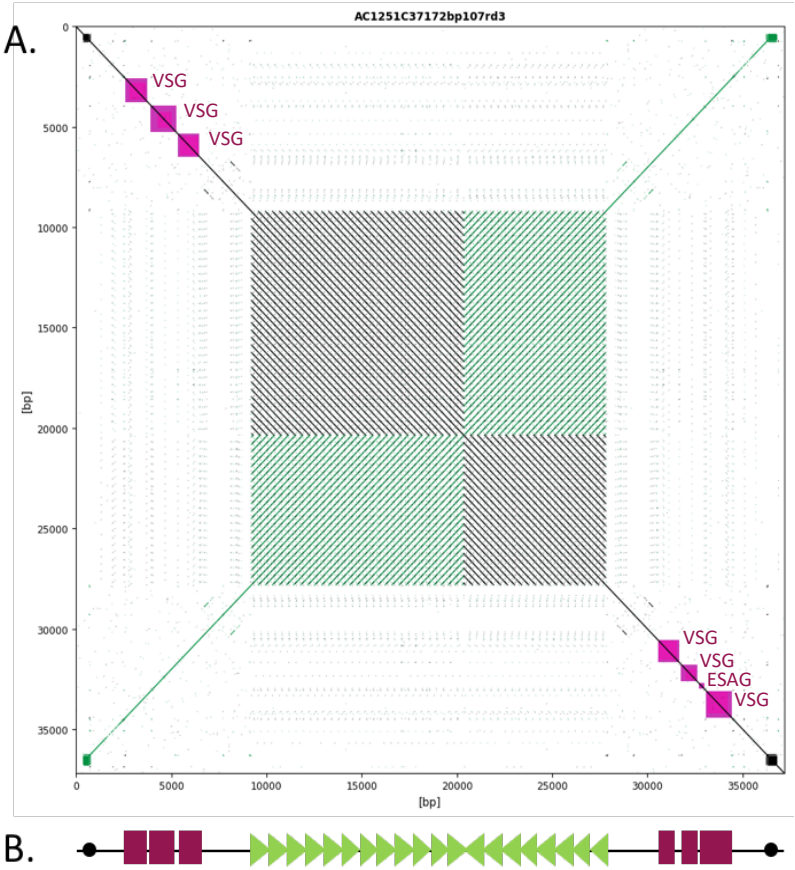

Figure 7 *T. congolense* Savannah minichromosome that matched existing models of MC structure represented by A: a dot plot and B: a representative drawing.

The total number of whole MCs recovered was the greatest for *T. congolense* Savannah GAM2 (260) and *T. simiae* ERA C2 (220). For *T. congolense* Savannah, most whole MCs were recovered by assembly of the PacBio reads, whilst the majority of *T. simiae* MCs were recovered simply by filtering the PacBio reads. Considering that the size range of *T. simiae* ERA C2 MCs visualised by PFGE analysis was 25-50 kbp (Table 4), it is possible that all MCs for this species have been recovered as MCs extracted covered this size range (Table 5). For *T. congolense* Savannah GAM2, the sizes of MCs extracted bioinformatically (24.1 – 64.1 kbp) do not cover the known size ranges from PFGE analysis (33 – 100 kbp), so it is unlikely that all MCs for this taxon have been uncovered, despite the high number of whole MCs extracted.

3.3.3 Length variation of MC

The lengths of whole MCs were compared between taxa (Table 5, Figure 6). *T. congolense* Kilifi had the largest MCs with a median of 74.0 kbp, closely followed by *T. godfreyi* (median = 70.0 kbp), whilst *T. congolense* Savannah GAM2 had the smallest MC (median = 29.3 kbp). Although *T. simiae* TV008 appeared to have the smallest MC (median = 8.5 kbp), the findings for this isolate may not be representative of the species due to the poor-quality sequence data. *T. simiae* ERA C2 MCs had a median length of 40.0 bp, which corresponds more closely to the PFGE findings for *T. simiae* MC length (= 25.0 – 50.0 kbp, Table 4).

These MC length findings are in broad agreement with the species-specific size patterns identified in the PFGE analysis (chapter 2.3). However, the MCs identified in raw read and sub-assembled data do not represent the larger end of the size range for all species except *T. simiae.* In PFGE images, *T. congolense* Kilifi and *T. godfreyi* clearly have multiple minichromosome bands between 100 – 200 kbp in size (Fig 4, chapter 2.3). However, for both species, all MCs assembled were < 100 kbp in length (Figure 6).

Table 5 Statistical measurements of minichromosome length, core repeat region length and subtelomeric arm length. Species abbreviations are: Tbb = *T. brucei brucei;* Tcf = *T. congolense* Forest; Tcs = *T. congolense* Savannah; Tck = *T. congolense* Kilifi; Tgo = *T. godfreyi;* Tsi = *T. simiae;* Tst = *T. simiae* Tsavo.

Species	MC length (kbp)		Core length (kbp)		Arm length (kbp)	
	Median	Range	Median	Range	Median	Range
Tbb 427	37.3	21.3 – 53.6	15.8	0.5 – 36.4	8.4	2.1 – 40.1
Tcf ANR3	32.3	21.8 – 44.5	10.3	1.0 – 27.1	9.6	6.2 – 16.4
Tcs GAM2	29.3	24.1 – 61.3	14.1	0.7 – 68.0	7.9	5.1 – 30.3
Tck WG84	74.0	42.4 – 112.3	23.5	1.6 – 56.7	30.1	0.8 – 54.5
Tgo KEN7	70.0	43.4 – 91.2	20.4	2.0 – 53.0	25.8	9.7 – 63.1
Tsi ERAC2	40.0	5.0 – 76.0	33.8	0.5 – 69.4	3.5	0.6 – 9.6
Tst 114	34.7	8.0 – 79.7	25.6	0.5 – 100.9	3.5	0.9 – 7.5

Statistical analyses of MC lengths were carried out for all taxa, excluding *T. simiae* TV008. Species-specific differences in MC lengths were statistically significant at the $p<0.05$ level ($F_{7, 663} = 79.48$, $P = <2e\text{-}16$; Appendix B, Table A2). Post hoc comparisons indicated that *T. godfreyi* and *T. congolense* Kilifi MC lengths were not significantly different from each other at the $p<0.05$ level but they

were significantly different from all other species, meaning that these two species can certainly be classed as having larger MCs than all other species within the subgenus. *T.b. brucei* and *T. congolense* Forest and Savannah MCs were not significantly different from each other in length but there were some significant differences between these species and *T. simiae* and *T. simiae* Tsavo, which were also not significantly different from each other in terms of MC length (Appendix B, Table A3).

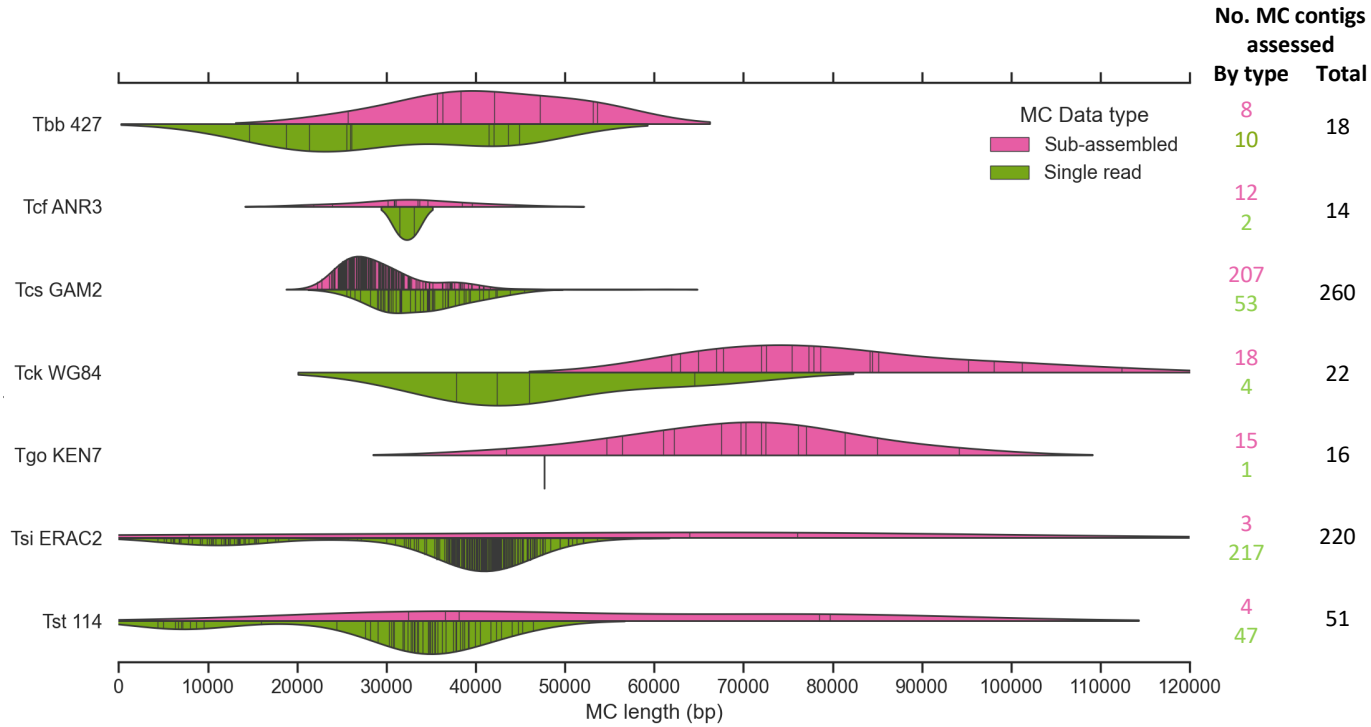

Figure 8 Species-dependent distributions of minichromosome length. Split violin plots encompass both sub-assembled (pink) and single read (green) whole minichromosome data. The sub-assembly protocol was ineffective for *T. simiae* and *T. simiae* Tsavo, despite the fact that minichromosomes were clearly present in the read pools for these species. Species abbreviations are: Tbb = *T. brucei brucei*; Tcf = *T. congolense* Forest; Tcs = *T. congolense* Savannah; Tck = *T. congolense* Kilifi; Tgo = *T. godfreyi*; Tsi = *T. simiae*; Tst = *T. simiae* Tsavo.

3.3.4 Core repeat region length variation

Having established that there are substantial length differences in the MC from different taxa, the next question was to what extent the length of core repeat regions contributes to these size differences. Single read minichromosomes and complete and partial (H+; Figure 5) were included in this analysis. *T. simiae* ERA C2 MCs had the largest core repeat region with a median of 33.8 kbp (Table 5), while the smallest core repeat regions were those belonging to *T. congolense* Forest ANR3, with a median length of 10.3 kbp (Table 5; Figure 9).

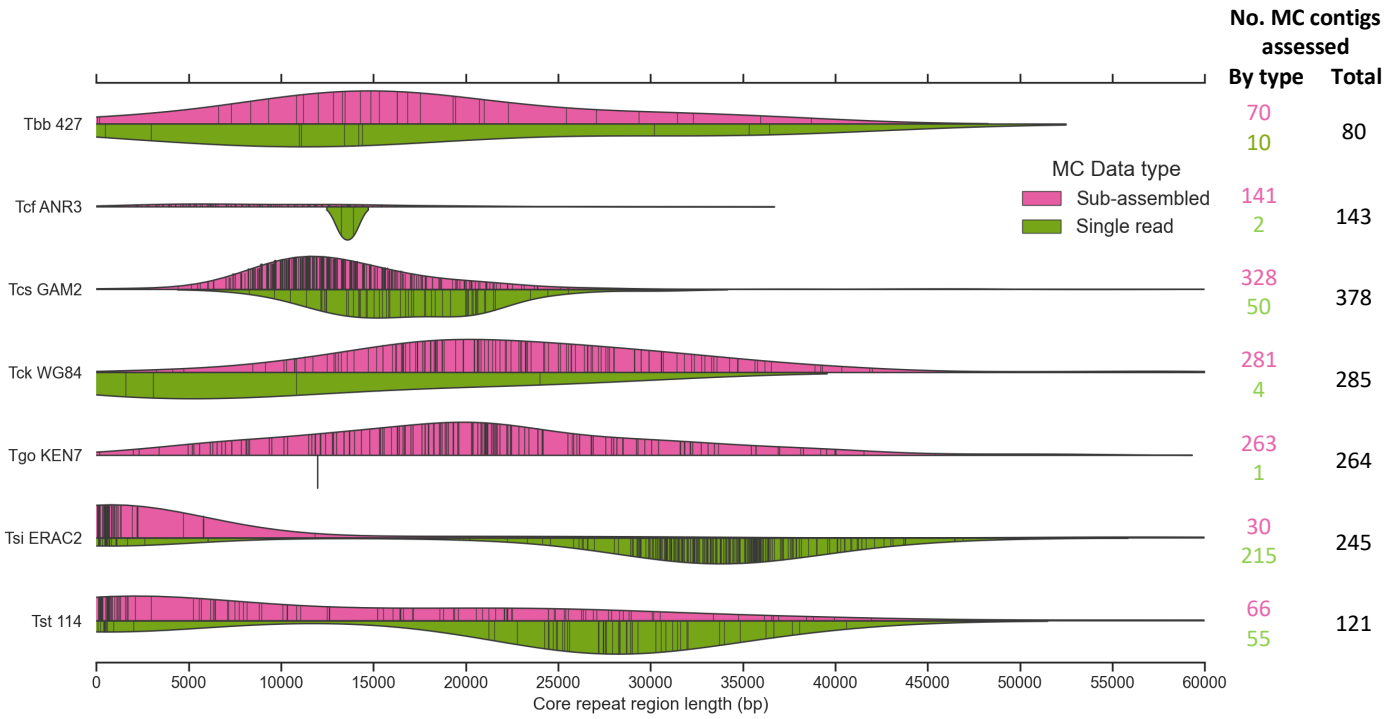

Figure 9 Species-dependent distributions of core repeat region length. Split violin plots encompass both sub-assembled (pink) and single read (green) whole minichromosome data. Species abbreviations are: Tbb = *T. brucei brucei*; Tcf = *T. congolense* Forest; Tcs = *T. congolense* Savannah; Tck = *T. congolense* Kilifi; Tgo = *T. godfreyi*; Tsi = *T. simiae*; Tst = *T. simiae* Tsavo.

Some species- and sub-species-specific differences in core repeat region lengths were statistically significant at the p<0.05 level ($F_{7, 1423}$= 158.6, *P* = <2e-16; Appendix C, Table A4). Post hoc tests revealed that these differences existed between all three *T. congolense* sub-types (Appendix C, Table A5), however findings for *T. congolense* Forest were limited due to the previously discussed limitations with the sequence data. The mean core repeat region length for *T. simiae* Tsavo was significantly different to that of *T. simiae* ($F_{7, 1423}$= 158.6, *P*= 0.000) but not significantly different to that of *T. congolense* Kilifi ($F_{7, 1423}$= 158.6, *P*= 0.935). These differences were not consistent with those uncovered by the pairwise comparisons of MC length for these species, in which *T. simiae* and *T. simiae* Tsavo had no significant differences. This suggests that there is no general evolutionary relationship between MC length and core repeat region length across all species, but does not determine whether such a relationship exists within each species.

A limited number of assembled contigs containing the entire core repeat region existed for *T. simiae* and *T. simiae* Tsavo compared to the other species within subgenus *Nannomonas*. The repeat region-containing contigs that were assembled for *T. simiae* and *T. simiae* Tsavo did not reflect the

distributions in repeat region size reflected in the single read MC data (Figure 7). The fact that these species have the greatest average core repeat region lengths (Table 5) may explain the difficulty in assembling these contigs. There may be a limit in the length of the repeat region that can be assembled by the sub-assembly pipeline, however many contigs containing core repeat regions >30 kbp were assembled for *T. congolense* Kilifi and *T. godfreyi.* Another potential issue with repeat region sub-assembly in *T. simiae* and *T. simiae* Tsavo is that the lengths of the core repeat units identified were ~175 bp longer than those identified for all other species and sub-groups within subgenus *Nannomonas* (Table 7), so it may be that the length of the repeat unit itself determines the success of core repeat region assembly.

For each species tested, there was a significant positive correlation between total MC length and core repeat region length at the p <0.05 significance level (Table 6, Figure 10). This relationship was strong for *T. simiae* and *T. simiae* Tsavo (r^2 = ~0.9) and moderate for the rest of the species within subgenus *Nannomonas* (r^2 = ~0.5-0.7). The relationship was weaker for *T.b. brucei* (r^2 = 0.3). These results prove that MC length and core repeat region length are related at the within-species level. However, without clear evidence of the mechanism of MC genesis within the nucleus, it is not possible to conclude that core repeat region length is a causal determinant of total MC length.

Table 6 Pearson's correlation values for MC length vs. repeat region length. Significant results (p <0.05) in bold text. For all species tested, there was a significant relationship between MC length and repeat region length.

Species	MC length vs repeat region length		
	r	r^2	p-value
Tbb 427	0.508962	0.259042	**0.030995**
Tcf ANR3	0.844602	0.713353	**0.000144**
Tcs GAM2	0.826108	0.682454	**1.823790e-66**
Tck WG84	0.835873	0.698683	**0.000001**
Tgo KEN7	0.720486	0.5191	**0.001642**
Tsi ERAC2	0.983001	0.966291	**6.826878e-195**
Tst 114	0.955545	0.913067	**4.696260e-34**

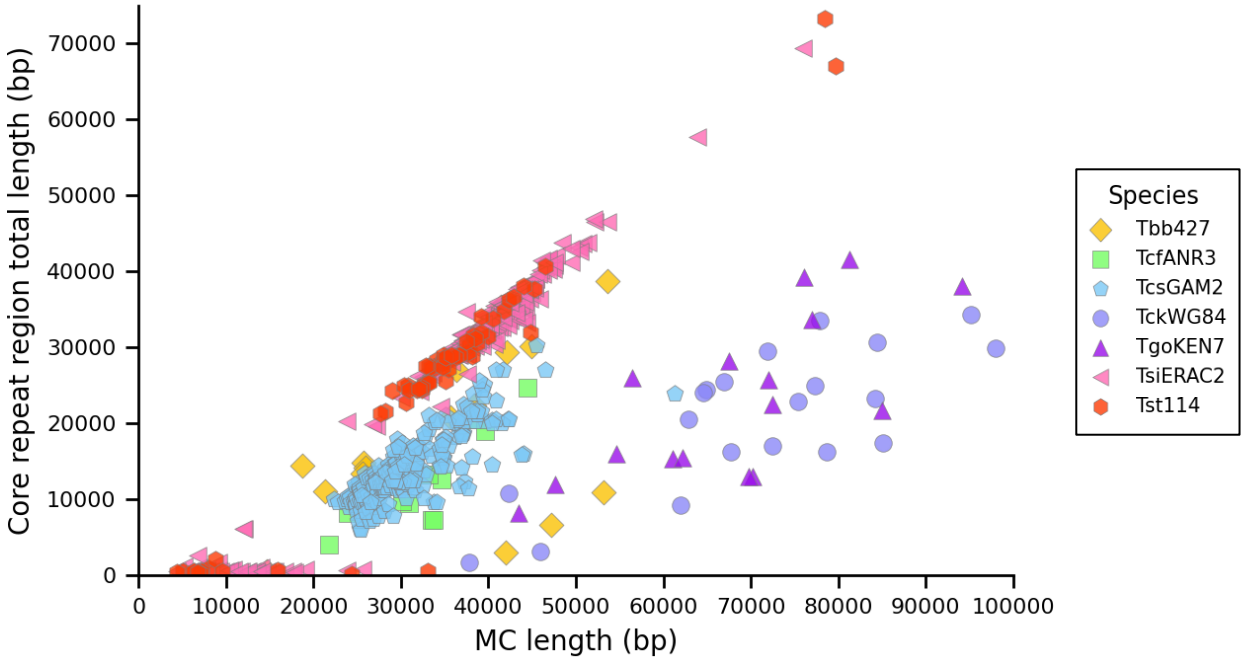

Figure 10 Correlations between MC length and core repeat region length for *T. brucei* and all species amongst subgenus *Nannomonas.* All correlations are significantly positive, with the strongest being in the *T. simiae* isolates and weaker for *T. congolense* Kilifi and *T. godfreyi.* Tbb = *T. brucei brucei*; Tcf = *T. congolense* Forest; Tcs = *T. congolense* Savannah; Tck = *T. congolense* Kilifi; Tgo = *T. godfreyi*; Tsi = *T. simiae*; Tst = *T. simiae* Tsavo.

3.3.5 Core repeat region composition and structure

Considering that the core repeat sequences used to filter 'core repeat reads' out of the read pools were discovered *de novo* in this study, it was necessary to compare these sequences with the published satellite repeat sequence for each species known to be present in MCs from PFGE hybridization studies (Sloof et al., 1983; Gibson et al., 1987; Gibson et al., 1988; Majiwa et al., 1987; Masiga et al., 1992; Majiwa et al., 1993; Masiga et al., 1996). For all species within subgenus *Nannomonas*, the core

Table 7 Comparison of MC repeat unit sequences from this study with published sequences. Sequence alignments revealed no significant differences between the existing and novel sequences, however, the lengths of MC repeat units were different for some species in the present analysis.
* There was no existing MC repeat sequence for Tst114 – the comparison for this isolate is of one of the 19 bp primers developed as a species identification probe with the MC repeat unit discovered in this study.

	Repeat unit size (bp)		GC content (%)		Sequence similarity	
Species	**Species-specific PCR primers**	**This study**	**Published results**	**This study**	**E-value**	**% identity**
Tbb427	177 (Sloof et al., 1983)	177	29 (Sloof et al., 1983)	29	2e-70	94.92
TcoANR3	350 (Gibson et al., 1988)	359	35 (Masiga et al., 1992)	35	0.0	97.76
TcoGAM2	369 (Gibson et al., 1987)	369	33 (Gibson et al., 1987)	31	4e-170	96.48
TcoWG84	400 (Gibson et al., 1988)	370	35 (Masiga et al., 1992)	33	0.0	99.46
TgoKEN7	373 (Masiga et al., 1996)	175	55 (Masiga et al., 1996)	60	2e-142	95.21
TsiERAC2	521-550 (Masiga et al., 1992; Majiwa et al., 1987)	544	55 (Masiga et al., 1992)	60	1e-05	95.45
Tst114	600 (Majiwa et al., 1993)	543	Not measured	62	7e-05*	89.47%*

repeat sequences identified in this study shared at least 95% sequence identity with species identification probes, with the exception of *T. simiae* Tsavo for which the complete MC satellite sequence had not previously been published. For this sample, a 19 bp primer sequence that is used for species identification was compared with the core repeat unit characterised in the present study. The lengths of sequences presented here were the same as the species identification probes for *T.b. brucei and T. congolense* Savannah but different for all other taxa (Table 7), with most being only slightly refined. All repeat unit sequences for *T. congolense* possessed an abundance of A and T homopolymer tracts, whereas *T. godfreyi, T. simiae* and *T. simiae* Tsavo had a greater proportion of G and C tracts. Masiga et al. (1996) described the satellite repeat sequence for *T. godfreyi* as a 373 bp sequence consisting of two ~170 bp imperfect internal repeats with a 25 bp sequence in between them. In the present study, only one 175 bp unit was identified, which was tandemly repeated throughout the core repeat region.

Several *T. congolense* Savannah, *T. simiae* and *T. simiae* Tsavo minichromosomes adhered to the existing *T.b. brucei* model of core repeat region structure (Wickstead et al., 2004), with the exception that the inversion point of the core repeat region was slightly off centre rather than completely central (Figure 7). However, this was by no means the consensus structure within these species or any others. A range of MC structures existed both within and between species (Figure 11).

The majority of *T.b brucei's* MCs had an inverted repeat region that met the criteria of the palindromic MC model described by Wickstead et al (2004) but some had off-centre inversion points or arms that were asymmetrical in length. All three *T. congolense* subtypes and *T. godfreyi* had no or a few (0-6.25%) inverted repeat region containing MCs, with the majority of them having core repeat regions in which the core repeat unit ran consistently in one direction. For *T. simiae* and *T. simiae* Tsavo, all complete MCs contained inverted core repeat sequences. Most of these contained one inversion point that was either central or off centre and some had multiple inversion points (Figure 11).

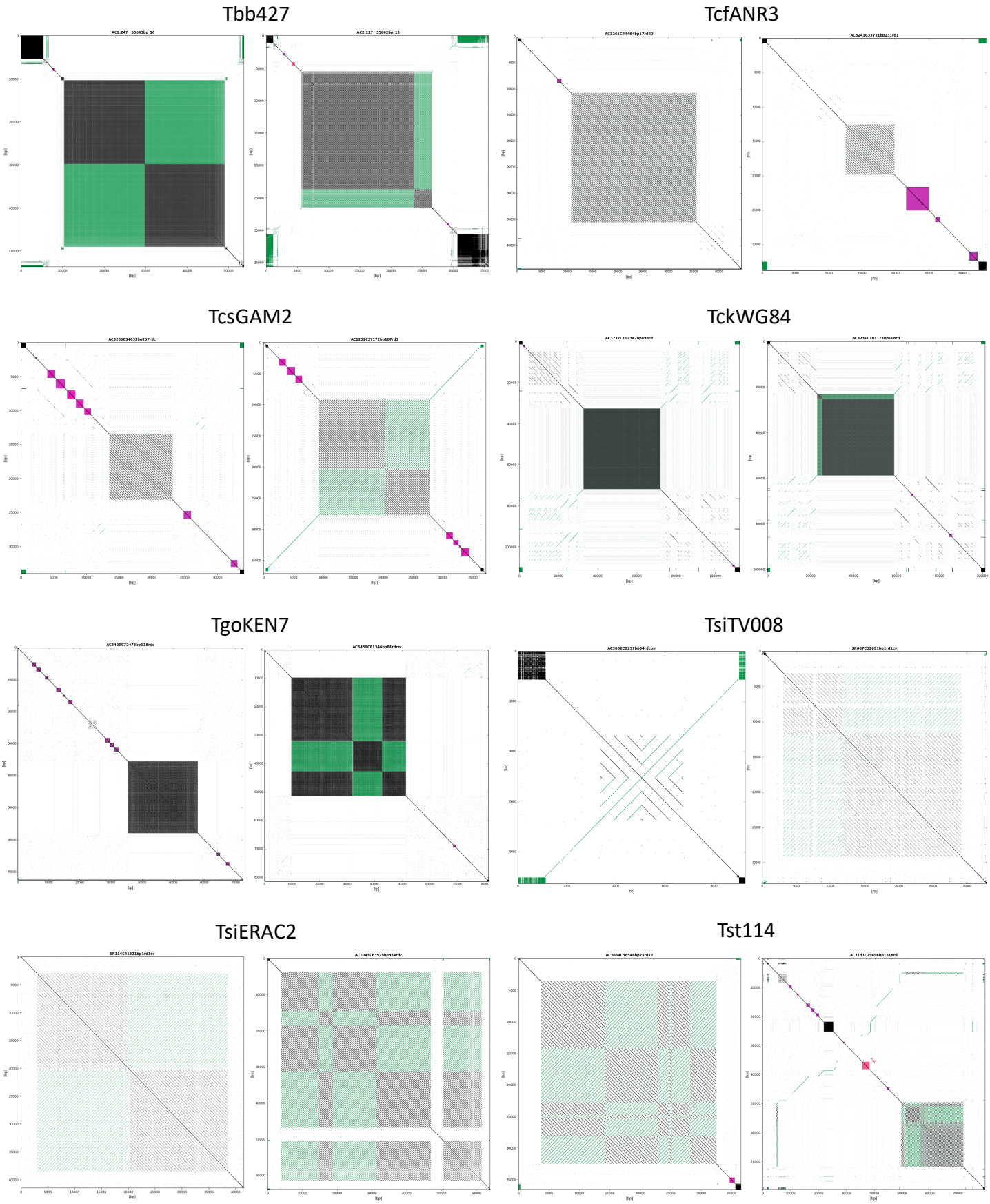

Figure 11 Self dot plots of two representative minichromosomes for *T. brucei* and all species within subgenus *Nannomonas.* There is great diversity in structure within and between each species, with one characteristic structure being predominant for each species (plot on the left for each species). Pink = annotated regions

3.3.6 Sub-telomeric arm length variation

MC arm lengths were the longest for *T. congolense* Kilifi and *T. godfreyi* (median = 25.0-30.0 kbp) and shortest for *T. simiae* and *T. simiae* Tsavo (median = ~3.5 kbp). Arm length differences were statistically significant at the p< 0.05 level ($F_{7,2217}$ = 1198, *P* = <2e-16; Appendix D, Table A6) and post hoc testing (Appendix D, Table A7) revealed that the only pairwise comparisons with no significant arm length differences within subgenus *Nannomonas* were those between *T. simiae* and *T. simiae* Tsavo ($F_{7,2217}$ = 1198, *P* = 0.99). The difference between *T. congolense* Kilifi and *T. godfreyi* mean arm lengths was significant ($F_{7,2217}$ = 1198, *P* = 0.00) despite both species having significantly longer MCs than all other species, as was the difference between *T. congolense* Forest and *T. congolense* Savannah mean arm length ($F_{7,2217}$ = 1198, *P* = 0.02). It is interesting to note that the MCs of most species within subgenus *Nannomonas* had narrow ranges in arm length variation, with standard deviations of ~1.0-2.5 kbp (Table 5), with the exceptions of *T. godfreyi* and *T. congolense* Kilifi, which had a relatively even distribution of a wide range of arm lengths (Figure 12), with standard deviations of ~ >9.0 kbp (Table 5). The consistent short measurements of *T. simiae* and *T. simiae* Tsavo arm lengths (Figure 12) is also

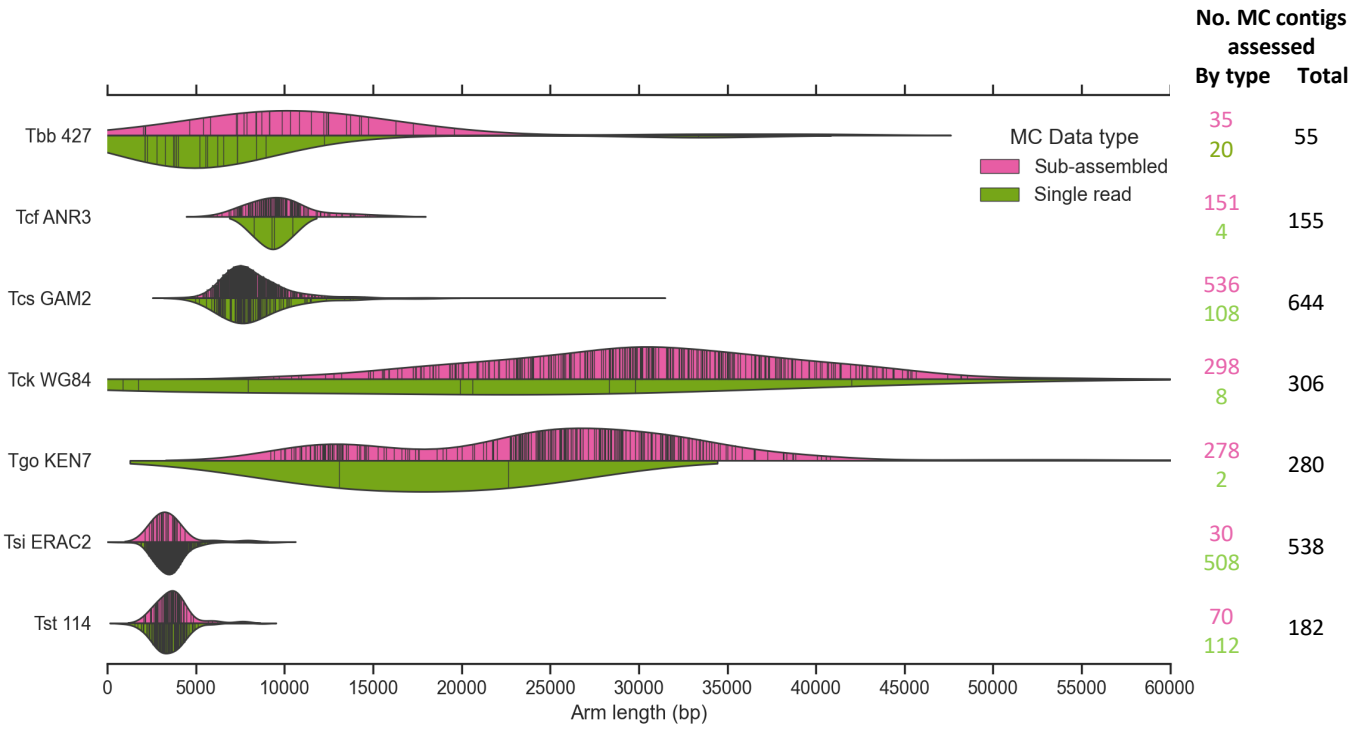

Figure 12 Species-dependent distributions of minichromosome 'arm' length for *T. brucei* and all species within subgenus *Nannomonas*. Split violin plots encompass both sub-assembled (pink) and single read (green) whole minichromosome data. Whole MC contigs were assessed twice so that each arm could be included in the analysis, hence the high total number of contigs for several groups. *Tbb* = T. brucei brucei; *Tcf* = T. congolense *Forest; Tcs* = T. congolense *Savannah; Tck* = T. congolense *Kilifi; Tgo* = T. godfreyi; *Tsi* = T. simiae; *Tst* = T. simiae *Tsavo.*

interesting considering the large range of variation that exists in the lengths of these species' core repeat regions. Their short arm lengths may also be a factor that limits assembly success, combined with the fact that the core repeat regions are extremely large in comparison to the arms.

The only species for which there was a significant correlation between arm length and core repeat region length was *T.b. brucei* (r_{55} = -0.34, p <0.05) (Table 8). The relationship between these two variables was negative, meaning that the longer the core repeat region was, the shorter the mini-chromosome arms were. This was however a weak relationship, with a coefficient of determination of 0.11. There was no significant correlation between arm length and core repeat region length for all species within subgenus *Nannomonas* (Figure 13).

Table 8 Pearson's correlation values for MC arm length vs. repeat region length, with significant results (p <0.05) in bold text.

Species	Arm length vs repeat region length		
	r	r^2	p-value
Tbb 427	-0.336201	0.113031	**0.012087**
Tcf ANR3	0.091915	0.008448	0.420447
Tcs GAM2	-0.006826	0.000047	0.866287
Tck WG84	0.08958	0.008025	0.40111
Tgo KEN7	-0.175597	0.030834	0.110111
Tsi ERAC2	-0.051344	0.002636	0.278172
Tst 114	0.128205	0.016437	0.150875

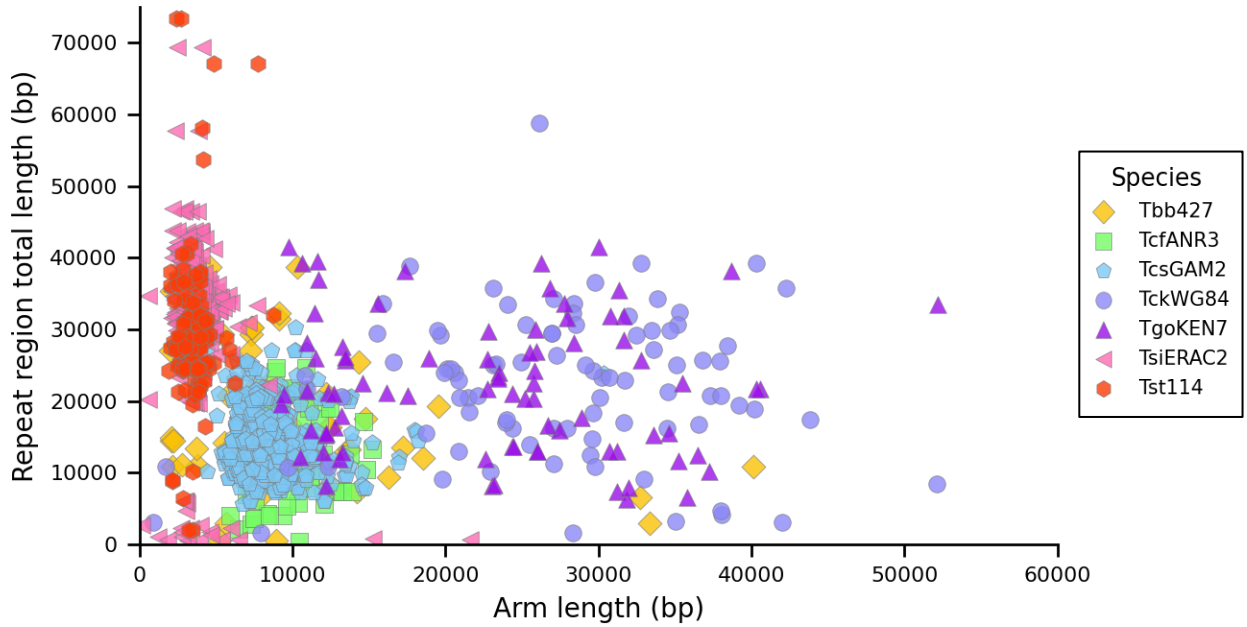

Figure 13 Correlations between MC arm length and repeat region length for *T. brucei* and all species amongst subgenus *Nannomonas.* Tbb = *T. brucei brucei*; Tcf = *T. congolense* Forest; Tcs = *T. congolense* Savannah; Tck = *T. congolense* Kilifi; Tgo = *T. godfreyi*; Tsi = *T. simiae*; Tst = *T. simiae* Tsavo.

Whole MCs belonging to *T. congolense* Forest had a significant positive correlation (r_{14} = 0.57, p<0.05) between the lengths of each arm per MC (Table 9, Figure 14), suggesting that for this species, individual MCs mostly contain arms that are relatively equal in length to each other. In contrast, *T. godfreyi* displayed a significant negative correlation between the two arm lengths per MC (r_{14} = -0.53 , p <0.05), suggesting that the majority of MCs for this species contain arms of unequal length. There were no significant relationships between the arm lengths of MCs for the rest of the species tested, which means that for these species there is no set pattern in the way that within-MC arm lengths relate to each other and that there may be a range of symmetrical and asymmetrical MCs within the populations.

Table 9 Pearson's correlation values for arm 1 vs. arm 2 length (within each MC contig) with significant results (p <0.05) in bold text.

Species	Arm 1 length vs arm 2 length		
	r	r^2	p-value
Tbb 427	-0.222257	0.049398	0.375391
Tcf ANR3	0.574356	0.329885	**0.031703**
Tcs GAM2	-0.010999	0.000121	0.859634
Tck WG84	-0.186891	0.034928	0.417259
Tgo KEN7	-0.533814	0.284957	**0.033192**
Tsi ERAC2	0.051956	0.002699	0.412451
Tst 114	0.077503	0.006007	0.563083

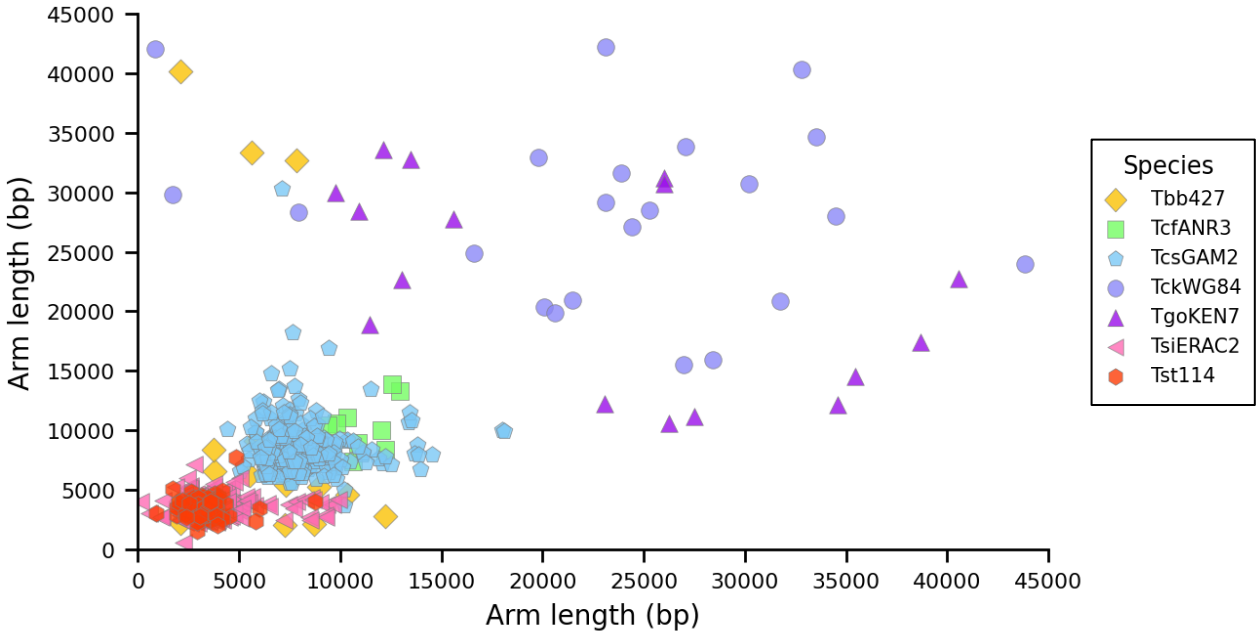

Figure 14 Correlations between the lengths of the two arms within each complete MC contig. Tbb = *T. brucei brucei*; Tcf = *T. congolense* Forest; Tcs = *T. congolense* Savannah; Tck = *T. congolense* Kilifi; Tgo = *T. godfreyi*; Tsi = *T. simiae*; Tst = *T. simiae* Tsavo.

3.3.7 Gene content of subtelomeric arms

Whilst MCs have traditionally been viewed simply as repositories of VSG genes (Van der Ploeg et al., 1984), recent analysis of MC sequences from the PacBio data of two *T. congolense* isolates revealed the presence of other protein coding genes, such as *DEAH-box RNA helicase* and the retrotransposon *Ingi* (Abbas et al., 2018). No data exist on the MC gene content of *T. godfreyi* and *T. simiae*. It was therefore of interest in the present study to test the hypothesis that MCs are VSG gene repositories, to define the gene content of the subtelomeric arms of all MCs recovered from the genomic datasets and to draw comparisons between the species.

T. congolense Savannah, *T. congolense* Kilifi and *T. godfreyi* yielded the greatest numbers of predicted genes (Table 10), probably due to a combination of the quality of the sequence data for these species relative to *T. congolense* Forest and the high number of MC contigs assembled relative to that of *T. simiae* ERA C2 and *T. simiae* Tsavo 114. Few annotated features were identified for the two *T. simiae* isolates. The MCs that were annotated for this species were mostly those captured within single long reads and so were likely to have a greater error rate than assembled contigs, decreasing the possibility of accurate annotation transfer and open reading frame (ORF) prediction by the Companion pipeline.

Table 10 Genes present on the MCs of six groups of trypanosomes from subgenus *Nannomonas* and one from subgenus *Trypanozoon* based on Companion annotations. VSGs were the most abundant genes identified for all species. Pseudogenes are those that possessed internal stop codons and/or significant frameshift mutations. These were all *Ingi* elements, except in *T.b brucei,* in which there were two RHSP and two ESAG pseudogenes. VSG = variable surface glycoprotein, HP = hypothetical protein, ABC = ABC transporter, C2H2 = zinc finger 2, PAG4 = procyclin associated gene 4, CPC = cysteine peptidase, DEAH = DEAH-box helicase, RHSP = retrotransposon hotspot protein, Pseudo = pseudogenes.

Species	*VSG*	*HP*	*ESAG*	*ABC*	*C2H2*	*PAG4*	*CPC*	*DEAH*	*Ingi*	*RHSP*	Total	*Pseudo*
Tbb427	13	35	5	0	0	0	0	0	7	2	62	7
TcoANR3	35	18	1	0	0	0	0	5	8	3	70	0
TcoGAM2	306	120	115	0	1	6	2	22	45	2	619	0
TcoWG84	263	75	25	0	5	0	0	0	1	0	369	0
TgoKEN7	600	211	33	135	0	0	0	0	40	0	1,019	7
TsiERAC2	19	1	2	0	0	0	1	0	2	0	25	1
Tst114	22	15	0	0	0	0	0	0	1	0	38	1

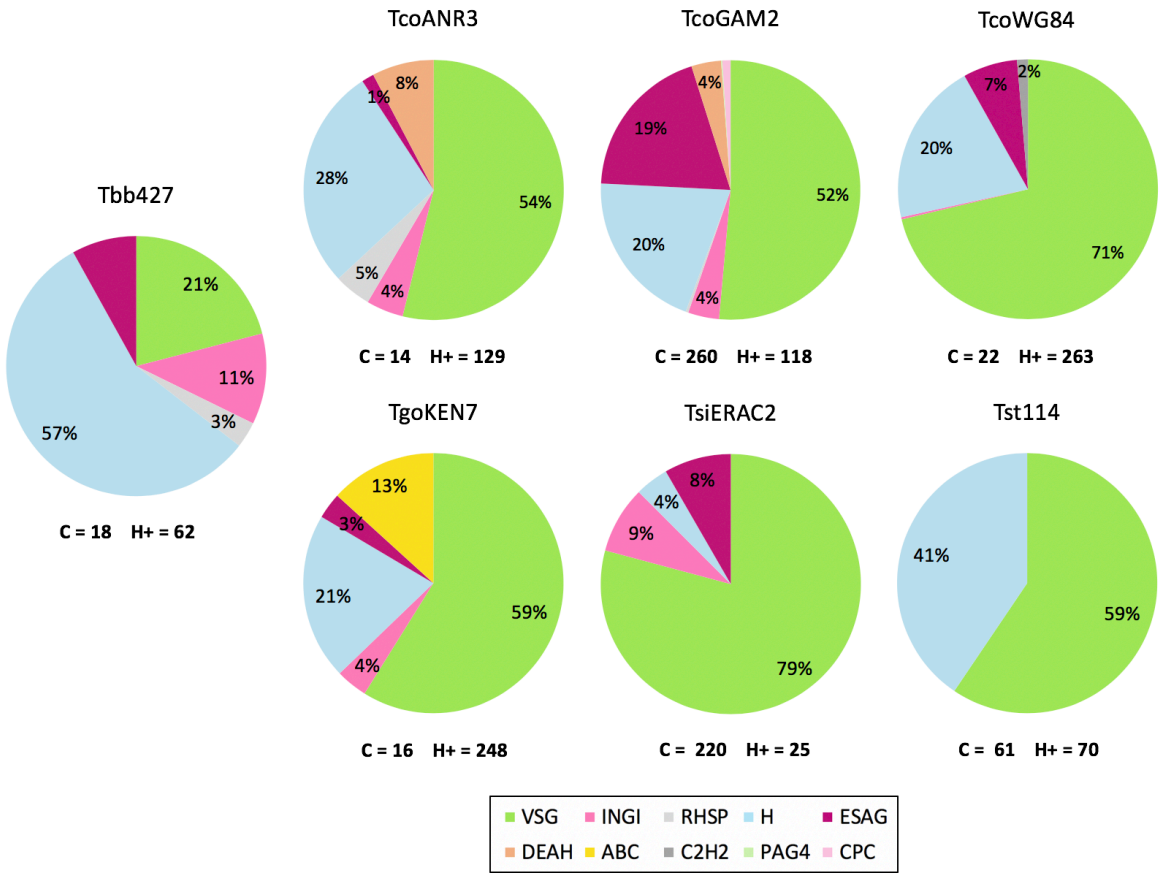

Figure 15 Relative percentages of predicted proteins in all species. Labels under charts: C = whole MC contig, H+ = more than half partial MC contig (Figure 5). Legend labels: VSG = variable surface glycoprotein, RHSP = retrotransposon hotspot protein, H = hypothetical protein, ESAG = expression site associated gene, DEAH = DEAH-box helicase, ABC = ABC transport protein, C2H2 = zinc finger, PAG4 = procyclin associated gene 4, CPC, cysteine peptidase C.

For all species within subgenus *Nannomonas,* the most abundant genes were VSG-encoding genes, which represented at least 50% of the total number of genes identified (Figure 11). For some species, previously uncharacterised ORFs labelled as 'hypothetical proteins' by Companion were designated as *VSG*s based on clustering analysis. Approximately 1-2 VSGs overall were identified per MC for all *T. congolense* subspecies, however it is likely that these numbers are under representative of the true number of *VSGs* per MC as the calculation included single read MC contigs for which minimal annotations were generated due to the error rate of these sequences. Visual inspection of dot plots revealed that many *T. congolense* Savannah GAM2 MC contigs each contained up to four *VSGs.* A significant limitation of the present annotations is that they were derived only from the *T.b. brucei* TREU 927 and *T. congolense* IL3000 proteomic databases. To gain a more accurate idea of the number and nature of all MC coding regions and pseudogenes, it would be optimal to annotate the MCs using

the methods presented here in conjunction with the annotated major chromosome data for each species and subgroup when it becomes available, especially considering that MCs are thought to be born from the major chromosomes (Wickstead et al., 2004). Hidden Markov Models (HMMs) constructed using the *VSG*s identified here, as well as any that may be newly characterised on the major chromosomes of each species, may reveal a more representative number of the true number of *VSG*s contained within the MC populations. Despite these limitations, the number of *VSG*s identified on *T. godfreyi* MCs was notably high compared to other isolate, averaging at approximately four per MC. This means that if VSG gene number is conserved across trypanosome species, based on previous predictions of approximately 2,500 VSG genes per whole *T. congolense* IL3000 genome (Jackson et al., 2012; Cross et al., 2014) 24% of all genomic VSG genes are located on MCs in *T. godfreyi*.

In addition to VSG genes, several other coding regions were identified on the MCs of all species. *Ingi* elements were identified within the subtelomeric regions of a small portion of MCs for all species, being more abundant in *T. congolense* Savannah and *T. godfreyi* compared to the other isolates (Table 10). *Ingi* is a diverse clade of non-long terminal repeat retrotransposons, which is present throughout trypanosome genomes (Bringaud et al., 2009), being most abundant in subtelomeric regions in *T.b brucei* (Bringaud et al., 2004). Full length *Ingi* sequences can encode functional transcription machinery proteins (Bringaud et al., 2009) that result in coding strand switches (Berriman et al., 2005). In this way, *Ingi* has initiated genomic rearrangements in *T. brucei,* such as the insertion of an ESAG on major chromosome 11 of *T.b gambiense* relative to *T.b brucei* (Jackson et al., 2010). *Ingi* has previously been identified in a small number of MCs (Alsford et al., 2000; Abbas et al., 2018). A few Ingi genes identified in the present analysis were classed as pseudogenes due to the presence of internal stop codons (Table 10). It is an appealing possibility that MC-specific *Ingi* elements may be involved in the rearrangement of MCs and possibly in their recombination with the other chromosomal classes, though this may be unlikely considering *ingi* was present on such a limited number of MC contigs. If these sequences were essential to the maintenance of the biological function of MCs, it would be expected that they would be present on all MCs. It is possible that these sequences have hitch-hiked onto the MCs with VSGs from subtelomeric regions of the major chromosomes during VSG

transposition and remain transcriptionally inactive alongside them. These speculations may be proven by the comparative analysis of these sequences with *Ingi* sequences present on the subtelomeric regions of the major chromosomes.

Another common feature for several species was the presence of *ESAG*s on a small number of MCs. *ESAG3* was present on *T.b brucei* MCs, whilst *ESAG2* was present on the MCs of species amongst subgenus *Nannomonas.* In *T. congolense* Savannah GAM2, *ESAG2* sequences were located adjacent to VSG genes, with eight MCs displaying tandem arrays of an *ESAG2* with a *VSG* on either side (Figure 16). In *T. brucei,* ESAG2 is a VSG-like surface protein without known antigenic function (Gadelha et al., 2015). Whilst not much is known regarding *T. congolense* expression sites and the associated genes, phylogenetic studies have characterised *T. congolense ESAG2* as a b-VSG-like gene belonging to Fam13, of which some members still encode functional variant antigens in this species, unlike in *T. brucei* (Jackson et al., 2012). Further analysis of the *ESAG2* sequences presented here for all subgenus *Nannomonas* species is required to determine whether these sequences should be classed as additional VSGs or not. It was recently discovered that *T. congolense* genomes possess very

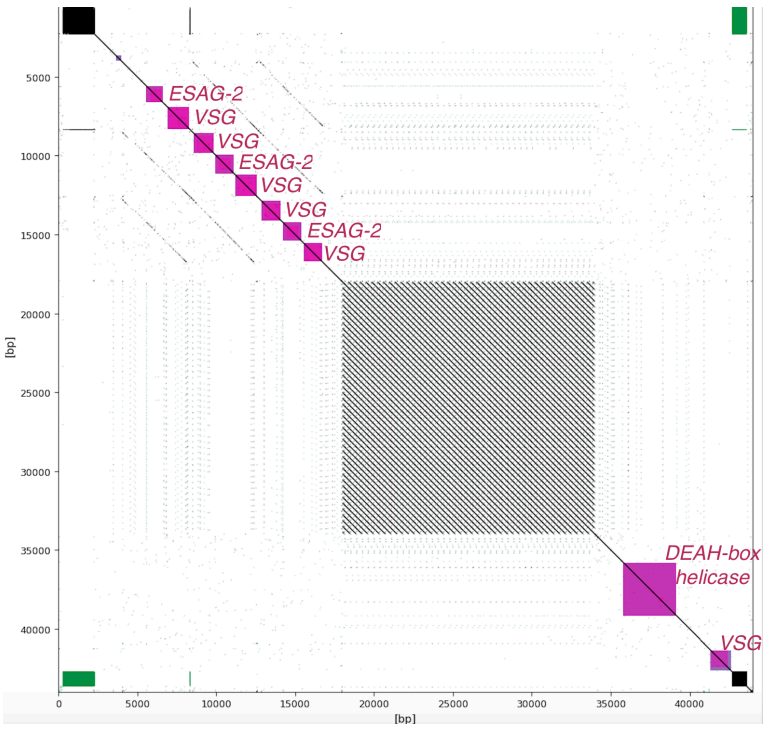

Figure 16 A *T. congolense* Savannah MC with gene annotations. This MC contains an ESAG2/VSG gene array on the subtelomeric region of one arm and a DEAH-box helicase gene located halfway between the core repeat region and a telomere-proximal VSG on the second arm.

few *ESAG*s relative to *T. brucei* and those that do exist are members of gene families that are widespread rather than associated explicitly with telomeric regions and expression sites (Abbas et al., 2018). *ESAG2* is therefore not an MC-specific gene but its relatively common presence on MCs, especially in *T. congolense* Savannah, is curious and remains to be explained.

DEAH-box RNA helicase genes that shared 70-75% sequence identity with a sequence on *T. congolense* IL3000 (TCIL3000_6_290) chromosome 6 (Jackson et al., 2012) were present on a limited number of *T. congolense* Savannah and Forest MCs (Figure 16). DEAH-box helicases are ATP-dependent enzymes that catalyse RNA unfolding in trypanosomes (Zinoviev et al., 2012) and those present in *T. congolense* represent an expansion of single copy genes present in other trypanosomes (Abbas et al., 2018). If *T. congolense* MCs do contain active VSG expression sites as suggested by Abbas et al. (2018), it may be that DEAH-box helicase genes are transcribed and that their products are involved in the unfolding of expressed VSG mRNA. Similarly to the situation with *Ingi* sequences, this functional proposition may be unlikely considering that these sequences were not conserved across all MCs, which would be expected if they were of functional importance.

Other protein-coding genes that were infrequently present on the MCs of certain species were: *retrotransposon hotspot protein (RHSP),* which is known to provide a hotspot for *Ingi* insertion in *T. brucei* (Florini et al., 2018); the *zinc finger (C2H2)* transcription factor (Fedotova et al., 2017)*; cysteine peptidase C (CPC)* and *procyclin-associated gene (PAG4)* (Table 10; Figure 15). These genes were present in such limited numbers (Table 10) that they could not be classed as conserved features of MCs as VSGs and possibly ESAGs could for all species.

A totally unprecedented finding was the existence of a significant percentage (13% of all coding sequences identified) of possible ABC transporter genes on *T. godfreyi* MCs, which were present on 42% (111) of all MC contigs analysed. These sequences were not identified on the MCs of any other species and they shared 25-35% sequence identity with a putative *ABC transporter/ATP-binding cassette protein* on *T. congolense* IL3000 chromosome 10 (TCIL3000_10_9180). These genes were located

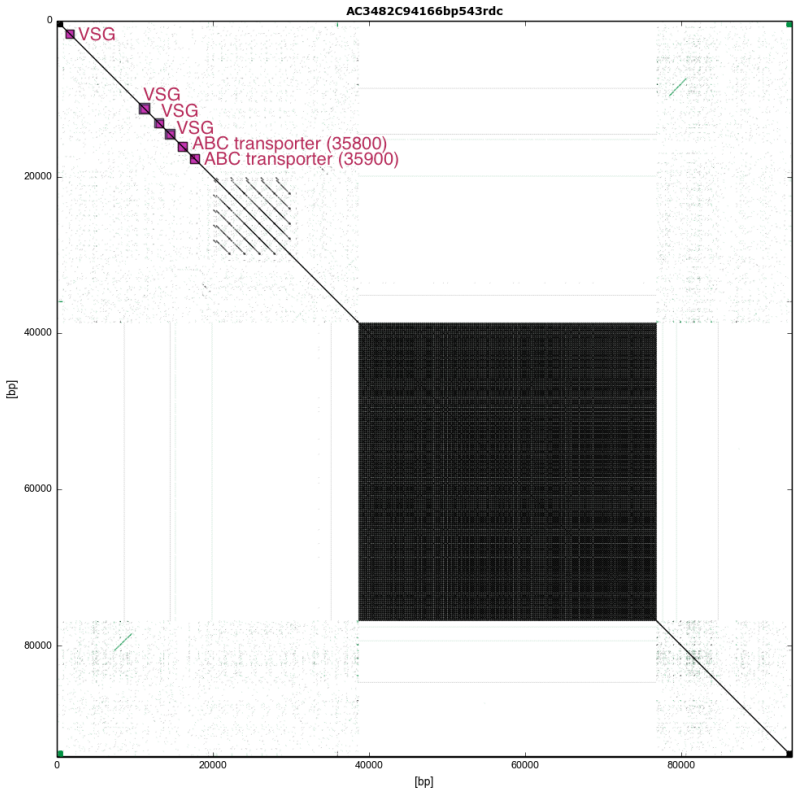

Figure 17 A *T. godfreyi* minichromosome containing ABC transporter genes, which are located between subtelomeric VSG genes and a 10 kbp repetitive region located downstream from the main core repeat region.

between subtelomeric VSGs and the core repeat region (Figure 17). On complete MC contigs, ABC transporter genes existed only on the longer arm of asymmetrical MCs, however this could not be concluded for the partial contigs that contained only one arm. Positioned between the ABC transporter genes and the core repeat regions were additional repeat sequence regions of either 5 – 10 kbp in length with a ~2 kbp repeat unit or 2 – 3 kbp in length with a 415 bp repeat unit. When multiple ABC transporter genes were present on one MC arm, they were located adjacent to each other and were not from the same cluster (<90% identity) according to CD-HIT, suggesting that they did not duplicate in-situ on the MC and so are likely to have originated from elsewhere in the genome. ABC transporter genes from the same cluster did exist on different MCs, which may indicate recombination events between these MCs. ABC transporters/ATP cassette binding proteins are involved in various survival mechanisms for trypanosomes, including glycolysis in *T. brucei* (Yernaux et al., 2005), haemoglobin uptake (Yamasaki et al., 2016) and resistance against trypanocidal drugs in *T. cruzi* (Franco et al., 2015). Though the ABC transporter genes identified on *T. godfreyi* MCs were labelled as such by Companion, this does not necessarily mean they encode functional ABC transporters, especially

considering the relatively low (25-35%) sequence identity the sequences shared with that on *T. congolense* IL3000 chromosome 10. *ABC transporters* labelled in both the *T.b. brucei* Lister 427 and *T. congolense* IL3000 genome assemblies have not been characterised in detail, with most having been labelled as 'putative'. Considering the high prevalence of potential *ABC transporter* sequences on *T. godfreyi* MCs compared to those of other species within subgenus *Nannomonas*, it would be worth comparing these with *ABC transporter* sequences present on the *T. godfreyi* major chromosomes and indeed on those of the other species tested. The abundance of apparent *ABC transporters* on the *T. godfreyi* MCs may be a signpost of a whole genome expansion of these gene families within *T. godfreyi* compared to *T. congolense* and *T. simiae.* Thus, a full genomic and proteomic investigation of ABC transporters across the subgenus may be worthwhile, especially considering the implication of these genes being involved in drug resistance (Baker et al., 2013; Franco et al., 2015)

Within each species, there were instances in which different MCs contained annotated regions with 100% shared identity as determined by clustering analysis, which suggests that recombination occurred between them or that they both underwent recombination with the same region of a major chromosome. To draw definitive conclusions regarding recombination events between MCs, intermediate and major chromosomes, complete and annotated genomic data of all of these sequences is required.

For all species, manual inspection of MC contigs revealed that a few of the largest assembled "H+" contigs appeared to represent the telomeric ends of some major chromosomes. These contigs contained many housekeeping genes that closely matched those present on the major chromosomes of *T. congolense* IL3000 (Jackson et al., 2012), as well as very small stretches (1-2 copies) of the MC core repeat sequence. Though these contigs were discarded as they appeared to be larger chromosome material, the existence of the MC core repeat sequence on them confirmed the possibility that this sequence is used as a recombination hotspot not only between the MCs but also between the different chromosome classes for species within subgenus *Nannomonas*, as is already known to be the case for *T.b. brucei* (Wickstead et al., 2004).

Chapter 4. General discussion

4.1 Thesis summary

This thesis ultimately represents an early stage of a complex endeavour to understand reasons for the formation and maintenance of MCs within the trypanosome genomic environment. It has been demonstrated that contrary to previous attempts, MCs are in fact amenable to computational assembly methods and this generates compelling possibilities in terms of investigating these previously neglected genomic components. A total number of 601 complete and 915 partial MC sequences were recovered from PacBio genome data of six taxa from subgenus *Nannomonas* and one from subgenus *Trypanozoon*, enabling novel insights to be uncovered regarding the species-specific nature of MCs in terms of their structure and gene content.

4.2 Insights into the evolution and genesis of MCs

Novel findings have been generated regarding the MCs of species for which MC sequence had not previously been studied beyond the knowledge of the core repeat unit sequence, including *T. godfreyi* and *T. simiae,* which opens up the possibility for evolutionary comparison between the MCs of different species. Dot plots of *T. simiae* MCs revealed that some of them possessed core repeat regions with multiple inversion points – a trait that has been associated with MCs on only one occasion before on a *T. congolense* IL3000 MC generated from single read analysis (Abbas et al., 2018). *T. congolense* and *T. godfreyi* MCs were rarely palindromic and most often had a core repeat region with the repeat unit sequence running in one direction only. These findings are significant when considering the palindromic model of the formation and replication of MCs in *T.b. brucei* proposed by Wickstead et al. (2004). In this model, researchers stated that all *T.b. brucei* MCs have a single inversion point in the core repeat region, which they implicated was involved in replication and centromeric function. The lack of an inversion point in the core repeat regions of *T. congolense* and *T. godfreyi* MCs may challenge the proposal that inversion points signify origins of replication, or it may suggest that the MCs

of these species undergo different genesis mechanisms to those of *T.b brucei.* The existing model is more likely to apply to *T. simiae* MCs as many of these were palindromic, however those that contained multiple inversion points may be subject to a more complex method of genesis. Future study is required to investigate these processes at the molecular level but it is clear that these mechanisms vary at the subgenus- or even species-level.

Clustering analyses identified the presence of conserved annotated features between different MCs within each species, which is congruent with existing evidence that MCs undergo homologous recombination with each other (Abbas et al., 2018). On some MCs there were annotated features that also displayed significant levels of sequence identity with genes present on the *T. congolense* IL3000 major chromosomes, though definitive conclusions regarding recombination between the different chromosome classes cannot be drawn without analysis of major chromosome genome data for the specific species tested in this analysis.

4.3 Conserved and divergent characteristics of MCs

In addition to divergent core repeat region structures, there were a number of statistically significant MC structure differences between the species tested, as well as differences in MC gene content. As a rule, *T. godfreyi* MCs were asymmetrical, meaning that one subtelomeric arm was a significantly different length to the other, whereas there was no specific pattern in the relative arm lengths for all other taxa within the subgenus, except for *T. congolense* Forest in which MC arms were strictly equal in length.

Several annotated features were conserved across all species, such as the presence of VSG genes, *ESAG*s and *Ingi* on MC subtelomeric arms, though *VSG*s were the only sequences present on all annotated contigs. Some species-specific differences in gene content were observed, most notably the presence of putative ABC transporter genes on almost half of the *T. godfreyi* MCs, which was a feature exclusive to this species.

4.4 Methodological limitations

Bioinformatic analyses illustrated that it is relatively simple to filter raw PacBio sequence data to uncover 'single read MCs' for many species, though not for those with MCs >50 kbp in length. Therefore, single read analysis is not a plausible method for the comparison of MCs across subgenus *Nannomonas,* which are up to 200 kbp in length in taxa such as *T. congolense* Kilifi and *T. godfreyi*. For many taxa, particularly *T. congolense* Kilifi and *T. godfreyi,* the assembly methods presented here were biased towards assembling MC at the smaller end of the size range specified by the PFGE analysis. Few complete MCs were assembled for *T. simiae* and *T. simiae* Tsavo, possibly due to the core repeat region constituting 74-85% of each MC, compared to 30-45% in *T. congolense.* In these cases, single-read MC analysis proved a useful method to fall back on, as more than two hundred MC sequences were uncovered in this way. Though single read MCs were difficult to annotate due to the inherently higher error rate of these sequences compared to MCs that have been assembled and corrected, they provided useful information on the size and structure of *T. simiae* MCs. For the reliable assembly of all MCs belonging to all species within subgenus *Nannomonas,* further improvement of the assembly pipeline so that larger and more repeat-region-rich MCs can be assembled is necessary.

Palindromy is susceptible to under-representation in genomic sequence data due to the fold-back of single stranded DNA during PCR (Wickstead et al., 2004), so it is possible that there are actually a greater number of palindromic MCs than has been presented here for species such as *T. congolense* and *T. godfreyi*. However, the fact that the majority of *T. simiae* MC contigs had inverted core repeat regions suggests that this limitation may not be such an issue for PacBio sequence data and that the relative species-specific levels of palindromy presented here are accurate.

The annotated sequences presented here are likely to be under-representative of the true number of coding regions present within the MC populations, especially for species in which representative MC contigs spanning all known MC size ranges were not generated. The annotation methods outlined in the Methods are likely to generate a greater number of findings for *T. simiae* once the assembly pipeline has been improved to yield more assembled contigs for annotation. However, to

gain a complete understanding of MC gene content at the within-species level, further annotation and characterisation of coding regions is necessary. Much of these investigations will rely on the availability of major chromosome data with de-novo annotations, which will enable the analysis of possible recombination events between the MCs and major chromosomes by clustering analysis.

Whilst the use of whole genome data for MC analysis is useful for drawing comparisons between species, it limits conclusions that can be drawn about MCs within individual trypanosomes. The MC numbers presented here for each species represent all MCs present in uncloned populations. To determine whether all trypanosomes within each isolate contain the same set of MCs or only a portion of them each, it would be necessary to analyse sequence data of several cloned trypanosomes for each species. Whilst this approach may seem costly and time consuming, it is likely to provide more detailed insight into the mechanisms and timescales of MC evolution within trypanosomes.

4.5 Future outlook

This investigation has opened up many valuable pathways for future study. MCs clearly vary in structure and content at the between-species level within subgenus *Nannomonas.* As well as studying the sequence data of cloned trypanosome MCs, it may be insightful to repeat both the PFGE and bioinformatic protocols with a wider range of isolates from each species, considering the within-species heterogeneity in MC size range that was observed in Chapter 2.

Additional insight into the origin and genesis of MCs is likely to be gained by 1) tracking the recombination history of MCs with each other and the major chromosomes and 2) comparing the relative distributions of MC protein-coding sequences across all chromosomal size classes. This may indicate whether elements such as *Ingi* and DEAH-box helicase accumulate on MCs by chance or if they serve an integral function of some kind. This is worth investigating considering the proposition by Abbas et al. (2018) that *T. congolense* MCs may possess active expression sites and the fact that several of the non-VSG coding genes identified in this study produce proteins involved in DNA rearrangement and RNA modification. Considering that MC structure often reflects the structure of the

major chromosome subtelomeric ends and some features of expression sites within these regions, the findings uncovered in this study may act as signposts towards sequences to look out for in search of the major chromosome subtelomeric regions and expression sites in genomic data for species within subgenus *Nannomonas.* For *T. godfreyi* in particular, this line of enquiry generates questions such as: On the whole genome level, does *T. godfreyi* have a greater number of VSGs and ABC transporters compared to the other species? What is the function of ABC transporters in *T. godfreyi,* how are they distributed throughout the genome and is there a reason they are found so frequently on MCs? Whilst the present study investigating MCs in isolation has provided useful insights, the next steps of analysis ought to be focused on placing them in context with the whole genome and on an evolutionary level to compare their involvement in antigenic variation mechanisms that quite possibly differ between species.

Conclusions

This study combined the classical chromosome analysis technique of PFGE with the newer technology of genome sequencing. The PFGE investigation provided detailed karyotype comparisons for *T.b. brucei* and all species within subgenus *Nannomonas*, revealing that that each species has its own MC size range and that there are within-species differences in *T. congolense* MC lengths. A semi-automated pipeline for the assembly and annotation of MCs from PacBio sequence data has been built, combining a range of widely available bioinformatic tools. Factors such as MC length, core repeat region composition and subtelomeric arm content vary between species and between groups within species, particularly within *T. congolense.* It was discovered that VSG genes are not always situated immediately adjacent to telomeres, that they can be present as subtelomeric arrays and that a significantly greater number of VSG genes existed on *T. godfreyi* MCs compared to those of other species within the subgenus. Other protein-coding genes were infrequently present on the MCs of all species and these included *Ingi, ESAG2, DEAH-box helicase* and uniquely to *T. godfreyi – ABC transporters.* The findings presented here contribute to the growing body of evidence that trypanosome species vary from one another in terms of their genomic architecture and content. The diversity in MC structures between species within subgenus *Nannomonas* may imply that the mechanisms for MC genesis and antigenic variation also vary between the species, though this is yet to be validated experimentally. The exploratory analyses presented here provide a template for future study and the paths this may take, emphasising the necessity of incorporating multiple species into models for processes such as MC genesis and evolution.

References

Adams, E.R. (2008). The Diversity, Stability and Prevalence of Trypanosome Infections from Wild-caught Tsetse Flies in Tanzania. *University of Bristol, Bristol.*

Adams, E.R., Hamilton, P.B., Malele, I.I. and Gibson, W. (2008). The identification, diversity and prevalence of trypanosomes in field caught tsetse in Tanzania using ITS-1 primers and fluorescent fragment length barcoding. *Infection, Genetics and Evolution* **8**, 439–444. doi:10.1016/j.meegid.2007.07.013.

Adams, E.R., Hamilton, P.B. and Gibson, W. (2010). African trypanosomes: celebrating diversity. *Trends in Parasitology* **26**, 324–328. doi:10.1016/j.pt.2010.03.003.

Alsford, S., Wickstead, B., Ersfeld, K. and Gull, K. (2001). Diversity and dynamics of the minichromosomal karyotype in Trypanosoma brucei, *Molecular and Biochemical Parasitology* **113**, 79-88.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410. doi:10.1016/S0166-6851(00)00388-1.

Auty, H., Torr, S.J., Michoel, T., Jayraman, S. and Morrison, L.J. (2015). Cattle trypanosomiasis: the diversity of trypanosomes and implications for disease epidemiology and control. *Revue Scientifique et Technique de l'OIE* **34(2)**, 587–598. doi:10.20506/rst.34.2.2382.

Aphasizheva, I., Alfonzo, J., Carnes, J., Cestari, I., Cruz-Reyes, J., Göringer H.U., Hajduk, S., Lukeš, J., Madison-Antenucci, S., Maslov, D.A., McDermott S.M., Ochsenreiter, T., Read, L.K, Salavati, R., Schnaufer, A., Schneider, A., Simpson, L., Stuart, K., Yurchenko, V., Zhou, Z.H., Zíková, Zhang, L., Zimmer, S. and Aphasizhev, R. (2020). Lexis and grammar of mitochondrial RNA processing in Trypanosomes. *Trends in Parasitology* **36**, 337–355. doi:10.1016/j.pt.2020.01.006.

Aslett, M., Aurrecoechea, C., Berriman, M., Brestelli, J., Brunk B.P., Carrington, M., Depledge, D.P., Fischer, S., Gajria, B., Gao, X., Gardner, M.J., Gingle, A., Grant, G., Harb, O.S., Heiges, M., Hertz-Fowler, C., Houston, R., Innamorato, F., Iodice, J., Kissinger, J.C., Kraemer, E., Li, W., Logan, F.J., Miller, J.A., Mitra, S., Myler, P.J., Nayak, V., Pennington, C., Phan, I., Pinney, D.F., Ramasamy, G., Rogers, M.B., Roos, D.S., Ross, C., Sivam, D., Smith, D.F., Srinivasamoorthy, G., Stoeckert, C.J., Subramanian, S., Thibodeau, R., Tivey, A., Treatman, C., Velarde, G. and Wang, H. (2010). TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Research* **38**. 457–462. doi:10.1093/nar/gkp851.

Bailey, M., Christoforidou, Z., Lewis, M.C. (2013). The evolutionary basis for differences between the immune systems of man, mouse, pig and ruminants. *Veterinary Immunology and Immunopathology* **152**, 13–9. doi:10.1016/j.vetimm.2012.09.022.

Baker, N., de Koning, H.P., Maser, P. and Horn, D. (2013). Drug resistance in African trypanosomiasis: the melarsoprol and pentamidine story, *Trends in Parasitology* **29(3)**, 110-118. doi:10.1016/j.pt.2012.12.005.

Bangs, J.D. (2018). Evolution of antigenic variation in African trypanosomes: variant surface glycoprotein expression, structure, and function. *BioEssays*. doi:10.1002/bies.201800181.

Barbet, A.F. and Kamper, S.M. (1993). The importance of mosaic genes to trypanosome survival. *Parasitology today* **9**, 63-66. 10.1016/0169-4758(93)90039-I.

Batista, J., Araújo Júnior, H., Moura, G., Góis, R., Paiva, K., Silva, J., Costa, W., Menezes, M., Nunes, F., Costa, K. and Medeiros, G. (2019). Cardiac involvement in trypanosomiasis in sheep experimentally infected by Trypanosoma vivax (Ziemman, 1905). *Experimental Parasitology* **205**, 107714. doi:10.1016/j.exppara.2019.05.008.

Bengaly, Z., Sidibe, I., Boly, H., Sawadogo, L. and Desquesnes, M. (2002) Comparative pathogenicity of three genetically distinct Trypanosoma congolense-types in inbred BALB/c mice. *Veterinary Parasitology* **105**, 111–118. doi:10.1016/s0304-4017(01)00609-4.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* **27(2)**, 573-80. doi:10.1093/nar/27.2.573.

Benvenuto, D., Giovanetti, M., Ciccozzi, A., Spoto, S., Angeletti, S. and Ciccozzi, M. (2020). The 2019-new coronavirus epidemic: evidence for virus evolution. *Journal of Medical Virology* **92**, 455–459. doi:10.1002/jmv.25688.

Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renauld, H., Bartholomeu, D. C., Lennard, N. J., Caler, E., Hamlin, N. E., Haas, B., Böhme, U., Hannick, L., Aslett, M. A., Shallom, J., Marcello, L., Hou, L., Wickstead, B., Alsmark, U. C. M., Arrowsmith, C., Atkin, R. J., Barron, A. J., Bringaud, F., Brooks, K., Carrington, M., Cherevach, I., Chillingworth, T. J., Churcher, C., Clark, L. N., Corton, C. H., Cronin, A., Davies, R. M., Doggett, J., Djikeng, A., Feldblyum, T., Field, M. C., Fraser, A., Goodhead, I., Hance, Z., Harper, D., Harris, B. R., Hauser, H., Hostetler, J., Ivens, A., Jagels, K., Johnson, D., Johnson, J., Jones, K., Kerhornou, A. X., Koo, H., Larke, N., Landfear, S., Larkin, C., Leech, V., Line, A., Lord, A., MacLeod, A., Mooney, P. J., Moule, S., Martin, D. M. A., Morgan, G. W., Mungall, K., Norbertczak, H., Ormond, D., Pai, G., Peacock, C. S., Peterson, J., Quail, M. A., Rabbinowitsch, E., Rajandream, M. A., Reitter, C., Salzberg, S. L., Sanders, M., Schobel, S., Sharp, S., Simmonds, M., Simpson, A. J., Tallon, L., Turner, C. M. R., Tait, A., Tivey, A. R., Van Aken, S., Walker, D., Wanless, D., Wang, S., White, B., White, O., Whitehead, S., Woodward, J., Wortman, J., Adams, M. D., Embley, T. M., Gull, K., Ullu, E., Barry, J. D., Fairlamb, A. H., Opperdoes, F., Barrell, B. G. and Donelson, J. E. (2005). The genome of the African trypanosome Trypanosoma brucei. *Science* **309**, 416–422. doi:10.1126/science.1112642.

Bhattacharya, S., Bakre, A. and Bhattacharya, A. (2002) Mobile genetic elements in protozoan parasites. *Journal of Genetics* **81**, 73–86. doi:10.1007/bf02715903.

Bio-rad Laboratories, Inc. (2011). Pulsed field gel electrophoresis. *Bio-Rad.* Available at:https://www.bio-rad.com/en-uk/applications-technologies/pulsed-field-gel-electrophoresis?ID=LUSORPDFX#1

Blum, M.L., Down, J.A., Gurnett, A.M, Carrington, M., Turner, M.J. and Wiley, D.C. (1993). A structural motif in the variant surface glycoproteins of Trypanosoma brucei. *Nature* **362**, 603-609. doi:10.1038/362603a0.

Borst, P. (1986). Discontinuous transcription and antigenic variation in trypanosomes. *Annual Review of Biochemistry* **55**, 701-732. doi:10.1146/annurev.bi.55.070186.003413.

Borst, P., Rudenko, G., Taylor, M.C., Blundell, P.A., Van Leeuwen, F., Bitter, W., Cross, M. and McCulloch, R. (1996). Antigenic variation in trypanosomes. *Archive of Medical Research* **27**, 379-388.

Borst, P. and Ulbert, S. (2001). Control of VSG gene expression sites. *Molecular and Biochemical Parasitology* **114**, 17-27. doi :10.1016/s0116-6851(01)00243-2.

Bossche, P.V. and Delespaux, V. (2011). Options for the control of tsetse-transmitted livestock trypanosomosis: an epidemiological perspective. *Veterinary Parasitology* **181**, 37-42. doi:10.1016/j.vetpar.2011.04.021.

Bringaud, F., Biteau, N., Zuiderwijk, E., Berriman, M., El-Sayed, N.M., Ghedin, E., Melville, S.,E., Hall, N. and Baltz, T. (2004). The ingi and RIME non-LTR retrotransposons are not randomly distributed in the genome of Trypanosoma brucei. *Molecular Biology and Evolution* **21**, 520–528. doi:10.1093/molbev/msh045.

Bringaud, F., Berriman, M. and Hertz-Fowler, C. (2009). Trypanosomatid genomes contain several subfamilies of ingi-related retroposons. *Eukaryotic Cell* **8**, 1532–1542. doi :10.1128/EC.00183-09.

Butikofer, P., Vassella, E., Boschung, M., Renggli, C.K., Brun, R., Pearson, T.W and Roditi, I. (2002). Glycosylphosphatidylinositol-anchored surface molecules of Trypanosoma congolense insect forms are developmentally regulated in the tsetse fly. *Molecular and Biochemical Parasitology* **119**, 7-16. doi:10.1016/s0166-6851(01)00382-6.

Claxton, J.R., Faye, J.A. and Rawlings, P. (1992). Trypanosome infections in warthogs (Phacochoerus aethiopicus) in the Gambia. *Veterinary Parasitology* **41**, 179–187. doi:10.1016/0304-4017(92)90077-M.

Coustou, V., Guegan, F., Plazolles, N., and Baltz, T. (2010). Complete In Vitro Life Cycle of Trypanosoma congolense: Development of Genetic Tools. *Plos Neglected Tropical Diseases* **4(3)**, e618, doi:10.1371/journal.pntd.0000618.

Cross, G.A.M. (1975). Identification, purification and properties of clone specific glycoprotein antigens constituting the surface coat of Trypanosoma brucei, *Parasitology* **71**, 393-417, doi :10.1017/S003118200004717X.

Cross, G.A.M., Kim, H.S. and Wickstead, B. (2014). Capturing the variant surface glycoprotein repertoire (the VSGnome) of Trypanosoma brucei Lister 427. *Molecular and Biochemical Parasitology* **195**, 59-73, doi:10.1016/j.molbiopara.2014.06.004.

De Lange, T., Liu, A.Y., Van der Ploeg, L.H., Borst, P., Tromp, M.C. and Van Boom, J.H. (1983). Tandem repetition of the 5' mini-exon of variant surface glycoprotein genes: a multiple promoter for VSG gene transcription?. *Cell* **34**, 891-900. doi:10.1016/0092-8674(83)90546-9.

Dennis, J.W., Durkin, S.M., Horsley Downie, J.E., Hamill, L.C., Anderson, N.E. and MacLeod, E.T. (2014). Sodalis glossinidius prevalence and trypanosome presence in tsetse from Luambe National Park, Zambia. *Parasites and Vectors* **7**, 378. doi:10.1186/1756-3305-7-378.

Desquesnes, M., Holzmuller, P., Lai, D. H., Dargantes, A., Lun, Z. R., and Jittaplapong, S. (2013). Trypanosoma evansi and surra: a review and perspectives on origin, history, distribution, taxonomy, morphology, hosts, and pathogenic effects. *BioMed research international,* 194176. doi:10.1155/2013/194176.

Desta, M., Beyene, D. and Haile, S. (2013). Trypanosome infection rate of Glossina pallidipes and trypanosomosis prevalence in cattle in Amaro Special District of Southern Ethiopia. *Journal of Veterinary Medicine and Animal Health* **5(6)**, 164–170. doi:10.5897/JVMAH2013.0199.

Dukes, P., Faye, J., McNamara, J.J., Snow, E.F., Rawlings, P., Dwinger, R.H. and Brun, R. (1989). Isolation and cultivation in vitro to the infective, metacyclic stage of Trypanosoma (Nannomonas) simiae from Glossina morsitans submorsitans. *Acta Tropica* **46(3)**, 191-203. DOI: 10.1016/0001-706x(89)90036-3.

El-Sayed, N.M., Hegde, P., Quackenbush, J., Melville, S.E., Donelson, J.E. (2000). The African trypanosome genome. *International Journal for Parasitology* **30**, 329-345. doi:10.1016/S0020-7519(00)00015-1.

Enyaru, J.C., Ouma, J.O., Malele, I.I., Matovu, E. and Masiga, D.K. (2010). Landmarks in the evolution of technologies for identifying trypanosomes in tsetse flies. *Trends in Parasitology* **26**, 388–394. doi:10.1016/j.pt.2010.04.011.

Ersfeld, K. and Gull, K. (1997). Partitioning of large and minichromosomes in Trypanosoma brucei. *Science* **276**. 611–614. doi:10.1126/science.276.5312.611.

Ersfeld, K., Melville, S.E. and Gull, K. (1999). Nuclear and genome organization of Trypanosoma brucei. *Parasitology Today* **15**, 58–63. doi :10.1016/S0169-4758(98)01378-7.

Esser, K.M., Schoenbechler, M.J. and Gingrich, J.B. (1982). Trypanosoma rhodesiense blood forms express all antigen specificities relevant to protection against metacyclic (insect form) challenge. *The Journal of Immunology* **129(4)**, 1715-1718.

Fedotova, A. A., Bonchuk, A. N., Mogila, V. A. and Georgiev, P. G. (2017). C2H2 zinc finger proteins: the largest but poorly explored family of higher eukaryotic transcription factors. *Acta Naturae* **9**, 47–58, doi:10.32607/20758251-2017-9-2-47-58.

Franco, J., Ferreira, R.C., Ienne, S. and Zingales, B. (2015). ABCG-like transporter of Trypanosoma cruzi involved in benznidazole resistance: gene polymorphisms disclose inter-strain intragenic recombination in hybrid isolates. *Infection, Genetics and Evolution* **31**, 198-208. doi:10.1016/j.meegid.2015.01.030.

Gadelha, C., Zhang, W.Z., Chamberlain, J.W., Chait, B.T., Wickstead, B. and Field, M.C. (2015). Architecture of a host-parasite interface: complex targeting mechanisms revealed through proteomics. *Molecular and Cellular Proteomics* **14**, 1911-1926. doi:10.1074/mcp.M114.047647.

Gardiner, P.R. (1989). Recent studies of the biology of Trypanosoma vivax. *Advances in Parasitology* **28**, 229–317. doi:10.1016/S0065-308X(08)60334-6.

Garside, L., Bailey, M. and Gibson, W. (1994). DNA content and molecular karyotype of trypanosomes of the subgenus Nannomonas. *Acta tropica* **57**, 21-8. doi:10.1016/0001-706X(94)90089-2.

Garside, L. H. and Gibson, W. C. (1995). Molecular characterisation of trypanosome species and subgroups within subgenus Nannomonas. *Parasitology* **111**, 301-312. doi:10.1017/S0031182000081853.

Gashumba, J. K., Gibson, W. C. and Opiyo, E. A. (1986a). A preliminary comparison of Trypanosoma simiae and T. congolense by isoenzyme electrophoresis. *Acta Tropica* **43**, 15-19.

Gashumba, J.K. (1986b). Two enzymically distinct stocks of Trypanosoma congolense. *Research in Veterinary Science* **40(3)**, 411-412. doi:10.1016/S0034-5288(18)30561-7.

Gashumba, J. K., Baker, R. D. and Godfrey, D. G. (1988). Trypanosoma congolense: The distribution of enzymic variants in East and West Africa. *Parasitology* **96**, 475–486. doi:10.1017/S0031182000080112.

González-de la Fuente, S., Peiró-Pastor, R., Rastrojo, A., Moreno, J., Carrasco-Ramiro, F., Requena, J.M. and Aguado, B. (2017). Resequencing of the Leishmania infantum (strain JPCM5) genome and de novo assembly into 36 contigs. *Scientific Reports* **7**, 18050. doi:10.1038/s41598-017-18374-y.

González-de la Fuente, S., Camacho, E., Peiró-Pastor, R., Rastrojo, A., Carrasco-Ramiro, F., Aguado, B. and Requena, J.M. (2019). Complete and de novo assembly of the Leishmania braziliensis (M2904) genome. *Memórias Do Instituto Oswaldo Cruz* **114**, e180438. doi:10.1590/0074-02760180438.

Gibson, W., Borst, P. (1986). Size fractionation of the small chromosomes of Trypanozoon and Nannomonas trypanosomes by PFGE**.** *Molecular and Biochemical Parasitology* **18**, 127-140. doi:10.1016/0166-6851(86)90033-2.

Gibson, W., Dukes, P., & Gashumba, J. (1988). Species-specific DNA probes for the identification of African trypanosomes in tsetse flies. *Parasitology* **97(1)**, 63-73. doi:10.1017/S0031182000066749.

Gibson, W. and Bailey, M. (1994). Genetic exchange in Trypanosoma brucei: evidence for meiosis from analysis of a cross between drug resistant transformants. *Molecular and Biochemical Parasitology* **64**, 241-252. doi:10.1016/0166-6851(94)00017-4.

Gibson, W., Stevens, J., Mwendia, C., Makumi, J., Ngotho, J. and Ndung'u, J. (2001). Unravelling the phylogenetic relationships of African trypanosomes of suids. *Parasitology* **122(6)***,* 625-631. doi:10.1017/S0031182001007880.

Gibson, W. (2007) Resolution of the species problem in African trypanosomes. *International Journal of Parasitology* **37(8-9)**, 829-838. doi:10.1016/j.ijpara.2007.03.002.

Gibson, W., Kay, C. and Peacock, L. (2017). Trypanosoma congolense: Molecular Toolkit and Resources for Studying a Major Livestock Pathogen and Model Trypanosome, *Advances in Parasitology* **98**, 283–309. doi: 10.1016/bs.apar.2017.03.002.

Gillingwater, K., Mamabolo, M.V. and Majiwa, P.A.O. (2010). Prevalence of mixed Trypanosoma congolense infections in livestock and tsetse in KwaZulu-Natal, South Africa. *Journal of the South African Veterinary Association* **81(4),** 219-223. doi:10.4102/jsava.v81i4.151.

Gottesdiener, K., Garcia-Anoveros, J., Lee, M.G. and Van der Ploeg, L.H. (1990). Chromosome organisation of the protozoan Trypanosoma brucei. *Molecular and Cellular Biology* **10**, 6079-6083. doi:10.1128/MCB.10.11.6079.

Guizetti, J., Martins, R.M., Guadagnini, S., Claes, A. and Scherf, A. (2013). Nuclear pores and perinuclear expression Sites of var and ribosomal DNA genes correspond to physically distinct regions in Plasmodium Falciparum. *Eukaryotic Cell* **12**, 697-702, doi:10.1128/EC.00023-13.

Günzl, A., Bruderer, T., Laufer, G., Schimanski, B., Tu, L.C., Chung, H.M., Lee, P.T and Lee, M.G.S. (2003) RNA polymerase I transcribes procyclin genes and variant surface glycoprotein gene expression sites in Trypanosoma brucei. *Eukaryotic Cell* **2**, 542-551. doi:10.1128/EC.2.3.542-551.2003.

Haag, J. (1998). The molecular phylogeny of trypanosomes: evidence for an early divergence of the Salivaria. *Molecular and Biochemical Parasitology* **91(1)**, 37-49. doi:10.1016/s0166-6851(97)00185-0.

Haile-Meskel, T.M. (2016). Trypanosomiasis costs 37 African countries USD 4.5 Billion yearly. *FAO, Food and Agriculture Organization of the United Nations*. Available at: http://www.fao.org/africa/news/detail-news/en/c/461166/

Hamilton, P., Stevens, J., Gaunt, M., Gidley, J. and Gibson, W. (2004). Trypanosomes are monophyletic: evidence from genes for glyceraldehyde phosphate dehydrogenase and small subunit ribosomal RNA, *International Journal of Parasitology* **34(12)**, 1393-1404. doi:10.1016/j.ijpara.2004.08.011

Hoare, C.A. (1972). The Trypanosomes of Mammals. *Journal of Small Animal Practice* **13**, 671-672. doi: 10.1111/j.1748-5827.1972.tb06818.x.

Hong, M. and Simpson, L. (2003). Genomic organization of Trypanosoma brucei kinetoplast DNA minicircles. *Protist* **154**, 265-279. doi:10.1078/143446103322166554.

Hovel-Miner, G., Mugnier, M.R., Goldwater, B., Cross, G.A.M. and Papavasiliou, F.N. (2016), A conserved DNA repeat promotes selection of a diverse repertoire of Trypanosoma brucei surface antigens from the genomic archive. *PLoS Genetics* **12**, e1005994. doi:10.1371/journal.pgen.1005994.

Huddleston, J., Ranade, S., Malig, M., Antonacci, F., Chaisson, M., Hon, L., Sudmant, P.H., Graves, T.A., Alkan, C., Dennis, M.Y., Wilson, R.K., Turner, S.W., Korlach, J. and Eichler, E.E. (2014). Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Research* **24**, 688-696. doi: 10.1101/gr.168450.113.

Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment, *Computing in Science & Engineering* **9**, 90-95. doi:10.1109/MCSE.2007.55.

Hutchinson, R. and Gibson, W. (2015). Rediscovery of Trypanosoma (Pycnomonas) suis, a tsetse-transmitted trypanosome closely related to T. brucei, *Infection, Genetics and Evolution* **36**, 381-388. doi:10.1016/j.meegid.2015.10.018.

FAO. (1990). Cost-benefit analysis for animal health programmes in developing countries. *FAO expert consultation, Rome*, 56.

Jackson, A.P., Sanders, M., Berry, A., McQuillan, J., Aslett, M.A., Quail, M.A., Chukualim., B., Capewell, P., MacLeod, A., Melville, S.E., Gibson, W., Barry, J.D., Berriman, M. and Hertz-Fowler, C. (2010). The genome sequence of Trypanosoma brucei gambiense, causative agent of chronic human african trypanosomiasis. *PLoS Neglected Tropical Diseases* **4**, e658. doi:10.1371/journal.pntd.0000658.

Jackson, A. P., Berry, A., Aslett, M., Allison, H. C., Burton, P., Vavrova-Anderson, J., Brown, R., Browne, H., Corton, N., Hauser, H., Gamble, J., Gilderthorp, R., Marcello, L., McQuillan, J., Otto, T. D., Quail, M. A., Sanders, M. J., van Tonder, A., Ginger, M. L., Field, M. C., Barry, J. D., Hertz-Fowler, C. and Berriman, M. (2012). Antigenic diversity is generated by distinct evolutionary mechanisms in African trypanosome species. *Proceedings of the National Academy of Sciences* **109**, 3416–3421. doi:10.1073/pnas.1117313109.

Jackson, A., Allison, H., Barry, J., Field, M., Hertz-Fowler, C. and Berriman, M. (2013). A Cell-surface Phylome for African Trypanosomes, *Plos Neglected Tropical Diseases* **7(3)**, e2121. doi:10.1371/journal.pntd.0002121.

Janssen, J.A.H.A. and Wijers, D.J.B. (1974). Trypanosoma simiae at the Kenya coast. A correlation between virulence and the transmitting species of Glossina. *Annals of Tropical Medicine and Parasitology* **68**, 5–19. doi:10.1080/00034983.1974.11686919.

Jiang, L., Mu, J., Zhang, Q., Ni, T., Srinivasan, P., Rayavara, K., Yang, W., Turner, L., Lavstsen., Theander, T.g., Peng, W., Wei, G., Jing, Q., Wakabayashi, Y., Bansal, A., Luo, Y., Ribeiro, J.M.C., Scherf, A., Aravind, L., Zhu, J., Zhao, K. and Miller, L. (2013). PfSETvs methylation of histone H3K36 represses virulence genes in Plasmodium falciparum. *Nature* **499**, 223–227, doi:10.1038/nature12361.

Kaare, M.T., Picozzi, K., Mlengeya, T., Fevre, E.M., Mellau, L.S., Mtambo, M.M, Cleaveland, S. and Welburn, S.C. (2007). Sleeping sickness - a re-emerging disease in the Serengeti?. *Travel Medicine and Infectious Disease* **5**, 117–124. doi:10.1016/j.tmaid.2006.01.014.

Katoh, K. and Standley, D.M. (2014). MAFFT multiple sequence alignment software version 7: improvements in performance and usability, *Molecular Biology and Evolution* **30**, 772-780, doi:10.1093/molbev/mst010.

Kazazian, H.H. (2004). Mobile elements: drivers of genome evolution. *Science* **303(5664)**, 1626-1632. doi:10.1126/science.1089670.

Kolev, N.G., Günzl, A. and Tschudi, C. (2017). Metacyclic VSG expression site promoters are recognized by the same general transcription factor that is required for RNA polymerase I transcription of bloodstream expression sites. *Molecular and Biochemical Parasitology* **216**, 52–55. doi:10.1016/j.molbiopara.2017.07.002.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R. and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research* **27(5)**, 722-736. doi:10.1101/gr.215087.116.

Kristjanson, P.M., Swallow, B.M., Rowlands, G.J., Kruska, R.L. and de Leeuw, P.N. (1999). Measuring the costs of African animal trypanosomiasis, the potential benefits of control and returns to research. *Agricultural Systems* **59**, 79-98. doi:10.1016/S0308-521X(98)00086-9.

Kukla, B. A., Majiwa, P. A. O., Young, J. R., Moloo, S. K. and Ole-Moiyoi, O. (1987). Use of species-specific DNA probes for detection and identification of trypanosome infections in tsetse flies. *Parasitology* **95**, 1–16.

Lehane, M., Msangi, A., Whitaker, C. and Lehane, S. (2000). Grouping of trypanosome species in mixed infections in Glossina pallidipes. *Parasitology* **120(6)**, 583-592. doi:10.1017/S0031182099005983.

Lenardo, M.J., Rice-Ficht, C., Kelly, G., Esser, K.M. and Donelson, J.E. (1984). Characterization of the genes specifying two metacyclic variable antigen types in Trypanosoma brucei rhodesiense. *Proceedings of the National Academy of Sciences USA* **8121**, 6642–6646. doi:10.1073/pnas.81.21.6642.

Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659. doi:10.1093/bioinformatics/btl158.

Lin, Y., Yuan, J., Kolmogorov, M., Shen, M.W., Chaisson, M. and Pevzner, P. (2016). Assembly of Long Error-Prone Reads Using de Bruijn Graphs. *PNAS,* E8396-E8405. doi:10.1073/pnas.1604560113.

Lloyd, L. and Johnson, W.B. (1984). The trypanosome infections of tsetse flies in northern Nigeria and a new method of estimation. *Bulletin of Entomological Research* **14**, 265-288. doi:10.1017/S0007485300028352.

Magona, J.W., Mayende, J.S.P., Olaho-Mukani, W., Coleman, P.G., Jonsson, N.N., Welburn, S.C., and Eisler, M.C. (2003). A comparative study on the clinical, parasitological and molecular diagnosis of bovine trypanosomosis in Uganda. *Onderstepoort Journal of Veterinary Research* **70(3)**, 213–218.

Majiwa, P.A.O., Young, J.R., Hamers, R. and Matthyssens, G. (1986). Minichromosomal variable surface glycoprotein genes and molecular karyotypes of Trypanosoma (Nannomonas) congolense. *Gene* **41**, 183-192. doi:183-192. 10.1016/0378-1119(86)90097-1.

Majiwa, P.A.O. and Otieno, L.H. (1990). Recombinant DNA probes reveal simultaneous infection of tsetse flies with different trypanosome species. *Molecular and Biochemical Parasitology* **40**, 245–54. doi:10.1016/0166-6851(90)90046-O.

Majiwa, P.A.O. (1992). The variability of Trypanosoma congolense. *Genome Analysis of Protozoan Parasites: Proceedings of a Workshop Held at ILRAD, Nairobi, Kenya, 11–13 November 1992* ed. S.P. Morzaria. Nairobi: The International Laboratory for Research on Animal Diseases, 1993. 94-99

Majiwa, P.A.O., Waitumbi, J., Mihok, S. and Zweygarth, E. (1993). Trypanosoma (Nannomonas) congolense: Molecular characterization of a new genotype from Tsavo, Kenya. *Parasitology* **106(2)**, 151-62. doi:10.1017/S0031182000074941.

Malele, I., Craske, L., Knight, C., Ferris, V., Njiru, Z., Hamilton, P., Lehane, S., Lehane, M. and Gibson, W. (2003). The use of specific and generic primers to identify trypanosome infections of wild tsetse flies in Tanzania by PCR. *Infection, Genetics and Evolution* **3**, 271–279. doi:10.1016/s1567-1348(03)00090-x.

Marcello, L. and Barry, J.D. (2007). Analysis of the VSG gene silent archive in Trypanosoma brucei reveals that mosaic gene expression is prominent in antigenic variation and is favoured by archive substructure. *Genome Research* **17**, 1344–1352. doi:10.1101/gr.6421207.

Masiga, D.K., Smyth, A.J., Hayes, P., Bromidge, T.J. and Gibson, W.C. (1992). Sensitive detection of trypanosomes in tsetse flies by DNA amplification. *International Journal of Parasitology* **22**, 909-918. doi:10.1016/0020-7519(92)90047-O.

Matthews, K.R. (1999). Developments in the differentiation of Trypanosoma brucei. *Parasitology Today* **15(2)**, 76-80. doi:10.1016/s0169-4758(98)01381-7.

McCoy, R.C., Taylor, R.W., Blauwkamp, T.A., Kelley, J.L., Kertesz, M., Pushkarev, D., Petrov, D.A. and Fiston-Lavier, A.S. (2014). Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. *PLoS One* **9(9)**, e106689. doi:10.1371/journal.pone.0106689.

Melville, S.E., Leech, V., Gerrard, C.S., Tait, A., Blackwell, J.M. (1998). The molecular karyotype of the megabase chromosomes of Trypanosoma brucei and the assignment of chromosome markers. *Molecular and Biochemical Parasitology* **94**, 155-173. doi:10.1016/S0166-6851(98)00054-1.

Melville, S.E., Gerrard, C.S. and Blackwell, J.M. (1999). Multiple causes of size variation in the diploid megabase chromosomes of African trypanosomes. *Chromosome Research* **7**, 191-203. doi:10.1023/a:1009247315947.

Melville, S.E., Leech, V., Navarro, M. and Cross, G.A.M. (2000). The molecular karyotype of the megabase chromosomes of Trypanosoma brucei stock 427. *Molecular and Biochemical Parasitology* **111**, 261–273. doi:10.1016/S0166-6851(00)00316-9.

McNamara, J.J., Dukes, P., Snow, W.F. and Gibson, W.C. (1989). Use of DNA probes to identify Trypanosoma congolense and T. simiae in tsetse flies from the Gambia. *Acta Tropica* **46**, 55-61. doi:10.1016/0001-706X(89)90016-8.

McNamara, J.J. and Snow, W.F. (1991). Improved identification of Nannomonas infections in tsetse flies from The Gambia. *Acta Tropica* **48**, 127-136. doi:10.1016/0001-706X(90)90052-2.

McNamara, J.J., Mohammed, G. and Gibson, W.C. (1994). Trypanosoma (Nannomonas) godfreyi sp. nov. from tsetse flies in the Gambia - biological and biochemical characterization. *Parasitology* **109**, 497–509. doi:10.1017/S0031182000080756.

Mori, H., Evans-Yamamoto, D., Ishiguro, S., Tomita, M. and Yachie, N. (2019). Fast and global detection of periodic sequence repeats in large genomic resources. *Nucleic Acids Research* **47(2)**, e8. doi: 10.1093/nar/gky890.

Morrison, L. J., Vezza, L., Rowan, T. and Hope, J. C. (2016). Animal African Trypanosomiasis: Time to Increase Focus on Clinically Relevant Parasite and Host Species. *Trends in Parasitology* **32**, 599–607, doi:10.1016/j.pt.2016.04.012.

Moser D. R., Cook G. A., Ochs D. E., Bailey C. P., McKane M. R. and Donelson J. E. (1989). Detection of Trypanosoma congolense and Trypanosoma brucei subspecies by DNA amplification using the polymerase chain reaction. *Parasitology* **99**, 57-66. doi:10.1017/S0031182000061023.

Mu, J., Awadalla, P., Duan, J., McGee, K.M., Keebler, J., Seydel, K., McVean, G.A.T. and Su, X. (2007). Genome-wide variation and identification of vaccine targets in the Plasmodium falciparum genome. *Nature Genetics* **39**. 126–130. doi:10.1038/ng1924.

Muhanguzi, D., Picozzi, K., Hatendorf, J., Thrusfield, M., Welburn, S.C., Kabasa, J.D. and Waiswa, C. (2014). Improvements on Restricted Insecticide Application Protocol for Control of Human and Animal African Trypanosomiasis in Eastern Uganda. *PLoS Neglected Tropical Diseases* **8(10)**. doi:10.1371/journal.pntd.0003284.

Muller, L.S.M., Cosentino, R.O., Forstner, K.U., Guizetti, J., Wedel, C., Kaplan, N., Janzen, C.J., Arampatzii, P., Vogel, J., Steinbiss, S., Otto, T.D., Saliba, A.E., Sebra, R.P. and Siegel, N. (2018). Genome organization and DNA accessibility control antigenic variation in trypanosomes. *Nature* **563**, 121-125. doi:10.1038/s41586-018-0619-8.

Murray, M., Trial, J. and D'ieteren, G. (1990). Trypanotolerance in cattle and prospects for the control of trypanosomiasis by selective breeding. *Revue Scientifique et Technique de l'OIE* **9(2)**, 369-386. doi:10.20506/rst.9.2.506.

Nimpaye, H., Njiokou, F., Njine, T. and Simo, G. (2011). Trypanosoma vivax, T. congolense “forest type” and T. simiae: Prevalence in domestic animals of sleeping sickness foci of Cameroon, *Parasite* **18**, 171-179. doi:10.1051/parasite/2011182171.

Njiokou, F., Simo, G., Nkinin, S. and Herder, S. (2004). Infection rate of Trypanosoma brucei s.l., T. vivax, T. congolense "forest type" and T. simiae in small wild vertebrates in south Cameroon. *Acta Tropica* **92**, 139-146. doi:10.1016/j.actatropica.2004.04.011.

Otto, T.D., Dillon, G.P., Degrave, W.S. and Berriman, M. (2011). RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Research* **39(9)**, e57. doi:10.1093/nar/gkq1268.

Pays, E., Tebabi, P., Pays, A., Coquelet, H., Revelard, P., Salmon, D. and Steinert, M. (1989). The genes and transcripts of an antigen gene expression site from T. brucei. *Cell* **57**, 835-845. doi:10.1016/0092-8674(89)90798-8.

Pays, E., Vanhamme, L., Berber, M. (1994). Genetic controls for the expression of surface antigens in African trypanosomes. *Annual Review of Microbiology* **48**, 25-52. doi:10.1146/annurev.mi.48.100194.000325

Pays, E. and Nolan, D.P (1998). Expression and function of surface proteins in Trypanosoma brucei, Molecular and Biochemical Parasitology **91**, 3-36. doi:10.1016/s0166-6851(97)00183-7.

Peacock, L., Ferris, V., Bailey, M. and Gibson, W. (2008). Fly transmission and mating of Trypanosoma brucei brucei strain 427. *Molecular and Biochemical Parasitology* **160 (2)**, 100-106. doi:10.1016/j.molbiopara.2008.04.009.

Peacock, L., Cook, S., Ferris, V., Bailey, M. and Gibson, W. (2012). The life cycle of Trypanosoma (Nannomonas) congolense in the tsetse fly. *Parasites and Vectors* **5**, 1–13. doi:10.1186/1756- 3305-5-109.

Pollard, M.O., Gurdasani, D., Mentzer, A.J., Porter, T., Sandhu, M.S. (2018). Long reads: their purpose and place. *Human molecular genetics* **27**, 234–241. doi:10.1093/hmg/ddy177.

Prucca, C.G., Slavin, I., Quiroga, R., Elías, E.V., Rivero, F.D., Saura, A., Carranza, P.G. and Luján, H.D. (2008). Antigenic variation in Giardia lamblia is regulated by RNA interference. *Nature* **456**, 750–754. doi:10.1038/nature07585.

Prucca, C. G. and Lujan, H. D. (2009). Antigenic variation in Giardia lamblia. *Cellular Microbiology* **11**, 1706–1715. doi:10.1111/j.1462-5822.2009.01367.x.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org/.

Rhoads, A. and Au, K.F. (2015). PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics* **13**, 278–289. doi:10.1016/j.gpb.2015.08.002.

Riou, G. and Pautrizel, R. (1969). Nuclear and kinetoplastic DNA from trypanosomes. *The Journal of Protozoology* **16**, 3. doi:10.1111/j.1550-7408.1969.tb02309.x.

Robinson, N. P., Burman, N., Melville, S. E. and Barry, J. D. (1999). Predominance of duplicative VSG gene conversion in antigenic variation in African trypanosomes. *Molecular and Cellular Biology* **19**, 5839–5846. doi:10.1128/MCB.19.9.5839.

Ryley, J.F. (1956). Studies on the metabolism of the protozoa. Comparative carbohydrate metabolism of eleven species of trypanosome. *Biochemical Journal* 62, 215–222, doi:10.1042/bj0620215.

Salim, B., Bakheit, M.A., Salih, S.E., Kamau, J., Nakamura, I, Nakao, R. and Sugimoto, C. (2011). An outbreak of bovine trypanosomiasis in the Blue Nile State, Sudan. *Parasites and Vectors* **4**, 74. doi:10.1186/1756-3305-4-74.

Schmidt, M. H. and Pearson, C. E. (2016). Disease-associated repeat instability and mismatch repair. *DNA Repair* **38**, 117–126. doi:10.1016/j.dnarep.2015.11.008.

Schwartz, D.C. and Cantor, C.R. (1984). Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell* **37**, 67-75. doi:10.1016/0092-8674(84)90301-5.

Seibt, K. M., Schmidt, T. and Heitkam, T. (2018). FlexiDot: highly customizable, ambiguity-aware dotplots for visual sequence analyses. *Bioinformatics* **34**, 3575–3577. doi:10.1093/bioinformatics/bty395.

Shaw A.P., Cecchi, G., Wint, G.R.W., Mattioli, R.C., Robinson, T.P. (2014). Mapping the economic benefits to livestock keepers from intervening against bovine trypanosomiasis in Eastern Africa. *Preventive Veterinary Medicine* **113**, 197-210. doi:10.1016/j.prevetmed.2013.10.024.

Solano, P., Guégan, J. F., Reifenberg, J. M. and Thomas, F. (2001). Trying to identify, predict and explain the presence of african trypanosomes in tsetse flies. *Journal of Parasitology* **87**, 1058–1063. doi:10.1645/0022-3395(2001)087[1058:TTPAET]2.0.CO;2.

Silvester, E., Ivens, A., & Matthews, K. R. (2018). A gene expression comparison of Trypanosoma brucei and Trypanosoma congolense in the bloodstream of the mammalian host reveals species-specific adaptations to density-dependent development. *PLoS neglected tropical diseases* **12(10)**, e0006863. doi:10.1371/journal.pntd.0006863.

Simpson, L. (1987). The mitochondrial genome of kinetoplastid protozoa: genomic organization, transcription, replication, and evolution. *Annual Reviews in Microbiol*ogy **41**, 363-382. doi: 10.1146/annurev.mi.41.100187.002051.

Simpson, A.G.B., Lukes, J. and Roger, A.J. (2002). The evolutionary history of kinetoplastids and their kinetoplasts. *Molecular Biology and Evolution* **19**, 2071–2083. doi:10.1093/oxfordjournals.molbev.a004032.

Sloof, P., Bos, J.L., Konings, A.F.J.M., Menke, H.H., Borst, P., Gutteridge, W.E. and Leon, W. (1983). Characterization of satellite DNA in Trypanosoma brucei and Trypanosoma cruzi. *Journal of Molecular Biology* **167**. 1-21, doi:10.1016/S0022-2836(83)80031-X.

Stephen, L. E. (1966). Pig Trypanosomiasis in Africa. *Commonwealth Agricultural Bureaux Farnham Royal* **8**, 65. doi:10.4269/ajtmh.1966.15.6.TM0150061010a.

Stevens, J. and Gibson, W. (1999). The molecular evolution of trypanosomes. *Parasitology Today* **15**, 432–437. doi:10.1016/s0169-4758(99)01532-x.

Swallow, B.M. (2000). Impacts of trypanosomiasis on African agriculture. *PAAT technical and scientific series* **2**, 52.

Taylor, K., and Mertens, B., (1999). Immune Response of Cattle Infected with African Trypanosomes. *Memórias Do Instituto Oswaldo Cruz* **94(2)**, 239-244. doi:10.1590/s0074-02761999000200022.

Tihon, E., Imamura, H., Dujardin, J.C., Van Den Abbeele, J. and Van den Broeck, F. (2017). Discovery and genomic analyses of hybridization between divergent lineages of Trypanosoma congolense, causative agent of Animal African Trypanosomiasis. *Molecular Ecology* **26(23),** 6524-6538. doi:10.1111/mec.14271.

Treangen, T.J. and Salzberg, S.L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics* **13(1)**, 36-46. doi:10.1038/nrg3117.

Truc, P., Büscher, P., Cuny, G., Gonzatti, M., Jannin, J., Joshi, P., Juyal, P., Lun, Z., Mattioli, R., Pays, E., Simarro, P., Teixeira, M., Touratier, L., Vincendeau, P. and Desquesnes, M. (2013). Atypical Human Infections by Animal Trypanosomes. *PLoS Neglected Tropical Diseases* **7(9)**, 2256. doi:10.1371/journal.pntd.0002256.

Turner, C.M. R., Melville, S.E. and Tait, A. (1997). A proposal for karyotype nomenclature in T. brucei. *Parasitology Today* **13**, 5-6. doi:10.1016/s0169-4758(96)20056-0.

Turner, C.M., and Barry, J.D., 1989. High frequency of antigenic variation in *Trypanosoma brucei* rhodesiense infections. *Parasitology* **99**, 67–75, doi:10.1017/S0031182000061035.

Vallat, R. (2018). Pingouin: statistics in Python. *The Journal of Open Source Software* **3(31)**, 1026. doi:10.21105/joss.01026.

Van der Ploeg, L., Schwartz, D., Cantor, C. and Borst, P. (1984). Antigenic variation in Trypanosoma brucei analyzed by electrophoretic separation of chromosome-sized DNA molecules. *Cell* **37 (1)**, 77-**84**. doi:10.1016/0092-8674(84)90302-7.

van Dijk, J.E., Zwart, D. and Leeflang, P. (1973). A Contribution to the Pathology of Trypanosoma simiae Infection in Pigs. *Zentralblatt für Veterinärmedizin Reihe B* **20**, 374-391. doi:10.1111/j.1439-0450.1973.tb01139.x.

van Dijk, E.L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. (2018). The third revolution in sequencing technology. *Trends in Genetics* **34**, 666–681. doi:10.1016/j.tig.2018.05.008.

Vickerman, K. and Preston, T.M. (1970). Spindle microtubules in the dividing nuclei of trypanosomes. *Journal of Cell Science* **6**, 365-383.

Wacher, T., Milligan, P., Rawlings, P. and Snow, W. (1994). Tsetse–trypanosomiasis challenge to village N'Dama cattle in The Gambia: field assessments of spatial and temporal patterns of tsetse–cattle contact and the risk of trypanosomiasis infection, *Parasitology* **109(2)**, 149–162, doi:https://doi.org/10.1017/S0031182000076265.

Walder, J.A., Eder, P.S., Engman, D.M., Brentano, S.T., Walder, R.Y., Knutzon, D.S., Dorfman, D.M. and Donelson, J.E. (1986). The 35-nucleotide spliced leader sequence is common to all trypanosome messenger RNA's. *Science* **233(4793)**, 569-571. doi:10.1126/science.3523758.

Weiden, M., Osheim, Y.N., Beyer, A.L. and Van der Ploeg, L.H. (1991). Chromosome structure: DNA nucleotide sequence elements of a subset of the minichromosomes of the protozoan Trypanosoma brucei. *Molecular Cell Biology* **11**, 3823-3834. doi:10.1128/mcb.11.8.3823.

Wickstead, B., Ersfeld, K. and Gull, K. (2003). The mitotic stability of the minichromosomes of Trypanosoma brucei. *Molecular and Biochemical Parasitology* **132**, 97-100. doi:10.1016/j.molbiopara.2003.08.007.

Wickstead, B., Ersfeld, K. and Gull, K. (2004). The small chromosomes of Trypanosoma brucei involved in antigenic variation are constructed around repetitive palindromes. *Genome Research* **14**, 1014-1024. doi:10.1101/gr.2227704

Williams, R.O., Young, J.R., and Majiwa. P.A.O. (1982). Genomic environment of 7. brucei VSG genes: presence of a mini-chromosome. *Nature* **299**, 417-421.

World Health Organisation WHO. (2020). Trypanosomiasis, human African (sleeping sickness). *World Health Organization*. Available at: https://www.who.int/news-room/fact-sheets/detail/trypanosomiasis-human-african-(sleeping-sickness)

Yamasaki, S., Suganuma, K., Yamagishi, J., Asada, M., Yokoyama, N., Kawazu, S. and Inoue, N. (2016). Characterization of an epimastigote-stage-specific hemoglobin receptor of Trypanosoma congolense. *Parasites and Vectors* **9(1)**, 299. doi:10.1186/s13071-016-1563-9.

Yang, X., Figueiredo, L.M., Espinal, A., Okubo, E. and Li, B. (2009). RAP1 Is Essential for Silencing Telomeric Variant Surface Glycoprotein Genes in Trypanosoma brucei. *Cell* **137**, 99-109. doi: 10.1016/j.cell.2009.01.037.

Yernaux, C., Fransen, M., Brees, C., Lorenzen, S., Michels, P.A.M. (2006). Trypanosoma brucei glycosomal ABC transporters: identification and membrane targeting. *Molecular Membrane Biology* **23**, 157-172. doi: 10.1080/09687860500460124.

Young, C.J. and Godfrey, D.G. (1983). Enzyme polymorphism and the distribution of Trypanosoma congolense isolates. *Annals of Tropical Medicine and Parasitology* **77**, 467–481. doi:10.1080/00034983.1983.11811740.

Young, R., Taylor, J.E., Kurioka, A., Becker, M., Louis, E.J., Rudenko, G. (2008). 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. doi:10.1186/1471-2164- 9-385.

Zhao, G. P. (2007). SARS molecular epidemiology: a Chinese fairy tale of controlling an emerging zoonotic disease in the genomics era. *Philosophical Transactions of the Royal Society B* **362**, 1063–1081. doi:10.1098/rstb.2007.2034.

Zinoviev, A., Akum, Y., Yahav, T. and Shapira, M. (2012). Gene duplication in trypanosomatids - two DED1 paralogs are functionally redundant and differentially expressed during the life cycle. *Molecular and Biochemical Parasitology* **185 (2)**, 127-136. doi:10.1016/j.molbiopara.2012.08.001.

Zweygarth, E. and Rötcher, D. (1987). The occurrence of Trypanosoma (Nannomonas) simiae in the cerebrospinal fluid of domestic pigs. *Parasitology Research* **73(5)**, 479-480. doi:10.1007/bf00538209.

Zweygarth, E., Mihok, S., Majiwa, P.A.O. and Kaminsky, R. (1994). A new Nannomonas-type trypanosome: Isolation, in vitro cultivation and partial characterisation. *First International Congress of the Parasitology and Tropical Medicine*, 188-190.

Appendices

Appendix A. Comparison of assemblers

Table A1 Comparison of the performance of assembly algorithms 'Flye' and 'Canu' for the assembly of mini-chromosomes. For all species tested, Canu generated at least 2x the number of whole MC contigs and a significantly greater number of partial MC contigs (except for TCs GAM2) than Flye did. Species abbreviations: Tbb = *T. brucei brucei;* Tcf = *T. congolense* Forest; Tcs = *T. congolense* Savannah; Tck = *T. congolense* Kilifi; Tgo = *T. godfreyi;* Tsi = *T. simiae;* Tst = *T. simiae* Tsavo.

Species	PacBio read pool (gbp)	Assembler			
		Flye		Canu	
		No. whole MCs	No. partial MCs	No. whole MCs	No. partial MCs
Tbb 427	Unknown	Not tested	Not tested	8	62
Tcf ANR3	2.1	2	18	12	129
Tcs GAM2	18.1	101	155	207	118
Tck WG84	9.8	4	32	18	263
Tgo KEN7	6.0	0	20	15	248
Tsi TV008	2.5	0	12	1	30
Tsi ERAC2	22.3	1	21	3	25
Tst 114	10.2	0	17	4	70

Appendix B. Statistical outputs for between-taxa differences in MC length

Table A2 Univariate ANOVA results for MC length.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Taxa	7	6.37E+10	9.10E+09	79.48	<2e-16
Residuals	663	7.59E+10	1.15E+08		

Table A3 Statistical outputs of post-hoc Tukey HSD test for pairwise comparisons of MC length between seven trypanosome taxa. Bold text indicates significant p values.

Taxa	diff	lwr	upr	p adj
Tco_ANR3-Tbb_427	-2660.59524	-14253.236	8932.046	0.9970467
Tco_GAM2-Tbb_427	-4826.21073	-12754.021	3101.6	0.5852375
Tco_WG84-Tbb_427	38718.60606	28379.332	49057.881	**0**
Tgo_KEN7-Tbb_427	32840.83333	21663.174	44018.493	**0**
Tsi_ERA_C2-Tbb_427	231.54545	-7693.358	8156.449	1
Tst_Tsavo_114-Tbb_427	-2736.20635	-11430.687	5958.274	0.9800725
Tco_GAM2-Tco_ANR3	-2165.61549	-11090.236	6759.005	0.9958125
Tco_WG84-Tco_ANR3	41379.2013	30257.184	52501.218	**0**
Tgo_KEN7-Tco_ANR3	35501.42857	23596.021	47406.836	**0**
Tsi_ERA_C2-Tco_ANR3	2892.14069	-6029.898	11814.179	0.9764123
Tst_Tsavo_114-Tco_ANR3	-75.61111	-9687.721	9536.499	1
Tco_WG84-Tco_GAM2	43544.81679	36322.621	50767.013	**0**
Tgo_KEN7-Tco_GAM2	37667.04406	29288.524	46045.564	**0**
Tsi_ERA_C2-Tco_GAM2	5057.75618	2218.104	7897.408	**0.0000024**
Tst_Tsavo_114-Tco_GAM2	2090.00438	-2476.561	6656.569	0.8610879
Tgo_KEN7-Tco_WG84	-5877.77273	-16566.552	4811.007	0.7054398
Tsi_ERA_C2-Tco_WG84	-38487.06061	-45706.066	-31268.055	**0**
Tst_Tsavo_114-Tco_WG84	-41454.81241	-49511.11	-33398.515	**0**
Tsi_ERA_C2-Tgo_KEN7	-32609.28788	-40985.057	-24233.518	**0**
Tst_Tsavo_114-Tgo_KEN7	-35577.03968	-44684.367	-26469.712	**0**
Tst_Tsavo_114-Tsi_ERA_C2	-2967.7518	-7529.269	1593.765	0.4974423

Appendix C. Statistical outputs for between-taxa differences in core repeat region length

Table A4 Univariate ANOVA results for core repeat region length.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Taxa	7	7.774e+10	1.111e+10	158.6	<2e-16
Residuals	1423	9.967e+10	7.004e+07		

Table A5 Statistical outputs of post-hoc Tukey HSD test for pairwise comparisons of core repeat region length between seven trypanosome taxa. Bold text indicates significant p values.

Taxa	diff	lwr	upr	p adj
Tco_ANR3-Tbb_427	-8317.001	-12861.201	-3772.8012	**0.0000009**
Tco_GAM2-Tbb_427	-4522.397	-8827.979	-216.8142	**0.0315474**
Tco_WG84-Tbb_427	5370.391	595.7744	10145.0075	**0.0151411**
Tgo_KEN7-Tbb_427	2162.138	-2208.8851	6533.1614	0.8068591
Tsi_ERA_C2-Tbb_427	13958.574	9511.4729	18405.6752	**0**
Tst_Tsavo_114-Tbb_427	6676.155	1956.7538	11395.5568	**0.0004957**
Tco_GAM2-Tco_ANR3	3794.604	1509.4685	6079.74	**0.0000144**
Tco_WG84-Tco_ANR3	13687.392	10608.3061	16766.4775	**0**
Tgo_KEN7-Tco_ANR3	10479.139	8072.97	12885.3082	**0**
Tsi_ERA_C2-Tco_ANR3	22275.575	19733.8208	24817.329	**0**
Tst_Tsavo_114-Tco_ANR3	14993.156	12000.4055	17985.9069	**0**
Tco_WG84-Tco_GAM2	9892.788	7178.1867	12607.3885	**0**
Tgo_KEN7-Tco_GAM2	6684.535	4766.7395	8602.3301	**0**
Tsi_ERA_C2-Tco_GAM2	18480.971	16395.5923	20566.3491	**0**
Tst_Tsavo_114-Tco_GAM2	11198.552	8582.2863	13814.8175	**0**
Tgo_KEN7-Tco_WG84	-3208.253	-6025.4965	-391.0091	**0.0130956**
Tsi_ERA_C2-Tco_WG84	8588.183	5654.2904	11522.0758	**0**
Tst_Tsavo_114-Tco_WG84	1305.764	-2026.4675	4637.9962	0.9350476
Tsi_ERA_C2-Tgo_KEN7	11796.436	9579.0935	14013.7783	**0**
Tst_Tsavo_114-Tgo_KEN7	4514.017	1791.3989	7236.6354	**0.000015**
Tst_Tsavo_114-Tsi_ERA_C2	-7282.419	-10125.571	-4439.2666	**0**

Appendix D. Statistical outputs for between-taxa differences in subtelomeric arm length

Table A6 Univariate ANOVA results for MC arm length.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Taxa	7	2.162e+11	3.088e+10	1198	<2e-16
Residuals	2217	5.713e+10	2.577e+07		

Table A7 Statistical outputs of post-hoc Tukey HSD test for pairwise comparisons of subtelomeric arm length between seven trypanosome taxa. Bold text indicates significant p values.

Taxa	diff	lwr	upr	p adj
Tco_ANR3-Tbb_427	-102.84401	-2516.027	2310.3388	1
Tco_GAM2-Tbb_427	-1631.88838	-3795.179	531.4023	0.3002633
Tco_WG84-Tbb_427	20138.02943	17882.974	22393.0849	**0**
Tgo_KEN7-Tbb_427	15534.04406	13263.193	17804.8949	**0**
Tsi_ERA_C2-Tbb_427	-6322.659	-8502.732	-4142.5863	**0**
Tst_Tsavo_114-Tbb_427	-6390.19593	-8758.487	-4021.9049	**0**
Tco_GAM2-Tco_ANR3	-1529.04437	-2899.437	-158.6513	**0.0165712**
Tco_WG84-Tco_ANR3	20240.87344	18729.773	21751.9743	**0**
Tgo_KEN7-Tco_ANR3	15636.88807	14102.315	17171.461	**0**
Tsi_ERA_C2-Tco_ANR3	-6219.81499	-7616.55	-4823.0804	**0**
Tst_Tsavo_114-Tco_ANR3	-6287.35192	-7962.748	-4611.9556	**0**
Tco_WG84-Tco_GAM2	21769.91781	20702.303	22837.5323	**0**
Tgo_KEN7-Tco_GAM2	17165.93244	16065.347	18266.5183	**0**
Tsi_ERA_C2-Tco_GAM2	-4690.77062	-5589.237	-3792.3039	**0**
Tst_Tsavo_114-Tco_GAM2	-4758.30755	-6048.008	-3468.6069	**0**
Tgo_KEN7-Tco_WG84	-4603.98537	-5875.495	-3332.4754	**0**
Tsi_ERA_C2-Tco_WG84	-26460.68843	-27561.911	-25359.466	**0**
Tst_Tsavo_114-Tco_WG84	-26528.22536	-27966.55	-25089.901	**0**
Tsi_ERA_C2-Tgo_KEN7	-21856.70306	-22989.919	-20723.487	**0**
Tst_Tsavo_114-Tgo_KEN7	-21924.23999	-23387.204	-20461.276	**0**
Tst_Tsavo_114-Tsi_ERA_C2	-67.53693	-1385.193	1250.1194	0.9999999