Genomic insights into *Trypanosoma brucei* and *Leishmania major*: compartmentalised DNA replication, modified base detection by Nanopore sequencing and nucleotide composition analysis

Marija Krasiļņikova

BSc (Hons)

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

School of Infection and Immunity

College of Medical, Veterinary and Life Sciences

University of Glasgow

Abstract

The genomes of the early-diverging eukaryotic parasites *Trypanosoma brucei* and *Leishmania major* are unusual in both their organisation and dynamics. The majority of transcribed genes in both vector-transmitted parasites are organised in polycistronic transcription units (PTUs), and the boundaries between these units act as transcription start and termination sites. The genome of *T. brucei* is further shaped by the parasite's immune evasion strategy – antigenic variation. One variant surface glycoprotein (VSG) is expressed on the parasite's surface at any given time, despite the trypanosome harbouring >2500 of *VSG* (pseudo)genes in its genome. This monoallelic gene expression is achieved whereby only one *VSG* is expressed at a time from a dedicated telomere-proximal site – the bloodstream-form expression site (BES), of which the parasite has ~15. The remainder of *VSG* genes are located in transcriptionally silent arrays in the subtelomeric compartments of the larger, so-called megabase chromosomes of the parasites, as well as the smaller mini- and intermediate chromosomes.

Analysis of DNA replication patterns using marker frequency analysis coupled with sequencing (MFAseq) in *T. brucei* showed that in the core of the larger chromosomes some PTU boundaries, as well as the annotated centromeric regions, co-localise with early S phase DNA replication initiation; this pattern was consistent between two *T. brucei* strains, TREU 927 and Lister 427, as well as between mammalian stage bloodstream-form (BSF) cells and insect-stage procyclic cells (PCF). Curiously, the active BES was also an early-replicating region, but only in the BSF cells; the origin – telomeric or upstream – of this replication could not be determined. In addition, complete, genome-wide analysis of DNA replication dynamics was not possible at the time due to incomplete genome assembly. Until recently, the genome of this parasite remained poorly assembled outside of the megabase chromosome cores, despite its small <50 Mb size. In 2018, a number of major improvements in the genome assembly of *T. brucei* were achieved using PacBio long-read assembly, assisted by Hi-C DNA interaction data, but the chromosome 'core' sequences remained separate from *VSG*-containing subtelomeric sequences and BES. In addition, none of the centromeric regions, which co-localise with the earliest S phase DNA replication sites, have been fully resolved. Moreover, MFA-seq mapping to mini-

and intermediate chromosomes was compromised by much of their content comprising 177 bp repeats.

In chapter 3 we discuss *de novo* long-read assembly of the genome of *T. brucei brucei* Lister 427 using Nanopore sequencing to improve contiguity along chromosome compartments (core, subtelomeres, BES), as well as improve repetitive region and sub-megabase chromosome assembly. This was motivated by a wish to provide a complete understanding of DNA replication, by expanding MFAseq mapping to the subtelomere compartments of the megabase chromosomes, across the core and subtelomere boundaries, across the entirety of the telomeric BESs, and within the submegabase chromosomes. Long-read assembly resulted in at least 1 bridging of previously separate genome sequences in 10/11 megabase chromosomes, overall improved contiguity, as well as assembly of full-length centromeric, 50bp, 177bp and 70bp repeats. Additionally, smaller, sub-megabase chromosomes of this parasite with the characteristic 177bp repeat region were at least partially assembled; the assembly of these smaller chromosomes showed that they contain more genes than previously thought. The improved genome assembly allowed mapping of MFAseq across the various genome compartments, and showed that megabase chromosome subtelomeres, surprisingly, contain no detectable early-replicating regions outside of centromeric repeats. We also show that 177bp repeats act as sites of DNA replication initiation in the submegabase chromosomes and that they are found in centromeric regions of the megabase chromosomes, revealing these repeats to be widespread, sequence conserved origins of replication. In addition, we demonstrate that early replication if the active BES in BSF cells initiates from the telomere. Finally, we show that, in addition to genome compartmentalisation being evident in gene content, organisation, transcription and DNA replication in *T. brucei*, genome stability is also compartmentalised, with the subtelomeric regions of megabase chromosomes showing pronounced genomic instability compared to the cores and, indeed, sub-megabase chromosomes.

Kinetoplastid parasite genomes also harbour an unusual DNA modification – a thymidine modification termed base J. It is thought to be generated in a two-step process – first, a thymidine is modified to 5-hydroxymethyluracil (5hmU), followed by glycosylation. Previous work has indicated that this modification is present in the *T. brucei* genome at very low levels, and that it is primarily detected at repetitive DNA, as well as PTU boundaries. Only mammalian stage parasites appear to harbour easily detectable levels of this base, as in the insect vector stage only very low or no base J has been previously detected. While both base J and 5hmU have previously been mapped genome-wide in *T. brucei*, the existing datasets have not been re-examined in light of improved genome assemblies, nor have they been evaluated together. In addition to offering long-read DNA sequencing, the two main long-read sequencing technologies at the time of writing – Nanopore and PacBio – also possess the ability to detect modified bases in DNA and RNA. In chapter 4, we provide more comprehensive analysis of base J and 5hmU ChIPseq datasets in *T. brucei*, along with newly generated modified base detection data using Nanopore sequencing and Tombo software. We find that Nanopore-generated data recapitulates many base J ChIPseq enrichment patterns, specifically, at polycistronic unit boundaries, in repetitive regions and around coding sequence, while also offering strand-specific and base-resolution data. To our surprise, insect and mammalian stage parasites show similar levels and patterns of DNA modification based on Nanopore data, arguing that the two lifecycle stages may, in fact, have similar DNA modification distribution patterns.

In our analysis of modified DNA distribution in Chapter 4, it became clear that broader canonical base distribution genome-wide has not been fully described in *T. brucei*. DNA strand asymmetries in prokaryotes and eukaryotes often arise as a consequence of directional processes, such as transcription and DNA replication, as these processes are asymmetrical with regards to the two DNA strands. This often leads to overabundances of certain nucleotides on one DNA strand relative to the other, termed nucleotide skews. In many eukaryotic and prokaryotic genomes, analysis of nucleotide skews can even be used to detect origins in DNA replication. In chapter 5 we present efforts in describing the nature, and elucidating the potential contributors of nucleotide skews in *T. brucei*, and comparing these skews to those observed in *L. major*, as well as a broad range of trypanosomatid genomes. We found that *T. brucei* and *L. major* display clear and distinct skews associated with transcription direction, but similar skews associated with DNA replication, meaning differential processes are responsible for the distinct nature of transcription-associated skews. Additionally, through the inclusion of analysis in other trypanosomatids, we show that nucleotide composition and skew differences observed in *T. brucei* and *L. major* can be explained through their evolutionary divergence.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I am deeply grateful to my supervisor Richard McCulloch for his unwavering support, understanding and encouragement throughout my PhD. Despite all the setbacks encountered on the way and my frustratingly slow progress with writing, Richard's remarkable patience and understanding have been essential. I consider myself incredibly fortunate to have had Richard as my supervisor.

I would also like to thank my assessors Kathryn Crouch and Jane Munday for taking the time to go through my work with me and offering support when it was most needed.

Kathryn Crouch and Dario Beraldi both helped me immensely with data analysis, and without their help and expertise I would have been quite lost.

Jeziel and Cat for all the suggestions they've made and the endless questions of mine they've answered.

Past and present colleagues and members of the McCulloch lab for creating an amazing environment that I would love to stay in for longer.

Lastly, my partner Joe for not letting me despair and for accepting my non-participation in trips and holidays due to work demands. To Anna for her support and much-needed positive distractions. And to Hercules for his warm, calming and purring presence in the evenings.

Author's Declaration

I declare that, except where explicit reference is made to the contribution of others, that this dissertation is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Marija Krasiļņikova

Abbreviations

2D-TLC	two-dimensional thin layer chromatography
5hmU	5-hydroxymethyluracil
AAT	Animal African trypanosomiasis
BES	bloodstream-form expression site
BLAST	basic local alignment search tool
BSF	bloodstream-form
CDS	coding sequence
ChIPseq	chromatin immunoprecipitation coupled with sequencing
CNS	central nervous system
CNV	copy number variation
cSSR	convergent strand-switch region
dJ	base J, ß-D-glucosyl-hydroxymethyluracil
DMOG	dimethyloxalylglycine
dNTP	deoxynucleotide triphosphate
DRIPseq	DNA-RNA hybrid immunoprecipitation coupled with sequencing
DSBR	double-stranded break repair
dsDNA	double-stranded DNA
dSSR	divergent strand-switch region
dT	deoxythymidine
ESAG	expression site associated gene
ESB	expression site body
ESB1	expression site body-specific protein 1
FACS	fluorescence-activated cell sorting
G4	g-quadruplex secondary structures
GO	gene ontology
H3	histone 3
H3.V	histone 3 variant
H4K10ac	histone 4 acetylation of lysine 10
HAT	Human African trypanosomiasis
HOMedU	5-hydroxymethyluracil
HR	homologous recombination
htSSR	head-to-tail strand-switch region
htTTS-TSS	head-to-tail transcription termination site and transcription start site
IPD	inter-pulse duration
JGT	J-specific glycosyltransferase
kDNA	kinetoplast DNA
KKT	kinetoplastid kinetochore protein
LC/MS-MS	liquid chromatography with tandem mass spectrometry
MES	metacyclic expression site
MFAseq	marker frequency analysis coupled with sequencing

MYA	million years ago
NGS	next generation sequencing
ONT	Oxford Nanopore Technologies
ORC	origin recognition complex
ori	origin of replication
PCF	procyclic-form
PFGE	pulsed field gel electrophoresis
PDS	pyridostatin
PTU	polycistronic transcription unit
QC	quality control
RDC	read depth coverage
RHS	retrotransposon hotspot protein
RNAP II	RNA polymerase II
RNaseH1	ribonuclease H1
ROI	region(s) of interest
RPKM	reads per kilobase per million reads mapped
SL	spliced leader
SNSseq	short nascent strand sequencing
ssDNA	single-stranded DNA
SSR	strand-switch region
TC-NER	transcription-coupled nucleotide excision repair
TSS	transcription start site
TTS	transcription termination site
UTR	untranslated region
VSG	variant surface glycoprotein
WCIP	Wellcome Centre for Integrative Parasitology
WHO	World Health Organisation
WT	wild-type

1 Introduction

1.1 Trypanosomatids

Eukaryotic parasites *Trypanosoma brucei* and *Leishmania major* belong to the protozoan family Trypanosomatida and order Kinetoplastea (Table 1). This diverse group of organisms shares a distinctive feature: the kinetoplast – an unusual mitochondrial organelle housing the cell's mitochondrial genome, known as the kinetoplast DNA (kDNA). The kDNA is uniquely arranged in a condensed, disc-like network composed of thousands of interlocked circles of DNA, all housed within a single organelle (Shapiro and Englund, 1995). Kinetoplastids are a part of the Discoba grouping – an early-diverging group of eukaryotes (Figure 1) (Al Jewari and Baldauf, 2023), and they comprise by far the most well-studied organisms within Euglenozoa due to their medical and veterinary importance (Kostygov *et al.*, 2021). Both *T. brucei* and *L. major* are vector-transmitted agents of disease causing Human and Animal African trypanosomiases (HAT and AAT) and leishmaniases, respectively. The Kinetoplastids are, however, a very diverse group of organisms, both free-living and parasitic (reviewed in Kostygov *et al.*, (2021)).

Table 1 Taxonomic classification of *T. brucei* and *L. major* (Simpson, Stevens and Lukeš, 2006; Burki *et al.*, 2020).

Domain:	Eukaryota	
Supergroup:	Excavata	
Group:	Discoba	
Phylum:	Euglenozoa	
Class:	Kinetoplastea	
Order:	Trypanosomatida	
Family:	Trypanosomatidae	
Genus:	Trypanosoma	Leishmania
Species:	Trypanosoma brucei	Leishmania major

The African trypanosome *Trypanosoma brucei* is endemic to sub-Saharan Africa and is transmitted through the bite of haematophagous tsetse flies belonging to the *Glossina* genus (Figure 2). Once in the mammalian host, the non-replicative metacyclic trypomastigotes transform into proliferating long slender forms (Langousis and Hill, 2014). These parasites establish and maintain the extracellular infection in the bloodstream, as well as other blood-derived fluids such as lymph and interstitial fluids in the skin, adipose tissue and major organs such as kidney and heart, among others (Capewell *et al.*, 2016; Trindade *et al.*, 2016). Density-dependent mechanisms described to be similar to quorum sensing in bacteria initiate the differentiation process into cell cycle arrested non-replicating short stumpy cells (Briggs *et al.*, 2021; Matthews, 2021). These can differentiate into procyclic trypomastigotes in the tsetse fly midgut following another bloodmeal from the infected host (Langousis and Hill, 2014). In the fly, a midgut infection is established, followed by migration to the proventriculus where asymmetric cell division leads to differentiation into short and long epimastigotes. Short epimastigotes replicate in the salivary glands, attached to the epithelial wall, and further asymmetric division leads to the emergence of metacyclic trypomastigotes that reside freely in the salivary gland lumen and are adapted for survival in the mammal (Langousis and Hill, 2014). A non-obligatory sexual stage with meiosis and the generation of haploid gametes may take place in the salivary glands of the tsetse fly (Peacock *et al.*, 2011, 2014).

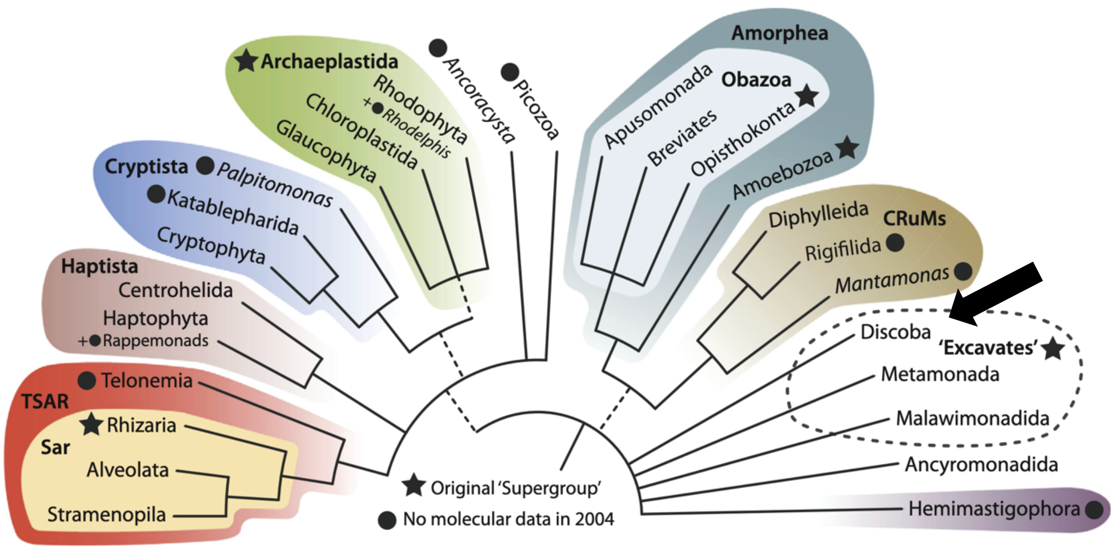
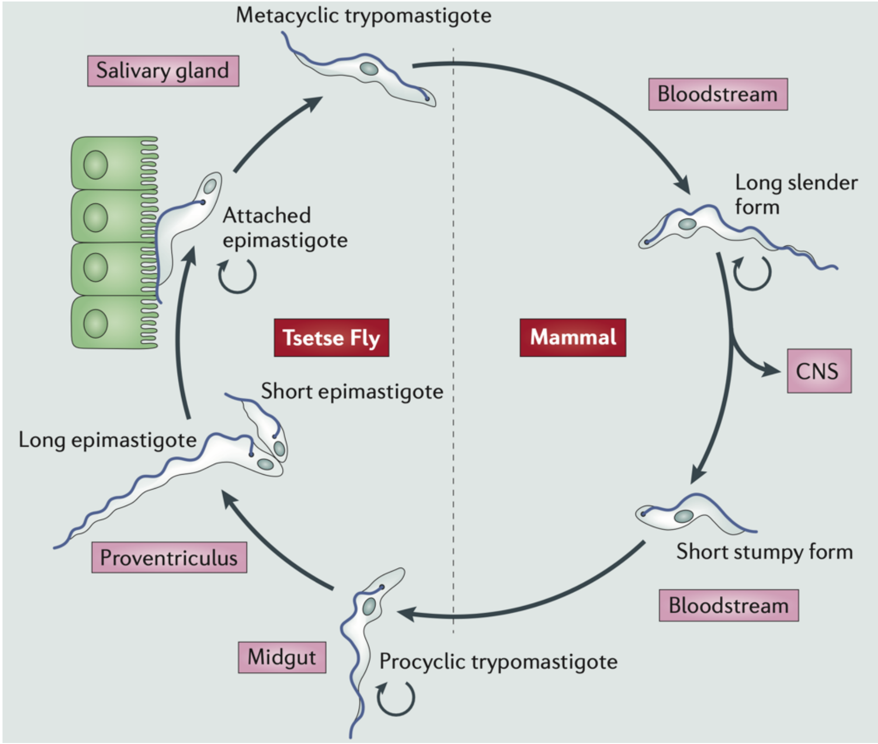


Figure 1 The eukaryotic tree based on recent phylogenetic studies.
Coloured groupings in this diagrammatic representation represent supergroups, with star symbols highlighting taxa that used to be considered as supergroups in previous iterations of the tree. Dashed lines indicate uncertainty about monophyly, and unresolved branching orders are depicted with multifurcations, while circles highlight taxa that lacked molecular data for phylogenetic placement in the previous iteration of this tree in 2004 (Simpson and Roger, 2004; Burki *et al.*, 2020). The location of Discoba, to which trypanosomatids belong, is indicated using a black arrow. Figure from Burki *et al.*, 2020 (reproduced with permission, license number 1590998-1).

Similarly to a *T. brucei* infection, non-dividing procyclic *Leishmania* promastigotes enter the mammalian host through the bite of sandflies (Kaye and Scott, 2011). However, unlike the trypanosome, *Leishmania* parasites do not reside extracellularly – instead they are phagocytosed by host cells. It is here, in the phagolysosome, that the parasites remain, differentiating into replicative amastigotes; as host cells accumulate parasites and subsequently rupture, they release amastigotes that can infect nearby phagocytes. During another bloodmeal, sandflies ingest infected phagocytes, and amastigotes begin differentiation into promastigotes in the insect host (Kaye and Scott, 2011). *Leishmania* parasites have a wide global distribution, ranging from Central and South America, to North Africa, Southern Europe and the Middle East, through to parts of Asia (Akhoundi *et al.*, 2016).

As mentioned above, both parasites cause disease in humans and other mammals, including livestock, contributing to morbidity and economic burden, particularly to some of the poorest and underprivileged communities of the world. While the reported incidence and disease burden of HAT, which is caused by two subspecies of *T. brucei* – *T. b. gambiense* and *T. b. rhodesiense,* has fallen substantially over the past two decades, with fewer than 1000 cases of HAT reported by the World Health Organization (WHO) in 2018 (compared to over 35 000 cases in 1998) (Gao *et al.*, 2020), AAT prevalence has not reduced significantly between the first and second decades of this century, affecting, on average, 13-20% of cattle in endemic areas (systemic review - Okello *et al.*, (2022)). As for leishmaniasis, the estimated annual incidence of 0.7 to 1 million new human infections contributes to significant disease burden, particularly in malnourished and socioeconomically disadvantaged communities (World Health Organization, 2023).

A



B

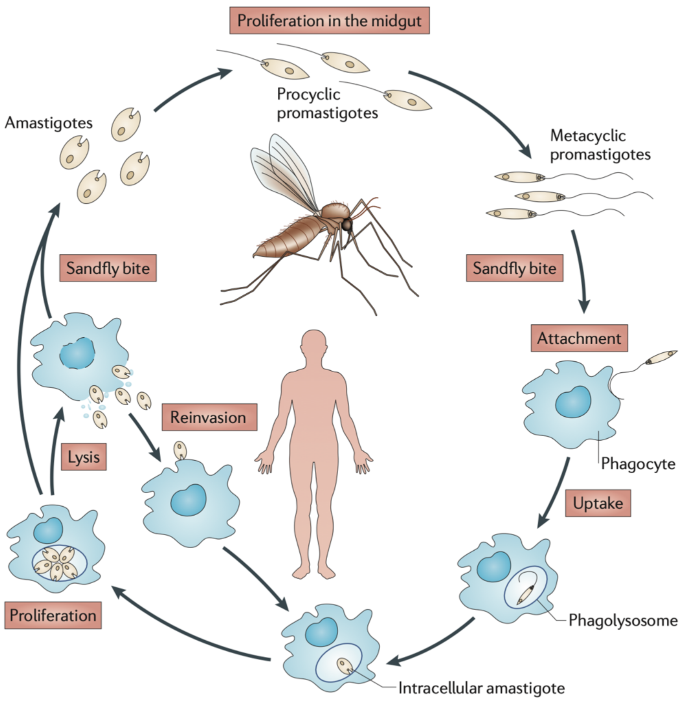


Figure 2 Lifecycles of *Trypanosoma brucei* (A) and *Leishmania major* (B).

A. Thicker arrows indicate the direction of life cycle progression, small circular arrows indicate division by binary fission. CNS – central nervous system. B. Arrows indicate the direction of life cycle progression. Figures taken from Langousis and Hill, 2014 (A, reproduced with permission, license number 1590993-1) and Kaye and Scott, 2011 (B, reproduced with permission, license number 5993701391917).

1.2 Unusual nuclear genome organisation, transcription and replication

Perhaps, due to their early divergence from most studied model organisms, kinetoplastids possess unusual nuclear genomes – both in terms of their organisation and function. Broadly diploid (but see below) and with modest genome sizes below 50 Mb (Berriman *et al.*, 2005; Ivens *et al.*, 2005), *T. brucei* and *L. major* are highly syntenic, sharing a total of 110 blocks of synteny, which correspond to 19.9 and 30.7 Mb, respectively (El-Sayed, Myler, Blandin, *et al.*, 2005) (Figure 3 A). Homologous genes share 40-45% amino acid identity between *L. major* and *Trypanosoma* (Ghedin *et al.*, 2004). Despite this similarity, some unusual aspects of their genomes are common between the parasites, while others differ.

The majority of transcribed genes in both *T. brucei* and *Leishmania*, and perhaps in all kinetoplastids (Jackson *et al.*, 2016; Schmid-Hempel *et al.*, 2018), are organised in polycistronic transcription units (PTUs) that are transcribed by RNA polymerase II in one direction from a single upstream transcription initiation site (Wedel *et al.*, 2017; Cordon-Obras *et al.*, 2022). The boundaries between PTUs are often called strand switch regions (SSRs), and adjacent PTU boundaries can be arranged in three ways depending on transcription direction. Convergent SSRs contain PTU termination sites, divergent SSRs – two transcription start sites , and head-to-tail PTU boundaries contain both a transcription start and termination site (Figure 3 B). Unlike in bacteria, which also display polycistronic gene organisation, the PTUs of kinetoplastids are not arranged by gene function (Berriman *et al.*, 2005). Curiously, when looking at synteny between *T. brucei* and *L. major* genomes, 43% synteny breaks occur at or close to SSRs (El-Sayed *et al.*, 2005).

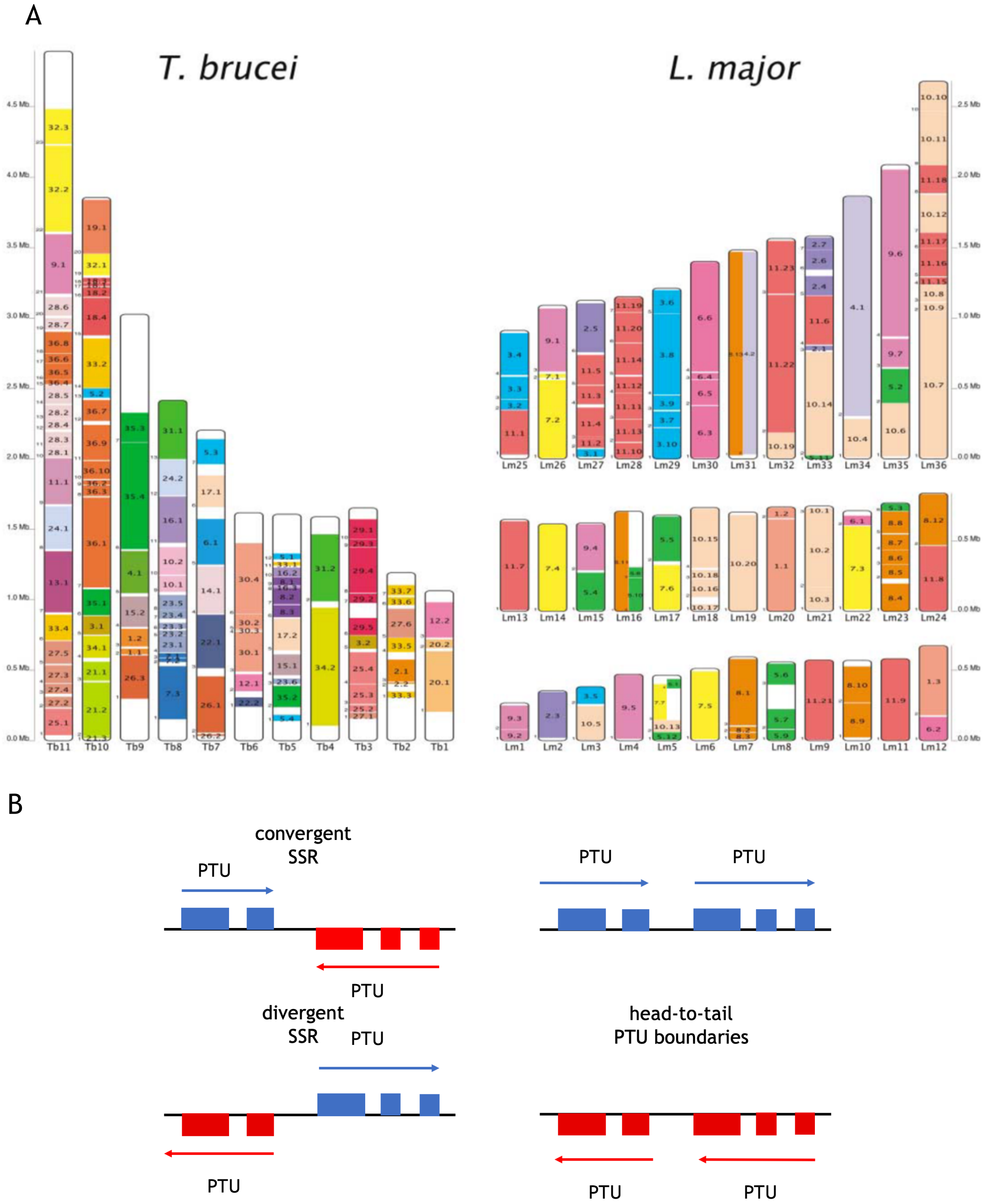


Figure 3 Synteny and gene organization in *T. brucei* and *L. major*.

A. Synteny blocks shared by *T. brucei* and *L. major* nuclear genomes, specifically, the first published reference genomes (Berriman *et al.*, 2005; Ivens *et al.*, 2005). Colours and numbers indicate the chromosome on which the syntenic sequence is found in the other organism; for example, yellow blocks in the *L. major* chromosomes (labelled 7.1 to 7.7) can be found on chromosome 7 of *T. brucei* (Tb7). A total of 7974 *T. brucei* and 7466 *L. major* genes, organized in 110 blocks, are represented here. Figure reproduced with permission from El-Sayed *et al.*, 2005 (license number 1591503-1). B. Diagram representing the organization of polycistronic transcription units (PTUs) in *T. brucei* and *L. major*. Blocks represent individual genes, blue colour represents the top (forward) strand and red – the reverse (bottom) strand. Arrows show the direction of transcription of each PTU. SSR – strand switch unit. Convergent SSRs contain two transcription termination sites (TTS), divergent SSRs – two transcription start sites (TSS), and a head-to-tail boundary is mixed, containing both a TSS and a TTS. Adapted from Siegel *et al.*, (2011).

As a PTU is transcribed from the transcription start site (TSS) to the transcription termination site (TTS) at the next PTU boundary, it is thought that most gene expression control in these parasites is post-transcriptional (reviewed in Clayton, (2019)). Trans-splicing of a short 5' spliced leader (SL) sequence and polyadenylation are responsible for the production of individual gene mRNAs from nascent RNA, and almost all genes are intronless, composed of just one exon/CDS (Berriman *et al.*, 2005; Ivens *et al.*, 2005; Clayton, 2019). While the precise determinants of TSS or TTS are unknown in these parasites, various epigenetic marks have been associated with these regions in *T. brucei*. At polycistronic TSS, modified or variant histones H4K10ac, H2AZ and H2BV (Siegel *et al.*, 2009) are enriched, as well RNA-DNA hybrids, specifically- R-loops (Briggs, Hamilton, *et al.*, 2018), whereas at TTS modified histones H3V and H4V (Siegel *et al.*, 2009; Schulz *et al.*, 2016), along with cohesin (Müller *et al.*, 2018) and a modified nucleobase base J (beta-D-glucosyl-hydroxymethyluracil) (Schulz *et al.*, 2016), are enriched.

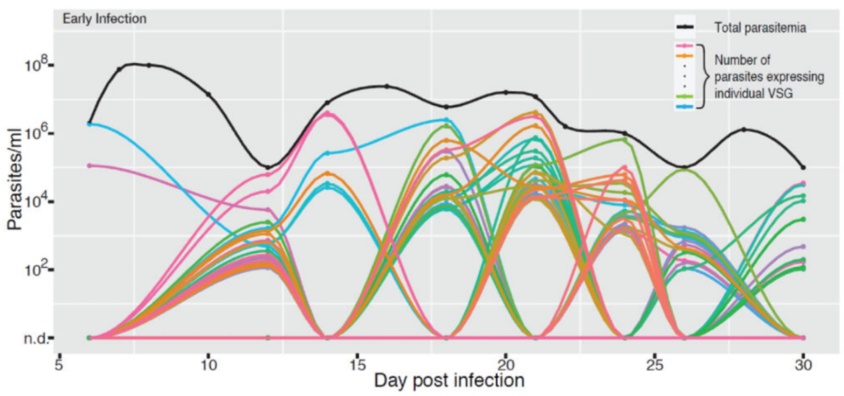
1.2.1 Antigenic variation and genomic organisation of *Trypanosoma brucei*

Genome organisation of *T. brucei* is, arguably, shaped by a core and unique aspect of its biology – antigenic variation. As the parasite is extracellular within the mammalian host, one of the key mechanisms of immune evasion it deploys is antigenic variation – specifically, changing of its dense protein 'coat' by switching from expressing one variant surface glycoprotein (*VSG*) to another, creating waves of parasitaemia dominated by specific *VSGs* (Figure 4 A) (Cross, 1978; Morrison *et al.*, 2005; Mugnier, Cross and Papavasiliou, 2015).

Only one *VSG* is expressed at a time through the mechanism of monoallelic expression, and this occurs in a dedicated genomic compartment – the bloodstream-form expression site (BES), of which the parasite has ~ 15 (Figure 4 B). Only one BES is transcriptionally active at a time; while transcription appears to initiate at all BES, elongation and production of mature messenger RNA (mRNA) only occurs in the active one (Vanhamme *et al.*, 2000; Kassem, Pays and Vanhamme, 2014). The active BES, and the *VSG* in particular, show high chromatin interaction with the SL locus (Faria *et al.*, 2021), and associates with an extranucleolar structure in the cell - the expression site body (ESB), that also contains RNA polymerase I and is thought to act as a transcriptional hub for *VSG* expression (Navarro and Gull, 2001), as well as a spatial separator of active and silent BES, with the latter found at nuclear periphery instead (Chaves *et al.*, 1998; Navarro and Gull, 2001; Landeira and Navarro, 2007). While the full details of the determinants and mechanisms of *VSG* monoallelic expression in *T. brucei* remain to be uncovered, some of the control elements have recently been discovered (reviewed by Faria *et al.*, (2022) and Barcons-Simon, Carrington and Siegel, (2023)). ESB-specific protein 1 (ESB1), in particular, has been shown to act as a developmentally regulated BES transcriptional regulator (López-Escobar *et al.*, 2022), and may prove to be one of the key determinants in BES activation or repression in the future.

BES are 40-90 kb long regions located at the ends of chromosomes, immediately upstream of telomeric repeats (Hertz-Fowler *et al.*, 2008; Müller *et al.*, 2018). Unlike most PTUs in the parasite's genome, BES are transcribed by RNA polymerase I (Günzl *et al.*, 2003) and contain between 2 and 17 genes and pseudogenes - expression site associated genes (ESAGs) and *VSG* genes and pseudogenes (Figure 4 B) (Hertz-Fowler *et al.*, 2008; Müller *et al.*, 2018). Repetitive elements are also a prominent part of the BESs – with so-called 50bp repeat regions just upstream of the expression site (Glover *et al.*, 2013), 70bp regions immediately upstream of the telomere-proximal *VSG* (and thus separating the *VSG* from *ESAGs*), and the telomeric TTAGGG downstream of the distal *VSG*.

A



B

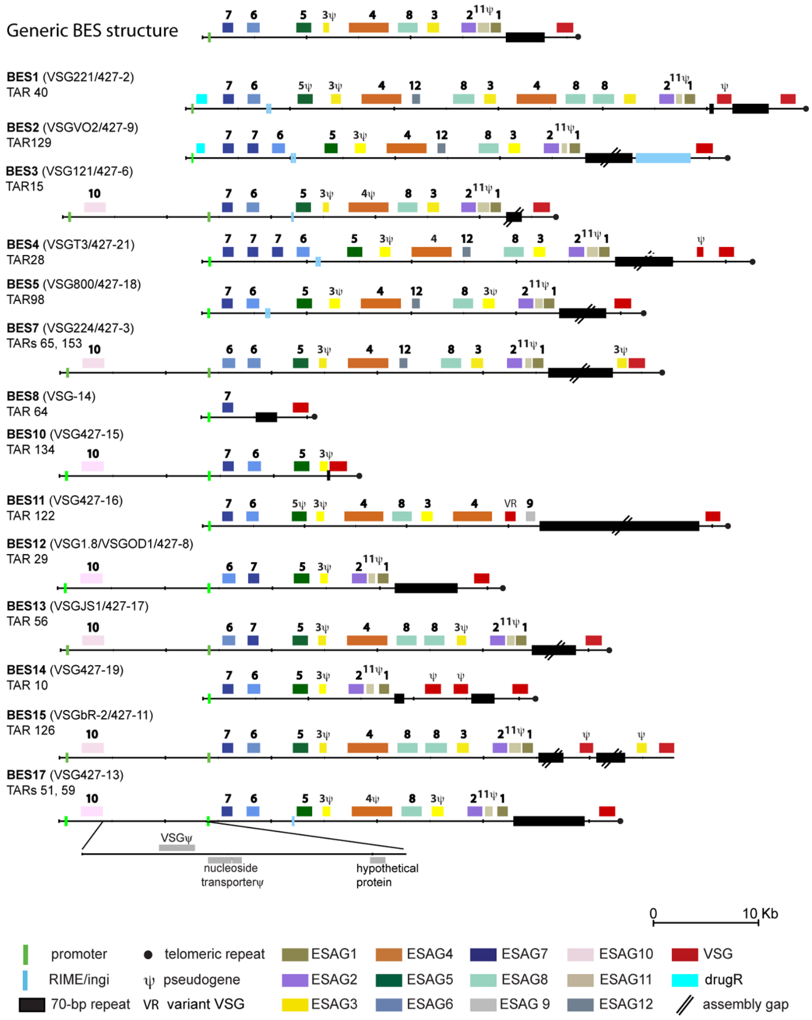


Figure 4 *T. brucei* variant surface glycoprotein (VSG) expression dynamics and expression sites.

A – dynamics of *VSG* expression *in vivo* during a mouse infection (*T. brucei* strain EATRO1125). Each coloured line represents the number of cells expressing a particular *VSG*, data for one mouse shown. Figure reproduced with permission, Mugnier, Cross and Papavasiliou, (2015) (license number 1591507-1). B – overview of *T. brucei* bloodstream-form expression site (BES) structure (*T. brucei* strain Lister 427). Figure source - Hertz-Fowler *et al.*, (2008) (reproduced under Open Access, Creative Commons CC BY 4.0).

Only one of the BES is transcriptionally active at a time; *VSG* switching occurs either by transcriptional activation of another BES, or homologous recombination - the translocation of another *VSG* into an active BES (Figure 5).

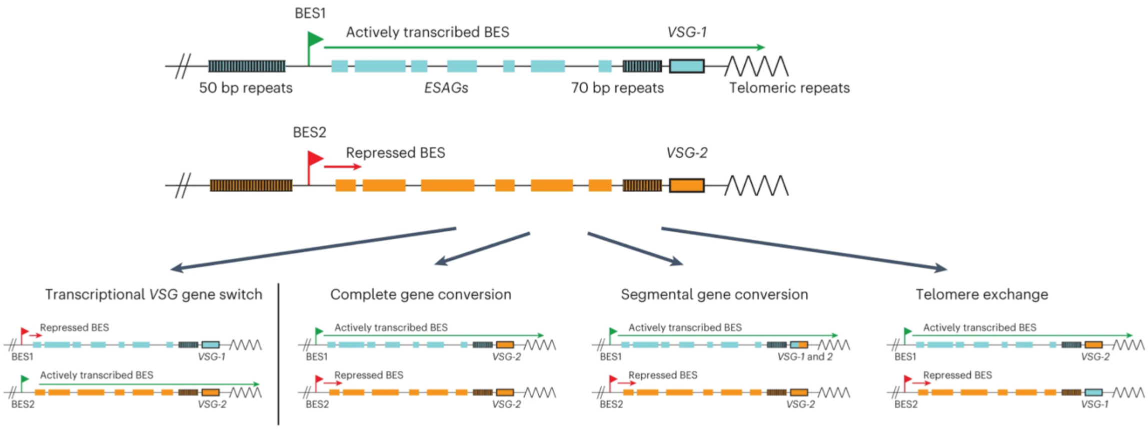


Figure 5 Mechanisms of variant surface glycoprotein (*VSG*) switching in *Trypanosoma brucei.*

A diagram showing the main mechanisms by which *VSG* switching occurs in *T. brucei*. In transcriptional VSG switching, a different bloodstream-form expression site (BES) becomes transcribed while the previously active one is silenced. Three recombinational mechanisms can result in VSG switching as well – complete gene conversion where just the *VSG* gene is inserted, segmental gene duplication where only a fragment of *VSG* is inserted, and telomere exchange, where two chromosome ends are swapped between two chromosomes (reviewed in McCulloch, Morrison and Hall, (2015)). Figure from Barcons-Simon, Carrington and Siegel, (2023) (reproduced with permission, license number 1591507-1).

The *T. brucei* genome contains an extensive archive of silent *VSGs* - the most complete reference genome contains over 2500 VSG genes and pseudogenes (strain Lister 427), most of them in dedicated genomic compartments – the subtelomeric sequences of megabase chromosomes, as well as mini- and intermediate chromosomes of this parasite (see below) (Berriman *et al.*, 2005; Marcello and Barry, 2007; Cross, Kim and Wickstead, 2014; Müller *et al.*, 2018). In addition to the wide selection of existing *VSG* (pseudo)genes, *T. brucei* is also capable of expanding its antigenic repertoire by generating novel mosaic *VSG* sequences via segmental gene conversion, resulting in antigenically distinct *VSGs* (Hall, Wang and Barry, 2013).

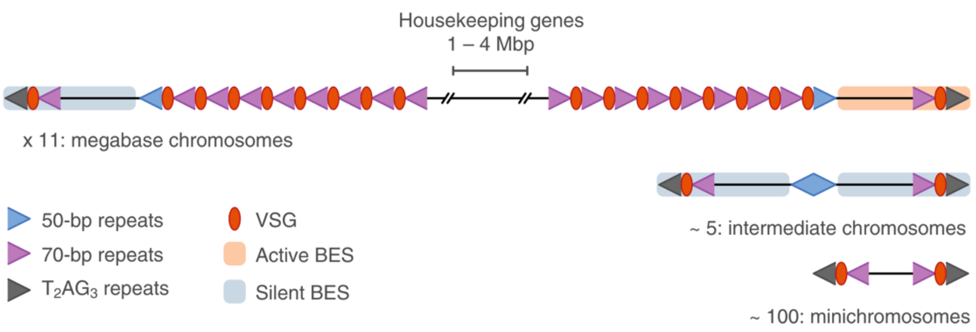
In addition to BES, the parasite also has ~5 dedicated metacyclic expression sites (MES) (highlighted in Figure 7 B)– these lack *ESAGs* and 70bp repeat regions, and are much shorter than BES (3-6 kbp). MES can be utilised, similarly to BES, by BSF cells, though, more commonly, MES are utilised by metacyclic cells in the salivary glands of the vector in preparation for infection (Kolev, Günzl and Tschudi, 2017).

1.2.1.1 Chromosomes and compartmentalisation of the *T. brucei* genome

The chromosomes of *T. brucei* have been historically categorised into three broad groups based on their size, somy, and functional distinction (Figure 6). Genome assemblies of this parasite have focused on so-called megabase chromosomes – 11 pairs of larger, 1-6 Mb chromosomes (Figure 7). The core part of these chromosomes contains most expressed genes, and this region is conserved between the two chromosome copies, with 4.2 variants per kilobase on average (Cosentino, Brink and Siegel, 2021); the more distal parts of the chromosomes, called subtelomeres, are transcriptionally inactive and contain arrays of *VSG*, *ESAG* and *RHS* (pseudo)genes, among others. Unlike the core region of these larger chromosomes, the subtelomeres are transcriptionally silent (Figure 7B), with more compacted chromatin conformation and no obvious polycistron-like organisation (Müller *et al.*, 2018). Interestingly, the subtelomeres of a given chromosome pair are hemizygous and can vary considerably in length and sequence (Figure 7) (Berriman *et al.*, 2005; Callejas *et al.*, 2006; Müller *et al.*, 2018).

The other two groups of *T. brucei* chromosomes have been historically understudied, perhaps even overlooked. The parasite is thought to have a small number of intermediate chromosomes (~5) and a much larger collection of minichromosomes (~100) – these estimates are based on band intensities of pulse field gel electrophoresis (Figure 6 B) (Wickstead, Ersfeld and Gull, 2004). Intermediate chromosomes are thought to be in the 150-500 kb size range, minichromosomes – 50-150 kb (El-Sayed *et al.*, 2000). Both groups are characterised by the presence of 177bp element tandem repeats and are thought to contain only repetitive sequence and *VSG* genes, though intermediate chromosomes can also have BESs. While there have been some limited descriptions of these non-megabase chromosomes over the years (Wickstead, Ersfeld and Gull, 2004; Cross, Kim and Wickstead, 2014), they have never been fully assembled or made it onto the reference genomes of this parasite (Berriman *et al.*, 2005; Müller *et al.*, 2018). As a result, no experimental evidence has been provided to explain how these chromosomes are duplicated during growth.

A



B

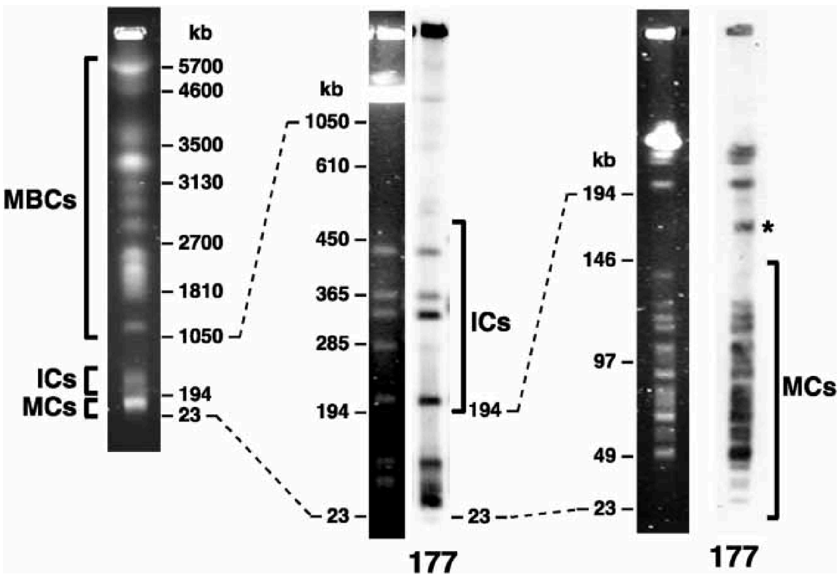


Figure 6 Chromosome types of *T. brucei* and their associated repeats.

A – a diagrammatic representation of *T. brucei* chromosomes. Mbp – megabasepairs, VSG – variant surface glycoprotein genes, BES – bloodstream-form expression sites, T_2AG_3 – telomeric repeats. Figure source - Glover *et al.*, (2013) (reproduced under Open Access, Creative Commons 3.0 CC BY license). B – the karyotype of *T. brucei* as assessed by pulsed-field gel electrophoresis. MBCs – megabase-sized chromosomes, ICs – intermediate chromosomes, MCs - minichromosomes. Hybridization to 177bp repeat sequence indicated with '177'. * denotes a circular extrachromosomal element of ~ 70kb. Figure source - Wickstead, Ersfeld and Gull, (2004), reproduced under Open Access, Creative Commons CC-BY-NC 4.0 license.

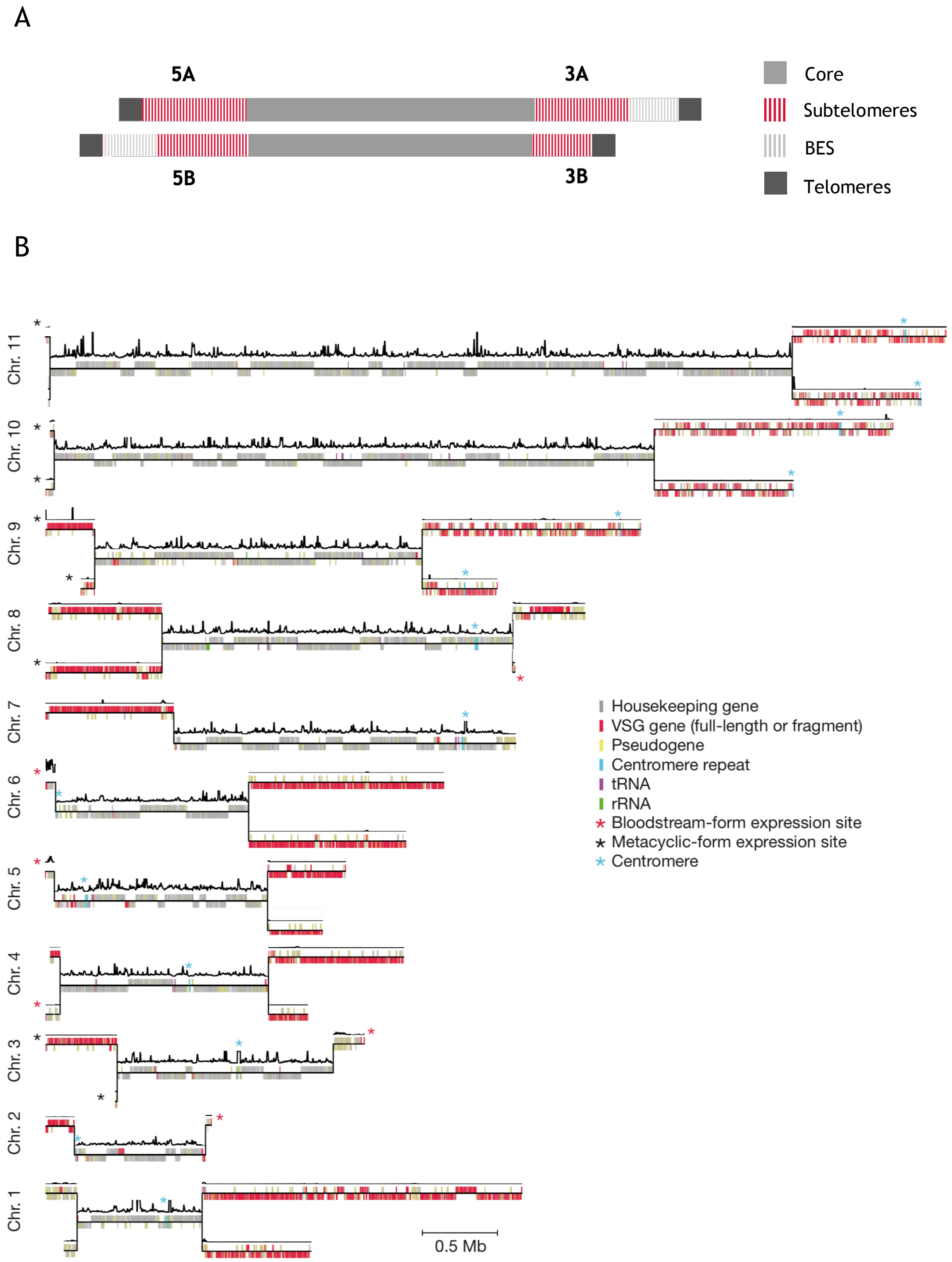


Figure 7 Organization of megabase chromosomes of *Trypanosoma brucei*.

A – a diagrammatic depiction of a pair of megabase chromosomes, highlighting general genomic regions. 5A, 5B, 3A and 3B refer to the variable subtelomeric arms of the chromosomes, BES – bloodstream-form expression site. A pair of megabase chromosomes may or may not have BES sequences, and the relative lengths of the genomic compartments varies. Adapted from Müller *et al.*, (2018**)**. B – Structure, content and transcription levels of the 11 megabase chromosome pairs; heterozygous subtelomeric chromosome arms depicted for both copies of the chromosome, and core only for one copy. Relative transcript levels shown as a black line above the chromosomes. Figure source - Müller *et al.*, (2018) (reproduced under Open Access, Creative Commons CC BY 4.0 license).

The *T. brucei* genome is distinctly compartmentalised, both in terms of function and content. In addition to structural, functional and transcriptional differences between these genomic compartments (discussed above), chromatin conformation capture (Hi-C) indicated that core-subtelomere boundaries, as well as centromeres, act as most prominent intra-chromosomal DNA-DNA interaction boundaries in *T. brucei* megabase chromosomes (Müller *et al.*, 2018), suggesting spatial and epigenetic separation of these compartments. The same analysis also indicated the *T. brucei* subtelomeres are also more compacted compared to the core genome (Müller *et al.*, 2018), likely reflecting the transcriptionally inactive state of these genomic regions.

1.2.2 Genome organisation of *L. major*

The *Leishmania* genomes are arguably arranged in a more conventional fashion compared to *T. brucei*. *Leishmania major*, specifically, has 36 chromosome pairs that range in size from 269 kb to 2.7 Mb (Figure 3 A). Within these chromosomes there is no evidence for compartmentalisation or any largely non-transcribed regions, like those of subtelomeres in *T. brucei*, and the chromosome pairs show fewer heterozygous polymorphisms than those of *T. brucei* and *T. cruzi* (Ivens *et al.*, 2005). In addition, there is no evidence of mitotically stable equivalents of the intermediate and minichromosomes seen in *T. brucei*. However, unlike *T. brucei* (Almeida *et al.*, 2018; Reis-Cunha *et al.*, 2024), most likely all *Leishmania* species are prone to genome-wide content variation. One form of this variation is mosaic aneuploidy – between different species, cell populations and individual cells one or more chromosome can be often be found to be non-disomic (Rogers *et al.*, 2011; Sterkers *et al.*, 2011; Lachaud *et al.*, 2014). In addition to varying chromosome copy number, copy number variation (CNV) of genes and larger chromosomal regions are also frequently observed in *Leishmania* (Ubeda *et al.*, 2008; Laffitte *et al.*, 2016; Potvin *et al.*, 2023). A further manifestation of genomic plasticity in *Leishmania* is the presence of extrachromosomal DNA – many circular and linear extrachromosomal DNA fragments are often found in *Leishmania* parasites, and these are thought to serve as a dynamic source of variation that can be utilised in situations where selective pressure is applied, such as in response to drug pressure (Laffitte *et al.*, 2016).

Such genomic plasticity is thought to serve as an adaptive mechanism for these parasites in order to alter gene expression patterns in response to environmental stimuli (reviewed in Reis-Cunha, Valdivia and Bartholomeu, (2018)). The mechanisms behind genome plasticity in *Leishmania spp.*, as well as what functionally differentiates *Leishmania* parasites from *T. brucei* in this regard (Almeida *et al.*, 2018), remains unclear.

1.2.3 DNA replication in *T. brucei* and *L. major*

DNA replication is a fundamental cellular process that plays a critical role in the survival, functioning, and faithful transmission of genetic material across diverse organisms; it is a strictly regulated process that involves the duplication of genetic material exactly once during the S phase of the cell cycle (Burgers and Kunkel, 2017). DNA replication initiates at specific loci termed origins of replication; in small bacterial genomes typically one origin of replication per chromosome is found (Leonard and Méchali, 2013), whereas most eukaryotic organisms possess several origins of replication per chromosome, likely due to the large size of many eukaryotic genomes (Hu and Stillman, 2023).

In *T. brucei*, DNA replication dynamics have been assayed using marker frequency analysis coupled with sequencing (MFAseq) in the most popular lab strains of *T. brucei brucei* – TREU927 and Lister 427 (Tiengwe, Marcello, Farr, Dickens, *et al.*, 2012; Devlin *et al.*, 2016). All of the detectable origins of replication in *T. brucei* are located at PTU boundaries, indicating that PTU boundaries act in both transcription and replication. Due to the limitations of genome assemblies at the time, nothing is known about replication initiation sites or dynamics more broadly in the subtelomeres of the parasite. Chromatin immunoprecipitation of a key origin-recognition complex (ORC) protein in *T. brucei* - ORC1/CDC6 - mapped to PTU boundaries, particularly to transcription start sites, as well as other loci that were within relative proximity of PTU boundaries (Tiengwe, Marcello, Farr, Dickens, *et al.*, 2012). Curiously, RNAi knockdown of ORC1/CDC6 affects mRNA levels in the core genome and BES, indicating a link between replication initiation and transcription (Benmerzouga *et al.*, 2013). In the limited subtelomeric sequences available in the reference genome at the time, ORC1/CDC6 binding was more dispersed and frequent compared to the core, particularly within 5’ upstream regions and ORF sequences of *VSGs* (Tiengwe, Marcello, Farr, Dickens, *et al.*, 2012), pointing to the possibility of distinct replication dynamics in the core vs subtelomeres of *T. brucei*. Furthermore, the transcriptionally active BES in *T. brucei* was found to be early-replicating, but only in bloodstream-form cells – not procyclic cells (Figure 8); it wasn't possible to discern at the time which direction the active BES was replicated from – the upstream region, from within the BES or from the downstream telomere (Devlin *et al.*, 2016). While it is unclear how this BES-specific and lifecycle stage-specific DNA replication process is orchestrated, this observation further highlights the distinct compartmentalisation and transcription-associated nature of DNA replication in *T. brucei*. As noted above, whether compartmentalisation of DNA replication extends to the megabase chromosomes relative to the mini- and intermediate chromosomes is unknown.

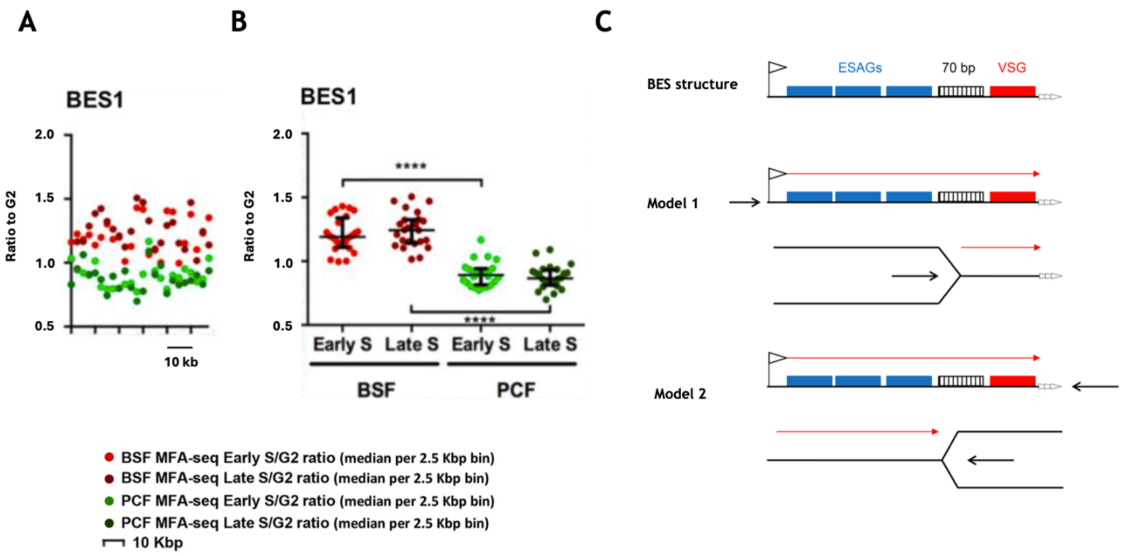


Figure 8 DNA replication in the active BES in bloodstream-form cells versus procyclic cells. A and B. The active BES in the cell line (BES1) is replicated earlier in BSF cells compared to PCF. C. The suggested models for direction of replication at the active BES; black arrows indicate the replication direction, red ones – direction of RNA polymerase I progression. Adapted from Devlin et al., (2016) under Open Access, Creative Commons license.

Two *Leishmania* species, *L. major* and *L. mexicana*, have also been analysed using MFAseq (Marques *et al.*, 2015). Unlike *T. brucei*, both *Leishmania* species were found to have only one predominant origin of replication per chromosome (Figure 9), which is highly unusual for a eukaryote. Based on the estimated replication fork progression in *T. brucei* (Calderano *et al.*, 2015) and length of S phase in *Leishmania* (Wheeler, Gluenz and Gull, 2011), the authors suggested that if these parasites, indeed, have only one origin of replication per chromosome, about half of the chromosomes would not have enough time to complete replication during S phase (Marques *et al.*, 2015). Another approach at detecting replication origins - small leading nascent strand purification coupled to next-generation sequencing (SNSseq) – suggested that the *L. major* genome contains ~5000 replication origins (Lombraña *et al.*, 2016); in addition, more recently, subtelomeric replication has also been detected in G1 and G2/M cells of *L. major* (Damasceno *et al.*, 2020). Both of the above-mentioned findings suggest that MFAseq-suggested single origins per chromosome might not reveal the full picture of replication dynamics in this parasite. In fact, recent in-depth work by Damasceno et al. (Damasceno *et al.*, 2024a) confirmed this – by detecting replication fork movement and prediction of initiation and termination sites in single DNA molecules of *L. major*, the authors showed that in addition to the single constitutive origin per chromosome as shown previously using MFAseq, this parasite deploys thousands of stochastic origins across the chromosomes. The work showed that the stochastic replication initiation sites have distinct sequence characteristics - high AT content and elevated levels of secondary DNA structures termed G-quadruplexes (Damasceno *et al.*, 2024a). As measured by MFAseq, replication timing in *L. major*, which, curiously, is related to chromosome size (with smaller chromosomes replicated earlier in S phase than larger ones) (Damasceno *et al.*, 2020) appears to also relate to genomic stability – the earlier-replicating constitutive initiation sites show higher SNP rates compared to later-replicating initiation sites (Damasceno *et al.*, 2024a).

In a separate publication, a relationship between R-loops and nuclear DNA replication was recently shown in *L. major* (Damasceno *et al.*, 2024b). R-loops are triple-stranded nucleic acid structures consisting of an RNA-DNA hybrid and a displaced DNA strand in the genome; these tend to accumulate in specific genomic regions, often associated with transcription and/or replication activity, and, in many cases, genomic instability (reviewed in Petermann, Lan and Zou, (2022)).In the genome of *L. major*, it was recently shown that R-loops accumulate in a pattern reflecting DNA replication timing: specifically, chromosomes that replicate later have higher levels of R-loop accumulation. These larger, later replicating chromosomes, in addition to accumulating higher levels of R-loops, also displayed increased chromatin accessibility (as measured by MNase-seq) and fewer G-quadruplex (G4) secondary DNA structures. RNase H1

– a factor that is involved in R-loop removal, appears to play a key role in this parasite's chromosome-size-related replication programme, as its loss abolishes it; beyond this, the loss of RNase H1 also leads to genomic instability in the form of copy number variation (CNV), particularly in the larger, usually more R-loop rich and later-replicating chromosomes (Damasceno *et al.*, 2024b), further highlighting the link between R-loops, DNA replication and genomic stability in *L. major*.

It remains unclear why the bimodal DNA replication programme described above - with single, constitutive and early-replicating loci per chromosome, supplemented with many stochastic loci across the chromosomes - is required to replicate a relatively small genome, or why the two kinetoplastids have evolved distinct DNA replication programmes. For *T. brucei*, DNA replication appears to be tightly connected to transcription, perhaps more so than in *L. major*, both in terms of co-localisation with numerous transcription initiation and termination sites (PTU boundaries) in the core genome and in the context of active/silent BESs in BSF and PCF cells (Tiengwe *et al.*, 2012; Devlin *et al.*, 2016); it remains to be explored in the arguably lesser-studied megabase chromosome subtelomeres or smaller chromosome categories, the replication timing and dynamics of which remain unknown.

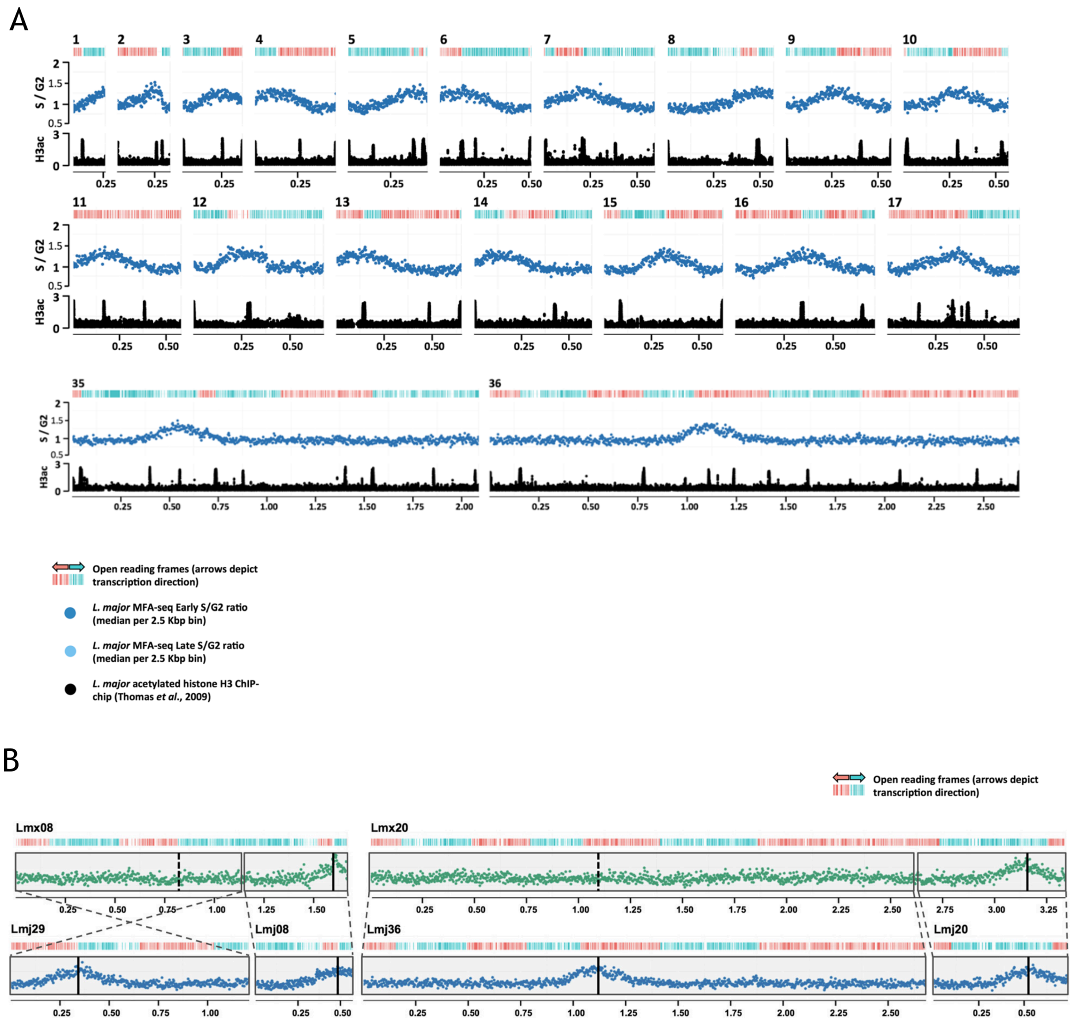


Figure 9 DNA replication dynamics and replication origins in *L. major* and *L. mexicana.*

A – MFAseq mapping across several chromosomes of *L. major*. MFAseq S phase signal was normalised against G2/M signal, and the ratio is shown as median values in windows of 2.5 kb across x axis. Acetylated histone 3 (H3ac) ChIPseq data is also shown (log2 values), highlighting transcription start sites. B – MFAseq mapping in syntenic chromosomes of *L. major* and *L. mexicana* that have undergone fusion or fission. Lmj – *L. major* chromosome, Lmx – *L. mexicana* chromosome, syntenic sequences and their orientation indicated with blocks and dashed lines. As above, S phase MFAseq signal was mapped relative to G2/M signal, and the ratios plotted (green for *L. mexicana* and blue for *L. major*). Approximate location of replication origins indicated with solid vertical lines, and vertical dashed lines indicate the location of 'missing' origins (present in the syntenic sequence). Figure adapted from Marques *et al.*, (2015) (reproduced under Open Access, Creative Commons CC BY 4.0 license).

1.3 Nanopore sequencing

So-called third generation sequencing comprises a group of newer nucleotide sequencing technologies, most notably PacBio and Oxford Nanopore Technologies (ONT, commonly referred to as Nanopore), among others. Unlike next generation sequencing, such as Illumina, where fragments of short length are generated, typically about 75-150 bp in length, PacBio and ONT can produce sequencing reads that may be as long as DNA fragments in the sample (Lu, Giordano and Ning, 2016). Nanopore, specifically, typically generates reads that are several or several tens of kilobases long, and the longest reads ever sequenced are above 2 Mb (Wang *et al.*, 2021). As such, it has been utilised in improving genome assemblies of various organisms, allowing researchers to overcome certain common issues, for instance, in resolving repetitive genomic regions, including centromeric repeats, paving the way to telomere-to-telomere chromosome and genome assembly (Jain *et al.*, 2018; Miga *et al.*, 2020; Chung, Kwon and Yang, 2021; Naish *et al.*, 2021; Nurk *et al.*, 2022). In a related trypanosome, *Trypanosoma cruzi*, long-read sequencing has been used extensively in order to improve genome assembly, resulting in resolved haplotypes, identification of novel repetitive sequences and improved understanding of expanded gene families in this parasite (Berná *et al.*, 2018; Callejas-Hernández *et al.*, 2018; Díaz-Viraqué *et al.*, 2019; Hakim *et al.*, 2023).

These newer sequencing technologies are not without fault, however. ONT, specifically, typically presents with lower base qualities compared to, for example, Illumina sequencing, and basecalling homopolymer sequences (consecutively repeated nucleotides) in particular represents a challenge – these tend to be incorrectly counted with older R9.4 and R9.5 pore technology (Wang *et al.*, 2021). The newer generation pores (R10.4.1) and kits (V14) that ONT now provide, coupled with newer basecalling tools such as dorado, have reduced these issues dramatically, however, as achieving base qualities of phred score 20 (99%) or above are routine (reviewed in Zhang *et al.*, (2024)).

1.4 Aims

The overarching theme of this thesis revolves around the exploration of genome organisation, stability and composition in *Trypanosoma brucei*. The three results chapters focus on the following:

- Genomic compartmentalisation, DNA replication and stability in *T. brucei*;

- Detection of modified DNA nucleobases in *Trypanosoma brucei* using Nanopore sequencing;

- Analysis of nucleotide patterns and skews in the *Trypanosoma brucei* genome, comparing these findings with *Leishmania major* and in the phylogenetic context of diverse trypanosomatids.

The three chapters share a strong emphasis on DNA sequence analysis and, in each case, their introduction provides a more complete explanation for the rationale of the investigations than is provided in the current section.

2 Materials and methods

2.1 Data generation

2.1.1 Parasite culture *in vitro*

2.1.1.1 Bloodstream-form *Trypanosoma brucei brucei*

Bloodstream form (BSF) *Trypanosoma brucei brucei* cells (strain Lister 427) were propagated at +37℃ and 5% CO_2 in HMI-9 medium with 10% foetal calf serum. Cell growth was analysed using a haemocytometer at 24h intervals. Every two days the cells were passaged, and the target cell density set to 1 x 10^4 cells mL^{-1}.

2.1.1.2 Procyclic-form *Trypanosoma brucei brucei*

Procyclic-form (PCF) *Trypanosoma brucei brucei* cells (strain Lister 427) were propagated in sealed flasks at +27℃ in SDM79 medium with 10% foetal calf serum and 0.2% hemin. Cell growth was analysed using a haemocytometer at 24h intervals. Every two days the cells were passaged, and the target cell density set to 5.5 x 10^5 cells mL^{-1}.

2.1.2 Parasite collection, DNA extraction and QC

For high molecular weight genomic DNA extraction of BSF cells, 200 mL of culture was harvested at approximate density of 1 x 10^6 cells mL^{-1}, whereas for PCF cells – 20 mL of culture at approximately 1 x 10^7 cells mL^{-1}. The cultures were centrifuged for 10 minutes at room temperature at 1500 RPM, the supernatant was then discarded, and the cell pellet was resuspended in 400 µL of 1X phosphate buffered saline (PBS) followed by centrifugation for 3 minutes at room temperature at 3000 RPM. The supernatant was removed and the cell pellet stored temporarily at -20℃.

For Illumina sequencing, DNA extraction was performed using a Qiagen DNeasy kit as per manufacturer’s instructions (Animal Blood or Cells Spin Column protocol), and the DNA library was prepared by Mr Craig Lapsley using a Qiagen QiaSeq FX DNA library kit. Illumina NextSeq 500 was used to perform paired-end whole genome sequencing at Glasgow Polyomics.

In order to preserve DNA fragment length, DNA extraction for ONT sequencing was carried out using the Qiagen MagAttract HMW DNA kit for genome assembly and the NEB Monarch HMW DNA Extraction Kit for Cells & Blood for modified base detection.

For assembly, most of the cell culture and DNA extraction was done by Dr Emma Briggs, and a minority by the author. Cell culture and DNA extraction for modified base detection was done by Dr Catarina de Almeida Marques.

DNA quality was assessed using NanoDrop 2000 for purity, Qubit 3.0 Fluorometer BR dsDNA kit for quantification and High Sensitivity DNA kit using a BioAnalyzer 2100 for fragment size estimation.

2.1.3 DNA sequencing library preparation and Nanopore sequencing

For genome assembly, the Oxford Nanopore Technologies (ONT) Ligation Sequencing Kit (SQK-LSK109) was used as per manufacturer's instructions. For modified base detection, the Ligation Sequencing (SQK-LSK109) library preparation kit with the Native Barcoding Expansion kit 1-12 (EXP-NBD104) were used for sequencing library preparation. R9.4.1 chemistry MinION flow cells were used. In both sets of sequencing experiments, care was taken to preserve the high molecular weight DNA - this was done by avoiding applying force, where possible, to avoid DNA shearing; examples of this include using wide-bore pipette tips, pipetting slowly and avoiding the introduction of bubbles to the samples. Mixing of the samples was performed using a HulaMixer rotating tube mixer rather than pipetting to allow more even mixing and avoid potential DNA shearing forces of pipetting.

2.2 Data analysis

2.2.1 Chapter 3 analysis

2.2.1.1 Basecalling and QC

Basecalling following sequencing was performed using guppy (version 3.3.3 for Linux CPU) using the high accuracy settings for DNA reads. Quality control and basic sequencing metrics on basecalled data were produced using NanoPlot (version 1.30.0).

2.2.1.2 Genome assembly and QC: *Trypanosoma brucei* genome

Long-read genome assembly was performed using canu with default settings and predicted genome size of 35 Mb. Briefly, canu operates in three stages – first, the longest 40x reads are corrected, improving the accuracy of bases, second – individual reads are trimmed so that only the highest quality bases are retained for assembly, and third – assembly, where reads are ordered into contigs and consensus sequence is generated (Koren *et al.*, 2017). The correction step involves the assembler overlapping reads, generating a consensus based on the overlaps and correcting the individual reads – this is necessary because the older generation ONT reads contained significant errors (in the forms of single nucleotide polymorphisms, insertions and deletions), which are 'corrected' as part of this process. In the second stage – trimming – regions of the corrected reads that do not align with the consensus are either trimmed or split into multiple reads; this is also when chimeric reads are identified and trimmed. Finally, the corrected and trimmed reads are once again overlapped, any final sequence errors are corrected, and an assembly graph is constructed and consensus contigs are generated (Figure 10, further detail on the genome assembly process that canu utilises can be found in the corresponding article).

After assembly using canu, four iterations of polishing with pilon (version 1.23) were performed in order to polish the resulting assembly using paired-end Illumina data (2x75bp); this is done in order to correct error-prone ONT data with higher base quality Illumina data. In the polishing process short reads were mapped to the assembly using bwa mem in paired-end mode (bwa version 0.7.17), the output converted into sorted bam format using samtools view -Sbu and sorted using samtools sort (samtools version 1.9). The resulting bam file is then indexed using samtools index and used as input for polishing using pilon. The resulting corrected assembly in fasta format is indexed using bwa index and used in the next iteration of polishing as described above, replacing the assembly fasta file in the process. Four iterations of polishing were performed, and the resulting genome assembly was used in all subsequent analyses.

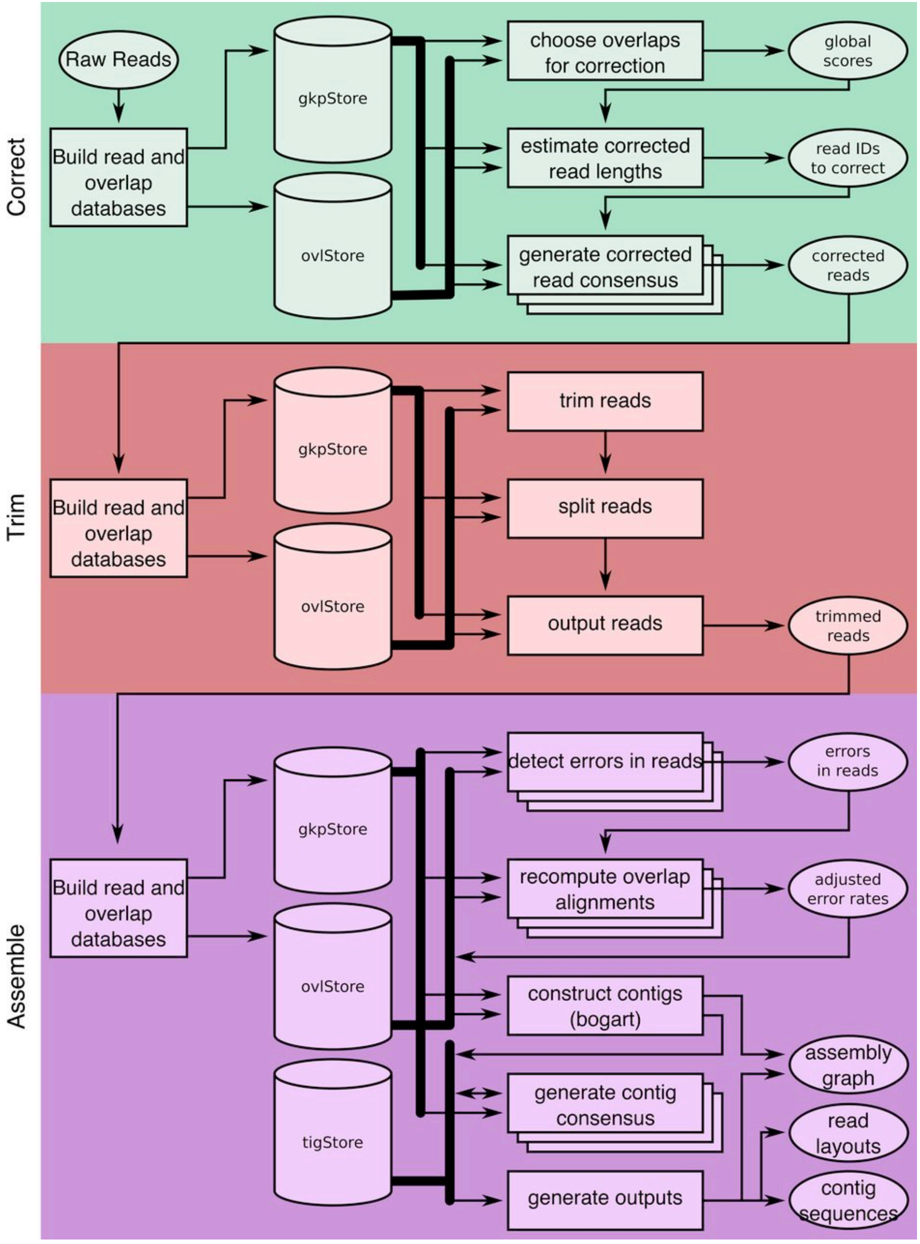


Figure 10 Canu genome assembly: stages and breakdown of steps.
The diagram shows the stages and individual steps of genome assembly using canu. gkpStore – read database, ovlStore – read overlap database, tigStore – unitig database. Figure from Koren et al., (2017), reproduced under Open Access, Creative Commons CC-BY 4.0 license.

Quality control and assessment was done using BUSCO (v. 3.5.2, database - eukaryota_odb10) and QUAST (v. 5.0.2.). Briefly, BUSCO provides a score that is based on the presence of expected core genes within a group of organisms (Manni *et al.*, 2021), in this case - eukaryotes, whereas QUAST provides a number of genome assembly metrics such as N50, assembly size and others.

2.2.1.3 Genome annotation and comparison

Companion (https://companion.gla.ac.uk/) was used for general genome annotation (Steinbiss *et al.*, 2016). *VSG* genes and pseudogenes in the assembly were identified using a custom script written by former WCIP bioinformatician Dario Beraldi (available on https://github.com/glaParaBio/TbruceiGenomics). Briefly, the short script extracted *VSG* sequences from the TriTrypDB reference genome for *Trypanosoma brucei* Lister 427 2018 (build version 47), created a Basic Local Alignment Search Tool (BLAST) database of *VSGs* using makeblastdb and allowed the user to query any DNA sequence (in this case, the resulting long-read assembly) in fasta format in order to identify and annotate *VSG* sequences. This was also used, more broadly, to assess the completeness of *VSG*-containing regions of the genome, such as subtelomeres.

Assembly vs assembly comparisons with the published reference (v. 46 *T. brucei* 427 2018 reference from TriTrypDB) were performed by running reciprocal alignments with minimap2 (command: minimap2 -x asm5 ref.fa assembly.fa > aln.paf) and visualised using Circos. Reciprocal assembly comparison was used in order to identify BES, confirm contig identity, assess potential rearrangements and gap bridging, among others.

Strand switch region (SSR) annotation was obtained by former WCIP bioinformatician Dr Kathryn Crouch from transcription start and termination sites, made available by Müller *et al.*, (2018) with the publication of the *T. brucei* Lister 427 2018 genome sequence.

In order to characterise any novel putative genes in the Nanopore assembly, we performed reciprocal BLASTp searches between the protein sequences from the TriTrypDB version of the Muller genome and the new assembly. Any protein sequences found not to have hits in the reference genome were further analysed in VEuPathDB's OrthoMCL pipeline that maps protein sequences to known OrthoMCL groups using DIAMONDp (https://orthomcl.org/orthomcl/app/workspace/map-proteins/new).

To identify regions of the ONT genome assembly that did not align to the 2018 reference genome, QUAST was re-run (version 5.2) and the coordinates of the unaligned regions were deposited on Zenodo (DOI: 10.5281/zenodo.14653477).

2.2.1.4 Repeat identification and motif isolation

Initial repeat identification in the genome was performed using Tandem repeats finder (TRF, version 4.09) using the recommended settings. The localisation of repeats reported by TRF was then investigated in the assembly vs assembly alignment in order to identify the likely category of said repeats - centromeric, 70bp, 50bp, 177bp, or telomeric. Motif sequences identified by TRF were subsequently used to further investigate underlying repetitive region structure using FIMO of the MEME suite of tools (Grant, Bailey and Noble, 2011), identifying the individual occurrences of the motifs in the region and genome more broadly; filtering based on maximum p-value ($<10^{-9}$) was applied to FIMO searches. Individual occurrences of the motif were extracted from FIMO output and uploaded to WebLogo (Crooks *et al.*, 2004) in order to generate motif logos.

2.2.1.5 Inter- and intra-region similarity assessment of repetitive genomic regions

In order to assess the similarity and uniformity of repeat regions, we created identity heatmaps using StainedGlass (Vollger *et al.*, 2022). Briefly, the input sequence is binned into fragments of a specified size (here, 500bp) and all possible pairwise sequence alignments between the fragments are made using minimap2. Taking into account the number of mismatches, insertions and deletions, the pairwise identity of the fragments is calculated and transformed into a score, which is then used to create the heatmaps (Mitchell R. Vollger *et al.*, 2022). Because the 70bp regions, on average, were shorter than the other repeat types, StainedGlass was not informative for visualisation and dotter (Sonnhammer and Durbin, 1995) was used instead (sliding window size parameter, W, set to 500).

2.2.1.6 Repeat region nucleotide content determination

Overall AT content of repetitive regions was determined using seqtk comp, by subtracting the number of G and C nucleotides from the total length of the region. Genome-wide GC content was calculated using nuc from the BEDtools suite of tools (Quinlan and Hall, 2010) with 50bp bin size; bedtools makewindows was used to create the genome-wide bed file containing all non-overlapping 50bp intervals.

2.2.1.7 Base quality assessment at 70bp and telomeric repeats

To evaluate base qualities at 70bp and telomeric repeats, >10kb ONT reads were extracted using samtools; utilising pysam, numpy and matplotlib python packages, the reads spanning the 70bp and telomeric repeat regions with +- 20kb flanking sequence (where available) were extracted, the median and interquartile base quality values in 100bp bins across the regions were calculated and plotted. To compare the base quality values between the two repeat region types, the base qualities were plotted as violin plots and then differences compared using a Mann-Whitney U test from the python package scipy.

2.2.1.8 Analysis of centromeric repeat composition

In order to locate centromere-associated repeats, makeblastdb was used to create a blastn database containing flanking regions of assembly scaffold gaps in putative centromeric loci in the Lister 427 assembly (Müller *et al.*, 2018). The database was used to query the new assembly with blastn; individual matches were manually checked to discard genic sequences, compared to TRF output, and the final coordinates, refined using TRF, were used in subsequent analysis. Motifs were identified by querying the identified regions using MEME using the following settings: -mod anr -nmotifs 10 -minw 10 -maxw 200 -revcomp -minsites 50 -maxsites 1500 (Bailey *et al.*, 2015). Based on the sequence and length of prevalent motifs within a given centromeric repeat region, it was assigned to a 'motif group'.

2.2.1.9 KKT2, KKT3, RNAseq and DRIPseq datasets

Raw DRIPseq (R-loop) data (*T. brucei brucei* strain Lister 427) (Briggs, Hamilton, *et al.*, 2018) was provided by Dr Emma Briggs. Kinetoplastid kinetochore proteins (KKT) KKT2 and KKT3 datasets (strain TREU927) were obtained from SRA using accession number PRJNA223204 (Akiyoshi and Gull, 2014).

The DRIPseq data was mapped using bwa mem, KKT2 and KKT3 – using bowtie2. In both cases, output sam files were transformed to bam format using samtools view -Sbu, sorted using samtools sort, indexed using samtools index, and duplicate reads removed using samtools rmdup. The resulting bam files were used with deepTools bamCompare in order to identify areas of enrichment (settings: minimum mapping quality 1, scaling method – SES). Metaplots for these data were generated using deepTools computeMatrix and plotHeatmap, using the scale-regions approach and only the figure formatting settings adjusted.

Previously published RNAseq data (strain Lister 427) (Briggs *et al.*, 2018) was trimmed using trim galore, aligned using hisat2 (v.2.2.1) (settings: --no-spliced-alignment), with further processing to sorted bam file as above, with additional samtools filtering retaining only reads that map once in the genome (samtools view -Sbu -d NH:1 file.sam > file.bam). DeepTools bamCoverage was used to assess expression across the genome, with RPKM used for normalisation.

2.2.1.10 MFAseq data processing

Marker frequency analysis coupled with sequencing (MFAseq) data was originally produced and published by Devlin *et al.*, (2016) (strain Lister 427, accession number PRJEB11437). More recently, Dr Catarina de Almeida Marques remapped the raw data to the more complete *Trypanosoma brucei* Lister 427 reference genome (Müller *et al.*, 2018) and the new long read Nanopore assembly (described in Chapter 3) as described by Devlin *et al.*, (2016) with the following changes: 1kb bins were used instead of 2.5kb, and bins with fewer than 100 reads were excluded from analysis, as well as those that were longer or shorter than 1kb. Origins were identified manually based on signal levels in the resulting bigwig files for both early and late S phase data. Metaplots of MFAseq data were plotted using deepTools' computeMatrix and plotHeatmap.

2.2.1.11 Genome stability assessment through read depth coverage change during *in vitro* growth

Illumina reads were trimmed and filtered using trim galore in paired-end mode with fastQC enabled (v 0.6.10), aligned to TriTrypDB reference genome *T. brucei*

brucei Lister 427 release 46 and the ONT genome assembly using bwa mem in paired-end mode (v. 0.7.17-r1188), formatted to bam, sorted and indexed using samtools (v. 1.19.2). DeepTools bamCoverage and bamCompare were used to analyse read depth coverage (RDC) and changes in RDC, respectively (normalisation using read count for bamCompare and RPKM for bamCoverage; minimum mapQ 1, bin size 50bp).

2.2.1.12 Analysis of sub-megabase chromosome content

To analyse gene content of the submegabase chromosomes, protein sequences from Companion annotation were used as input for the OrthoMCL (OG6_r20) pipeline on VEuPathDB's Galaxy (https://veupathdbprod.globusgenomics.org); the gene product description for the matching *T. brucei* gene on TriTrypDB was used to describe putative genes on the smaller chromosomes.

Further analysis of gene content of sub-megabase chromosomes, focusing on protein domains specifically, was performed by former WCIP Bioinformatician Dr Dario Beraldi (available on https://github.com/glaParaBio/TbruceiGenomics). Briefly, repeats were masked using RepeatMasker (https://www.repeatmasker.org/), genome annotation was carried out using BRAKER2 ('proteins of any evolutionary distance' mode), further identified protein annotation with Pfam protein domains was carried out using hmmer. Pfam protein domain enrichment plots were generated using ggplot2 in R.

2.2.2 Chapter 4 analysis

2.2.2.1 Reference genome versions

Genome versions used in this chapter - TriTrypDB release 46 of *Trypanosoma brucei* (Lister 427 strain) 2018 version (Müller *et al*., 2018) and the long-read *Trypanosoma brucei brucei* (Lister 427 strain) assembly generated in Chapter 3.

2.2.2.2 Demultiplexing, QC and modified base detection using Tombo

The following work – demultiplexing, QC and modified base detection using Tombo – was carried out by former WCIP bioinformatician Dr Dario Beraldi, and the relevant commands and scripts are available on GitHub (https://github.com/glaParaBio/TbruceiGenomics). Briefly, guppy (version 6.0.1 for Linux 64 CPU) was used for demultiplexing the two samples – BSF and PCF cells. NanoPlot (version 1.40.2) was used for QC, and for this purpose basecalling using guppy was carried out using fast basecalling settings. For signal correction (re-squiggling) and modified base detection, Tombo (version 1.5.1) was used in *de novo*, model sample and level sample modes.

2.2.2.3 Processing of 5hmU and base J ChIPseq data

5hmU ChIPseq data (strain EATRO1125 AnTat1.1, accession number PRJNA745154), originally published by Ma *et al*., (2021), were obtained from SRA and processed as described for DRIPseq data in 2.2.1.9 above. Base J ChIPseq data (strain Lister 427), originally published by Cliffe *et al*., (2010), was kindly shared by Dr Robert Sabatini in Solexa fastq format. The fastq files were mapped using bowtie2 (settings: --very-sensitive -solexa-quals –integer-quals), followed by samtools view -Sbu, samtools sort and samtools index in order to produce the sorted and indexed bam files used in subsequent analysis. As the base J ChIPseq data did not include an input sample, to normalise the dataset a wild-type (WT) Illumina-sequenced sample from *T. brucei* 427 Lister (BSF) produced in our lab was used. DeepTools' bamCompare (settings: scaling factors method – SES, minimum mapping quality 1) was used to generate the bigwig and bedgraph files used in downstream analysis.

2.2.2.4 Genomic region annotation sources

Gene and CDS annotations were obtained from the TriTrypDB annotation file in gff format, as well as using Companion (https://companion.gla.ac.uk/) (Steinbiss *et al.*, 2016) where indicated. *VSG* gene coordinates were extracted from the TriTrypDB gff annotation file using grep based on the annotation 'variant surface glycoprotein'.

PTU boundary and replication origin annotation was obtained as described in 2.2.1.3 and 2.2.1.10, respectively. PTU annotations were extrapolated from PTU boundary annotations manually, excluding the annotated boundary region from PTU coordinates.

Repetitive genomic region annotations used in this chapter were described in 2.2.1.4

2.2.2.5 Visualisation of 5hmU, base J and Tombo data

Several figures (metaplot profiles and heatmaps) in this chapter were produced using DeepTools' computeMatrix and plotHeatmap or plotProfile functions, either using reference-point mode or scale-regions mode, in many cases including flanking regions. Beyond the above-mentioned settings, only figure formatting settings were changed from the default. Correlograms (scatterplots) were plotted using python (seaborn package); briefly, all the datasets assayed were merged based on coordinates - for this, all datasets were re-mapped using 50bp windows. The merged data were then grouped based on genomic compartment (core, subtelomeric, BES or other) and scatterplot matrices were plotted between the different datasets.

2.2.2.6 Base-resolution data analysis

In order to analyse nucleotide-level Tombo data, all T, A, C and G residues were extracted from the chromosome 1 core of the TriTrypDB *T. brucei* Lister 427 2018 reference genome by, firstly, creating a bed file of 1bp bins across this region using bedtools makewindows, and secondly, running bedtools nuc to extract the nucleotide identity of each position in the sequence. The resulting coordinates for A, T, G and C were used to extract Tombo values for each nucleotide using deepTools' computeMatrix and plotProfile. Further analysis of base-resolution data was carried out using a custom python script (packages used: seaborn, matplotlib.pyplot, pandas, os). Briefly, the following data were merged based on base-resolution coordinates: read depth coverage (RDC) data from Tombo, *de novo* mode Tombo fraction of modified bases for BSF and PCF, GC and AT skews calculated in 10bp windows, and the nucleotide identity (A, T, C, G, N) of the position. Positions with read depth coverage below 10 and with 'N' nucleotide identity were removed, and the remaining data was used to interrogate modifications of A, C, T and G residues separately, including with varying levels of modification fraction (for brevity, only *de novo* BSF data on the forward (top) strand was used).

Nucleotide skews mentioned above around T and G residues on chromosome 1 were calculated using bedtools makewindows to create 10bp windows across the reference sequence, followed by bedtools nuc. For each 10 bp window, AT and GC skews were calculated as follows: AT skew = (A-T)/(A+T), and GC skew = (G-C)/(G+C), where A, T, G and C = instances of the corresponding nucleotide within a given window.

In order to determine whether apparently modified G positions were within ± 6bp of a highly modified T, we extracted the locations of T and G residues showing modified fraction levels of 0.30 of higher (with minimum RDC 10), and calculated the shortest distance between each such 'highly modified' G and any adjacent T's with >0.30 modification fraction; the resulting minimum distances for each G were then grouped based on whether they fall within ± 1bp ('immediately adjacent') or ± 6bp of a 'highly modified' T.

Lastly, in order to assay the nucleotide composition of the flanking sequence of apparently highly modified T and G residues, as above, and using thresholds of 0.16, 0.18, 0.30 and 0.50 fraction, ± 4bp flanking regions were extracted for each position using bedtools, the fasta sequences extracted, and used as input in WebLogo (https://weblogo.berkeley.edu/logo.cgi) (Crooks *et al.*, 2004), using both nucleotide frequencies and conservation (bits) settings.

2.2.2.7 Active and silent BES analysis

For the comparison of active vs silent BES elements (resident *VSG* and 70bp repeat region), BES *VSG* coordinates were taken from the TriTrypDB reference genome gff annotation file, whereas for the 70bp repeat region the newly generated long-read genome assembly was used – these regions were identified and annotated as described in 2.2.1.4. BES1 was designated as the ‘active’ expression site, though this might not be accurate for the cell line used in 5hmU ChIPseq experiments.

2.2.3 Chapter 5 analysis

2.2.3.1 Reference genome versions

Reference genome versions used in Chapter 5 are TriTrypDB release 58 of *Leishmania major* (Friedlin strain) and release 46 of *Trypanosoma brucei* (Lister 427 strain) 2018 version (Müller *et al.*, 2018). For G-quadruplex mapping in *T. brucei* a different genome reference was used – TriTrypDB release 62 of *Trypanosoma brucei* (TREU 927 strain). Lastly, the long-read assembly generated in Chapter 3 was used for visualising nucleotide skew changes at core-subtelomere genomic boundaries.

2.2.3.2 Nucleotide composition and skew calculations

Genome size, GC fraction, coding GC fraction and gene density were obtained using the genome annotation tool companion (https://companion.gla.ac.uk/). In order to calculate genome-wide GC and AT skews, non-overlapping 1kb windows were delineated using bedtools makewindows function; these were then used to profile the nucleotide content genome-wide in 1kb windows for both *L. major* and *T. brucei* using bedtools nuc. From the output of bedtools nuc, AT and GC skews were calculated as follows:

- AT skew = (A-T)/(A+T)
- GC skew = (G-C)/(G+C), where
    - A, T, G and C = instances of the corresponding nucleotide within a given window.

In addition, coding and non-coding AT and GC skews were calculated for both parasites. For this, instead of using 1 kb windows, the positions of all 'protein_coding_gene' annotations for *L. major* and 'gene' annotations for *T. brucei* were extracted from the corresponding gff file on TriTrypDB, and these were used as windows for bedtools nuc. In order to extract the positions of non-coding sequences, bedtools complement was used to generate a bed file that contains all genomic regions not contained in the 'coding' bed mentioned above, and this bed file was then used to run bedtools nuc and calculate nucleotide skews as above.

In order to calculate the % of coding A, T, G and C nucleotides, as above, 'protein_coding_gene' annotations for *L. major* and 'gene' for *T. brucei* were extracted from the corresponding gff files. Retaining the strand information for each interval, bedtools nuc was used separately for forward (+) and reverse (-) strand sequences. The sum of A, C, T and G nucleotide occurrences were calculated for the genes on the forward strand, whereas for the reverse strand the complement sum was calculated, i.e. each occurrence of A counted as T, T as A, G as C and C as G – this was done because bedtools nuc analysed the reference or forward strand without taking strand information into account. The combination of the forward strand nucleotide sums with the complement of reverse strand sums yielded the number of occurrences of each of A, T, C and G, and this was used to extrapolate the % coding.

GC fraction genome-wide was calculated, as above, using 1kb windows and bedtools nuc, except the GC fraction was calculated as GC fraction = G+C.

2.2.3.3 Genomic region annotation sources

Gene, CDS and *VSG* annotations were extracted from the corresponding genome version's gff file from TriTrypDB. Core, subtelomeric and BES genomic regions annotations were obtained from the *T. brucei* reference genome fasta file. Strand-switch region annotations for *L. major* were provided by Jeziel Damasceno, originally published by Lombraña *et al.*, (2016). For *T. brucei*, PTU boundary and replication origin sites were obtained as described in 2.2.1.3 and 2.2.1.10.

VSG annotation and sequence comparison to identify genomic loci in the long-read genome assembly was carried out as described in 2.2.1.3.

2.2.3.4 Visualising AT and GC skew data

AT and GC skew was visualised primarily in metaplot format for various regions of interest (ROI). For this, AT and GC skew bedgraph files generated as described above were converted to bigwig format using kentUtils bedGraphToBigWig, and the resulting bigwig files were then used by the deepTools tool computeMatrix in order to process the data for plotting with deepTools plotHeatmap.

2.2.3.5 G-quadruplex and R-loop data

Published G-quadruplex (Marsico *et al.*, 2019) mapping data for *L. major* and *T. brucei* (strain EATRO1125) was used in bedgraph format obtained from SRA (accession number PRJNA434023).

R-loop (DRIPseq) data for *L. major* was generated and provided by Jeziel Damasceno (Damasceno *et al.*, 2024b). Briefly, DNA/RNA hybrid antibody S9.6 was used to immunoprecipitate DNA/RNA hybrids, and a no-antibody control was used as an input sample for normalisation. For the treated sample, RNase H treatment was carried out before immunoprecipitation. DeepTools bamCompare was used to generate bedgraph files used here.

The *T. brucei* R-loop data processing and mapping was described in 2.2.1.9.

2.2.3.6 Codon usage assessment

Codon usage tables for the corresponding genome reference version were downloaded from TriTrypDB. Both frequency and abundance were plotted for each codon for *L. major* and *T. brucei* as is, grouping the codons based on G-C balance. In addition, a score was introduced in order to compare codon preference in the two parasites and its skewing potential. For each codon, +1 was added to the score for each G, and -1 for each C. The codon score was then multiplied by the abundance of said codon (number of codons per 1000 codons) for each parasite. In order to assess how the two parasites compare, the resulting weighted score of *L. major* for each codon was subtracted from the corresponding *T. brucei* score.

2.2.3.7 Identification of syntenic sequences between *L. major* and *T. brucei*

Example syntenic regions between *L. major* and *T. brucei* were identified using the TriTrypDB Genome Browser by selecting the corresponding species genome reference and selecting ‘Syntenic Sequences and Genes’ track.

2.2.3.8 Analysis of nucleotide composition and skews in trypanosomatids (and *Bodo saltans*) more broadly

To evaluate the observed skews in a phylogenetic context, analysis of nucleotide composition and skews has been expanded to include 37 more trypanosomatid species and *Bodo saltans*. For this, species were chosen to represent a broad range of trypanosomatids for which genome sequences are available on TriTrypDB or Genbank, aiming to use the framework for analysing traits in a phylogenetic context suggested in Kostygov et al. (2024) (Figure 77 A and Table 23). Genome annotations were done using companion, and the overall and coding GC % presented here was obtained from companion output. As described above (2.2.3.2), nucleotide skews were calculated using bedtools nuc in 1kb bins. To assess nucleotide skews in CDS and inter-CDS sequences, the gff annotations for 'CDS' were used, and where two neighbouring CDS were on the same strand, inter-CDS regions were extracted. Where the CDS sequences were on the negative (bottom) strand, the skew values were inverted. Median, Q1 and Q3 AT and GC skew values were calculated for each species and plotted using matplotlib.

3 Genomic compartmentalisation of *Trypanosoma brucei*: sequence, replication and genome stability

3.1 Introduction

3.1.1 Genome structure and compartmentalisation

The genome of *T. brucei* is shaped by unique processes governing infection by this parasite. While the parasite's modest <60 Mb diploid genome consists of mini (50-150kb), intermediate (150-700kb) and megabase (1Mb+) (Van der Ploeg *et al.*, 1984) chromosomes, it carries over 2500 variant surface glycoprotein (*VSG*) (pseudo)genes (Müller *et al.*, 2018). In bloodstream-form cells, only one VSG is expressed on the cell surface at a time during an infection (Cross, Kim and Wickstead, 2014). The transcriptionally active *VSG* gene normally resides in an active bloodstream-form expression site (BES); only one of ~15 BES is active at a time. BES are positioned immediately adjacent to the telomeres on some megabase and intermediate chromosomes and contain a number of co-transcribed genes, termed expression site associated genes (*ESAG*s), in addition to the resident *VSG* (Figure 11). Just upstream of the expressed *VSG* lie so-called 70-bp repeats (Bernards, Kooter and Borst, 1985; Kooter *et al.*, 1987; Hertz-Fowler *et al.*, 2008), and upstream of the BES lie 50-bp repeats (Zomerdijk *et al.*, 1990, 1991).

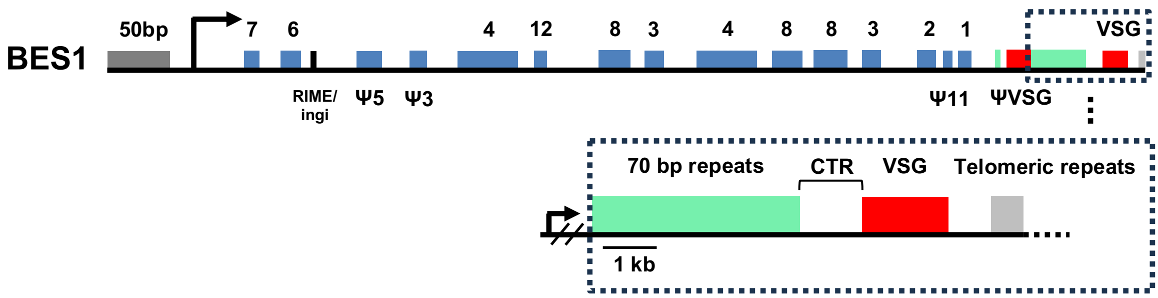


Figure 11 *T. brucei* bloodstream form expression site structure.
Diagrammatic representation of the structure of BES1 as an example of a bloodstream-form expression site. The black arrow on the lefthand side represents the location and direction of the expression site transcription promoter region and transcription direction; the numbered blue boxes represent expression site associated genes (*ESAGs*), with the number corresponding to the *ESAG* group. The blow-up provides detail of the *VSG* and surrounding repeats, as well as the so-called co-transposed region (CTR) between the *VSG* and 70 bp repeats. 50bp = so-called 50 bp repeat region upstream of the BES.

Most of the remainder of *VSG* genes and pseudogenes are located in the subtelomeric regions of 11 megabase chromosome pairs – these are mostly transcriptionally silent and variable genomic regions that occupy the space between the core, transcribed genome and telomeres (Figure 7). In addition, *VSGs* are also located on mini- and intermediate chromosomes that are thought to only contain repetitive elements, expanded gene families and, in the case of intermediate chromosomes, BES (Glover *et al.*, 2013). The subtelomeric regions of the larger chromosomes are hemizygous and can vary considerably in length within a pair of homologous chromosomes, as well as between strains and isolates (Callejas *et al.*, 2006), whereas the core of these chromosomes – where the bulk of housekeeping genes reside - is thought to be much more conserved within a pair (Figure 7) (Müller *et al.*, 2018; Cosentino, Brink and Siegel, 2021).

3.1.1.1 Current genome assembly

Deciphering the complete genome structure of *T. brucei* through sequencing has proven difficult; until 2018, the genome assemblies of this parasite were incomplete and contained subtelomeric sequences that could not be assigned to the respective chromosome 'cores' and remained highly fragmented (Berriman *et al.*, 2005; Müller *et al.*, 2018). The newest assembly, based on PacBio long-read whole-genome DNA sequencing and aided by Hi-C DNA interaction data, has provided major improvements to previous assemblies - subtelomeric sequences have been assigned to their respective chromosome cores, and many BES have been assigned to chromosomes, too (Müller *et al.*, 2018). The original and the more recent genome assemblies were done in different strains of *T. brucei brucei* – TREU 927 and Lister 427, respectively (Berriman *et al.*, 2005; Müller *et al.*, 2018).

In addition to the core, subtelomere and BES contigs, the more recent Lister 427 strain reference genome contains numerous, shorter, unassigned sequences – unitigs, a total of 260. These are between 1 and 142 kb in length (median 24.785 kb, mean 28.586 kb) and vary in composition – some contain repetitive sequences, others contain *VSGs* and other genes. In total, unitigs represent 14.8% of the reference genome and contain 2115 genes. It is unclear which genome compartments or chromosome sizes these sequences belong to.

Despite the significant advance of the 2018 genome assembly, some uncertainties and gaps in knowledge remain. The subtelomere and core contigs remain as separate contigs and we have no sequence information as to how these genomic compartments – core, subtelomere and BES – are bridged, as well as what the genomic context of unitigs is. In addition, the assembly is rather fragmented, comprising 317 contigs in the latest version, as well as 49 scaffold gaps (Müller *et al.*, 2018).

3.1.1.2 Genomic compartmentalisation: beyond structure

Genome compartmentalisation in *T. brucei* is not solely evident when looking at its organisation. Beyond the known transcriptional dichotomy of the megabase chromosome cores and subtelomeres (Müller *et al.*, 2018), genome-wide chromosome conformation capture (Hi-C, specifically) has further highlighted the partition between the megabase chromosome core and subtelomeres by showing that core-subtelomere boundaries also act as intrachromosomal DNA interaction domain boundaries. Consistent with their transcriptional inactivity, the subtelomeres also show a more compacted chromatin conformation (Müller *et al.*, 2018).

The active BES is found in a dedicated subnuclear compartment called the expression site body (ESB) (Chaves *et al.*, 1998; Navarro and Gull, 2001; Günzl *et al.*, 2003; Navarro, Peñate and Landeira, 2007), consistent with the high VSG expression rates of bloodstream-form parasites. The smaller chromosomes of this parasite also appear to be frequently confined within a portion of the nucleus during interphase (Chung *et al.*, 1990). These observations highlight that the genome partitioning is evident not merely on the sequence level, but also reflected in the spatial organisation in the nucleus.

3.1.2 Known repetitive elements of the *T. brucei* genome

In addition to having three classes of chromosomes with distinct structures and size, as well as harbouring over 2500 *VSG* genes and pseudogenes in a small genome, a further complication in genome assembly for *T. brucei* is the presence of largely uncharacterised repetitive regions. Such regions are associated with functionally distinct parts of the genome.

3.1.2.1 70bp repeats

So-called 70bp repeats lie immediately upstream of the telomere-adjacent *VSG* in BES (Figure 11), as well as upstream of 90% of subtelomeric *VSGs* (Marcello and Barry, 2007). The function of these repeats it thought to be two-fold – first, they act as homology tracts for recombination, second – due to their repetitive nature these regions increase DNA fragility and thus promote double-stranded DNA breaks (DSB) (Boothroyd *et al.*, 2009; Glover, Alsford and Horn, 2013; Hovel-Miner *et al.*, 2016). Furthermore, in plasmids TTA·TAA repeats, which are a prominent feature of the 70bp repeats, have been shown to adopt non-base-paired conformations under certain conditions – a property not shared with other triplet repeats and one likely to lead to instability (Ohshima *et al.*, 1996b, 1996a). Double-stranded DNA breaks would necessitate DSB repair, thereby recruiting the homologous recombination (HR) machinery which can then facilitate gene conversion – the replacement of the resident *VSG* with a donor from another BES or elsewhere in the genome (Glover *et al.*, 2013). *In vitro* experiments with artificially induced DSBs indicate that the precise location of the break may dictate whether or not VSG switching will be triggered (Boothroyd *et al.*, 2009; Glover, Alsford and Horn, 2013; Thivolle *et al.*, 2021).

Repeat sequence and structure analysis of these regions has been limited (Shah *et al.*, 1987; Hovel-Miner *et al.*, 2016), and it was carried out prior to the most recent version of the genome assembly being published in 2018, where all but two assembly gaps in the expression site 70bp sequences have been closed (Müller *et al.*, 2018). Earlier Sanger-based sequence analysis suggested variability in the repeats (Shah *et al.*, 1987), whereas the more recent analysis suggested that, contrary to the name (70bp repeats), there is a conserved and AT-rich 75-77bp repeat sequence underlying these regions (Figure 12) (Hovel-Miner *et al.*, 2016). This disagreement has yet to be resolved.

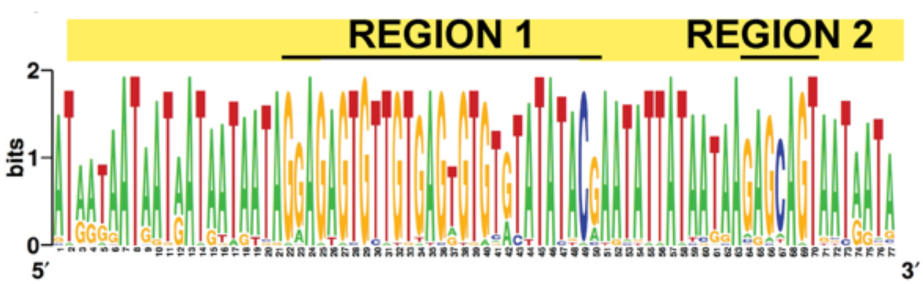


Figure 12 Repeat consensus sequence for BES1 70bp regions.

Repeat consensus obtained from 42 repeat units from *T. brucei brucei* Lister 427 BES1. Region 1 and region 2 highlight the more G-rich portions of the motif. Figure source – Hovel-Miner *et al.*, (2016) (reproduced under Open Access license, Creative Commons CC BY 4.0 license).

3.1.2.2 50bp repeats

Less is known about 50bp repeats that are positioned upstream of bloodstream-form expression sites (Figure 11). The limited published information suggests that these 'imperfect' repeat regions are likely to span at least 10kb in length, contain two *Rsa*I restriction sites (5' GT|AC 3') per repeat unit, and are present in both megabase and intermediate chromosomes (Zomerdijk *et al.*, 1990). These repeats have been found only in association with expression sites, specifically, the region upstream of the BES promoters (Zomerdijk *et al.*, 1990, 1991; Leeuwen *et al.*, 1997). The modified base J also localises to these repeat regions (Leeuwen *et al.*, 1997). As these repeats lie at the boundary of genomic compartments (core or subtelomere and BES), they serve as a barrier to sequence contiguity in genomic assembly; in the context of analysing MFAseq data to examine genome replication dynamics, for example, it was not possible to discern the direction or origin of replication at the active expression site due to this issue (Tiengwe *et al.*, 2012; Devlin *et al.*, 2016).

3.1.2.3 177bp repeats

Another group of repeats - the 177bp repeats - has been specifically associated with smaller, sub-megabase chromosomes (Sloof *et al.*, 1983; Van der Ploeg *et al.*, 1984; Van der Ploeg *et al.*, 1984; Wickstead, Ersfeld and Gull, 2004). Curiously, at interphase, 177bp repeat-containing DNA appears to be contained in less than 50% of the nucleus in 72.1% ± 1.1% of cells in contrast with DNA containing the telomeric repeats TTAGGG, where this figure was significantly lower – 20.5% ± 4.9%, suggesting the possibility of the 177bp repeats acting as centromeres for the smaller chromosomes (Chung *et al.*, 1990). Tandem stretches of 177bp repeats have been shown to contain a characteristic inversion, and these repetitive regions form the core of mini- and intermediate chromosomes, prompting further speculation that these regions may act as centromeres and origins of replication (Wickstead, Ersfeld and Gull, 2004). However, due to limited and incomplete 177bp-containing genome assemblies, it has not been possible to analyse DNA replication dynamics using techniques like MFAseq. Thus, there has been no test of how the submegabase chromosomes of *T. brucei* are duplicated.

3.1.2.4 Centromere-associated repeats

Among the 49 scaffold gaps of the current reference genome are centromeres of *T. brucei* – AT-rich, highly repetitive loci with repeat units varying from 30 to 147 bp (Obado *et al.*, 2007). Long-distance restriction mapping for chromosomes 1 to 8, where centromeres localise in the chromosome cores, indicated their size to be between 20 to 120 kbp, depending on the chromosome (Echeverry *et al.*, 2012). Centromeres of megabase chromosomes 9 to 11 were not found in the chromosome cores and therefore had not been analysed at the time (Echeverry *et al.*, 2012). The latest assembly assigned centromeres of chromosomes 9-11 to the subtelomeres, and provides a few kilobases of flanking sequence for the annotated centromeres (Müller *et al.*, 2018). When MFAseq data analysis was performed for *T. brucei,* the regions annotated as centromeres for chromosomes 1-8 co-localised with the most prominent MFAseq peak in these chromosomes early in S phase DNA replication, suggesting centromeres may act early in replication (Tiengwe *et al.*, 2012; Devlin *et al.*, 2016). Subtelomeric sequences were not available for this analysis at the time; it remains to be seen if centromeric regions localised in subtelomeric regions of chromosomes 9-11 show distinct sequence features, and, more importantly, if these also act in early DNA replication, similarly to those of chromosomes 1-8.

3.1.3 Genome stability

Genome stability is the principal goal and focus of DNA replication and repair mechanisms; in *T. brucei*, genome stability has been examined in two broad research contexts – comparative and evolutionary genomics (Callejas *et al.*, 2006; Sistrom *et al.*, 2014; Jackson *et al.*, 2016; Almeida *et al.*, 2018; Cosentino, Brink and Siegel, 2021; Reis-Cunha *et al.*, 2024) and in the study of DNA repair mechanisms, particularly in relation to antigenic variation (Robinson *et al.*, 2002; Hartley and McCulloch, 2008). Across a variety of isolates and lab strains, *T. brucei* tends to display stable ploidy and somy (Almeida *et al.*, 2018; Cosentino, Brink and Siegel, 2021; Reis-Cunha and Jeffares, 2024; Reis-Cunha *et al.*, 2024); at the same time, *T. brucei* subspecies and strains show pronounced variation in terms of DNA content and/or karyotype (Gibson and Garside, 1991; Gottesdiener *et al.*, 1991; Kanmogne, Bailey and Gibson, 1997; Melville *et al.*, 1998; Mulindwa *et al.*, 2021), at least some of which is attributed to differences in subtelomeric sequences of megabase chromosomes (Melville, Gerrard and Blackwell, 1999; Callejas *et al.*, 2006; Müller *et al.*, 2018). Focusing on short-term genomic stability of *T. brucei*, rather than genomic variability across evolutionary time, to our best knowledge, this has only been examined genome-wide in the context of isolate adaptation to *in vitro* culture (Mulindwa *et al.*, 2021) and in relation to two factors involved in DNA double-stranded break repair: MRE11 (Robinson *et al.*, 2002) and BRCA2 (Hartley and McCulloch, 2008). For the latter two studies, the nature of the genome stability experiments included prolonged passage *in vitro* of *T. brucei* cell lines that are heterozygous or homozygous null mutants for the respective factors (MRE11 and BRCA2), alongside WT cells, in order to assess gross genomic stability over ~550 and ~290 generations, respectively. In both cases this was assessed using pulsed field gel electrophoresis (PFGE) as NGS was not available or accessible at the time. Gross chromosomal changes, consistent with shortening of megabase chromosomes, were observed for MRE11-null and BRCA2-null cells, and to a lesser degree in BRCA2+/- cell lines (Robinson *et al.*, 2002; Hartley and McCulloch, 2008). In both cases, loss of select *VSG* genes was observed (detected using Southern blotting), whereas changes in intermediate or mini chromosome length were not easily detectable.

The unusual genome organization and compartmentalisation in *T. brucei*, as most recently evidenced by Hi-C interaction and DNA accessibility data (Müller *et al.*, 2018), leads to questions of what drives this unusual genomic partitioning and how it's maintained. Equally, the consequences of this partitioning haven't been fully elucidated. In the context of genome stability, the work mentioned above (Robinson *et al.*, 2002; Hartley and McCulloch, 2008) has hinted at the potential differential outcomes in the core vs subtelomeres of *T. brucei* megabase chromosomes, with the latter possibly suffering from higher instability. Nevertheless, the question of how genomic partitioning affects genomic stability, and whether DNA replication may play a role in this, has not been investigated. Crucially, no work has yet mapped DNA replication dynamics across the *T. brucei* subtelomeres.

3.1.4 A broader view of genomic compartmentalisation

In several other organisms, physical and functional genomic compartmentalisation has been examined and at least partially elucidated. In a recent publication investigating genome compartmentalisation in *Sulfolobus* species of archaea, the relationship between transcriptional activity, DNA replication, genome stability and three-dimensional nuclear conformation was examined. These organisms harbour distinct genomic compartments in relation to gene expression levels and topological domains, and it is in the domains with higher gene expression levels that the archaean origins of replication are located (Takemata, Samson and Bell, 2019). In subsequent work, the authors showed that the less actively transcribed regions, located further from origins of replication, accumulate higher levels of mutations (in the form of SNPs), and are also more densely compacted (Badel, Samson and Bell, 2022). In eukaryotes, similar observations have been made – DNA replication timing, chromatin architecture and three-dimensional nuclear conformation are inter-related aspects of eukaryotic genomes (reviewed by Chen and Buonomo, (2023)), and in yeast and human genomes these characteristics are often associated with gene expression levels as well (Müller and Nieduszynski, 2017; van Steensel and Belmont, 2017; Wang *et al.*, 2021).

With the above in mind, we wondered whether the arguably more extreme gene expression and functional split in the partitioned megabase chromosomes of *T. brucei* might also be reflected in and associated with DNA replication and genomic stability. To address this question, this chapter compares DNA replication dynamics in the megabase chromosome core and subtelomeric compartments.

3.1.5 Chapter aims and objectives

The aim of this chapter, in broad terms, was to improve our understanding of DNA replication dynamics and genome stability of *Trypanosoma brucei* in the context of a highly compartmentalised genome.

To achieve this, Oxford Nanopore Technologies long read DNA sequencing of the Lister 427 strain *Trypanosoma brucei brucei* parasites was utilised in *de novo* genome assembly with the aim of improving assembly contiguity, particularly across genomic compartment boundaries; this assembly was also described and compared to the most recent reference genome. The new genome assembly was leveraged to expand marker frequency analysis sequencing data (MFAseq) analysis to include core-subtelomere boundaries and subtelomeric regions of megabase chromosomes, as well as smaller, 177bp-repeat-containing sub-megabase chromosomes of the parasite. Next, genome stability was analysed across genomic compartments using sequencing data from *in vitro* longitudinal growth experiments of bloodstream-form *T. brucei*, in the context of DNA replication dynamics.

Lastly, as a more minor point, we took advantage of ONT sequencing in spanning additional repetitive sequences to characterise repetitive genomic elements in more detail than previously published, highlighting the potential of genome assembly to serve as a resource for a variety of analyses and applications.

3.2 *De novo* genome assembly and evaluation

3.2.1 DNA sequencing

A combined 13 runs of MinION and Flongle R9.4.1 flow cells were used to generate 319 075 reads totalling 3.56 gigabases of genomic DNA sequence of bloodstream-form *Trypanosoma brucei brucei* Lister 427 (Table 2, Figure 13).

Figure 13 Nanopore sequencing read length distribution.
Histogram depicting the sequenced read length distribution of the combined Nanopore sequencing runs. Note the prominent peak of reads at around 1 kb, which are derived from the mitochondrial minicircles.

Table 2 Nanopore sequencing output summary.
Data extracted from NanoPlot (De Coster *et al.*, 2018) output.

SEQUENCING STATISTICS	
Total number of reads sequenced	319 075
Total bases sequenced	3 558 525 709
Mean read length (bp)	11 152.6
Median read length (bp)	4307
Read length N50	28 458
Longest reads (bp)	345 688
	277 178
	252 902
	241 633
	240 202

3.2.2 Genome assembly and polishing

The resulting reads were used for *de novo* genome assembly using Canu. Briefly, Canu operates in three stages: first, the longest 40x reads are corrected, improving the accuracy of bases; second, individual reads are trimmed so that only the highest quality bases are retained for assembly; and third – assembly, where reads are ordered into contigs and consensus sequence is generated (Koren et al., 2017). Additional sequence correction following assembly was carried out using four iterations of Pilon (Walker *et al.*, 2014) using Illumina paired-end (2 x 75bp) sequencing reads generated from the same parasite strain.

3.2.3 Overall assembly metrics

In order to assess the quality of our assembly, a number of genome metrics were considered in comparison to the currently most complete and up-to-date reference genome for this strain, TriTrypDB version 46 of the *T. brucei brucei* Lister 427 genome, published in 2018 (Müller *et al.*, 2018). The assembly described here offers an additional 5.5 Mb of sequence, while reducing the contig number from 317 to 166 (Table 3). Furthermore, a number of other key metrics were improved; the minimum contig length needed to cover half of the genome - N50 – increased from 1 412 180 to 2 194 184, and the minimum number of contigs needed to produce half of the assembly - L50 - reduced from 11 to 9. Overall, the new assembly offers improved contiguity compared to the 2018 Lister 427 reference genome; indeed, there were also significant changes relative to the older Lister 427 reference genome (Table 3, Figure 14).

Genome completeness was measured using BUSCO; briefly, BUSCO provides a score that is based on the presence of expected core genes within a group of organisms (Manni *et al.*, 2021), in this case eukaryotes. The BUSCO score of our assembly was comparable, or perhaps slightly reduced, compared to that of the reference (Table 3), as 139, instead of 141, genes were detected. A higher degree of duplication was evident in our assembly compared to the reference (11 duplicated genes instead of 3), suggesting possible overlap between contigs or haplotype phasing. The total number of bases aligned to the 427 2018 reference was 52 565 861, and 1 914 208 bases were unaligned; of these, 396 939 bp represent 54 fully unaligned contigs and 1 517 269 bp - parts of 72 contigs (representing partially unaligned contigs), based on QUAST output (unaligned regions provided as data on Zenodo, DOI:10.5281/zenodo.14653477).

Table 3 General genome assembly metrics.
Overview of select genome assembly metrics generated by QUAST (Gurevich *et al.*, 2013), a custom script for the *VSG* numbers, and BUSCO for the BUSCO scores (Manni *et al.*, 2021). 'CANU 427' – current genome assembly; '427 2018' – the most up-to-date genome reference initially published by (Müller *et al.*, 2018);'427' – the previous generation of *T. brucei brucei* Lister 427 assembly available on TriTrypDB (tritrypdb.org).

Genome statistics	CANU 427	427 2018	427
# contigs	166	317	32
# contigs (>= 10000 bp)	104	303	32
# contigs (>= 25000 bp)	100	183	31
# contigs (>= 50000 bp)	97	72	24
largest contig	5080222	4633729	4977113
total length	55332974	50081021	26754408
total length (>= 25000 bp)	55128177	47671040	26743874
total length (>= 50000 bp)	54988747	43945746	26465291
N50	2194184	1412180	2482252
N75	538565	539199	1619978
L50	9	11	4
L75	22	25	7
GC (%)	43.85	43.71	46.71
# N's	0	49000	1050844
# N's per 100 kbp	0	97.84	3927.74
# VSG genes	3511	3524	387
BUSCO score	C: 139 [S: 128, D: 11], F: 23, M: 93, n: 255	C: 141 [S: 138, D: 3], F: 22, M: 92, n: 255	C: 140 [S: 137, D: 3], F: 22, M: 93, n: 255

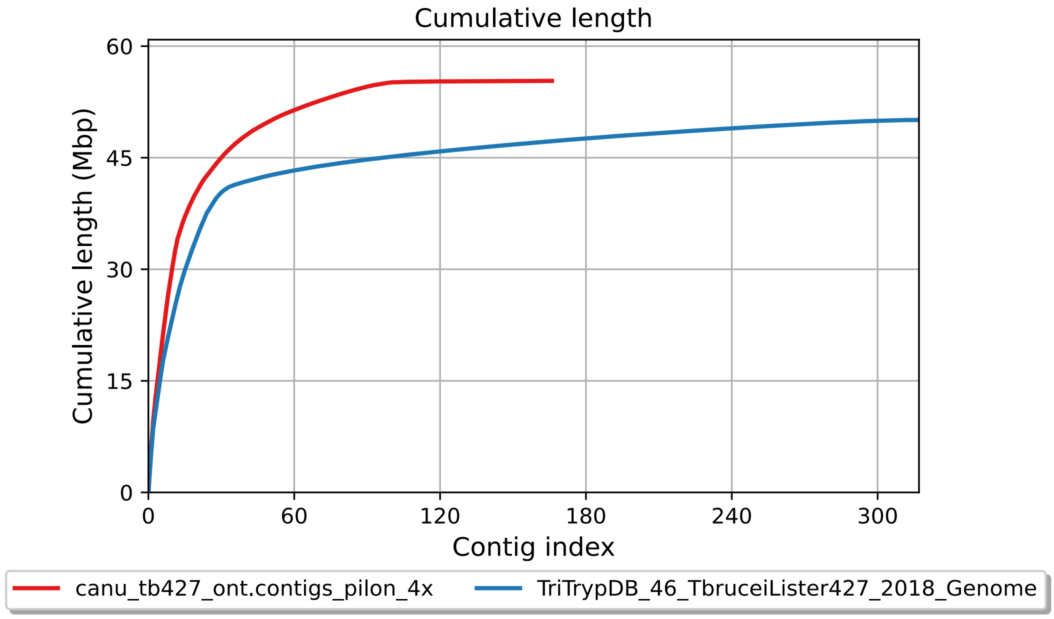


Figure 14 Cumulative assembly length.
Contigs ordered from longest to shortest; figure obtained from QUAST output (Gurevich *et al.*, 2013) .

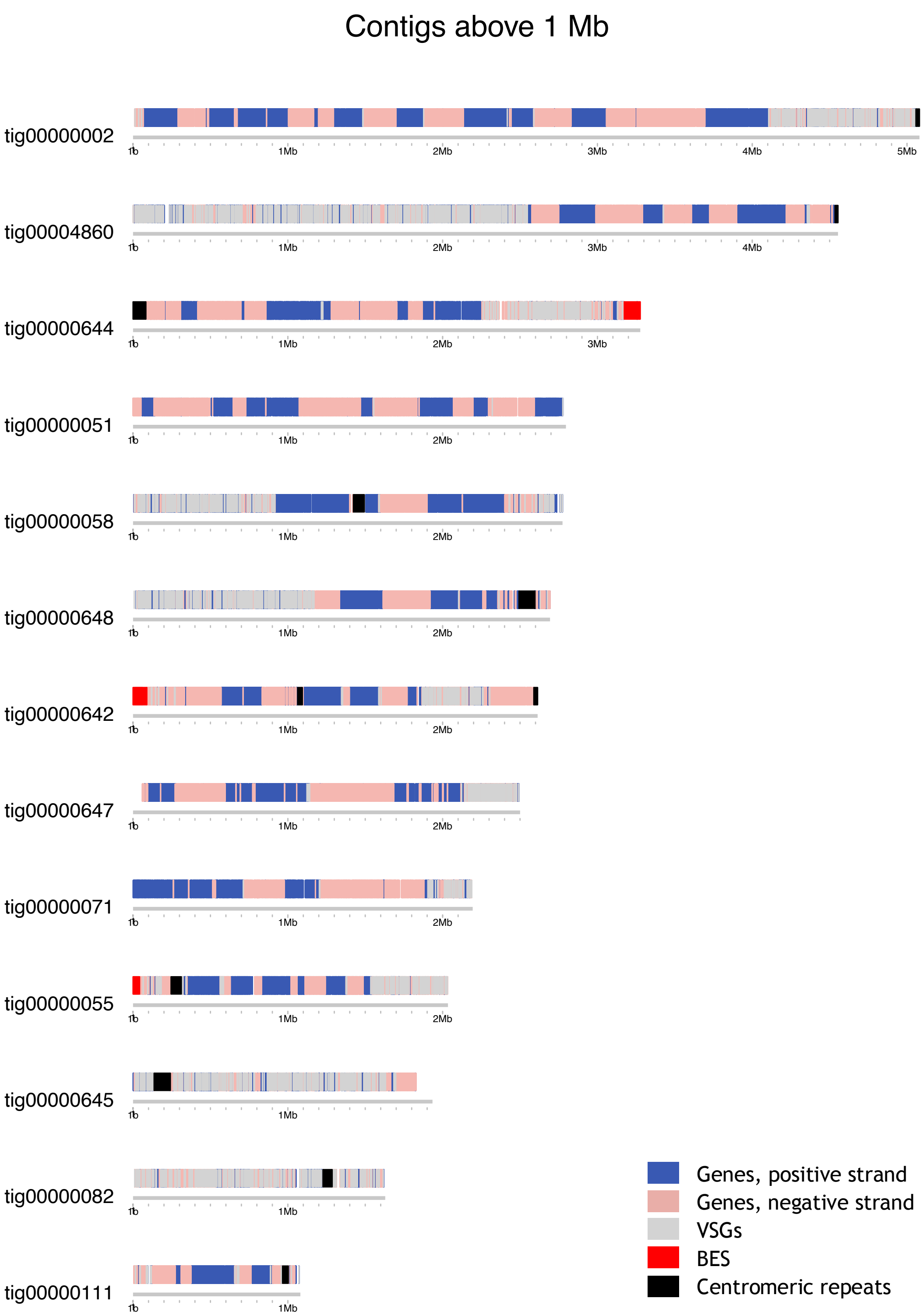


Figure 15 Overview of contigs above 1 Mb in the new assembly.

BES – bloodstream-form expression site, VSG – variant surface glycoprotein.

3.3 Contiguation evaluation

3.3.1 Bridging the gaps: megabase chromosomes

In order to further assess how the Nanopore assembly and the 2018 Lister 427 reference compare, overlaps between the new assembly and the reference were visualised and analysed. 13 contigs in the Nanopore assembly were >1 Mb in size, with the longest being ~5 Mb (Figure 15). Four of these contigs extended from the core or subtelomere to a BES, five encompassed centromeres, and nine included both core and subtelomere compartments of the genome. For 10 out of 11 megabase chromosomes, the new assembly bridged, in sequence, at least two previously separate contigs (Table 4); while for chromosome 2 this bridging was not observed, the core contig did, however, connect with numerous other previously unassigned genomic fragments to form two large contigs – one over 300 kb in length, the other over 1 Mb. Below, a selection of chromosomes is shown and examined in relation to the reference assembly in more detail.

Table 4 **Summary of contigs bridged in the new assembly for megabase chromosomes.** 5A, 5B, 3A and 3B refer to subelomeric arms of the chromosomes as annotated by Müller *et al.*, (2018), chr – chromosome, BES – bloodstream-form expression site.

Chr	Contigs within the reference					Contigs bridged in the new assembly
	5A	5B	3A	3B	core	
1	✓	✓	✓	✓	✓	3A & core 3B & core 5B & core 5A & core
2	✓				✓	
3	✓	✓	✓		✓	5A & core & 3A 5B & core
4	✓	✓	✓	✓	✓	5A & 5B & core & 3A 3B & core
5			✓	✓	✓	3A & core (& BES5) 3B & core
6			✓	✓	✓	3B & core 3A & core
7	✓				✓	5A & core
8	✓	✓	✓	✓	✓	5B & core 3A & core 5A & core
9	✓	✓	✓	✓	✓	3B & core 3A & core 5A & core
10	✓	✓	✓	✓	✓	3B & core 3A & core 5B & core
11	✓	✓	✓	✓	✓	3B & core

3.3.1.1 Chromosome 4

In the reference genome, chromosome 4 is represented by four subtelomere contigs (5A, 5B, 3A, 3B) and one core contig (Figure 16).

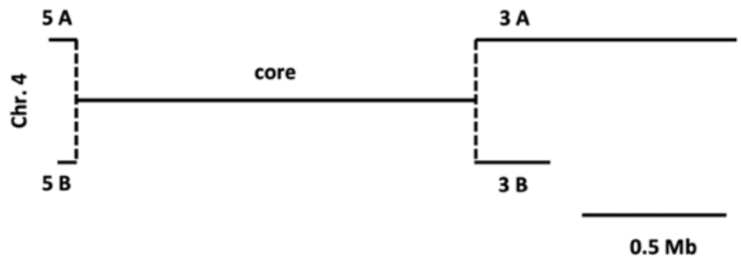


Figure 16 A digrammatic representation of the structure of chromosome 4 in the reference genome.
5A, 5B, 3A, and 3B – the subtelomeric contigs. Figure adapted from Müller *et al.*, (2018). Created with BioRender.com

In the new assembly, all the previously separate contigs have been bridged and this resulted in two contigs: one spanning from subtelomere to subtelomere (5A-5B-core-3A) (Figure 17 A), and one bridging across from the core of the chromosome into 3B (Figure 17 B). An additional third contig, which may represent the other 5' subtelomere, has also been identified (Figure 17 C), but it does not bridge two 'known' chromosome 4 contigs. Based on the new assembly, it appears that 5A and 5B subtelomere contigs are part of the same subtelomere, bridged by a *VSG*-containing genomic region that was not previously identified as belonging to chromosome 4 (instead, a number of short, unassigned contigs in the 2018 assembly map to the region).

In addition, the centromeric repeats were expanded in the assembly (Figure 17 A, in black) and this region no longer contains a scaffold gap. The second contig, spanning the core-3B subtelomere boundary (Figure 17 B), showed that the full sequence of this region was already present in the reference genome and assigned to the matching chromosome, but was split at the core-subtelomere boundary into two contigs.

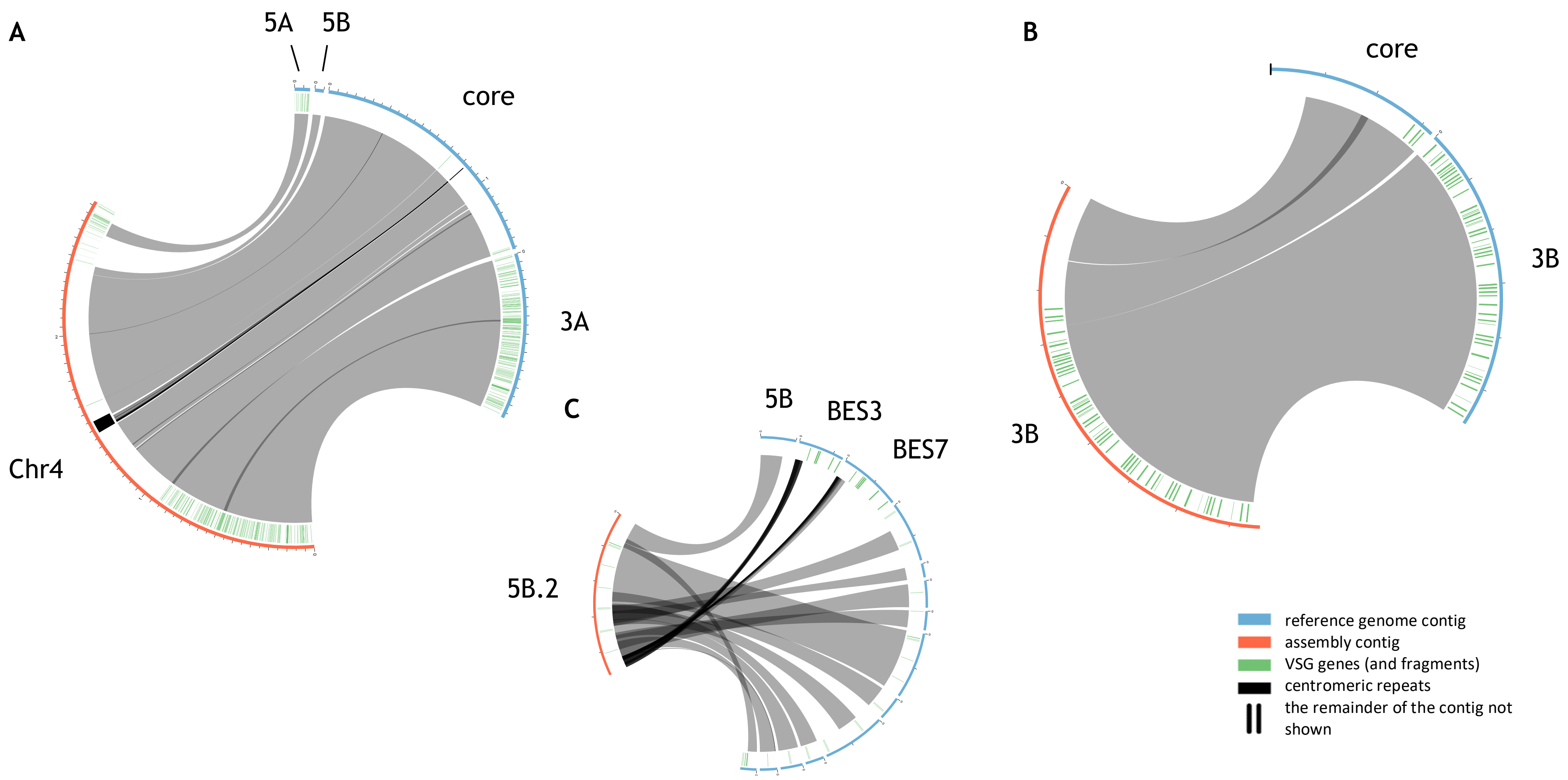


Figure 17 Contig structure of chromosome 4 in the new assembly.

Circos plots highlighting the overlap between the 2018 Lister 427 reference genome (Müller *et al.*, 2018) and the new assembly. Grey ribbons represent overlaps, black ribbons – multiple overlaps within a region (repetitive sequence).

3.3.1.2 Chromosome 10

Chromosome 10 is similarly represented by four subtelomere contigs (5A, 5B, 3A, 3B) and a core contig in the reference genome (Figure 18).

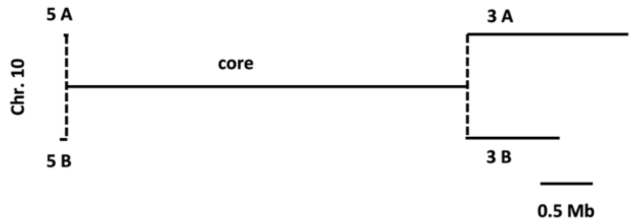


Figure 18 A diagrammatic representation of the structure of chromosome 10 in the reference genome.

Figure adapted from Müller et al., (2018). Created with BioRender.com.

This chromosome was captured in three contigs in the new assembly – a subtelomere-to-subtelomere contig (5A-core-3B) (Figure 19 A), a core-5B boundary-spanning contig, and a core-3A boundary-spanning contig (Figure 19 B). An additional contig containing almost the entire sequence of 3A was also identified (‘3A.2’ in Figure 19 B), but, as above, as it does not bridge two known chromosomal compartments for this chromosome, and therefore it remains uncertain whether it belongs to this chromosome. However, what would have been the boundary of the two 3A contigs (3A.1 and 3A.2 in Figure 19 B) consists of a long stretch of AT-rich repetitive sequence on both contigs of the new assembly - this may have resulted in difficulty during assembly of these sequences if there were too few long reads spanning this region in its entirety. The nature of these repeats has not yet been examined in detail, but their presumed genomic location suggests they do not fit the description of any other previously characterised repeats (such as the 50-bp or 70-bp repeats). As for chromosome 4 above, a centromeric repeat region was expanded in the new assembly in 3A.2 (Figure 19 B), eliminating the scaffold gap, and a second centromere-associated repetitive sequence was found at the end of the larger 3B-containing contig (Figure 19 A).

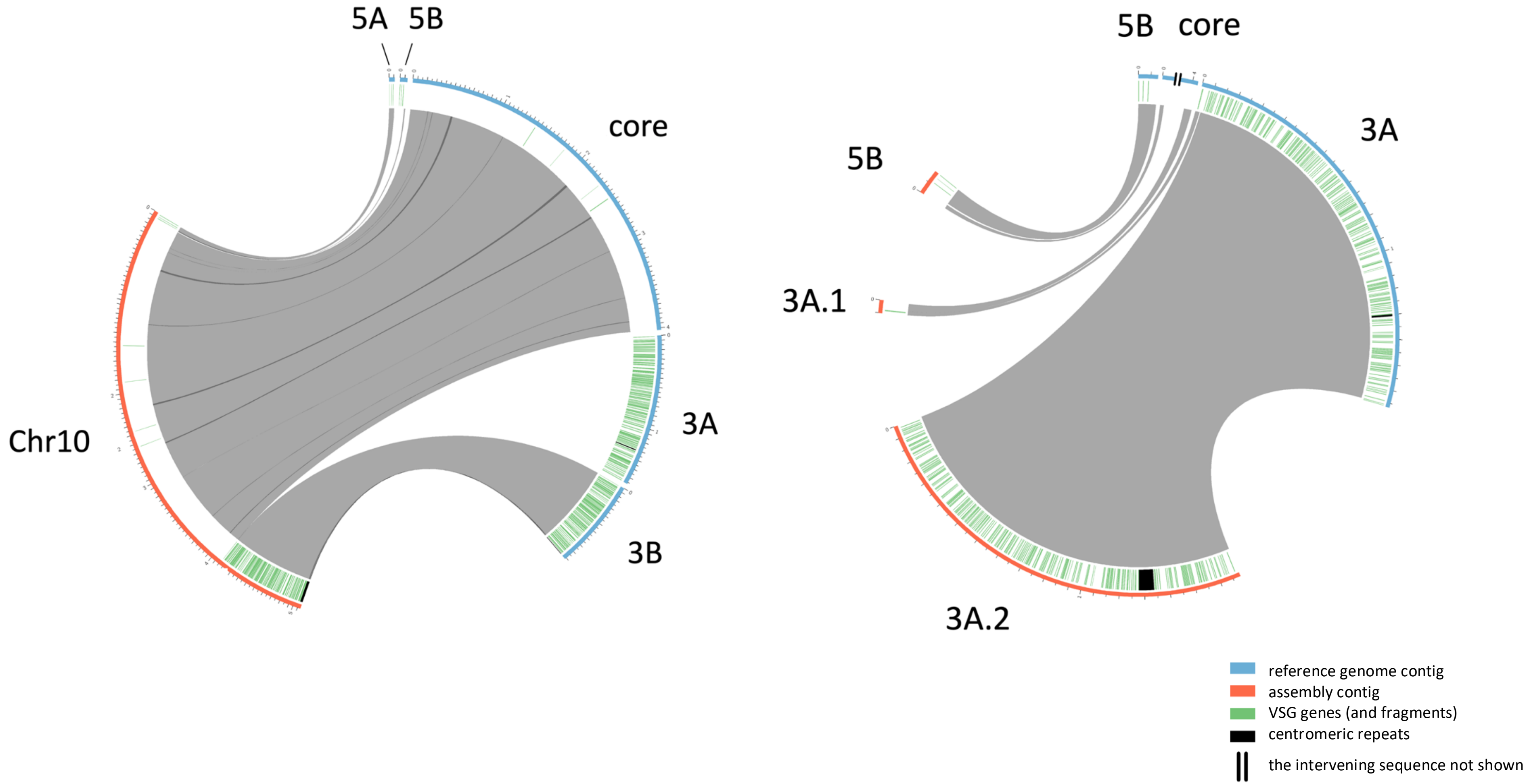


Figure 19 Contig structure of chromosome 10 in the new assembly.

Circos plots highlighting the overlap between the 2018 Lister 427 reference genome (Müller *et al.*, 2018) and the new assembly. Grey ribbons represent overlaps, black ribbons – multiple overlaps within a region (repetitive sequence).

3.3.1.3 Chromosome 5

Chromosome 5 is represented by 4 contigs in the reference assembly – subtelomeres 3A and 3B, a bloodstream-form expression site, and a core contig (Figure 20).

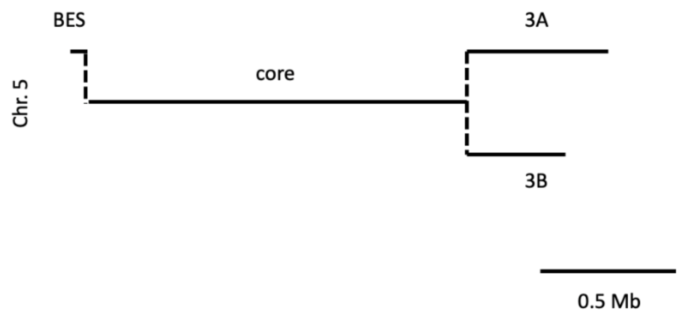


Figure 20 A diagrammatic representation of chromosome 5 structure in the reference genome.

Figure adapted from Müller *et al.*, (2018). Created with BioRender.com.

Consistent with the reference assembly, in the new assembly there was a contig that bridges the core contig with 3A on one end, and BES5 on the other (Figure 21 A). There was also a second contig that bridges the core of this chromosome with the 3B subtelomere (Figure 21 B). The former contig closes three scaffold gaps present in the reference, as well as the two 'gaps' between the separate contigs (BES, core, subtelomere); among these gap closures we find an expanded, non-truncated and thus possibly complete centromeric repeat region (Figure 21 A).

The second contig joined the 3' end of the core contig with the other subtelomere - 3B (Figure 21 B). As above, several gap closures relative to the reference can be seen here – two scaffold gap closures as well as the closure of the core-subtelomere 'gap'. In all three gap closures, as well as the distal contig end (Figure 21 B, regions indicated by black triangles), synteny with the reference was interrupted, as a variety of other DNA sequences were incorporated in the gaps – some from unitigs, others from other chromosome cores, and subtelomere contigs (Table 19 in Appendices).

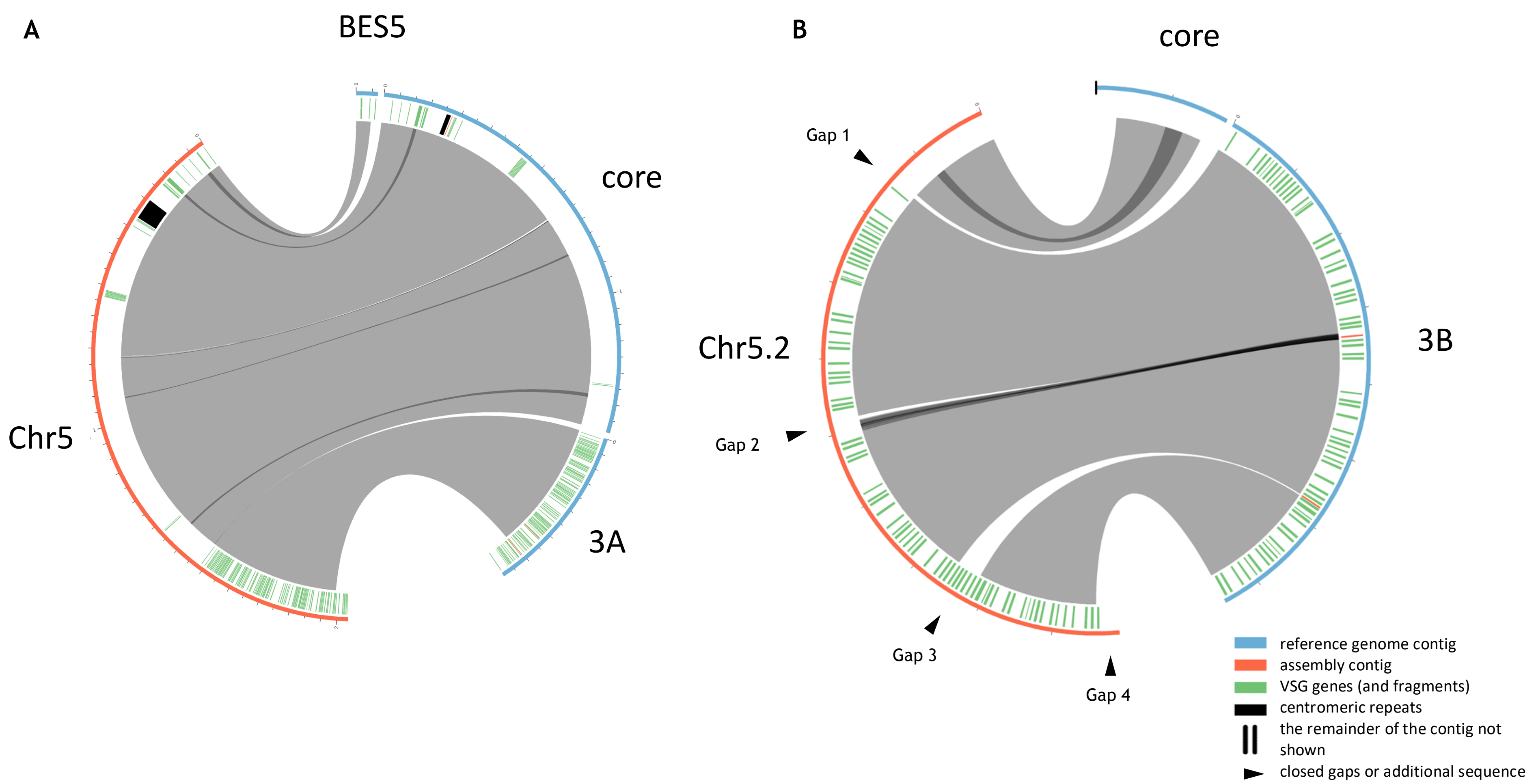


Figure 21 Contig structure of chromosome 5 in the new assembly

Circos plots highlighting the overlap between the 2018 Lister 427 reference genome (Müller *et al.*, 2018) and the new assembly. Grey ribbons represent overlaps, black ribbons – multiple overlaps within a region (repetitive sequence). For gaps 1-4 overlapping reference sequence detailed in Table 19 in the Appendices.

3.3.2 More subtelomeric sequences for chromosomes 2 and 7

In addition to bridging known and previously assigned subtelomeric and core regions, the genome assembly presented here also recovered a couple of subtelomeres not previously assigned. For chromosome 2, tig111 contains sequences flanking the core on both sides, and neither sequence matches the sole 5A subtelomere annotated for this chromosome in the 2018 genome (Figure 22). In keeping with the 2018 genome nomenclature, this '5B' sequence extends to the telomeric repeats, whereas the sequence flanking the core on the other end of the contig does not.

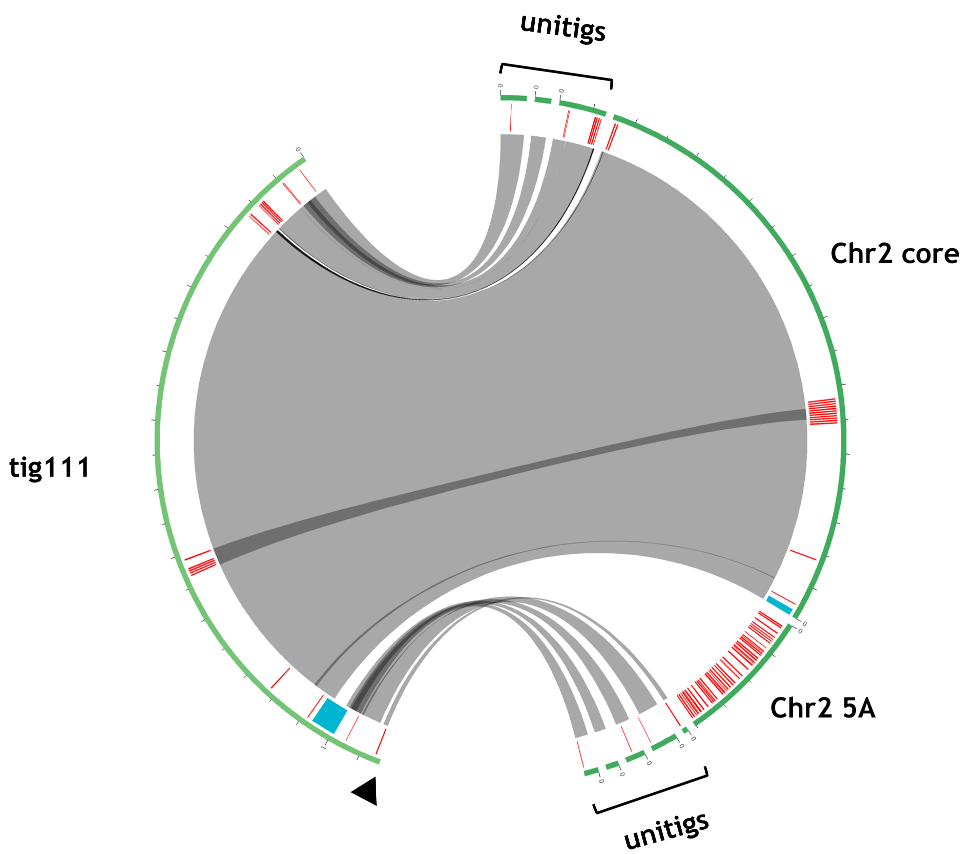


Figure 22 Chromosome 2 contig contains full-length centromeres and previously unassigned sequence.
A circos plot showing the overlap between tig111 in the new genome assembly and respective contigs in the 2018 reference genome (Müller *et al.*, 2018). Putative *VSG* gene locations are shown in red, and the overlaps (links) between the two genomes are shown as grey bands. The black triangle indicates the location of telomeric repeats, the light blue – of centromeric repeats.

For chromosome 7, an alternate subtelomere has been assembled in contig 4860 - what would be subtelomere 5B. It contains a portion of 5A, but also fragments of many other subtelomeres; this contig is described in more detail in section 3.3.5.1 and Figure 29 below.

3.3.3 Bloodstream-form expression site assembly and incorporation

3.3.3.1 BES recovery overview

Assembly-vs-assembly comparison also led to the recovery of 15 known bloodstream form expression site sequences in the new assembly (Table 5). Out of these, 10 contained full-length 50bp repeat regions at the 5' end, as well as some sequence upstream, thus providing a genomic and sequence context for these expression sites. On the 3' end, 3/15 contained telomeric sequences. Only two expression sites contained the entire BES sequence, as well as flanking regions on both ends (1 and 2 in Table 5, Figure 23). Surprisingly, a large proportion on the BES sequences in the new assembly terminated within the 70bp repeat regions – 7/15 terminate within these regions, and a further 3/15 within the adjoining CTR region, immediately upstream of the telomere-adjacent VSG. Additionally, 3/15 contigs with BES also contained 177bp repeats, indicating the contig may be that of non-megabase chromosomes (more on that later). While most of the recovered BES are incomplete in their sequence compared to previous work (Hertz-Fowler *et al.*, 2008; Müller *et al.*, 2018), the upstream sequence provides context and utility for examining DNA replication dynamics. Further analysis suggesting a possible reason for the difficulty in spanning the 70bp repeats in this genome assembly is discussed later in 3.4.

Table 5 Bloodstream-form expression sites and their sequence elements.
A summary of the elements present in the uncovered BES sequences in the new assembly. UP – presence of sequence upstream of the 50bp repeat region; BES body – genes located between the 50bp repeats and the telomere-adjacent *VSG*; CTR – co-transposed region immediately upstream of the telomere-adjacent *VSG*; TTAGGG – presence of the telomeric repeats at the 3’ end of the BES; ID – most likely identity of the BES in relation to the nomenclature used in the reference genome; CHR – the chromosome the corresponding BES appears to be on; CHR PUB – the correspondence of said BES to a chromosome in the reference genome or published literature. Int – intermediate chromosomes (based on the presence of 177bp repeats).

NO	CONTIG	UP	50BP	BES BODY	70BP	CTR	BES VSG	TTAGGG	ID	CHR	CHR PUB	NOTES ON THE BES OR CONTIG
1	tig00000137	✓	✓	✓		✓	✓	✓	BES10	7 or int	int	no 70bp repeats
2	tig00000644	✓	✓	✓	✓	✓	✓	✓	BES8?	8	8	very few genes in BES body
3	tig00000653		✓	✓	✓	✓	✓	✓	BES1		6	
4	tig00000114		✓	✓	✓	✓	✓		BES12		2	
5	tig00000642	✓	✓	✓	✓	✓	✓		BES15	3	3	two 70bp repeat regions
6	tig00004876	✓	✓	✓	✓	✓			BES11	int	int	177bp repeats
7	tig00004877		✓	✓	✓	✓			BES11		int	
8	tig00004878		✓	✓	✓	✓			BES13		int	
9	tig00000055	✓	✓	✓	✓				BES5	5	5	
10	tig00000156	✓	✓	✓	✓				BES14		7	
11	tig00000157	✓	✓	✓	✓				BES4		int	
12	tig00000658	✓	✓	✓	✓				BES7	int	6	177bp repeats
13	tig00004879	✓	✓	✓	✓				BES13	int	int	177bp repeats and telomere at the 5’ end
14	tig00000116		✓	✓	✓				BES3		4	
15	tig00000652	✓	✓	✓					BES1		6	

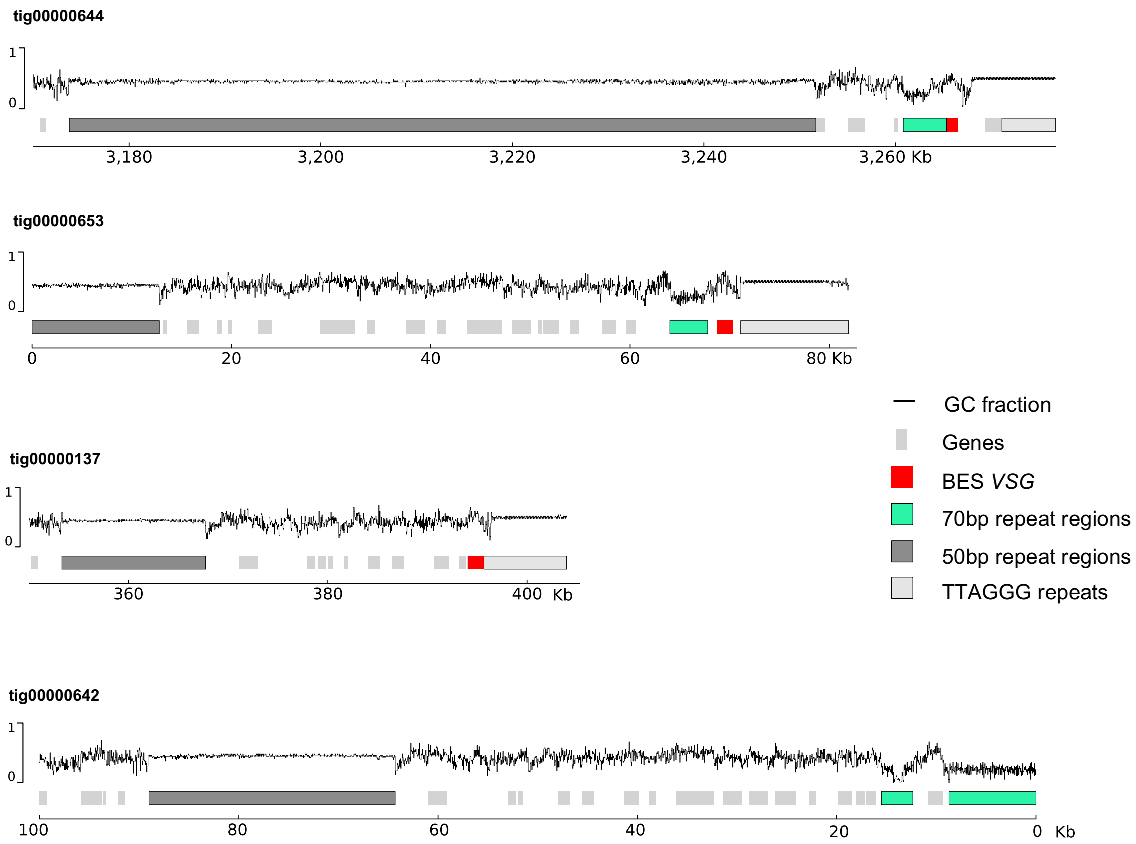


Figure 23 Examples of BES structure.

The four most complete BES sequences uncovered in the new assembly are shown. Note: for tig642, the orientation of the contig and associated annotations has been switched left-to-right for consistency.

3.3.3.2 Genomic context of the identified expression sites

The Nanopore assembly extends our understanding of the genomic context of the VSG expression sites. In the reference genome, the BES are isolated separate sequences, and, for several of them, their association with particular chromosomes has been elucidated through Hi-C data; according to the authors of the current reference, BES8 was the only expression site that was attached (through sequence) to a chromosome in their assembly (Müller *et al.*, 2018). In the current assembly, we managed to identify four BES sequences as belonging to specific megabase chromosomes, and three belonging to intermediate chromosomes, based on the fact that they also contain 177bp repeats, which are a hallmark for these shorter chromosomes. For BES5, BES8, BES11, BES13 and BES15 the chromosome identity/type matches the previously published data

(Hertz-Fowler *et al.*, 2008; Müller *et al.*, 2018), whereas for BES7 and BES10 there appears to be a disagreement.

3.3.3.3 BES7

The contig containing BES7 in our assembly also contains 177bp repeats, indicating that it is likely an intermediate chromosome sequence; previously published data suggests this expression site is located on a megabase chromosomes exceeding 3.1 Mb in size (Hertz-Fowler *et al.*, 2008), whereas the authors of the reference assembly have placed it with chromosome 6 (Müller *et al.*, 2018). In our assembly the contig it is located on, tig00000658, is just over 384 kb in size and contains no known sequences belonging to megabase chromosomes. It is likely that this expression site, therefore, has translocated in the cells used for assembly here.

3.3.3.4 BES10

The location of BES10 in our assembly is somewhat ambiguous, though it is likely to be either on chromosome 7 or an intermediate chromosome; previously published data suggests it's located on an intermediate chromosome of approximately 450kb in size (Hertz-Fowler *et al.*, 2008). Tig00000137, harbouring this expression site, is just under 404 kb, and contains an approximately 153 kb overlap with the chr 7 core region. The remainder of the contig consists of 12 previously unassigned sequences (unitigs). It remains unclear whether this contig, harbouring BES10, is a previously unidentified subtelomeric sequence of chromosome 7, or whether this contig represents an intermediate chromosome that contains a translocation from it. There are no 177bp repeats on this sequence, and therefore it remains unclear whether this could be an intermediate chromosome.

3.3.3.5 BES1, BES3, BES4 and BES11-14

The remainder of the BES sequences recovered cannot be assigned to any particular chromosome or type of chromosome with confidence. BES3, BES12 and one of the copies each of BES1, BES11 and BES13 do not contain sequences upstream of the 50bp repeat areas, and therefore their localisation remains unknown. As for BES4, BES14 and the other copy of BES1 – the contigs that contain these, similarly to that of BES10 discussed above, are predominantly composed of unitigs, with very small overlap with known megabase sequences and no 177bp repeats.

3.3.4 Repetitive sequence identification and analysis

A major advantage of third-generation sequencing technologies for genome assembly is the potential to resolve repetitive regions by offering very long sequence reads. In the current assembly, some of the major repetitive regions have been incorporated; these include the centromeric repeats, the sub-megabase chromosome-associated 177bp repeats, as well as the two groups of repeats associated with bloodstream-form expression sites – so called 50bp and 70bp repeats.

3.3.4.1 70bp repeats

As discussed in 3.1.2.1, the so-called 70bp repeats sit immediately upstream of the *VSG*, including in bloodstream form expression sites. They are known to be overall AT-rich sequences, with shorter GC-rich regions, that separate the majority of the BES genes from the arguably most important one – the *VSG* - and their sequence is said to be a conserved motif of 75-77bp in length (Figure 12) (Hovel-Miner *et al.*, 2016). In the current assembly, a total of 14 BES-associated 70bp repeat regions have been recovered, first using assembly-vs-assembly sequence comparison with the current Lister 427 strain reference from 2018, then using tandem repeats finder (Benson, 1999) for a more precise and detailed breakdown of the repetitive elements.

Table 6 Length of 70bp repeat regions in the new assembly.
Q1 – lower quartile , Q3 – upper quartile.

	Full-length, bp	Partial, bp
	n = 7	n = 7
min	2154	2387
max	17302	21189
median	4358	8904
mean	7579.7	10027
Q1	3152	4435
Q3	17155	15170

Based on the presence of flanking sequences or lack thereof, 7 of these regions appear full-length, whereas the other 7 are partial repetitive regions, as these are truncated in the assembly. The full-length regions were on average 7580 bases long (mean, median 4358 bases), whereas the partial length regions were on average 10027 bases long (mean, median 8904 bp) (Table 6). Though the latter seem to be larger, the difference is not statistically significant (p = 0.50) (Figure 24). The GC content of the 70bp repeat regions was found to be between 17% and 24%, consistent with these regions being AT rich.

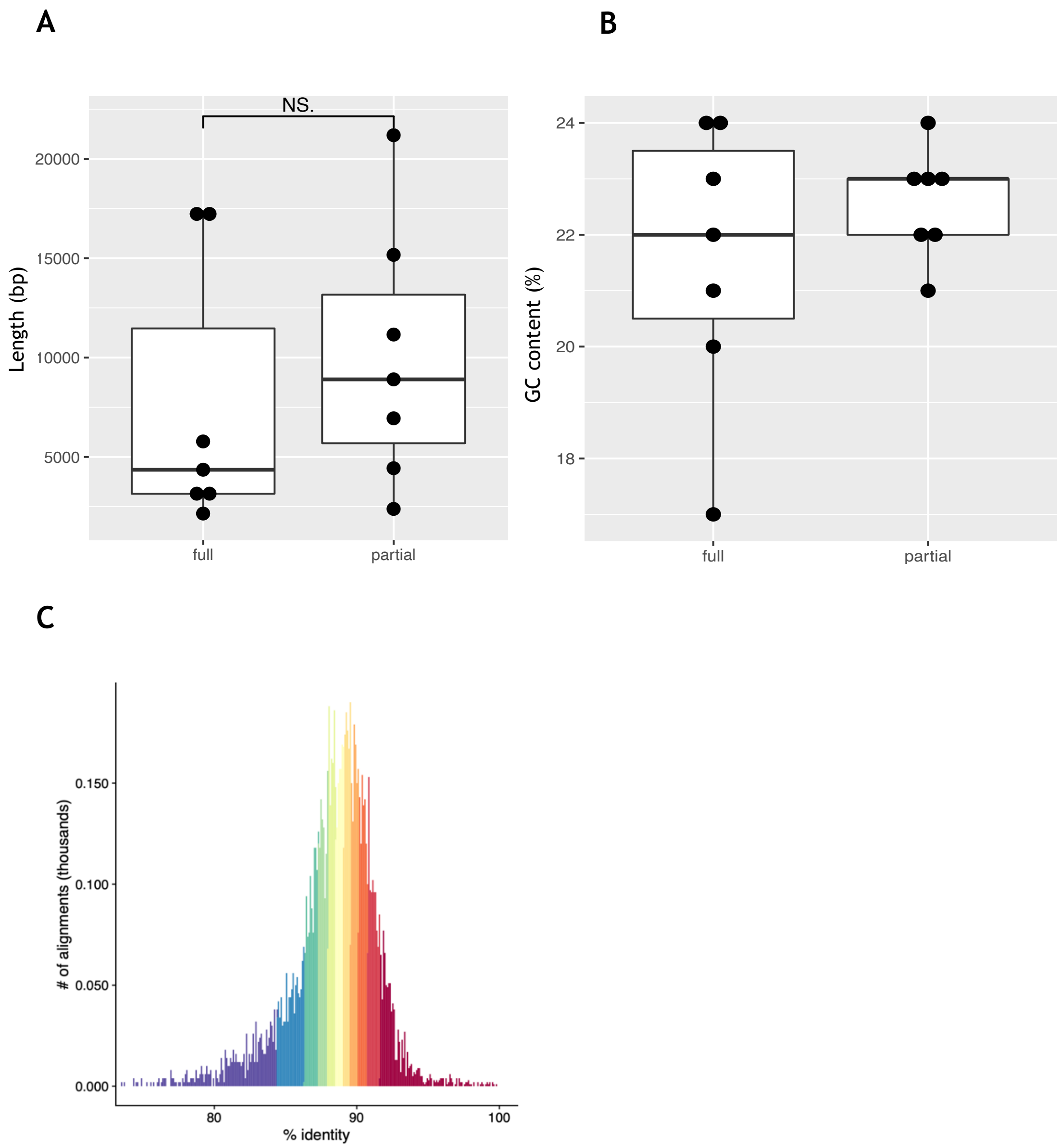


Figure 24 Length, GC content and intra-region pairwise sequence identity of complete and partial 70bp repeat region sequences.

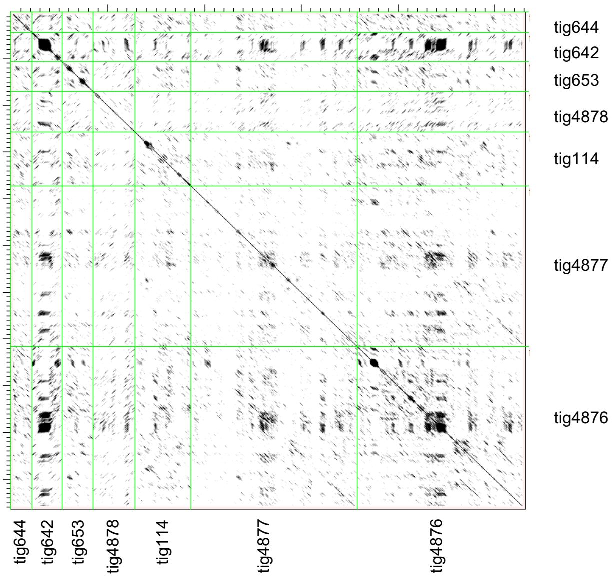
A – length of full-length and partial (truncated) 70bp repeat regions identified in the assembly. B – GC% of 70bp regions. C – sequence identity in intra-region pairwise comparison using StainedGlass (Vollger *et al.*, 2022).

Repeat structure analysis was performed using Tandem repeats finder (TRF) – a tool that scans a given fasta sequence (in this case, the entire assembly) for any tandemly repeated sequences (Benson, 1999). Unlike some other repetitive sequences that will be analysed later, the 70bp repeat regions appear highly complex and varied, being a composite of multiple repeats, and thus any meaningful visualisation is problematic. In order to examine any possible underlying structure to the 70bp repeat regions, we used Dotter (Sonnhammer and Durbin, 1995) to show pairwise mapping within and between the repeat regions (Figure 25). Out of the seven full-length 70bp repeat regions examined, dotter highlighted the presence of a more conserved core in three of the repetitive regions (on contigs 642, 4876 and 4877), but not others.

To investigate whether the previously published (Hovel-Miner *et al.*, 2016) so-called 70-bp repeat region motif is present in the assembly, FIMO (Grant, Bailey and Noble, 2011), a tool from the MEME suite of tools, was used to query the assembly for that motif specifically. With a significance cut-off of $p < 10^{-9}$, a total of 15782 occurrences of this motif were found using this approach, spanning each of the identified 70bp repeat regions.

The two tools, TRF and FIMO, do not appear to be in agreement regarding the structure of the 70bp repeat regions, as FIMO predicts the aforementioned 77bp motif is present across the repeat regions with few gaps in between, whereas, as mentioned before, TRF suggests there are many and varied repeat elements throughout the region, without forming an obvious pattern. One possible explanation for this discrepancy is that FIMO may be more permissive to mismatches and gaps than TRF.

A



B

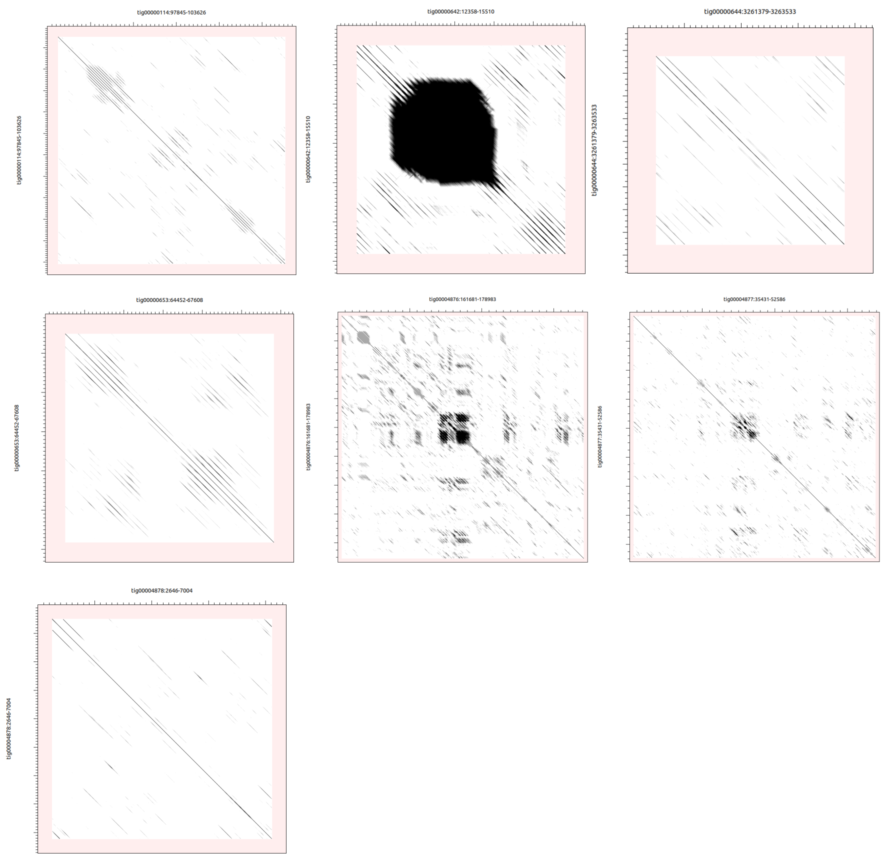


Figure 25 Full-length 70bp repeat region structure as visualised using Dotter.
Dotter (Sonnhammer and Durbin, 1995) images showing pairwise 70bp repeat region alignment and similarity, with the darker regions highlighted the regions with the highest similarity. A. Comparing all full-length 70bp repeat regions (all vs all), B. intra-region similarity for each full length 70bp repeat region.

3.3.4.2 50bp repeats

The 50bp repeats lie upstream of the BES and are predicted to form a boundary between the expression sites and the rest of the chromosome (Sheader *et al.*, 2003). In the ONT assembly, 10 full-length and 8 truncated 50bp repeat regions were identified (Table 5); 5 of the truncated regions are on contigs that contain BES sequence (truncated at the 5’ end), whereas the remaining 3 are truncated before reaching the BES (at the 3’ end). As above in the case of 70bp repeats, the term ‘full-length’ here is used based on the presence of flanking non-repetitive sequence.

The full-length regions were on average 37477 bases long (mean, median 40387), whereas the partial length regions were on average 28441 bases long (mean, median 28013) (Table 7, Figure 26). The full-length 50bp regions vary considerably in length, ranging from 4866 bp to 77923 bp.

Table 7 Length of full-length and partial 50bp repeat region sequences.
Q1 – lower quartile, Q3 – upper quartile.

	Full-length, bp	Partial, bp
	n = 10	n = 8
min	4866	1728
max	77923	65486
median	40387	28013.5
mean	37477.1	28441.25
Q1	25251.25	17316.25
Q3	48887.5	32946

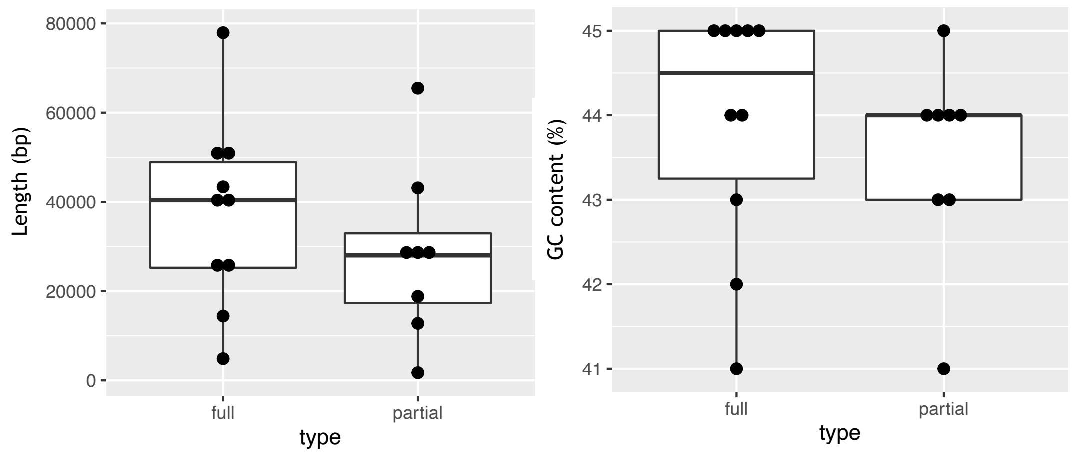


Figure 26 Length and GC content of complete and truncated 50bp repeat region sequences.

The GC content of the 50bp repeats was found to be 41-45% - a more narrow range than the 70bp regions (17-24%). In addition, again unlike the 70bp repeats, the 50bp repeat motif and structure was found to be highly conserved – a single 50bp long motif is tandemly repeated with average identity of 91-95% across these regions (Figure 27). Intra-region identity heatmaps (Figure 28) suggest varying, albeit high, conservation across a given repeat region, with most displaying an ultra-conserved core (above 96% identity, shown in maroon); the position of this ‘core’ doesn’t appear to be dependent on the relative proximity to the expression site.

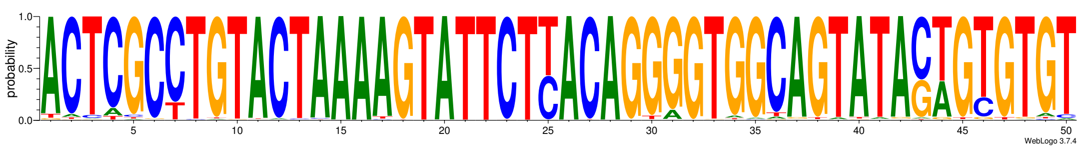


Figure 27 50bp motif of *T. brucei*.

Summary 50bp motif sequence of the full-length 50bp regions with p value below 10^{-9} according to FIMO (Grant, Bailey and Noble, 2011). Figure produced using WebLogo (Crooks *et al.*, 2004). This motif is consistent with the previously reported sequence (Zomerdijk *et al.*, 1990).

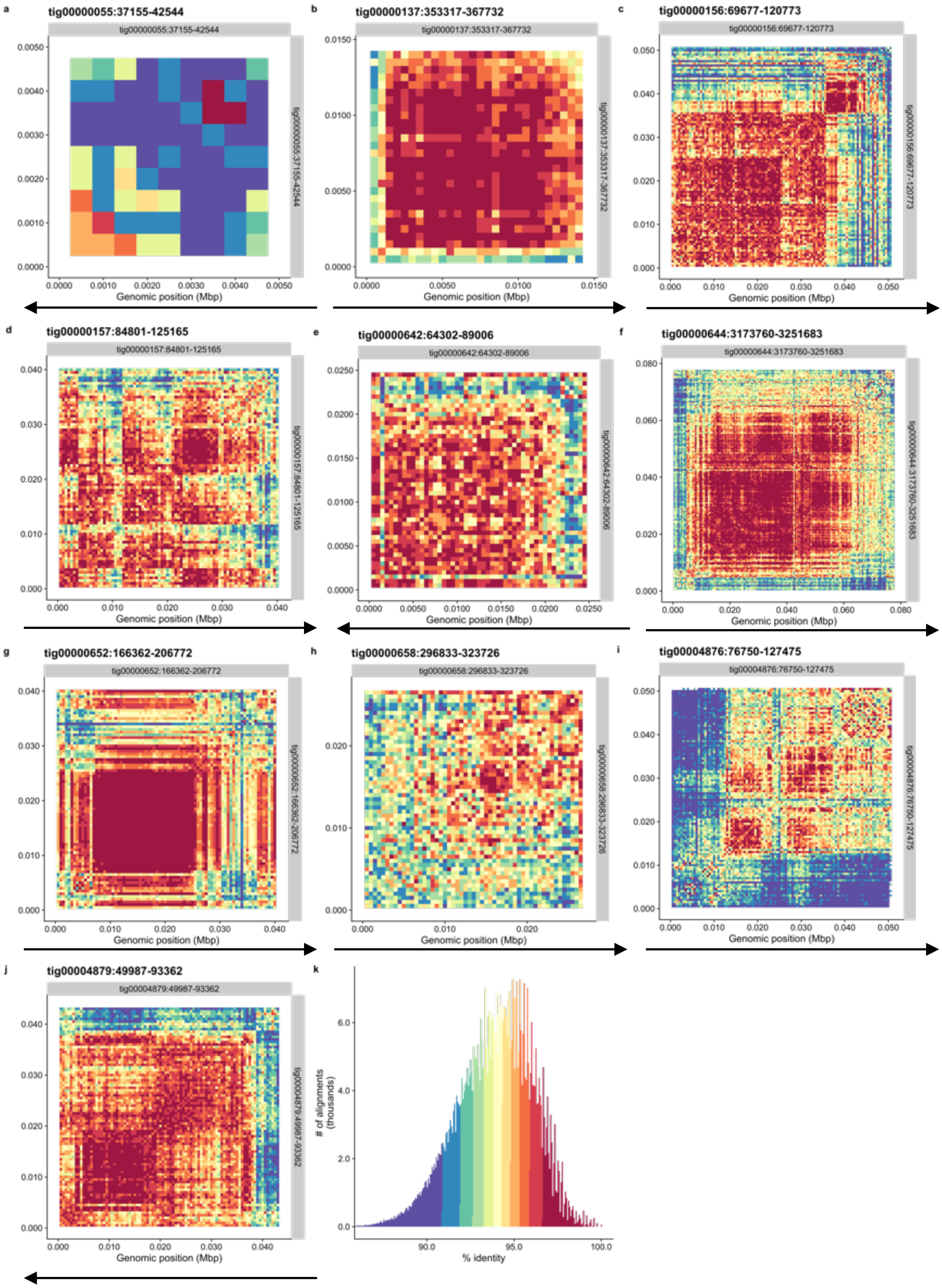


Figure 28 50bp repeat region intra-region identity.
Intra-region pairwise sequence identity heatmaps of all full-length 50bp repeat regions. Arrows below the heatmaps indicate the direction towards the nearest BES. Figure produced using StainedGlass (Vollger *et al.*, 2022).

3.3.5 A note on scaffold gap closure

For completeness, we also attempted to evaluate scaffold gap coverage in the new genome assembly relative to the 2018 genome. In order to do this, we extracted the locations of all stretches of N's in the reference fasta and assessed the reciprocal mapping of the two genomes around the identified loci. In total, 49 scaffold gaps were identified in the Muller genome – 2 in BES contigs, 12 in megabase chromosome 'core' contigs, and 35 in subtelomeric contigs of megabase chromosomes. 43/49 of these gaps were closed in the new genome assembly, and the remaining 6 were either not closed or difficult to evaluate due the complex nature of reciprocal mapping and contiguity breaks at those sites (Table 20). Below, we present tig4860 (chromosome 7) as an example of scaffold gap closure scenarios encountered (Figure 29, Table 8).

3.3.5.1 Chromosome 7 (tig4860)

One of the longer contigs in the new genome assembly is 4860; this contig overlaps 6 scaffold gaps in the reference genome (numbered 1-6 in Table 8 and Figure 29). The 1st gap, originating in the core contig, represents the centromeric repeat region, and is greatly expanded in both contigs which partially cover the gap, but neither contig contains the full centromeric repeat region and both flanking sequences. Gaps 3 and 5 appear closed via the incorporation of unitigs and a small fragment of a subtelomeric sequence from chromosome 6 subtelomere 3A. Gap 4 bridges two subtelomeric contigs, acting as a breakpoint, whereas gap 2 appears to contain a 19bp insertion. Based on assembly vs assembly comparisons, gap 6, similarly to gap 2, would be expected to also contain just a short stretch of sequence, as there is no interruption in the assembly-to-assembly mapping. However, unlike for gap 2, Clustal Omega multiple sequence alignment (MSA) suggested a possible ~1kb insertion, though the sequence similarity surrounding the scaffold gap was relatively low, providing little confidence in this observation.

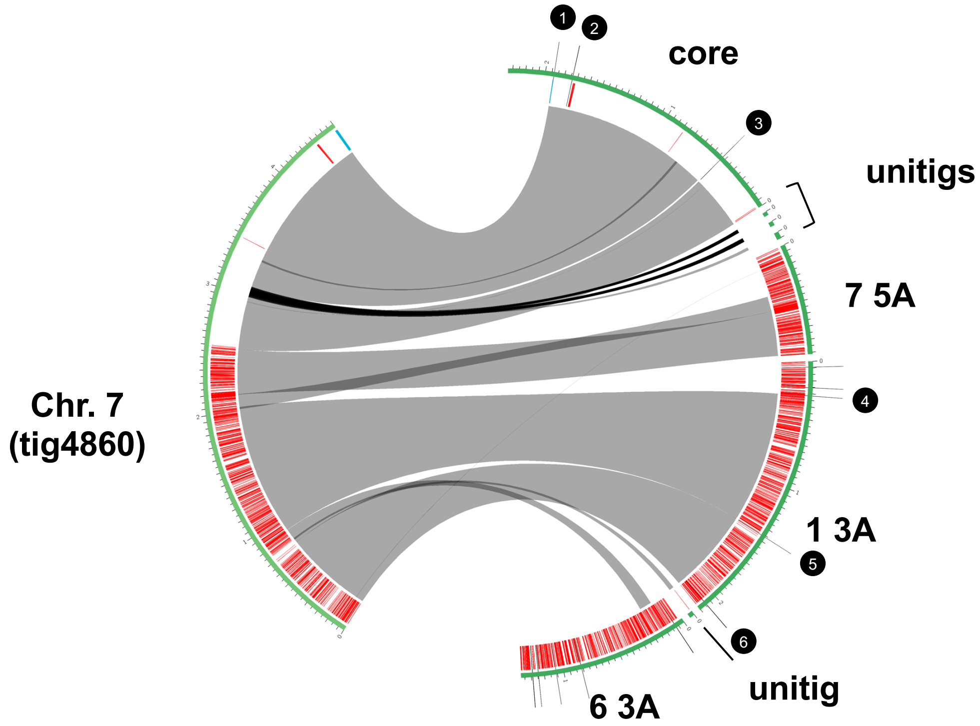


Figure 29 Scaffold gap closure examples for tig4860 (chromosome 7).
A circos plot showing the overlap between tig4860 in the new genome assembly and respective contigs in the 2018 reference genome (Müller *et al.*, 2018). Putative *VSG* gene locations are shown in red, and the overlaps (links) between the two genomes are shown as grey bands. Scaffold gaps mentioned are highlighted with numbers (1-6), and other scaffold gaps (not covered by tig4860) are highlighted as thin black lines on the outer edges of the plot.

Table 8 Scaffold gap resolution in tig4860 in relation to the 2018 reference genome.

#	Reference contig	Gap start coordinates	Contigs covering the gap	Closed?	Notes
1	Chr7_core_Tb427v10	1933353	tig00004860; tig00000127	No	Centromeric repeat region extended in both contigs
2	Chr7_core_Tb427v10	1789177	tig00004860	Closed	Short 19bp sequence insertion
3	Chr7_core_Tb427v10	473368	tig00004860	Closed	Unitigs incorporated
4	Chr1_3A_Tb427v10	264716	tig00000069; tig00004860	Closed	In tig4860, two subtelomeric contigs bridged
5	Chr1_3A_Tb427v10	1409580	tig00004860	Closed	Unitig and Chr6 3A subtelomere fragment inserted
6	Chr1_3A_Tb427v10	2090279	tig00004860	Closed	Possible ~ 1kb sequence insertion

3.4 Base quality evaluation at 70bp repeats and chromosome ends

As noted in 3.3.4.1 and 3.3.2, many of the recovered BES sequences truncated in or around the 70bp repeats, despite their relatively modest size. We decided to investigate whether this might be explained by lower base quality at these repeats. To do this, ONT reads >10kb in length and spanning the 70bp regions, as well as 20kb flanking regions upstream and downstream (where available), were extracted and the median base quality, in 100bp bins, evaluated. To our surprise, 70bp repeats did not display an apparent drop in base quality relative to the upstream sequence, even where read depth coverage was reduced (see Figure 30 A and B for examples). However, a drop in base quality was noted further downstream – at telomeric repeats (Figure 30 B); in fact, this was the case at telomeric repeats that are not associated with BES as well (Figure 30 C). The reasons for lower base quality at telomeric repeats of *T. brucei* are unknown, although some likely culprits are presented in the discussion for this chapter; this lower base quality might explain two unexpected hurdles encountered in this genome assembly – difficulty sequencing across 70bp repeats and assembly of minichromosomes (see Discussion).

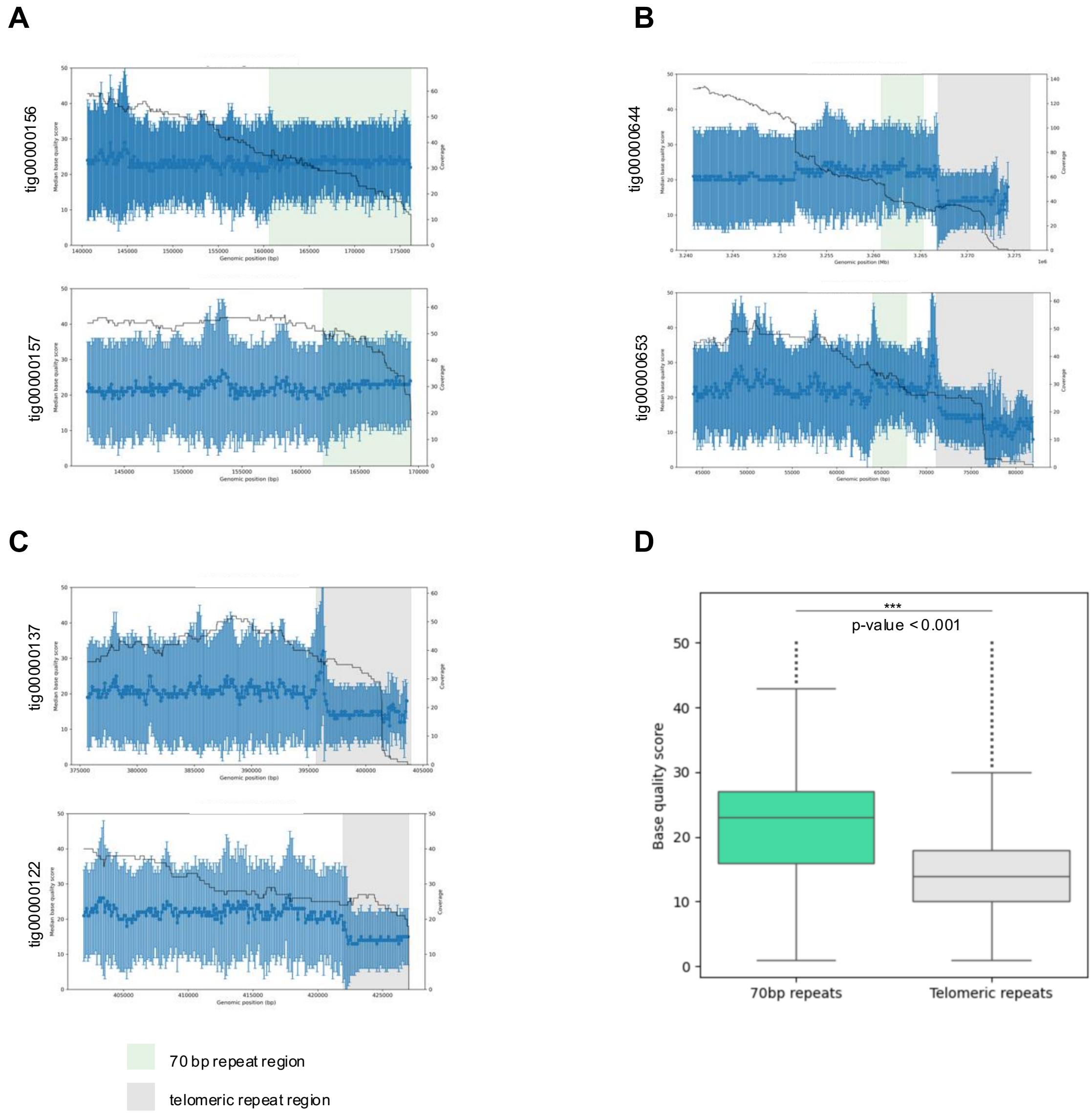


Figure 30 – Telomeric repeats of *T. brucei* represent a challenge for Nanopore DNA sequencing.

Median base quality, along with the interquartile range, is plotted for representative 70 bp-and -telomeric repeat-containing genomic regions; only >10 kb reads represented. A. Two examples of truncated 70 bp regions (green) with flanking sequence. B. Examples of full-length 70 bp repeat regions (green) as well as telomeric repeats (grey). C. Examples of other telomeric repeat-containing genomic regions (grey). D. Boxplots showing the distribution of base qualities across all 70 bp regions and all telomeric regions; p-value < 0.001, significance tested using a Mann-Whitney U test (scipy package, v.1.10.1).

3.5 Sub-megabase chromosome recovery and characterisation

3.5.1.1 Overall structure

In addition to the so-called megabase chromosomes, the assembly contains a number of contigs representing likely sub-megabase chromosomes; we have decided to categorize the identified sequences as such instead of following the established nomenclature (mini chromosomes and intermediate chromosomes) as we cannot ascertain which group these sequences belong to. Sub-megabase chromosomes have been largely overlooked in published assemblies. A total of 8 candidate sub-megabase contigs were identified based on the incorporation of the characteristic 177bp repeats (Figure 31 B). None of the candidates are telomere-to-telomere assembled, therefore it is not possible to tell whether or not the 8 sequences represent 8 distinct chromosomes or, perhaps, are parts of larger chromosomes that are fragmented.

Curiously, all eight sequences contain long stretches of predicted protein coding genes, not dissimilar to megabase chromosomes, a finding that is contrary to the *status quo*, as it has been long asserted that the smaller chromosomes of *T. brucei* do not contain protein coding genes other than *VSGs* or those in BES. The assumption arose from Southern blotting-based experiments that indicated that there are no rRNA genes on sub-megabase chromosomes (Van der Ploeg *et al.*, 1984). However, in later detailed mapping, non-repetitive regions within 100 kb of the mapping centre were detected but not characterised (Wickstead, Ersfeld and Gull, 2004), though previous BAC analysis suggested they may correspond to BES sequence (Berriman *et al.*, 2002). It has been long thought that the smaller chromosomes act as *VSG* archives and contain only *VSGs*, repetitive sequences and BES (Wickstead, Ersfeld and Gull, 2004); the current assembly shows that, at least for the chromosomes captured here, that is not the case.

The largest four sub-megabase chromosomes captured contain BES and/or the associated 50bp repeats (Figure 31 B). As mentioned above, none of the assembled sequences are full-length, or telomere-to-telomere. However, it appears there is a particular structure inherent to these shorter chromosomes, with 177bp repeat regions positioned immediately adjacent to a telomere on one end, followed by a section of protein-coding genes and, in some cases, a BES at the other end of the chromosome (Figure 31 A). Assembly versus assembly comparisons showed no significant overlap with known megabase chromosome sequences; instead, the vast majority of the sub-megabase chromosomes mapped to unitigs - short and unidentified sequences in the reference assembly (Figure 32), further indicating that these are, in fact, likely sub-megabase chromosomes.

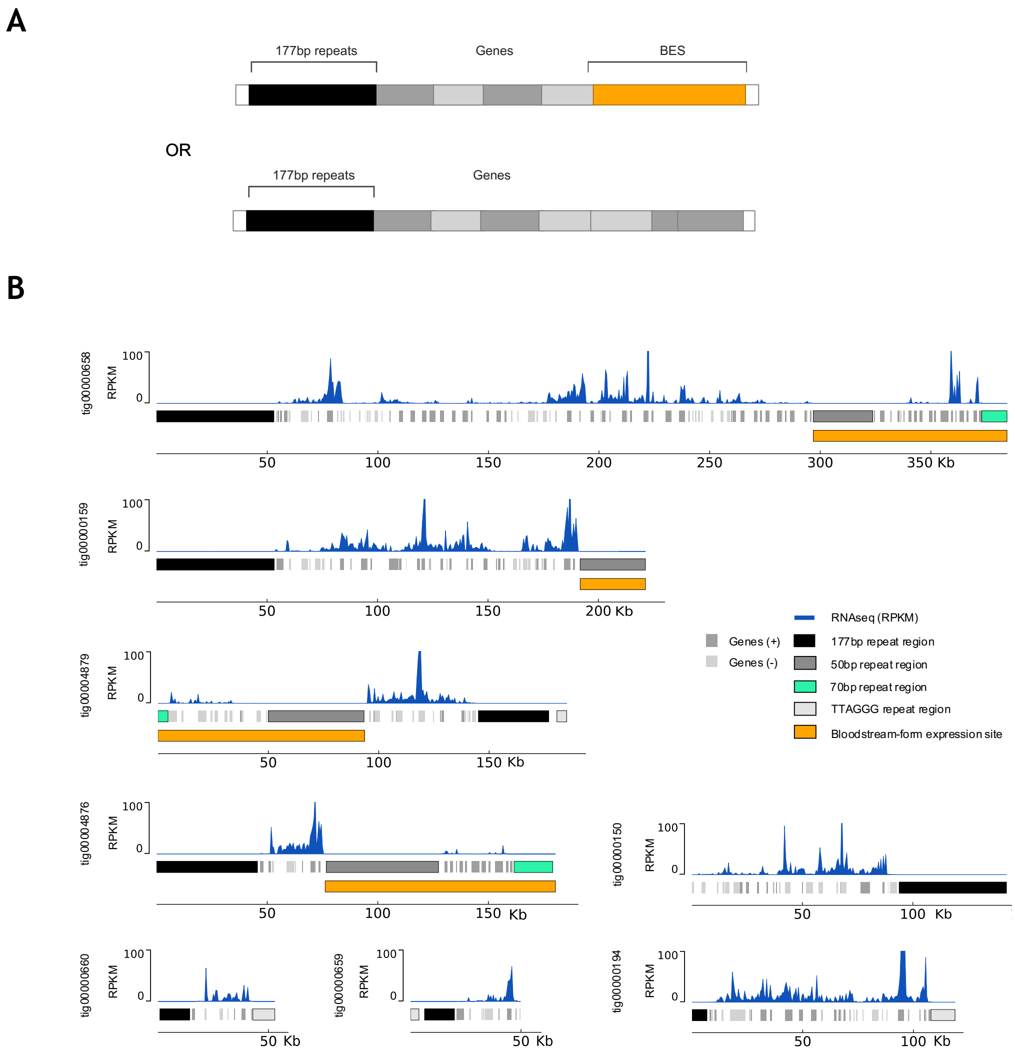


Figure 31 Structure of identified sub-megabase chromosomes in the new assembly
A – proposed general structures of the sub-megabase chromosomes. B – Structure of the identified sub-megabase chromosomes, with RNAseq mapping shown in blue. Elements of panel A were created with BioRender.com.

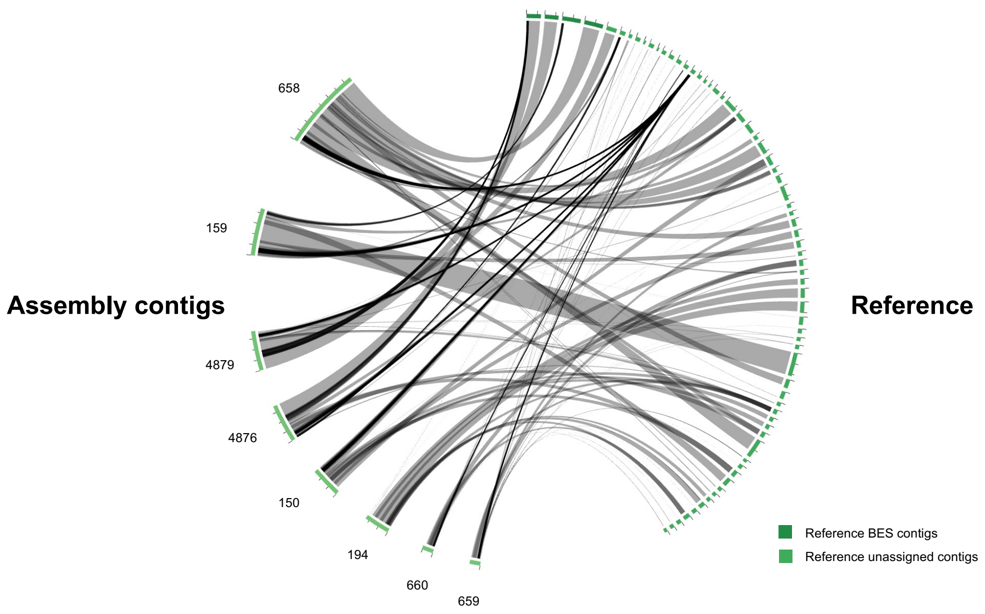


Figure 32 Newly assembled sub-megabase chromosomes encompass many formerly unassigned unitigs.
Correspondence of the newly assembled sub-megabase chromosomes to unassigned sequences in the reference genome – unitigs – is shown. Grey ribbons show overlaps between the assemblies, black ribbons – overlaps that contain several sequences (usually repetitive DNA).

3.5.1.2 Sub-megabase chromosome gene content

In order to analyse the gene content of sub-megabase chromosome contigs, we analysed the differential distribution of protein (pfam) domains and performed orthology-based searches using OrthoMCL (Li, Stoeckert and Roos, 2003; Alvarez-Jarreta *et al.*, 2024). First, looking at orthology-based assignments, hypothetical protein coding genes comprised the biggest single group – 107 genes, representing a broad range of orthogroups (Figure 33 A) - while expression-site associated genes (*ESAGs*) and retrotransposon hotspot (*RHS*) proteins were also numerous, representing a further 131 genes. Surprisingly, *VSGs* were not in abundance, and the lower density of *VSGs,* relative to the rest of the genome, was also reflected in the pfam protein domain analysis (Figure 33 B marked with asterisks, Table 21 in Appendices), where three groups of domains found in *VSG* genes were significantly less likely to be found on the sub-megabase chromosome contigs. In addition, there appeared to be a statistically significant enrichment in retrotransposon hotspot (*RHS*) protein domains in the sub-megabase chromosome contigs, consistent with the orthogroup analysis.

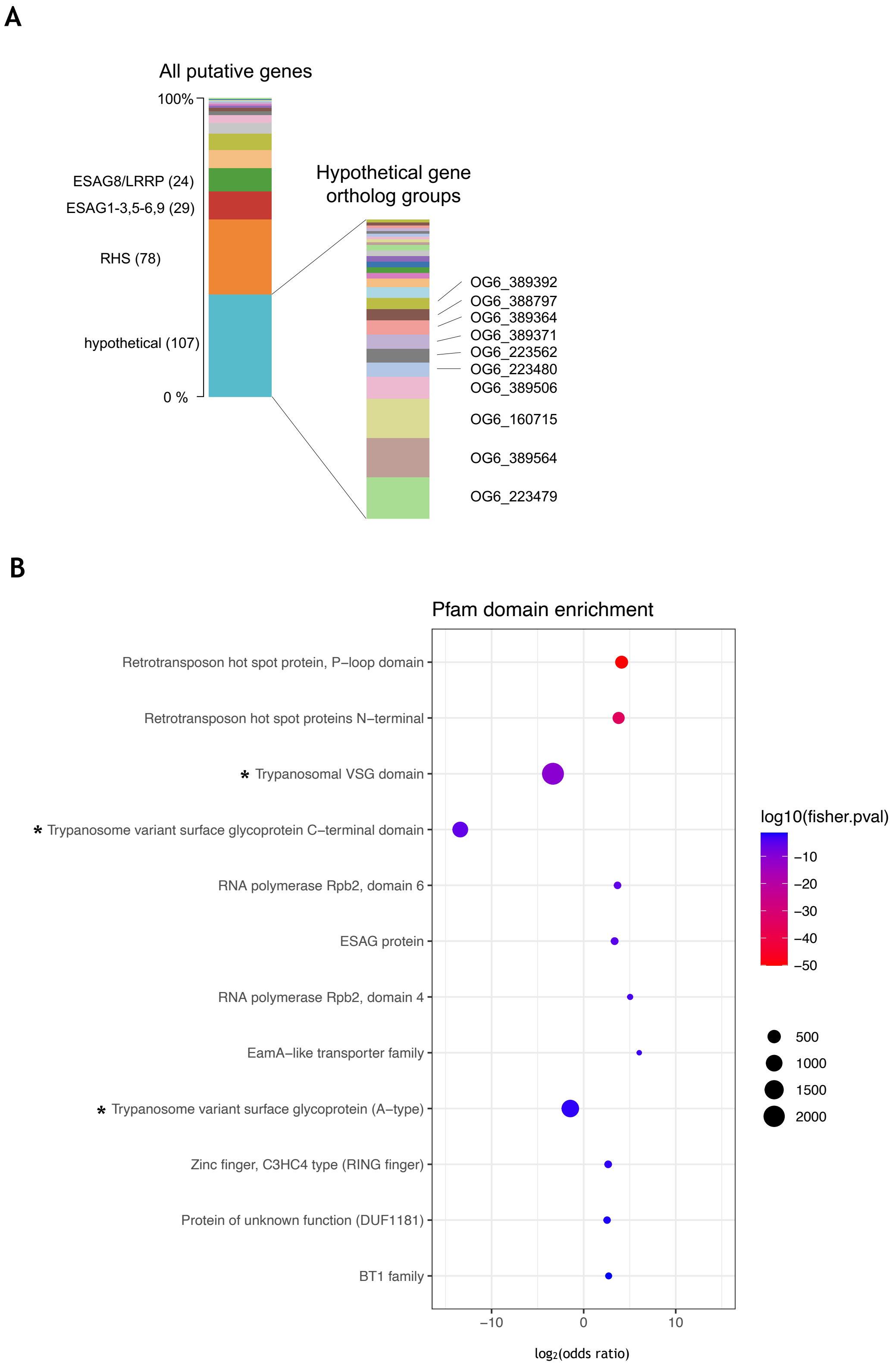


Figure 33 Gene content of sub-megabase chromosomes: orthogroups and pfam domains
A – Orthology group assignment of putative genes predicted by companion (Steinbiss *et al.*, 2016) and analysed using OrthoMCL (Li, Stoeckert and Roos, 2003; Alvarez-Jarreta *et al.*, 2024). B – Pfam protein domain enrichment in the sub-megabase contigs relative to other contigs. The size of the datapoints indicate the total number of sequences present in the assembly for a given protein domain.

To examine the transcriptional status of these smaller chromosome contigs, we mapped RNAseq data (Briggs *et al.*, 2018) to the new assembly, only retaining uniquely mapping reads (those that map only once in the genome) (Figure 32). To our surprise, gene expression was evident in these contigs, albeit at lower levels than in the core of megabase chromosomes, suggesting these contigs are likely distinct from the transcriptionally silent subtelomeric regions of megabase chromosomes, despite some similarity in gene content.

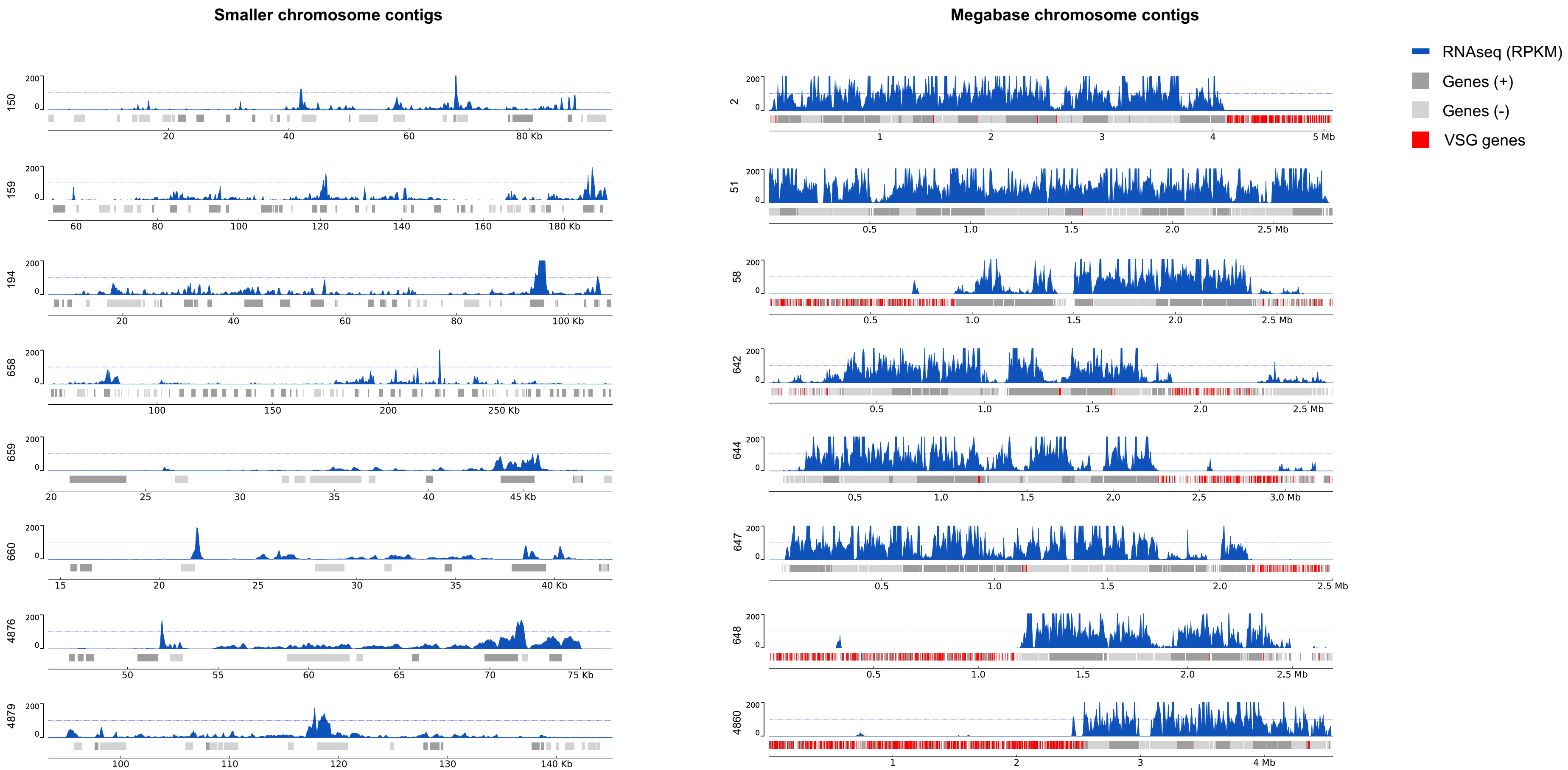


Figure 34 Sub-megabase chromosome display gene expression as measured by RNAseq.
Only uniquely mapping reads were retained and plotted, normalised using RPKM.

3.5.1.3 177bp repeat region structure

Of the 8 sub-megabase chromosome candidates, three contained what appear to be full length 177bp repeat regions – contigs 659, 660 and 4879; the regions were 13.5kb, 13.5 kb and 30 kb in length, respectively. Like the sub-megabase chromosomes, the 177bp repeat regions appear to have an inherent structure, composed of three groups of elements: the 177bp motif, interspersed with 18-20 bp elements, anchored by a short, tandemly repeated region composed of a 59 bp motif at the telomere-proximal end of the repeat region (Figure 35 A). Throughout the 177bp repetitive regions, sequence identity is exceptionally high – above 96% across the majority of the regions, and over 98% in the conserved core regions (Figure 35 B).

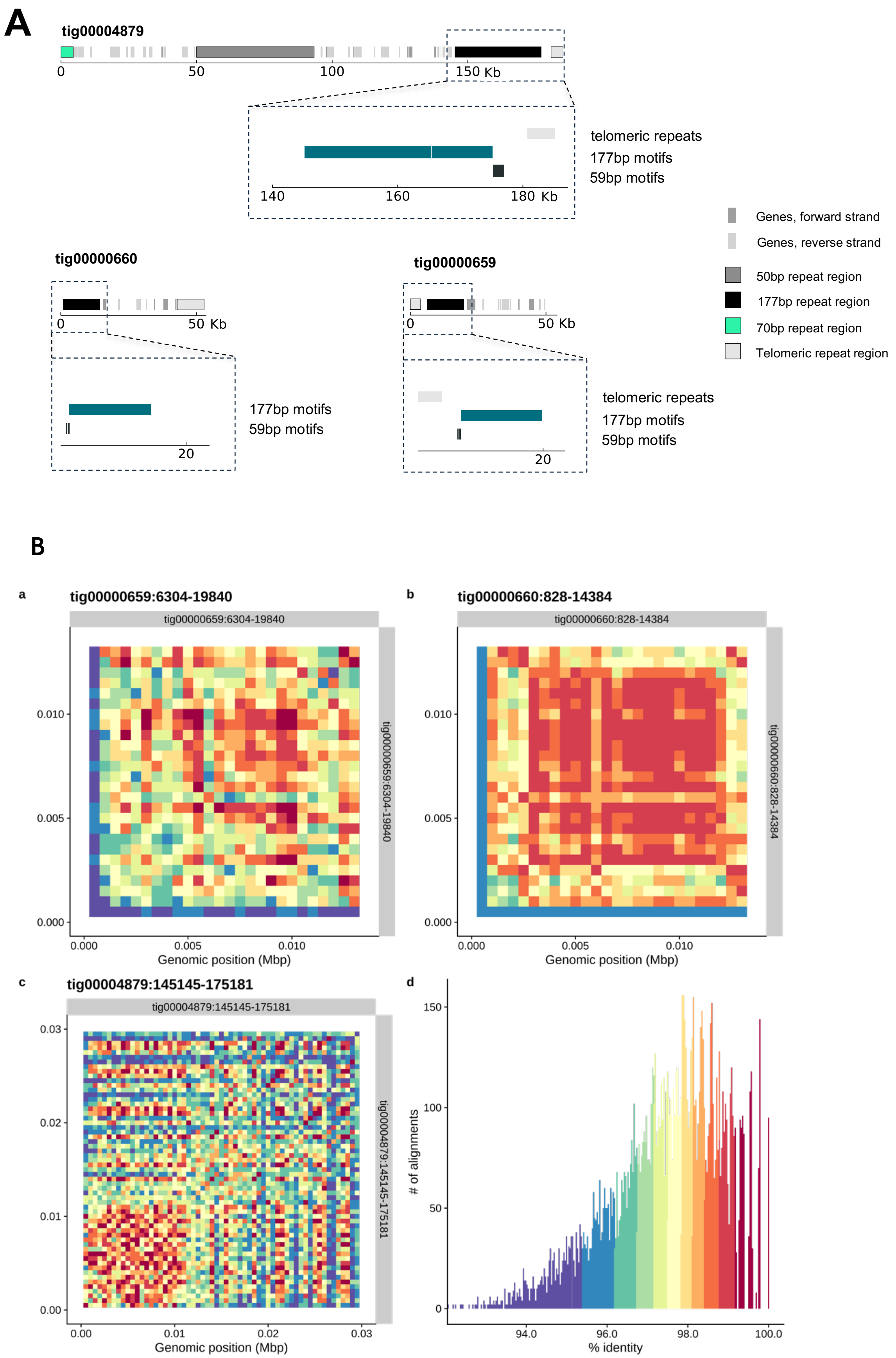


Figure 35 Full-length 177bp repeat region structure.
A – Full-length 177bp region repeat structure breakdown by the length of the repetitive element as reported by tandem repeats finder (Benson, 1999). Bottom row shows the repetitive elements overlapped. ‘Others’ - repetitive elements identified by TRF that do no fall into any of the other size categories. B – Intra-region pairwise identity heatmaps for the full-length 177bp repeat regions, produced by StainedGlass (Vollger *et al.*, 2022).

3.6 Centromere assembly and characterisation

In order to localise potential centromeric repeat regions, we initially created a blastn database containing flanking sequences surrounding gaps in the annotated centromeres from the reference genome and used blastn to localise matching sequences in the new assembly. The results were further manually refined based on sequence composition (GC fraction), synteny with the reference and tandem repeats finder results.

In total, 23 centromeric repeat candidates were retrieved (Table 9), 9 of which are likely full-length regions based on the presence of flanking regions. The length of the candidate regions varied considerably – for full length regions, between 30.2 and 105.2 kb; and for partial ones, 6.1 kb and 95.7 kb (Table 9, Figure 36). Both the size and the AT content of these repeats is not dissimilar from previously published estimates for another *T. brucei brucei* strain – TREU 927 (Table 9) (Obado *et al.*, 2007; Echeverry *et al.*, 2012), although it should be noted that the published AT content is based on analysing short fragments of centromeric repeats rather than their entirety, as full sequences were not available at the time (Obado *et al.*, 2007).

3.6.1.1 Centromeric repeat inter-region sequence identity

Focusing on the full-length centromeric repeat candidates exclusively, their inter- and intra-region similarity and therefore underlying structure was assessed through pairwise mapping and visualised as identity heatmaps (Figure 37, Figure 38). When assessing inter-region similarity, we can see that a given centromeric repeat region only has sequence identity above 80% with up to one other region; this is the case for 6 out of 8 sequences and the pairing is reciprocal – for example, contig 111 has regions above 80% identity with contig 126 and vice-versa (Figure 37). Interestingly, this high reciprocal sequence identity co-occurs with identified motif groups (discussed in 3.6.1.5 below). Only centromeres from contigs 642 (chromosome 3) and 648 (chromosome 6) do not display sequence identity above 80% with any of the other regions; in the case of the former this is unsurprising, given the chromosome 3 centromere has an exceptionally low AT content (50%) compared to the rest (61-75%), presumably due to unique sequence composition (Table 9) (Obado *et al.*, 2007).

Table 9 Summary of identified centromeric repeat candidates.
CHR – chromosome identity based on assembly vs assembly comparisons, published size is based on (Echeverry *et al.*, 2012), published AT content based on (Obado *et al.*, 2007).

NO	CONTIG	CHR	SIZE (KB)	PUBLISHED SIZE (KB)	FULL-LENGTH	AT CONTENT (%)	PUBLISHED AT CONTENT (%)	MOTIF GROUP
1	tig00000654	1	20.4	20 & 65	No	70	66	
2	tig00000069	1	21.3		No	69		
3	tig00000111	2	37.2	30 & 55	Yes	73	66	1
4	tig00000126	2	36.6		Yes	72		1
5	tig00000642a	3	30.2	75 & 80	Yes	50	49	2
6	tig00000058	4	70.6	70	Yes	62	61	3
7	tig00000055	5	66.5	50 & 75	Yes	62	62	3
8	tig00000648	6	103.7	55	Yes	70	71	4
9	tig00000127	7	17.0	100 & 120	No	73	75	
10	tig00004860	7	19.3		No	73		
11	tig00000032	8	39.5	100	No	62	59	
12	tig00000644	8	83.1		No	62		
13	tig00004861	8	44.4		No	62		
14	tig00000643	8	103.0		Yes	61		3
15	tig00000642b	8	24.5		No	61		
16	tig00000645	9	105.2	Unknown	Yes	71	60	4
17	tig00000168	9	29.8		No	68		
18	tig00000002	10	22.9	Unknown	No	75	61	
19	tig00000082	10	59.5		Yes	73		4
20	tig00000152	11	20.9	Unknown	No	71	61	
21	tig00000133	11	6.1		No	70		
22	tig00000115	2 or 7	14.6		No	74		
23	tig00000027	unitig_2133 or 8	95.7		No	62		

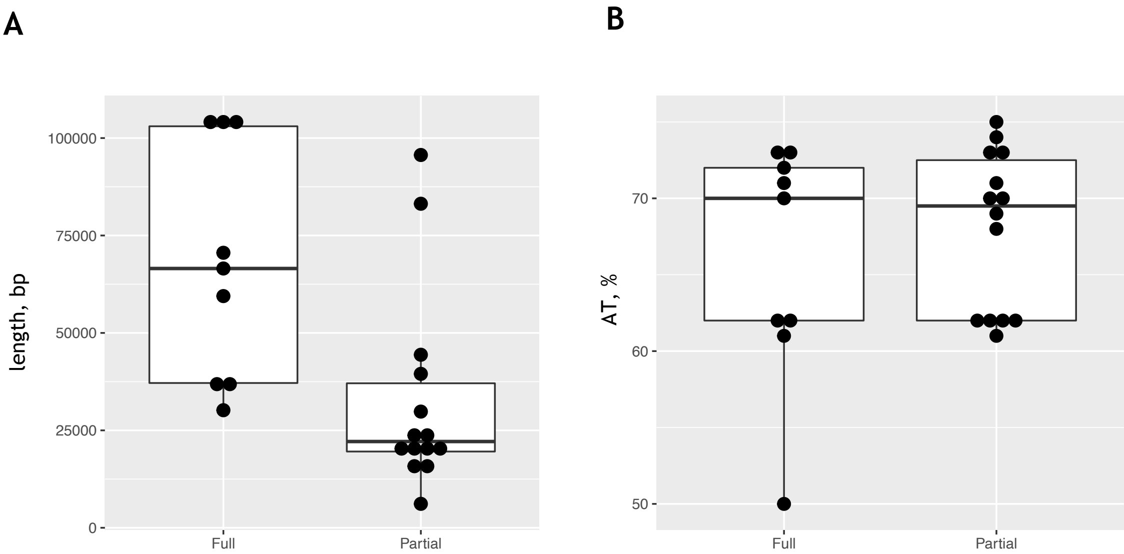


Figure 36 Full and partial length centromeric repeat candidate length (A) and sequence composition (B).

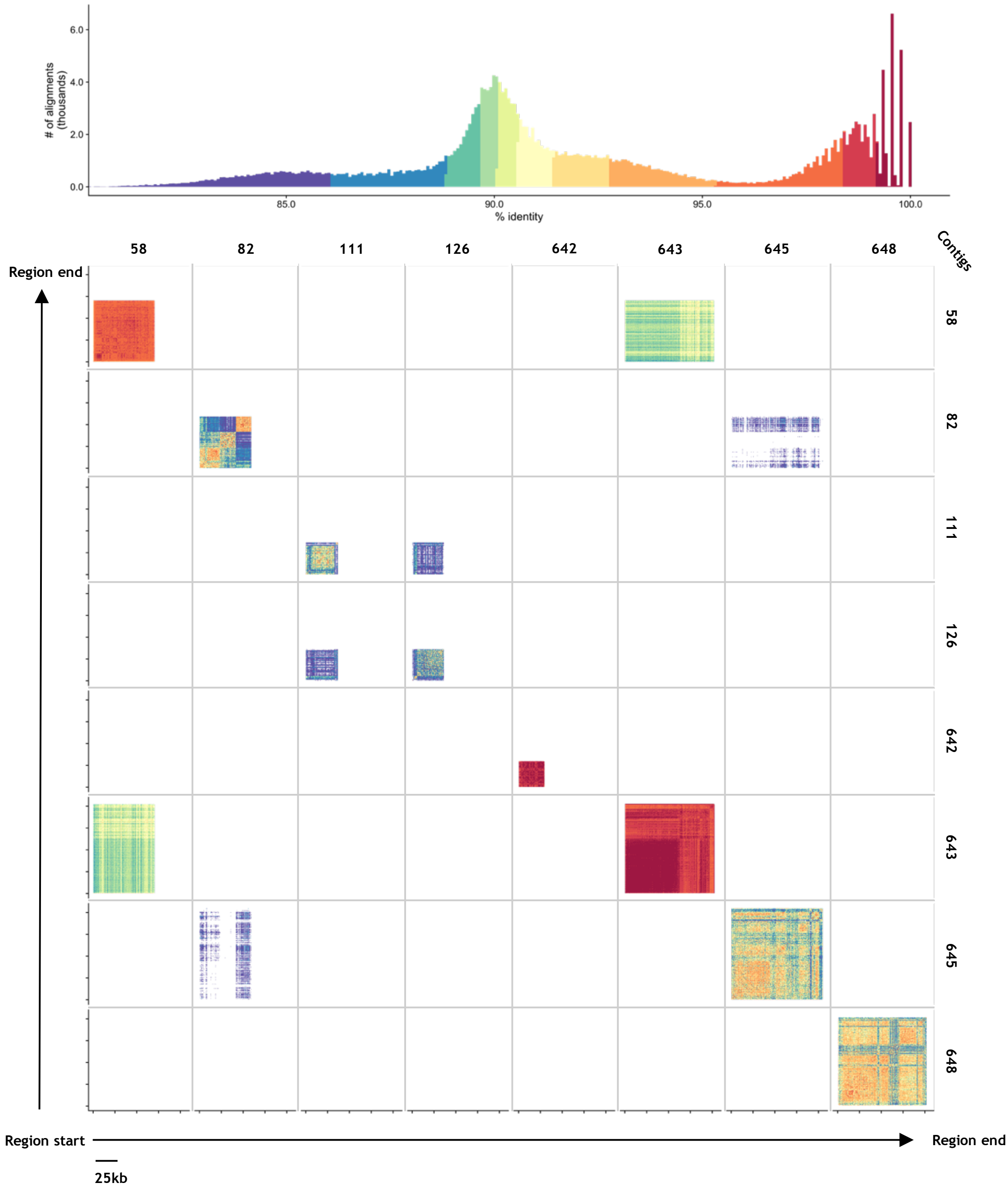


Figure 37 Inter-region pairwise sequence identity heatmaps for centromeric repeat candidates.
Pairwise sequence identity comparison between all eight full-length centromeric repeat candidates, as produced by StainedGlass (Vollger *et al.*, 2022).

Table 10 Summary of pairwise sequence comparison between full-length centromeric repeat candidates.

CONTIG	>85% IDENTITY WITH CONTIG	RECIPROCAL?	MOTIF GROUP
tig00000111	tig00000126	yes	1 & 1
tig00000058	tig00000643	yes	3 & 3
tig00000082	tig00000645	yes	4 & 4
tig00000642	None	N/A	2
tig00000648	None	N/A	4

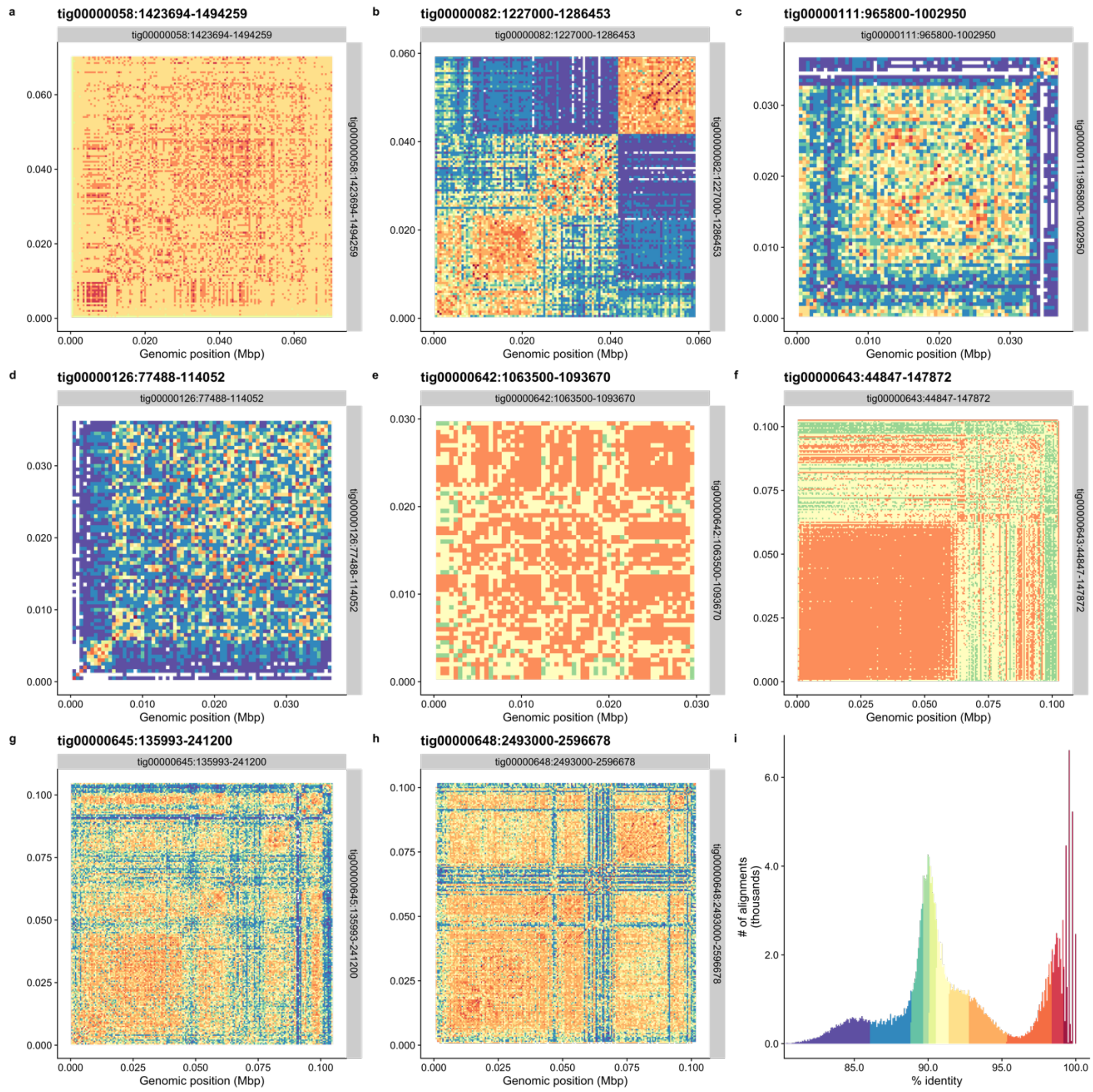


Figure 38 Intra-region sequence identity heatmaps for full-length centromeric repeat region candidates.

3.6.1.2 Centromere structure and intra-region sequence identity

Similarly, repetitive sequence structure and evolution can be assessed through pairwise sequence identity heatmaps (Logsdon *et al.*, 2021; Vollger *et al.*, 2022). Looking at full-length centromere candidate sequences only, we can see that the candidates vary considerably in terms of both underlying sequence identity level and structure (Figure 38). In contigs 111 and 126 (chromosome 2, motif group 1), there appears to be comparable, more conserved core repeat region, flanked by less conserved, 80-85% identity regions. While centromeres from contigs 58 and 643 (chromosomes 4 and 8, respectively, motif group 3) may share sequence identity, their intra-region structure is dissimilar, and the same can be said about centromeres of contigs 82 and 645 (chromosomes 10 and 9, respectively, motif group 4). Centromeric repeats of contigs 645 and 648 (chromosomes 9 and 6, respectively, motif group 4) are remarkably similar in terms of intra-region sequence identity and structure, despite sharing less than 80% sequence identity with each other.

3.6.1.3 Centromeric repeat candidate QC: do they act in early S phase DNA replication?

Centromeres in *T. brucei* are coincident with early-acting DNA replication origins (Tiengwe, Marcello, Farr, Dickens, *et al.*, 2012; Devlin *et al.*, 2016), a timing shared with fungi but not with mammals (Sreekumar *et al.*, 2021; Massey and Koren, 2022). As a quality control measure, we asked if the more complete centromere assembly made available by Nanopore sequencing matches the previous work in showing DNA replication dynamics consistent with early S replication, and if it revealed variation in replication timing, perhaps due to the considerable variation in centromere sequence composition or positioning of centromeric repeats in different genome compartments. To do this, we mapped MFAseq data (Devlin *et al.*, 2016) to the centromeric regions and the flanking 200 kb of upstream and downstream sequence (Figure 39).

Pronounced signal enrichment was seen in early S phase cells across the centromeric regions and a gradual decrease in signal further away from these regions, consistent with bidirectional progression of the replication fork. The signal from late S phase cells was still elevated relative to G2M, but was less pronounced, and G1 stage signal relative to G2M was mostly centred around 1, indicative of no replication. Of note, in the PCF cells the signal at the centromeric repeats did not seem to correspond to the surrounding signal – instead, it appeared decreased in early S phase cells and elevated in G1 cells. There might be a variety of reasons for this, including cell sorting or ‘gating’ differences during FACS, or drift in sequence in PCF relative to BSF and therefore read mapping issues. Despite this, the overall MFAseq pattern (amplitude and width) surrounding the centromeric repeat candidates was consistent between BSF and PCF cells. In addition, despite some apparent variation in signal intensity between the centromeric candidates, metaplots suggested broadly consistent MFAseq patterns for all centromeres, indicating comparable timing of DNA replication initiation.

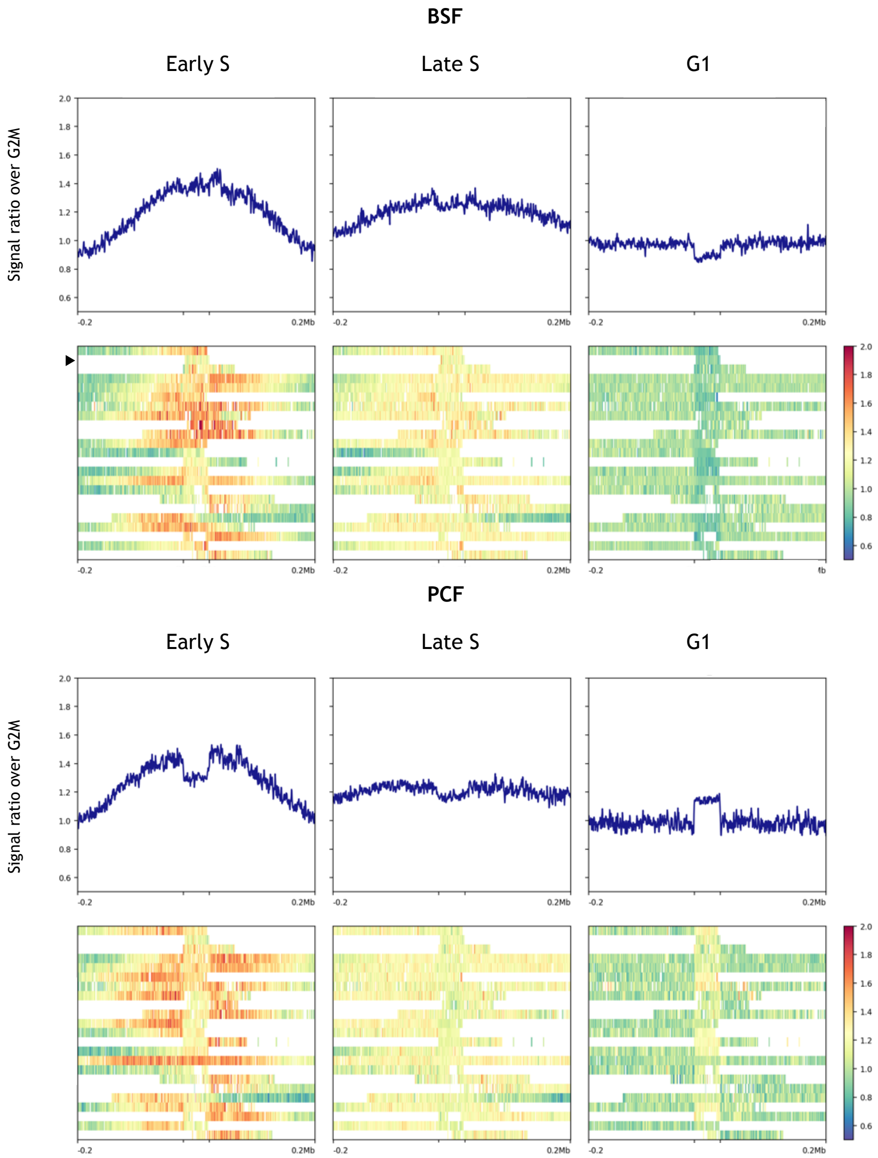


Figure 39 MFAseq signal across all identified centromeric repeat region candidates. Metaplot and heatmaps of MFAseq signal across centromere candidates and up to 200kb upstream and downstream. BSF – bloodstream-form cells, PCF – procyclic cells. Signal is a ratio of read depth relative to G2M. In the heatmaps where there is not enough sequence data upstream or downstream, the region is coloured white. Data from (Devlin *et al.*, 2016), mapped and processed by Dr Catarina Marques.

3.6.1.4 Centromeric repeat candidate QC: mapping of centromere-associated factors – KKT2, KKT3 and R-loops

Some factors have been previously associated with and mapped to centromeres, or rather fragments thereof, in *T. brucei*, including the kinetoplastid kinetochore proteins 2 and 3 (KKT2, KKT3) (Akiyoshi and Gull, 2014; Müller *et* al., 2018) and R-loops (Briggs, Hamilton, *et al.*, 2018). To determine whether this association was maintained or altered for the assembled centromeric candidates identified here, we remapped these data (Figure 40). For the kinetoplastid kinetochore proteins KKT2 and KKT3, mapping was not uniform across the centromeres: levels of enrichment varied across the different centromeres; signal enrichment was pronounced for some centromeric repeat candidates, but was less clear or undetectable for others; and in some cases the mapping flanked the repetitive region instead. R-loop mapping, on the other hand, was more uniform across the centromeres in both WT and *Tbrh1-/-* cells, though enrichment was again not equivalent for all centromeres. Consistent with previous reports, R-loop enrichment was greater in the RNaseH1-null cell line compared to the WT. The mapping of these factors in relation to sequence motifs is evaluated below in section 3.6.1.5.

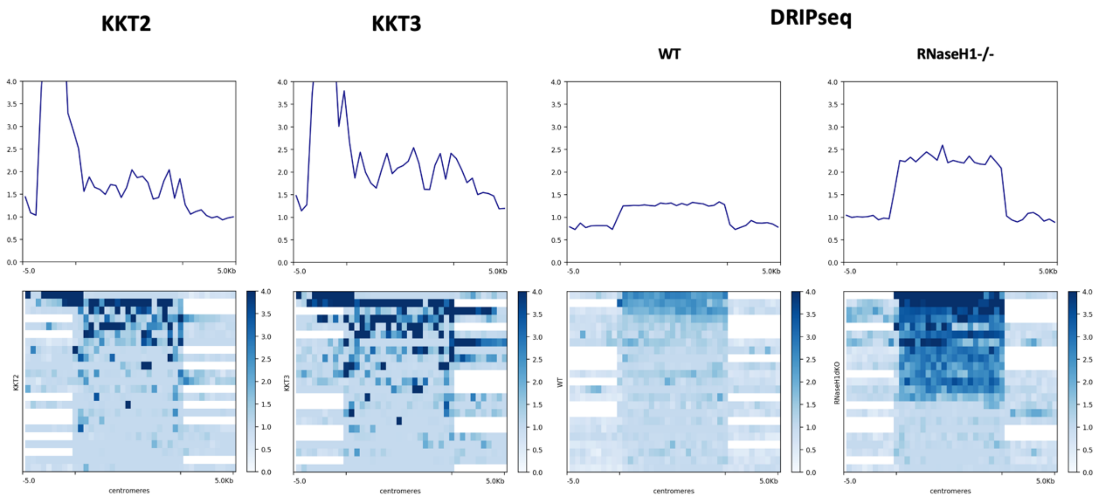


Figure 40 KKT2, KKT3 and R-loop mapping to centromere candidates.
ChIPseq data for KKT2 and KKT3 (Akiyoshi and Gull, 2014), as well as DRIPseq data (R-loops) (Briggs, Hamilton, *et al.*, 2018) mapped across all centromeric repeat region candidates and up to 5kb upstream and downstream. Signal – ratio relative to corresponding input samples. In the heatmaps where there is not enough sequence data upstream or downstream, the region is coloured white. R-loop data is presented for both wild-type cells (WT) and RNaseH1-null cells. Regions sorted based on signal intensity.

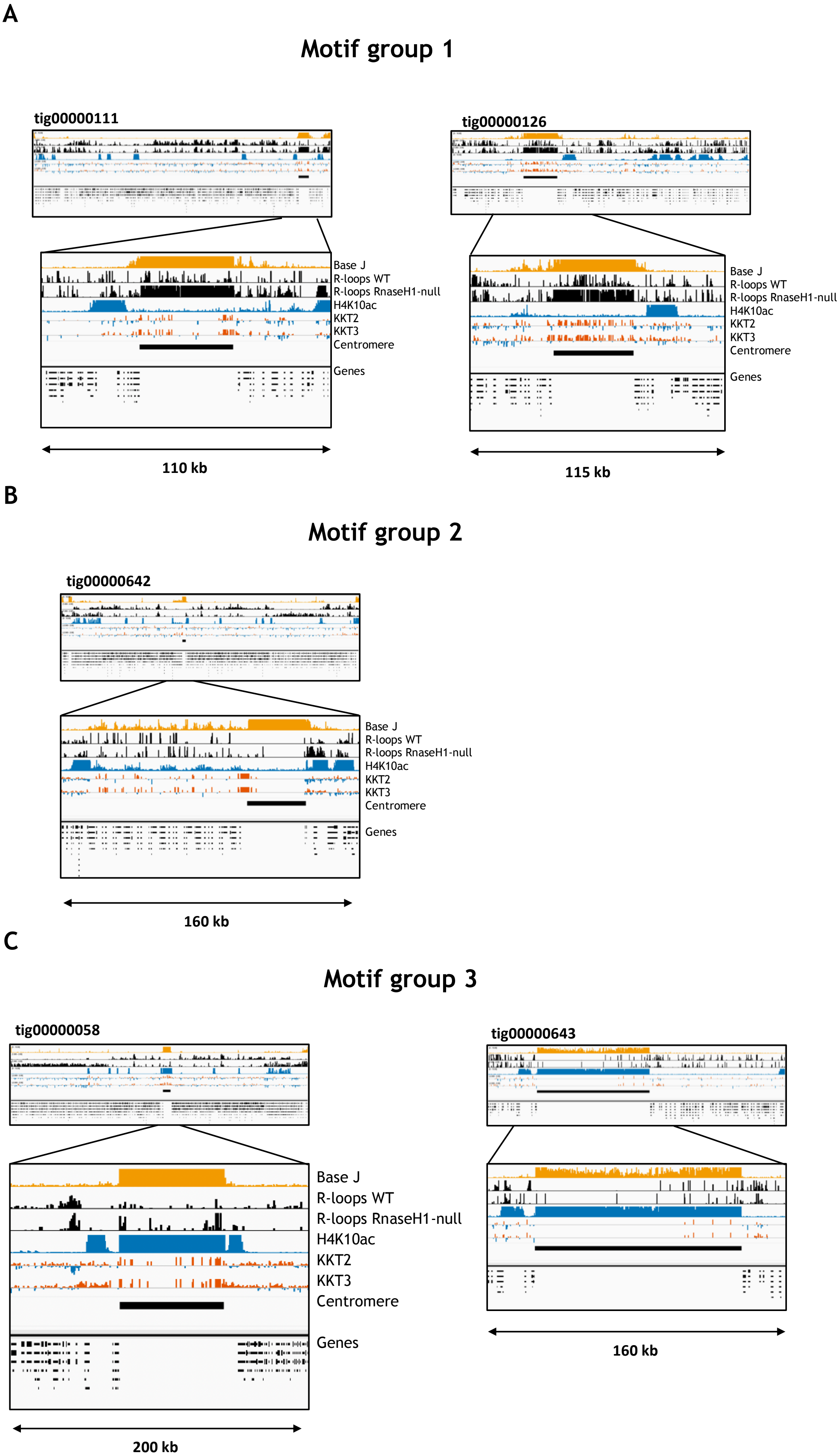
3.6.1.5 Centromeric repeat motif analysis

From published literature (Obado *et al.*, 2007) and the initial attempts at motif analysis here, it is clear that there is a lot of variety in motif sequence and length between different centromeres. Here, we attempted to loosely categorise the full-length centromeric repeat motifs based on the dominant repetitive element sequence and length, as well as note association with certain factors (Table 11, Figure 41, Table 22 in Appendices). While there appears to be motif-based overlap in terms of KKT2, KKT3, base J, R-loop and H4K10ac mapping, the small number of regions falling into each category (between 1 and 3) prevented us deriving any clear sense of organisation. From the above-mentioned factors, the most reliable association with centromeric repeat regions appears to be that of base J. Nonetheless, these data reinforce the considerable variability of *T. brucei* centromere repeat organisation between chromosomes. Consensus main motif sequences for the full-length centromeric repeat candidates, as determined using Tandem Repeats Finder (Benson, 1999), can be found in Table 22 in the Appendices.[1]

Table 11 Summary of centromeric motif group characteristics

NO	MOTIF LENGTH, BP	CONTIGS	CHR	KKT2	KKT3	BASE J	R-LOOPS WT	R-LOOPS TBRH1-/-	H4K10AC
1	26-30	111,126,	2, 7	Yes	Yes	Yes	Some	Yes	No
2	120 (98 + 19 + 3)	642a	3	Adjacent	Adjacent	Yes	No	Adjacent	No
3	147 (70 + 70 + 7)	27, 32, 55, 58, 642b, 643, 644, 4861	4,5,8	Yes/No	Yes/No	Yes	No	No/Some	Yes
4	complex	82,645,648	1,6,9,10, 11	Yes/Some	Yes/Some	Yes	Yes	Yes	No

[1] Due to the complexity of many of the centromeric repeat regions, we have not been able to generate a satisfactory motif sequence conservation logo for each repeat region (as is commonplace in the field), as our attempts at this resulted in misleading representations.


A
Motif group 1
tig00000111
tig00000126
Base J
R-loops WT
R-loops RnaseH1-null
H4K10ac
KKT2
KKT3
Centromere
Genes
110 kb
115 kb
B
Motif group 2
tig00000642
160 kb
C
Motif group 3
tig00000058
tig00000643
200 kb
160 kb

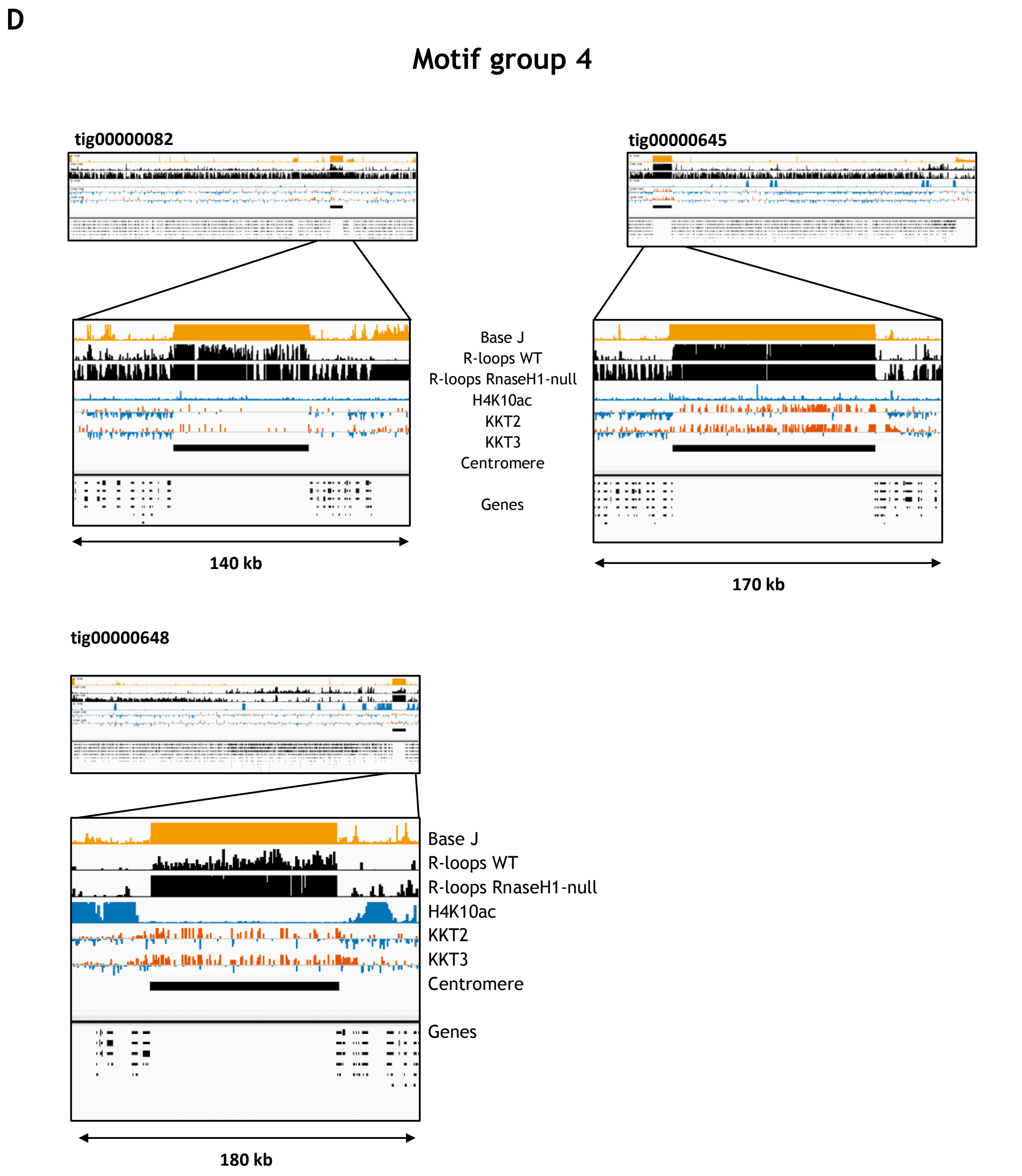


Figure 41 Detailed look at base J, R-loop , KKT2 ChIPseq, KKT3 ChIPseq and H4K10ac ChIPseq mapping across full-length centromeric repeat region candidates.

3.6.1.6 Overlap with 177bp motif

One unexpected outcome of the centromere repeat analysis was the realisation that, when analysing the genome-wide distribution of the 177bp and associated 59bp repeats characteristic of sub-megabase chromosomes, we found that these motifs are also present in the centromeric repeat candidates for chromosomes 1, 2, 6 and 9-11 (Table 12), suggesting a possible relationship between centromeres of megabase chromosomes and 177bp repeats in the smaller chromosomes. Interestingly, all of the centromeric repeat regions that have been found to contain the 177bp motif are either located at the boundary of core-subtelomeric compartments (chromosomes 2 and 6), or in the subtelomeric compartments (chromosomes 9, 10 and 11).

Table 12 177bp repeat region-associated motif presence in megabase chromosome centromeres.
Sequences identified from significant (p value < 10^{-9}) hits in FIMO (Grant, Bailey and Noble, 2011) output data searching the corresponding motifs – 177bp and 59bp motifs found in sub-megabase chromosome 177bp repeat regions.

Contig	Chromosome (subtelomere arm)	Number of 177bp repeats, p value < 10^{-9}	Number of 59bp repeats, p value < 10^{-9}
69	1	0	20
654	1	0	1
111	2	2	0
648	6	1307	244
645	9 (3A)	51	0
168	9 (3B)	58	200
82	10 (3A)	284	5
2	10 (3B)	1	0
152	11 (3A)	59	404
4875	11 (3A)	3	107
133	11 (3B)	5	28

3.7 DNA replication and genomic stability across genomic compartments

Previously in *T. brucei*, replication dynamics have been mapped genome-wide using marker frequency analysis coupled with whole genome next generation sequencing (MFAseq) – a method that captures DNA read depth across each chromosome in replicating cells relative to non-replicating cells (Tiengwe, Marcello, Farr, Dickens, *et al.*, 2012; Devlin *et al.*, 2016).

For the dataset analysed here, FACS-sorted *T. brucei* cells have been separated according to their DNA content, which should reflect the cell cycle stage – G1 (2C), S stage (between 2C and 4C) and G2M (4C); the 2C-4C cells were further split into ‘early S’ (ES) and ‘late S’ (LS) groups (Devlin *et al.*, 2016). The G1, ES and LS whole genome DNA sequencing is then plotted as a ratio against G2M values for normalisation; this was done both for bloodstream-form (BSF) and procyclic (PCF) *T. brucei* Lister 427 cells (Devlin *et al.*, 2016). We decided to focus on the MFAseq data in core and subtelomeric regions of megabase chromosomes, as well as sub-megabase chromosomes, as this analysis hasn’t previously been possible due to lack of comprehensive reference genomes at the time (Tiengwe, Marcello, Farr, Dickens, *et al.*, 2012; Devlin *et al.*, 2016).

3.7.1 Differential replication dynamics and genomic stability of megabase chromosome compartments

3.7.1.1 Mapping of DNA replication dynamics across megabase chromosome compartments

The above sections provide an evaluation of the Nanopore genome assembly, which was now used to extend our understanding of *T. brucei* DNA replication and genome stability. We first looked at DNA replication by plotting available *T. brucei* 427 MFAseq data onto select 1Mb+ megabase chromosome contigs that contain long stretches of subtelomeric regions (Figure 42). Overall, we saw the same consistency between BSF and PCF cells that has been previously reported (Devlin *et al.*, 2016), with similar peaks and troughs in both life cycle stages. However, a striking observation was the lack of apparent DNA replication initiation sites in the *VSG*-dense subtelomeric regions – apart from putative centromeric regions, which, in chromosomes 1-8 (where centromeres are located in the core compartments), are known to act early in DNA replication, there were no clearly discernible MFAseq peaks in this genomic compartment (Figure 42). In some cases, this resulted in over 1 or 2 megabase long stretches of chromosomes remaining under-replicated in early and late S phase, as detected using MFAseq.

It should be noted that MFAseq signal also appeared more diffuse in the subtelomeric regions relative to the cores, for reasons that are unclear. Because in the current most-complete reference genome the core and subtelomeric sequences are split into separate contigs (Müller *et al.*, 2018), allowing us to contrast data between the two compartments, we decided to additionally plot the core and subtelomere MFAseq overview using that reference genome for visual clarity (Figure 43). The same pattern was evident as shown above - while the core contigs (left hand side of Figure 43) contained clear MFAseq peaks throughout, the subtelomeres were devoid of detectable peaks, apart from at the centromeres in chromosomes 9, 10 and 11 (right hand side of Figure 43). *In toto*, in contrast with 47 MFAseq origins in the ~23 Mb of the core genome (Devlin *et al.*, 2016), only 6 are apparent in ~19 Mb the subtelomeres of *T. brucei* megabase chromosomes, further highlighting the compartmentalisation of this parasite's genome. Combined, these observations reveal two novel aspects about *T. brucei* genome replication dynamics. First, DNA replication dynamics in the core are distinct from those of the subtelomeres, and the subtelomeres are notably devoid of origins detectable by MFAseq. Second, the centromeric repeats, when situated in the subtelomeres of megabase chromosomes, similar to those in the core, act early in DNA replication.

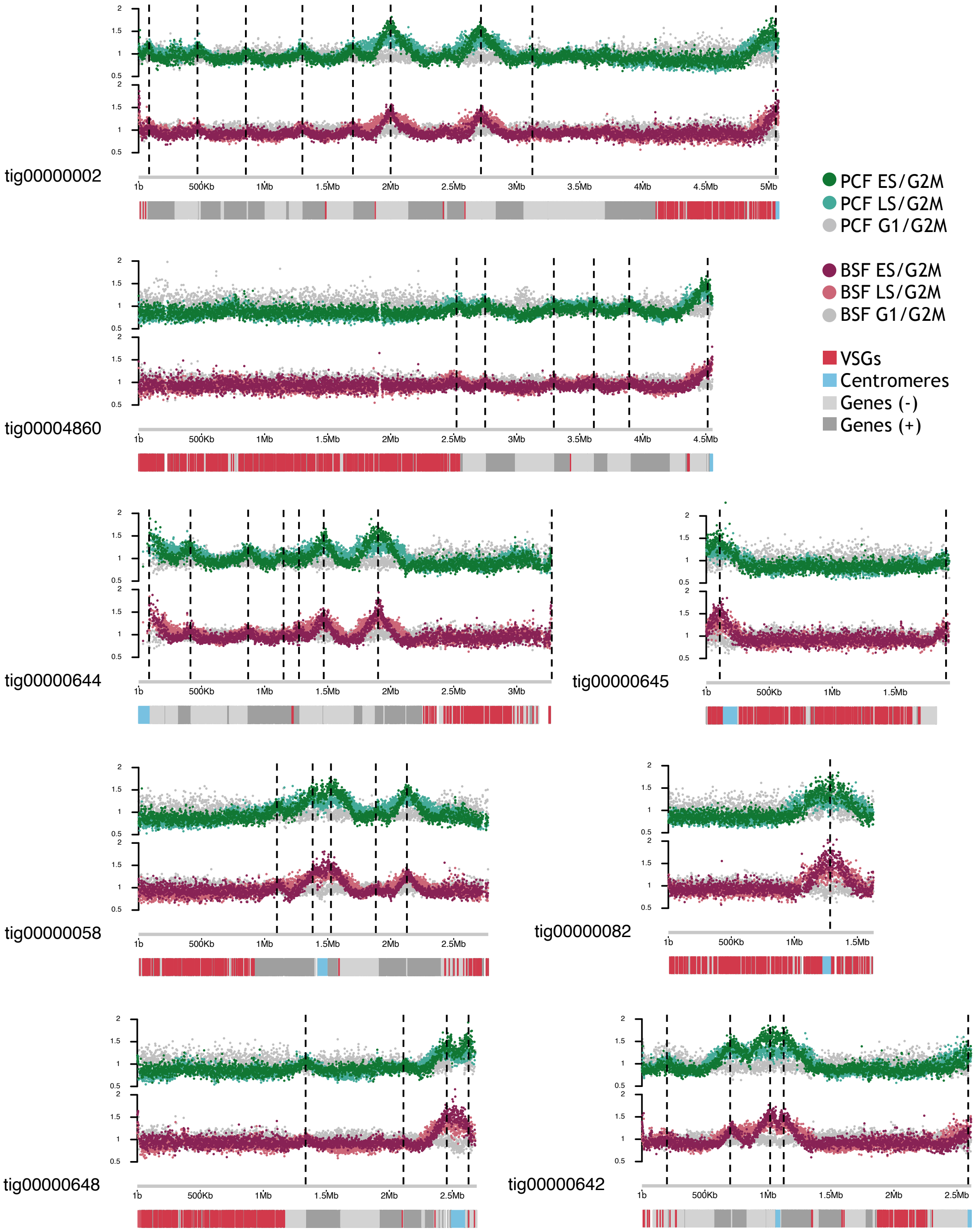


Figure 42 MFAseq mapping across megabase chromosomes.
Signal relative to G2M, BSF – bloodstream-form cells, PCF – procyclic cells. Vertical dashed lines added to highlight predicted MFAseq signal peaks.

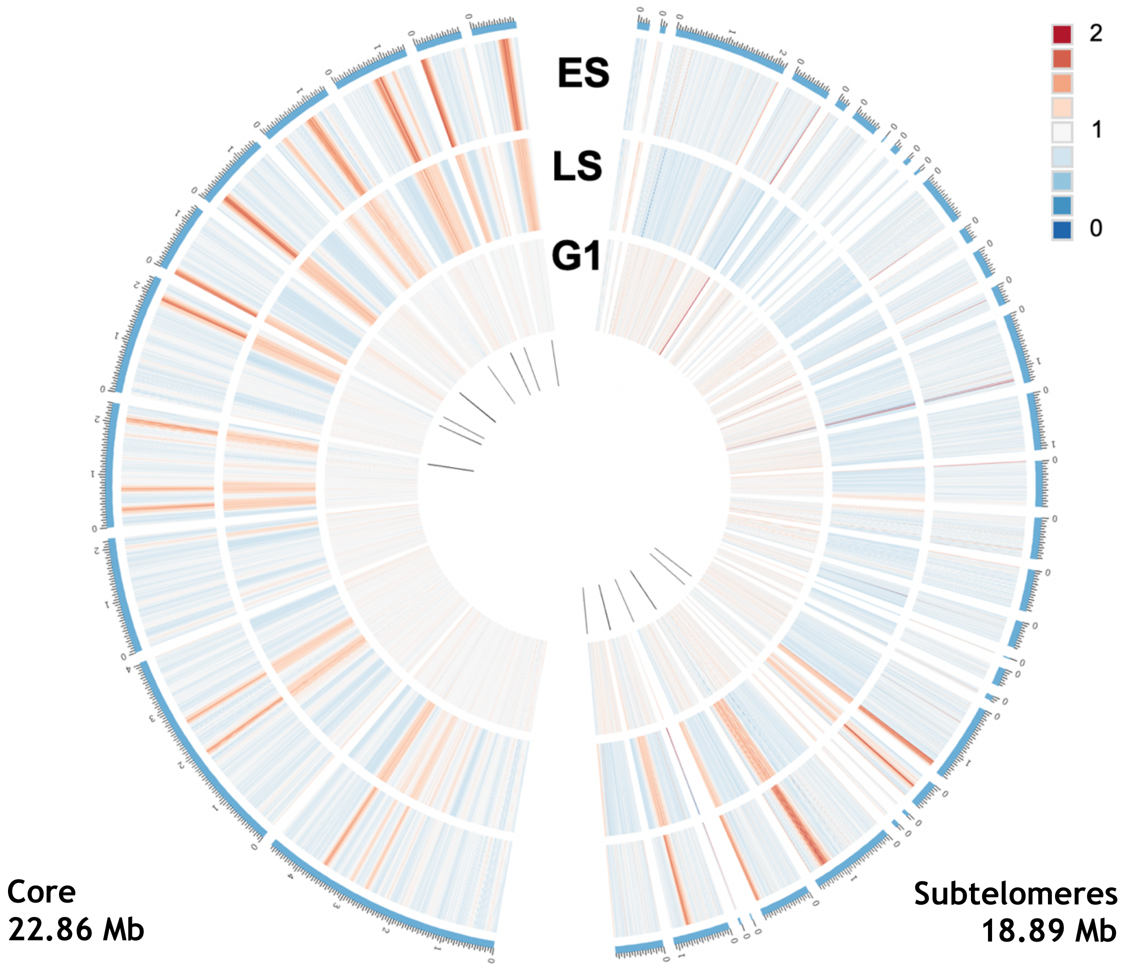


Figure 43 MFAseq in the core vs subtelomeric genome compartments.
Core genomic regions plotted separately on the lefthand side, subtelomeric regions – on the righthand side. Total sequence length for each compartment shown at the bottom. ES – early S stage, LS – late S stage, G1 – G1 cell cycle stage. Ratios 0-2 - read depth coverage in the respective cell cycle stage relative to that of G2M. Black lines in the centre – centromeric repeat regions.

3.7.1.2 Genomic stability of megabase chromosomes

As noted in the introduction to this chapter (3.1.4), in a range of organisms differential DNA replication timing can be reflected in the stability of said genomic regions. Additionally, previous work (Hartley and McCulloch, 2008) indicated that *T. brucei* cells lacking BRCA2, a key player in double-stranded DNA break repair (DSBR) in homologous recombination (HR), display gross chromosomal changes consistent with sequence loss during cell passage *in vitro*. To investigate the potential relationship between genomic compartmentalisation, replication timing, genomic stability and what role HR may play in these processes, we subcloned wild-type, RAD51-/- and BRCA2-/- cells of *T. brucei* Lister 427 bloodstream-form cells and grew two clones of each for 23 passages. Genomic DNA was collected at the start of the passaging (P0 or passage 0), at the end (P23), and further subcloning was also performed from P23 samples, collecting 3 subclones from each P23 sample. The DNA was sequenced using short-read technologies (Illumina and DNBSEQ) and mapped to both the 2018 reference genome (Müller *et al.*, 2018) and the assembly presented in this chapter.

Looking at the change in read depth coverage (RDC) of the sequenced samples over the time course, it is evident that in all three genotypes, albeit to varying degree, large-scale genomic instability is not equally spread across megabase chromosomes, as more regions of reduced RDC were seen in subtelomeres compared to the cores (Figure 44, Figure 45, Figure 46). The extent of the large-scale genomic changes was variable, being most pronounced in RAD51-/- cells, although BRCA2-/- samples, particularly when looking at P23 subclones (Figure 46), also displayed pronounced instability. These data extend previous analysis of *VSG* loss in BRCA2-/- cells and indicate that compartmentalisation between the core and subtelomeres is not merely seen in DNA replication activity but so in levels of instability.

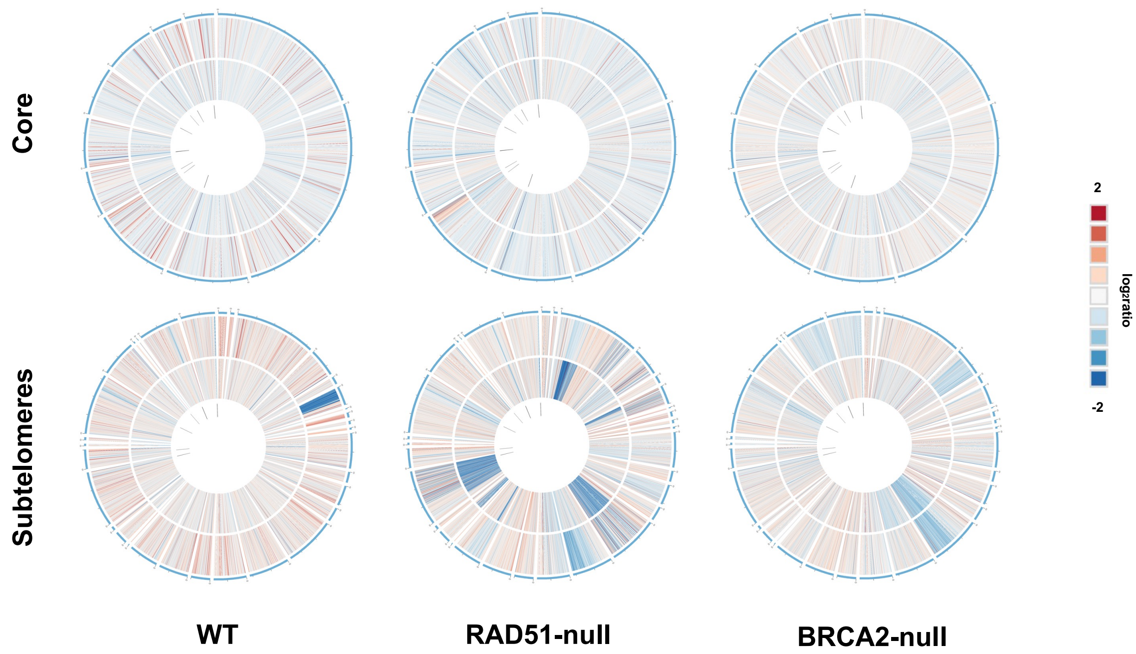


Figure 44 Differential genomic stability of megabase chromosome compartments. Change in read depth mapping across the core and subtelomeric compartments of two clones (separate circles) of wild type (WT), RAD51 null mutant and BRCA2 null mutant cells after 23 passages of growth in culture; data are plotted as log2 ratios and locations of centromeres (Müller *et al.*, 2018) are indicated as black lines .

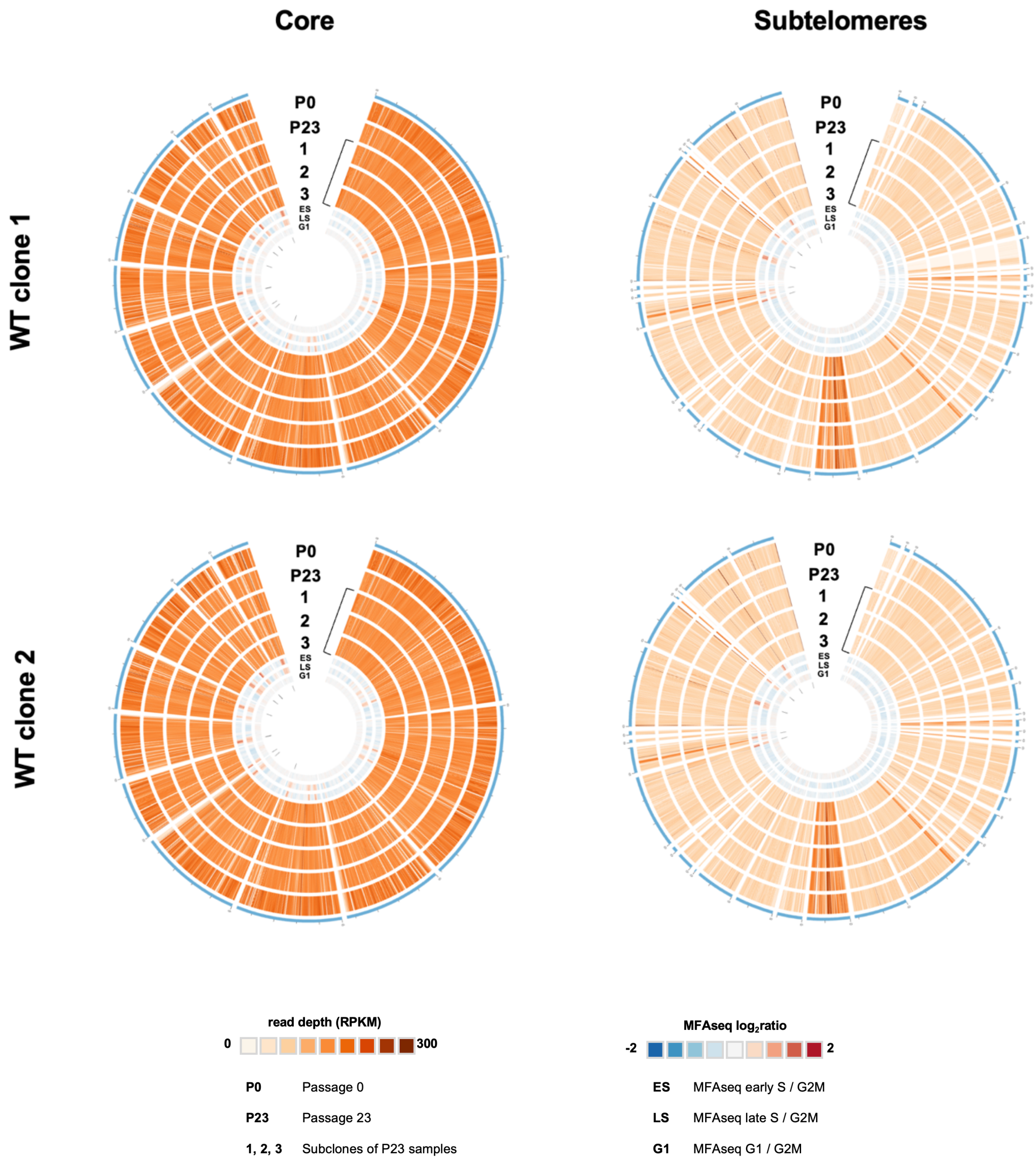


Figure 45 Stability between the cores and subtelomeres of the *T. brucei* megabase chromosomes (WT).
Read depth mapping is shown across the core and subtelomeric compartments in two clones of wild type (WT) cells as follows: as a population before growth (P0), as a population after 23 passages of growth in culture (P23), and in three clones (1, 2, 3) generated from the populations after 23 passages of growth; data is show as a heatmap of RPKM (reads per kb per million reads mapped). The innermost circles show MFAseq data (Devlin *et al.*, 2016), and the black lines indicate centromeric region locations (Müller *et al.*, 2018).

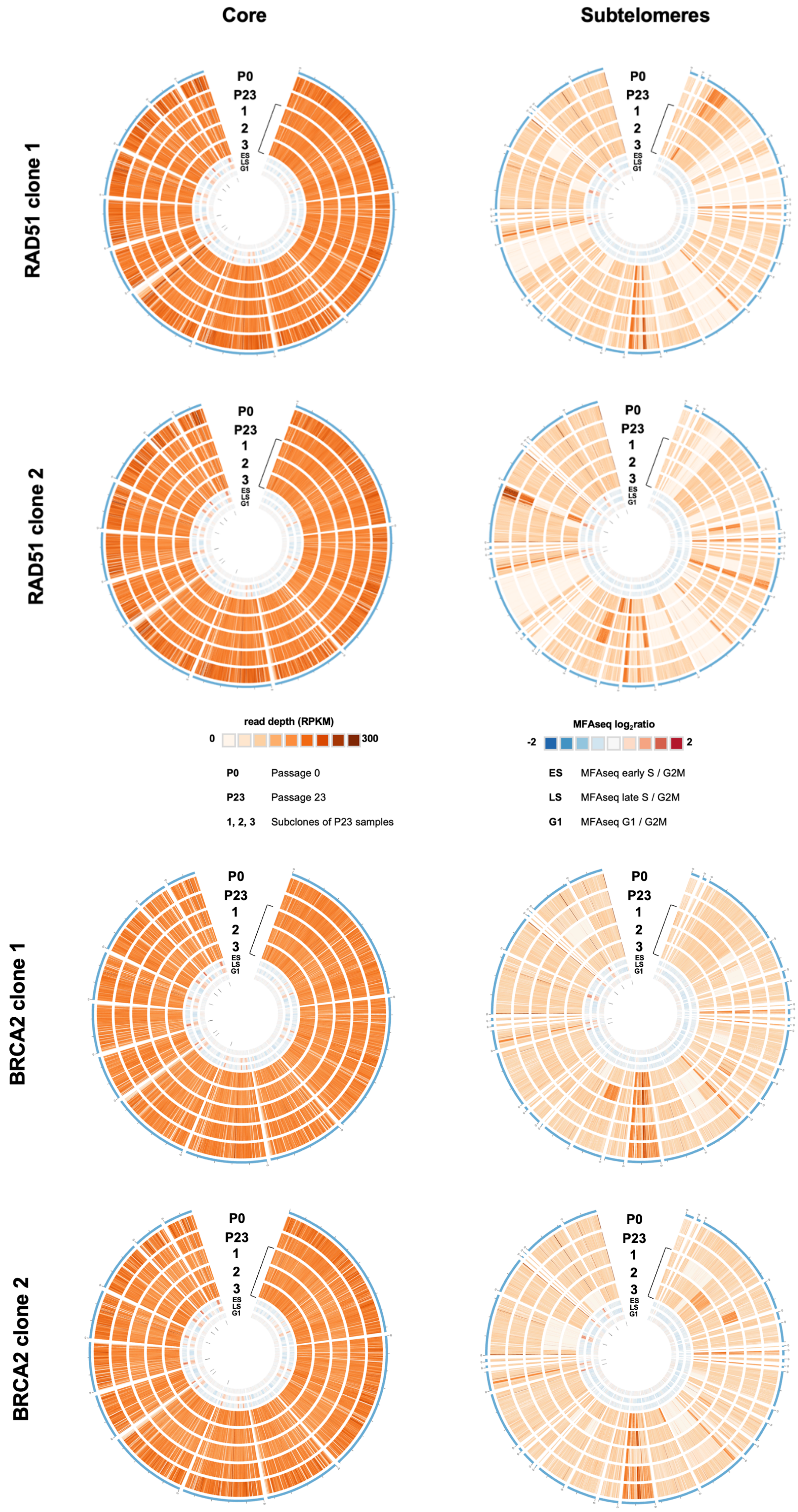


Figure 46 Stability between the cores and subtelomeres of the *T. brucei* megabase chromosomes (RAD51-/- and BRCA2-/-).
Read depth mapping is shown across the core and subtelomeric compartments in two clones of RAD51-/- or BRCA2-/- cells as follows: as a population before growth (P0), as a population after 23 passages of growth in culture (P23), and in three clones (1, 2, 3) generated from the populations

after 23 passages of growth; data is show as a heatmap of RPKM (reads per kb per million reads mapped). The innermost circles show MFAseq data (Devlin *et al.*, 2016), and the black lines indicate centromeric region locations (Müller *et al.*, 2018).

3.7.2 DNA replication dynamics and genomic stability of sub-megabase chromosomes

The second, previously unexplored, region of interest for replication mapping are the sub-megabase contigs. No origins of replication have been mapped to the smaller chromosomes of *T. brucei*; based on electron microscopy experiments, replication of these is thought to take place from a single bidirectional origin in minichromosomes (Weiden *et al.*, 1991). No centromeric sequences have been identified, though 177bp repeats have been speculated to act as centromeres in both mini- and intermediate chromosomes (Weiden *et al.*, 1991; Wickstead, Ersfeld and Gull, 2004).

Mapping MFAseq to these contigs, we can see that overall, as above, the enrichment of S/G2 reads was mostly consistent between BSF and PCF cells (Figure 32A). Interestingly, there appeared to be increased MFAseq signal across the 177bp repeats, which gradually decreased away from these regions. This signal is most clearly seen in metaplots of the available contigs (Figure 32B), where in the BSF cells there was a clear separation between G1 signal relative to G2 compared with ES/G2 and LS/G2 signal. In PCF cells, intriguingly, where ES/G2 and LS/G2 appeared comparable to BSF cells, elevated G1/G2 signal was seen, for reasons that are unclear (Figure 47 B). Similar to megabase chromosome cores, submegabase chromosomes displayed stability during the growth time course, even, remarkably, in HR mutants (Figure 48).

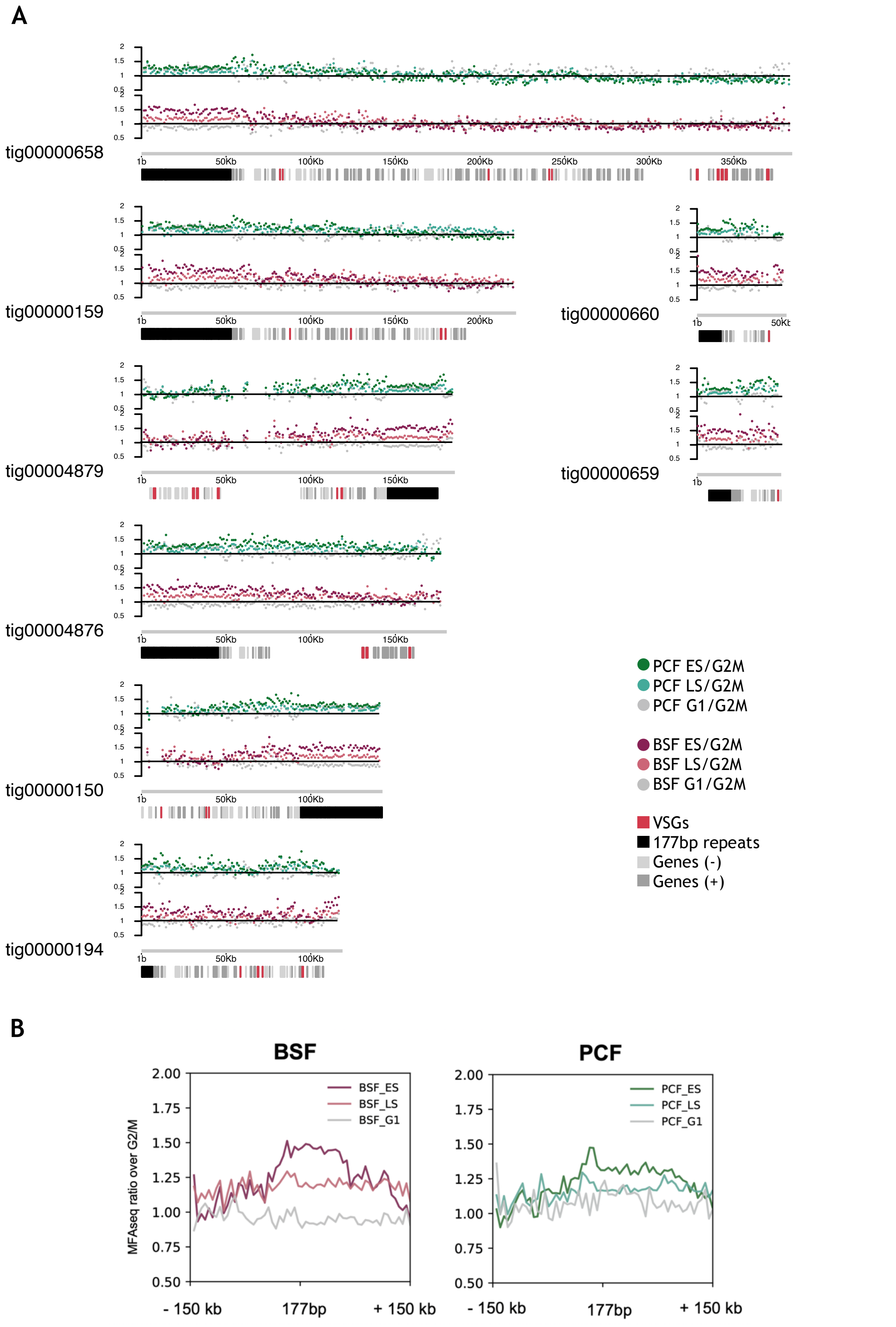


Figure 47 MFAseq mapping across sub-megabase chromosomes.
A - MFAseq signal plotted across identified sub-megabase contigs. B – Metaplot of MFAseq signal across 177bp repeat regions, as well as 150kb upstream and downstream. ES – early S stage, LS – late S stage, G1 – G1 stage. Signal relative to G2M. BSF – bloodstream-form cells, PCF – procyclic cells.

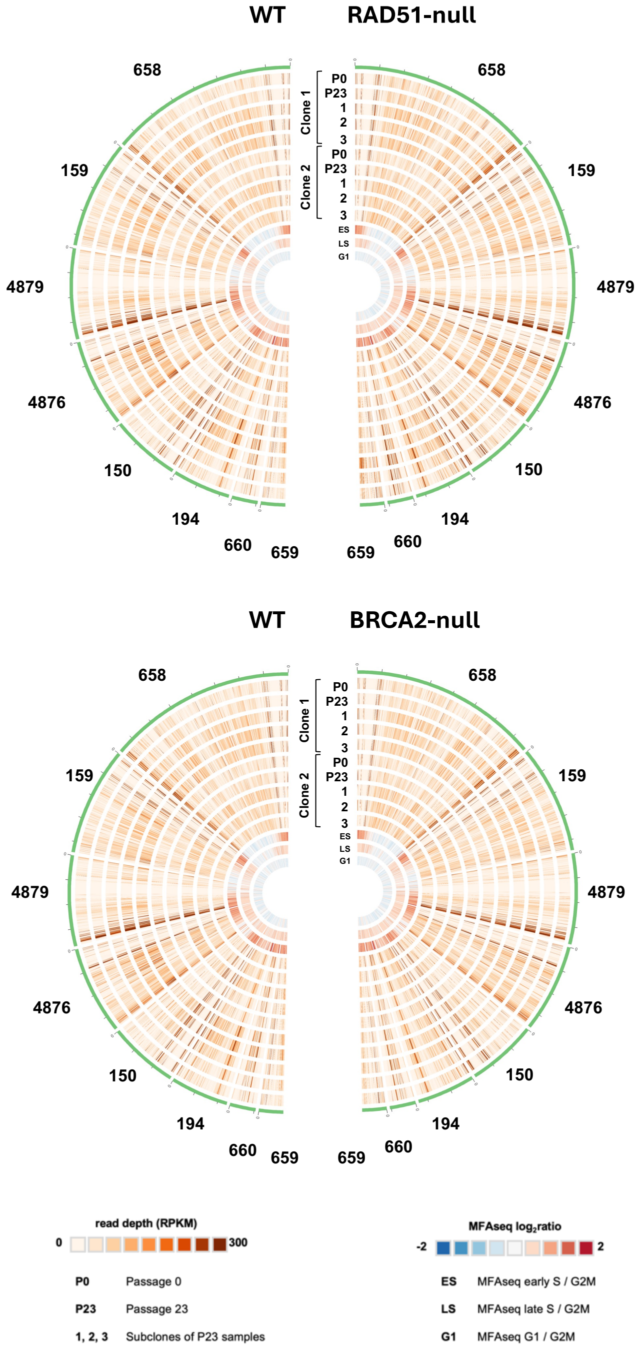


Figure 48 Sub-megabase chromosome contigs display remarkable stability during *in vitro* passage of WT, RAD51-/- and BRCA2-/- cells.
Read depth mapping is shown across identified sub-megabase contigs in two clones of WT, RAD51-/- and BRCA2-/- cells as follows: as a population before growth (P0), as a population after 23 passages of growth in culture (P23), and in three clones (1, 2, 3) generated from the populations after 23 passages of growth; data is show as a heatmap of RPKM (reads per kb per million reads mapped). The innermost circles show MFAseq data (Devlin *et al.*, 2016).

3.7.3 Telomere-proximal replication mapping

DNA replication initiation at telomeric/subtelomeric regions has been described in *Leishmania major* (Damasceno *et al.*, 2020), but not in *T. brucei*, though ORC has been reported to interact with the telomere in the latter and its loss shown to lead to altered VSG expression (Benmerzouga *et al.*, 2013). Whether this is related to the selective early replication of the active BES in BSF cells is unknown. In order to analyse whether there may be DNA replication at *T. brucei* telomeres, we plotted MFAseq data at identified TTAGGG stretches at contig ends and upstream flanking sequence (Figure 49).

In BSF cells, there was a clear enrichment in MFAseq signal at the telomeric repeats, which was most pronounced in early S phase, less so in late S, and with some enrichment in G1. These MFAseq signals relative to G2 did not obviously extend far into the sequences upstream of the telomere. Importantly, such MFAseq signal at the telomere repeats was not seen in PCF cells, ruling out a cross-mapping problem and indicating lifecycle stage-specific DNA replication.

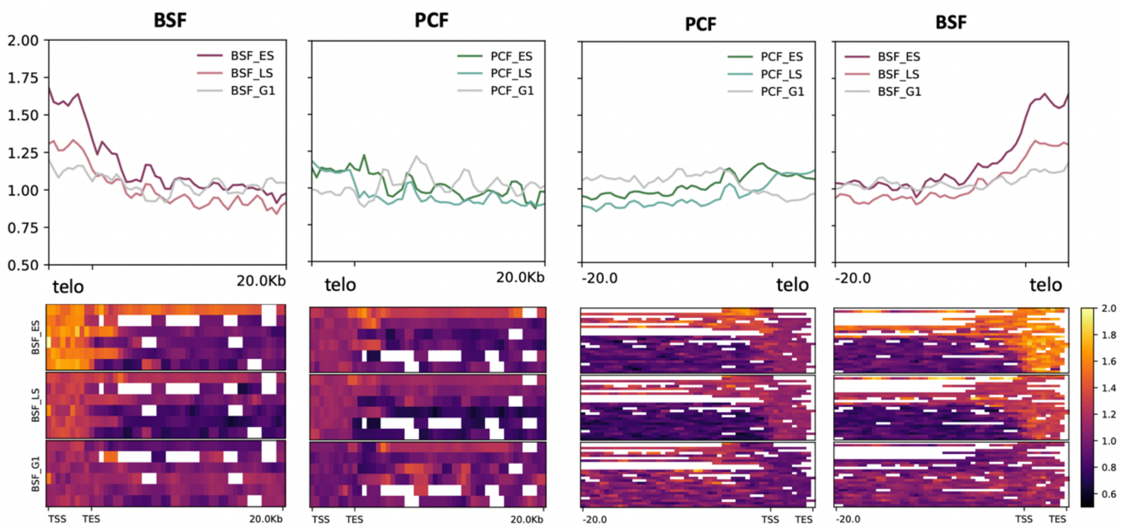


Figure 49 MFAseq signal at telomeres.
Metaplots and heatmaps of MFAseq signal mapped across telomeric (TTAGGG) repeat regions and preceding 20kb sequence. Panels on the lefthand side – for sequences where the telomeric sequence is on the 5' end of the contig, panels on the right – for those where the telomeric sequences are located on the 3' end of the contig. Missing data is plotted in white. ES – early S stage, LS – late S stage, G1 – G1 stage. Signal is read depth relative to G2 cells. BSF – bloodstream-form cells, PCF – procyclic cells.

3.7.4 Bloodstream-form expression site replication

It has been previously shown that the active BES is replicated early in S phase, but only in bloodstream-form cells – not procyclic stage cells (Devlin *et al.*, 2016). At the time, due to limitations of the underlying reference sequence, it was not possible to discern the direction of replication in the active BES – whether it originated upstream, downstream (from the telomere), or internally (Figure 8). As the active BES – BES1 – in the present assembly included the full 50bp repeat region and sequence upstream of it, MFAseq mapping across the contig (tig652), as well as its less contiguous counterpart (tig653), was examined to assess the direction of DNA replication (Figure 50).

First, differences between BSF and PCF signal were assessed across both BES1 copies, and, consistent with the aforementioned published results (Devlin *et al.*, 2016), BSF cells displayed significantly elevated signal in both early and late S phase (p-values < 0.0001) compared to PCF samples (Figure 50 B). Next, MFAseq levels were assessed in both lifecycle stages at different genomic locations in relation to the BES (Figure 50 C) – region immediately upstream of the BES ('2' in Figure 50 A) and a region even further upstream ('1' in Figure 50 A); this analysis showed gradually elevating signal in early S BSF cells in the direction of the BES, while in PCF cells both early and late S cells showed the opposite – a decrease in signal in the BES relative to upstream regions. Comparing MFAseq signal in the BES regions to the downstream telomeric repeats, the telomeric regions displayed elevated signal in both early S samples (BSF and PCF), with the telomeric repeats in early S BSF cells showing the highest levels of all (median ~ 1.6) (Figure 50 D). Combined, these data show that early S phase replication of the active BES in BSF cells originates from the telomeric repeats downstream of the BES.

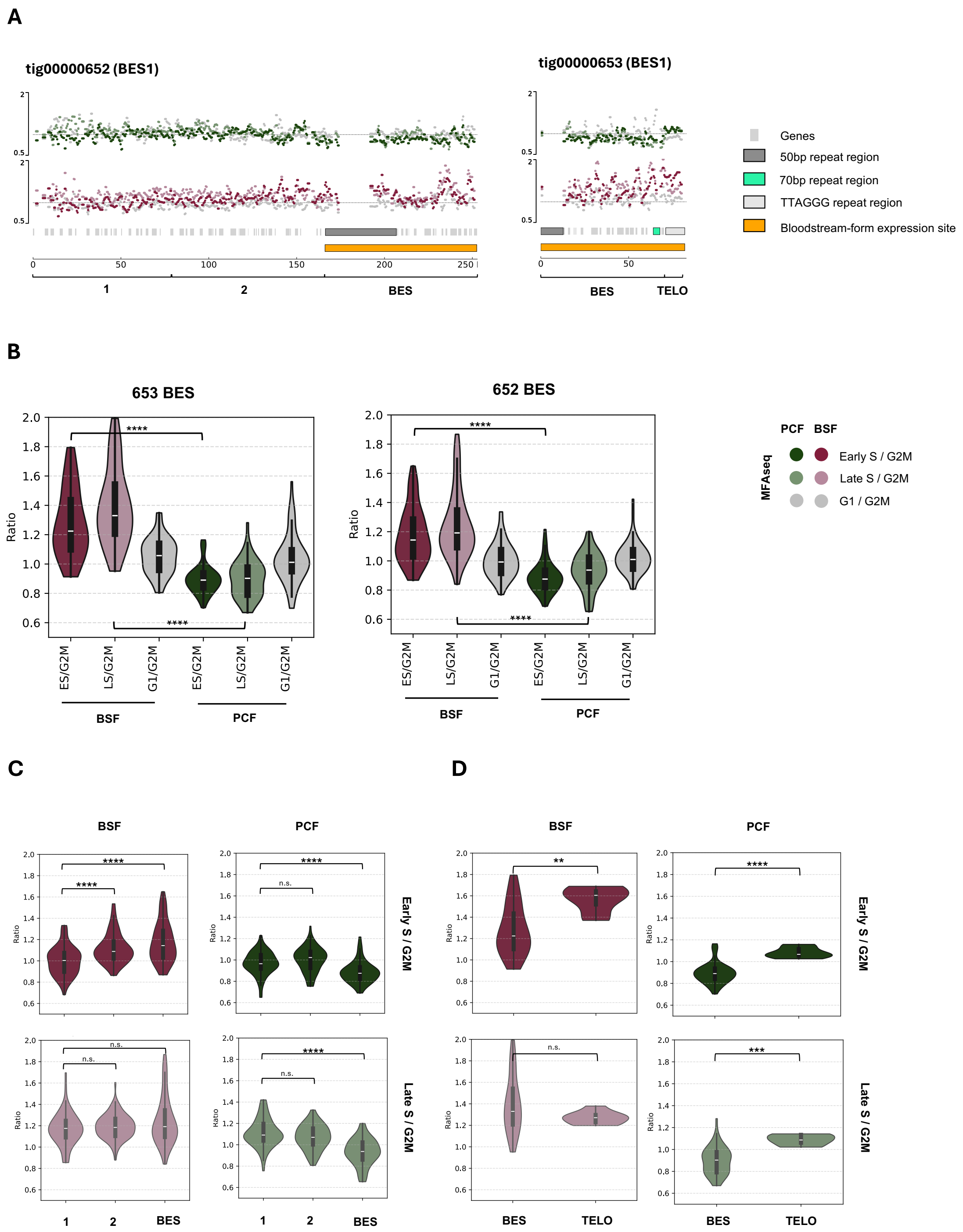


Figure 50 Early replication of the active BES stems from the telomeric repeats. MFAseq signal plotted in and upstream of the active BES1. A. MFAseq mapping shown for contigs 652 and 653, both of which contain BES1. Upstream regions in 652 divided into BES1-sized regions that are further analysed in C. B. MFAseq signal of BES regions specifically plotted as violinplots across the BSF and PCF samples. C. MFAseq signal of upstream regions and the BES region (as shown in A) plotted as violinplots for early and late S samples of both BSF and PCF samples (tig652). D. MFAseq signal across tig653 plotted as violinplots for BES regions and telomeric repeats for early and late S BSF and PCF samples. Statistical differences and pairwise comparisons were carried out using a Kruskal-Wallis test (p value < 0.05) followed by a Dunn's post-hoc test from scipy.stats and scikit.posthocs python packages (p-value adjustment for multiple comparisons using Bonferroni). Asterisks indicate p-values (**** - p-value < 0.0001, *** - p-value < 0.001, ** - p-value < 0.01, n.s. – not significant).

3.8 R-loops across genomic compartments

The improved contiguity and assembly of sub-megabase chromosomes in the new assembly allows for re-evaluation of existing datasets. In recent work, Jeziel Damasceno (Damasceno *et al.*, 2024b) showed that genome-wide R-loop accumulation in *Leishmania major*, a related trypanosomatid, is associated with replication timing as measured by MFAseq, and displays a chromosome size dependence. Furthermore, RNase H1 – a protein that resolves R-loops – displays a similar accumulation pattern across the genome, and loss of this factor perturbs the chromosome size-associated replication timing and leads to increased genomic instability, highlighting the role for R-loops and RNase H1 in DNA replication in this parasite (Damasceno *et al.*, 2024b).

In *T. brucei*, R-loops have been mapped in both WT and RNaseH1-null cells (Briggs, Hamilton, *et al.*, 2018), where the deletion of RNaseH1 resulted in increased R-loop accumulation at inter-CDS regions within PTUs, transcription start sites, as well as BES (Briggs, Crouch, *et al.*, 2018; Briggs, Hamilton, *et al.*, 2018). As was the case for *T. brucei* MFAseq data discussed in 3.7, analysis of subtelomeric sequences was limited at the time this research was carried out (prior to the 2018 genome assembly).

3.8.1 R-loop mapping across the core and subtelomeric genome regions

We mapped DRIP-seq (R-loop) data from WT and RNaseH1-null (*Tbrh1-/-) T. brucei* Lister 427 cells (Briggs, Hamilton, *et al.*, 2018) to the longer contigs of the new genome assembly that contain significant stretches of both core and subtelomeric genome regions – contigs 2 (chr10), 4860 (chr7), 644 (chr8), 58 (chr4), 648 (chr6) (Figure 51 A), as well as full-length regions of 50bp, 70bp and centromeric repeats (Figure 51 B).

When mapped to the new assembly, there was a clear distinction between core and subtelomeric regions in both WT and *Tbrh1-/-* (RNaseH1-null) cells (Figure 51 A). In the WT cells, there was consistent, albeit uneven, mapping across the core regions, with little or no signal in the subtelomeric regions (indicated by the presence of putative *VSG* genes in the bottom panel). In the *Tbrh1-/-* cell line, however, there was a similar signal pattern, though weaker, in the core genome, but a very pronounced and consistent signal across the subtelomeric regions, in contrast with the WT. Considering the role of RNase H1 in the removal of R-loops (Briggs, Crouch, *et al.*, 2018), it is likely that the low signal level in subtelomeres of the WT cells is due to R-loop removal by RNase H1.

R-loop mapping in repetitive regions, specifically, in the 70bp, 50bp and centromeric repeats, appears to be dependent on the nature of the repeat region (Figure 51 B). Across 50bp regions, there appeared to be no clear enrichment of R-loops in the WT or *Tbrh1-/-* cells, although there was a slightly higher signal at the boundaries of the regions in *Tbrh1-/-* cells. Across 70bp regions, there was pronounced R-loop signal enrichment in the *Tbrh1-/-* cells, whereas in the WT cells there appeared to be a dip in R-loop signal across those regions; this is consistent with previously published research (Briggs, Crouch, *et al.*, 2018). Across the centromeres, there was enrichment in R-loop signal in both WT and *Tbrh1-/-* cells, with more pronounced enrichment in the null cell line – also consistent with the previously published research (Briggs, Hamilton, *et al.*, 2018).

Curiously, R-loop signal at centromeric repeats varied depending on the genomic compartment the centromeric repeat is positioned in; the R-loop signal in both cell lines was highest in the subtelomeric centromeres and lowest in the core ones, with boundary centromeric repeats displaying intermediate levels, consistent with their positioning ‘between’ core and subtelomeric compartments (Figure 51 C). Despite the overall pattern of core regions displaying higher R-loop levels in the WT and lower R-loop levels in RNaseH1-null cells (Figure 51 A), centromeric repeats displayed higher R-loop levels in RNaseH1 mutant cells regardless of the genomic compartment (core, subtelomere or ‘boundary’).

The differential enrichment of R-loops in the core and subtelomeres of the megabase chromosomes further extends the evidence for pronounced compartmentalisation between these genomic compartments. In addition, R-loop mapping hints at differences between the highly sequence variable centromeres.

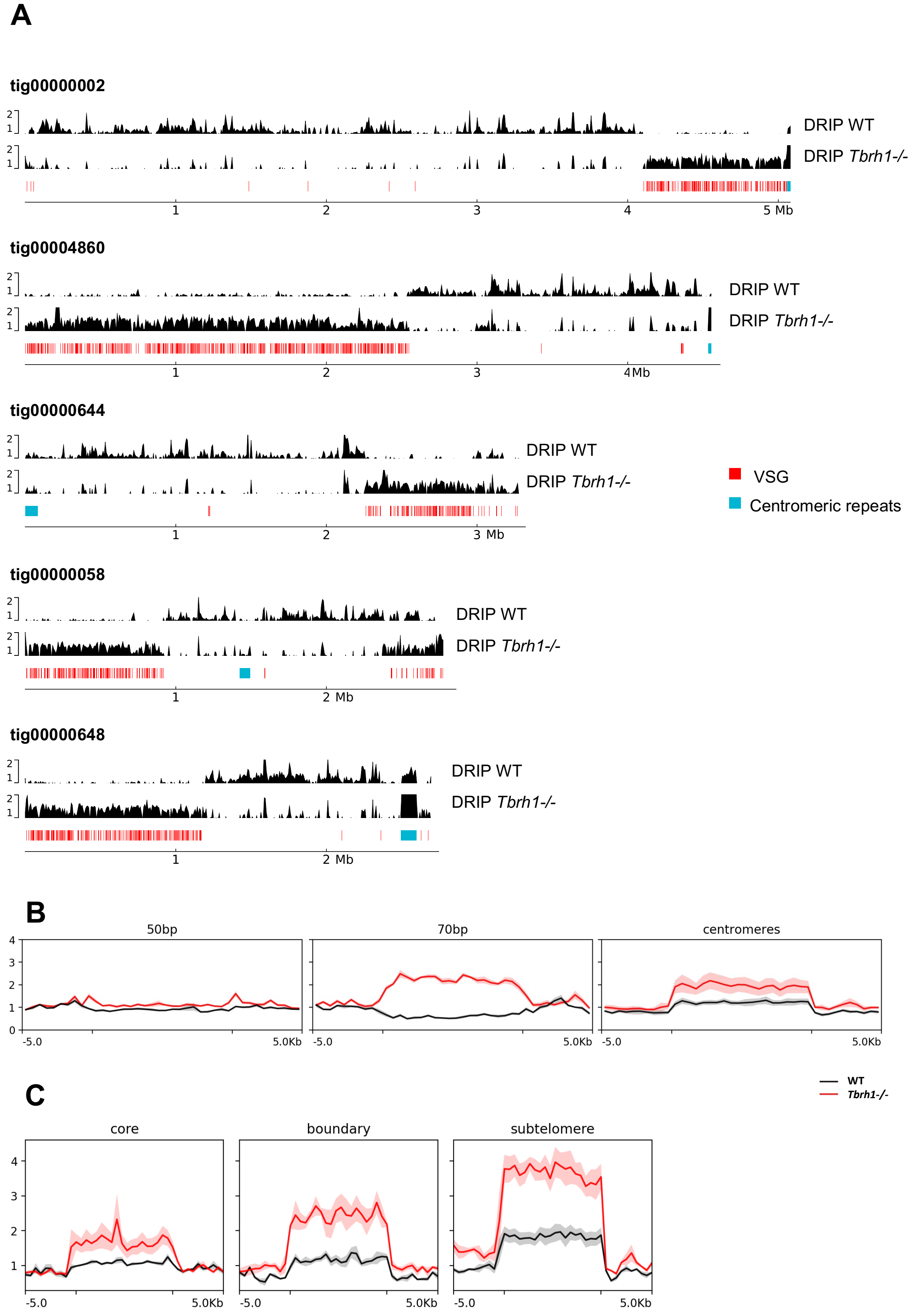


Figure 51 R-loop mapping across select megabase chromosomes, repetitive elements and centromeric regions.
A - R-loop mapping across some of the longest megabase contigs in the new assembly, in WT cells and in RNaseH1-null cells. Red triangle indicates the position of centromeric repeats. B – metaplot of full-length regions for 50bp, 70bp and centromeric repeat regions. C – metaplot of DRIPseq signal across full-length centromeric repeat regions depending on their position – core

genome (chromosomes 4, 5, 8), boundary of core-subtelomeric regions (chromosomes 2 and 6) and subtelomeric (chromosomes 9 and 10). Signal ratio relative to input.

3.8.2 R-loop mapping in sub-megabase chromosomes

In sub-megabase chromosomes, R-loop signal enrichment across the 177bp repeats signal was seen in *Tbrh1-/-* cells, but not in the WT (Figure 52). Similarly to the megabase chromosomes, there was a pronounced enrichment in R-loop signal across 70bp repeat regions in the null cell line. Throughout the coding region of these chromosomes there appeared to be similar, slightly higher signal in the RH1-null cell line compared to the WT, in contrast with the gene-coding core of the megabase chromosomes, where R-loop signal was, on average, higher in the WT than *Tbrh-/-* cells (Figure 52 A). It is possible that this reflects differences in levels of transcription.

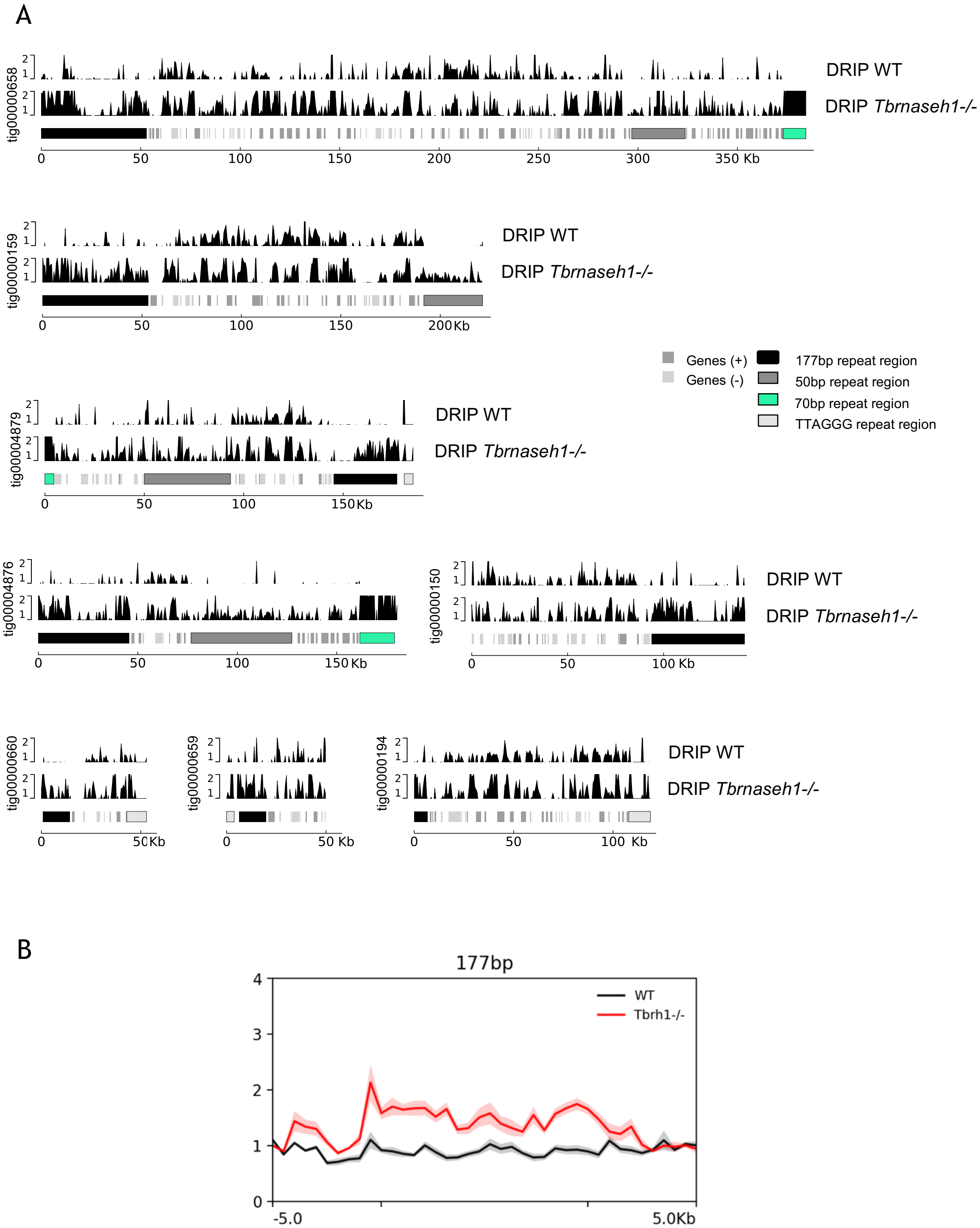


Figure 52 R-loop mapping across sub-megabase chromosomes.
A R-loop mapping across all eight identified sub-megabase chromosome contigs for WT and RNaseH1-null cells. 177bp, 50bp and 70bp regions annotated for clarity. Signal ratio relative to input. B Metaplot of DRIPseq R-loop signal across 177bp regions of sub-megabase chromosomes, signal relative to input, in WT and RNaseH1-null cells.

3.9 Discussion

In this chapter, we performed *de novo* genome assembly of *Trypanosoma brucei brucei* Lister 427 strain using Nanopore long-read sequencing to improve the understanding of DNA replication dynamics in the compartmentalised genome of the parasite. The broad aim of the project was to improve assembly contiguity to expand published analysis of replication dynamics using MFAseq across genomic compartment boundaries (core, subtelomere, BES), repetitive genomic regions and the sub-megabase chromosomes. In order to achieve this, we combined data from 13 Oxford Nanopore Technologies MinION sequencing runs of bloodstream-form *T. brucei brucei* (strain Lister 427), performed *de novo* long-read genome assembly, polished it using short Illumina sequencing reads, and then analysed the resulting assembly. While the assembly does not represent a telomere-to-telomere representation of the entire genome, it provides a number of novel insights into the organisation, sequence, DNA replication timing and compartmentalisation of the *Trypanosoma brucei* genome, as well as, more broadly, a basis for further data mining and exploration.

3.9.1 Overall genome assembly metric improvement

The new assembly shows improved contiguity in relation to the 2018 Lister 427 reference genome that was assembled using PacBio and Hi-C data (Müller *et al.*, 2018); this is reflected both in the general genome assembly metric comparison as well as localised bridging of previously separate sequences. When assessing genome assembly metrics, the new assembly is less fragmented (166 contigs instead of 317), despite showing higher levels of sequence duplication. More of the assembly sequence is part of longer contigs – 97/166 contigs (54.99/55.33Mb) in the new assembly are above 50kb in length, in contrast with 72/317 (43.95/50.08Mb) for the reference. In addition to improved N50 and L50 values the new assembly, unlike the reference, does not contain scaffold gaps.

The new assembly and the 2018 reference show similar completeness in terms of BUSCO scores, and comparable numbers of predicted *VSG* genes and pseudogenes – 3511 in the Nanopore assembly and 3524 in the published genome. It should be noted that the gene predictions are just that – *in silico* predictions - and the data has not been validated using, for example, transcript evidence.

Improved overall contiguation, N50, L50 and comparable BUSCO scores and very similar GC% all suggest that the current assembly is, overall, of comparable quality and in certain ways a potential improvement over the currently used reference. Hence, despite the *T. brucei* genome being extremely well understood through a combination of PacBio and Hi-C analysis, Nanopore sequencing has still been able to enhance understanding.

3.9.2 Bridging of genomic compartments

For 10/11 megabase chromosome pairs, between one and four previously separate contigs for each chromosome pair have been joined in the new assembly. The majority of these newly formed connections are between core and subtelomeric contigs; for previously isolated BES sequences, 10/15 have also been attached to identified (7/10) or unidentified (3/10) chromosomes.

We attempted to also investigate the number of scaffold gaps that have been closed in this assembly, although in some cases this proved to be complex and ambiguous. In several cases we identified additional sequence that would have likely previously been masked by the scaffold gap in the reference assembly, but it is followed by either an apparent translocation or a truncation, thereby adding complexity to the seemingly simple question of ‘is this gap closed in the new assembly?’. The reference assembly was assembled with the assistance of Hi-C chromatin conformation capture data (Müller *et al.*, 2018), which was not carried out for the genome assembly here; any scaffolding of sequences is not possible without further experimental work.

3.9.3 Assembly of unitigs into sub-megabase chromosomes

The unbiased nature of the *de novo* genome assembly approach allowed for recovery of eight likely mini or intermediate (here collectively called sub-megabase) chromosome sequences. This was done on the basis of the following two factors: presence of extensive tracts of 177bp repeats and lack of mapping to known megabase chromosome sequences. Contrary to long-held assertions

(Sloof *et al.*, 1983; Van der Ploeg *et al.*, 1984; Van der Ploeg *et al.*, 1984; Berriman *et al.*, 2005; Glover *et al.*, 2013; Müller *et al.*, 2018), these smaller chromosome candidates contain long coding regions with very few *VSG* genes – this is evident from orthology-based protein searches and protein domain analyses, where we see statistically significant depletion of three VSG-associated protein domains. Moreover, mapping of RNAseq data showed that these chromosomes contain expressed genes, further distinguishing them from mere silent gene archives as they have often been portrayed in the past. While their content, gene-wise, is reminiscent of that of megabase chromosome subtelomeres, their apparent expression (based on RNAseq data) highlights the distinct nature of this genomic compartment.

Whether the sub-megabase chromosomes recovered here are representative of the general structure of mini or intermediate chromosomes is unclear. Having all eight sequences in view, we propose these chromosomes are all similarly structured, with 177bp repeat regions positioned near a telomere at one end of the chromosome, followed by a gene coding region, and in at least some cases a BES at the opposing end of the chromosome. This organisation does not correspond to the description of mini or intermediate chromosomes of *T. brucei* that has been perpetuated since the initial experiments that designated the separation in size and content among chromosomes of this parasite (Sloof *et al.*, 1983; Van der Ploeg *et al.*, 1984; Chung *et al.*, 1990; Zomerdijk *et al.*, 1991). It does, however, more closely resemble the later description which included the identification of some genic sequences on the smaller chromosomes (Wickstead, Ersfeld and Gull, 2004). Considering the characteristic 177bp repeat unit is also present in several centromeres of megabase chromosomes, it is possible that the contigs that have been assigned as sub-megabase here are, in fact, previously unidentified sequences of megabase chromosomes. However, the distinct - highly structured and conserved – nature of the 177bp repeat region in the sub-megabase contigs as opposed to the instances of 177bp repeats found in known megabase chromosomes, coupled with the overwhelming mapping to previously uncharacterised reference sequences (unitigs) and the <1Mb size of the recovered candidates, hints at these being chromosomes distinct from the 11 megabase chromosome pairs. Furthermore, no wider centromeric sequences were found on these contigs.

3.9.4 Repetitive region assembly and analysis: overview

Third generation sequencing technologies, particularly Nanopore, have the advantage over earlier, short-read next generation sequencing of producing long sequencing reads that can span complex repetitive regions and therefore resolve problematic sequences (McGinty *et al.*, 2017; Jain *et al.*, 2018). This was particularly helpful for the assembly of the *T. brucei* genome, as it has distinct localised repetitive elements that are associated with specific functions. For example, 50bp and 70bp repetitive regions are associated with the bloodstream-form expression sites (Zomerdijk *et al.*, 1991; Hovel-Miner *et al.*, 2016) that are key to the parasite's ability to establish an infection in mammals, and the centromeric repeats are associated with early-acting DNA replication initiation sites (Tiengwe, Marcello, Farr, Dickens, *et al.*, 2012; Devlin *et al.*, 2016) as well as providing the location for assembly of the unconventional *T. brucei* kinetochore (Akiyoshi and Gull, 2014). In the new assembly, we managed to recover 7 full-length 70bp repeat regions, 10 full-length 50bp repeat regions, 9 complete centromeric repeat regions and 3 full 177bp repeat region sequences. The four repetitive sequence types are distinct in terms of structure, genomic localisation, sequence composition, complexity and conservation (Figure 53).

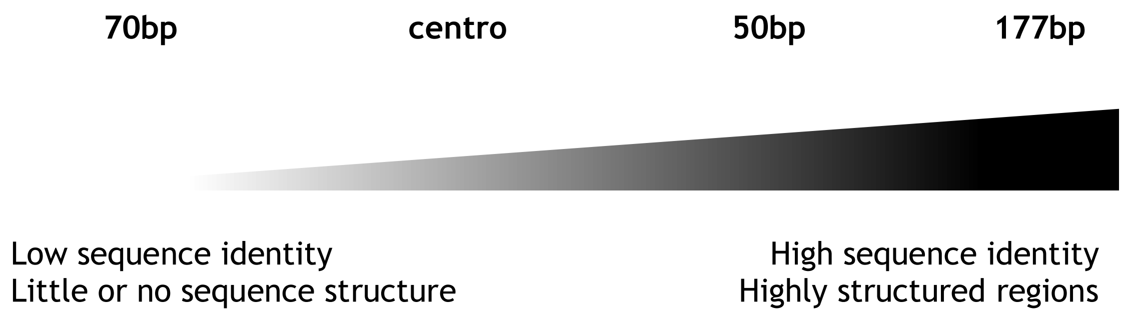


Figure 53 Summary of *T. brucei* repetitive region sequence identity and structure

3.9.4.1 70 bp regions are not as conserved as posited

BES-associated 70 bp repeat regions appear to be the most complex of the four categories; with GC content of 17-25% and variable size between 2 kb and over 20 kb, these regions don't appear to have any clear structure and show the lowest sequence conservation. While we found the previously published 77bp repeat unit (Hovel-Miner *et al.*, 2012) throughout the 70bp regions, the notion of this being a conserved sequence may be overstated, at least within the context of recognised repetitive element groups of this parasite (Figure 53); the 70bp repeat regions are not a mere tandem repetition of a motif, nor do they show easily detectible structure. It has been proposed that the 70bp repeat sequence is fragile, and that DSBs may be a contributing factor, or even trigger, in VSG switching in trypanosomes (Boothroyd *et al.*, 2009; Glover, Alsford and Horn, 2013; Li, 2015; Thivolle *et al.*, 2021). It is alluring to think that the lack of structure and conservation observed here is suggestive of repeated cycles of DNA breaks and imperfect repair in the past. Certainly, the fact that only half of the recovered 70bp regions appear full-length, despite being relatively short (compared to centromeric or 50bp repeat regions), may be considered as more evidence of sequence fragility complicating sequencing attempts. Recent work by Mark Girasol, using BLISS to map endogenous double stranded DNA breaks in *T. brucei* Lister 427 showed that the region where an overwhelming majority of DSBs appear to occur is around the 70bp repeats in the transcriptionally active BES (Girasol *et al.*, 2023). Whether the mapped breaks are inherent to the sequence or chromatin topology or, instead, enzymatically-induced is not known, and indeed it is unclear if the DSB mapped by BLISS initially has this break conformation or is a distinct lesion, such as a single-strand break or stalled replication fork, that is processed to a DSB during the analysis. Nonetheless, the presence of frequent breaks would necessitate repair either by homologous recombination or MMEJ, and both of these processes can be error-prone (Glover, McCulloch and Horn, 2008; Glover, Jun and Horn, 2011). We propose that, over time, the accumulation of errors would contribute to loss of sequence fidelity and produce variation, unless there is strong selection for the retention of certain motifs, such as for factor binding.

Further analysis of the stability of the 70bp repeat sequences may be carried out through, for example, longitudinal serial passaging of cells *in vitro* or throughout an infection, coupled with whole genome or targeted sequencing of these repeat regions.

3.9.4.2 50bp repeats show remarkable sequence fidelity

In contrast with 70bp repeats, the other BES-associated group of repeats – the 50bp repeat regions - are highly conserved. While the size of the region also varies widely, from around 5kb to at least 78 kb, the sequence composition of these repeats is more conserved: GC content varies little (41-45%), and the 50bp long motif unit is identical for each repeat region and tandemly repeated with only a few basepairs of intervening sequence. The function of these repeats remains unstudied and unknown; the high apparent sequence fidelity may be suggestive of strong selection for sequence preservation. Reverse chromatin precipitation using the 50bp motifs as a probe would be of interest, as it might yield an insight into the function of these repeats by highlighting binding factors. An additional way of investigating the function and essentiality of the 50bp repeat regions would be via targeted mutagenesis or deletion of these regions, for example, using CRISPR/Cas9, and subsequent assessment of the viability, infectivity, gene expression and *VSG* switching of the resulting cells.

The 50bp repeat region positioning at the interface between a BES and the rest of the chromosome may act as a boundary between functionally distinct genomic compartments or barrier for transcription machinery (Sheader *et al.*, 2003), as, unlike the rest of the genome, the BES genes are transcribed by RNA polymerase I (Palenchar and Bellofatto, 2006). Alternatively, the 50bp region may act as a tether for spatial separation of BES from the rest of the genome in the nucleus; it is known that the active BES in bloodstream-form cells localises into a dedicated extra-nucleolar compartment termed an expression site body (ESB) (Chaves *et al.*, 1998; Navarro and Gull, 2001; Budzak *et al.*, 2019, 2022). In addition, BES-BES interactions have been found to be relatively high in HiC chromatin conformation capture data (Müller *et al.*, 2018), suggesting spatial proximity.

3.9.5 Centromere-associated repeats vary among chromosomes

We found a total of 23 centromeric repeat candidates, and 9 appeared to be full-length based on the presence of flanking non-repetitive sequence. The full-length regions range from 30 to over 100kb in length, are low in GC% - mostly between 25 and 39%, with one exception (chromosome 3) where it is 50%. All of the above-mentioned are consistent with previously published data (Obado *et al.*, 2007; Echeverry *et al.*, 2012). Centromeric repeats appear to have varying levels of conservation and structure – we identified 4 major groups of repeats based on the dominant motif. Intra-region sequence identity and structure are variable, and there is limited inter-region similarity, as full-length centromeric regions share >80% sequence identity with no more than with one other identified region. This poses a question: with limited sequence identity, do the centromeric repeat regions have differential affinity for DNA replication and repair, chromosome segregation and transcription factor binding? This appears to be the case for KKT2, KKT3 and H4K10ac ChIPseq data, as well as R-loop DRIPseq data, as the mapping of these factors varies between the four motif groups, and in some cases among individual centromeric repeat regions within a given motif group. Mapping of short sequencing reads, such as those of ChIPseq and DRIPseq datasets, to repetitive regions can be problematic (Li, Ruan and Durbin, 2008), as there are often multiple points a given read can map to; nevertheless, we would expect to see consistency in mapping profiles between the centromeric regions, and that is not the case. The potential consequences, if any, of such variation are unknown, though mapping MFAseq data did not appear to reveal variation in level or timing of DNA replication initiation.

In *T. brucei*, centromere-associated repeats are located in strand switch regions which co-localise with either polycistronic transcription start sites, termination sites or both (Tiengwe, Marcello, Farr, Dickens, *et al.*, 2012; Tiengwe, Marques and McCulloch, 2014; Devlin *et al.*, 2016); characterisation of SSR type has not been carried out for this assembly as it would be beyond the scope of the project. However, investigating whether there is an association between the SSR type and motif group might provide an insight into whether the centromeric repeat sequences are dependent on the transcriptional activity of a given region. Centromeric repeats or their flanking sequences also act as early-acting DNA replication initiation sites (Tiengwe, Marcello, Farr, Dickens, *et al.*, 2012; Devlin *et al.*, 2016), which brings further complexity to the topic. Reverse chromatin immunoprecipitation experiments for the various motif sequences would be a valuable first step in investigating whether the different centromeric repeat motif groups have varying affinity for DNA replication, chromosomal segregation and transcription factors.

3.9.6 R-loop accumulation is associated with genome compartmentalisation

DRIPseq data from Emma Briggs' experiments (Briggs, Hamilton, *et al.*, 2018) has been mapped to the new assembly in order to assess R-loop levels across either previously unassembled sequences (such as repetitive sequences) or previously separate sequences that have been joined in the current assembly (core and subtelomeric regions). We see high R-loop signal across subtelomeric regions in the *Tbrh1-/-* cells deficient in RNaseH1 – a key protein that resolves R-loops (Briggs, Crouch, *et al.*, 2018). In contrast, in WT cells, there is very little R-loop mapping across subtelomeric regions. We can speculate that RNaseH1 in *T. brucei* may have a dedicated role in these genomic compartments, or, alternatively, subtelomeres are particularly prone to R-loop formation. The source of the RNA in such structures is unclear, however, as these regions are not thought to be transcribed. In *Leishmania major*, recent work (Damasceno *et al.*, 2024b) showed that later replicated parts of the genome accumulate higher levels of R-loops compared to earlier-replicated regions in WT cells and showed no obvious relationship with nascent RNA or gene expression levels. Based on the work carried out here, in *T. brucei*, curiously, the opposite pattern is evident as the earlier-replicating regions – megabase chromosome cores – accumulate higher levels of R-loops in the wild-type cells. Thus, R-loop accumulation in *T. brucei* is, perhaps, more clearly associated with transcriptional activity rather than replication (unlike in *L. major*).

3.9.7 R-loops in repetitive DNA: differential levels depending on repeat type

R-loop accumulation across repetitive regions appears to be repeat-type specific (Table 13). 50bp repeats show minor enrichment only in *Tbrh1-/-* cells at the repeat region flanking sites and no clear enrichment in the WT. 70bp repeat regions, consistent with Emma Briggs' data (Briggs, Hamilton, *et al.*, 2018), are highly enriched in R-loops in RNaseH1-null cells, but depleted in the WT. For centromeric regions, there is overall R-loop enrichment in both cell lines, but the signal is higher in *Tbrh1-/-* cells – though it should be noted that not all centromeric regions display this pattern. In 177bp repeat regions we see enrichment in *Tbrh1-/-* cells but not in the wild-type. Taking these data together, it seems that the mere presence of repetitive elements is not predictive of R-loop accumulation or resolution by RNaseH1 in *T. brucei*. One explanation for this variation may be levels of transcription in the repeats.

Table 13 R-loop enrichment or depletion across repetitive elements: summary

Cell line	70bp	50bp	Centro	177bp
WT	⇓	-	⇑	-
Tbrh1-/-	⇑⇑	⇑ at 5' and 3'	⇑⇑	⇑

Furthermore, in the case of centromeric repeats, their localisation is predictive of R-loop levels – centromeric repeats that localise to the megabase chromosome cores show the lowest R-loop levels (both in WT and RNaseH1-null cells), whereas those in subtelomeric regions, the highest. Strikingly, for chromosomes 2 and 6, where centromeric repeats are located in the boundaries between the core and subtelomeric compartments, R-loops in WT and RNaseH1-null cells are present at intermediate levels, consistent with their localisation 'between' these genomic compartments.

The above observation highlights two things. First, R-loop accumulation at centromeric repeats defies the broad R-loop level accumulation pattern in the respective genome compartments – regardless of the genomic compartment (core, subtelomere or boundary), R-loop levels in RNaseH1-null cells is consistently higher than in the WT at centromeric repeat regions. Second, despite no obvious replication timing difference between centromeres in the core vs subtelomeres, R-loop accumulation at these sites does vary in a predictable manner, with the regions showing the fewest origins of replication – subtelomeres – also showing the highest levels of R-loops. Combined, we believe this highlights the distinction between genic region and centromeric R-loop level patterns (Figure 54), suggesting possible differential regulation.

In model organisms, where the source and function of R-loops has been more extensively investigated than in kinetoplastids (reviewed in Petermann, Lan and Zou, (2022)), similarities with the observed R-loop accumulation patterns in *T. brucei* can be drawn – the co-transcriptional R-loops, most clearly associated with mRNA processing in this parasite, were previously shown in the original published work (Briggs *et al.*, 2018) and it is these that we see predominantly across the RNA pol II-transcribed core of megabase chromosomes. While this pattern is slightly exaggerated in the RNaseH1-null mutant (Briggs *et al.*, 2018), the more striking change in R-loop abundance is evident in the subtelomeres and some, though not all, repetitive genomic regions (70bp, 177bp and centromeric repeats, but not 50bp repeats), suggesting that these regions are likely acted upon by RNase H1 in *T. brucei*. Furthermore, the centromeric repeats display R-loop levels that relate to the genomic compartment said centromeres are positioned in; considering the pronounced differences in genomic stability and replication in the two biggest compartments of the megabase chromosomes (discussed futher in sections below), we speculate that this differential R-loop pattern is related to processes other than transcription, but rather DNA replication and/or the maintenance of genomic stability.

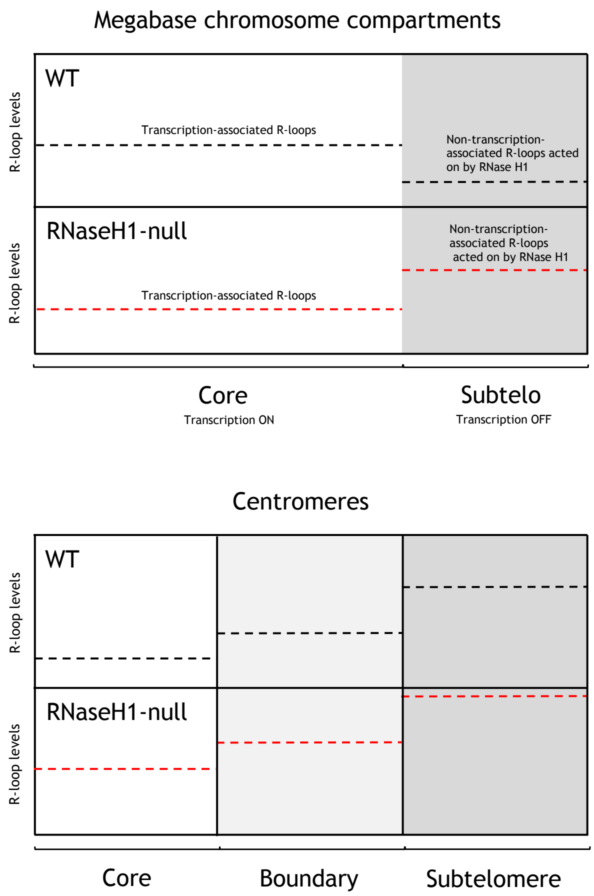


Figure 54 Summary of diverging R-loop patterns across the compartmentalised genome of *T. brucei*.

3.9.8 Compartmentalised, transcription-associated genome replication across *T. brucei* megabase chromosomes and BES

Finally, the genome assembly allowed us to expand the understanding of DNA replication dynamics in *T. brucei*. Previous work had identified population-level DNA replication timing across the core regions of the megabase chromosomes, showing that transcription unit boundaries, as well as centromeres, act in early S phase DNA replication. Additionally, the work showed that this pattern holds true and is consistent between at least two common lab strains of *T. brucei* (927 and 427), as well as in two distinct lifecycle stages – insect stage procyclics and mammalian bloodstream-form cells (Tiengwe *et al.*, 2012; Devlin *et al.*, 2016). One key difference between PCF and BSF was noted, however – the active BES in BSF cells (here, BES1 harbouring *VSG221*) was replicated early, whereas in PCF it was not, and this pattern was recapitulated in a cell line that had switched to expressing a different BES (BES3 harbouring *VSG121*), highlighting that the transcriptional state of the BES, rather than the sequence identity, was a likely determinant in this lifecycle stage-specific replication process (Devlin *et al.*, 2016).

Questions remained, however, about other genomic compartments of this parasite – namely, the subtelomeres of megabase chromosomes and the smaller, arguably lesser-studied, chromosomes. Additionally, while the centromeres (that are very prominent early-acting DNA replication initiation sites) had been mapped in megabase chromosomes 1-8 (Obado *et al.*, 2007; Echeverry *et al.*, 2012), it remained to be seen whether those of chromosomes 9-11 display similar replication patterns or were similarly composed of long repetitive DNA (as it is for chromosomes 1-8, but not, for instance, in *Leishmania* species). Furthermore, the direction (and hence origin) of DNA replication in the active BES in BSF could not be established at the time.

The novel 2018 Lister 427 genome assembly by Müller *et al.* (2018) provided some answers – subtelomeric regions were assigned to specific chromosomes and centromeric regions were pinpointed in the subtelomeres of chromosomes 9-11. However, the centromeric repeat regions contained scaffolding gaps, and the core and subtelomeric contigs remained as separate contigs. Additionally, the

BES sequences included in the assembly were also separate contigs, limiting analysis aimed at elucidating the direction of replication in the active BES.

To overcome these barriers, we utilised the long-read genome assembly generated and described here and mapped MFAseq data from PCF and BSF *T. brucei* Lister 427 cells (Devlin *et al.*, 2016).

This mapping has revealed several new features of DNA replication in *T. brucei*. First, we found that centromeric repeat regions of chromosomes 9-11, located in the subtelomeric regions, do indeed act early in S phase DNA replication, similarly to those of chromosomes 1-8, and display similar magnitude and replication timing. Second, outside of the centromeric repeats in the subtelomeres of chromosomes 9-11, strikingly, MFAseq does not reveal any detectable replication initiation sites in any subtelomeres. Subtelomeric DNA forms a significant fraction of the megabase chromosomes, in some cases comprising more than half of the chromosome, and so the lack of detectable replication is perplexing, especially because ORC has been shown to localise to this compartment (Tiengwe, Marcello, Farr, Gadelha, *et al.*, 2012; Maree *et al.*, 2017). Whether the subtelomeres are replicated from the core replication initiation sites or employ a replication initiation programme that is distinct from the core and undetectable via MFAseq is unclear from existing data.

In *Leishmania major* and *Leishmania mexicana*, MFAseq data had previously shown only one early-replicating region per chromosome – a unique finding in eukaryotes (Marques *et al.*, 2015). More recent work by Jeziel Damasceno et al. (2024a) investigated DNA replication dynamics in *L. major* in finer detail, examining DNA replication activity in single DNA molecules using Nanopore sequencing and analysis using DNAscent (Damasceno *et al.*, 2024a). Briefly, this technique relies on detecting the incorporation of thymidine analogs, such as BrdU, as the cells replicate, using Nanopore sequencing technology which can detect the presence of unconventional DNA bases, including analogs. The incorporation of the base analog is then mapped onto the genome, allowing the analysis of enrichment genome-wide. The software can detect fork direction, allowing predictions of replication initiation sites and replication termination sites (Boemo, 2021), and further analysis can be carried out to establish origin efficiency metrics, predominance of right- or left-moving forms, initiation and termination zones, among others. The work in *Leishmania major* confirmed that MFAseq-predicted early-replicating regions are sites of constitutive DNA replication activation in early S phase, but also revealed thousands of predicted DNA replication initiation sites that are spread throughout the rest of the chromosomes. The latter, lower-frequency and putative stochastic initiation events evaded detection by MFAseq; is it possible that *T. brucei* uses a similar approach in replicating origin-distal areas of the genome such as subtelomeres? The same experimental approach is currently – at the time of writing – being carried out in *T. brucei* by Gabriel Almeida da Silva in our lab; perhaps this work will clarify the mechanisms of subtelomere replication.

An alternative mechanism that might explain subtelomere replication dynamics relates to another finding of the work presented here – *T. brucei* telomeres display elevated MFAseq levels in S phase cells, indicative of these regions acting in DNA replication. While this is particularly prominent in BSF cells, PCF cells, at least in some cases, display higher MFAseq levels in the telomeric repeats compared to the upstream sequence. This finding might help explain how the subtelomeres are replicated – if there is replication originating from the core on one side and from the telomere at the other, perhaps, this is sufficient to replicate the entirety of the subtelomeres in this parasite.

Finally, the work presented here revealed that the early-replicated active BES (here, BES1) in bloodstream-form cells is replicated from the distal, telomeric end of the chromosome – the telomeric repeats, rather than from the upstream 50bp regions or from within the core or subtelomeric regions of the chromosome. In combination with the above observation that telomeric repeats, more generally, appear to act in early S replication, it is possible that telomere-directed DNA replication is initiated across all chromosome ends, but is extended further into only a subset of regions, such as the actively transcribed BES, and not into silent regions such as subtelomeres or inactive BESs.

Taken altogether, *T. brucei* appears to replicate its genome in a temporal fashion that reflects the transcriptional activity of the replicated region, with actively transcribed regions replicated first (Figure 55).

3.9.9 Compartmentalised, transcription-associated genome stability across *T. brucei* megabase chromosomes

To ask if the pronounced genomic compartmentalisation of the *T. brucei* genome extends to genomic stability, we grew *T. brucei* Lister 427 bloodstream-form cells in vitro for 23 passages (~140 cell doublings), sequencing whole genome DNA at the start and end of the experiment, as well as from final passage subclones. Previous work had shown BRCA2-/- *T. brucei* BSF cells grown *in vitro* displayed apparent chromosome shortening and loss of select *VSG* sequences (Hartley and McCulloch, 2008); to expand on that work, in addition to WT cells we also performed the *in vitro* experiments with RAD51-/- and BRCA2-/- cells. Pronounced genomic compartmentalisation was evident, particularly in RAD51-/- and BRCA2-/- cells, in terms of stability – while the cores of the megabase chromosomes displayed stable read depth coverage across the passage process, the subtelomeres showed abundant gross chromosomal changes, up to tens or hundreds of kilobases long, consistent with both deletions and copy number gain. Further work would be needed to establish whether there is any direct causal link between the compartmentalised DNA replication timing and genomic stability, however, the mere presence of such compartmentalisation is intriguing. There are striking parallels with recent works in *Sulfolobus* species of archaea (as mentioned in 3.1.4), where genome compartmentalisation in terms of transcription and replication is similar to that of *T. brucei,* albeit less extreme, and a causal relationship between the less actively transcribed and replicated regions displaying higher levels of genomic instability has been shown (Takemata, Samson and Bell, 2019; Badel, Samson and Bell, 2022).

3.9.10 177bp repeats act as DNA replication initiation sites

The assembly of longer 177bp repeat-containing contigs allowed us to examine DNA replication dynamics across the smaller, sub-megabase chromosomes of *T. brucei* for the first time. MFAseq data showed that DNA replication progresses outward from the 177bp repeats (Figure 55), highlighting their role as DNA replication initiation sites. Curiously, some megabase chromosome centromeres also contain a variable amount of the 177bp repeat motif, highlighting a likely evolutionary relationship between the two. Whether the 177bp repeat regions in the smaller chromosomes also act as centromeres remains to be seen, but given their apparent evolutionary relationships with megabase chromosome centromeres that are also early-replicated origins of DNA replication, this is likely to be the case.

While the genome assembly process only recovered, in all likelihood, a subset of the sub-megabase chromosomes present in the cell line, their apparent genomic stability throughout *in vitro* passage experiments, even in RAD51-/- and BRCA2-/- mutants, aligns with the previously reported mitotic stability of minichromosomes (Zomerdijk, Kieft and Borst, 1992; Wickstead, Ersfeld and Gull, 2003). Additionally, the genomic stability of these contigs is more similar to those of the cores of megabase chromosomes, rather than their subtelomeres; coupled with the fact that these contigs show evidence of gene expression (based on RNAseq data), the sub-megabase chromosomes, despite their subtelomere-like gene content, appear to more closely align with megabase chromosome cores (Figure 55).

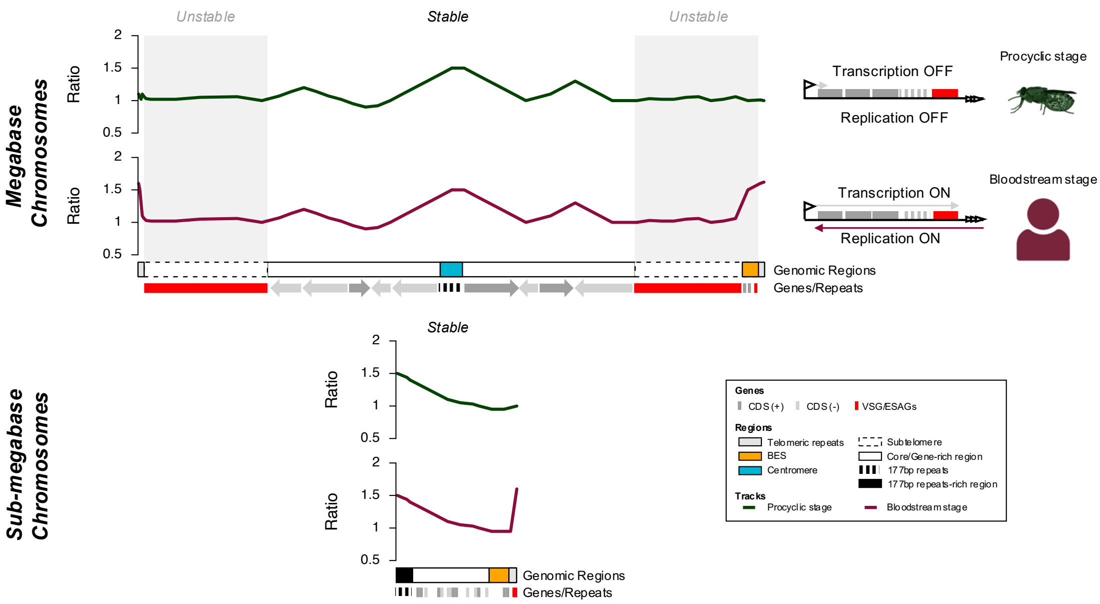


Figure 55 Summary of observed replication dynamics by MFAseq in *T. brucei* early S phase cells.
Figure created by Catarina de Almeida Marques. Created with BioRender.com

3.9.11 Limitations and future prospects

3.9.11.1 Improving assembly contiguity further

The assembly presented here has been produced solely through whole genome sequencing – long read Nanopore data supplemented with short read Illumina data for assembly polishing; there has been no validation of the presented data using other molecular biology techniques. As with any genome, there is some likelihood of misassembly. When the ONT sequencing for this assembly was carried out (mostly in 2019), obtaining a large amount of data or good quality long reads proved problematic, which is why a total of 13 MinION runs were used here. Since then, both our expertise on the topic and the quality of both the flow cells and reagents for MinION sequencing, as well as HMW DNA extraction, have improved, and as things stand currently, 2x or more data can be reliably obtained from a single MinION run for *Trypanosoma brucei*. Additionally, ONT have released the Ultra-Long DNA Sequencing Kit (SQK-ULK001) that can be utilised for very long sequencing read recovery. All of these factors combined mean it is now possible and realistic to obtain the foundation of a good assembly – a large amount of long reads from a single flask of cells - in a single sequencing run.

When performing genome assembly, there are a number of parameters that can be adjusted depending on the particular genome at hand; to simplify, the choice is between having a more compact assembly that does not preserve haplotypes, or a potentially largely duplicated one to achieve haplotype recovery. For the latter, a large amount of long reads is needed, up to 2x the amount needed for a 'regular' assembly (Koren *et al.*, 2017). As core regions of the megabase chromosomes of *T. brucei* are mostly conserved between the two copies of a homologous chromosome pair, whereas subtelomeric regions are not, assembly parameter choice is not an easy task for this parasite. Here, due to a limited amount of long read data, we chose the default assembly parameters when using canu (Koren *et al.*, 2017), as there wasn't enough data for good quality haplotype phasing.

As it has become easier to generate long ONT reads for *T. brucei*, haplotype preservation in assembly is a possibility now, with the goal of assembling telomere-to-telomere haplotype-phased chromosomes. Telomere-to-telomere assemblies have been produced in recent years, including in humans (Miga *et al.*, 2020; Nurk *et al.*, 2022), mainly through combining of several technologies, such as short read data (e.g. Illumina), more than one long-read data technology, e.g. ONT and PacBio, as well as chromatin conformation capture and optical mapping of chromosomes (Miga *et al.*, 2020; Nurk *et al.*, 2022). It is likely that improved ONT sequencing alone will not be sufficient for telomere-to-telomere assembly of all megabase chromosomes of *T. brucei*, let alone the unknown number of mini or intermediate ones; the obvious next step in assembly improvement following more ONT sequencing would be incorporation of chromatin conformation capture data and/or optical mapping such as Bionano, as these can aid in scaffolding and correcting existing assemblies (Miga *et al.*, 2020).

3.9.11.2 Low base quality at telomeric repeats of *T. brucei*

We were surprised to see that many 70bp repeat regions were truncated in the ONT genome assembly and we may have found the reason this is the case – base quality at telomeric repeats, which are within a few kilobases of 70bp regions, displayed markedly lower base quality compared to that of 70bp repeats or intervening or upstream sequences, regardless of the read depth coverage at those regions. While it remains unclear why *T. brucei* telomeric repeats as sequenced by ONT are lower quality than other sequences, there are a few possibilities that might explain this. First, telomeric repeats – TTAGGG – contain homopolymers – stretches of DNA sequence with repeated nucleotides; the key weakness of ONT sequencing lies in this exact nature of sequence (Wang *et al.*, 2021; Searle *et al.*, 2023). However, this did not appear to be an issue in the 70bp repeats which also harbour many repeated nucleotides (Figure 12). It is possible that a related, but more specific issue causes this - base quality issues in ONT sequencing specifically relating to telomeric repeats have been previously reported in telomere-to-telomere human chromosome assembly (Tan *et al.*, 2022), as well as in a range of other genomes that contain TTAGGG or very similar repeats. Analysis showed that this issue is likely related to ONT basecalling methods, as training a basecaller with TTAGGG-containing reads improved telomeric sequence base quality; this explanation might also apply to base quality issues at telomeres of *T. brucei*.

Another possible reason, which might also explain a curious observation in Chapter 4, is a high prevalence of modified bases at telomeres, as these could impede basecalling and lead to lower base quality. In *T. brucei*, modified DNA bases have been described, including specifically in the telomeres (discussed in the following chapter), and it is possible that either these or a different base modification are present in the telomeric repeats of this parasite. In a variety of other organisms, including plants and humans, telomeres accumulate a guanine (G) modification due to oxidative stress – 8-oxoguanine (8-*oxoG*) (An *et al.*, 2015; Castillo-González *et al.*, 2022). While this is speculative and 8-oxoG has not been described in *T. brucei*, the pronounced base quality issues at telomeres seen here and a possible G base modification identified in the following chapter are also consistent with 8-*oxoG* presence at *T. brucei* telomeres.

3.9.11.3 Utility of any single reference in a complex and repetitive genome

The genome presented here is a single attempt of assembling a complex and repetitive genome of *T. brucei*; it could be argued that the utility of any single reference in this context is limited due to the changing nature of this parasite's genome, particularly in the subtelomeric compartments, and possibly repetitive elements. However, in order to even begin to study the stability of certain genomic features, for example, there needs to be a starting 'reference' point which can be referred to – a representative and broad collection of genomic features and repetitive motifs. While any particular fully-assembled repetitive region might change from one sequencing experiment to another, the utility of this assembly lies in providing a comprehensive reference point for any future analysis, be it through sequencing experiments, reverse ChIPseq experiments or probe-base analysis, all of which require good quality DNA sequence.

3.9.11.4 Limitations of using MFAseq to assess DNA replication dynamics

MFAseq is a technique that relies on population-level data; as discussed above, in *Leishmania major* MFAseq had predicted the localisation of the predominant, constitutive origins in the population, however, more recent techniques (DNAscent) provided molecule-level detail that highlighted the replication dynamics across the whole of the genome, not just at the one-per-chromosome constitutive origins (Damasceno *et al.*, 2024a). As mentioned, the same approach is currently being applied to *T. brucei* in our lab, and the results are yet to be seen. A further step in assessing DNA replication dynamics would be the analysis of single cells, rather than relying on population-level data; it is technically possible, albeit challenging, to perform MFAseq or DNAscent on single cells – this would provide insight into how an individual cell, rather than a population, performs DNA replication. This approach would, in a way, mimic older DNA combing methods, except it would provide crucial detail – sequence and cell identity.

4 Genome-wide detection of DNA modifications in the *Trypanosoma brucei* genome using Nanopore sequencing.

4.1 Introduction

4.1.1 Known modified DNA bases of *T. brucei*

In addition to conventional DNA bases - adenine, cytosine, guanine and thymidine - *Trypanosoma brucei* possesses the hypermodified DNA base ß-D-glucosyl-hydroxymethyluracil, also known as base J. The initial indications of this parasite harbouring a modified base came from an observation that some of its genomic sequences cannot be fully cleaved by restriction enzymes PstI and PvuII (Bernards, van Harten-Loosbroek and Borst, 1984). It was later shown that the reason for this is the presence of a modified version of thymidine that was not amenable to digestion by restriction enzymes (Gommers-Ampt *et al.*, 1993).

Base J is thought to be synthesised in a two-step process (Figure 56). First, a thymidine nucleoside (T) is modified by thymidine hydroxylases to form 5-hydroxymethyluracil (5hmU, also known as base V and HOMedU), followed by glycosylation by a ß-glucosyl-transferase to form base J (Borst and Sabatini, 2008; Bullard *et al.*, 2014).

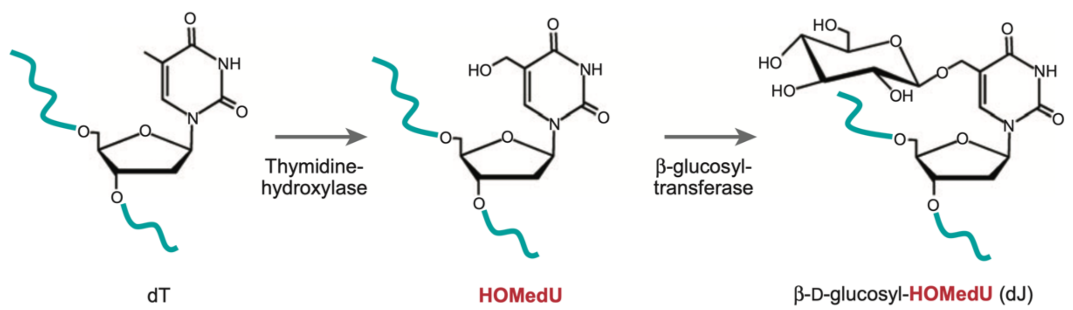


Figure 56 Presumed two-step process of base J synthesis from thymidine.

Arrows indicate the direction of the chemical reaction, teal lines represent the DNA backbone. dT-deoxythymidine, HOMedU – 5-(hydroxymethyl)uridine, dJ – base J. Figure reproduced with permission from Borst and Sabatini, (2008) (license number 1591516-1).

Not all thymidines and not all 5hmU become base J – in fact, only a fraction of thymidines become modified (see below), both genome-wide and at any given position in a cell population (Borst and Sabatini, 2008). This means that the parasite harbours more than one modified DNA base, as demonstrated by 2D thin layer chromatography, quantitative mass spectrometry, as well as immunoprecipitation assays (Table 14) (Gommers-Ampt, Lutgerink and Borst, 1991; Fred van Leeuwen *et al.*, 1998; Liu *et al.*, 2014). In addition to highlighting the presence of two modified bases in the parasite's genome, these early experiments suggested that base J is only present in the mammalian-infective bloodstream-form cells (BSF) and not in insect-stage procyclic cells (PCF) (Figure 57) (Gommers-Ampt, Lutgerink and Borst, 1991; Fred van Leeuwen *et al.*, 1998). Furthermore, the estimated fraction of bases corresponding to base J is small – up to 0.09% or 0.12% of total DNA in BSF and <0.00025% or none at all in PCF based on early 2D thin layer chromatography (2D-TLC) experiments (Gommers-Ampt, Lutgerink and Borst, 1991; Fred van Leeuwen *et al.*, 1998). More recent quantitative mass spectrometry data indicated that 0.15% of all bases correspond to base J in BSF (Liu *et al.*, 2014), and anti-J immunoblot experiments estimated about 1.6% of all DNA to be base J in *T. brucei* BSF cells (Fred van Leeuwen *et al.*, 1998). While there may not be agreement in how much of the parasite's DNA is modified (Table 14), it appears that overall the modification is not prevalent in its genome, but at very low levels in BSF cells and not detectable in the insect-stage cells.

In addition to base J and base V, more recently 5-formyluracil (5fU) has been identified in the DNA of *T. brucei* (Bullard *et al.*, 2014). From the very limited information that is known about this base and its role in the parasite's biology, it appears that it might be present at 1.5 to 2 x higher levels than 5hmU. In the study, the authors suggested that 5hmU may act as a precursor to 5fU, but this has not been further explored (Bullard *et al.*, 2014).

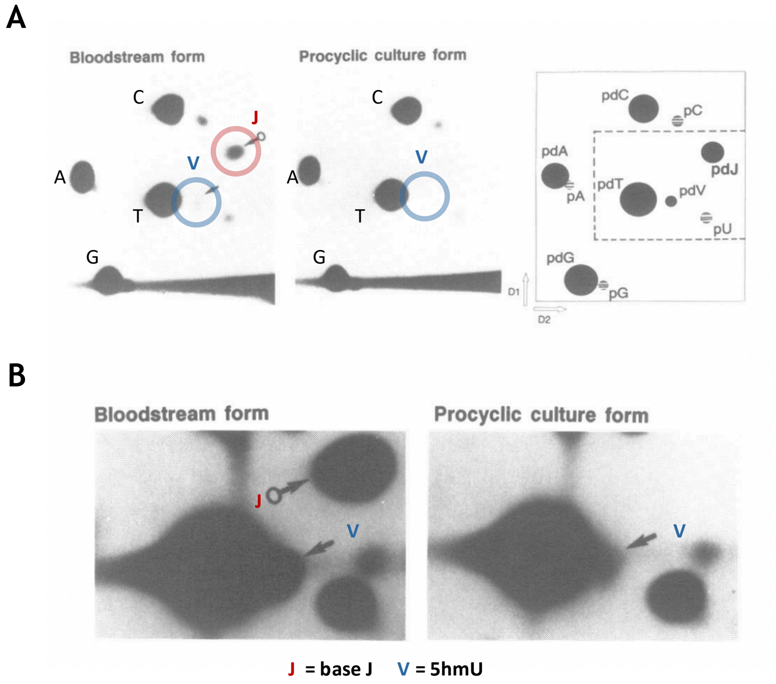


Figure 57 Autoradiograms showing nucleobases detected in *T. brucei* using two-dimensional thin layer chromatography (2D-TLC).

A - ^{32}P-postlabeled nucleotides from *T. brucei* minichromosomes in bloodstream-form and procyclic-form cells. B - Longer exposure of autoradiograms in A; only region indicated by dashed lines on the right-hand panel in A shown. J – base J, V – 5hmU. Adapted from Gommers-Ampt, Lutgerink and Borst, (1991) (reproduced with permission, license number 5995370273526**)**.

Table 14 Summary of base J and V quantification results in *T. brucei* over the years.

2D – two dimensional, LC-MS/MS – liquid chromatography with tandem mass spectrometry, BSF – bloodstream-form cells, PCF – procyclic cells, J – base J, V – 5hmU.

Method	% J in BSF	% V in BSF	% J in PCF	% V in PCF	Source
^{32}P post-labelled 2D thin layer chromatography coupled with Cerenkov assay	0.04-0.09	-	<0.00025	-	Gommers-Ampt, Lutgerink and Borst, (1991)
^{32}P post-labelled 2D thin layer chromatography coupled with phosphorimager analysis	0.12*	0.04	0	0.02	Fred van Leeuwen, Taylor, *et al.*, (1998)
anti-J immunoprecipitation	1.6	0.2	0	0	Fred van Leeuwen, Taylor, *et al.*, (1998)
quantitative mass spectrometry (LC-MS/MS)	0.15	0.00425	-	-	Liu *et al.*, (2014)

4.1.2 Base J and V distribution genome-wide

The distribution of base modifications across the genome appears to be non-random. Repetitive elements - including telomeric TTAGGG, 50bp, 177bp, 70bp repeats - all appear to be enriched in base J to varying degrees (van Leeuwen *et al.*, 2000; Cliffe *et al.*, 2010). Whole genome anti-J immunoprecipitation coupled with sequencing showed enrichment in base J in WT cells at telomeric and subtelomeric sequences (Cliffe *et al.*, 2010), including silent bloodstream-form expression sites (BES) and the 50bp and telomeric repeats surrounding both active and inactive BES (Leeuwen *et al.*, 1997). In addition, most, though not all, polycistronic transcription unit (PTU) boundaries, also known as strand switch regions (SSRs), have been showed to harbour base J (Cliffe *et al.*, 2010). More recently, enzyme-mediated bioorthogonal labeling coupled with sequencing was performed for genome-wide mapping of 5hmU in *T. brucei* (Ma *et al.*, 2021). In contrast with base J genome-wide localisation, 5hmU enrichment was seen predominantly across coding sequences (62.15% of detected peaks); only limited analysis of 5hmU genome-wide distribution has been performed, as the work was focused on the methodology rather than in-depth analysis of 5hmU distribution in the *T. brucei* genome (Ma *et al.*, 2021).

4.1.3 Base J may function in transcriptional repression or termination.

Broadly speaking, base J is thought to be involved in transcriptional repression or termination in *T. brucei*. Unlike in several closely related *Leishmania* species, base J is not essential for *T. brucei* (Cliffe *et al.*, 2009). As mentioned above, one of the key sites of base J localisation in *T. brucei* are SSRs (Cliffe *et al.*, 2010), which can be a combination of polycistronic Pol II-driven transcription start and termination sites (head-to-tail PTU boundaries), bidirectional transcriptional start sites (divergent SSRs) or solely termination sites (convergent SSRs) (Clayton, 2019). Elimination of base J at convergent SSRs (cSSRs) appears to lead to read-through transcription in *T. brucei* as detected using total RNAseq (Schulz *et al.*, 2016), though the same was not observed when using only small RNAseq, despite the latter being effective at highlighting read-through transcription at SSRs caused by base J loss in *Leishmania* (Reynolds *et al.*, 2014).

In addition to PTU boundaries, base J is also present in PTU-internal regions in *T. brucei*, and the elimination of base J at these sites upregulates transcription of downstream genes within the same PTU (Reynolds *et al.*, 2014). Thus, the function of base J at these PTU-internal transcription termination sites is thought to act as developmental expression regulation of downstream genes, as base J is thought to be absent in procyclic cells (Reynolds *et al.*, 2014). Similarly, localisation of base J to repetitive elements and inactive BES is speculated to act in repressing transcription at these regions (Leeuwen *et al.*, 1997; Borst and Sabatini, 2008), though this has not been explored.

Base J appears to co-localise to some degree with histone 3 variant (H3.V) in *T. brucei* and play similar repressive roles in gene expression (Cliffe *et al.*, 2010; Schulz *et al.*, 2016). Both base J elimination and H3.V deletion, separately, led to increased expression of a subset of previously silenced VSG genes, and this effect was exacerbated in double null mutants (ΔH3.VΔJ) (Schulz *et al.*, 2016).

Whether 5hmU has a function, beyond acting as a precursor for base J, is unknown. The glycosylation of this modified base by J-specific glycosyltransferase (JGT) to form base J is proposed to be independent of sequence specificity of this protein *in vitro* (Bullard *et al.*, 2015); it is unknown what dictates thymidine or 5hmU modification at any given locus. Exogenous 5hmU incorporation appears to lead to base J formation in genomic regions that normally lack base J (F. van Leeuwen *et al.*, 1998); lack of JGT sequence specificity *in vitro* might suggest that thymidine modification to form 5hmU – the first step to thymidine modification to base J - may predetermine base J positioning in the genome, rather than 5hmU modification to base J (Bullard *et al.*, 2015), in which case broad patterns of 5hmU and base J localisation would overlap, however this is speculative. Certainly, published 5hmU quantification assays (Table 14) suggest that genome-wide 5hmU levels are very low compared to base J levels, suggesting that most 5hmU bases become further modified to form base J. The limited genome-wide assays to determine base J and 5hmU localisation, at first glance, don't clarify this picture as the published results (Cliffe *et al.*, 2010; Ma *et al.*, 2021) focus on different genomic elements.

4.1.4 Detecting modified bases using third generation sequencing technologies.

A relatively novel approach to elucidating the presence of modified nucleobases in DNA or RNA is by utilising third generation sequencing technologies. Both major technologies at the time of writing – Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT or Nanopore) – possess this ability. The two technologies use different metrics in order to detect the presence of modified bases in DNA or RNA (Figure 58), and both offer single base resolution and strand-specific data (Searle *et al.*, 2023).

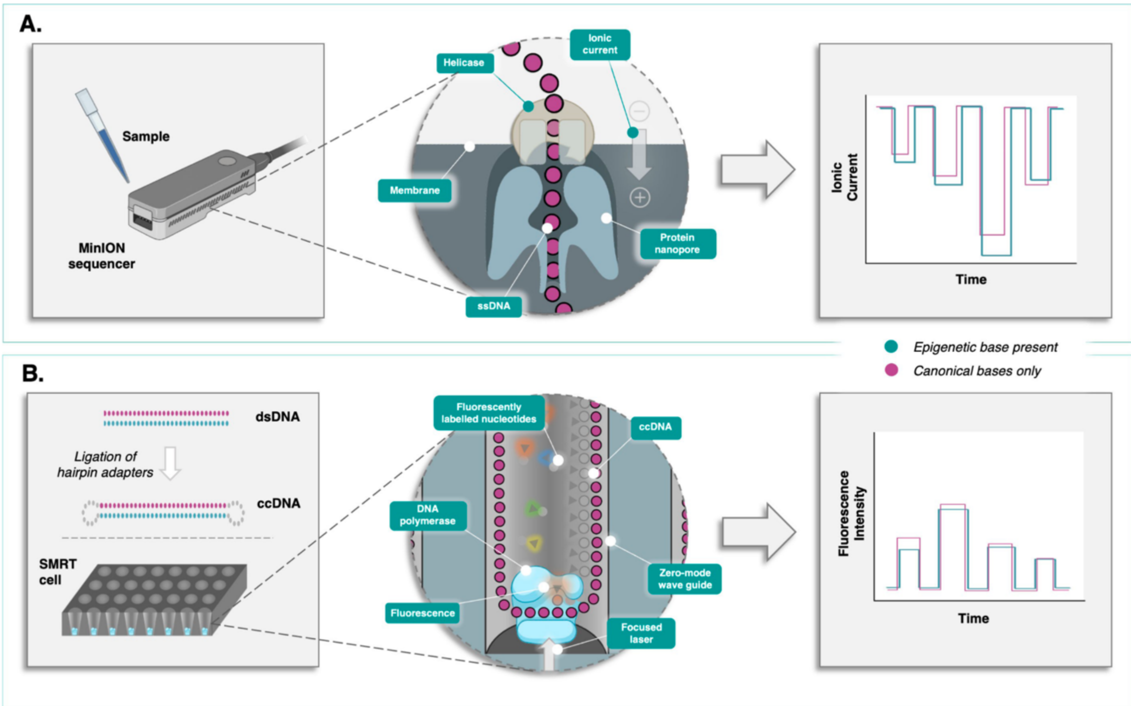


Figure 58 Overview of modified nucleobase base detection using third generation sequencing technologies Nanopore and PacBio SMRT.

A - An example of Nanopore sequencing using a MinION device. Nanopore proteins are suspended in a membrane in the sequencing flow cell; as a tether protein with a helicase attaches to input DNA fragment, it becomes unwound and a single strand DNA (ssDNA) passes through the pore. The passing of the DNA changes the ion flow through the pore channel – the deviations in the ion current are detected and recorded for each hexamer of sequence. The deviation of the ion current from the canonical nucleobases (A, T, G and C) is used to assess the presence of modified bases. B - In PacBio SMRT sequencing, DNA is circularised, and a DNA polymerase attaches a fluorescently labelled complement nucleotide – this process is monitored both in terms of the wavelength of the light emission and the speed of polymerase processivity. Figure reproduced from Searle *et al.*, (2023) (Open Access license, Creative Commons CC-BY-ND 3.0).

PacBio SMRT sequencing, similar to next generation sequencing (NGS) technologies such as Illumina, works by incorporating fluorescent dNTPs complementary to a template strand and capturing the emitted fluorescence. The dNTP incorporation dynamics can be utilised to detect modified bases on the template strand: in particular, measured inter-pulse duration (IPD) is altered in the presence of modified bases due to polymerase pausing (Searle *et al.*, 2023).

PacBio sequencing has been utilised in *Leishmania tarentolae*, a related kinetoplastid, in order to investigate whether base J can be detected using this technology and to characterise base J-containing sequences (Genest *et al.*, 2015). Unlike *T. brucei*, in *Leishmania* species studied to date the majority of base J is reported to be telomeric, and, unsurprisingly, base J was detected within DNA segments containing *L. tarentolae* telomeric repeats. Specifically, two T residues that are on opposite strands and 12 bases apart appear to be the most prominent telomeric base J modification pattern in these parasites (Genest *et al.*, 2015), consistent with previous observations in *T. brucei* (van Leeuwen *et al.*, 1996). PacBio SMRT sequencing does not appear to have been used to investigate broader base J deposition patterns.

Nanopore sequencing is distinct from Sanger, Illumina or PacBio sequencing as it does not rely on fluorescently labelled dNTPs. ONT sequence capture is carried out by passing a single DNA or RNA strand through a dedicated nanopore on a synthetic membrane that is within the sequencing flow cell; as the fragment passes through a given nanopore, the ionic current is altered, and this fluctuation in signal is recorded (White and Hesselberth, 2022). Each nucleobase has a specific signature, which is used in the process of basecalling. Thus, modified bases can be sequenced as well, and they produce a unique signature in the current, which can then, in principal, be assigned to a given modification (Searle *et al.*, 2023). Characterisation of the signal for common base modifications, such as CpG methylation, has been carried out and many tools exist that can identify specific bases (Searle *et al.*, 2023). *De novo* base modification detection is also a possibility.

4.1.5 Aim

While genome-wide mapping of the modified bases in *T. brucei* has been performed earlier, these datasets, in our view, have not been fully evaluated across the different genomic features of the parasite. Even so, expanding the analysis of these datasets will be limited by the short read sequencing technologies employed, as these cannot provide base-level or strand-specific localisation of modified bases. The aim of this chapter is:

1. Expand the analysis of published base J and 5hmU datasets to provide a more detailed and broad overview of modified base distribution in the *T. brucei* genome.

2. Investigate whether *de novo* modified base detection is possible in *T. brucei* using Nanopore sequencing and existing software packages.

3. Provide more comprehensive modified base analysis in *T. brucei* through combining published datasets and *de novo* modified base detection using Nanopore presented here (if the data allows this).

4.2 Results

In order to detect base modifications in *T. brucei*, we sequenced *T. brucei brucei* Lister 427 bloodstream-form and procyclic cells on a MinION ONT sequencer. A total of 5.62Gb and 483.67 thousand reads were sequenced (BSF and PCF data combined), with read N50 – 40.77 kb. The raw electric current signal data, in fast5 format, was re-squiggled and further processed using Tombo - a dedicated tool for modified base detection using Nanopore sequence data (Oxford Nanopore Technologies, 2018). Briefly, the re-squiggling process aligns the raw signal and associated basecalls to the reference sequence; this alignment is then used to run the modified base detection commands. Three Tombo modes were used here – *de novo*, model sample, and level sample. *De novo* mode identifies signal deviation in a given sample relative to the canonical expected base model, and this is done in each read at every genomic position. Model sample and level sample Tombo modes differ in that they require a 'no-modification' control set of reads to compare the signal against. In the model sample mode, the control sample is used to adjust the canonical base model; this adjusted model is then used in what is effectively *de novo* mode to identify signal deviations in the modified sample. The third Tombo mode, level sample mode, operates slightly differently – it compares two sets of reads to identify signal variation at a given reference position (Oxford Nanopore Technologies, 2018).

For our purposes, we used procyclic-form (PCF) *T. brucei brucei* Nanopore sequencing data as control reads in the two comparison modes, as well as performed *de novo* modified base identification in both bloodstream-form (BSF) and PCF cells, independently (Table 15). This was carried out using two genome sequences as references – the TriTrypDB *T. brucei* Lister 427 2018 genome (build version 46) (Müller *et al.*, 2018), as well as the Nanopore-based *T. brucei* Lister 427 genome long-read assembly described in Chapter 3.

Table 15 Summary of *T. brucei* datasets analysed in this chapter.

BSF – bloodstream-form, PCF – procyclic-form, 5hmU – 5-hydroxymethyluracil, N/A – not applicable.

Dataset	Strain	Lifecycle stage	Post-infection?	Replicates	Sequencing	Input data available	Mode	Output	Strand specific	Base resolution
5hmU ChIPseq	Antat 1.1	BSF	Yes, mouse	2	Illumina, PE 150bp	Yes	N/A	N/A	No	No
Base J ChIPseq	Lister 427	BSF	No	1	Solexa 32-36bp	No	N/A	N/A	No	No
Tombo *de novo* PCF	Lister 427	PCF	No	1	ONT (MinION, R9.4.1.)	N/A	*De novo*	Fraction of bases modified	Yes	Yes
Tombo *de novo* BSF	Lister 427	BSF	No	1			*De novo*		Yes	Yes
Tombo model sample	Lister 427	BSF / PCF	No	1			Model sample		Yes	Yes
Tombo level sample	Lister 427	BSF / PCF	No	1			Level sample	Test statistic	Yes	Yes

4.2.1 Overview of genome-wide patterns

For simplicity and reproducibility, the majority of the analysis described in this chapter was carried out using the published TriTrypDB reference sequence. The reasons for this are three-fold. Firstly, because some regions of interest (ROI) have more complete annotations – for example for the polycistronic transcription unit boundaries, which have not been annotated in the Nanopore assembly. Secondly, the splitting of contigs into core, subtelomeric and BES contigs in the reference makes it convenient to compare mapping across genome compartments. And, lastly, it becomes easier to interpret for a reader that might be accustomed to the more clearly organised and well annotated reference genome.

Looking at an overview mapping of the existing datasets – 5hmU ChIPseq, base J ChIPseq, along with the new Tombo-based data (Figure 59) - overlap in peaks was evident in the genome core between BSF and PCF Tombo samples in *de novo* mode, as well as the comparison (BSF/PCF) model sample mode, with peaks co-occurring in the three datasets. The signal on the reverse strand (in red) often followed the signal on the forward strand (in blue), although this was not always the case. The other comparison method – level sample - provides a test statistic instead of fraction of reads modified, and the significantly different regions didn't follow the signal patterns of the other two modes or the ChIPseq datasets mentioned above. In the base J ChIPseq data, enrichment was largely co-localised with putative transcription termination and start sites (TTS/TSS) and centromeres. The two 5hmU ChIPseq datasets were consistent in their signal pattern, though not in enrichment values, and the data didn't clearly colocalise with any of the other datasets or genomic features annotated here.

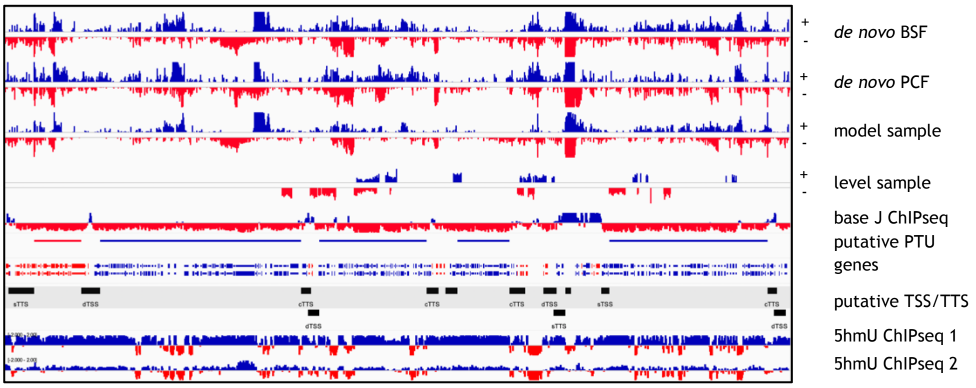
4.2.1.1 Genomic compartments and possible base modifications

To investigate whether there is any difference in signal distribution between genomic compartments, we mapped the datasets to the core, subtelomeric and BES sequences of the TriTrypDB genome (Figure 60). Consistent with previously published data (Reynolds *et al.*, 2016), the base J ChIPseq data suggests the subtelomeric compartments are enriched in base J relative to the core genome. This pattern was reversed in the 5hmU ChIPseq datasets, although it should be noted that the 5hmU experiments were performed in a different strain of *T. brucei brucei* (AnTat 1.1 as opposed to Lister 427), and we don't know how different the subtelomeric compartments of this strain are compared to Lister 427; certainly, other strains display considerable variation in subtelomere content and length (Callejas *et al.*, 2006; Müller *et al.*, 2018) and so it is possible that the observed signal was lower because of poor mapping to the Lister 427 reference genome.

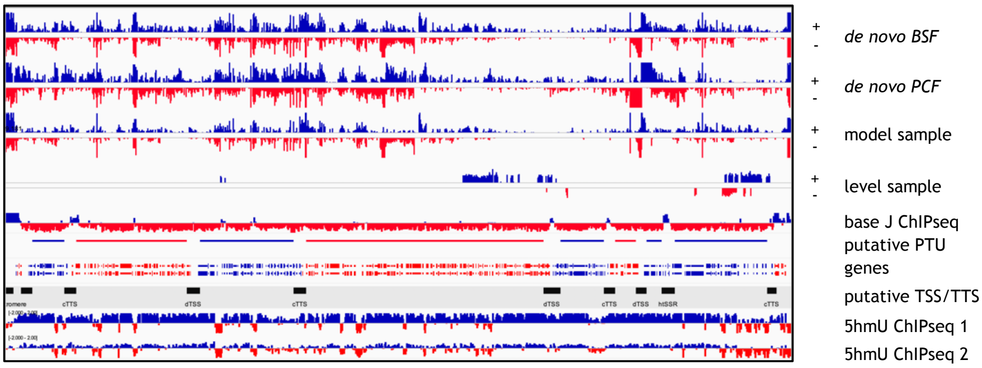
Tombo-based modified base detection didn't suggest clear overall signal differences between genomic compartments. If the 5hmU and base J ChIPseq datasets are representative of the levels of the respective modified bases in the genome, the level of base J was higher in the subtelomeres, whereas the level of 5hmU was higher in the core, it is possible that what we are seeing in the Tombo datasets is the sum of both modifications, and the overall level of base modification is very similar in these compartments. Because Tombo doesn't make a distinction between different modifications, or, indeed, suggest what the modifications are, the signal that is picked up is representative of the overall level of signal divergence from the canonical bases at a given locus and not representative of the nature or identity of the modification. The other possibility is that Tombo was only picking up one of the modifications and that modification was present in both lifecycle stages and at similar levels regardless of the genomic compartment, although this is arguably less likely.

Interestingly, some strand-specific divergence was apparent from high-level overviews: specifically, at the 5' region of the BES contigs, where 50bp repeats reside, the signal was more pronounced on the forward (plus) strand in *de novo* BSF and PCF datasets, as well as model sample mode. Tombo level sample mode showed very few significantly modified regions outwith the core genome compartment, and high-level overview mapping was not informative for this dataset.

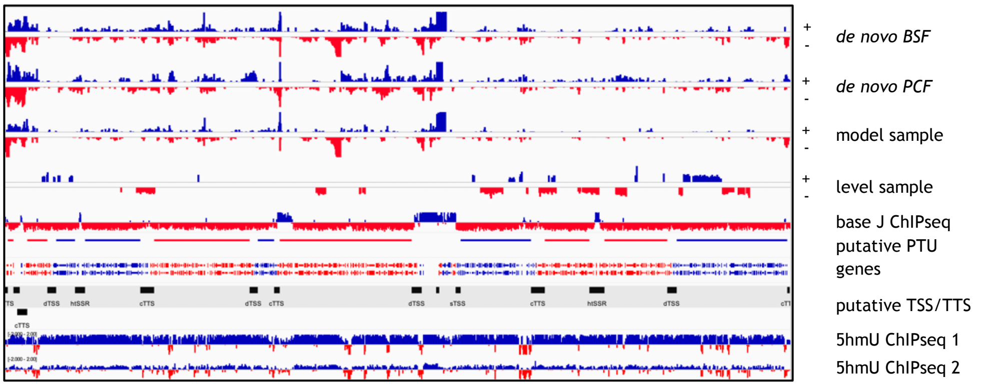
Chr1



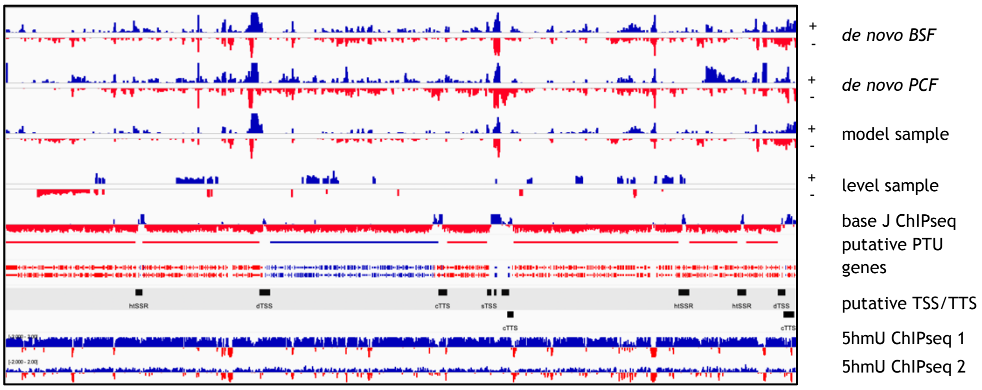
Chr2



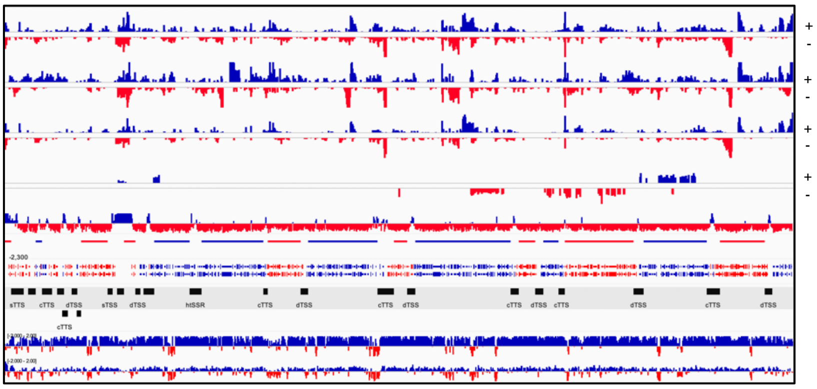
Chr3



Chr4

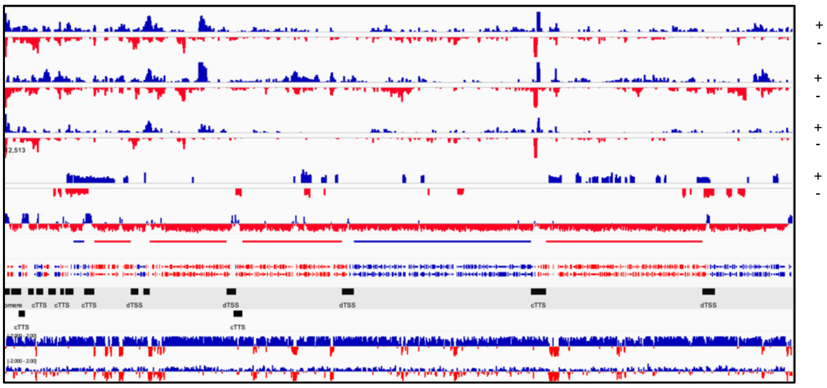


Chr5



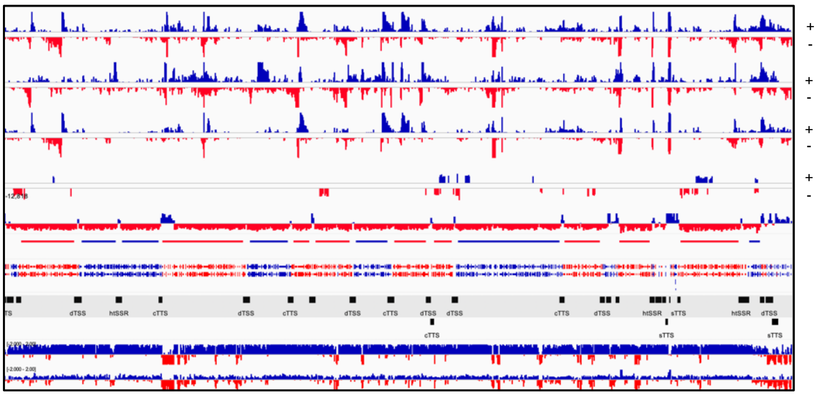
de novo BSF
de novo PCF
model sample
level sample
base J ChIPseq
putative PTU
genes
putative TSS/TTS
5hmU ChIPseq 1
5hmU ChIPseq 2

Chr6



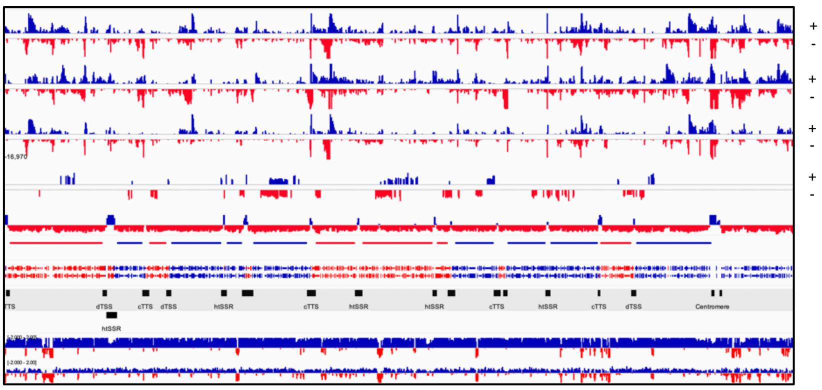
de novo BSF
de novo PCF
model sample
level sample
base J ChIPseq
putative PTU
genes
putative TSS/TTS
5hmU ChIPseq 1
5hmU ChIPseq 2

Chr7



de novo BSF
de novo PCF
model sample
level sample
base J ChIPseq
putative PTU
genes
putative TSS/TTS
5hmU ChIPseq 1
5hmU ChIPseq 2

Chr8



de novo BSF
de novo PCF
model sample
level sample
base J ChIPseq
putative PTU
genes
putative TSS/TTS
5hmU ChIPseq 1
5hmU ChIPseq 2

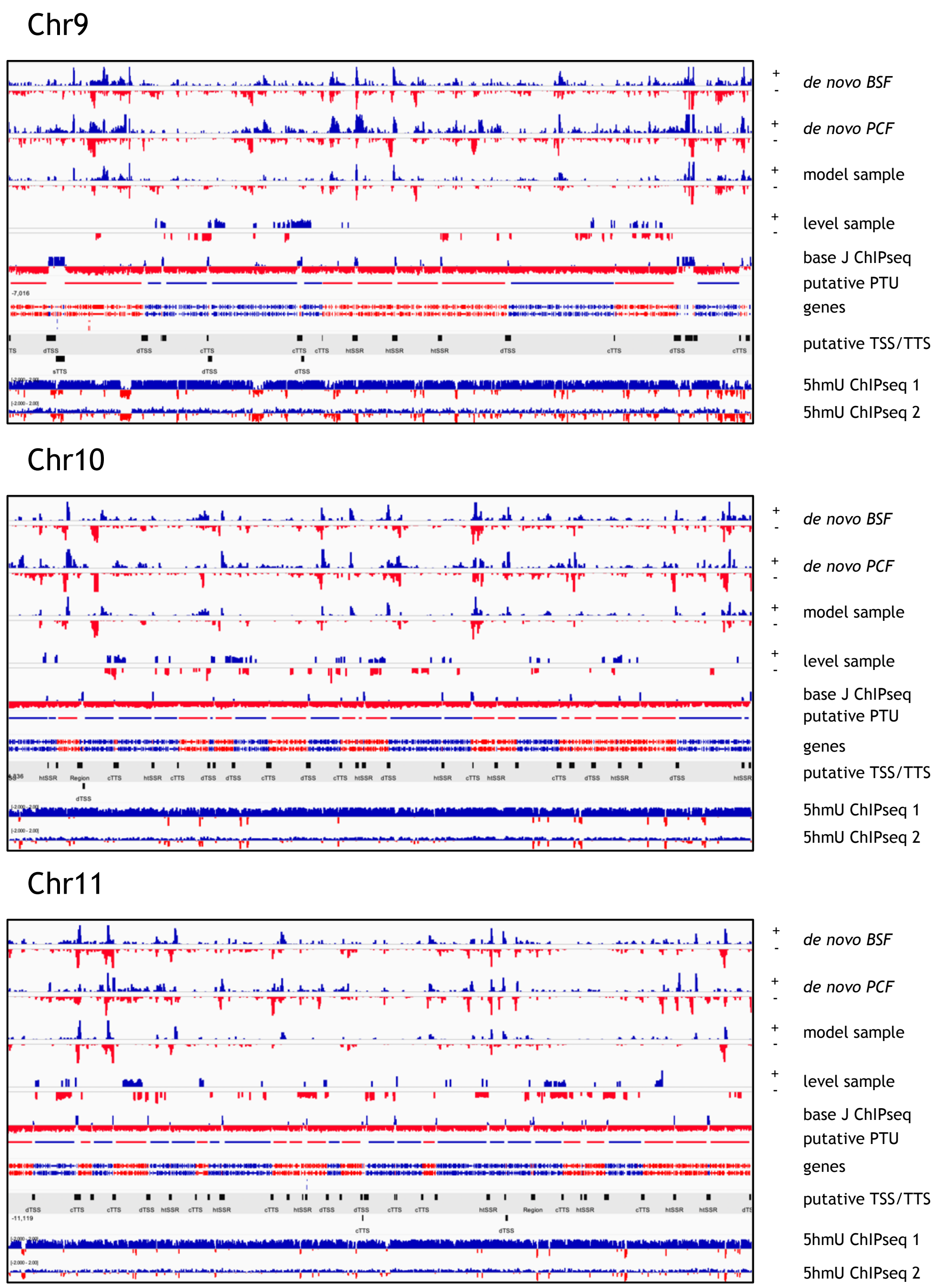


Figure 59 Overview of modified base mapping across the core regions of the *T. brucei* genome.

Chr – chromosome, '+' forward or top strand, '-' reverse or bottom strand, TTS – transcription termination site, TSS – transcription start site, PTU – polycistronic transcription unit, BSF – bloodstream-form cells, PCF – procyclic cells. *De novo*, model sample and level sample refer to the different outputs of Tombo.

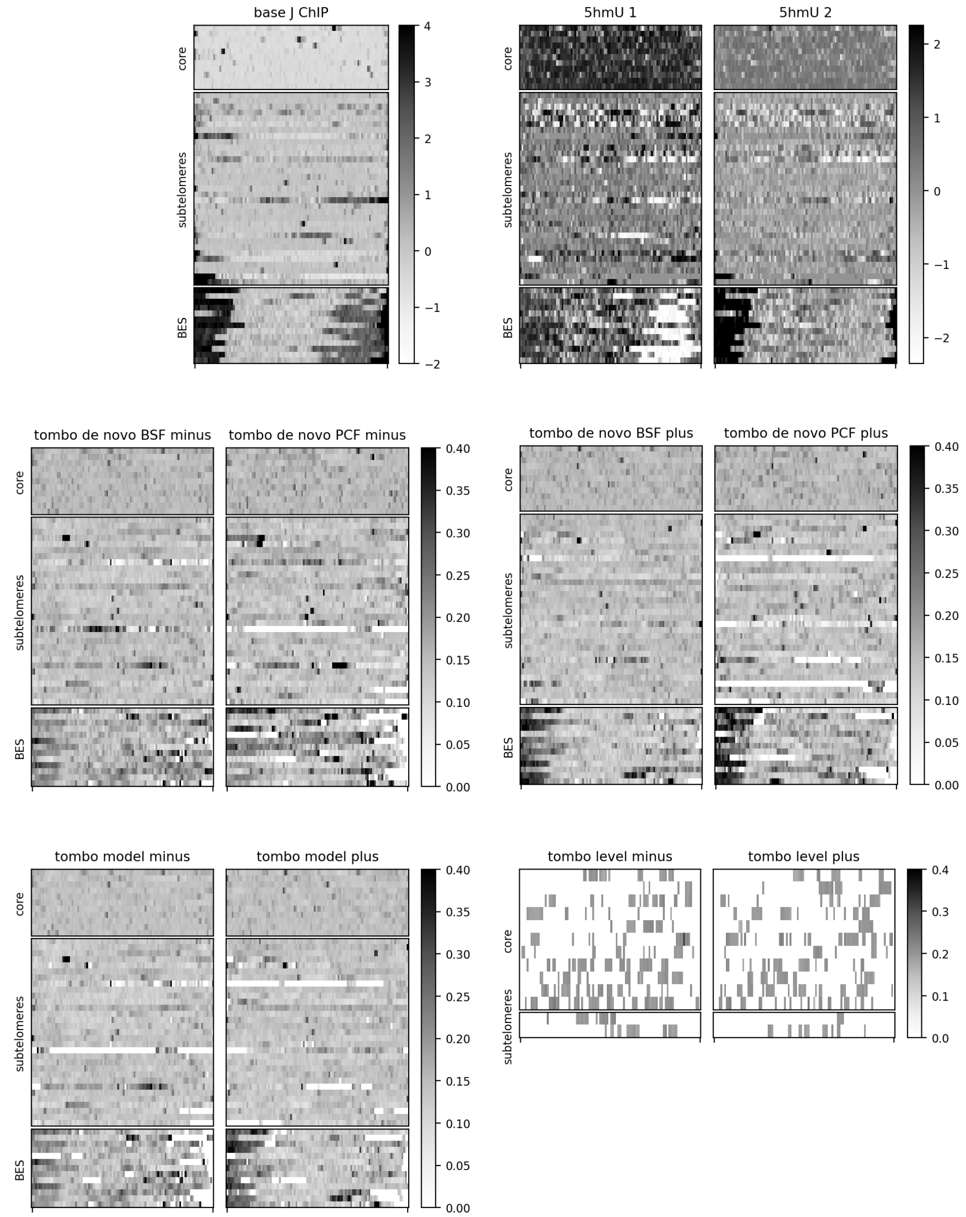


Figure 60 Modified base distribution across three genomic compartments of *T. brucei* – core, subtelomeric and BES regions.

Base J ChIPseq, 5hmU ChIPseq and Tombo datasets mapped to the Lister 427 2018 reference genome from TriTrypDB. Each line represents a full reference contig. Minus – reverse or bottom strand, plus – forward or top strand, BES – bloodstream-form expression sites.

4.2.1.2 Transcription-associated genomic regions

In previously published literature, one of the key genomic features that base J has been associated with is transcription termination sites (TTS) and strand switch regions more broadly (Cliffe *et al.*, 2010; Schulz *et al.*, 2016). In light of this, we decided to investigate base modification patterns across putative TTSs and TSSs (Figure 61), as well as across PTUs more generally (Figure 62).

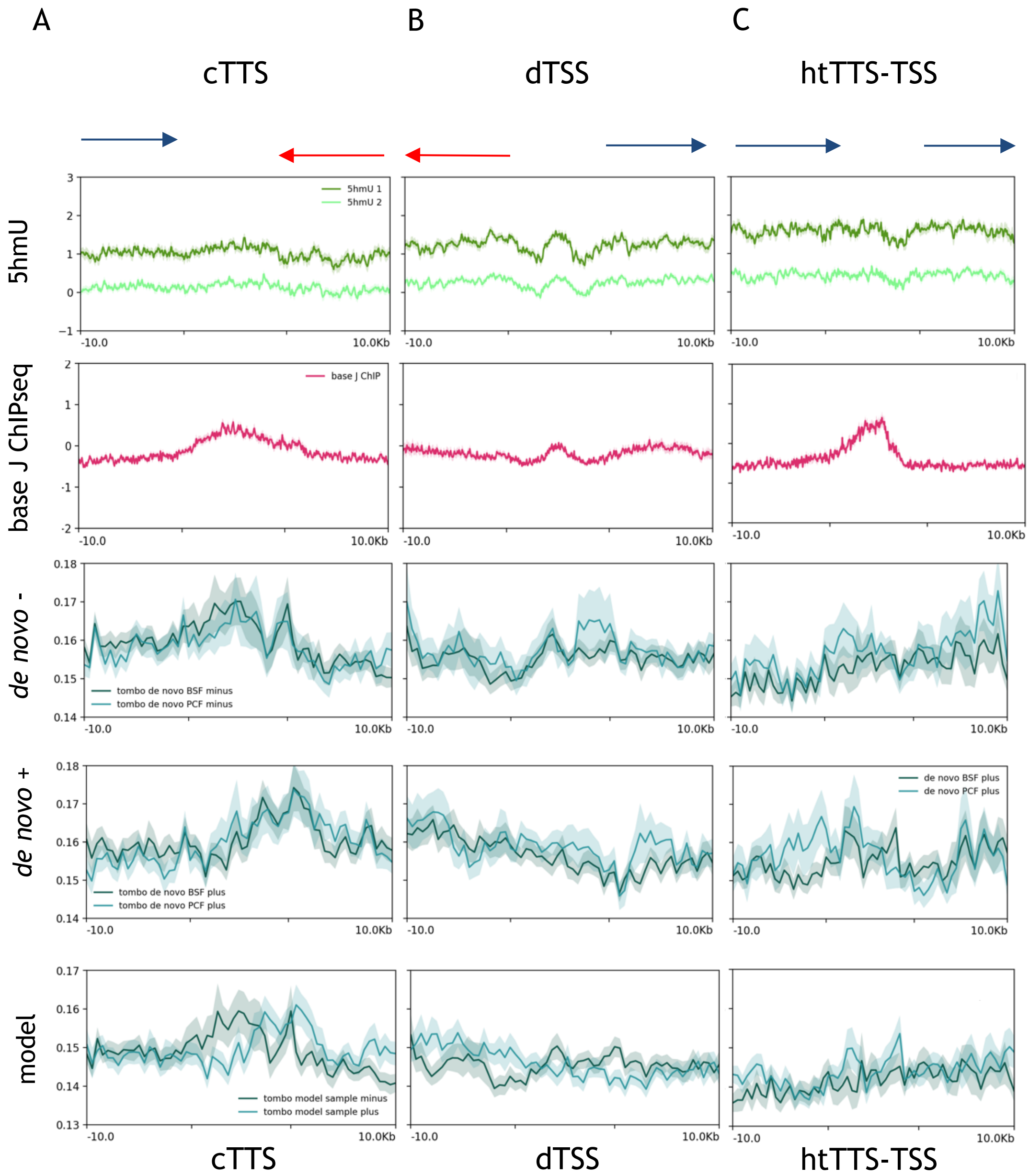


Figure 61 Detection of modified bases at polycistronic transcription start and termination sites of *T. brucei.*

Mapping of base J ChIPseq, 5hmU ChIPseq replicates and newly generated Tombo data to *T. brucei* polycistronic unit boundaries ± 10kb. A – convergent transcription termination sites (cTSS), B – divergent transcription start sites (dTSS), C – head-to-tail transcription termination and start sites (htTTS-TSS). Arrows indicate the direction of transcription. '+' or 'plus' refer to the forward (top) strand, '-' or 'minus' refer to the reverse (bottom) strand of the reference sequence. BSF – bloodstream-form cells, PCF – procyclic cells, TSS – transcription start site, TTS – transcription termination site, model and *de novo* refer to different Tombo outputs.

At convergent transcription termination sites (Figure 61 A), enrichment of base J ChIPseq signal was apparent, as was Tombo signal for *de novo* and model sample datasets. As before, Tombo signal was very similar among *de novo* BSF, *de novo* PCF and model sample modes, and it provided strand-specific detail, which revealed enrichment in modified bases further downstream of the TTS on each strand. The opposite pattern was seen at divergent transcription start sites (Figure 61 B): Tombo data suggested there is depletion in modified bases at a TSS, just upstream of the respective PTU. Accordingly, the non-strand-specific 5hmU and base J ChIPseq data signal dipped just upstream of the PTUs, consistent with a depletion in modified DNA bases at PTU TSSs. At head-to-tail PTU boundaries located on the same strand (Figure 61 C), the picture was more complex; 5hmU ChIPseq data showed a small dip in signal just upstream of the TSS and base J ChIPseq data showed an increase in signal at the TTS, as in the previous two examples of PTU boundary type, and this was consistent with Tombo data on the forward strand of the *de novo* datasets. However, the reverse strand data and model sample data suggested there is a steady increase in modified base signal from the TTS to the TSS and into the following PTU.

When examining the datasets across full-length PTUs, a slightly different picture emerged (Figure 62). In the 5hmU and base J ChIPseq data the signal was lower within the PTU than just upstream (base J) or at the TSS (5hmU), and there was also enrichment in base J level downstream of the PTU at the TTS. Tombo data across PTUs, on the other hand, offered a less clear picture with no pronounced enrichment at TSS or TTS. It is feasible that the varying nature of PTU boundaries (namely, the combination of transcription start or termination sites) introduces complexity, therefore obscuring any PTU-wide patterns. Another possibility is that a different modified base (for example, 5fU) is prevalent within polycistrons and it was detected by Tombo, but not the other methods; however, from these data it is not possible to establish what is responsible for this divergent pattern.

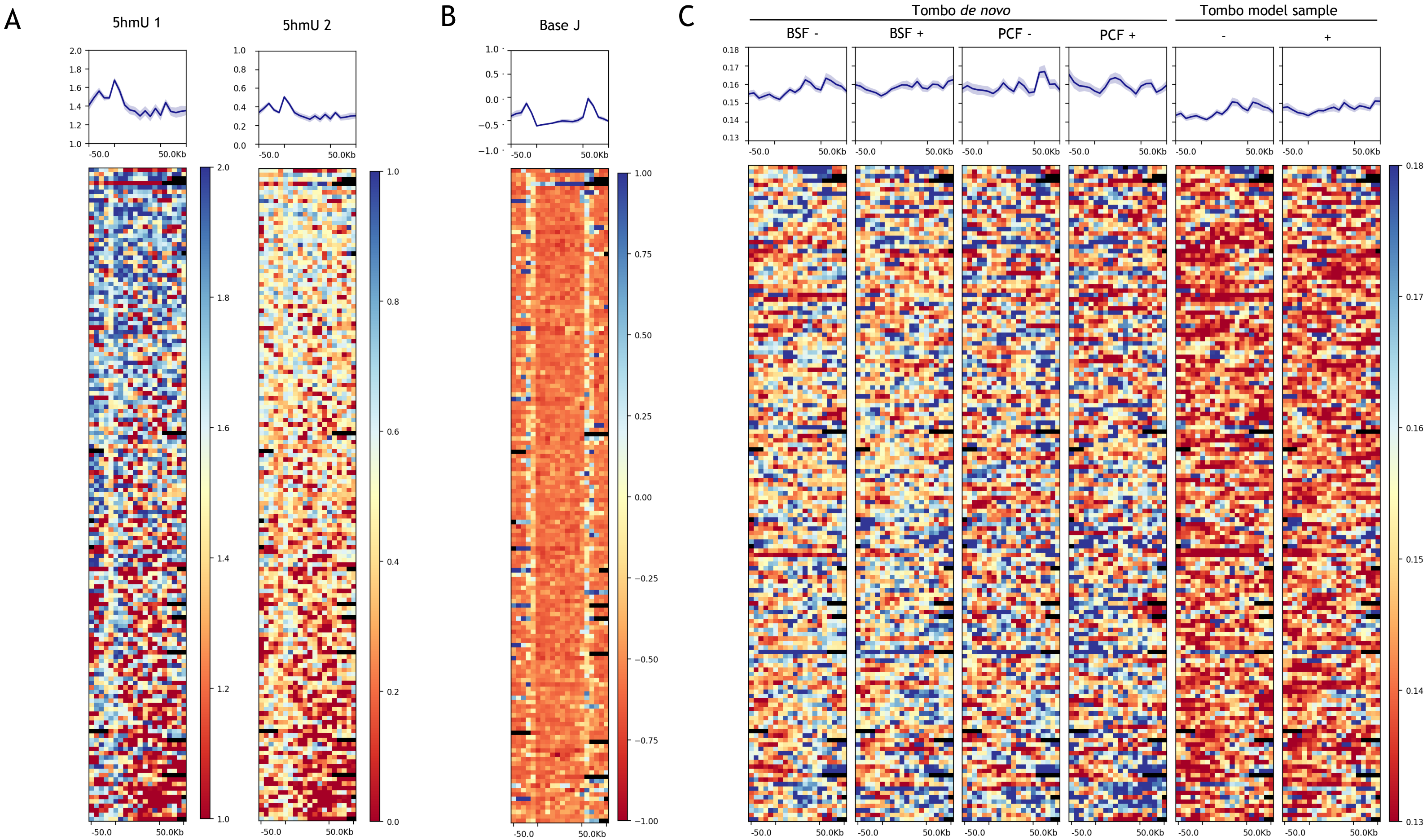


Figure 62 Mapping of base modification across polycistronic transcription units of *T. brucei*.

Heatmaps and overall profile of A - 5hmU ChIPseq, B - base J ChIPseq and C - Tombo data across putative polycistronic transcription units and ± 50 kb flanking regions in *T. brucei*. BSF – bloodstream-form cells, PCF – procyclic-form cells, '+' - forward or top strand, '-' - reverse or bottom strand.

4.2.1.3 Repetitive genomic regions

Next, we examined the other set of regions associated with base J deposition (van Leeuwen *et al.*, 1996, 2000; F. van Leeuwen *et al.*, 1998; Cliffe *et al.*, 2010): repetitive DNA, specifically - centromeric, 70bp, 50bp, 177bp and telomeric repeats. We decided to focus this part of the analysis mostly on the newly assembled Nanopore-based genome sequence (described in Chapter 3), as the repetitive regions are more complete in this assembly compared to the TriTrypDB Lister 427 reference sequence.

First, centromere-associated repeats (Figure 63) displayed very strong enrichment of base J ChIPseq signal, as well as *de novo* and model sample Tombo data for both the forward and reverse strand, in both lifecycle stages. 5hmU mapping was not quite as clear, as in one of the ChIPseq datasets (‘5hmU 2’) no clear signal enrichment or depletion was evident, whereas in the other one (‘5hmU 1’) the enrichment in signal was instead surrounding the centromeric repeats.

Similarly, across 70bp and 50bp repeats clear enrichment in base J ChIPseq and Tombo datasets was evident, whereas 5hmU was not consistent between the two replicates, although for 50bp repeats some enrichment was seen in both 5hmU datasets (Figure 64 A). Base J ChIPseq signal was also clearly enriched at 177bp repeats and telomeric repeats, and this was also the case for one of the 5hmU pulldowns (‘5hmU 2’). Tombo data, on the other hand, appeared somewhat noisy and incomplete at 177bp and telomeric repeats, and therefore difficult to interpret, although some enrichment, preceded by a dip in signal, was evident towards the end of the telomeric repeats in the *de novo* Tombo data on the forward (‘+’) strand. A clearer picture of the Tombo signal across 177bp regions can be gleaned by looking at the individual regions rather than metaplots (Figure 64 B) – these showed clear enrichment in modified base signal for *de novo* and model Tombo data as well as some localised enrichment in level sample mode Tombo data.

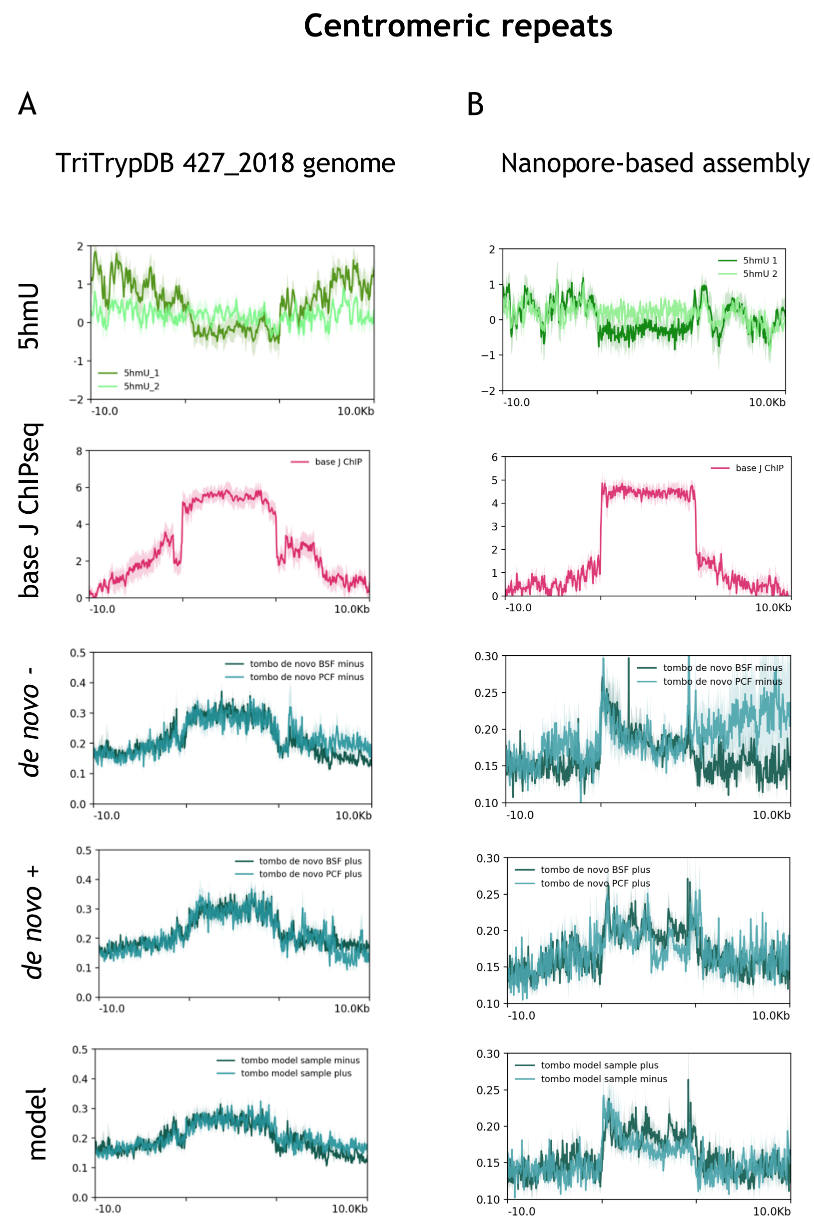


Figure 63 Mapping of 5hmU, base J and Tombo data across centromere-associated repeats.

Metaplots of 5hmU ChIPseq, base J ChIPseq and Tombo dataset mapping to centromere-associated repetitive DNA ± 10 kb flanking sequence in A - the Lister 427 2018 reference genome from TriTrypDB (Müller *et al.*, 2018), B - the Nanopore-based assembly generated in Chapter 3**.** Minus and '-' refer to the reverse or bottom strand, whereas plus and '+' – to the forward or top strand. PCF – procyclic cells, BSF – bloodstream-form cells. Model (or model sample) and *de novo* refer to different Tombo outputs.

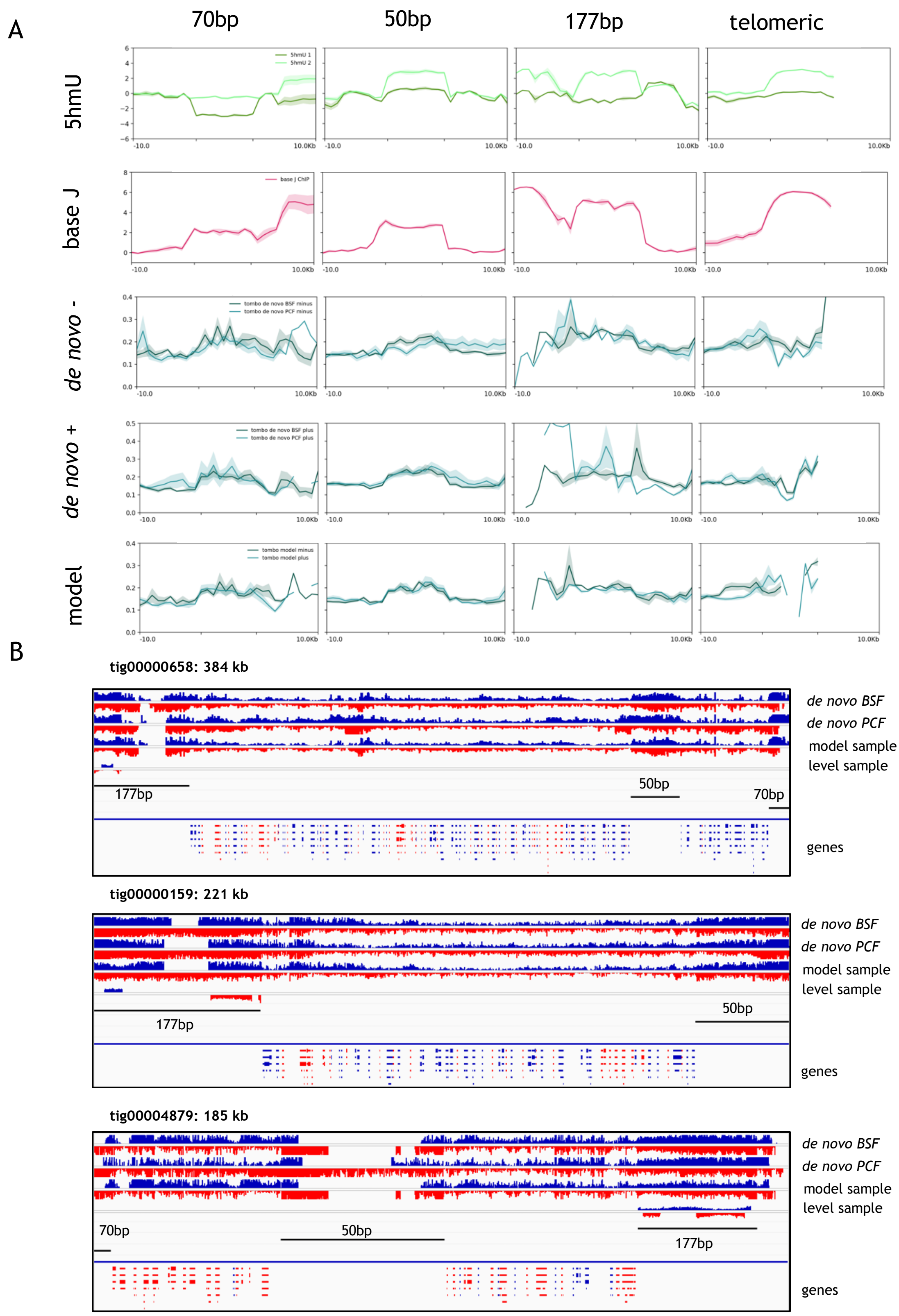


Figure 64 Mapping of 5hmU, base J and Tombo data across *T. brucei* repetitive DNA elements.

A – Metaplots of 5hmU ChIPseq, base J ChIPseq and Tombo data mapping across 70bp, 50bp, 177bp repeats and telomeric regions ± 10 kb flanking regions of the *T. brucei* genome. Telomeric repeat regions have been oriented with the chromosome arm on the left. B - Tombo dataset mapping across three *T. brucei* contigs containing the 177bp repeat regions. BSF – bloodstream-form cells, PCF – procyclic cells, '+' – forward or top strand, '-' – reverse or bottom strand. In blue – data of the forward or top stand, in red – data on the reverse or bottom strand. Level, model and *de novo* refer to different outputs of Tombo.

4.2.1.4 Modified base distribution around protein-coding sequences

Perhaps the most unexpected finding was modified base distribution at and surrounding coding sequences – a previously unexplored area. We mapped 5hmU and base J ChIPseq data (Figure 65), as well as Tombo data (Figure 66), to all gene, CDS and *VSG* annotations. First, assessing the 5hmU and base J ChIPseq data for all genes and CDS (Figure 65), elevated 5hmU signal within the gene and CDS regions was seen, as well as a corresponding dip in base J signal, followed by a small increase in base J signal downstream of the 3’ boundary (Figure 65 B). This pattern was more pronounced for *VSG* genes, and a gradual increase in 5hmU signal upstream of the gene, peaking across the gene region, followed by a sharp drop in signal, was also evident.

Unfortunately, as the ChIPseq datasets examined here did not offer strand-specific information, the distribution of ChIPseq signal between the DNA strands could not be discerned. For this, Tombo data provided potential insight (Figure 66); the broad signal pattern was consistent with the base J ChIPseq data – higher signal surrounding CDS and genes, as well as higher signal 3’ of *VSG* genes. However, for CDS and genes the mentioned pattern only held true for the strand on which the gene/CDS was located. For *VSG* genes, similar enrichment patterns, consistent with base J ChIPseq data, on both strands were seen, highlighting differential DNA modification patterns around coding sequences depending on the type of coding sequence.

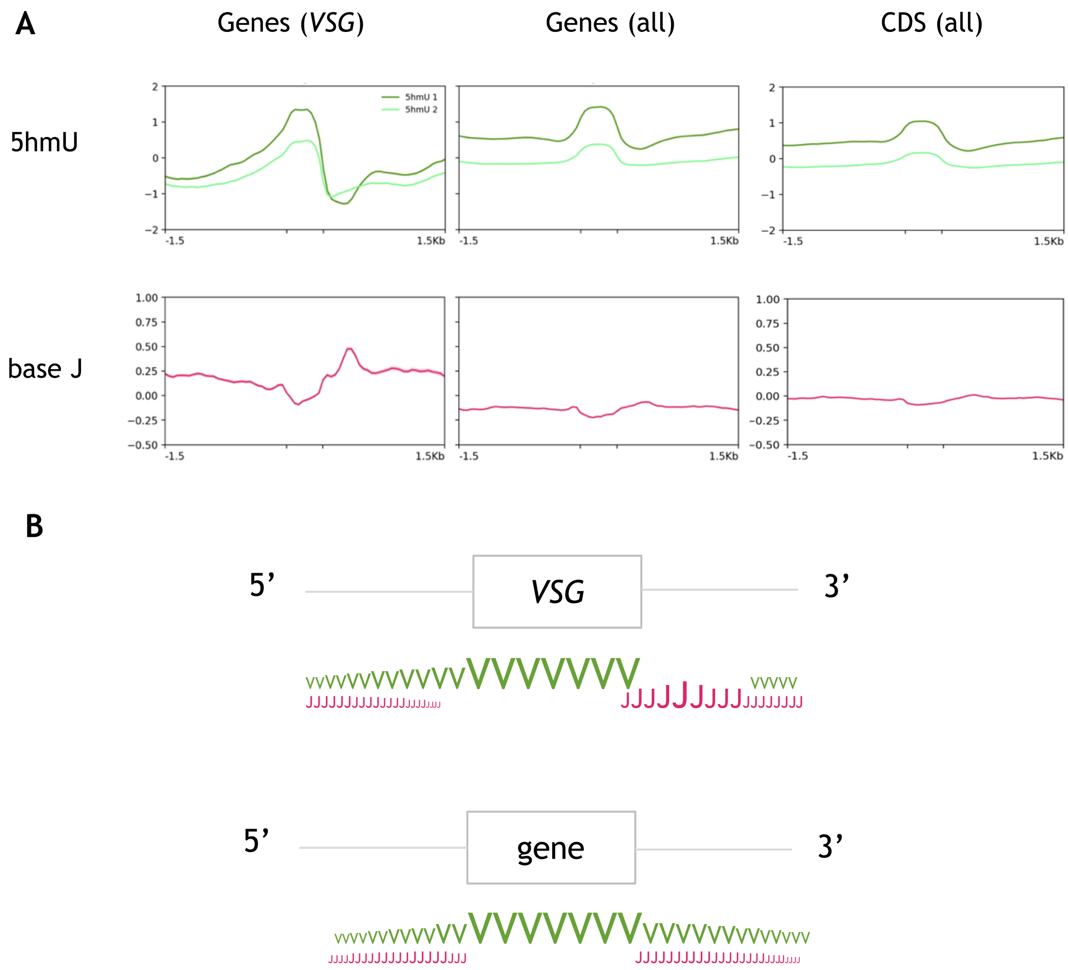


Figure 65 Base J and 5hmU enrichment patterns around coding sequences of the *T. brucei* genome.

A – 5hmU ChIPseq and base J ChIPseq datasets mapped across all annotated *VSG* genes, all non-VSG genes, or all non-VSG CDS ± 1.5 kb of flanking sequence. B - diagram summarising the modified base accumulation pattern around genes (VSG or non-VSG) based on the 5hmU and base J ChIPseq mapping results in A. Grey-bordered boxes represent the respective genes, whereas the grey lines on both sides – the flanking sequence (which includes some or all of the untranslated regions (UTRs). VSG – variant surface glycoprotein, CDS – coding sequence annotations. V – 5hmU, J – base J, increasing size of the letters indicates the higher proportion of bases modified at a given position.

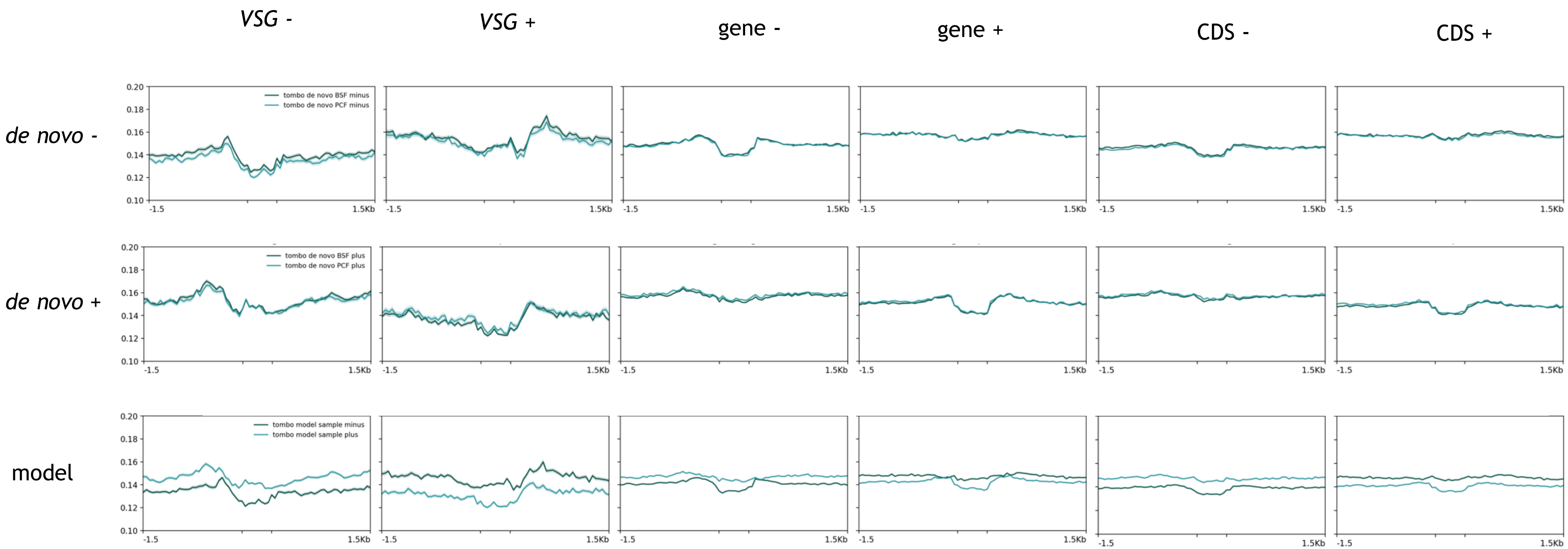


Figure 66 Strand-specific DNA base modifications as detected by Tombo around protein-coding sequences of *T. brucei*.

Tombo mapping across *VSG* genes, non-VSG genes, and non-VSG CDS of the *T. brucei* genome. For clarity, the sequences have been separated based on reference orientation, '+' for the forward or top strand, '-' for the reverse or bottom strand. Model and *de novo* refer to the Tombo output, BSF – bloodstream-form cells, PCF – procyclic cells. VSG – variant surface glycoprotein, CDS – coding sequence annotations. Minus or '-'– reverse or bottom strand data, plus or '+'– forward or top strand data.

4.2.1.5 Relationship between modified base ChIPseq data and Tombo signal

Many of the genomic regions analysed above were suggestive of similar enrichment patterns between, for example, base J ChIPseq data and Tombo data; to assess whether this holds true on a larger scale, we decided to investigate whether there is a clear linear relationship between existing ChIPseq datasets and the Tombo data presented here. In order to achieve this, correlograms were plotted for both groups of datasets, comparing enrichment values at the same loci, genome-wide (Figure 67). First, within the base J and 5hmU ChIPseq datasets (Figure 67 A), no clear linear relationship between base J and 5hmU signal levels was found genome-wide; 5hmU ChIPseq replicates ('5hmU1' and '5hmU2') showed a weak linear relationship, but it was not as pronounced as one might expect from replicate experiments.

No clear linear relationship between the Tombo and ChIPseq datasets was found (Figure 67 B). Within Tombo datasets, however, a linear relationship between *de novo* BSF and model mode data for both strands was evident (Figure 67 C), which is consistent with more localised similarities between the *de novo* BSF and model mode signal patterns above.

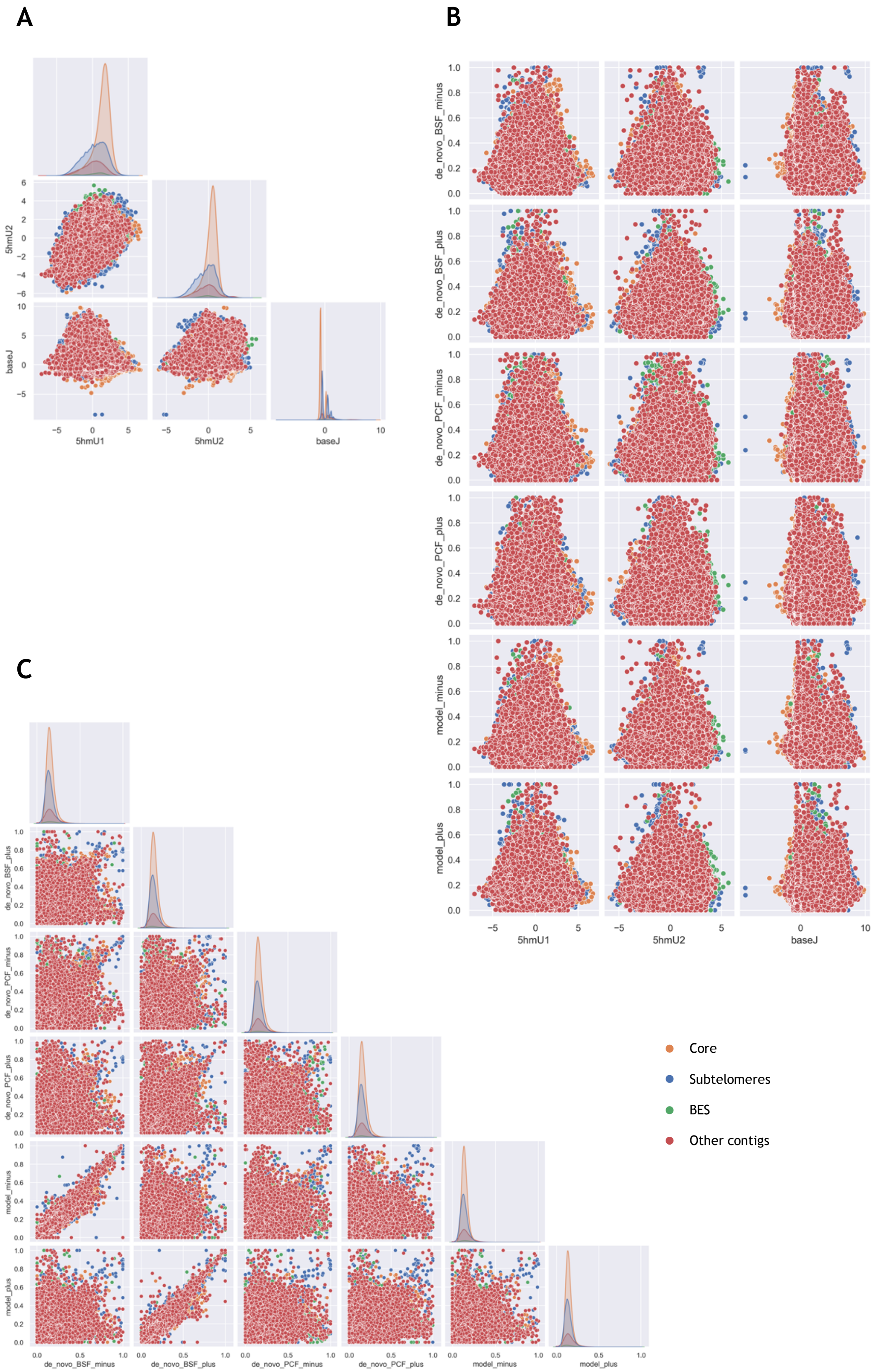


Figure 67 Correlograms assessing possible linear relationships between different methods of detecting modified bases in *T. brucei*.

A - assessing the two 5hmU ChIPseq values and base J ChIPseq, B – assessing the presence of a linear relationship between 5hmU ChIPseq and base J ChIPseq datasets with Tombo datasets, C – assessing the presence of a linear relationship among Tombo datasets.

4.2.1.6 Base-resolution mapping of Tombo signal

While Tombo data cannot provide information on the identity or nature of any base modification detected, the base resolution and strand specificity of the data can be used to assess modified signal level across and surrounding canonical bases of the reference sequence – A, T, G and C. Due to computational constraints, we limited this analysis to the core of chromosome 1 (Figure 68).

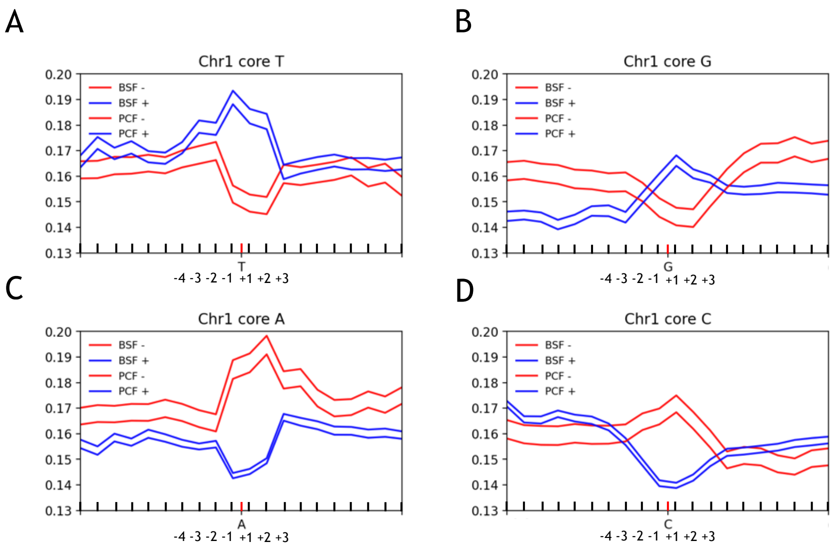


Figure 68 Nucleotide-level Tombo *de novo* mode signal surrounding A, T, G and C residues of chromosome 1 core regions in *T. brucei.*

Nucleotide-level Tombo *de novo* signal at and around A, T, G and C residues, ± 10 bp of flanking sequence, on the core of chromosome 1 of *T. brucei*. BSF – bloodstream-form cells, PCF – procyclic cells, '+' – forward or top strand data, '-' – reverse or bottom strand data.

Consistent with known base modifications of thymidine - 5hmU, base J and 5fU - being present in the *T. brucei* genome, pronounced enrichment at and around T residues on the forward strand and the reverse strand of A residues in both lifecycle stages of the parasite was observed (Figure 68 A and C). Unexpectedly, a smaller increase in signal on the forward strand of G residues and negative strand of C residues, consistent with a possible modification of G residues, was also detected.

The metaplots in Figure 68 also highlighted the fact that the modified base signal extends beyond the single nucleobase – up to 3-4 bases upstream and downstream also showed altered signal; it is possible that the apparent signal elevation seen at G residues was a result of 'carryover' signal from an adjacent thymidine residue rather than an inherent base G modification. This could occur if, for example, G residues are more likely to be found within 3-4 nucleobases of a modified T than A or C residues, or, if T positioning near a G might be more easily or more likely modified. Due to inherent non-random distribution of nucleotides throughout the genome (Chapter 5) these hypotheses would be difficult to adequately address, however.

To gain more insight into the possibility of guanine modifications in the *T. brucei* genome, we analysed the distribution of Tombo modified base signal levels among the four reference nucleobases (Figure 69) as well as the frequency of nucleotides surrounding T and G residues of the core of chromosome 1 (Figure 70).

Compared to A and C residues, G and T displayed higher frequency of elevated Tombo signal (Figure 69), suggesting greater modification rates for these nucleotides. Looking at nucleotide frequencies surrounding T and G residues of varying Tombo signal levels, it became evident that with increasing T modification, the frequency of G residues immediately flanking the residue also increased (Figure 70 A). In addition, more modified T's were found in T-rich sequences.

Focusing on apparently modified G residues (Figure 70 B), with increased G modification, an increase in the frequency of flanking T residues was also observed. Furthermore, when we isolated G residues with a modification fraction of 0.30 or over, 62.7% (21734/34658) of these were immediately adjacent to a T with a modification fraction of 0.30 or higher, and 78.6% (27231/34658) were found within ± 6bp of such T.

Taking these findings together, it is likely that the apparent high rates of G modification seen in our Tombo data was indeed a result of proximity to modified T bases, rather than a true modification of G residues, although it cannot be excluded from these data.

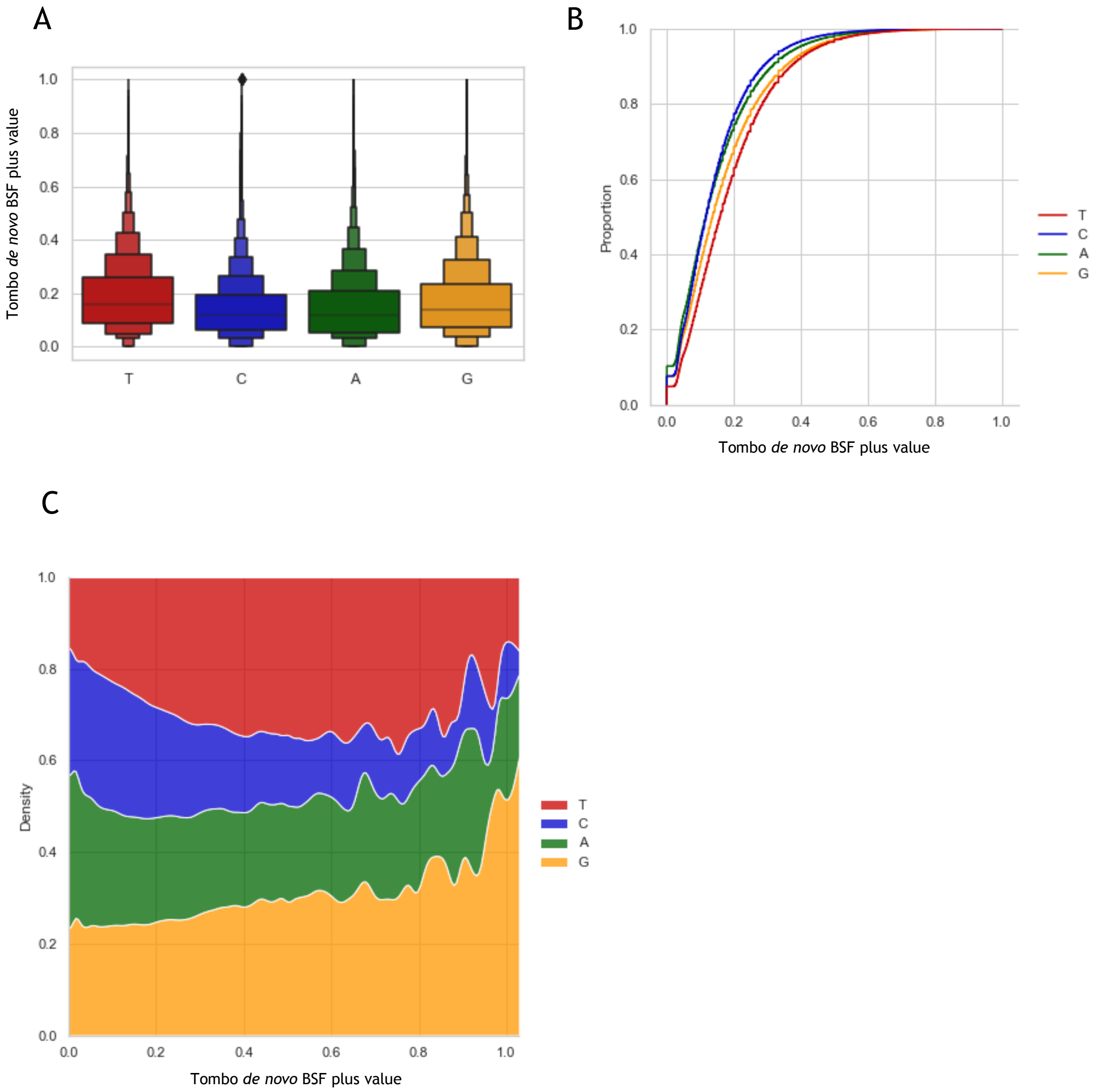


Figure 69 Distribution of *de novo* BSF Tombo signal among A, T, C, G residues in the core of chromosome 1 in *T. brucei*.

A - boxenplots describing the distribution of Tombo *de novo* BSF values on the forward strand for each of the four canonical reference bases, B - the cumulative distribution of *de novo* Tombo values on the positive strand in the BSF cells, C - kernel density distribution of *de novo* Tombo values on the positive strand in the BSF cells. BSF – bloodstream-form cells.

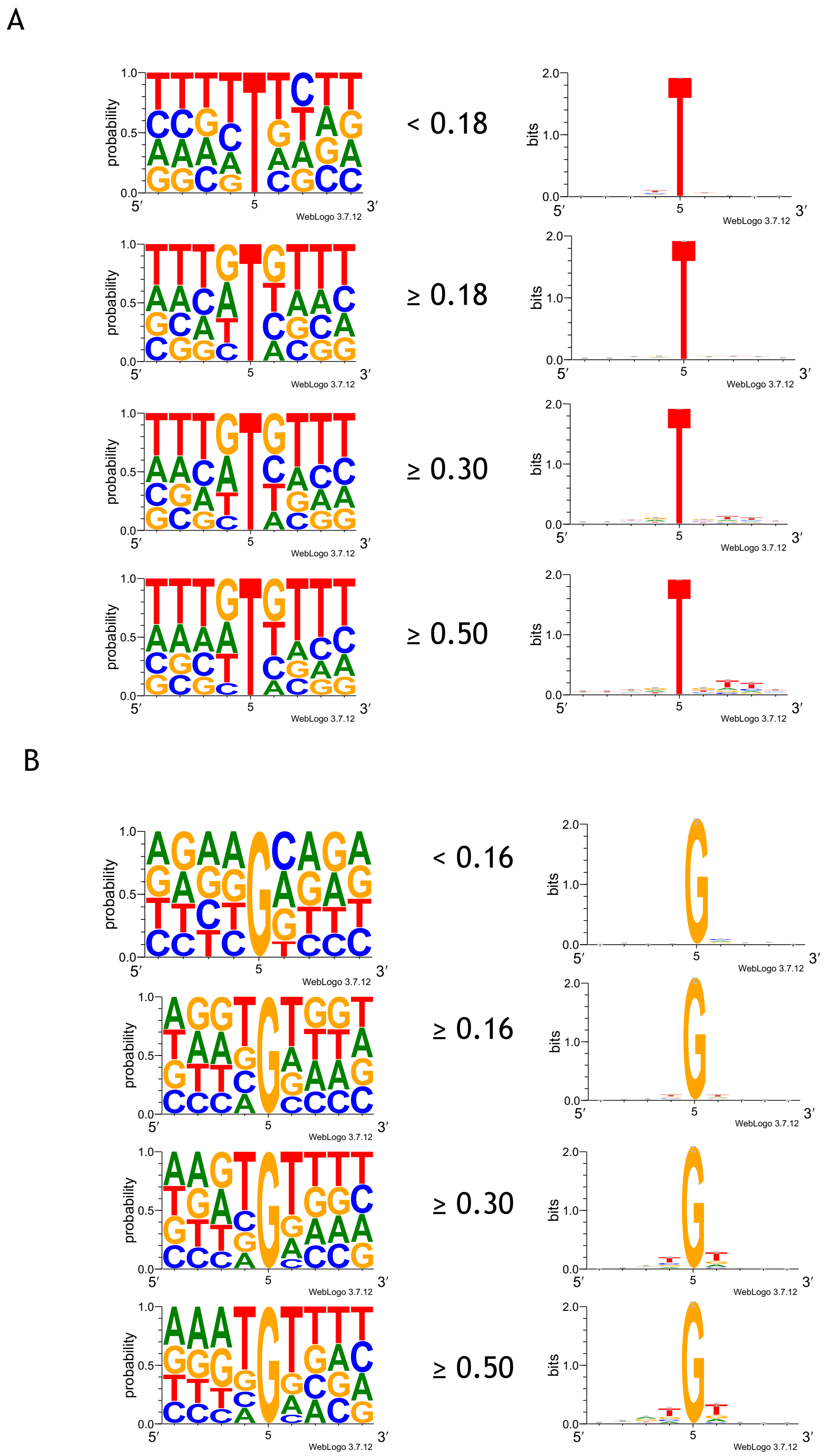


Figure 70 Sequence composition surrounding highly modified T and G residues of chromosome 1 based on Tombo *de novo* values in BSF cells (forward strand only).

Sequence composition around residues with varying levels of modification as detected by Tombo in bloodstream-form cells on the forward strand, A – around T residues, B – around G residues. On the lefthand side, panels showing the frequencies of A, T, C and G nucleotides surrounding T residues with varying levels of modification, as indicated in the middle. On the right-hand panels, the same positions are shown, except instead of surrounding nucleotide frequencies the sequence conservation in bits is shown instead. Figures generated using WebLogo (Crooks *et al.*, 2004).

4.2.1.7 Sequence composition around modified sites

To further assess the sequence composition surrounding apparently modified residues, AT and GC nucleotide skews were plotted around T and G residues with various Tombo signal values (Figure 71). Briefly, AT and GC skews highlight local DNA composition deviations from A=T and G=C 'equilibrium'; higher AT skew values represent higher number of A residues compared to T, and higher GC skew values are indicative of overabundance of G residues over C. With increasing Tombo signal at T residues, a decrease in AT and an increase in GC skews was seen, mostly downstream of the modified site. A similar pattern was seen for apparently modified G residues (as determined by Tombo); notably, the group of G residues with the lowest Tombo value showed a much lower GC skew than the other groups, or, indeed, T.

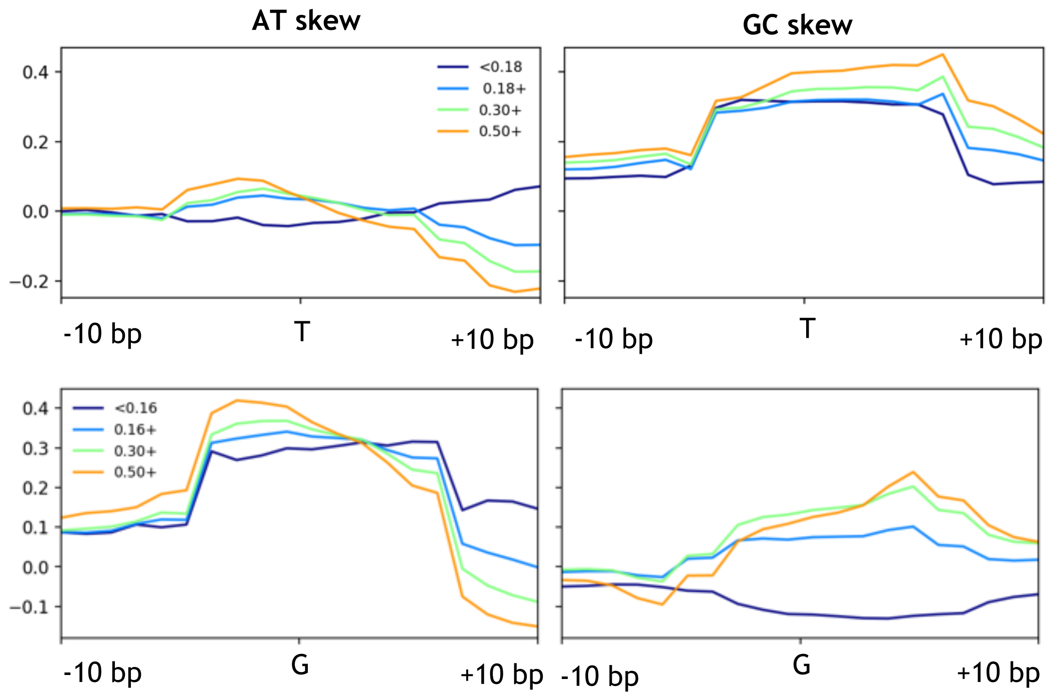


Figure 71 Nucleotide skews around T and G residues with varying modification level, as determined by Tombo.

Nucleotide skews surrounding T (top panels) and G (bottom panels) residues showing varying Tombo signal level on the forward strand using *de novo* mode in bloodstream-form cells (± 10bp flanking regions plotted). Nucleotide skews were calculated in 10 bp bins across the genome.

4.2.2 Localised modified base interrogations

4.2.2.1 Active and silent bloodstream-form expression sites

In addition to genome-wide patterns of base J deposition, a few reports have mentioned more localised examples of base J enrichment (Leeuwen *et al.*, 1997; Reynolds *et al.*, 2014, 2016). First, several loci in actively transcribed and silent bloodstream-form expression sites (BES) in bloodstream-form (BSF) *T. brucei* were examined by Leeuwen *et al.*, (1997). In their targeted anti-J immunoprecipitation experiments the actively transcribed VSG was not bound by anti-J antibodies, whereas the silenced ones were, suggesting differential modification of BES-resident *VSGs* depending on their transcriptional activity. Additionally, 70bp repeats in the active BES, specifically, did not appear to be modified (Leeuwen *et al.*, 1997).

Here, upon examination of the BES-resident *VSGs* and 70bp repeat regions (Figure 72 and Figure 73), base J ChIPseq data clearly showed differential base J deposition between *VSGs* in active versus silent expression sites, with the active BES *VSG* apparently depleted of base J. 5hmU and Tombo data, on the other hand, did not show striking differences in modified base deposition across BES *VSGs,* although *de novo* BSF data of the forward (+) strand was suggestive of a small effect. Across the 70bp repeat regions just upstream of the BES VSG, no clear difference was evident in any of the datasets analysed here (Figure 73), although, as above, perhaps a small effect could be discerned in the Tombo *de novo* BSF data on the forward strand.

4.2.2.2 Genes at the end of some polycistronic transcription units

In *Leishmania tarentolae* and *Leishmania major*, depletion of base J leads to read-through transcription at PTU termination sites (van Luenen *et al.*, 2012; Reynolds *et al.*, 2014); while in *T. brucei* loss of base J does not seem to cause this effect, it does appear to result in increased transcription of genes downstream of base J peaks towards the end of select PTUs (Reynolds *et al.*, 2014). To examine whether the reported loci show signs of base J enrichment in the datasets examined here, we plotted two such loci, picked at random, alongside the original published figures (Figure 74).

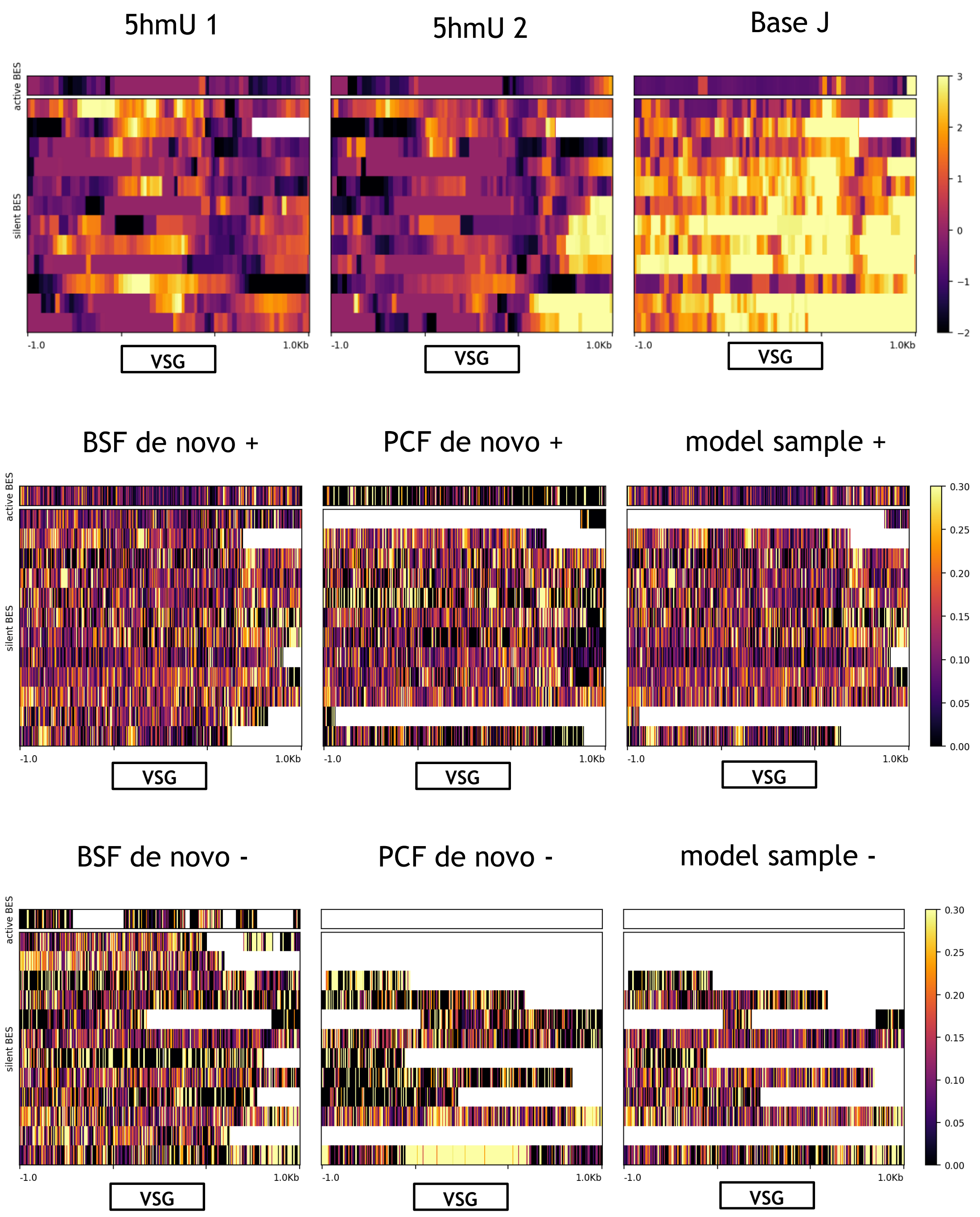


Figure 72 Modified DNA bases spanning *VSGs* in the active vs silent bloodstream-form expression sites (BES).

5hmU ChIPseq, base J ChIPseq and Tombo data mapping across bloodstream-form expression site (BES)*VSGs* (± 1kb flanking sequence). Active expression site – BES1. Missing data (due to insufficient coverage) indicated in white.

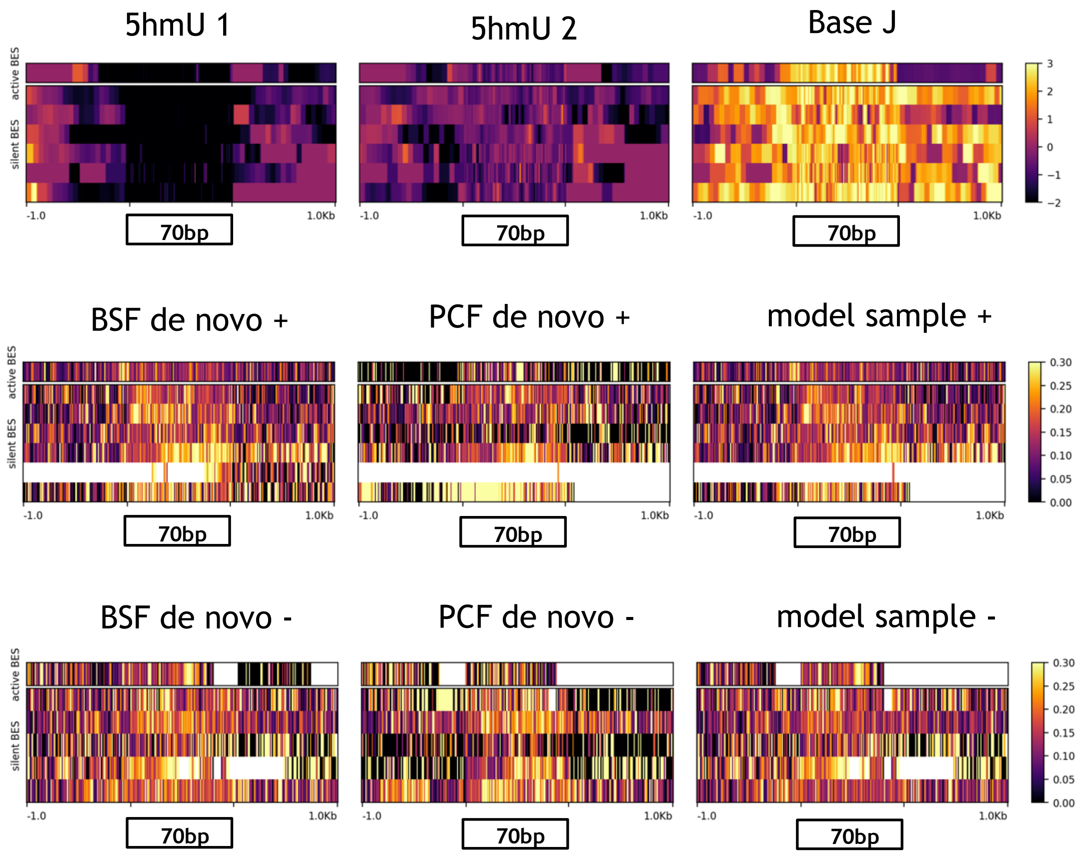


Figure 73 Modified base mapping across 70bp repeats in active versus silent bloodstream-form expression sites.

5hmU ChIPseq, base J ChIPseq and Tombo data mapping across bloodstream-form expression site (BES) 70bp repeat regions (± 1kb flanking sequence). Active expression site – BES1. Missing data due to insufficient coverage indicated in white.

First, we analysed a putative convergent TTS-TTS site (cSSR) found on chromosome 5 (Figure 74 A). In the original figure, a base J peak can be seen between genes 1 and 2; in the datasets analysed here, enrichment in base J ChIPseq data just upstream and on gene 2 was evident, but no clear signal enrichment was seen in the Tombo datasets. In the second example, another putative cSSR on chromosome 8, a sharp base J peak is evident in the original data just upstream of the last gene of the PTU on the forward strand (Figure 74 B). Here, no such clear corresponding base J ChIPseq enrichment was detected, but a corresponding increase in signal in the Tombo data on both strands in the *de novo* PCF data and the reverse strand of BSF were evident.

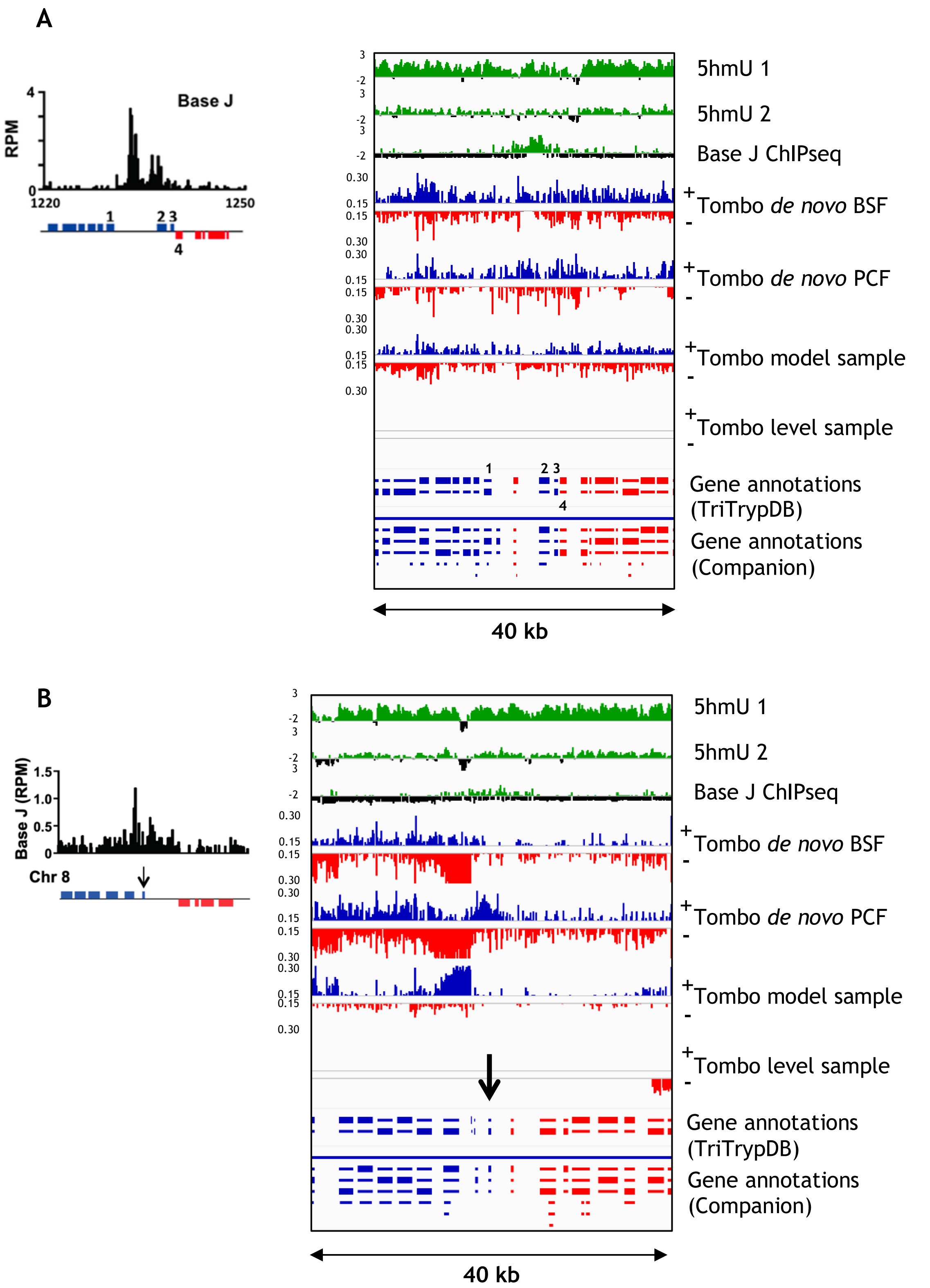


Figure 74 Modified base distribution at two PTU boundaries previously reported to have high levels of base J.

A – A convergent polycistronic unit boundary on chromosome 5, B - Another convergent polycistronic unit boundary on chromosome 8. The arrow on the right-hand side panel in B indicates the same region as on the left-hand side panel. '+' – forward strand, '-' – reverse strand, BSF – bloodstream-form cells, PCF – procyclic cells. The left-hand panels in both A and B were taken from Reynolds *et al.*, (2014) under Open Access, Creative Commons CC BY license.

4.3 Discussion

The genome of *Trypanosoma brucei* is known to contain three modified bases - ß-D-glucosyl-hydroxymethyluracil, also known as base J, its precursor 5-hydroxymethyluracil (sometimes called V), and 5-formyluracil (Gommers-Ampt, Lutgerink and Borst, 1991; Gommers-Ampt *et al.*, 1993; Bullard *et al.*, 2014). Quantitative mass spectrometry (Bullard *et al.*, 2014; Liu *et al.*, 2014) and 2D thin layer chromatography experiments (Gommers-Ampt, Lutgerink and Borst, 1991; Fred van Leeuwen *et al.*, 1998; F. van Leeuwen *et al.*, 1998) aimed at interrogating the nucleotide composition of *T. brucei* DNA, have not indicated the presence of other modified DNA bases. Early work with targeted probe-based methods showed that base J is present in tandemly repeated DNA sequences, such as telomeric TTAGGG, 50bp, 70bp and 177bp repeats, as well as silent BES (van Leeuwen *et al.*, 1996, 2000; Leeuwen *et al.*, 1997; Fred van Leeuwen *et al.*, 1998). Later, more global interrogations using anti-J antibodies indicated base J is also present at PTU boundaries and subtelomeric regions (van Leeuwen *et al.*, 1996, 2000; Leeuwen *et al.*, 1997; Fred van Leeuwen, Taylor, *et al.*, 1998). The genome-wide distribution of the putative precursor for base J, 5hmU, which was mapped using an enzyme-mediated bioorthogonal labeling method, has not been extensively analysed (Ma *et al.*, 2021); overall, it appears that 5hmU enrichment is mostly seen within genes and intergenic sequences. Despite the publication of the above-mentioned datasets, they have, in our view, remained incompletely analysed. Nothing is known about the distribution of 5fU in the *T. brucei* genome.

Oxford Nanopore Technologies provide a relatively novel way to analyse DNA and RNA modifications. Because the technology allows sequencing of native (non-amplified) nucleic acids, it provides the ability to detect the presence and position of nucleotide modifications during downstream analysis (Searle *et al.*, 2023). Several tools have been developed for this purpose, many focusing on detecting specific and common modifications in raw ONT data, such as 5-methylcytosine (5mC) and N6-methyldeoxyadenosine (6mA) (White and Hesselberth, 2022). To our knowledge, no current tools are able to detect base J, 5hmU or 5fU specifically. The tool Tombo, on the other hand, offers *de novo* identification of DNA modifications in ONT data, without specifying a particular type of modification (Oxford Nanopore Technologies, 2018). Here, we used Tombo, alongside the published ChIPseq datasets for base J and 5hmU, to gain a better understanding of the distribution of DNA modifications in *Trypanosoma brucei*.

4.3.1.1 Repetitive DNA sequences, PTU boundaries and T residues

At specific genomic regions, we found that Tombo data in *de novo* and model sample modes is consistent with what has been published in relation to base J genome-wide distribution: enrichment in repetitive sequences, at polycistronic transcription unit boundaries, and at T residues. The third Tombo mode – level sample – rarely correlated with the other modes or, indeed, the ChIPseq datasets, therefore we conclude that it is likely an inappropriate route towards detecting base modifications in *Trypanosoma brucei*.

Centromeric, 177bp, 70bp and 50bp repetitive elements all showed enrichment in base J ChIPseq data and Tombo data (*de novo* and model sample mode). 5hmU ChIPseq data was also enriched at 50bp, 177bp and telomeric repeats (1/2 replicates only) and showed conflicting signal at centromeric and 70 bp repeats, with one of the replicates suggesting depletion of signal, and another showing signal consistent with that of the flanking sequences. Centromeric base J localisation has not to date been described, but reanalysis of the ChIPseq data using a Nanopore genome assembly and Tombo data indicate that the repeats of this genome feature are, like all other *T. brucei* repeats analysed, a site of J and perhaps 5hmU accumulation.

At PTU boundaries, base J ChIPseq data showed some level of enrichment at all three main types of boundaries – convergent, divergent and head-to-tail PTU boundaries. Similarly, 5hmU ChIPseq data showed enrichment at divergent strand switch regions, but not as clearly at the other boundary types. Tombo data showed clear enrichment at convergent TTS-TTS SSRs (*de novo* and model sample modes) with strand-specific effects: an apparent peak further downstream of the TTS on the coding strand.

We also noted strand-specific signal patterns at divergent TSS-TSS sites in Tombo data, with the lowest signal level right at the TSS on the corresponding coding strand. At head-to-tail TTS-TSS the picture was a bit different: as the region is a combination of transcription termination and initiation, it shows signs of both cTTS and dTSS - a signal peak at the TTS side and a dip in signal at the TSS on the coding strand. Taking these data together, it appears that transcription initiation sites are J-depleted, whereas termination sites are J enriched; the signal pattern of the base J ChIPseq dataset appears to mirror the combined amount of enrichment seen in Tombo data on both strands (Figure 75).

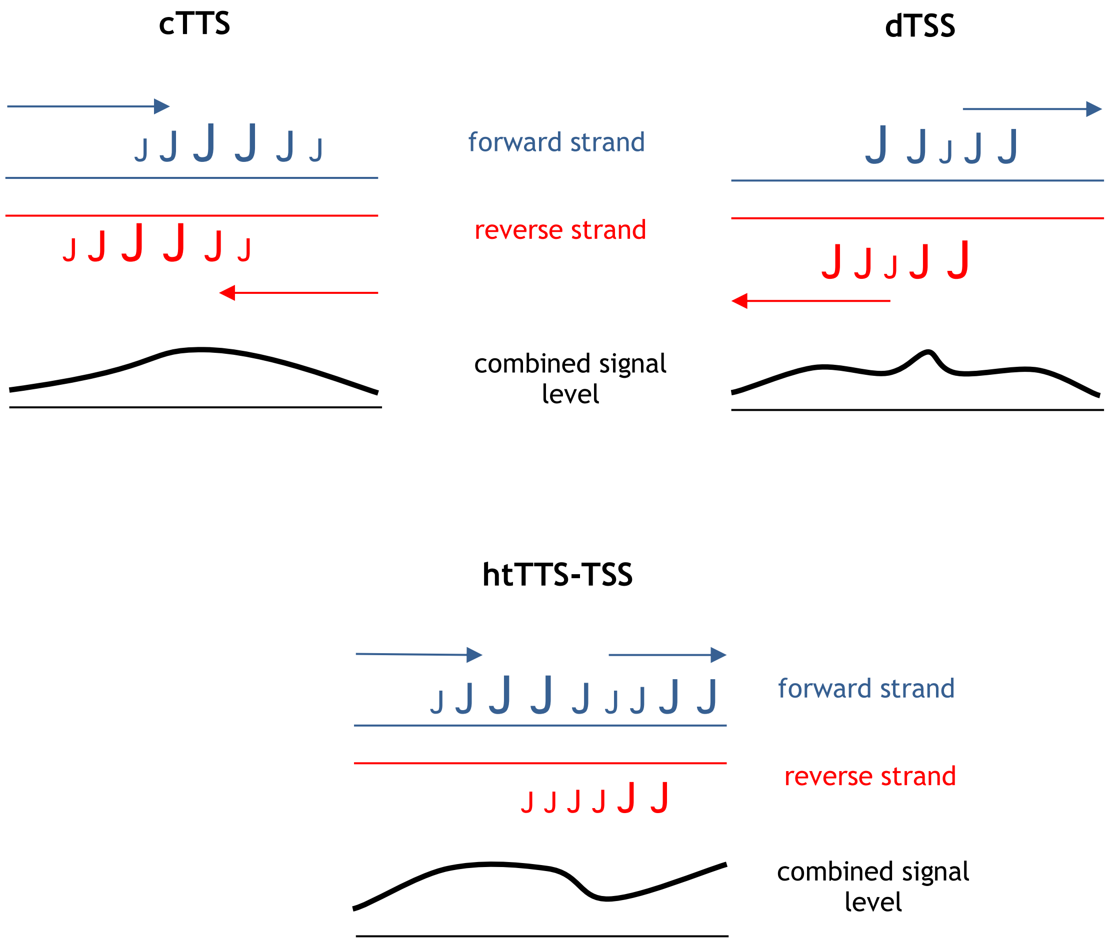


Figure 75 Putative Tombo base J deposition patterns between two DNA strands mirrors non-stranded base J ChIPseq data at different PTU boundary types.

A diagram highlighting our observation that the combined Tombo *de novo* mode signal of both DNA strands corresponds to base J ChIPseq enrichment at polycistronic transcription unit boundaries in *T. brucei*. Arrows indicate the direction of transctioption, cTTS – convergent transcription termination site, dTSS – divergent transcription start site, htTTS-TSS – head-to-tail transcription termination and start site, J – base J, with larger letters representing higher fraction of modified T residues at a given position.

Lastly, when looking at nucleotide-level data, albeit just for the core region of chromosome 1, T residues showed the highest modification levels on the forward strand, as well as A residues on the reverse, which is consistent with a modification at a T residue, such as 5hmU or J.

More globally, however, looking at chromosome-wide patterns, or, indeed, directly correlating the enrichment of base J ChIPseq, 5hmU ChIPseq and Tombo data across the genome (Figure 59, Figure 67), the agreement between these data fades – while base J ChIPseq data mostly presents as discrete peaks at PTUs and repetitive regions, Tombo data is more dispersed and variable (further discussed in 4.3.2).

Taking the above observations together, and the fact that 5hmU is present in the genome at a much lower frequency than base J based on previous work (see Table 14), we propose that the observed signal in Tombo data is likely most reflective of base J deposition in bloodstream-form parasites. It seems possible, therefore, that Nanopore sequencing could be used in organisms not so far explored for the presence of base J for its detection and analysis of its distribution. In other words, Nanopore sequencing could provide a combined *de novo* genome assembly and analysis of base J, bypassing the need for ChIPseq or other methods of base J detection.

4.3.1.2 Broad similarity in Tombo signal between procyclic-form and bloodstream-form *T. brucei*

Immunoprecipitation and 2D thin layer chromatography experiments performed in PCF and BSF *T. brucei* suggested that procyclic-form cells have very low or undetectable levels of both 5hmU and base J (Table 14). An unexpected finding with Tombo data is that, in *de novo* mode, procyclic-form and bloodstream-form *T. brucei* data are broadly consistent: PTU boundaries, repetitive sequences, coding sequences and flanking regions, nucleotide resolution level enrichment all show are very similar predicted signal between the two lifecycle stages both in amplitude (fraction of modified bases) and pattern. Two separate aspects are worth considering here – the technical limitations of the technology, and the potential biological implications of the data.

First, the sequencing technology and Tombo. Looking at nucleotide level resolution data, most bases showed some level of modification according to

Tombo; less than 10% of all nucleotides assessed here showed no indication of modification, and between 60 and 80% residues showed up to 20% modification. The median modification level is between 11.5% and 15.9%, depending on the nucleotide (A, T, C or G); this seems unlikely to be representative of the underlying biology (see below).

In ONT sequencing, signal from a DNA fragment going through the pore is captured for 6bp at a time, rather than at a single nucleotide level (White and Hesselberth, 2022); in the Tombo data, this is reflected in the nucleotide-level signal, as signal is elevated not only at a T residue, for example, but within a ± 3-4 bp flanking area. In this context, the adjacent nucleotides have non-zero values, and this may explain a large proportion of the observed non-zero values. It is also feasible that there is inherent noise in the data, and that apparent 'baseline modification' values are ubiquitous. In addition, whether or not the 'fraction modified' measure is biologically meaningful or represents a value that can be compared between samples is unclear. Nonetheless, the apparent overestimation of modified base levels should be considered if Nanopore sequencing was used for predicting such bases in an unsequenced genome.

In terms of trypanosome genome biology, the similarity of signal between the two lifecycle stages is intriguing. While previous studies have either failed to identify any base J or 5hmU in procyclic-form parasites or detected lower levels of these modifications compared to BSF parasites, it is possible that the true modification levels fell below the detection threshold of the methods used.

For a long time, it has been thought that PCF trypanosomes lack base J (Borst and Sabatini, 2008). While we cannot determine the chemical nature of DNA modifications detected by Tombo, the signal correspondence to the base J ChIPseq in BSF cells is suggestive of a couple of possibilities. First, previous work has shown that JGT - the protein that is responsible for converting 5hmU to base J - is specific to 5hmU (Bullard, Kieft and Sabatini, 2017) and can perform this conversion in a sequence context-independent manner (Bullard *et al.*, 2015). It has been hypothesised, therefore, that most deposited 5hmU residues become converted to base J in bloodstream-form cells (Bullard *et al.*, 2015); if we assume that PCF cells do not harbour base J at any meaningful level, it is possible that in the procyclic cells 5hmU are deposited at specific loci, and many of these loci then become converted to base J in BSF cells. In this scenario, modified base deposition patterns would be similar in PCF and BSF cells, even though the chemical modifications are different: 5hmU in PCF cells, both base J and 5hmU in BSF cells. Another possibility is that both lifecycle stages harbour both modifications and with a similar broad genome-wide pattern, but at lower frequencies in PCF and the output metric from Tombo is not necessarily comparable between samples. From the data available to us we cannot make any conclusions regarding this.

In *Leishmania major*, a related kinetoplastid parasite that also carries base J, biotinylation-based approaches have been used to map 5hmU and base J; in this parasite, over half of the detected enrichment peaks (112/211 peaks, 199/301 kb) overlapped between the two modified bases (Kawasaki *et al.*, 2017). Interestingly, despite the high level of overlap, 5hmU was found to be associated with specific sequence contexts (namely, T-rich), whereas base J did not; the authors suggested that 5hmU and base J have distinct localisations and that 5hmU residues are not necessarily converted to base J. For both *T. brucei* and *L. major*, it appears that 5hmU and base J localisation in the genome is non-random and partially overlapping (Cliffe *et al.*, 2009; Kawasaki *et al.*, 2017; Ma *et al.*, 2021). We could then speculate that there is some mechanism by which certain residues become base J, while others remain 5hmU. Furthermore, nothing is known about the genomic distribution of 5fU modifications in *T. brucei*; one possibility is that in procyclic parasites the loci that become base J during in BSF cells are modified from 5hmU to form 5fU in the insect-stage cells, or during differentiation, and the combined pattern of 5hmU and 5fU deposition in procyclic cell Tombo data is what gives rise to the very similar pattern between the two lifecycle stages.

4.3.1.3 Novel insights around protein coding sequences

Another curious finding is the profile of signal within and surrounding annotated genes and coding sequences, which appear to be enriched in 5hmU and slightly depleted in base J ChIPseq datasets. Tombo data highlighted the fact that modified base deposition is strand-specific around coding sequences, only showing up on the same strand as the gene/CDS. Interestingly, there is a distinct and much more pronounced pattern at and surrounding *VSG* genes, and, unlike at non-VSG genes, it is evident on both strands of DNA. Tombo data matches base J ChIPseq enrichment pattern at these loci, with a pronounced peak just 3' of *VSG* genes. While from these data we cannot conclude what roles *VSG*-adjacent base J may play, it is feasible that base J deposition around *VSG* genes acts to prevent spurious transcription at non-transcribed *VSGs*. This might be achieved through various mechanisms: for example, by directly impeding RNA polymerase progression, or by playing a role in maintaining an epigenetic landscape consistent with lack of transcription. It is also possible that a modified base like base J can impede sequence-specific factor binding, thereby hindering processes that rely on binding DNA at specific loci.

4.3.1.4 Potential G modification

In addition to putative DNA modifications centred at T residues, we have also noted an increased signal level in Tombo data at G residues. Whether this finding represents a genuine G modification in the *T. brucei* genome is uncertain – to our knowledge, there have been no indications of DNA modifications beyond base J, 5hmU and 5fU in this parasite, all of which are modified versions of thymidine. The more likely explanation for this observation, in our view, is the carry-over of signal from adjacent T residues, a result of the technical limitation of the ONT nanopores, as signal is captured in hexamers rather than at single nucleotide resolution. However, without further experimental investigation, the possibility of a previously unknown G modification in *T. brucei* is worth exploring in our view.

4.3.2 Limitations

There are several aspects of the datasets and analyses performed here that must be considered when interpreting the results discussed above.

4.3.2.1 Cell lines, strains, and experimental conditions

First, the datasets analysed here – 5hmU ChIPseq, base J ChIPseq and Tombo data – have been obtained from different strains and cell lines, as well as experimental conditions (Table 15). Whether modified base deposition varies by strain is unknown, as is whether a murine infection might impact this process. *T. brucei* subtelomeres, in particular, show significant sequence/arrangement variation by strain (Callejas *et al.*, 2006; Müller *et al.*, 2018), so these genomic compartments might be particularly problematic to assess. While the base J ChIPseq dataset was obtained from the same strain as Tombo data, the sequencing reads were very short and lacked an input sample for optimal normalisation.

In a hypothetical future experiment, these data can be improved by using the same cell lines and applying the same experimental conditions (e.g. only post-infection parasites) and sequencing methods with appropriate input samples; this would reduce the number of variables needing consideration during data interpretation.

For instance, performing base J, 5hmU and 5fU ChIPseq experiments, in combination with a panel of ChIPseq experiments for other known nucleotide modifications present in eukaryotes, along with quantitative mass spectrometry in PCF and BSF cell lines from the same strain, would greatly clarify the presence, distribution and quantification of modified bases in this parasite. Additionally, as ONT sequencing and basecalling are improving, with one basecaller, dorado (https://github.com/nanoporetech/dorado), now providing base modification calling for four modified DNA bases (4mC, 5mC, 5hmC and 6mA) and three modified RNA bases (m5C, m6A, pseU), it is likely that base-level and strand-specific modified base calling using third generation sequencing will soon provide more robust base modification detection (e.g. by calling 5hmU specifically, or calling base modifications for several different, though still unknown, modified bases).

4.3.2.2 Limitations of Tombo software and current data

Tombo detects deviations from canonical base (A, C, G, T) signal in raw Nanopore sequencing data. Based on the tool's documentation, in *de novo* and model sample modes it relies predominantly on comparison of signal between an ONT read and the expected signal at the reference position it maps to, and any deviation of signal between the reference and ONT sequences is reported as a modification. This is problematic, as in these modes variants would also be considered modifications. Level sample mode, on the other hand, compares the signal between two samples (in this case, BSF and PCF cells) at a given reference position; while this mitigates any issues with variants relative to the reference genome, it does not take into account any variants that may exist between the two cell lines, as they have been cultured separately for decades and are thus likely no longer genetically identical.

One way to mitigate this would be using a version of the reference genome that has been corrected using high quality sequencing reads from the same cell line as the experimental data. However, in a complex genome such as *T. brucei*'s this might not be a trivial task. While SNPs might be easy to 'correct' in the reference genome, larger rearrangements or variants may be problematic to resolve correctly. Alternatively, whole genome amplification may be used as a control sample: amplification would remove any modified bases from the DNA sample. This approach would offer a different set of challenges, however, as PCR artefacts may become introduced, and the resulting DNA fragments would likely be shorter compared to native DNA used as input for ONT sequencing, thus complicating the analysis of non-unique genomic regions.

Mapping to the long read assembly (described in Chapter 3) generated from the same cell line as was used for Tombo showed similar results to mapping to the reference, suggesting variant-related false positive modifications are unlikely to have a dramatic effect on modified base detection, though it should be noted that not all genomic regions have been compared due to limited annotation of the long read assembly.

In addition to the complexity introduced by genetic variants, Tombo is also unable to classify or distinguish between modified bases. We are aware of three DNA modifications present in the *T. brucei* genome, one of which hadn't been discovered before the work done by Bullard *et al.* in 2014, and it is possible that there are more. Tombo output only provides information on the presence/absence and level of modification, not its identity. Based on the fact that base J is more abundant than 5hmU and 5fU (Bullard *et al.*, 2014), and that many of the patterns we see in the data match base J ChIPseq patterns, it is likely that most of the signal we see in Tombo output is from base J, but not with certainty. With currently available tools, we cannot disentangle 5hmU, base J and 5fU localisation in ONT data. This limitation could be a key consideration if Nanopore sequencing was used to predict base J in novel organism with increasing evolutionary distance from trypanosomatids.

4.3.2.3 Modified base deposition – stable or stochastic process?

Lastly, not much is known about how stable or stochastic modified base deposition is in *T. brucei*. The available 5hmU ChIPseq data, with two replicates, is not entirely consistent; whether that is due to technical differences or biological, is unclear, but it highlights the fact that a single experiment might not be reliable. Base J ChIPseq and Tombo datasets do not have replicates at all, and therefore we cannot determine with certainty which deposition patterns might be stable or transient, for example. Again, global similarity between base J ChIPseq and Tombo data (*de novo* and model sample mode) points towards the likelihood of stably modified genomic regions (repetitive elements, PTU boundaries), but more specific loci or loci with lower levels of modification may show more variation in modification levels between experiments.

4.3.3 Future work suggestion

There are a number of ways modified base detection in *T. brucei* can be improved in the future, even with current tools and techniques. First, as mentioned above, repeating the 5hmU and base J ChIPseq experiments using the same cell lines and sequencing approaches, as well as appropriate input samples. Tombo data may be further improved by using a ‘no-modification’ whole genome amplified sample as a control, as well as correcting the reference genome sequence to account for variants present in the cell lines used relative to it.

In addition, Tombo offers model training using sequences with known DNA modifications; in theory, synthetically modified oligonucleotides with one or more of the parasite’s DNA modifications can be used to train the tool to be able to recognise it specifically. This approach is more likely to work with 5hmU than base J, as 5hmU is present in a wider range of organisms (Gommers-Ampt and Borst, 1995; Olinski, Starczak and Gackowski, 2016), and therefore more likely to be available to order as synthetic oligonucleotides needed for training.

As it has been long suggested that base J acts in transcriptional repression in *T. brucei* (Schulz *et al.*, 2016), an additional experiment that would complement the ones suggested above would be performing RNA sequencing on the same samples as those sequenced for Tombo with the aim of evaluating any link between expression and modification levels of DNA sequences (genes or otherwise).

5 Nucleotide skews in the *Trypanosoma brucei* and *Leishmania major* genomes

5.1 Introduction

An organism's genomic sequence is inherently non-random. Across both prokaryotic and eukaryotic lineages the prescribed nature of codons and regulatory sequences translates into sequence nucleotide composition patterns and biases (Sueoka, 1962; Bennetzen and Hall, 1982; Ikemura, 1985; Sharp and Li, 1987; Bernardi *et al.*, 1988; Akashi, Kliman and Eyre-Walker, 1998), and these can be further perturbed by mutagenic biological processes. In particular, DNA replication and transcription have been found to introduce, in some cases, significant nucleotide biases, usually referred to as skews, which result in the over- or underabundance of a certain nucleotide on a given DNA strand relative to their complementary strand (Beletskii and Bhagwat, 1996; Lobry, 1996; Kano-Sueoka, Lobry and Sueoka, 1999; Pavlov, Newlon and Kunkel, 2002; Touchon *et al.*, 2004, 2005; Xia, 2012); such strand asymmetries are born out of the asymmetrical nature of DNA replication and transcription.

5.1.1 Transcription- and replication-associated mutagenesis

During transcription, RNA polymerase progresses directionally across the template DNA strand, synthesising the complementary nascent RNA (Figure 76 A) (Jinks-Robertson and Bhagwat, 2014). A dedicated transcription-coupled nucleotide excision repair (TC-NER) DNA repair mechanism operates during transcription, resolving DNA damage preferentially on the template strand to allow RNA polymerase elongation (Svejstrup, 2002). The non-template (coding) strand, on the other hand, remains single-stranded during transcription and is not preferentially repaired by TC-NER, thereby accumulating the bulk of transcription-associated mutations (Spivak and Ganesan, 2014; Moeckel, Zaravinos and Georgakopoulos-Soares, 2023).

A couple of studies have interrogated transcription-coupled nucleotide excision repair (TC-NER) and its machinery in *T. brucei* (Lecordier *et al.*, 2007; Lee *et al.*, 2007; Lee, Jung and Günzl, 2009; Badjatia *et al.*, 2013; Machado *et al.*, 2014). The work suggested that TC-NER in *T. brucei* is uncoupled from transcription factor II H (TFIIH) (Machado *et al.*, 2014) – a general transcription factor complex that is involved in NER in humans (Fousteri and Mullenders, 2008). The authors also suggested that TC-NER in *T. brucei*, and possibly other kinetoplastids, plays a dominant role in nucleotide excision repair, due to multigenic transcription (Machado *et al.*, 2014).

Similarly to transcription, DNA replication is an asymmetric process (Figure 76 B), continuous on the leading strand and discontinuous on the opposite (lagging) strand, with distinct machinery involved in each (reviewed in MacNeill, (2012)). In all domains of living organisms, as well as viruses, strand asymmetries stemming from origins of replication have been observed; in fact, nucleotide skews can be used to predict the location of origins of replication (Lobry, 1996; Pavlov, Newlon and Kunkel, 2002; Touchon *et al.*, 2005; Shinbrot *et al.*, 2014).

The exact cause of this ubiquitous replication-associated strand asymmetry remains unclear, but it is likely at least in part due to intrinsic properties of the machinery involved in this process. In human and yeast cells, distinct DNA polymerases with proofreading capabilities are responsible for replicating the leading and lagging strands – ε and δ, respectively (Morrison *et al.*, 1990; Pursell *et al.*, 2007; Nick McElhinny *et al.*, 2008). In both human and mouse cells, particularly cancer cells, deficiencies in the proofreading and DNA repair machinery more broadly have been associated with high mutational burden (Venkatesan *et al.*, 2007; Albertson *et al.*, 2009; Li *et al.*, 2018). Such heightened mutational burden often results in specific mutations that can lead to pronounced skew patterns (see Figure 76 C for examples) (Shinbrot *et al.*, 2014; Robinson *et al.*, 2021). While the precise nature of mutations and strand asymmetries varies among different organisms, cell types and genomic contexts (reviewed by Moeckel, Zaravinos and Georgakopoulos-Soares, (2023)), inherent chemical properties of DNA often play a key role in these mutational processes (see below).

In particular, cytosine deamination is thought to be the dominant spontaneous DNA mutation in *Escherichia coli* and mammalian cells (de Jong, Grosovsky and Glickman, 1988; Halliday and Glickman, 1991; Douglas *et al.*, 1994; Bhagwat *et al.*, 2016). Cytosine is the most unstable of the canonical bases in nucleic acids, particularly in single-stranded DNA (ssDNA), where cytosine deamination occurs ~ 140x more frequently compared to double-stranded DNA (dsDNA) (Frederico, Kunkel and Shaw, 1990). If left unrepaired, cytosine deamination results in C>T

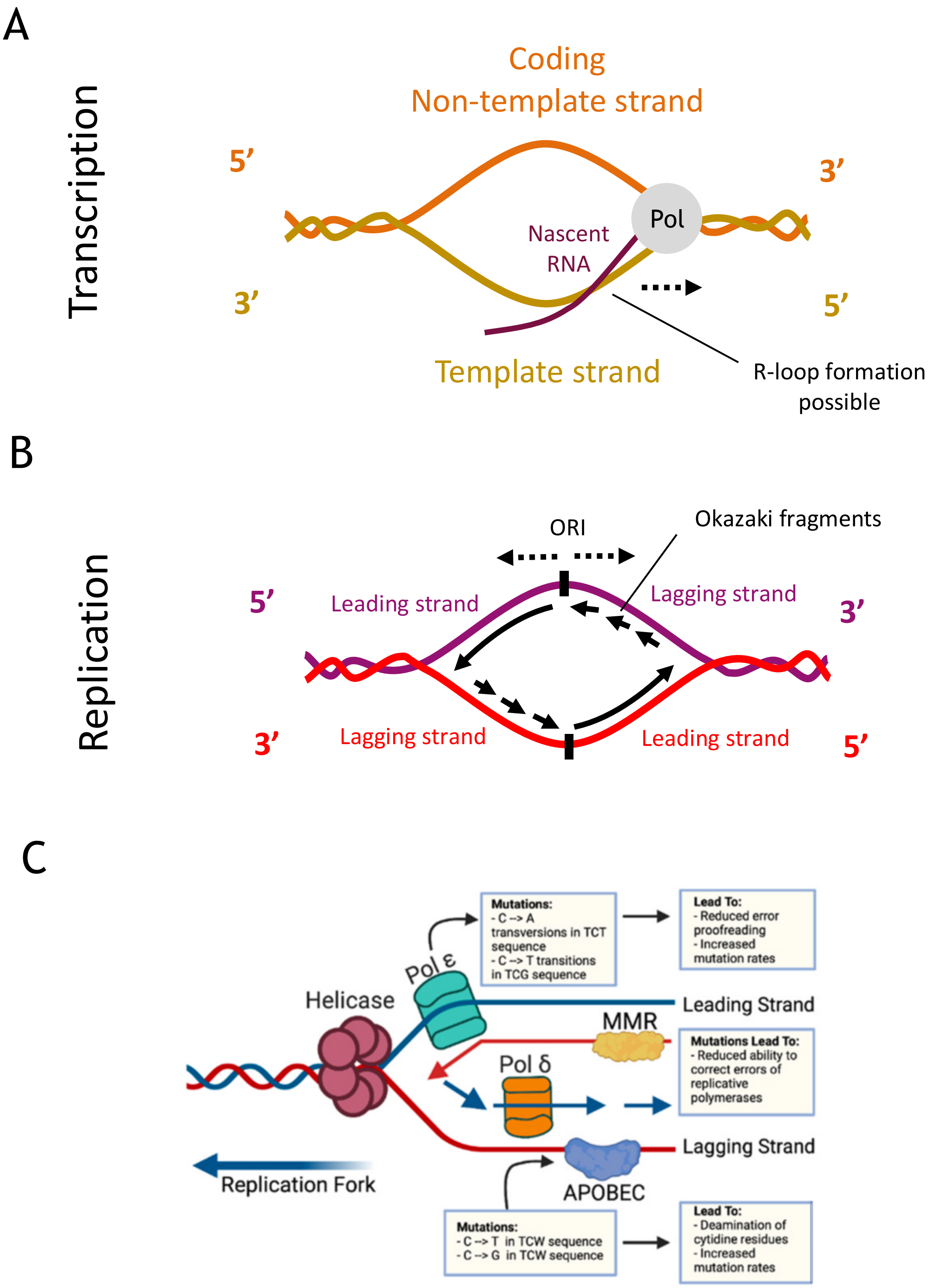


Figure 76 Strand asymmetries associated with transcription and replication.

A – Diagram depicting a simplified transcription fork with an RNA polymerase enzyme traversing the nucleic acid left to right (indicated by dashed arrow direction). Pol – RNA polymerase. B – Diagram depicting a simplified replication bubble, with leading and lagging strands indicated. ORI – origin of replication, short arrows represent Okazaki fragments, longer arrows – the newly replicated DNA on the leading strand. C – a diagram highlighting some examples of the processes and machinery involved in DNA replication and repair in human cells, and how mutations in the corresponding genes can lead to mutations. MMR – mismatch repair, APOBEC - apolipoprotein B mRNA editing catalytic polypeptide-like family of cytidine deaminases, Pol – polymerase, TCW – TCA or TCT sequence. Adapted from Moeckel, Zaravinos and Georgakopoulos-Soares, (2023) under Open Access, Creative Commons CC BY 4.0.

mutations – this signature is particularly common in contexts where a DNA strand remains single stranded, such as during transcription in the non-template (coding) strand (Beletskii and Bhagwat, 1996). In human and model eukaryotic genomes, C>T deamination can explain the observed enrichment of G and T residues on the non-template strand (Touchon *et al.*, 2004; Moeckel, Zaravinos and Georgakopoulos-Soares, 2023). Interestingly, the rate of C deamination increases with increased temperature (Lewis *et al.*, 2016).

Repeated mutational processes lead to nucleotide skews and can affect the overall genome nucleotide composition and even organisation. In both prokaryotes and eukaryotes, more genes are often found on the leading strand of DNA (Huvet *et al.*, 2007; Srivatsan *et al.*, 2010; Merrikh *et al.*, 2012); it is thought that the reason for this is the more 'favourable' lower mutation rate of the leading strand due to co-directional replication and transcription. Head-on collision of transcription and replication machinery, which can occur if these processes are not co-directional, can lead to genomic instability (Touchon *et al.*, 2004); while collisions still occur in co-directional replication and transcription environments, they are typically less disruptive (reviewed in Brambati *et al.*, (2015)).

In *Trypanosoma brucei*, there has been limited study of transcription-replication conflicts. However, the existing data in *T. brucei* suggests that transcription occurs throughout the cell cycle, including the S phase (da Silva *et al.*, 2019). Transcriptional arrest in *T. brucei* drastically reduces genome-wide R-loop accumulation, as well as γH2A levels, a histone modification frequently used as an indicator of DNA damage (da Silva *et al.*, 2019), suggesting that either transcription-related processes or conflicts between replication and transcription are responsible for R-loop and DNA damage accumulation. To our knowledge, similar experiments have not been performed in *Leishmania* parasites.

5.1.2 Nucleotide skews can lead to formation of secondary structures

Nucleotide skews may favour the formation of R-loops – triple-stranded RNA-DNA hybrid regions (Figure 76 A) that have been found to have a major impact on replication progression and transcription, as well as play a role in genomic instability and double stranded DNA break formation (Crossley, Bocek and Cimprich, 2019). In unmethylated CpG islands and promoters in the human genome, GC skew appears to be correlated with R-loop formation (Ginno *et al.*, 2012, 2013); in *Arabidopsis*, both AT and GC skews are associated with R-loop formation (Xu *et al.*, 2017). In addition to R-loops, G-quadruplexes – secondary G-rich DNA structures that typically form within a single strand - can also be associated with GC skew, specifically on the non-template G-rich strand of DNA that exhibits positive GC skew (Duquette *et al.*, 2004). A set of experiments mapping G-quadruplexes in several organisms, including *Leishmania major* and *Trypanosoma brucei*, showed that G- and GG-richness, as well as G/C fraction ratio (G fraction / C fraction) were strongly and positively associated with G-quadruplex abundance between species (Marsico *et al.*, 2019).

5.1.3 Nucleotide skews in trypanosomatids

In *Trypanosoma* and *Leishmania* species relatively little is known about nucleotide skews. Limited skew analysis in *T. cruzi*, *T. brucei* and *L. major* was performed (Nilsson and Andersson, 2005) before the publication of the first whole genome sequences of these parasites (Berriman *et al.*, 2005; El-Sayed, Myler, Bartholomeu, *et al.*, 2005; Ivens *et al.*, 2005). The two trypanosome species were shown to exhibit positive GC and negative AT skews with transcription direction, whereas in *L. major*, as had been reported earlier (McDonagh, 2000), both skews were negative with transcription direction (Nilsson and Andersson, 2005). Limited genome annotation and sequence was available at the time, and only one chromosome per organism was used in the analyses. However, the reversal of GC skew pattern in *L. major* relative to the trypanosomes was a peculiar finding given the perceived evolutionary proximity of these species. While *Leishmania* and *Trypanosoma* are in the same family – Trypanosomatidae - they are estimated to have diverged more than 200 million years ago (MYA) (Figure 77) (Lake *et al.*, 1988; Douzery *et al.*, 2004), by some estimates even 400-600 MYA (Overath *et al.*, 2001) – for context, human ancestors and dinosaurs (Dinosauria) diverged approximately 319 MYA (timetree.org). In fairness, in the case of the trypanosomatids, a definitive timing is difficult to achieve as it relies entirely on sequence divergence – no

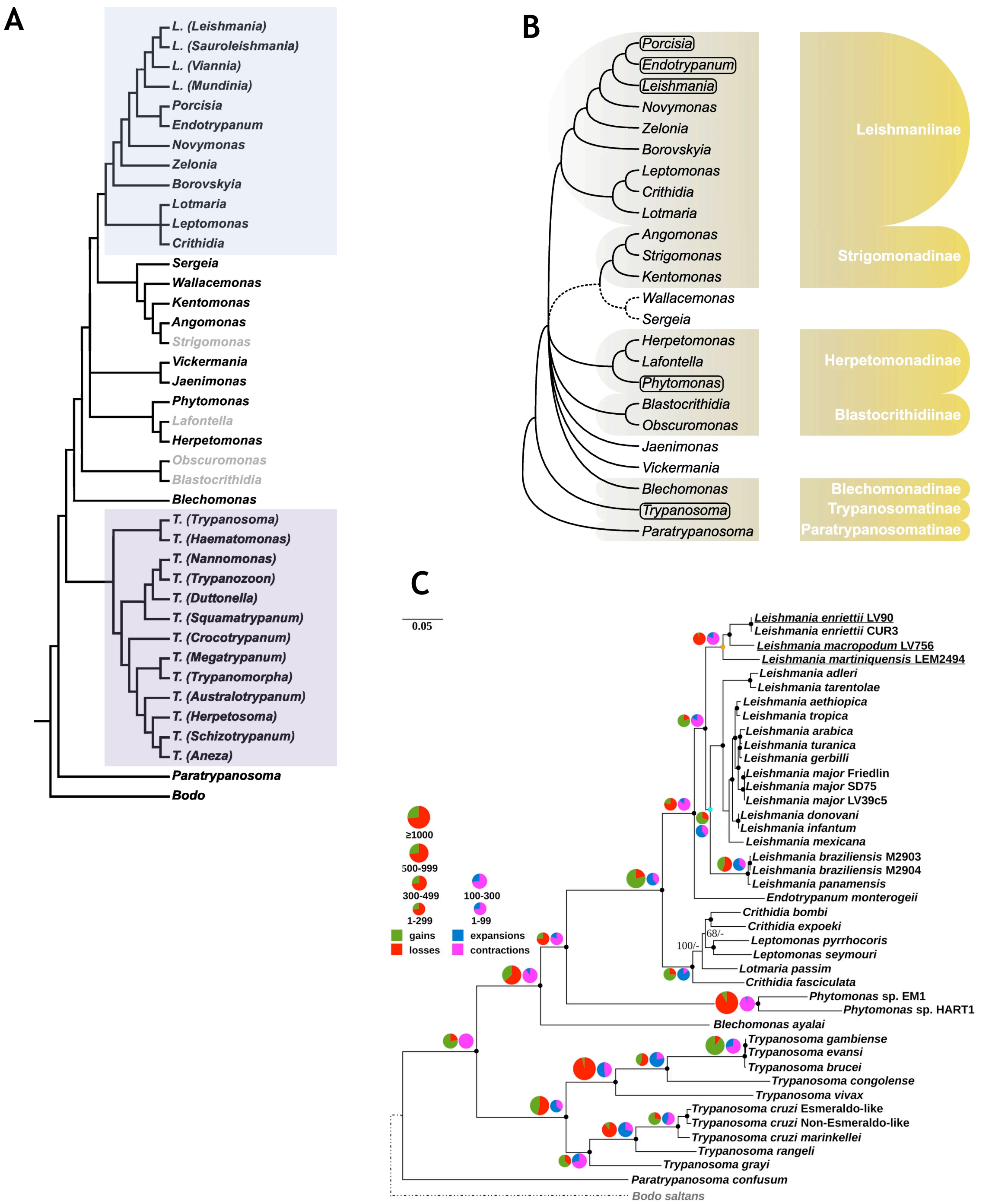


Figure 77 Evolutionary relationships among Trypanosomatidae.
A. Most recent and complete phylogenomic tree of Trypanosomatidae, focusing on genera and subgenera (adapted from Kostygov et al., 2024, license number 5995351449296). The tree was developed, specifically, to serve as a phylogenetic framework for analysis of various traits in an evolutionary context; the authors used a combination of independent analyses and data from a previously published tree (Albanaz et al., 2021) that focused on *Endotrypanum* and *Porcisia*, specifically, as basis for it (Kostygov et al., 2024). Leishmaniinae subfamily highlighted in light blue, Trypanosomatinae – in purple. The genera in grey highlight those that have been excluded from analysis in this chapter. B. A recent cladogram depicting the relationships among trypanosomatids (based on the available literature). Dixenous genera displayed in boxes, subfamilies indicated on the righthand side on yellow background. Figure adapted from Albanaz *et al.*, (2023), reproduced under Open Access, Creative Commons CC-BY 4.0. C. A phylogenetic tree of trypanosomatids and *Bodo saltans* based on 92 conserved proteins. The pie charts at the nodes highlight changes in ortholog groups (gains – green, losses – red, expansions – blue, contractions – magenta). The

scale bar represents 0.05 substitutions per site. Figure from Butenko *et al.*, 2019 (reproduced under Open Access, Creative Commons CC-BY 4.0).

fossil record is readily available; for this reason, the time estimations are just that – estimations.

Further nucleotide skew analyses have been performed in *L. major*, specifically at early replicating origins of DNA replication; these regions showed a highly localised reversal of skew polarity for both GC and AT skews and a corresponding localised peak of G-quadruplexes on the G-rich strand (Marsico *et al.*, 2019; Damasceno *et al.*, 2020). Such analysis has not been performed in *T. brucei*.

Studies in diverse organisms have indicated that replication-associated nucleotide skews tend to be more prevalent compared to those associated with transcription (Moeckel, Zaravinos and Georgakopoulos-Soares, 2023). Given the polycistronic nature of transcription and the frequent co-localisation of replication and transcription machinery and genomic features at polycistronic transcription unit boundaries in *L. major* and *T. brucei* (Tiengwe, Marcello, Farr, Dickens, *et al.*, 2012; Marques *et al.*, 2015), further analysis of nucleotide patterns and skews in these parasites may provide crucial insight into how replication and transcription are orchestrated in these unusual eukaryotic genomes.

5.1.4 Aim

To date, nucleotide skew analysis has only been performed in *T. brucei* in one chromosome, and only in the context of transcription. The aim of this short chapter was to provide more comprehensive, genome-wide characterisation and comparison of *Trypanosoma brucei* and *Leishmania major* genomic nucleotide composition, as well as determine the potential contribution of transcription and replication to any observed nucleotide skews. By also analysing nucleotide skews in other trypanosomatids for which genome sequences are available, we attempted to discern phylogenetic associations in skew patterns.

5.2 Results

5.2.1 Overall nucleotide composition of *T. brucei* and *L. major* genomes

We chose to focus our analysis on the strains and genome versions that are most complete and most utilised by the research community and our lab – the updated *T. brucei brucei* Lister 427 assembly from 2018 (Müller *et al.*, 2018) and the *L. major* Friedlin assembly, both from TriTrypDB. First, assessing genome-wide nucleotide patterns of the two parasites (Table 16), *T. brucei* displayed a larger difference between overall and coding GC % (43.71 vs 50.44) than *L. major* (59.72 vs 62.49). Looking further at specific nucleotide abundance differences between overall and coding DNA, it is clear that some nucleotides show more drastic shifts than others – in particular, T and G abundances vary considerably in *T. brucei*, with higher G and lower T levels in coding sequence compared to the rest of the genome (Table 16). A-to-T and G-to-C abundance ratios within coding sequences were not 1:1 in *T. brucei* or *L. major*, and the ratios were more extreme in the *T. brucei* genome compared to *L. major*, likely to be reflected in more pronounced coding sequence AT and GC skews.

Table 16 Nucleotide composition and general characteristics of *T. brucei* and *L. major* reference genomes

	Trypanosoma brucei brucei	*Leishmania major*
Genome version	TriTrypDB release 46 TbruceiLister427_2018	TriTrypDB release 58 LmajorFriedlin
Genome size, bp	50,081,021	32,855,095
Gene density, genes per megabase	267.65	271.68
GC, %	43.71	59.72
Coding GC, %	50.44	62.49
Total A, %	27.57	19.97
Total coding A, %	28.17	19.24
Total T, %	28.66	20.31
Total coding T, %	22.45	18.27
Total C, %	21.79	29.99
Total coding C, %	23.03	31.04
Total G, %	21.88	29.73
Total coding G, %	26.35	31.45
Total N, %	0.10	<0.01

5.2.2 AT and GC skews, as well as G-quadruplexes, follow transcription direction

AT and GC skews are commonly used to assess nucleotide biases in genomic sequences, particularly in relation to processes such as transcription and replication (Moeckel, Zaravinos and Georgakopoulos-Soares, 2023). As *T. brucei* displays more clearly defined genome compartmentalisation, we decided to first assess the AT and GC skews of the main genomic compartments – the core genome, subtelomeres, and the bloodstream-form expression sites (Figure 78). Two key elements emerged from this analysis. First, subtelomeric contigs displayed more consistent AT and GC skew values within a given contig – this coincided with the inferred direction of transcription based on CDS directions and in many of the subtelomeric contigs a dominant strand (with respect to gene content) could be observed (Figure 78 B). Similarly, as all BES contigs in the reference are in the forward strand orientation (transcribed left-to-right), BES sequences showed a clearer overall pattern compared to the core genome, as the latter consists of PTUs that run in both directions.

More broadly, AT and GC skews were pronounced in both parasites and followed transcription direction (Figure 79 and Figure 80), though not in the same way. In the *T. brucei* genome, a negative AT skew and a positive GC skew was associated with transcription direction, whereas in *L. major* both skews showed negative values; putting it another way, an overabundance of T and G residues was present on the forward strand in *T. brucei*, and an overabundance of T and C in *L. major*.

Furthermore, differential experimental G-quadruplex mapping was evident in both species and coincided with transcription direction, though this was much more pronounced in *L. major* than *T. brucei* (Figure 81 and Figure 82). A higher density of G-quadruplex formations was seen on the forward strand in *L. major* and the reverse in *T. brucei* – in both cases the strand with a negative GC skew (lower number of G residues relative to C). Even when interrogating genome regions that are syntenic, these trends held true (Figure 83).

Curiously, in *T. brucei* a different skew pattern to the core genome was present in the subtelomeres and BES – whereas in the core a positive AT and negative

GC skews were associated with transcription direction, in the subtelomeres and BES both skews were positive (Figure 78 B and Figure 84). The switch in AT skew occurred around the first *VSG* array genes in the subtelomeres (Figure 84).

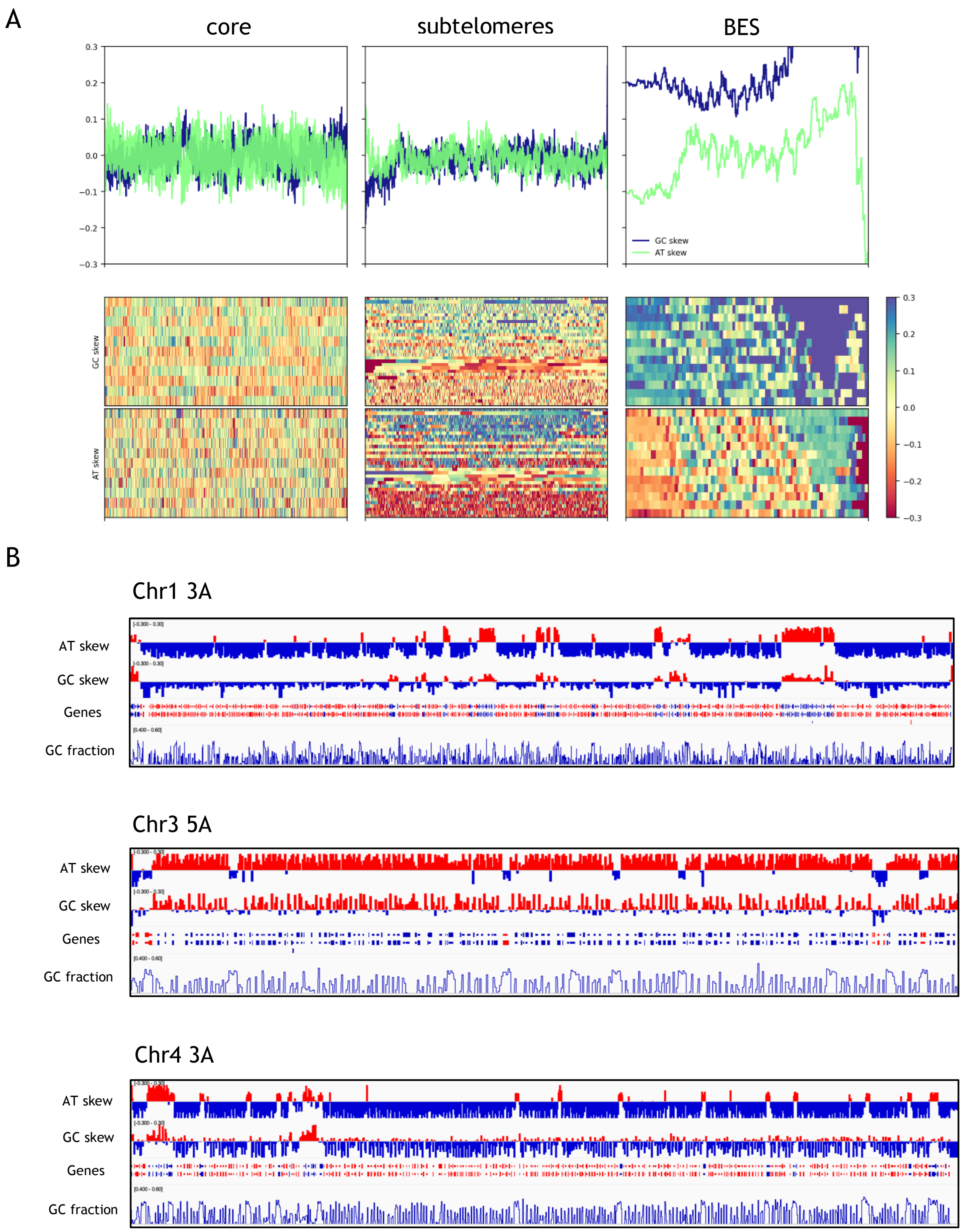


Figure 78 AT and GC skews across genomic compartments of *T. brucei.*

A – Metaplots and heatmaps showing GC and AT skews in the genome compartments of *T. brucei* – core, subtelomeres and BES, as designated in the reference genome assembly (Müller *et al.*, 2018). B – examples of GC and AT skews at subtelomeric sequences of *T. brucei* that show a predominant strand with respect to gene localisation. Genes – annotations in the reference gff file from TriTrypDB, with blue representing annotations on the forward strand, and red – those on the reverse strand.

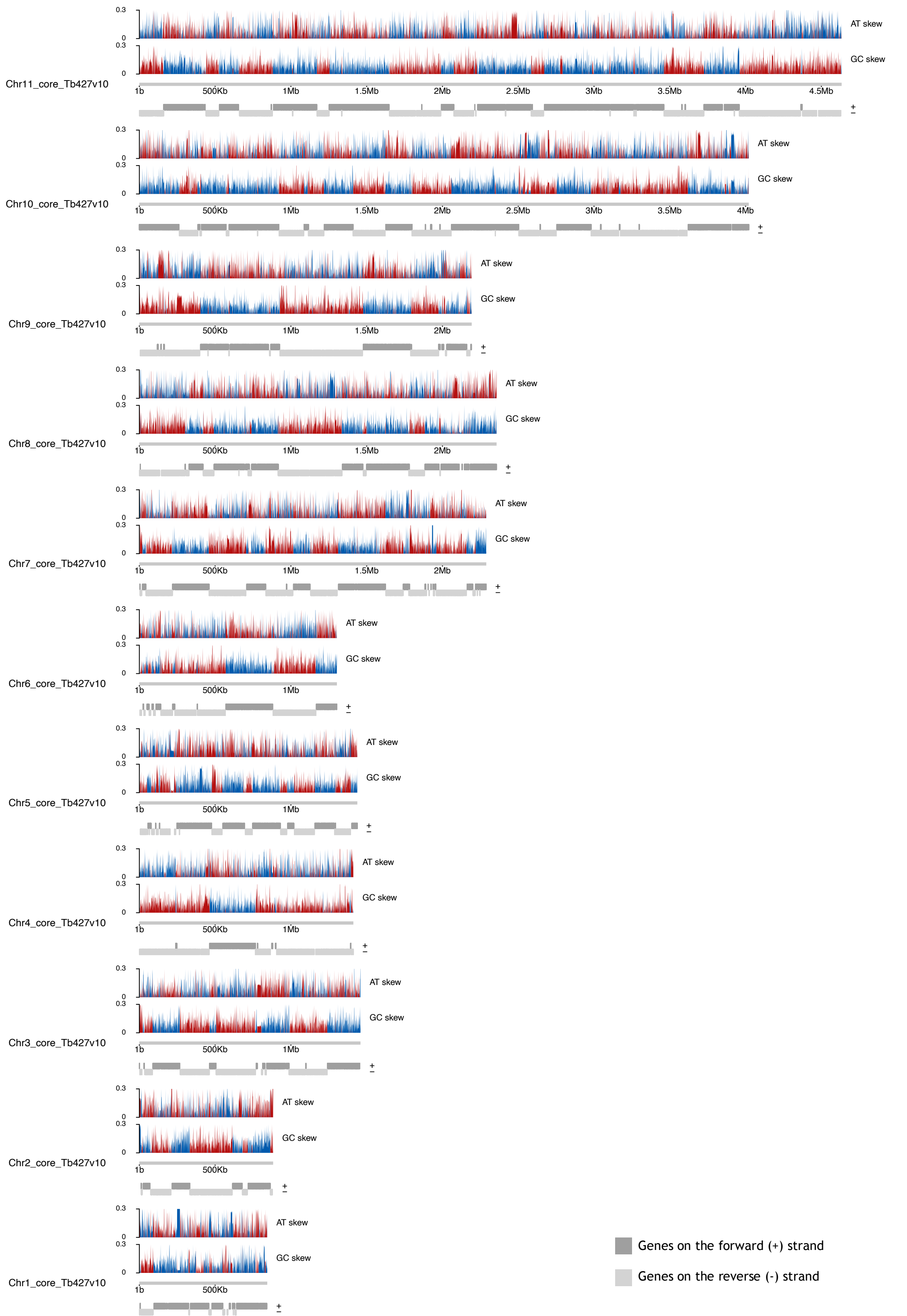
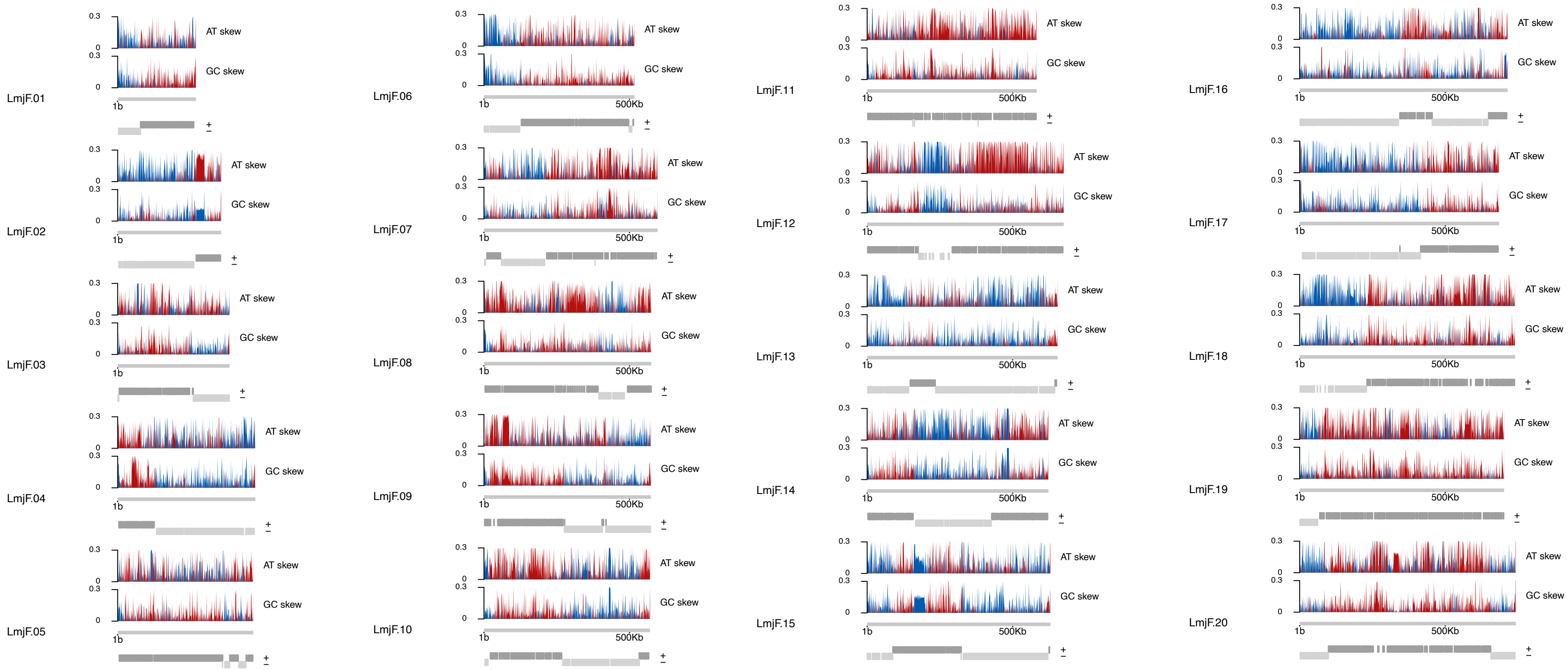


Figure 79 Nucleotide skews follow transcription direction in *T. brucei.*

Mapping of AT and GC skews along the chromosome core regions of 11 megabase chromosomes of *T. brucei*. Red – negative skew values, blue – positive skew values. Chr – chromosome, '+' – genes on the forward strand, '-' – genes on the reverse strand.



Genes on the forward (+) strand

Genes on the reverse (-) strand

(continued on the following page)

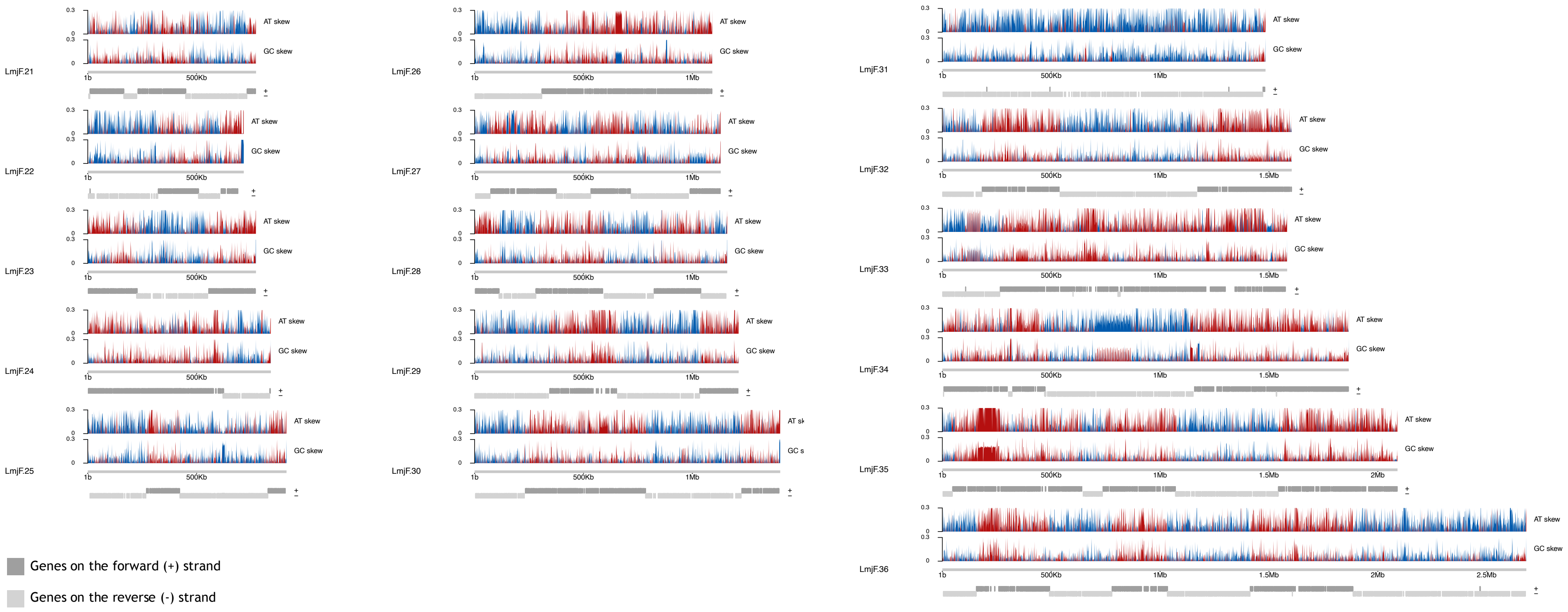


Figure 80 Nucleotide skews follow transcription direction in *L. major.*

Mapping of AT and GC skews along all 36 chromosomes of *L. major*. Red – negative skew values, blue – positive skew values. '+' – genes on the forward strand, '-' – genes on the reverse strand.

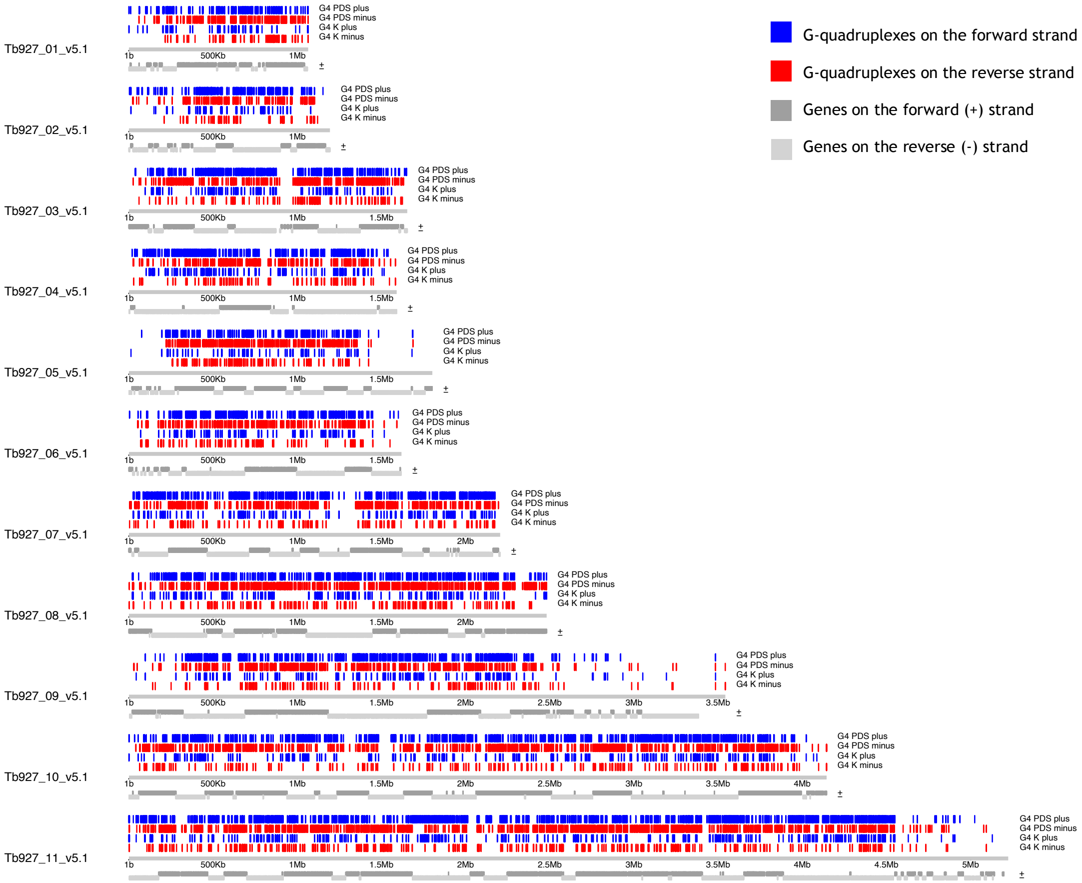


Figure 81 G-quadruplex distribution in the core genome of *T. brucei (*strain TREU927).

Mapping of experimentally determined G-quadruplex structures in *T. brucei* (Marsico *et al.*, 2019). Two different sequencing buffers were used to stabilise G-quadruplexes in *T. brucei* DNA – here referred to as K and PDS. 'K' refers to the standard buffer that is intended to resemble physiological conditions (described by Chambers *et al.*, (2015)), whereas 'PDS' refers to the addition of a G-quadruplex structure stabilizer – pyridostatin instead of Na^{+}.

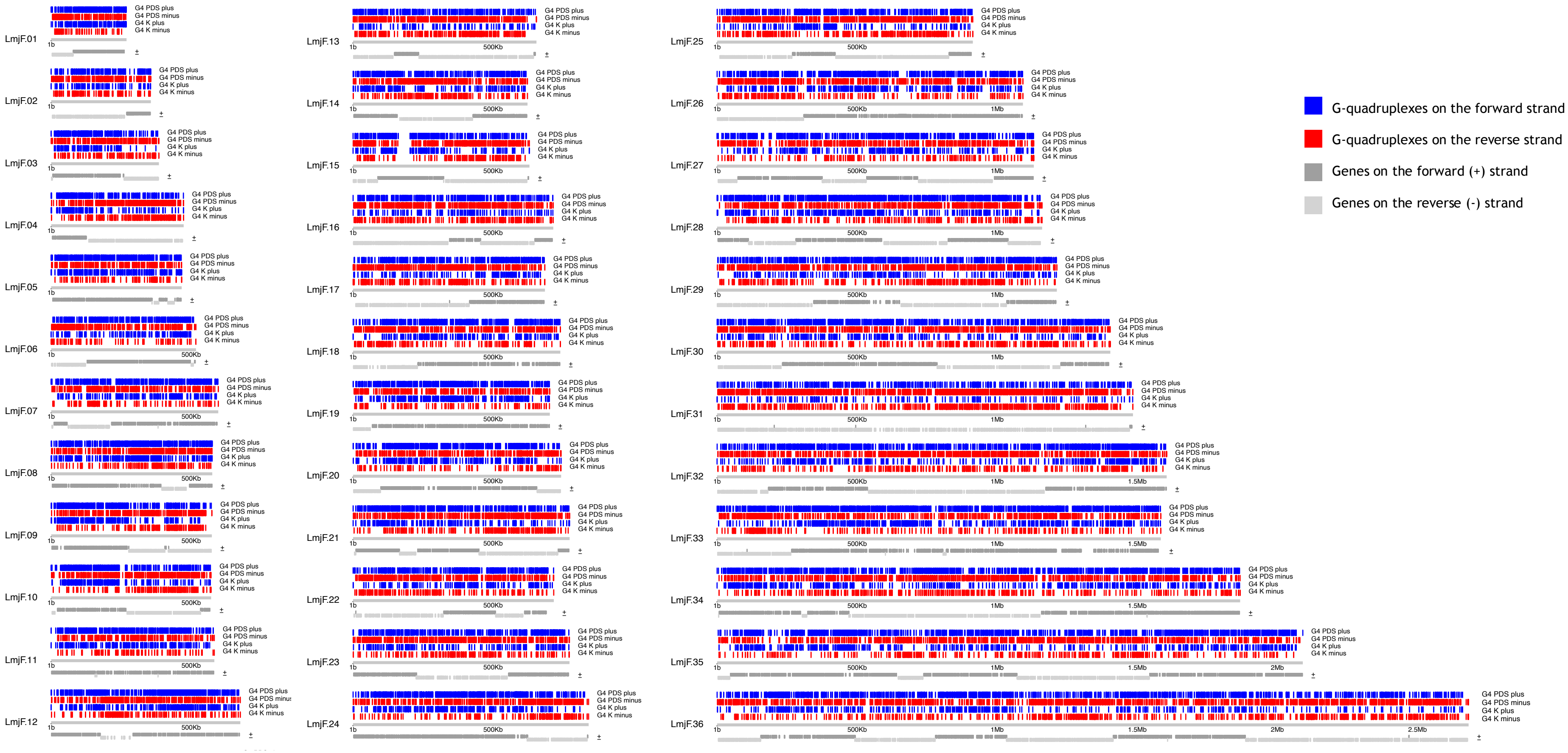


Figure 82 Global G-quadruplex deposition in *L. major*.

Mapping of experimentally determined G-quadruplex structures in *T. brucei* (Marsico *et al*., 2019). Two different sequencing buffers were used to stabilise G-quadruplexes in *L. major* DNA – here referred to as K and PDS. ‘K’ refers to the standard buffer that is intended to resemble physiological conditions (described by Chambers *et al*., (2015)), whereas ‘PDS’ refers to the addition of a G-quadruplex structure stabilizer – pyridostatin instead of Na^+.

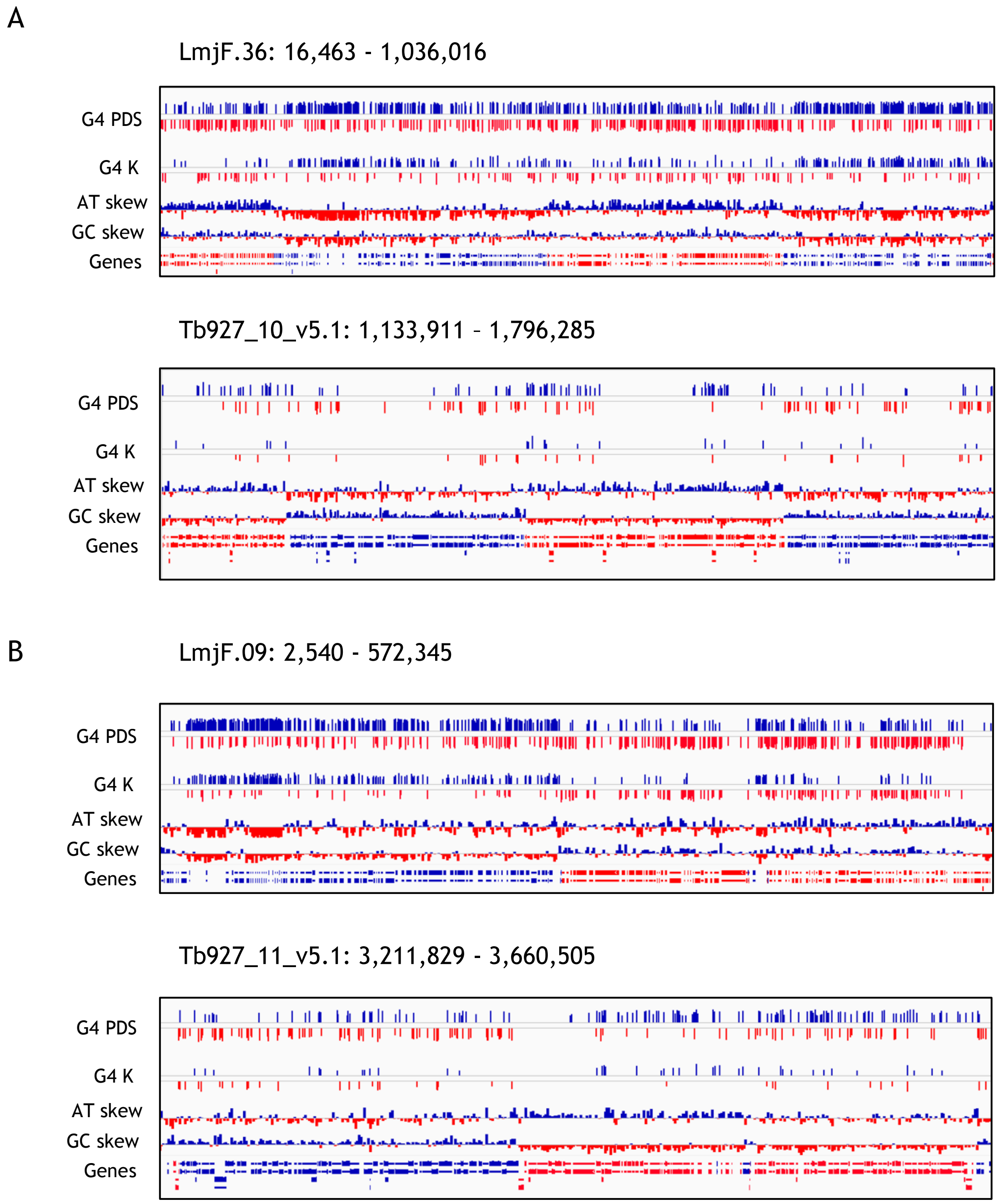


Figure 83 Syntenic regions between *L. major* and *T. brucei* show opposite GC skews with transcription direction.

Two examples (A and B) of syntenic regions between *L. major* and *T. brucei* that highlight the divergent GC skew and G-quadruplex patterns in the two parasites. G4 K – G-quadruplex mapping using the standard sequencing buffer that is formulated to mimic physiological conditions, blue – G-quadruplex mapping on the forward strand, red – on the reverse. G4 PDS – same as above, except the sequencing buffer includes PDS instead of Na^+ for stabilization of G-quadruplex structures (G-quadruplex data mapped by Marsico *et al.*, (2019). Genes – annotations from the corresponding reference gff annotation file from TriTrypDB. Syntenic sequences were identified using the TriTrypDB genome browser track that highlights syntenic sequences between specified genomes.

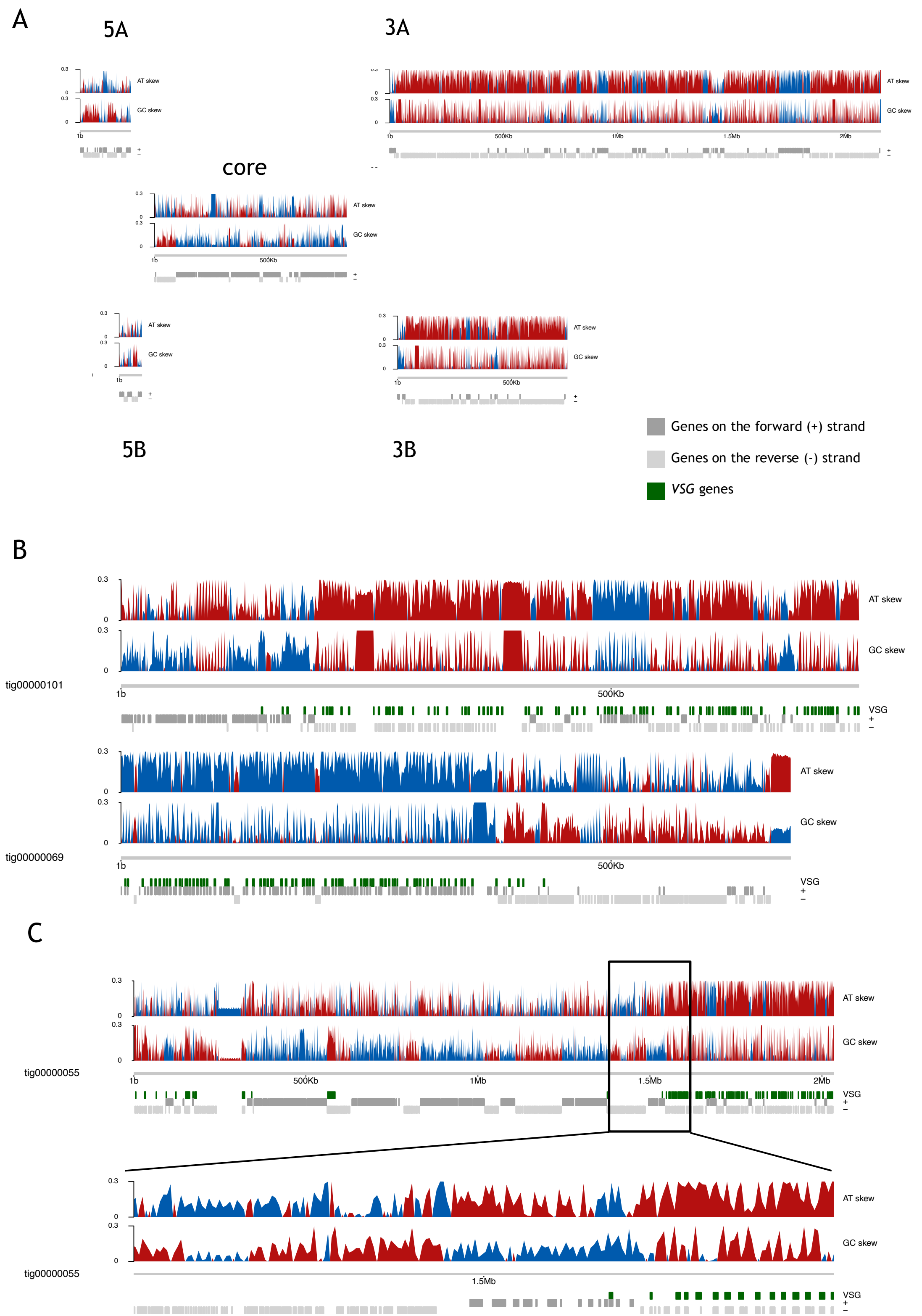


Figure 84 Subtelomeric sequences of *T. brucei* show divergent GC skew patterns.

A – AT and GC skews of chromosome 1, as annotated in the reference genome (5A, 5B, 3A and 3B represent subtelomeric arms of the chromosome). B – AT and GC skews of two contigs containing both core and subtelomeric sequences, identified in the long-read genome assembly described in Chapter 3, with *VSG* annotations indicating the subtelomeres. C – similarly to B, AT and GC skews of a contig containing both core and subtelomeric genomic regions (tig00000055, corresponding to chromosome 5 core and subtelomere 3A). Skews in blue indicate positive skew values, red – negative.

5.2.3 Codon usage and coding sequence skews

We decided to investigate whether codon usage and coding sequence skew, specifically, could explain the differences between the two parasites, as well the AT skew 'flip' seen at the core-subtelomere boundary in *T. brucei*. First, codon usage preferences in the form of frequency and abundance did not highlight any clear differences between *T. brucei* and *L. major* in the use of G-skewing or C-skewing codons (Figure 85). To assess codon usage contribution to G skew further, we assigned each codon a score depending on the GC content – for each G residue in the codon, +1 was added to the score, and -1 was subtracted for each C; positive values would contribute positively to GC skew, whereas negative ones – negatively. Then, we multiplied the score with the corresponding codon abundance for each parasite and subtracted the *L. major* score from the *T. brucei* one; positive values indicate a G-skewing codon is more abundant in *T. brucei* than *L. major*, or a C-skewing codon is more abundant in *L. major,* and vice-versa for the negative values. Overall, more G-skewing codons were found to be used by *T. brucei* (Figure 86), which also matches the coding G and C % discrepancies we saw earlier (Table 16).

Next, to evaluate whether AT skew direction reversal at the core genome – subtelomere boundary in *T. brucei* can be explained via differential skew patterns within the coding sequences, we compared nucleotide composition of *VSG* and non-*VSG* genes (Figure 87). Curiously, for both non-*VSG* and *VSG* genes, the median AT skew value was positive, albeit drastically different - 0.378 for *VSG* genes and 0.032 for non-*VSG* genes on the forward strand (-0.379 and -0.044 on the reverse, respectively), indicating that non-coding sequence is the likely contributor to AT skew differences between the core and subtelomeric compartments (more on coding and non-coding skews later).

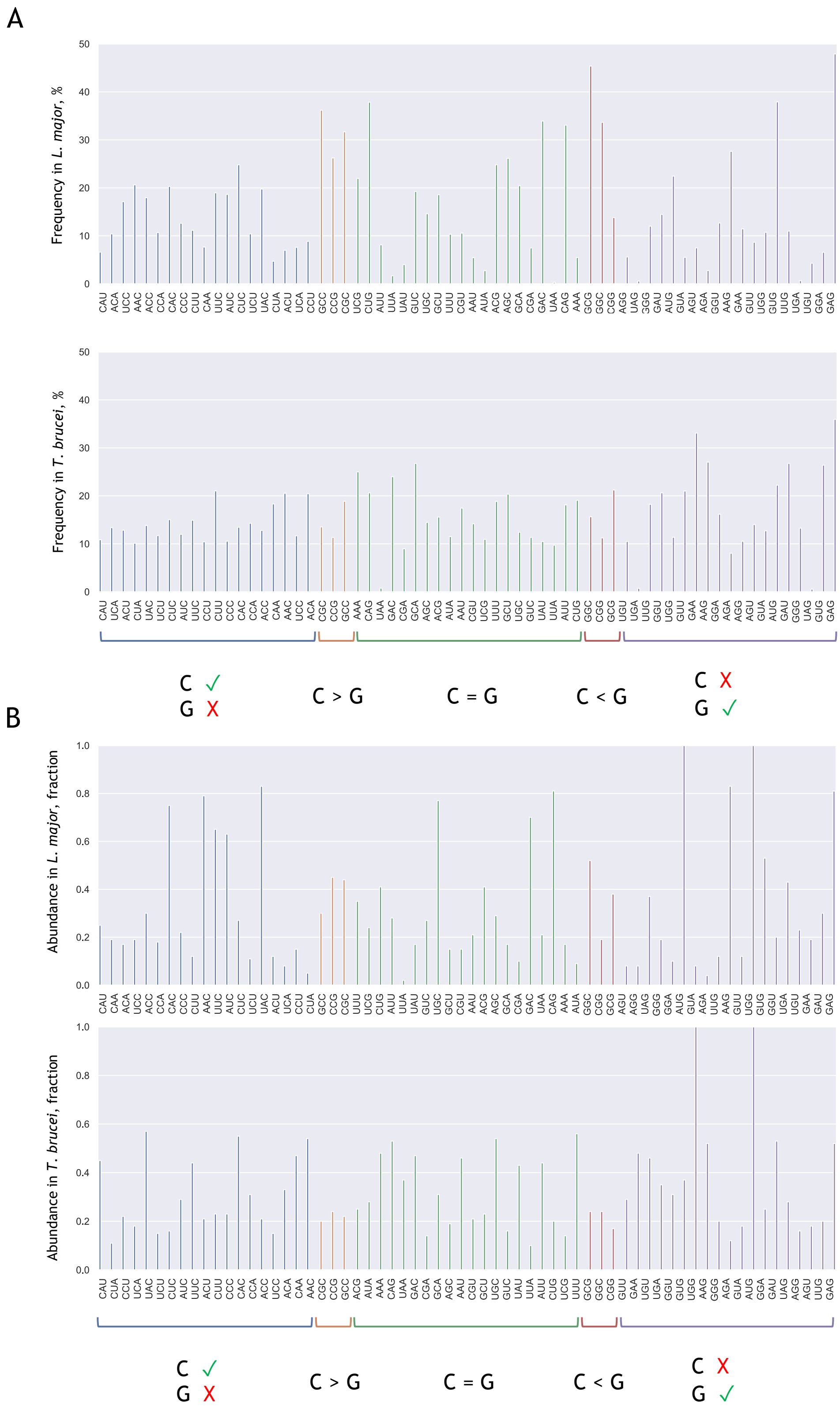


Figure 85 Codon usage in *Leishmania major* and *Trypanosoma brucei*.

Codon frequencies (A) and abundances (B) for *L. major* and *T. brucei* were obtained from the TriTrypDB codon usage table for the respective reference genomes and plotted in groups with increasing G-skewing from left to right.

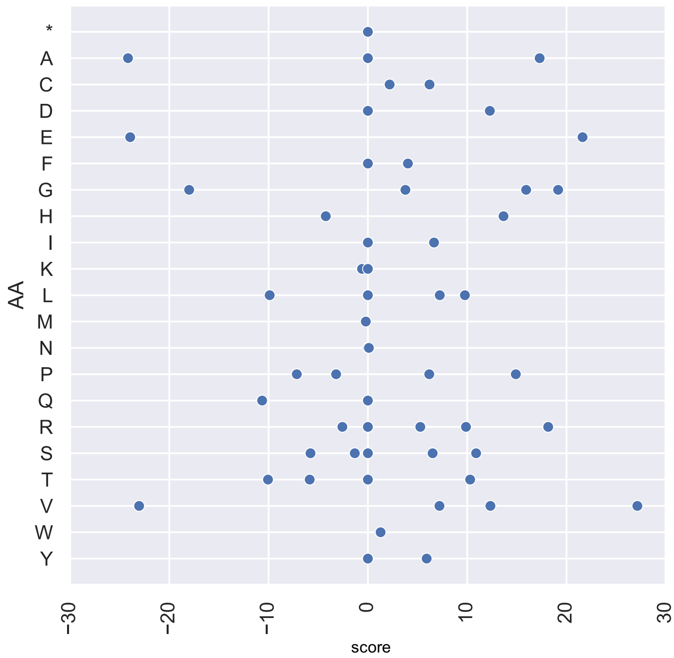


Figure 86 More G-skewing codons are used by *T. brucei* compared to *L. major*.

Each codon used by *T. brucei* and *L. major* was assigned a score depending on the GC content – for each G residue in the codon, +1 was added to the score, and -1 was subtracted for each C; positive values would contribute positively to GC skew, whereas negative ones – negatively. Then, we multiplied the score with the corresponding codon abundance for each parasite and subtracted the *L. major* score from the *T. brucei* one. Positive values on the x axis indicate a G-skewing codon is more abundant in *T. brucei* than *L. major*, or a C-skewing codon is more abundant in *L. major,* and vice-versa for the negative values.

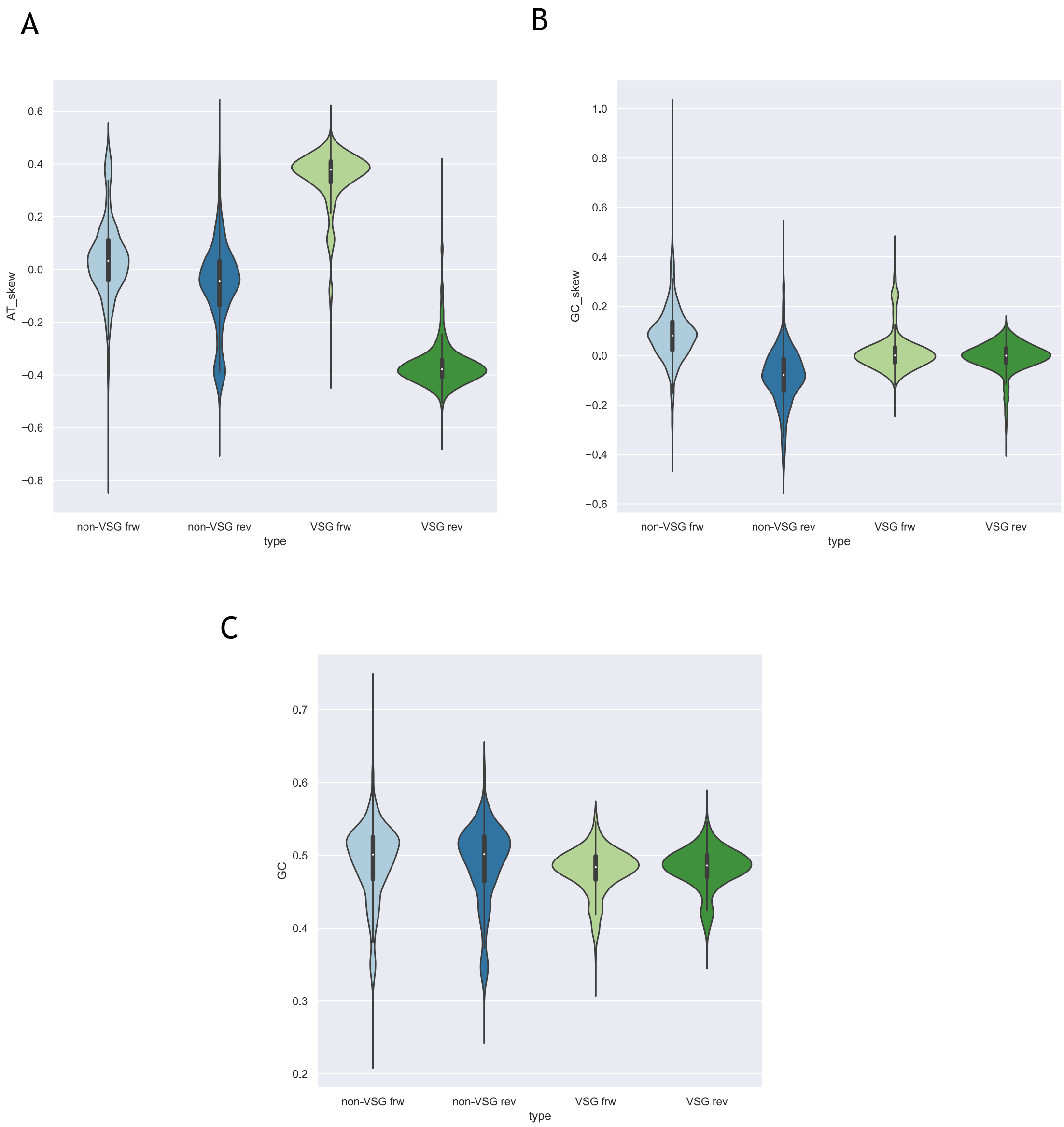


Figure 87 Nucleotide skews in *VSG* and non-*VSG* gene sequences of *T. brucei.*

Distribution of AT and GC skews (A and B), as well as GC fraction (C), of *VSG* and non-*VSG* gene coding sequences of T. *brucei*. Frw indicates genes on the forward strand, rev – genes on the reverse strand.

5.2.4 Nucleotide skews at polycistronic unit boundaries and replication sites

Polycistronic unit boundaries in *T. brucei* and *L. major* play a major role in both transcription and DNA replication in these parasites – not only are these sites of RNA polymerase II (RNAP II) transcription initiation and/or termination, but in some cases they also co-localise with putative replication initiation sites, as mapped by MFAseq (Tiengwe, Marcello, Farr, Dickens, *et al.*, 2012; Marques *et al.*, 2015; Devlin *et al.*, 2016), ORC1/CDC6 mapping in *T. brucei* (Tiengwe, Marcello, Farr, Dickens, *et al.*, 2012) and, more recently in our lab, DNAscent (Damasceno *et al.*, 2024a).

To investigate whether DNA replication contributes to nucleotide skew in the two parasites, we plotted AT and GC skews for PTUs with and without putative DNA replication origins in *T. brucei* (Figure 88) and *L. major* (Figure 89 and Figure 90). As discussed above, in *T. brucei*, a negative AT skew and positive GC skew appears to be characteristic on the coding strand of the core genome, and the corresponding reversal of these trends was seen with strand switching at convergent or divergent PTU boundaries; it was particularly pronounced for the GC skew. In the PTU boundaries that contain putative origins of replication, both the AT and GC skews were more exaggerated compared to the PTU boundaries that, to our knowledge, do not contain origins – this difference can be clearly seen in the divergent strand switch regions (top set of panels in Figure 88) where the transcription and replication would both act bidirectionally. The picture was slightly different at head-to-tail PTU boundaries (bottom set of panels in Figure 88) where transcription and replication are not fully co-directional. First, there appeared to be a dip in GC skew values right at the boundary and, second, the exaggerated skews were only seen where transcription and replication are codirectional (highlighted in Figure 88).

In *L. major* (Figure 89, Figure 90), the expected negative AT and GC skews associated with transcription direction could be seen. Similar to *T. brucei*, replication appeared to contribute a positive GC skew and negative AT skew in *L. major*. In addition, origin-containing PTU boundaries, as well as some non-origin ones that do not contain a base J peak, appeared to also contain very localised and sharp AT and GC skew fluctuations.

Figure 88 Nucleotide skews at PTU boundaries of *T. brucei*.

Nucleotide skews at convergent, divergent and head-to-tail polycistronic transcription unit (PTU) boundaries, with ± 50kb flanking regions, were plotted for *T. brucei,* further dividing these regions based on whether they contain putative origins of replication. Grey arrows indicate the presumed direction of replication progression, whereas blue and red arrows indicate the direction of transcription, with red indicating transcription on the reverse (bottom) strand, whereas blue – on the forward (top) strand. Ori – origin-containing boundaries, non-ori – boundaries that don't contain origins of replication, cTTS – convergent transcription termination sites, dTSS – divergent transcription start sites, htSSR – head-to-tail PTU boundaries, l-r – left-to-right, r-l – right-to-left, referring to the direction of flanking PTU transcription.

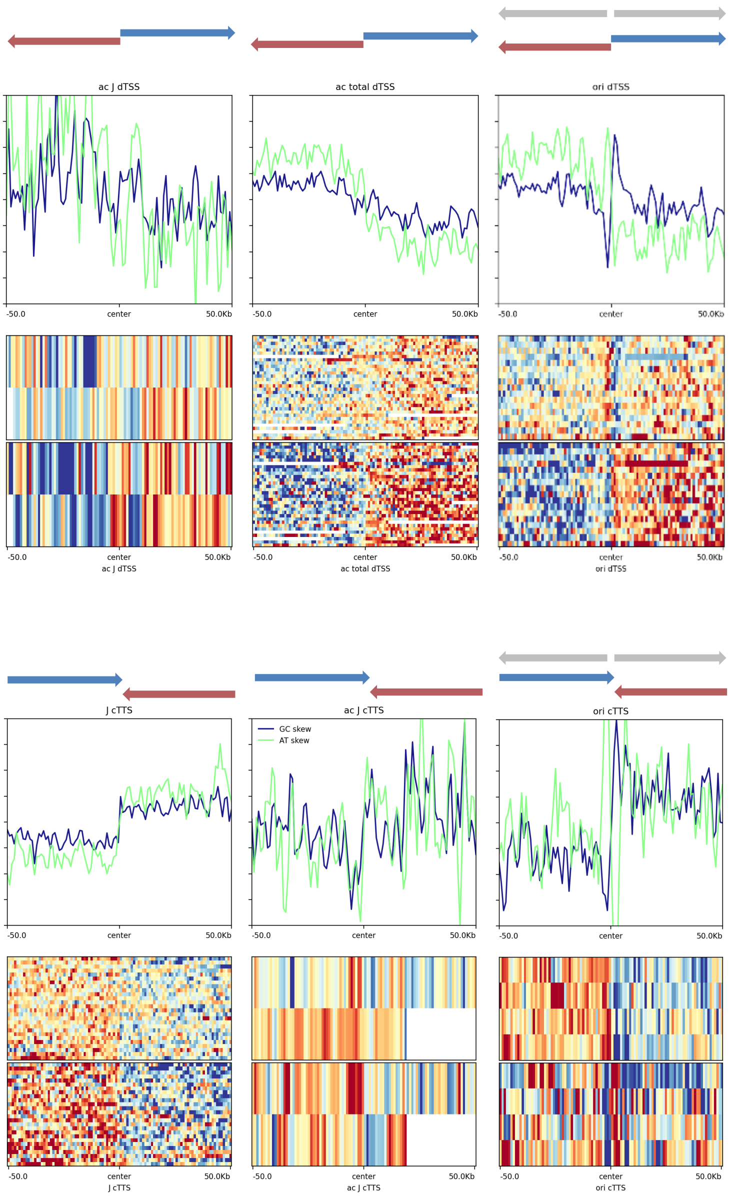


Figure 89 Nucleotide skews at convergent and divergent PTU boundaries of *L. major.*

Nucleotide skews at convergent and divergent polycistronic transcription unit (PTU) boundaries, with ± 50kb flanking regions, were plotted for *L. major,* further dividing these regions based on whether they contain putative origins of replication. Grey arrows indicate the presumed direction of replication progression, whereas blue and red arrows indicate the direction of transcription, with red indicating transcription on the reverse (bottom) strand, whereas blue – on the forward (top) strand. Ori – origin-containing boundaries, non-ori – boundaries that don't contain origins of replication, cTTS – convergent transcription termination sites, dTSS – divergent transcription start sites. PTU boundaries that do not contain origins of replication were further divided based on whether they contained histone 3 (H3) acetylation ('ac' or 'ac total') or base J ('J') peaks in Lombraña *et al.*, (2016).

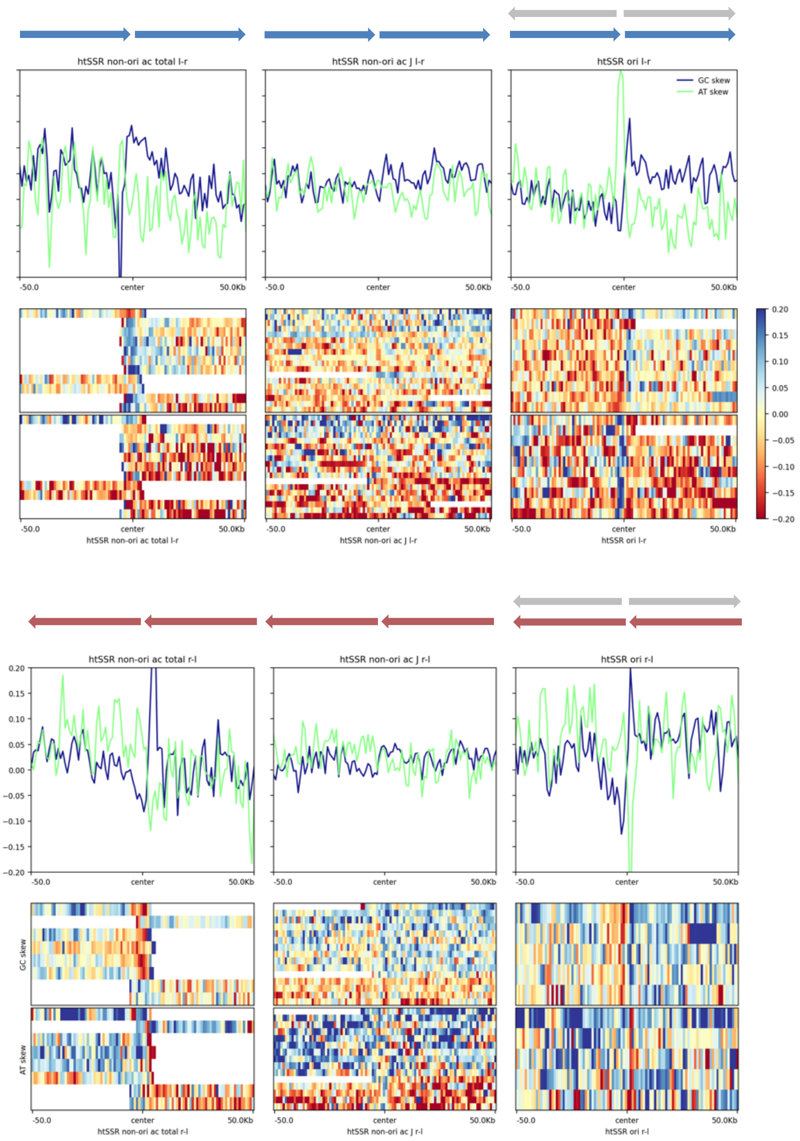


Figure 90 Nucleotide skews at head-to-tail PTU boundaries of L. major.

Nucleotide skews at head-to-tail polycistronic transcription unit (PTU) boundaries, with ± 50kb flanking regions, were plotted for *L. major,* further dividing these regions based on whether they contain putative origins of replication. Grey arrows indicate the presumed direction of replication progression, whereas blue and red arrows indicate the direction of transcription, with red indicating transcription on the reverse (bottom) strand, whereas blue – on the forward (top) strand. Ori – origin-containing boundaries, non-ori – boundaries that don't contain origins of replication, htSSR – head-to-tail PTU boundary, l-r – left-to-right, r-l – right-to-left, referring to the direction of flanking PTU transcription. PTU boundaries that do not contain origins of replication were further divided based on whether they contained histone 3 (H3) acetylation ('ac' or 'ac total') or base J ('J') peaks in Lombraña *et al.*, (2016).

5.2.5 Coding vs non-coding sequence nucleotide skews

To investigate whether nucleotide skews in *L. major* and *T. brucei* differ in coding versus non-coding sequences, we plotted these separately for both origin-containing and non-origin-containing PTU boundaries (Figure 91 for *T. brucei* and Figure 92 for *L. major*). As above, data at convergent and divergent PTU boundaries appeared less noisy and more easily interpreted; for GC skews in *T. brucei*, the overall pattern of positive skews in both coding and non-coding sequences could be seen, with slightly different signal pattern between the two (Figure 91).

AT skews, on the other hand, were much more pronounced in non-coding compared to the coding sequences. Furthermore, at convergent transcription termination sites (cTTS) they too showed opposite patterns - higher AT skew values were seen on the forward strand of coding sequences compared to the reverse strand, while the opposite pattern was seen in non-coding sequences. At dTSS, coding AT skew did not change noticeably at the PTU boundary, however, non-coding AT skew displayed a sharp polarity switch at the boundary.

The *L. major* genome showed even more dramatic differences in coding versus non-coding sequence nucleotide skews (Figure 92). Not only were the non-coding sequence skews more pronounced, they also displayed opposite patterns in all but one case - cTTS GC skew. In fact, our previous assertion that *L. major* shows negative GC and AT skews with transcription direction only appears to hold true in non-coding sequence.

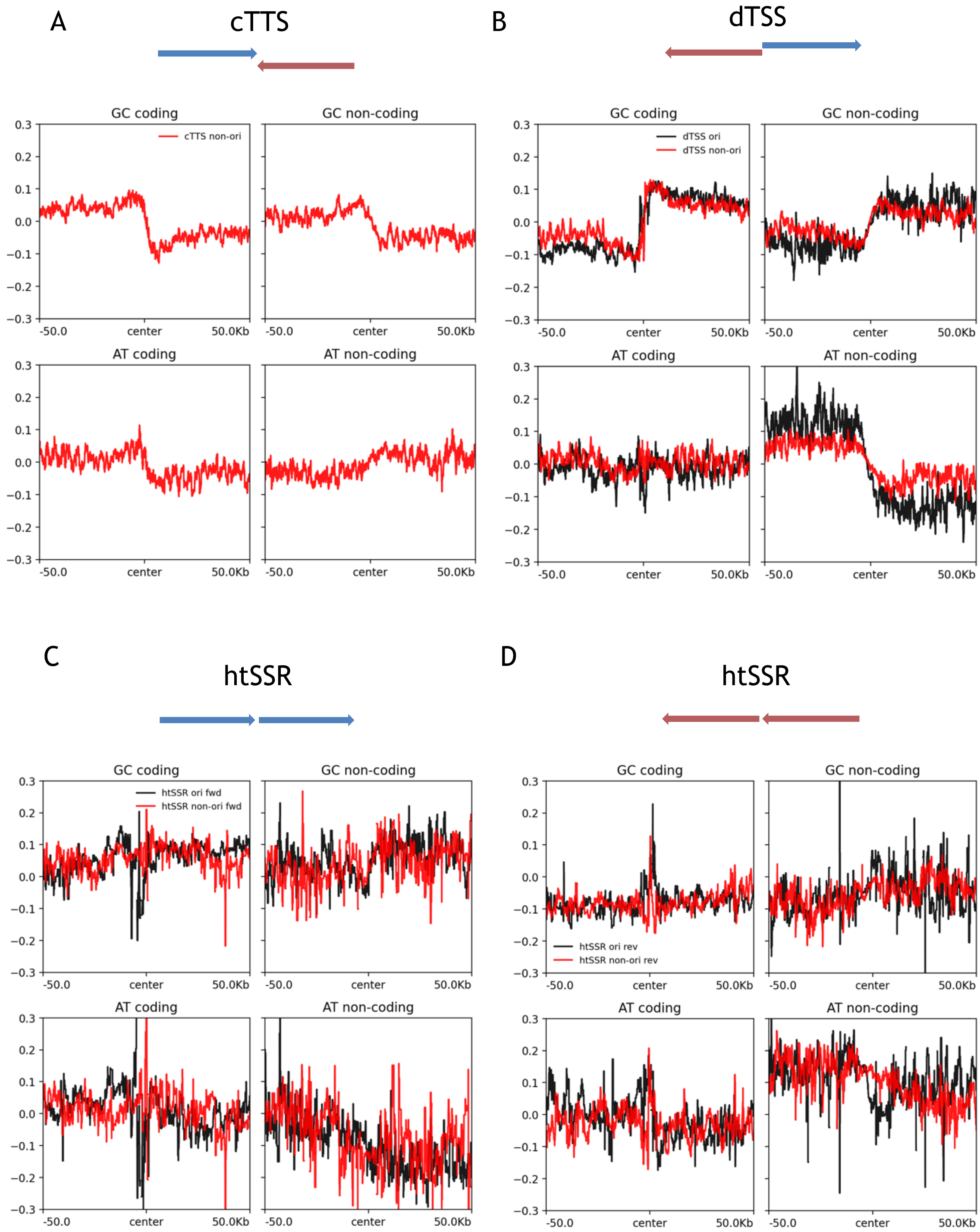


Figure 91 Coding and non-coding AT and GC skews at *T. brucei* PTU boundaries.

AT and GC nucleotide skews were calculated separately for *T. brucei* coding and non-coding sequences, and these were plotted at convergent, divergent and head-to-tail polycistronic transcription unit (PTU) boundaries ± 50kb of flanking sequence. The PTU boundaries were further divided based on whether they contain putative replication origin sites. A - cTTS – convergent transcription termination site, B - dTSS – divergent transcription start site, C and D - htSSR – head-to-tail PTU boundary. Ori – origin-containing site, non-ori – PTU boundary that does not contain a replication origin, fwd – flanking transcription units are on the forward (top) strand (C), rev – flanking transcription units are on the reverse (bottom) strand (D).

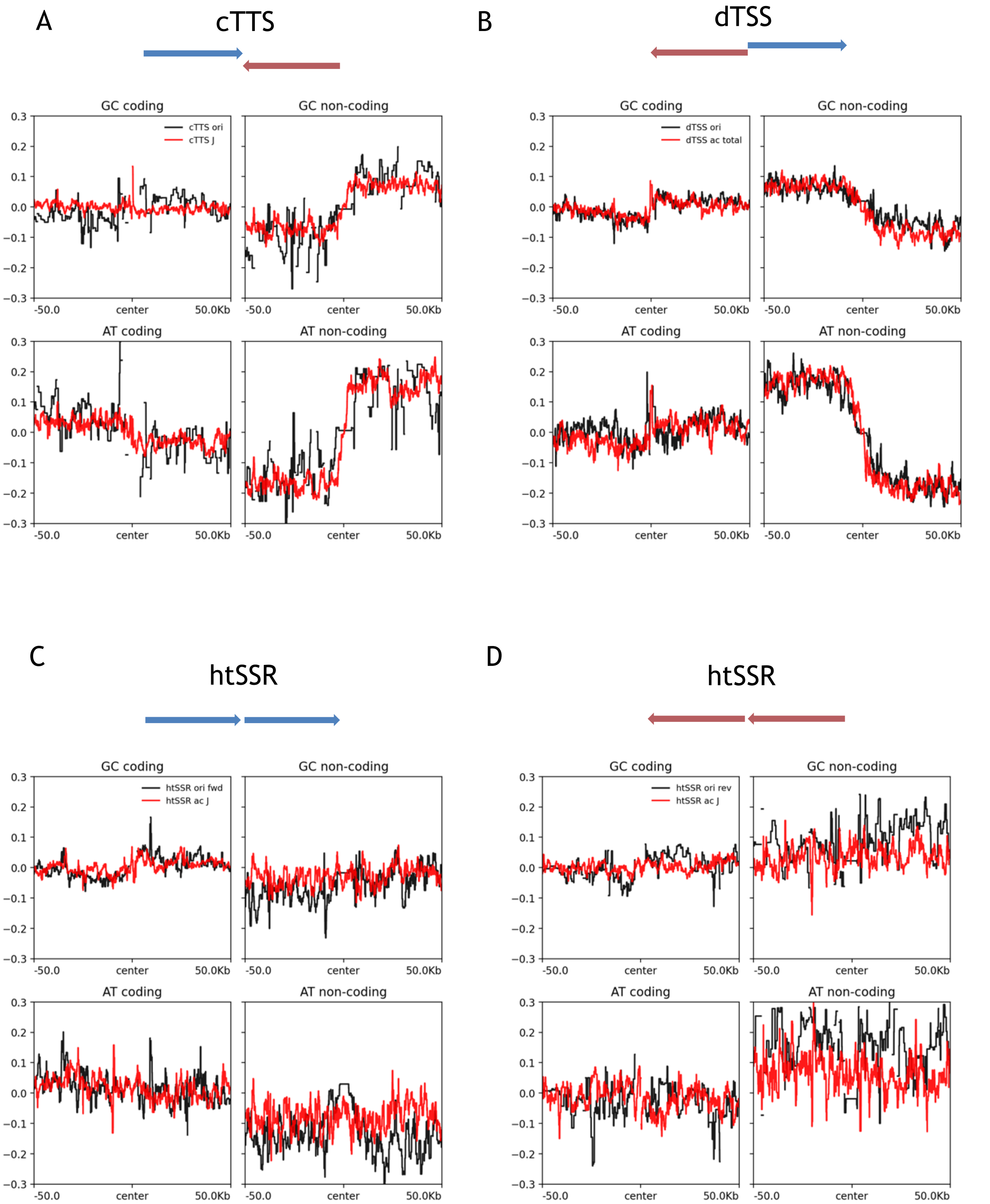


Figure 92 Nucleotide skews of coding and non-coding sequences at L. major PTU boundaries.

AT and GC nucleotide skews were calculated separately for *L. major* coding and non-coding sequences, and these were plotted at convergent, divergent and head-to-tail polycistronic transcription unit (PTU) boundaries ± 50kb of flanking sequence. The PTU boundaries were further divided based on whether they contain putative replication origin sites. A - cTTS – convergent transcription termination site, B - dTSS – divergent transcription start site, C and D - htSSR – head-to-tail PTU boundary. Ori – origin-containing site, non-ori – PTU boundary that does not contain a replication origin, fwd – flanking transcription units are on the forward (top) strand (C), rev – flanking transcription units are on the reverse (bottom) strand (D). PTU boundaries that do not contain origins of replication were further divided based on whether they contained histone 3 (H3) acetylation (‘ac’ or ‘ac total’) or base J (‘J’) peaks in Lombraña *et al.*, (2016).

5.2.6 Genome-wide patterns in *L. major* that change with chromosome size

Previously, *Leishmania spp.* parasites have displayed genome-wide patterns that vary with chromosome size. In *L. mexicana*, long-term passage *in vitro* resulted in increase in copy number of smaller chromosomes and decrease in copy number of larger ones (Campbell, 2019) and in *L. major*, DNA replication timing appears to be influenced by chromosome size, with smaller ones replicating earlier than larger ones (Damasceno *et al.*, 2020). Furthermore, R-loop (DRIPseq) and SNSseq (mapping nascent strands of DNA replication) (Lombraña *et al.*, 2016) also display chromosome-size dependent patterns (Damasceno *et al.*, 2024b).While the reason for the chromosome size-dependant patterns is unclear, we decided to investigate whether nucleotide composition patterns or skews also vary with chromosome size.

First, looking at GC fraction patterns genome-wide in both parasites, a statistically significant (p-value < 0.0001) decrease in GC content with increased chromosome size could be seen in *L. major* (Figure 93 A). In *T. brucei*, while the effect size was smaller (R^2=0.6017 compared to R^2=0.7020 in *L. major*), an association between chromosome size and GC content (p-value 0.005) could be seen, but the relationship was reversed – the larger chromosomes showed higher GC content compared to the smaller ones (Figure 93 B). For nucleotide skew values, on the other hand, chromosome size was responsible for a smaller fraction of the variation (Figure 94), and the values were statistically significant only in *L. major* (Figure 94 A). To see if R-loop data shows chromosome size-dependant accumulation in *T. brucei*, as it does in *L. major* (Damasceno *et al.*, 2024b), both datasets were plotted (Figure 95); while there did appear to be a lesser relationship in *T. brucei* that was, again, opposite of that seen in *L. major* (R^2=0.3196 in *T. brucei* vs R^2=0.7607 in *L. major*), it was not found to be statistically significant (p-value 0.0699).

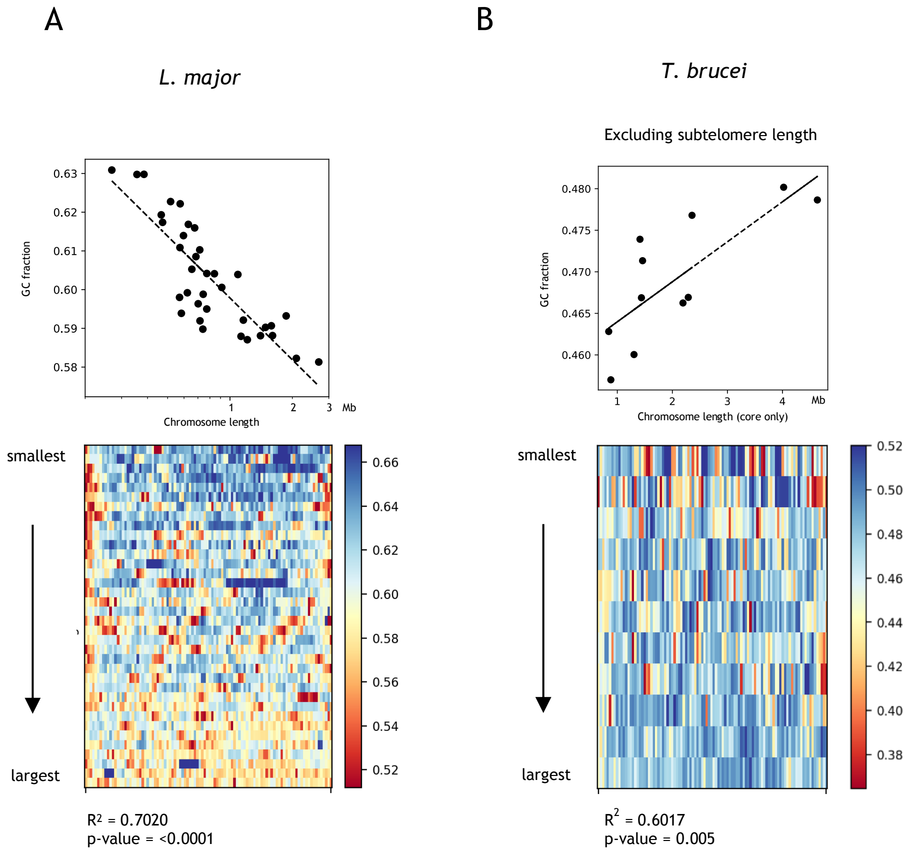


Figure 93 Sequence GC fraction with increasing chromosome size in *L. major* (A) and *T. brucei* (B).

Average (mean) GC fraction was calculated for each chromosome in *L. major* (A) and *T. brucei* megabase chromosome core regions (B), and a linear regression model was applied using python package sklearn module linear_model (linear regressor 'LinearRegression'); R^2 and p-values were obtained using the same package. Heatmap plots were generated using deepTools' computeMatrix and plotHeatmap with sorting of chromosomes (core region only for *T. brucei*) from smallest to largest.

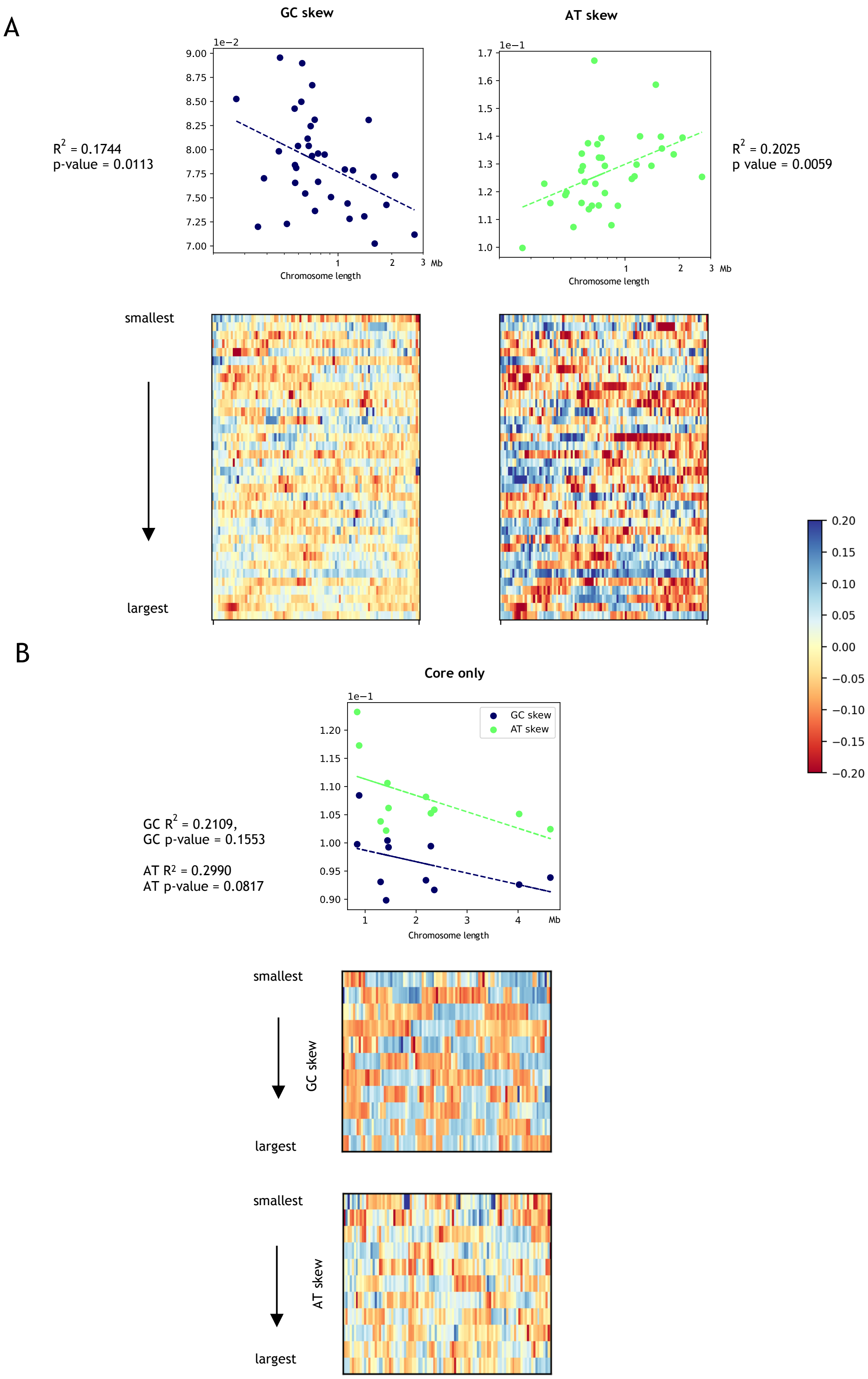


Figure 94 Nucleotide skews with increasing chromosome size in *L. major* (A) and *T. brucei* (B).

Average absolute AT and GC skew values were calculated for each chromosome in *L. major* (A) and *T. brucei* megabase chromosome core regions (B), and a linear regression model was applied using python package sklearn module linear_model (linear regressor 'LinearRegression'); R^2 and p-values were obtained using the same package. Heatmap plots were generated using deepTools' computeMatrix and plotHeatmap with sorting of chromosomes (core region only for *T. brucei*) from smallest to largest.

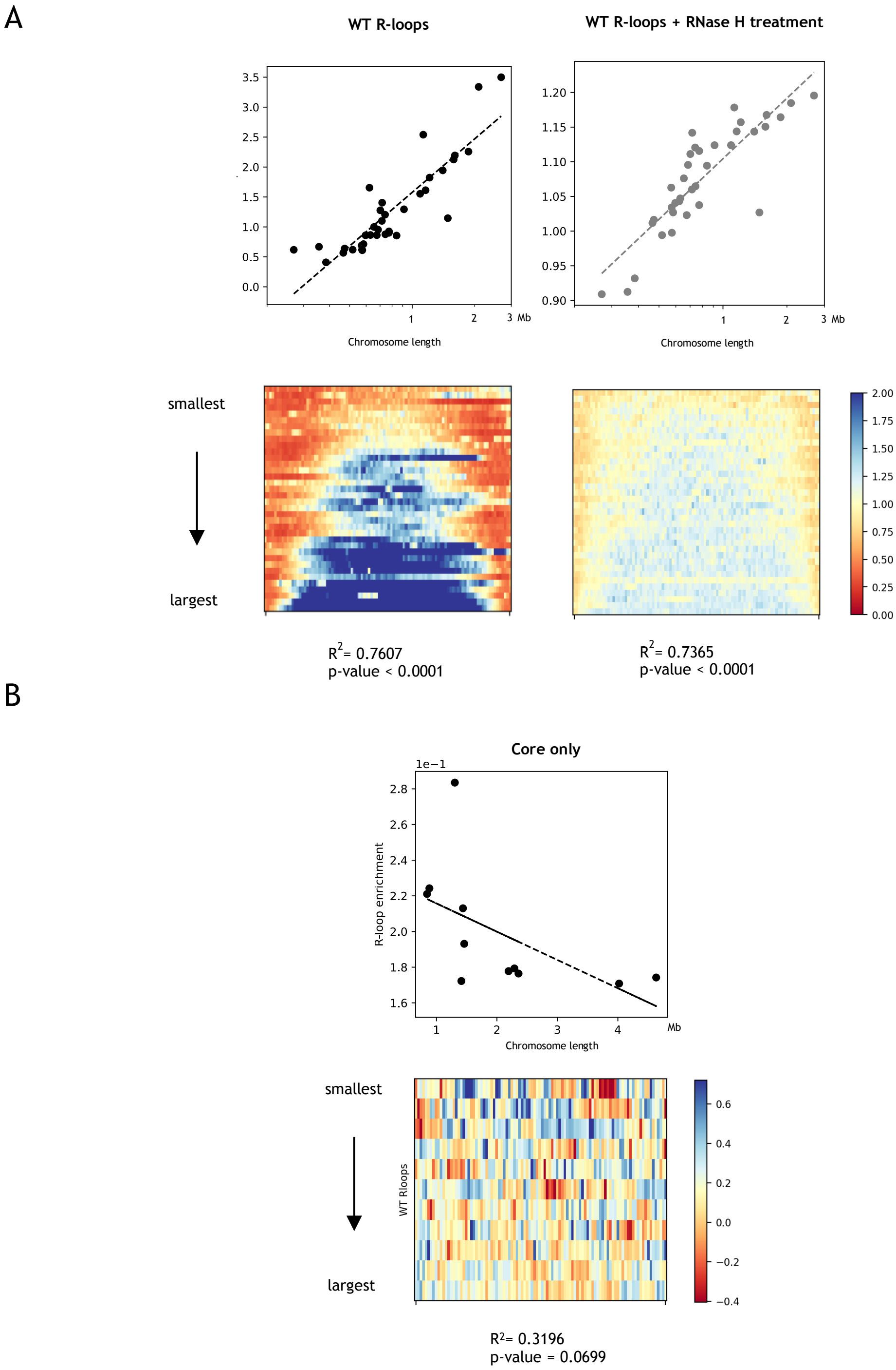


Figure 95 R-loop abundance varies with chromosome size in *L. major* (A) and *T. brucei* (B).

Average R-loop abundance values were calculated for each chromosome in *L. major* (A) and *T. brucei* megabase chromosome core regions (B), and a linear regression model was applied using python package sklearn module linear_model (linear regressor 'LinearRegression'); R^2 and p-values were obtained using the same package.. Heatmap plots were generated using deepTools' computeMatrix and plotHeatmap with sorting of chromosomes (core region only for *T. brucei*) from smallest to largest.

5.2.7 Coding and inter-CDS sequence skews across trypanosomatids more broadly

Lastly, to put the *T. brucei* and *L. major* observations into a wider phylogenetic context, nucleotide composition and skews were also examined across other available trypanosomatid genomes. First, we identified which trypanosomatid genomes are available on TriTrypDB and Genbank (Table 23), and annotated those that did not have associated annotation files using companion (Steinbiss *et al.*, 2016). Following this, nucleotide skew and content tracks were generated using bedtools; this was also done for *Bodo saltans* – a non-trypanosomatid kinetoplastid protozoan that is commonly used as an outgroup (Jackson, Quail and Berriman, 2008; Jackson *et al.*, 2016).

Nucleotide composition, in the form of coding and overall GC %, showed that the change in genome nucleotide content when looking at *T. brucei* and *L. major* is broadly reflected in the gradual GC content changes throughout the phylogenetic tree of trypanosomatids (Figure 96). With regards to nucleotide skews in CDS and inter-CDS sequences, broadly speaking, *Leishmania* species, along with closely related species such as those of *Crithidia*, *Porcisia* and *Leptomonas* genera (all belonging to the subfamily Leishmaniinae), among others, consistently showed negative AT and GC skews in both CDS and inter-CDS sequences, whereas *Trypanosoma* species showed positive GC skew in both (Figure 97). With regards to the other, lesser-studied trypanosomatid genera, the picture is more mixed and less clear.

Paratrypanosoma confusum, an early-branching trypanosomatid thought to represent an evolutionary link between the free-living non-trypanosomatid bodonids, such as *Bodo saltans*, and the parasitic trypanosomatids (Flegontov *et al.*, 2013) (Figure 77), displays positive GC skew (similar to *Trypanosoma* species) and mixed AT skew – positive in the CDS sequences, and negative in the inter-CDS sequences. *Bodo saltans*, on the other hand, showed less pronounced skew values, all of which were negative, reminiscent more of skews in Leishmaniinae.

Figure 96 Genome GC composition in trypanosomatid and *Bodo saltans* genomes. Coding and overall genome nucleotide composition, in the form of % GC; data obtained from companion (Steinbiss *et al.*, 2016) annotation output.

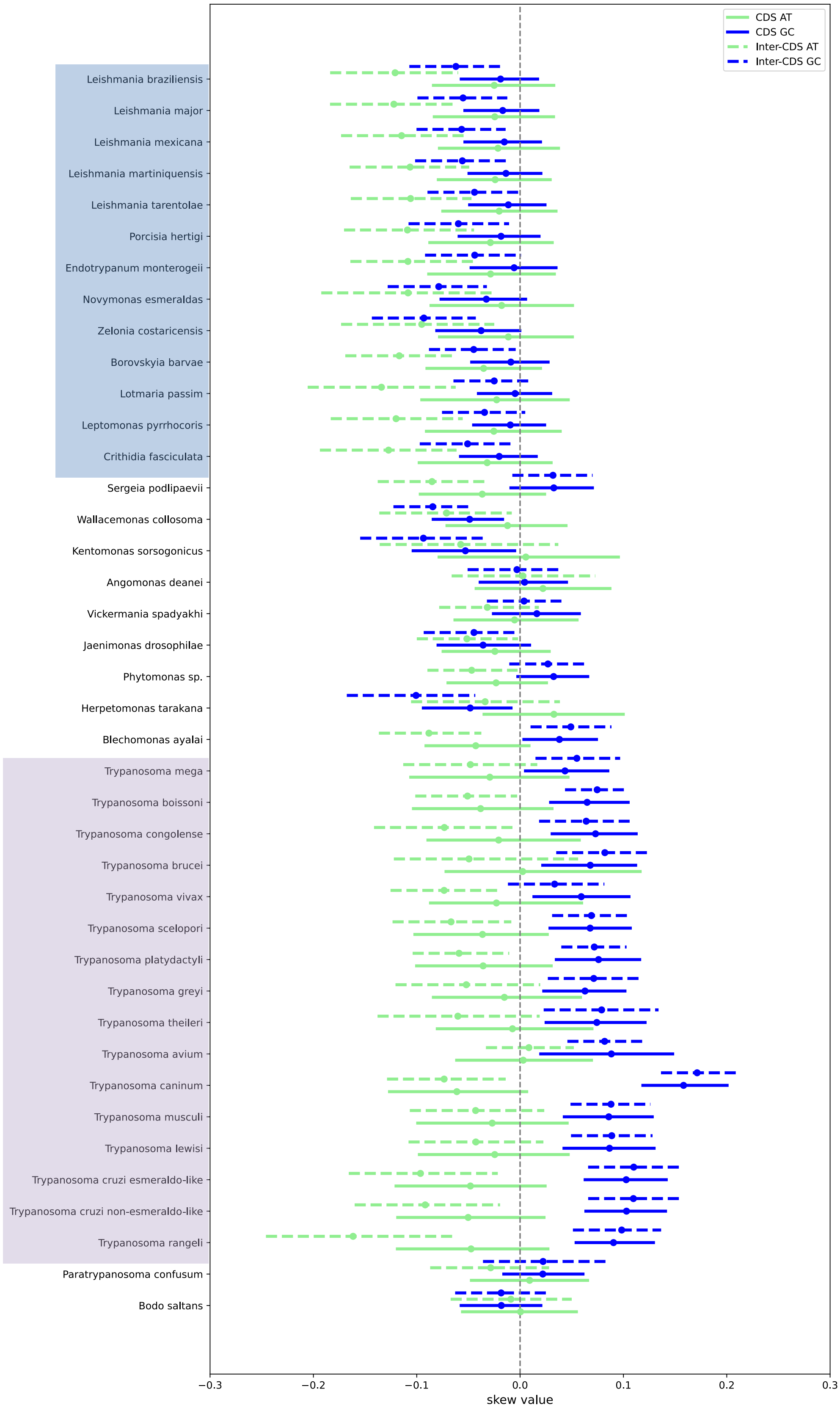


Figure 97 Nucleotide skews in CDS and inter-CDS regions across trypanosomatid and *Bodo saltans* genomes.
The values represent AT or GC skews in either CDS or inter-CDS regions, based on annotation gff file data. The range represents the interquartile range with the median values highlighted as dots. Inter-CDS regions were only included where neighbouring CDS annotations are on the same strand.

5.3 Discussion

For clarity, a summary of core findings in this chapter can be found in Table 17, Table 18 and Figure 99 below.

5.3.1 Divergent skew patterns in *T. brucei* and *L. major*

In this chapter, we described the nature and attempted to elucidate the potential contributors of nucleotide patterns and skews present in the genomes of *Trypanosoma brucei* and *Leishmania major*. Despite the high levels of synteny between the two parasites, their evolutionary divergence is ancient (Figure 77), and this is, arguably, evident in the nucleotide composition of their genomes (Table 17).

Table 17 Summary of AT and GC skew patterns, as well as G-quadruplex enrichment, in *T. brucei* and *L. major* genomes.

'+' – positive skew values, '-' – negative skew values, BES – bloodstream-form expression site, N/A – not applicable. * - skew value just above 0.

	T. brucei		*L. major*	
	GC	AT	GC	AT
Transcription				
Coding strand	+	-	-	-
Template strand	-	+	+	+
Replication				
Lagging	+	-	+	-
Leading	-	+	-	+
Sequence function				
Coding sequence	+	+ *	+	+
Intergenic sequence	+	-	-	-
Genomic compartments				
Core	+	-	N/A	
Subtelo	+	+		
BES	+	+		
G-quadruplex enrichment	Template strand		Coding strand	

Coding and template strand nucleotide skews are the most apparent in these kinetoplastids, with an overabundance of G and T on the coding strand for *T. brucei* and C and T for *L. major*. While the same holds true for intergenic sequence, coding DNA displays distinct patterns – both GC and AT skews are positive in both parasites. This represents a particularly striking AT skew change when comparing coding and intergenic sequence. Overall, AT skew in both genomes and GC skew in *L. major* are more pronounced in intergenic DNA compared to coding DNA, which also explains why the coding strand skews match the intergenic sequence rather than coding sequence.

Transcription-associated skews in *T. brucei* match the replication-associated skews in both parasites, which are very similar between the two kinetoplastids, with only the *L. major* transcription-associated GC skew displaying a divergent pattern. Codon preference does not explain this departure, as GC skew deviation in *L. major* is seen in intergenic regions, rather than stemming from coding DNA. G-quadruplex mapping showed clear strand asymmetry in both parasites, enriched on the template strand in *T. brucei* and coding strand of *L. major*. This pattern follows the opposite GC skews observed in these protozoa; however, to our surprise, in both cases the G-quadruplex enrichment is found on the stand of DNA that contains fewer G residues compared to Cs. Whether G-quadruplex enrichment on the G-poor strand is representative of underlying biology, or, perhaps (and more likely in our view), a mistake in the data deposited to Sequence Read Archive (SRA) (Marsico *et al.*, 2019), is unclear.

In *T. brucei*, subtelomeric and BES regions display divergent AT skew patterns relative to that of the core genome – the skew is overall positive at subtelomeres and BESs, and negative in the core genome. *VSG* gene coding sequence is very strongly AT-skewed but displays almost no GC skew; it appears that the GC skew of the subtelomeres originates from the intergenic sequence, whereas the AT skew arises from the *VSG* coding sequence.

5.3.2 Processes that shape *T. brucei* and *L. major* genome nucleotide composition

Overall, the transcription-associated nucleotide skews in *T. brucei* are consistent with what is typically found in eukaryotes, where cytosine deamination on the single-stranded coding strand results in C>T mutations, contributing to positive GC skew and negative AT skew (Jinks-Robertson and Bhagwat, 2014). Whether or not this is, indeed, a significant contributor to the skew patterns observed would have to be determined by further experimental investigation. In *L. major*, unlike *T. brucei*, the coding strand shows a negative GC skew – overabundance of cytosine residues over guanine; this overall pattern appears to stem from intergenic sequences, as coding sequences show positive GC skew in this parasite. The reason for this deviation is unknown, but we discuss a couple of speculative suggestions below.

First, a difference in transcription-associated skew patterns points to differential mutational and/or selective pressures that act on transcription. Overall, on the coding strand and in intergenic sequences, specifically, GC skew is negative in *L. major* - this is contrary to what is found in *T. brucei* and cannot be explained by C>T deamination on ssDNA of the coding strand during transcription. In addition, G-quadruplexes (G4s) appear to accumulate on the coding strand. It is possible that coding strand secondary structure formation affects the accessibility of this DNA strand to destabilizing agents, as well as binding of various factors. This could affect the accumulation C>T of mutations, as well as the ability to repair any DNA lesions that occur. It may also lead to a different mutational profile. Furthermore, perhaps G4 accumulation carries a selective advantage to the parasite, and the maintenance of a negative GC skew is needed to sustain these secondary DNA structures. What such an advantage might be is unclear.

Second, C>T mutation is a multi-step process with varying outcomes (Owiti, Stokdyk and Kim, 2019). Cytosine is first methylated to 5-methylcytosine, and its deamination leads to the formation of thymidine. The resulting T:G basepair is then repaired either by replacing T, thereby restoring the C:G basepair, or the G, resulting in a T:A basepair – the latter is what typically leads to positive GC and negative AT skews on the coding strand (Jinks-Robertson and Bhagwat, 2014). Cytosine can also become deaminated without the initial methylation, as cytosine hydrolysis leads to deamination, forming uracil (Figure 98); as uracil isn’t a canonical constituent of DNA, it is typically removed using dedicated enzymatic base excision repair (BER) mechanisms, restoring the C:G basepair (Owiti, Stokdyk and Kim, 2019). However, if this repair is not carried out, the resulting U:G may lead to U:A basepairing, and in subsequent rounds of replication – T:A. Uracil can also become incorporated in DNA instead of thymidine, as DNA polymerases can’t always distinguish the two (Owiti, Stokdyk and Kim, 2019). Both of these uracil incorporation scenarios typically lead to BER, which may result in an abasic site replacing the U and this can lead to further genomic instability (Sedwick, Brown and Glickman, 1986; Guillet, Van Der Kemp and Boiteux, 2006).

While the consequences harbouring genomic uracil and/or abasic sites vary (Figure 98), they can lead to the involvement of DNA repair-associated translesion DNA polymerases – error-prone enzymes that often incorporate particular dNTPs in place of an abasic site and allow replication to progress (reviewed in Jain, Aggarwal and Rechkoblit, (2018)). To our knowledge, translesion DNA polymerases have not been studied in *L. major*, it is possible that the parasite resolves abasic or U sites using a translesion polymerase that, through its error-prone repair activity (Castillo-Acosta *et al.*, 2012; Boiteux and Jinks-Robertson, 2013; Sakofsky *et al.*, 2015), leads to the divergent GC skew pattern.

Other likely contributors might be considered – the effect of base J DNA modification (see Chapter 4) and differential TC-NER activity on coding and template strands in *L. major*, but little is known about either in relation to DNA damage and repair in this parasite.

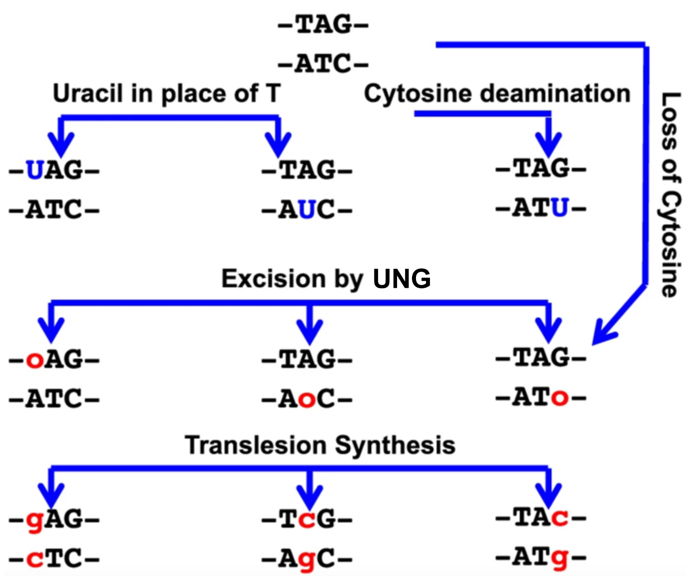


Figure 98 Consequences of cytosine deamination and uracil incorporation in DNA.

UNG – unspecified uracil-DNA glycosylase, o – abasic site. Figure adapted from Owiti, Stokdyk and Kim, (2019), reproduced under Open Access, Creative Commons .

5.3.3 Nucleotide composition in *L. major* and *T. brucei* varies with chromosome size

In addition to transcription- and replication-associated skews in *T. brucei* and *L. major*, nucleotide patterns can also be observed varying with chromosome size (Table 18).

Table 18 Summary of nucleotide and secondary DNA structure patterns that change with increasing chromosome size in *T. brucei* and *L. major.*

↑ - increases with chromosome size, ↓ - decreases with chromosome size. * - based on data analysed by Jeziel Damasceno (Damasceno *et al.*, 2024b).

	T. brucei		*L. major*	
GC skew	↓	R^2 = 0.2109, p-value = 0.1553	↓	R^2 = 0.1744 p-value = 0.0113
AT skew	↓	R^2 = 0.2990 p-value = 0.0817	↑	R^2 = 0.2025 p value = 0.0059
GC fraction	↑	R^2 = 0.6017 p-value = 0.005	↓	R^2 = 0.7020 p-value = <0.0001
R-loops	↓	R^2= 0.3196 p-value = 0.0699	↑	R^2= 0.7607 p-value < 0.0001*
G4-quadruplexes			↓	*

These chromosome size-dependant features are currently being further analysed and investigated by Jeziel Damasceno. As above, we don't know what factors may influence chromosomes differentially in a size-dependant manner. Considering the difference in replication timing between smaller and larger chromosomes of *L. major* (Damasceno *et al.*, 2020), we speculate that that the observed chromosome size-dependant nucleotide and secondary structures might either dictate or stem from differential replication timing, which there might be selective pressure for. Alternatively, these size-dependant patterns might reflect the spatial organisation of the nucleus: perhaps chromosomes are spatially separated in the nucleus based on size as there is some advantage for such organisation.

Without further work, our speculations on this matter remain just that. Unfortunately, nucleotide and base excision repair mechanisms and the activity of DNA translesion polymerases remain poorly described in both *T. brucei* and *L. major;* more broadly, transcription machinery and dynamics remain incompletely resolved (reviewed in Clayton, (2019)). Despite the well-accepted dogma that these parasites initiate the majority of their transcription from TSSs at PTU boundaries (Clayton, 2019), it remains to be understood how RNA polymerases progress across the template DNA, how any encountered DNA lesions are dealt with and what other mutational and selective pressures affect the evolution of their genomes.

5.3.4 Phylogenetic context of trypanosomatid nucleotide skews

To put the observed skew differences in *T. brucei* and *L. major* into phylogenetic context, we analysed coding sequence (CDS) and inter-CDS skews across 37 more trypanosomatid genomes, as well as *Bodo saltans*, a free-living non-trypanosomatid kinetoplastid (Figure 97 and summary Figure 99). The broad consensus the analysis showed is that genera belonging to the Leishmaniinae subfamily (*Leishmania*, *Crithidia*, *Leptomonas*, *Porcisia* and others) showed negative AT and GC skews, whereas *Trypanosoma* species showed positive GC skews; other subfamilies and genera showed more varied coding sequence skew patterns. Our results support the conclusion that the skew differences observed in *T. brucei* and *L. major* are reflective of their evolutionary distance.

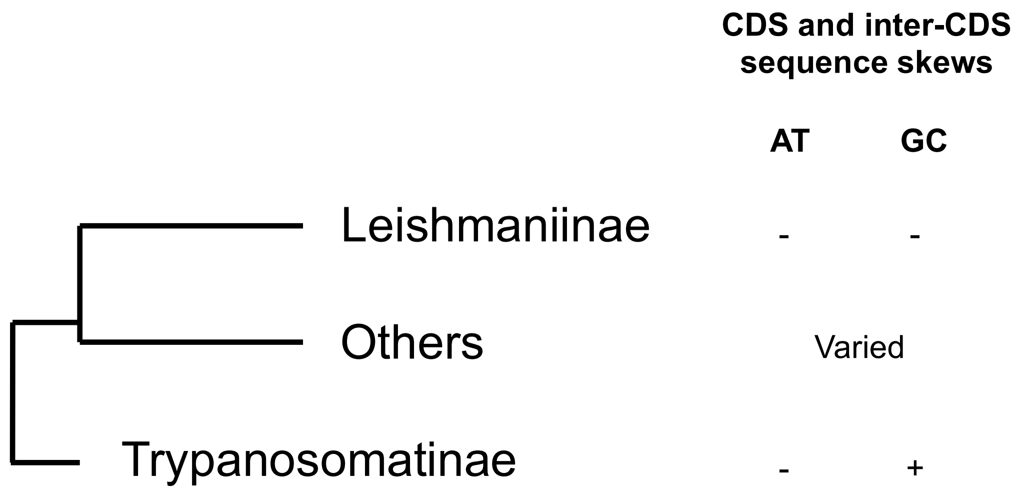


Figure 99 Summary of CDS and inter-CDS nucleotide skews observed across analysed trypanosomatid genomes.

Unfortunately, DNA replication-related or more broadly transcription-related skews cannot be discerned in other trypanosomatids as the relevant datasets (showing DNA replication dynamics or transcription initiation or termination sites) are not available, and therefore phylogenetic analyses of these traits cannot be performed at this time.

More broadly, taking both the observed nucleotide composition and skews into account (Figure 96, Figure 97), these traits are variable between Trypanosomatid genera, but more similar among congeneric species (at least for *Leishmania* and *Trypanosoma*). A departure from the other genera can be seen in *Trypanosoma* in terms of GC skew of both coding and inter-CDS sequences, as well as coding and overall GC content, suggesting that, early in trypanosomatid evolution, a variation in some aspect of genome stability maintenance occurred in *Trypanosoma* and became fixed, leading to the divergent GC skews and composition we see today.

Appendices

Table 19 Overlaps between chromosome 5 contig tig00000084 in the new assembly and the reference genome.

Chromosome 5 contig overlaps highlighted in bold, gap identity (refer to Figure 21) indicated in right-most column.

#	START	END	STRAND	REFERENCE CONTIG	SIZE, BP	START	END	GAP
1	**19**	**48608**	**+**	**Chr5_core_Tb427v9**	**1443501**	**1370298**	**1419315**	-
2	42523	74994	+	unitig_1923_Tb427v9	32483	29	32471	Gap1
3	**72242**	**236318**	**+**	**Chr5_3B_Tb427v9**	**374853**	**18**	**163935**	-
4	236512	242485	-	Chr9_5B_Tb427v9	101244	51338	57217	Gap 2
5	237778	238757	+	Chr4_3B_Tb427v9	285418	3787	4821	
6	238726	240980	+	Chr10_3A_Tb427v9	1616207	815037	817311	
7	241294	242686	+	Chr4_3B_Tb427v9	285418	3787	4826	
8	241800	242791	+	Chr4_3B_Tb427v9	285418	3787	4826	
9	242350	243499	+	Chr4_3B_Tb427v9	285418	3787	4826	
10	**242667**	**245728**	**+**	**Chr5_3B_Tb427v9**	**374853**	**166688**	**169581**	-
11	243166	244267	+	Chr4_3B_Tb427v9	285418	3787	4826	Gap 2
12	243200	244525	+	Chr4_3B_Tb427v9	285418	3782	4826	
13	243451	244603	+	Chr4_3B_Tb427v9	285418	3787	4610	
14	**243480**	**246438**	**+**	**Chr5_3B_Tb427v9**	**374853**	**166688**	**169247**	-
15	**245353**	**246270**	**+**	**Chr5_3B_Tb427v9**	**374853**	**168557**	**169422**	-
16	**246419**	**247497**	**+**	**Chr5_3B_Tb427v9**	**374853**	**168076**	**169422**	-
17	**247064**	**248415**	**+**	**Chr5_3B_Tb427v9**	**374853**	**168076**	**169247**	-
18	**248396**	**249505**	**+**	**Chr5_3B_Tb427v9**	**374853**	**168076**	**169247**	-
19	**249486**	**250549**	**+**	**Chr5_3B_Tb427v9**	**374853**	**168076**	**169247**	-
20	**250530**	**251491**	**+**	**Chr5_3B_Tb427v9**	**374853**	**168076**	**169247**	-
21	**251472**	**252502**	**+**	**Chr5_3B_Tb427v9**	**374853**	**168076**	**169212**	-
22	**252583**	**253575**	**+**	**Chr5_3B_Tb427v9**	**374853**	**168144**	**169225**	-
23	**253647**	**254605**	**+**	**Chr5_3B_Tb427v9**	**374853**	**168144**	**169247**	-
24	**254586**	**373082**	**+**	**Chr5_3B_Tb427v9**	**374853**	**168076**	**286621**	-
25	357738	392183	-	Chr7_5A_Tb427v9	863729	474840	509238	Gap 3
26	**391395**	**478457**	**+**	**Chr5_3B_Tb427v9**	**374853**	**287636**	**374843**	-
27	478780	480965	+	Chr6_3B_Tb427v9	1075594	412016	414201	Gap 4
28	478805	480965	+	Chr1_3A_Tb427v9	2154586	311770	313936	
29	480981	484125	+	Tb427VSG-23_containing_unitig_Tb427v9	24484	19201	22262	
30	480981	482711	+	Tb427VSG-23_containing_unitig_Tb427v9	24484	19491	21258	
31	480981	482711	+	Tb427VSG-23_containing_unitig_Tb427v9	24484	19317	21113	
32	480981	482711	+	Tb427VSG-23_containing_unitig_Tb427v9	24484	19462	21200	
33	480981	482711	+	Tb427VSG-23_containing_unitig_Tb427v9	24484	19781	21432	
34	480981	482711	+	Tb427VSG-23_containing_unitig_Tb427v9	24484	19288	21026	
35	480981	482682	+	Tb427VSG-23_containing_unitig_Tb427v9	24484	19172	20834	
36	480981	482711	+	Tb427VSG-23_containing_unitig_Tb427v9	24484	19027	20718	
37	480981	482711	+	Tb427VSG-23_containing_unitig_Tb427v9	24484	19694	21345	

38	480981	482711	+	Tb427VSG-23_containing_unitig_Tb427v9	24484	19955	21605	
39	480981	482711	+	Tb427VSG-23_containing_unitig_Tb427v9	24484	18882	20554	
40	480981	482711	+	Tb427VSG-23_containing_unitig_Tb427v9	24484	18679	20351	Gap 4
41	480981	482711	+	Tb427VSG-23_containing_unitig_Tb427v9	24484	18853	20525	
42	480981	482711	+	Tb427VSG-23_containing_unitig_Tb427v9	24484	19085	20747	
43	480981	482711	+	Tb427VSG-23_containing_unitig_Tb427v9	24484	18737	20380	
44	480981	482711	+	Tb427VSG-23_containing_unitig_Tb427v9	24484	20071	21692	
45	480981	482682	+	Tb427VSG-23_containing_unitig_Tb427v9	24484	18969	20554	
46	480981	482569	+	Tb427VSG-23_containing_unitig_Tb427v9	24484	18795	20322	
47	480981	482569	+	Tb427VSG-23_containing_unitig_Tb427v9	24484	18911	20438	
48	480981	482569	+	Tb427VSG-23_containing_unitig_Tb427v9	24484	20100	21692	
49	480981	482513	+	Tb427VSG-23_containing_unitig_Tb427v9	24484	19143	20583	
50	480981	482398	+	Tb427VSG-23_containing_unitig_Tb427v9	24484	20245	21692	
51	482379	483568	+	Tb427VSG-717_containing_unitig_Tb427v9	14778	13431	14562	
52	482550	484125	+	BES2_Tb427v9	82592	80692	82272	
53	483703	484967	+	BES17_Tb427v9	86347	79568	80842	

Table 20 Gap closure in the ONT assembly relative to the 2018 *T. brucei* Lister 427 genome

	Reference contig	Start	Assembly contig	Closed?	Notes
1	BES17_Tb427v10	65823	tig00000116	closed	
2	BES2_Tb427v10	62606	tig00000156; tig00000016	ambiguous	
3	Chr1_3A_Tb427v10	43105	tig00000069	closed	
4	Chr1_3A_Tb427v10	202156	tig00000069	closed	
5	Chr1_3A_Tb427v10	264716	tig00000069; tig00004860	closed	closed in tig69; in tig4860, two subtelomeric fragments bridged
6	Chr1_3A_Tb427v10	1409580	tig00004860	closed	unitig and chromosome 6 subtelomere 3a fragment inserted
7	Chr1_3A_Tb427v10	2090279	tig00004860	closed	
8	Chr1_3B_Tb427v10	226721	tig00000101	closed	
9	Chr1_3B_Tb427v10	299541	tig00000101	closed	
10	Chr1_3B_Tb427v10	372293	tig00000101	closed	
11	Chr1_core_Tb427v10	609279	tig00000654; tig00000069	closed	
12	Chr10_3A_Tb427v10	1256835	tig00000082	closed	
13	Chr11_3A_Tb427v10	784536	tig00000152	closed	
14	Chr11_3A_Tb427v10	284701	tig00004874	closed	
15	Chr11_3A_Tb427v10	497646	tig00004874	closed	
16	Chr11_3B_Tb427v10	266410	tig00000071	closed	
17	Chr11_3B_Tb427v10	319049	tig00000133	closed	
18	Chr11_core_Tb427v10	263638	tig00000051	closed	
19	Chr3_5A_Tb427v10	303597	tig00000642; possibly tig00000069	closed	
20	Chr3_core_Tb427v10	820374	tig00000642	closed	
21	Chr4_core_Tb427v10	879615	tig00000058	closed	
22	Chr5_3A_Tb427v10	456755	tig00000055?	ambiguous	
23	Chr5_3A_Tb427v10	378648	tig00000055	closed	
24	Chr5_3B_Tb427v10	167071	tig00000084	closed	
25	Chr5_3B_Tb427v10	286617	tig00000084	closed	
26	Chr5_core_Tb427v10	223826	tig00000055	closed	
27	Chr6_3A_Tb427v10	861837	tig00000069; tig00004878; tig00000101; tig00000139	ambiguous	
28	Chr6_3A_Tb427v10	1057903	tig00000139	closed	
29	Chr6_3A_Tb427v10	1194770	tig00000139	closed	
30	Chr6_3A_Tb427v10	71012	tig00000648	closed	
31	Chr6_3A_Tb427v10	1239269	tig00000139; tig00004877	closed	
32	Chr6_3B_Tb427v10	1000300	tig00000037	closed	
33	Chr7_core_Tb427v10	473368	tig00004860	closed	
34	Chr7_core_Tb427v10	1789177	tig00004860	closed	
35	Chr7_core_Tb427v10	1933353	tig00004860; tig00000127	no	Centromeric repeat region extended in both contigs
36	Chr8_3A_Tb427v10	235236	tig00000069; tig00000030	ambiguous	
37	Chr8_3A_Tb427v10	215801	tig00000642	closed	
38	Chr8_5A_Tb427v10	666111	tig00000644	closed	
39	Chr8_5B_Tb427v10	335655	tig00000036	closed	

40	Chr8_5B_Tb427v10	113435	tig00000036; tig00000192	closed
41	Chr8_core_Tb427v10	2124030	tig00000642	ambiguous
42	Chr8_core_Tb427v10	316608	tig00000644	closed
43	Chr9_3A_Tb427v10	437845	tig00000645	closed
44	Chr9_3A_Tb427v10	617324	tig00000645	closed
45	Chr9_3A_Tb427v10	1341403	tig00000645	closed
46	Chr9_3B_Tb427v10	286774	tig00000168	closed
47	Chr9_5A_Tb427v10	211585	tig00000647	closed
48	Chr9_core_Tb427v10	2009384	tig00000646	closed
49	Chr9_core_Tb427v10	347612	tig00000647	closed

Table 21 Pfam protein domain enrichment on sub-megabase chromosomes relative to other contigs.

A – number of domains found on sub-megabase contigs, B – number of domains found on non-sub-megabase contigs.

NAME	DESCRIPTION	A	B	P-VALUE	LOG2 ODDS RATIO
RHSP	Retrotransposon hot spot protein, P-loop domain	71	447	9.79E-51	4.12478685
RHS_N	Retrotransposon hot spot proteins N-terminal	53	354	1.09E-35	3.79739103
VSG_B	Trypanosomal *VSG* domain	4	2149	3.28E-11	-3.3344238
TRYPAN_GLYCOP_C	Trypanosome variant surface glycoprotein C-terminal domain	0	846	2.51E-06	-13.403278
RNA_POL_RPB2_6	RNA polymerase Rpb2, domain 6	7	37	3.83E-06	3.6785574
ESAG1	ESAG protein	7	46	1.38E-05	3.36338659
RNA_POL_RPB2_4	RNA polymerase Rpb2, domain 4	3	6	0.00026756	5.04636643
EAMA	EamA-like transporter family	2	2	0.00133951	6.03645686
TRYPAN_GLYCOP	Trypanosome variant surface glycoprotein (A-type)	7	1161	0.00480193	-1.4532496
ZF-C3HC4	Zinc finger, C3HC4 type (RING finger)	4	42	0.00503784	2.66067598
DUF1181	Protein of unknown function (DUF1181)	3	34	0.01815943	2.54264729
BT1	BT1 family	2	20	0.04315907	2.71609175

Table 22 Motif sequences of full-length centromeric repeat region candidates.
Consensus motifs identified using tandem repeats finder (Benson, 1999), chromosome assigned based on overlap with reference (Müller *et al.*, 2018). Chr – chromosome.

CONTIG	CHR	MOTIF GROUP	CONSENSUS MOTIF LENGTH, BP	CONSENSUS CORE MOTIF
111	2	1	30	ACACGTAATAACACGCTTTTTCAACATAAA
126	2	1	29	AAAACAAGCAATAACACGATTTTGCACAC
642A	3	2	120	TGTAACAGGGTTTTGTGCGATAACACACCATTATGCGCCGTAGGGGCGA TAACCATGAAACATGGGCCATGATGTGCCACCAATGCATGTGATCAGTG GAAGGGCGTTACCCAATCACAC
58	4	3	145	GTATGATTGTGCAAAAACAGTGTAGCAACGCAACATGTAAAGTGTTTTGT GTAAAACACGCATTCTTGCATAACATGCACAATGTTGCATGTTTGTTGCA ATTGTGCACTATTGCGTATTTTCACGTCAAATACGCGTTTATGC
643	8	3	147	GTATTTTACGTGAAAATACGCAATAGTGCACATTTTGTGTACAAACATGC AACGTTGTGCATGTTGTGCAAGAATGCGTGTTTTAAACCAAAACACGTCA CATGTTGCATTGCAACACTGTTTTTGCACAATCACACGTATGAACGC
82	10	4	39	CAATAATAGGAAATAATGTGCATTTTACGCAATAATGTT
645	9	4	39	GCACATTATGGCTTAAAATGTGCATTATTGCCTATTATT
648	6	4	58	TGCACATTATTGCGTATAATTAAACAAATTGCACATTATTGCGTATAATT GCACAAAT

Table 23 Sources, strains and versions of trypanosomatid and *Bodo saltans* genome sequences used for nucleotide composition and transcription-associated skew analysis.

Species	Strain/version/isolate	Source	Genbank ID
Leishmania braziliensis	MHOM/BR/75/M2904	TriTrypDB-66	
Leishmania major	Friedlin	TriTrypDB-66	
Leishmania mexicana	MHOM/GT/2001/U1103	TriTrypDB-66	
Leishmania martiniquensis	LEM2494	TriTrypDB-66	
Leishmania tarentolae	Parrot-TarII	TriTrypDB-66	
Porcisia hertigi	C119	NIH	GCA_017918235.1
Endotrypanum monterogeii	LV88	TriTrypDB-66	
Novymonas esmeraldas	E262AT.01	NIH	GCA_019188245.1
Zelonia costaricensis	15EC	NIH	GCA_030849075.1
Borovskyia barvae	21EC	NIH	GCA_030849085.1
Lotmaria passim	422	NIH	GCA_034478905.1
Leptomonas pyrrhocoris	H10	TriTrypDB-66	
Crithidia fasciculata	Cf-Cl	TriTrypDB-66	
Sergeia podlipaevii	CER4	NIH	GCA_030849805.1
Wallacemonas collosoma	ATCC 30261	NIH	GCA_030849615.1
Kentomonas sorsogonicus	MF-08	NIH	GCA_030347455.1
Angomonas deanei	Cavalho ATCC PRA-265	TriTrypDB-66	
Vickermania spadyakhi	S13	NIH	GCA_030849815.1
Jaenimonas drosophilae	Fi-01.02	NIH	GCA_030849065.1
Phytomonas sp.	EM1	NIH	GCA_000582765.1
Herpetomonas tarakana	OSR18	NIH	GCA_030849825.1
Blechomonas ayalai	B08-376	TriTrypDB-66	
Trypanosoma mega	ATCC 30038	NIH	GCA_030849715.1
Trypanosoma boissoni	ITMAP 2211	NIH	GCA_030849725.1
Trypanosoma congolense	IL3000	TriTrypDB-66	
Trypanosoma brucei	Lister strain 427 2018	TriTrypDB-66	
Trypanosoma vivax	Y486	TriTrypDB-66	
Trypanosoma scelopori	H3-2	NIH	GCA_030849745.1
Trypanosoma platydactyli	RI-340	NIH	GCA_030849675.1
Trypanosoma avium	A1412	NIH	GCA_030849755.1
Trypanosoma caninum	crio 12064	NIH	GCA_036321205.1
Trypanosoma musculi	Partinico II	NIH	GCA_036321165.1
Trypanosoma lewisi	CPO02	NIH	GCA_036321185.1
Trypanosoma rangeli	SC58	TriTrypDB-66	
Trypanosoma cruzi	CL Brener Esmeraldo-like	TriTrypDB-66	
Trypanosoma cruzi	CL Brener Non-Esmeraldo-like	TriTrypDB-66	
Trypanosoma greyi	ANR4	TriTrypDB-66	
Trypanosoma theileri	Edinburgh isolate	TriTrypDB-66	
Paratrypanosoma confusum	CUL13	TriTrypDB-66	
Bodo saltans	Lake Konstanz	TriTrypDB-66	

Bibliography

Akashi, H., Kliman, R.M. and Eyre-Walker, A. (1998) ‘Mutation pressure, natural selection, and the evolution of base composition in Drosophila’, *Genetica*, 102–103(1–6), pp. 49–60.

Akhoundi, M. *et al.* (2016) ‘A Historical Overview of the Classification, Evolution, and Dispersion of Leishmania Parasites and Sandflies’, *PLOS Neglected Tropical Diseases*, 10(3), p. e0004349. Available at: https://doi.org/10.1371/journal.pntd.0004349.

Akiyoshi, B. and Gull, K. (2014) ‘Discovery of Unconventional Kinetochores in Kinetoplastids’, *Cell*, 156(6), pp. 1247–1258. Available at: https://doi.org/10.1016/j.cell.2014.01.049.

Al Jewari, C. and Baldauf, S.L. (2023) ‘An excavate root for the eukaryote tree of life’, *Science Advances*, 9(17), p. eade4973. Available at: https://doi.org/10.1126/sciadv.ade4973.

Albanaz, A.T.S. *et al.* (2021) ‘Genome Analysis of Endotrypanum and Porcisia spp., Closest Phylogenetic Relatives of Leishmania, Highlights the Role of Amastins in Shaping Pathogenicity’, *Genes*, 12(3), p. 444. Available at: https://doi.org/10.3390/genes12030444.

Albanaz, A.T.S. *et al.* (2023) ‘Shining the spotlight on the neglected: new high-quality genome assemblies as a gateway to understanding the evolution of Trypanosomatidae’, *BMC Genomics*, 24(1), p. 471. Available at: https://doi.org/10.1186/s12864-023-09591-z.

Albertson, T.M. *et al.* (2009) ‘DNA polymerase epsilon and delta proofreading suppress discrete mutator and cancer phenotypes in mice’, *Proceedings of the National Academy of Sciences of the United States of America*, 106(40), pp. 17101–17104. Available at: https://doi.org/10.1073/pnas.0907147106.

Almeida, L.V. *et al.* (2018) ‘Chromosomal copy number variation analysis by next generation sequencing confirms ploidy stability in Trypanosoma brucei subspecies’, *Microbial Genomics*, 4(10), p. e000223. Available at: https://doi.org/10.1099/mgen.0.000223.

Alvarez-Jarreta, J. *et al.* (2024) ‘VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center in 2023’, *Nucleic Acids Research*, 52(D1), pp. D808–D816. Available at: https://doi.org/10.1093/nar/gkad1003.

An, N. *et al.* (2015) ‘Nanopore Detection of 8-Oxoguanine in the Human Telomere Repeat Sequence’, *ACS Nano*, 9(4), pp. 4296–4307. Available at: https://doi.org/10.1021/acsnano.5b00722.

Badel, C., Samson, R.Y. and Bell, S.D. (2022) ‘Chromosome organization affects genome evolution in Sulfolobus archaea’, *Nature Microbiology*, 7(6), pp. 820–830. Available at: https://doi.org/10.1038/s41564-022-01127-7.

Badjatia, N. *et al.* (2013) ‘Trypanosoma brucei harbours a divergent XPB helicase paralogue that is specialized in nucleotide excision repair and conserved among kinetoplastid organisms', *Molecular Microbiology*, 90(6), pp. 1293–1308. Available at: https://doi.org/10.1111/mmi.12435.

Bailey, T.L. *et al.* (2015) 'The MEME Suite', *Nucleic Acids Research*, 43(W1), pp. W39–W49. Available at: https://doi.org/10.1093/nar/gkv416.

Barcons-Simon, A., Carrington, M. and Siegel, T.N. (2023) 'Decoding the impact of nuclear organization on antigenic variation in parasites', *Nature Microbiology*, 8(8), pp. 1408–1418. Available at: https://doi.org/10.1038/s41564-023-01424-9.

Beletskii, A. and Bhagwat, A.S. (1996) 'Transcription-induced mutations: Increase in C to T mutations in the nontranscribed strand during transcription in Escherichia coli', *Proceedings of the National Academy of Sciences*, 93(24), pp. 13919–13924. Available at: https://doi.org/10.1073/pnas.93.24.13919.

Benmerzouga, I. *et al.* (2013) 'Trypanosoma brucei Orc1 is essential for nuclear DNA replication and affects both VSG silencing and VSG switching', *Molecular Microbiology*, 87(1), pp. 196–210. Available at: https://doi.org/10.1111/mmi.12093.

Bennetzen, J.L. and Hall, B.D. (1982) 'Codon selection in yeast', *The Journal of Biological Chemistry*, 257(6), pp. 3026–3031.

Benson, G. (1999) 'Tandem repeats finder: a program to analyze DNA sequences', *Nucleic Acids Research*, 27(2), pp. 573–580. Available at: https://doi.org/10.1093/nar/27.2.573.

Berná, L. *et al.* (2018) 'Expanding an expanded genome: long-read sequencing of Trypanosoma cruzi', *Microbial Genomics*, 4(5). Available at: https://doi.org/10.1099/mgen.0.000177.

Bernardi, G. *et al.* (1988) 'Compositional patterns in vertebrate genomes: conservation and change in evolution', *Journal of Molecular Evolution*, 28(1–2), pp. 7–18. Available at: https://doi.org/10.1007/BF02143493.

Bernards, A., van Harten-Loosbroek, N. and Borst, P. (1984) 'Modification of telomeric DNA in Trypanosoma brucei; a role in antfgenic varation?', *Nucleic Acids Research*, p. 18.

Bernards, A., Kooter, J.M. and Borst, P. (1985) 'Structure and transcription of a telomeric surface antigen gene of Trypanosoma brucei', *Molecular and Cellular Biology*, 5(3), pp. 545–553. Available at: https://doi.org/10.1128/mcb.5.3.545-553.1985.

Berriman, M. *et al.* (2002) 'The architecture of variant surface glycoprotein gene expression sites in Trypanosoma brucei', *Molecular and Biochemical Parasitology*, 122(2), pp. 131–140. Available at: https://doi.org/10.1016/s0166-6851(02)00092-0.

Berriman, M. *et al.* (2005) 'The genome of the African trypanosome Trypanosoma brucei', *Science (New York, N.Y.)*, 309(5733), pp. 416–422. Available at: https://doi.org/10.1126/science.1112642.

Bhagwat, A.S. *et al.* (2016) ‘Strand-biased cytosine deamination at the replication fork causes cytosine to thymine mutations in Escherichia coli’, *Proceedings of the National Academy of Sciences of the United States of America*, 113(8), pp. 2176–2181. Available at: https://doi.org/10.1073/pnas.1522325113.

Boemo, M.A. (2021) ‘DNAscent v2: detecting replication forks in nanopore sequencing data with deep learning’, *BMC Genomics*, 22(1), p. 430. Available at: https://doi.org/10.1186/s12864-021-07736-6.

Boiteux, S. and Jinks-Robertson, S. (2013) ‘DNA Repair Mechanisms and the Bypass of DNA Damage in Saccharomyces cerevisiae’, *Genetics*, 193(4), pp. 1025–1064. Available at: https://doi.org/10.1534/genetics.112.145219.

Boothroyd, C.E. *et al.* (2009) ‘A yeast-endonuclease-generated DNA break induces antigenic switching in Trypanosoma brucei’, *Nature*, 459(7244), pp. 278–281. Available at: https://doi.org/10.1038/nature07982.

Borst, P. and Sabatini, R. (2008) ‘Base J: discovery, biosynthesis, and possible functions’, *Annual Review of Microbiology*, 62, pp. 235–251. Available at: https://doi.org/10.1146/annurev.micro.62.081307.162750.

Brambati, A. *et al.* (2015) ‘Replication and transcription on a collision course: eukaryotic regulation mechanisms and implications for DNA stability’, *Frontiers in Genetics*, 6. Available at: https://www.frontiersin.org/articles/10.3389/fgene.2015.00166 (Accessed: 23 September 2023).

Briggs, E., Hamilton, G., *et al.* (2018) ‘Genome-wide mapping reveals conserved and diverged R-loop activities in the unusual genetic landscape of the African trypanosome genome’, *Nucleic Acids Research*, 46(22), pp. 11789–11805. Available at: https://doi.org/10.1093/nar/gky928.

Briggs, E., Crouch, K., *et al.* (2018) ‘Ribonuclease H1-targeted R-loops in surface antigen gene expression sites can direct trypanosome immune evasion’, *PLOS Genetics*, 14(12), p. e1007729. Available at: https://doi.org/10.1371/journal.pgen.1007729.

Briggs, E.M. *et al.* (2021) ‘Single-cell transcriptomic analysis of bloodstream Trypanosoma brucei reconstructs cell cycle progression and developmental quorum sensing’, *Nature Communications*, 12(1), p. 5268. Available at: https://doi.org/10.1038/s41467-021-25607-2.

Budzak, J. *et al.* (2019) ‘Dynamic colocalization of 2 simultaneously active VSG expression sites within a single expression-site body in Trypanosoma brucei’, *Proceedings of the National Academy of Sciences of the United States of America*, 116(33), pp. 16561–16570. Available at: https://doi.org/10.1073/pnas.1905552116.

Budzak, J. *et al.* (2022) ‘An assembly of nuclear bodies associates with the active VSG expression site in African trypanosomes’, *Nature Communications*, 13(1), p. 101. Available at: https://doi.org/10.1038/s41467-021-27625-6.

Bullard, W. *et al.* (2014) 'Identification of the glucosyltransferase that converts hydroxymethyluracil to base J in the trypanosomatid genome', *The Journal of Biological Chemistry*, 289(29), pp. 20273–20282. Available at: https://doi.org/10.1074/jbc.M114.579821.

Bullard, W. *et al.* (2015) 'Base J glucosyltransferase does not regulate the sequence specificity of J synthesis in trypanosomatid telomeric DNA', *Molecular and Biochemical Parasitology*, 204(2), pp. 77–80. Available at: https://doi.org/10.1016/j.molbiopara.2016.01.005.

Bullard, W., Kieft, R. and Sabatini, R. (2017) 'A method for the efficient and selective identification of 5-hydroxymethyluracil in genomic DNA', *Biology Methods & Protocols*, 2(1), p. bpw006. Available at: https://doi.org/10.1093/biomethods/bpw006.

Burgers, P.M.J. and Kunkel, T.A. (2017) 'Eukaryotic DNA Replication Fork', *Annual Review of Biochemistry*, 86(1), pp. 417–438. Available at: https://doi.org/10.1146/annurev-biochem-061516-044709.

Burki, F. *et al.* (2020) 'The New Tree of Eukaryotes', *Trends in Ecology & Evolution*, 35(1), pp. 43–55. Available at: https://doi.org/10.1016/j.tree.2019.08.008.

Butenko, A. *et al.* (2019) 'Comparative genomics of Leishmania (Mundinia)', *BMC Genomics*, 20(1), p. 726. Available at: https://doi.org/10.1186/s12864-019-6126-y.

Calderano, S.G. *et al.* (2015) 'Single molecule analysis of Trypanosoma brucei DNA replication dynamics', *Nucleic Acids Research*, 43(5), pp. 2655–2665. Available at: https://doi.org/10.1093/nar/gku1389.

Callejas, S. *et al.* (2006) 'Hemizygous subtelomeres of an African trypanosome chromosome may account for over 75% of chromosome length', *Genome Research*, 16(9), pp. 1109–1118. Available at: https://doi.org/10.1101/gr.5147406.

Callejas-Hernández, F. *et al.* (2018) 'Genomic assemblies of newly sequenced Trypanosoma cruzi strains reveal new genomic expansion and greater complexity', *Scientific Reports*, 8(1), p. 14631. Available at: https://doi.org/10.1038/s41598-018-32877-2.

Campbell, S. (2019) *Understanding the genomic relationship between nuclear DNA replication and genome plasticity in kinetoplastid genomes*. PhD. University of Glasgow.

Capewell, P. *et al.* (2016) 'The skin is a significant but overlooked anatomical reservoir for vector-borne African trypanosomes', *eLife*. Edited by P. Sinnis, 5, p. e17716. Available at: https://doi.org/10.7554/eLife.17716.

Castillo-Acosta, V.M. *et al.* (2012) 'Increased uracil insertion in DNA is cytotoxic and increases the frequency of mutation, double strand break formation and VSG switching in Trypanosoma brucei', *DNA repair*, 11(12), pp. 986–995. Available at: https://doi.org/10.1016/j.dnarep.2012.09.007.

Castillo-González, C. *et al.* (2022) ‘Quantification of 8-oxoG in Plant Telomeres’, *International Journal of Molecular Sciences*, 23(9), p. 4990. Available at: https://doi.org/10.3390/ijms23094990.

Chambers, V.S. *et al.* (2015) ‘High-throughput sequencing of DNA G-quadruplex structures in the human genome’, *Nature Biotechnology*, 33(8), pp. 877–881. Available at: https://doi.org/10.1038/nbt.3295.

Chaves, I. *et al.* (1998) ‘Subnuclear localization of the active variant surface glycoprotein gene expression site in Trypanosoma brucei’, *Proceedings of the National Academy of Sciences*, 95(21), pp. 12328–12333. Available at: https://doi.org/10.1073/pnas.95.21.12328.

Chen, N. and Buonomo, S.C.B. (2023) ‘Three-dimensional nuclear organisation and the DNA replication timing program’, *Current Opinion in Structural Biology*, 83, p. 102704. Available at: https://doi.org/10.1016/j.sbi.2023.102704.

Chung, D., Kwon, Y.M. and Yang, Y. (2021) ‘Telomere-to-telomere genome assembly of asparaginase-producing Trichoderma simmonsii’, *BMC Genomics*, 22(1), p. 830. Available at: https://doi.org/10.1186/s12864-021-08162-4.

Chung, H.M. *et al.* (1990) ‘Architectural organization in the interphase nucleus of the protozoan Trypanosoma brucei: location of telomeres and mini-chromosomes’, *The EMBO journal*, 9(8), pp. 2611–2619.

Clayton, C. (2019) ‘Regulation of gene expression in trypanosomatids: living with polycistronic transcription’, *Open Biology*, 9(6), p. 190072. Available at: https://doi.org/10.1098/rsob.190072.

Cliffe, L.J. *et al.* (2009) ‘JBP1 and JBP2 are two distinct thymidine hydroxylases involved in J biosynthesis in genomic DNA of African trypanosomes’, *Nucleic Acids Research*, 37(5), pp. 1452–1462. Available at: https://doi.org/10.1093/nar/gkn1067.

Cliffe, L.J. *et al.* (2010) ‘Two thymidine hydroxylases differentially regulate the formation of glucosylated DNA at regions flanking polymerase II polycistronic transcription units throughout the genome of Trypanosoma brucei’, *Nucleic Acids Research*, 38(12), pp. 3923–3935. Available at: https://doi.org/10.1093/nar/gkq146.

Cordon-Obras, C. *et al.* (2022) ‘Identification of sequence-specific promoters driving polycistronic transcription initiation by RNA polymerase II in trypanosomes’, *Cell Reports*, 38(2), p. 110221. Available at: https://doi.org/10.1016/j.celrep.2021.110221.

Cosentino, R.O., Brink, B.G. and Siegel, T.N. (2021) ‘Allele-specific assembly of a eukaryotic genome corrects apparent frameshifts and reveals a lack of nonsense-mediated mRNA decay’, *NAR Genomics and Bioinformatics*, 3(3), p. lqab082. Available at: https://doi.org/10.1093/nargab/lqab082.

Crooks, G.E. *et al.* (2004) ‘WebLogo: A Sequence Logo Generator’, *Genome Research*, 14(6), pp. 1188–1190. Available at: https://doi.org/10.1101/gr.849004.

Cross, G.A.M. (1978) ‘Molecular basis of antigenic variation in trypanosomes’, *Trends in Biochemical Sciences*, 3(1), pp. 49–51. Available at: https://doi.org/10.1016/S0968-0004(78)93810-0.

Cross, G.A.M., Kim, H.-S. and Wickstead, B. (2014) ‘Capturing the variant surface glycoprotein repertoire (the VSGnome) of Trypanosoma brucei Lister 427’, *Molecular and Biochemical Parasitology*, 195(1), pp. 59–73. Available at: https://doi.org/10.1016/j.molbiopara.2014.06.004.

Crossley, M.P., Bocek, M. and Cimprich, K.A. (2019) ‘R-Loops as Cellular Regulators and Genomic Threats’, *Molecular Cell*, 73(3), pp. 398–411. Available at: https://doi.org/10.1016/j.molcel.2019.01.024.

Damasceno, J.D. *et al.* (2020) ‘Genome duplication in Leishmania major relies on persistent subtelomeric DNA replication’, *eLife*. Edited by C. Clayton et al., 9, p. e58030. Available at: https://doi.org/10.7554/eLife.58030.

Damasceno, J.D., Silva, G.L., *et al.* (2024a) ‘Nuclear DNA replication in Leishmania major relies on a single constitutive origin per chromosome supplemented by thousands of stochastic initiation events’. bioRxiv, p. 2024.11.14.623610. Available at: https://doi.org/10.1101/2024.11.14.623610.

Damasceno, J.D., Briggs, E.M., *et al.* (2024b) ‘R-loops acted on by RNase H1 are a determinant of chromosome length-associated DNA replication timing and genome stability in Leishmania’. bioRxiv, p. 2024.04.29.591643. Available at: https://doi.org/10.1101/2024.04.29.591643.

De Coster, W. *et al.* (2018) ‘NanoPack: visualizing and processing long-read sequencing data’, *Bioinformatics*, 34(15), pp. 2666–2669. Available at: https://doi.org/10.1093/bioinformatics/bty149.

Devlin, R. *et al.* (2016) ‘Mapping replication dynamics in Trypanosoma brucei reveals a link with telomere transcription and antigenic variation’, *eLife*, 5. Available at: https://doi.org/10.7554/eLife.12765.

Díaz-Viraqué, F. *et al.* (2019) ‘Nanopore Sequencing Significantly Improves Genome Assembly of the Protozoan Parasite Trypanosoma cruzi’, *Genome Biology and Evolution*, 11(7), pp. 1952–1957. Available at: https://doi.org/10.1093/gbe/evz129.

Douglas, G.R. *et al.* (1994) ‘Sequenc spectra of spontaneous lacZ gene mutations in trransgenic mouse somatic and germline tissues’, *Mutagenesis*, 9(5), pp. 451–458. Available at: https://doi.org/10.1093/mutage/9.5.451.

Douzery, E.J.P. *et al.* (2004) ‘The timing of eukaryotic evolution: Does a relaxed molecular clock reconcile proteins and fossils?’, *Proceedings of the National Academy of Sciences*, 101(43), pp. 15386–15391. Available at: https://doi.org/10.1073/pnas.0403984101.

Duquette, M.L. *et al.* (2004) ‘Intracellular transcription of G-rich DNAs induces formation of G-loops, novel structures containing G4 DNA’, *Genes & Development*, 18(13), pp. 1618–1629. Available at: https://doi.org/10.1101/gad.1200804.

Echeverry, M.C. *et al.* (2012) 'Centromere-associated repeat arrays on Trypanosoma brucei chromosomes are much more extensive than predicted', *BMC Genomics*, 13(1), p. 29. Available at: https://doi.org/10.1186/1471-2164-13-29.

El-Sayed, N.M. *et al.* (2000) 'The African trypanosome genome', *International Journal for Parasitology*, 30(4), pp. 329–345. Available at: https://doi.org/10.1016/s0020-7519(00)00015-1.

El-Sayed, N.M., Myler, P.J., Blandin, G., *et al.* (2005) 'Comparative genomics of trypanosomatid parasitic protozoa', *Science (New York, N.Y.)*, 309(5733), pp. 404–409. Available at: https://doi.org/10.1126/science.1112181.

El-Sayed, N.M., Myler, P.J., Bartholomeu, D.C., *et al.* (2005) 'The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease', *Science (New York, N.Y.)*, 309(5733), pp. 409–415. Available at: https://doi.org/10.1126/science.1112631.

Faria, J. *et al.* (2021) 'Spatial integration of transcription and splicing in a dedicated compartment sustains monogenic antigen expression in African trypanosomes', *Nature Microbiology*, 6(3), pp. 289–300. Available at: https://doi.org/10.1038/s41564-020-00833-4.

Faria, J. *et al.* (2022) 'Emergence and adaptation of the cellular machinery directing antigenic variation in the African trypanosome', *Current Opinion in Microbiology*, 70, p. 102209. Available at: https://doi.org/10.1016/j.mib.2022.102209.

Flegontov, P. *et al.* (2013) 'Paratrypanosoma Is a Novel Early-Branching Trypanosomatid', *Current Biology*, 23(18), pp. 1787–1793. Available at: https://doi.org/10.1016/j.cub.2013.07.045.

Fousteri, M. and Mullenders, L.H. (2008) 'Transcription-coupled nucleotide excision repair in mammalian cells: molecular mechanisms and biological effects', *Cell Research*, 18(1), pp. 73–84. Available at: https://doi.org/10.1038/cr.2008.6.

Frederico, L.A., Kunkel, T.A. and Shaw, B.R. (1990) 'A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy', *Biochemistry*, 29(10), pp. 2532–2537. Available at: https://doi.org/10.1021/bi00462a015.

Gao, J.-M. *et al.* (2020) 'Human African trypanosomiasis: the current situation in endemic regions and the risks for non-endemic regions from imported cases', *Parasitology*, 147(9), pp. 922–931. Available at: https://doi.org/10.1017/S0031182020000645.

Genest, P.-A. *et al.* (2015) 'Defining the sequence requirements for the positioning of base J in DNA using SMRT sequencing', *Nucleic Acids Research*, 43(4), pp. 2102–2115. Available at: https://doi.org/10.1093/nar/gkv095.

Ghedin, E. *et al.* (2004) 'Gene synteny and evolution of genome architecture in trypanosomatids', *Molecular and Biochemical Parasitology*, 134(2), pp. 183–191. Available at: https://doi.org/10.1016/j.molbiopara.2003.11.012.

Gibson, W. and Garside, L. (1991) ‘Genetic exchange in Trypanosoma brucei brucei: variable chromosomal location of housekeeping genes in different trypanosome stocks’, *Molecular and Biochemical Parasitology*, 45(1), pp. 77–89. Available at: https://doi.org/10.1016/0166-6851(91)90029-6.

Ginno, P.A. *et al*. (2012) ‘R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters’, *Molecular Cell*, 45(6), pp. 814–825. Available at: https://doi.org/10.1016/j.molcel.2012.01.017.

Ginno, P.A. *et al*. (2013) ‘GC skew at the 5’ and 3’ ends of human genes links R-loop formation to epigenetic regulation and transcription termination’, *Genome Research*, 23(10), pp. 1590–1600. Available at: https://doi.org/10.1101/gr.158436.113.

Girasol, M.J. *et al*. (2023) ‘RAD51-mediated R-loop formation acts to repair transcription-associated DNA breaks driving antigenic variation in Trypanosoma brucei’. bioRxiv, p. 2023.05.11.540369. Available at: https://doi.org/10.1101/2023.05.11.540369.

Glover, L. *et al*. (2013) ‘Antigenic variation in African trypanosomes: the importance of chromosomal and nuclear context in *VSG* expression control: Antigenic variation in African trypanosomes’, *Cellular Microbiology*, 15(12), pp. 1984–1993. Available at: https://doi.org/10.1111/cmi.12215.

Glover, L., Alsford, S. and Horn, D. (2013) ‘DNA Break Site at Fragile Subtelomeres Determines Probability and Mechanism of Antigenic Variation in African Trypanosomes’, *PLoS Pathogens*. Edited by K.L. Hill, 9(3), p. e1003260. Available at: https://doi.org/10.1371/journal.ppat.1003260.

Glover, L., Jun, J. and Horn, D. (2011) ‘Microhomology-mediated deletion and gene conversion in African trypanosomes’, *Nucleic Acids Research*, 39(4), pp. 1372–1380. Available at: https://doi.org/10.1093/nar/gkq981.

Glover, L., McCulloch, R. and Horn, D. (2008) ‘Sequence homology and microhomology dominate chromosomal double-strand break repair in African trypanosomes’, *Nucleic Acids Research*, 36(8), pp. 2608–2618. Available at: https://doi.org/10.1093/nar/gkn104.

Gommers-Ampt, J., Lutgerink, J. and Borst, P. (1991) ‘A novel DNA nucleotide in Trypanosoma brucei only present in the mammalian phase of the life-cycle’, *Nucleic Acids Research*, 19(8), pp. 1745–1751. Available at: https://doi.org/10.1093/nar/19.8.1745.

Gommers-Ampt, J.H. *et al*. (1993) ‘beta-D-glucosyl-hydroxymethyluracil: a novel modified base present in the DNA of the parasitic protozoan T. brucei’, *Cell*, 75(6), pp. 1129–1136. Available at: https://doi.org/10.1016/0092-8674(93)90322-h.

Gommers-Ampt, J.H. and Borst, P. (1995) ‘Hypermodified bases in DNA’, *FASEB journal: official publication of the Federation of American Societies for Experimental Biology*, 9(11), pp. 1034–1042. Available at: https://doi.org/10.1096/fasebj.9.11.7649402.

Gottesdiener, K. *et al.* (1991) ‘Characterization of VSG gene expression site promoters and promoter-associated DNA rearrangement events.’, *Molecular and Cellular Biology*, 11(5), pp. 2467–2480. Available at: https://doi.org/10.1128/MCB.11.5.2467.

Grant, C.E., Bailey, T.L. and Noble, W.S. (2011) ‘FIMO: scanning for occurrences of a given motif’, *Bioinformatics*, 27(7), pp. 1017–1018. Available at: https://doi.org/10.1093/bioinformatics/btr064.

Guillet, M., Van Der Kemp, P.A. and Boiteux, S. (2006) ‘dUTPase activity is critical to maintain genetic stability in Saccharomyces cerevisiae’, *Nucleic Acids Research*, 34(7), pp. 2056–2066. Available at: https://doi.org/10.1093/nar/gkl139.

Günzl, A. *et al.* (2003) ‘RNA polymerase I transcribes procyclin genes and variant surface glycoprotein gene expression sites in Trypanosoma brucei’, *Eukaryotic Cell*, 2(3), pp. 542–551. Available at: https://doi.org/10.1128/EC.2.3.542-551.2003.

Gurevich, A. *et al.* (2013) ‘QUAST: quality assessment tool for genome assemblies’, *Bioinformatics*, 29(8), pp. 1072–1075. Available at: https://doi.org/10.1093/bioinformatics/btt086.

Hakim, J.M.C. *et al.* (2023) ‘Whole Genome Assembly of a Hybrid Trypanosoma cruzi Strain Assembled with Nanopore Sequencing Alone’, *bioRxiv: The Preprint Server for Biology*, p. 2023.07.27.550875. Available at: https://doi.org/10.1101/2023.07.27.550875.

Hall, J.P.J., Wang, H. and Barry, J.D. (2013) ‘Mosaic VSGs and the Scale of Trypanosoma brucei Antigenic Variation’, *PLoS Pathogens*. Edited by D. Horn, 9(7), p. e1003502. Available at: https://doi.org/10.1371/journal.ppat.1003502.

Halliday, J.A. and Glickman, B.W. (1991) ‘Mechanisms of spontaneous mutation in DNA repair-proficient Escherichia coli’, *Mutation Research*, 250(1–2), pp. 55–71. Available at: https://doi.org/10.1016/0027-5107(91)90162-h.

Hartley, C.L. and McCulloch, R. (2008) ‘Trypanosoma brucei BRCA2 acts in antigenic variation and has undergone a recent expansion in BRC repeat number that is important during homologous recombination’, *Molecular Microbiology*, 68(5), pp. 1237–1251. Available at: https://doi.org/10.1111/j.1365-2958.2008.06230.x.

Hertz-Fowler, C. *et al.* (2008) ‘Telomeric Expression Sites Are Highly Conserved in Trypanosoma brucei’, *PLoS ONE*. Edited by N. Hall, 3(10), p. e3527. Available at: https://doi.org/10.1371/journal.pone.0003527.

Hovel-Miner, G. *et al.* (2016) ‘A Conserved DNA Repeat Promotes Selection of a Diverse Repertoire of Trypanosoma brucei Surface Antigens from the Genomic Archive’, *PLOS Genetics*. Edited by H.S. Malik, 12(5), p. e1005994. Available at: https://doi.org/10.1371/journal.pgen.1005994.

Hovel-Miner, G.A. *et al.* (2012) ‘Telomere Length Affects the Frequency and Mechanism of Antigenic Variation in Trypanosoma brucei’, *PLoS Pathogens*.

Edited by K.L. Hill, 8(8), p. e1002900. Available at: https://doi.org/10.1371/journal.ppat.1002900.

Hu, Y. and Stillman, B. (2023) ‘Origins of DNA replication in eukaryotes’, *Molecular Cell*, 83(3), pp. 352–372. Available at: https://doi.org/10.1016/j.molcel.2022.12.024.

Huvet, M. *et al.* (2007) ‘Human gene organization driven by the coordination of replication and transcription’, *Genome Research*, 17(9), pp. 1278–1285. Available at: https://doi.org/10.1101/gr.6533407.

Ikemura, T. (1985) ‘Codon usage and tRNA content in unicellular and multicellular organisms’, *Molecular Biology and Evolution*, 2(1), pp. 13–34. Available at: https://doi.org/10.1093/oxfordjournals.molbev.a040335.

Ivens, A.C. *et al.* (2005) ‘The genome of the kinetoplastid parasite, Leishmania major’, *Science (New York, N.Y.)*, 309(5733), pp. 436–442. Available at: https://doi.org/10.1126/science.1112680.

Jackson, A.P. *et al.* (2016) ‘Kinetoplastid Phylogenomics Reveals the Evolutionary Innovations Associated with the Origins of Parasitism’, *Current Biology*, 26(2), pp. 161–172. Available at: https://doi.org/10.1016/j.cub.2015.11.055.

Jackson, A.P., Quail, M.A. and Berriman, M. (2008) ‘Insights into the genome sequence of a free-living Kinetoplastid: Bodo saltans (Kinetoplastida: Euglenozoa)’, *BMC Genomics*, 9(1), p. 594. Available at: https://doi.org/10.1186/1471-2164-9-594.

Jain, M. *et al.* (2018) ‘Nanopore sequencing and assembly of a human genome with ultra-long reads’, *Nature Biotechnology*, 36(4), pp. 338–345. Available at: https://doi.org/10.1038/nbt.4060.

Jain, R., Aggarwal, A.K. and Rechkoblit, O. (2018) ‘Eukaryotic DNA polymerases’, *Current Opinion in Structural Biology*, 53, pp. 77–87. Available at: https://doi.org/10.1016/j.sbi.2018.06.003.

Jinks-Robertson, S. and Bhagwat, A.S. (2014) ‘Transcription-Associated Mutagenesis’, *Annual Review of Genetics*, 48(1), pp. 341–359. Available at: https://doi.org/10.1146/annurev-genet-120213-092015.

de Jong, P.J., Grosovsky, A.J. and Glickman, B.W. (1988) ‘Spectrum of spontaneous mutation at the APRT locus of Chinese hamster ovary cells: an analysis at the DNA sequence level.’, *Proceedings of the National Academy of Sciences*, 85(10), pp. 3499–3503. Available at: https://doi.org/10.1073/pnas.85.10.3499.

Kanmogne, G.D., Bailey, M. and Gibson, W.C. (1997) ‘Wide variation in DNA content among isolates of *Trypanosoma brucei* ssp.’, *Acta Tropica*, 63(2), pp. 75–87. Available at: https://doi.org/10.1016/S0001-706X(96)00600-6.

Kano-Sueoka, T., Lobry, J.R. and Sueoka, N. (1999) ‘Intra-strand biases in bacteriophage T4 genome’, *Gene*, 238(1), pp. 59–64. Available at: https://doi.org/10.1016/s0378-1119(99)00296-6.

Kassem, A., Pays, E. and Vanhamme, L. (2014) ‘Transcription is initiated on silent variant surface glycoprotein expression sites despite monoallelic expression in Trypanosoma brucei’, *Proceedings of the National Academy of Sciences*, 111(24), pp. 8943–8948. Available at: https://doi.org/10.1073/pnas.1404873111.

Kawasaki, F. *et al.* (2017) ‘Genome-wide mapping of 5-hydroxymethyluracil in the eukaryote parasite Leishmania’, *Genome Biology*, 18(1), p. 23. Available at: https://doi.org/10.1186/s13059-017-1150-1.

Kaye, P. and Scott, P. (2011) ‘Leishmaniasis: complexity at the host-pathogen interface’, *Nature Reviews Microbiology*, 9(8), pp. 604–615. Available at: https://doi.org/10.1038/nrmicro2608.

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, pp. 52–55. Available at: https://doi.org/10.1016/j.molbiopara.2017.07.002.

Kooter, J.M. *et al.* (1987) ‘The anatomy and transcription of a telomeric expression site for variant-specific surface antigens in T. brucei’, *Cell*, 51(2), pp. 261–272. Available at: https://doi.org/10.1016/0092-8674(87)90153-X.

Koren, S. *et al.* (2017) ‘Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation’, *Genome Research*, 27(5), pp. 722–736. Available at: https://doi.org/10.1101/gr.215087.116.

Kostygov, A.Y. *et al.* (2021) ‘Euglenozoa: taxonomy, diversity and ecology, symbioses and viruses’, *Open Biology*, 11(3), p. 200407. Available at: https://doi.org/10.1098/rsob.200407.

Kostygov, A.Yu. *et al.* (2024) ‘Phylogenetic framework to explore trait evolution in Trypanosomatidae’, *Trends in Parasitology*, 40(2), pp. 96–99. Available at: https://doi.org/10.1016/j.pt.2023.11.009.

Lachaud, L. *et al.* (2014) ‘Constitutive mosaic aneuploidy is a unique genetic feature widespread in the Leishmania genus’, *Microbes and Infection*, 16(1), pp. 61–66. Available at: https://doi.org/10.1016/j.micinf.2013.09.005.

Laffitte, M.-C.N. *et al.* (2016) ‘Plasticity of the Leishmania genome leading to gene copy number variations and drug resistance’, *F1000Research*, 5, p. 2350. Available at: https://doi.org/10.12688/f1000research.9218.1.

Lake, J.A. *et al.* (1988) ‘Evolution of parasitism: kinetoplastid protozoan history reconstructed from mitochondrial rRNA gene sequences.’, *Proceedings of the National Academy of Sciences*, 85(13), pp. 4779–4783. Available at: https://doi.org/10.1073/pnas.85.13.4779.

Landeira, D. and Navarro, M. (2007) ‘Nuclear repositioning of the VSG promoter during developmental silencing in Trypanosoma brucei’, *The Journal of Cell Biology*, 176(2), pp. 133–139. Available at: https://doi.org/10.1083/jcb.200607174.

Langousis, G. and Hill, K.L. (2014) ‘Motility and more: the flagellum of Trypanosoma brucei’, *Nature Reviews Microbiology*, 12(7), pp. 505–518. Available at: https://doi.org/10.1038/nrmicro3274.

Lecordier, L. *et al.* (2007) ‘Characterization of a TFIIH homologue from Trypanosoma brucei’, *Molecular Microbiology*, 64(5), pp. 1164–1181. Available at: https://doi.org/10.1111/j.1365-2958.2007.05725.x.

Lee, J.H. *et al.* (2007) ‘Spliced leader RNA gene transcription in Trypanosoma brucei requires transcription factor TFIIH’, *Eukaryotic Cell*, 6(4), pp. 641–649. Available at: https://doi.org/10.1128/EC.00411-06.

Lee, J.H., Jung, H.S. and Günzl, A. (2009) ‘Transcriptionally active TFIIH of the early-diverged eukaryote Trypanosoma brucei harbors two novel core subunits but not a cyclin-activating kinase complex’, *Nucleic Acids Research*, 37(11), pp. 3811–3820. Available at: https://doi.org/10.1093/nar/gkp236.

van Leeuwen, F. *et al.* (1996) ‘The telomeric GGGTTA repeats of Trypanosoma brucei contain the hypermodified base J in both strands’, *Nucleic Acids Research*, 24(13), pp. 2476–2482. Available at: https://doi.org/10.1093/nar/24.13.2476.

Leeuwen, F. van *et al.* (1997) ‘Localization of the modified base J in telomeric VSG gene expression sites ofTrypanosoma brucei’, *Genes & Development*, 11(23), pp. 3232–3241. Available at: https://doi.org/10.1101/gad.11.23.3232.

van Leeuwen, F. *et al.* (1998) ‘Biosynthesis and function of the modified DNA base beta-D-glucosyl-hydroxymethyluracil in Trypanosoma brucei’, *Molecular and Cellular Biology*, 18(10), pp. 5643–5651. Available at: https://doi.org/10.1128/MCB.18.10.5643.

van Leeuwen, Fred *et al.* (1998) ‘ß-d-Glucosyl-hydroxymethyluracil is a conserved DNA modification in kinetoplastid protozoans and is abundant in their telomeres’, *Proceedings of the National Academy of Sciences*, 95(5), pp. 2366–2371. Available at: https://doi.org/10.1073/pnas.95.5.2366.

van Leeuwen, F. *et al.* (2000) ‘Tandemly repeated DNA is a target for the partial replacement of thymine by beta-D-glucosyl-hydroxymethyluracil in Trypanosoma brucei’, *Molecular and Biochemical Parasitology*, 109(2), pp. 133–145. Available at: https://doi.org/10.1016/s0166-6851(00)00247-4.

Leonard, A.C. and Méchali, M. (2013) ‘DNA Replication Origins’, *Cold Spring Harbor Perspectives in Biology*, 5(10), p. a010116. Available at: https://doi.org/10.1101/cshperspect.a010116.

Lewis, C.A. *et al.* (2016) ‘Cytosine deamination and the precipitous decline of spontaneous mutation during Earth’s history’, *Proceedings of the National Academy of Sciences*, 113(29), pp. 8194–8199. Available at: https://doi.org/10.1073/pnas.1607580113.

Li, B. (2015) ‘DNA Double-Strand Breaks and Telomeres Play Important Roles in Trypanosoma brucei Antigenic Variation’, *Eukaryotic Cell*, 14(3), pp. 196–205. Available at: https://doi.org/10.1128/EC.00207-14.

Li, H., Ruan, J. and Durbin, R. (2008) 'Mapping short DNA sequencing reads and calling variants using mapping quality scores', *Genome Research*, 18(11), pp. 1851–1858. Available at: https://doi.org/10.1101/gr.078212.108.

Li, H.-D. *et al.* (2018) 'Polymerase-mediated ultramutagenesis in mice produces diverse cancers with high mutational load', *The Journal of Clinical Investigation*, 128(9), pp. 4179–4191. Available at: https://doi.org/10.1172/JCI122095.

Li, L., Stoeckert, C.J. and Roos, D.S. (2003) 'OrthoMCL: identification of ortholog groups for eukaryotic genomes', *Genome Research*, 13(9), pp. 2178–2189. Available at: https://doi.org/10.1101/gr.1224503.

Liu, S. *et al.* (2014) 'Quantitative mass spectrometry-based analysis of β-D-glucosyl-5-hydroxymethyluracil in genomic DNA of Trypanosoma brucei', *Journal of the American Society for Mass Spectrometry*, 25(10), pp. 1763–1770. Available at: https://doi.org/10.1007/s13361-014-0960-6.

Lobry, J.R. (1996) 'Asymmetric substitution patterns in the two DNA strands of bacteria', *Molecular Biology and Evolution*, 13(5), pp. 660–665. Available at: https://doi.org/10.1093/oxfordjournals.molbev.a025626.

Logsdon, G.A. *et al.* (2021) 'The structure, function and evolution of a complete human chromosome 8', *Nature*, 593(7857), pp. 101–107. Available at: https://doi.org/10.1038/s41586-021-03420-7.

Lombraña, R. *et al.* (2016) 'Transcriptionally Driven DNA Replication Program of the Human Parasite Leishmania major', *Cell Reports*, 16(6), pp. 1774–1786. Available at: https://doi.org/10.1016/j.celrep.2016.07.007.

López-Escobar, L. *et al.* (2022) 'Stage-specific transcription activator ESB1 regulates monoallelic antigen expression in Trypanosoma brucei', *Nature Microbiology*, 7(8), pp. 1280–1290. Available at: https://doi.org/10.1038/s41564-022-01175-z.

Lu, H., Giordano, F. and Ning, Z. (2016) 'Oxford Nanopore MinION Sequencing and Genome Assembly', *Genomics, Proteomics & Bioinformatics*, 14(5), pp. 265–279. Available at: https://doi.org/10.1016/j.gpb.2016.05.004.

Ma, C.-J. *et al.* (2021) 'An enzyme-mediated bioorthogonal labeling method for genome-wide mapping of 5-hydroxymethyluracil', *Chemical Science*, 12(42), pp. 14126–14132. Available at: https://doi.org/10.1039/D1SC03812E.

Machado, C.R. *et al.* (2014) 'Nucleotide excision repair in Trypanosoma brucei: specialization of transcription-coupled repair due to multigenic transcription', *Molecular Microbiology*, 92(4), pp. 756–776. Available at: https://doi.org/10.1111/mmi.12589.

MacNeill, S. (2012) 'Composition and Dynamics of the Eukaryotic Replisome: A Brief Overview', in S. MacNeill (ed.) *The Eukaryotic Replisome: a Guide to Protein Structure and Function*. Dordrecht: Springer Netherlands (Subcellular Biochemistry), pp. 1–17. Available at: https://doi.org/10.1007/978-94-007-4572-8_1.

Manni, M. *et al.* (2021) 'BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes', *Molecular Biology and Evolution*, 38(10), pp. 4647–4654. Available at: https://doi.org/10.1093/molbev/msab199.

Marcello, L. and Barry, J.D. (2007) 'Analysis of the VSG gene silent archive in Trypanosoma brucei reveals that mosaic gene expression is prominent in antigenic variation and is favored by archive substructure', *Genome Research*, 17(9), pp. 1344–1352. Available at: https://doi.org/10.1101/gr.6421207.

Maree, J.P. *et al.* (2017) 'Well-positioned nucleosomes punctuate polycistronic pol II transcription units and flank silent VSG gene arrays in Trypanosoma brucei', *Epigenetics & Chromatin*, 10(1). Available at: https://doi.org/10.1186/s13072-017-0121-9.

Marques, C.A. *et al.* (2015) 'Genome-wide mapping reveals single-origin chromosome replication in Leishmania, a eukaryotic microbe', *Genome Biology*, 16, p. 230. Available at: https://doi.org/10.1186/s13059-015-0788-9.

Marsico, G. *et al.* (2019) 'Whole genome experimental maps of DNA G-quadruplexes in multiple species', *Nucleic Acids Research*, 47(8), pp. 3862–3874. Available at: https://doi.org/10.1093/nar/gkz179.

Massey, D.J. and Koren, A. (2022) 'Telomere-to-telomere human DNA replication timing profiles', *Scientific Reports*, 12(1), p. 9560. Available at: https://doi.org/10.1038/s41598-022-13638-8.

Matthews, K.R. (2021) 'Trypanosome Signaling-Quorum Sensing', *Annual Review of Microbiology*, 75, pp. 495–514. Available at: https://doi.org/10.1146/annurev-micro-020321-115246.

McCulloch, R., Morrison, L.J. and Hall, J.P.J. (2015) 'DNA Recombination Strategies During Antigenic Variation in the African Trypanosome', *Microbiology Spectrum*, 3(2), pp. MDNA3-0016–2014. Available at: https://doi.org/10.1128/microbiolspec.MDNA3-0016-2014.

McDonagh, P.D. (2000) 'The unusual gene organization of Leishmania major chromosome 1 may reflect novel transcription processes', *Nucleic Acids Research*, 28(14), pp. 2800–2803. Available at: https://doi.org/10.1093/nar/28.14.2800.

McGinty, R.J. *et al.* (2017) 'Nanopore sequencing of complex genomic rearrangements in yeast reveals mechanisms of repeat-mediated double-strand break repair', *Genome Research*, 27(12), pp. 2072–2082. Available at: https://doi.org/10.1101/gr.228148.117.

Melville, S.E. *et al.* (1998) 'The molecular karyotype of the megabase chromosomes of Trypanosoma brucei and the assignment of chromosome markers', *Molecular and Biochemical Parasitology*, 94(2), pp. 155–173. Available at: https://doi.org/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 tyrpanosomes', *Chromosome Research: An International Journal on the Molecular,* *Supramolecular and Evolutionary Aspects of Chromosome Biology*, 7(3), pp. 191–203. Available at: https://doi.org/10.1023/a:1009247315947.

Merrikh, H. *et al.* (2012) ‘Replication-transcription conflicts in bacteria’, *Nature Reviews Microbiology*, 10(7), pp. 449–458. Available at: https://doi.org/10.1038/nrmicro2800.

Miga, K.H. *et al.* (2020) ‘Telomere-to-telomere assembly of a complete human X chromosome’, *Nature*, 585(7823), pp. 79–84. Available at: https://doi.org/10.1038/s41586-020-2547-7.

Moeckel, C., Zaravinos, A. and Georgakopoulos-Soares, I. (2023) ‘Strand asymmetries across genomic processes’, *Computational and Structural Biotechnology Journal*, 21, pp. 2036–2047. Available at: https://doi.org/10.1016/j.csbj.2023.03.007.

Morrison, A. *et al.* (1990) ‘A third essential DNA polymerase in S. cerevisiae’, *Cell*, 62(6), pp. 1143–1151. Available at: https://doi.org/10.1016/0092-8674(90)90391-q.

Morrison, L.J. *et al.* (2005) ‘Probabilistic order in antigenic variation of Trypanosoma brucei’, *International Journal for Parasitology*, 35(9), pp. 961–972. Available at: https://doi.org/10.1016/j.ijpara.2005.05.004.

Mugnier, M.R., Cross, G.A.M. and Papavasiliou, F.N. (2015) ‘The in vivo dynamics of antigenic variation in Trypanosoma brucei’, *Science*, 347(6229), pp. 1470–1473. Available at: https://doi.org/10.1126/science.aaa4502.

Mulindwa, J. *et al.* (2021) ‘In vitro culture of freshly isolated Trypanosoma brucei brucei bloodstream forms results in gene copy-number changes’, *PLoS neglected tropical diseases*, 15(9), p. e0009738. Available at: https://doi.org/10.1371/journal.pntd.0009738.

Müller, C.A. and Nieduszynski, C.A. (2017) ‘DNA replication timing influences gene expression level’, *Journal of Cell Biology*, 216(7), pp. 1907–1914. Available at: https://doi.org/10.1083/jcb.201701061.

Müller, L.S.M. *et al.* (2018) ‘Genome organization and DNA accessibility control antigenic variation in trypanosomes’, *Nature* [Preprint]. Available at: https://doi.org/10.1038/s41586-018-0619-8.

Naish, M. *et al.* (2021) ‘The genetic and epigenetic landscape of the Arabidopsis centromeres’, *Science*, 374(6569), p. eabi7489. Available at: https://doi.org/10.1126/science.abi7489.

Navarro, M. and Gull, K. (2001) ‘A pol I transcriptional body associated with VSG mono-allelic expression in Trypanosoma brucei’, *Nature*, 414(6865), pp. 759–763. Available at: https://doi.org/10.1038/414759a.

Navarro, M., Peñate, X. and Landeira, D. (2007) ‘Nuclear architecture underlying gene expression in *Trypanosoma brucei*’, *Trends in Microbiology*, 15(6), pp. 263-270. Available at: https://doi.org/10.1016/j.tim.2007.04.004.

Nick McElhinny, S.A. *et al.* (2008) ‘Division of Labor at the Eukaryotic Replication Fork’, *Molecular cell*, 30(2), pp. 137–144. Available at: https://doi.org/10.1016/j.molcel.2008.02.022.

Nilsson, D. and Andersson, B. (2005) ‘Strand asymmetry patterns in trypanosomatid parasites’, *Experimental Parasitology*, 109(3), pp. 143–149. Available at: https://doi.org/10.1016/j.exppara.2004.12.004.

Nurk, S. *et al.* (2022) ‘The complete sequence of a human genome’, *Science*, 376(6588), pp. 44–53. Available at: https://doi.org/10.1126/science.abj6987.

Obado, S.O. *et al.* (2007) ‘Repetitive DNA is associated with centromeric domains in Trypanosoma brucei but not Trypanosoma cruzi’, *Genome Biology*, 8(3), p. R37. Available at: https://doi.org/10.1186/gb-2007-8-3-r37.

Ohshima, K. *et al.* (1996a) ‘Cloning, characterization, and properties of seven triplet repeat DNA sequences’, *The Journal of Biological Chemistry*, 271(28), pp. 16773–16783. Available at: https://doi.org/10.1074/jbc.271.28.16773.

Ohshima, K. *et al.* (1996b) ‘TTA.TAA triplet repeats in plasmids form a non-H bonded structure’, *The Journal of Biological Chemistry*, 271(28), pp. 16784–16791. Available at: https://doi.org/10.1074/jbc.271.28.16784.

Okello, I. *et al.* (2022) ‘African Animal Trypanosomiasis: A Systematic Review on Prevalence, Risk Factors and Drug Resistance in Sub-Saharan Africa’, *Journal of Medical Entomology*, 59(4), pp. 1099–1143. Available at: https://doi.org/10.1093/jme/tjac018.

Olinski, R., Starczak, M. and Gackowski, D. (2016) ‘Enigmatic 5-hydroxymethyluracil: Oxidatively modified base, epigenetic mark or both?’, *Mutation Research/Reviews in Mutation Research*, 767, pp. 59–66. Available at: https://doi.org/10.1016/j.mrrev.2016.02.001.

Overath, P. *et al.* (2001) ‘The surface structure of trypanosomes in relation to their molecular phylogeny’, *International Journal for Parasitology*, 31(5), pp. 468–471. Available at: https://doi.org/10.1016/S0020-7519(01)00152-7.

Owiti, N., Stokdyk, K. and Kim, N. (2019) ‘The etiology of uracil residues in the Saccharomyces cerevisiae genomic DNA’, *Current Genetics*, 65(2), pp. 393–399. Available at: https://doi.org/10.1007/s00294-018-0895-8.

Oxford Nanopore Technologies (2018) *Tombo 1.5.1 documentation*. Available at: https://nanoporetech.github.io/tombo/ (Accessed: 28 July 2023).

Palenchar, J.B. and Bellofatto, V. (2006) ‘Gene transcription in trypanosomes’, *Molecular and Biochemical Parasitology*, 146(2), pp. 135–141. Available at: https://doi.org/10.1016/j.molbiopara.2005.12.008.

Pavlov, Y.I., Newlon, C.S. and Kunkel, T.A. (2002) ‘Yeast origins establish a strand bias for replicational mutagenesis’, *Molecular Cell*, 10(1), pp. 207–213. Available at: https://doi.org/10.1016/s1097-2765(02)00567-1.

Peacock, L. *et al.* (2011) ‘Identification of the meiotic life cycle stage of Trypanosoma brucei in the tsetse fly’, *Proceedings of the National Academy of*

Sciences, 108(9), pp. 3671–3676. Available at: https://doi.org/10.1073/pnas.1019423108.

Peacock, L. *et al.* (2014) ‘Meiosis and Haploid Gametes in the Pathogen Trypanosoma brucei’, *Current Biology*, 24(2), pp. 181–186. Available at: https://doi.org/10.1016/j.cub.2013.11.044.

Petermann, E., Lan, L. and Zou, L. (2022) ‘Sources, resolution and physiological relevance of R-loops and RNA–DNA hybrids’, *Nature Reviews Molecular Cell Biology*, 23(8), pp. 521–540. Available at: https://doi.org/10.1038/s41580-022-00474-x.

Potvin, J.-É. *et al.* (2023) ‘Increased copy number of the target gene squalene monooxygenase as the main resistance mechanism to terbinafine in Leishmania infantum’, *International Journal for Parasitology. Drugs and Drug Resistance*, 23, pp. 37–43. Available at: https://doi.org/10.1016/j.ijpddr.2023.09.001.

Pursell, Z.F. *et al.* (2007) ‘Yeast DNA polymerase epsilon participates in leading-strand DNA replication’, *Science (New York, N.Y.)*, 317(5834), pp. 127–130. Available at: https://doi.org/10.1126/science.1144067.

Quinlan, A.R. and Hall, I.M. (2010) ‘BEDTools: a flexible suite of utilities for comparing genomic features’, *Bioinformatics*, 26(6), pp. 841–842. Available at: https://doi.org/10.1093/bioinformatics/btq033.

Reis-Cunha, J.L. *et al.* (2024) ‘Ancestral aneuploidy and stable chromosomal duplication resulting in differential genome structure and gene expression control in trypanosomatid parasites’, *Genome Research*, 34(3), pp. 441–453. Available at: https://doi.org/10.1101/gr.278550.123.

Reis-Cunha, J.L. and Jeffares, D.C. (2024) ‘Detecting complex infections in trypanosomatids using whole genome sequencing’, *BMC Genomics*, 25(1), p. 1011. Available at: https://doi.org/10.1186/s12864-024-10862-6.

Reis-Cunha, J.L., Valdivia, H.O. and Bartholomeu, D.C. (2018) ‘Gene and Chromosomal Copy Number Variations as an Adaptive Mechanism Towards a Parasitic Lifestyle in Trypanosomatids’, *Current Genomics*, 19(2), pp. 87–97. Available at: https://doi.org/10.2174/1389202918666170911161311.

Reynolds, D. *et al.* (2014) ‘Regulation of transcription termination by glucosylated hydroxymethyluracil, base J, in Leishmania major and Trypanosoma brucei’, *Nucleic Acids Research*, 42(15), pp. 9717–9729. Available at: https://doi.org/10.1093/nar/gku714.

Reynolds, D. *et al.* (2016) ‘Histone H3 Variant Regulates RNA Polymerase II Transcription Termination and Dual Strand Transcription of siRNA Loci in Trypanosoma brucei’, *PLoS genetics*, 12(1), p. e1005758. Available at: https://doi.org/10.1371/journal.pgen.1005758.

Robinson, N.P. *et al.* (2002) ‘Inactivation of Mre11 Does Not Affect *VSG* Gene Duplication Mediated by Homologous Recombination in *Trypanosoma brucei*’, *Journal of Biological Chemistry*, 277(29), pp. 26185–26193. Available at: https://doi.org/10.1074/jbc.M203205200.

Robinson, P.S. *et al.* (2021) 'Increased somatic mutation burdens in normal human cells due to defective DNA polymerases', *Nature Genetics*, 53(10), pp. 1434–1442. Available at: https://doi.org/10.1038/s41588-021-00930-y.

Rogers, M.B. *et al.* (2011) 'Chromosome and gene copy number variation allow major structural change between species and strains of Leishmania', *Genome Research*, 21(12), pp. 2129–2142. Available at: https://doi.org/10.1101/gr.122945.111.

Sakofsky, C.J. *et al.* (2015) 'Translesion Polymerases Drive Microhomology-Mediated Break-Induced Replication Leading to Complex Chromosomal Rearrangements', *Molecular Cell*, 60(6), pp. 860–872. Available at: https://doi.org/10.1016/j.molcel.2015.10.041.

Schmid-Hempel, P. *et al.* (2018) 'The genomes of Crithidia bombi and C. expoeki, common parasites of bumblebees', *PLOS ONE*, 13(1), p. e0189738. Available at: https://doi.org/10.1371/journal.pone.0189738.

Schulz, D. *et al.* (2016) 'Base J and H3.V Regulate Transcriptional Termination in Trypanosoma brucei', *PLoS genetics*, 12(1), p. e1005762. Available at: https://doi.org/10.1371/journal.pgen.1005762.

Searle, B. *et al.* (2023) 'Third-Generation Sequencing of Epigenetic DNA', *Angewandte Chemie International Edition*, 62(14), p. e202215704. Available at: https://doi.org/10.1002/anie.202215704.

Sedwick, W.D., Brown, O.E. and Glickman, B.W. (1986) 'Deoxyuridine misincorporation causes site-specific mutational lesions in the lacI gene of Escherichia coli', *Mutation Research*, 162(1), pp. 7–20. Available at: https://doi.org/10.1016/0027-5107(86)90066-7.

Shah, J.S. *et al.* (1987) 'The 5' flanking sequence of a Trypanosoma brucei variable surface glycoprotein gene', *Molecular and Biochemical Parasitology*, 24(2), pp. 163–174. Available at: https://doi.org/10.1016/0166-6851(87)90103-4.

Shapiro, T.A. and Englund, P.T. (1995) 'The Structure and Replication of Kinetoplast Dna', *Annual Review of Microbiology*, 49(1), pp. 117–143. Available at: https://doi.org/10.1146/annurev.mi.49.100195.001001.

Sharp, P.M. and Li, W.H. (1987) 'The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications.', *Nucleic Acids Research*, 15(3), pp. 1281–1295.

Sheader, K. *et al.* (2003) 'Delineation of the regulated Variant Surface Glycoprotein gene expression site domain of Trypanosoma brucei', *Molecular and Biochemical Parasitology*, 128(2), pp. 147–156. Available at: https://doi.org/10.1016/S0166-6851(03)00056-2.

Shinbrot, E. *et al.* (2014) 'Exonuclease mutations in DNA polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication', *Genome Research*, 24(11), pp. 1740–1750. Available at: https://doi.org/10.1101/gr.174789.114.

Siegel, T.N. *et al.* (2009) ‘Four histone variants mark the boundaries of polycistronic transcription units in *Trypanosoma brucei*’, *Genes & Development*, 23(9), pp. 1063–1076. Available at: https://doi.org/10.1101/gad.1790409.

Siegel, T.N. *et al.* (2011) ‘Gene expression in Trypanosoma brucei: lessons from high-throughput RNA sequencing’, *Trends in Parasitology*, 27(10), pp. 434–441. Available at: https://doi.org/10.1016/j.pt.2011.05.006.

da Silva, M.S. *et al.* (2019) ‘Transcription activity contributes to the firing of non-constitutive origins in African trypanosomes helping to maintain robustness in S-phase duration’, *Scientific Reports*, 9(1), p. 18512. Available at: https://doi.org/10.1038/s41598-019-54366-w.

Simpson, A.G.B. and Roger, A.J. (2004) ‘The real “kingdoms” of eukaryotes’, *Current biology: CB*, 14(17), pp. R693-696. Available at: https://doi.org/10.1016/j.cub.2004.08.038.

Simpson, A.G.B., Stevens, J.R. and Lukeš, J. (2006) ‘The evolution and diversity of kinetoplastid flagellates’, *Trends in Parasitology*, 22(4), pp. 168–174. Available at: https://doi.org/10.1016/j.pt.2006.02.006.

Sistrom, M. *et al.* (2014) ‘Comparative genomics reveals multiple genetic backgrounds of human pathogenicity in the Trypanosoma brucei complex’, *Genome Biology and Evolution*, 6(10), pp. 2811–2819. Available at: https://doi.org/10.1093/gbe/evu222.

Sloof, P. *et al.* (1983) ‘Size fractionation of Trypanosoma brucei DNA: localization of the 177-bp repeat satellite DNA and a variant surface glycoprotein gene in a mini-chromosomal DNA fraction.’, *Nucleic Acids Research*, 11(12), pp. 3889–3901.

Sonnhammer, E.L.L. and Durbin, R. (1995) ‘A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis’, *Gene*, 167(1), pp. GC1–GC10. Available at: https://doi.org/10.1016/0378-1119(95)00714-8.

Spivak, G. and Ganesan, A.K. (2014) ‘The complex choreography of transcription-coupled repair’, *DNA Repair*, 19, pp. 64–70. Available at: https://doi.org/10.1016/j.dnarep.2014.03.025.

Sreekumar, L. *et al.* (2021) ‘Orc4 spatiotemporally stabilizes centromeric chromatin’, *Genome Research*, 31(4), pp. 607–621. Available at: https://doi.org/10.1101/gr.265900.120.

Srivatsan, A. *et al.* (2010) ‘Co-Orientation of Replication and Transcription Preserves Genome Integrity’, *PLOS Genetics*, 6(1), p. e1000810. Available at: https://doi.org/10.1371/journal.pgen.1000810.

van Steensel, B. and Belmont, A.S. (2017) ‘Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression’, *Cell*, 169(5), pp. 780–791. Available at: https://doi.org/10.1016/j.cell.2017.04.022.

Steinbiss, S. *et al.* (2016) ‘Companion: a web server for annotation and analysis of parasite genomes’, *Nucleic Acids Research*, 44(W1), pp. W29-34. Available at: https://doi.org/10.1093/nar/gkw292.

Sterkers, Y. *et al.* (2011) ‘FISH analysis reveals aneuploidy and continual generation of chromosomal mosaicism in Leishmania major’, *Cellular Microbiology*, 13(2), pp. 274–283. Available at: https://doi.org/10.1111/j.1462-5822.2010.01534.x.

Sueoka, N. (1962) ‘ON THE GENETIC BASIS OF VARIATION AND HETEROGENEITY OF DNA BASE COMPOSITION’, *Proceedings of the National Academy of Sciences of the United States of America*, 48(4), pp. 582–592.

Svejstrup, J.Q. (2002) ‘Mechanisms of transcription-coupled DNA repair’, *Nature Reviews Molecular Cell Biology*, 3(1), pp. 21–29. Available at: https://doi.org/10.1038/nrm703.

Takemata, N., Samson, R.Y. and Bell, S.D. (2019) ‘Physical and Functional Compartmentalization of Archaeal Chromosomes’, *Cell*, 179(1), pp. 165-179.e18. Available at: https://doi.org/10.1016/j.cell.2019.08.036.

Tan, K.-T. *et al.* (2022) ‘Identifying and correcting repeat-calling errors in nanopore sequencing of telomeres’, *Genome Biology*, 23(1), p. 180. Available at: https://doi.org/10.1186/s13059-022-02751-6.

Thivolle, A. *et al.* (2021) ‘DNA double strand break position leads to distinct gene expression changes and regulates VSG switching pathway choice’, *PLoS pathogens*, 17(11), p. e1010038. Available at: https://doi.org/10.1371/journal.ppat.1010038.

Tiengwe, C., Marcello, L., Farr, H., Dickens, N., *et al.* (2012) ‘Genome-wide Analysis Reveals Extensive Functional Interaction between DNA Replication Initiation and Transcription in the Genome of Trypanosoma brucei’, *Cell Reports*, 2(1), pp. 185–197. Available at: https://doi.org/10.1016/j.celrep.2012.06.007.

Tiengwe, C., Marcello, L., Farr, H., Gadelha, C., *et al.* (2012) ‘Identification of ORC1/CDC6-Interacting Factors in Trypanosoma brucei Reveals Critical Features of Origin Recognition Complex Architecture’, *PLoS ONE*. Edited by C.A. Nieduszynski, 7(3), p. e32674. Available at: https://doi.org/10.1371/journal.pone.0032674.

Tiengwe, C., Marques, C.A. and McCulloch, R. (2014) ‘Nuclear DNA replication initiation in kinetoplastid parasites: new insights into an ancient process’, *Trends in Parasitology*, 30(1), pp. 27–36. Available at: https://doi.org/10.1016/j.pt.2013.10.009.

Touchon, M. *et al.* (2004) ‘Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes’, *Nucleic Acids Research*, 32(17), pp. 4969–4978. Available at: https://doi.org/10.1093/nar/gkh823.

Touchon, M. *et al.* (2005) ‘Replication-associated strand asymmetries in mammalian genomes: Toward detection of replication origins’, *Proceedings of* *the National Academy of Sciences*, 102(28), pp. 9836–9841. Available at: https://doi.org/10.1073/pnas.0500577102.

Trindade, S. *et al.* (2016) 'Trypanosoma brucei Parasites Occupy and Functionally Adapt to the Adipose Tissue in Mice', *Cell Host & Microbe*, 19(6), pp. 837–848. Available at: https://doi.org/10.1016/j.chom.2016.05.002.

Ubeda, J.-M. *et al.* (2008) 'Modulation of gene expression in drug resistant Leishmania is associated with gene amplification, gene deletion and chromosome aneuploidy', *Genome Biology*, 9(7), p. R115. Available at: https://doi.org/10.1186/gb-2008-9-7-r115.

Van der Ploeg, L.H. *et al.* (1984) 'Chromosomes of kinetoplastida.', *The EMBO Journal*, 3(13), pp. 3109–3115.

Van der Ploeg, L.H.T. *et al.* (1984) 'Antigenic variation in trypanosoma brucei analyzed by electrophoretic separation of chromosome-sized DNA molecules', *Cell*, 37(1), pp. 77–84. Available at: https://doi.org/10.1016/0092-8674(84)90302-7.

Vanhamme, L. *et al.* (2000) 'Differential RNA elongation controls the variant surface glycoprotein gene expression sites of Trypanosoma brucei', *Molecular Microbiology*, 36(2), pp. 328–340. Available at: https://doi.org/10.1046/j.1365-2958.2000.01844.x.

van Luenen, H.G.A.M. *et al.* (2012) 'Glucosylated Hydroxymethyluracil, DNA Base J, Prevents Transcriptional Readthrough in Leishmania', *Cell*, 150(5), pp. 909–921. Available at: https://doi.org/10.1016/j.cell.2012.07.030.

Venkatesan, R.N. *et al.* (2007) 'Mutation at the Polymerase Active Site of Mouse DNA Polymerase δ Increases Genomic Instability and Accelerates Tumorigenesis', *Molecular and Cellular Biology*, 27(21), pp. 7669–7682. Available at: https://doi.org/10.1128/MCB.00002-07.

Vollger, M.R. *et al.* (2022) 'StainedGlass: interactive visualization of massive tandem repeat structures with identity heatmaps', *Bioinformatics*, 38(7), pp. 2049–2051. Available at: https://doi.org/10.1093/bioinformatics/btac018.

Walker, B.J. *et al.* (2014) 'Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement', *PLOS ONE*, 9(11), p. e112963. Available at: https://doi.org/10.1371/journal.pone.0112963.

Wang, Yuchuan *et al.* (2021) 'SPIN reveals genome-wide landscape of nuclear compartmentalization', *Genome Biology*, 22(1), p. 36. Available at: https://doi.org/10.1186/s13059-020-02253-3.

Wang, Yunhao *et al.* (2021) 'Nanopore sequencing technology, bioinformatics and applications', *Nature Biotechnology*, 39(11), pp. 1348–1365. Available at: https://doi.org/10.1038/s41587-021-01108-x.

Wedel, C. *et al.* (2017) 'GT-rich promoters can drive RNA pol II transcription and deposition of H2A.Z in African trypanosomes', *The EMBO Journal*, 36(17), pp. 2581–2594. Available at: https://doi.org/10.15252/embj.201695323.

Weiden, M. *et al*. (1991) ‘Chromosome Structure: DNA Nucleotide Sequence Elements of a Subset of the Minichromosomes of the Protozoan Trypanosoma brucei’, *Molecular and Cellular Biology*, 11(8), pp. 3823–3834. Available at: https://doi.org/10.1128/mcb.11.8.3823-3834.1991.

Wheeler, R.J., Gluenz, E. and Gull, K. (2011) ‘The cell cycle of Leishmania: morphogenetic events and their implications for parasite biology’, *Molecular Microbiology*, 79(3), pp. 647–662. Available at: https://doi.org/10.1111/j.1365-2958.2010.07479.x.

White, L.K. and Hesselberth, J.R. (2022) ‘Modification mapping by nanopore sequencing’, *Frontiers in Genetics*, 13. Available at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1037134 (Accessed: 12 September 2023).

Wickstead, B., Ersfeld, K. and Gull, K. (2003) ‘The mitotic stability of the minichromosomes of *Trypanosoma brucei*’, *Molecular and Biochemical Parasitology*, 132(2), pp. 97–100. Available at: https://doi.org/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(6), pp. 1014–1024. Available at: https://doi.org/10.1101/gr.2227704.

World Health Organization (2023) *Leishmaniasis*. Available at: https://www.who.int/news-room/fact-sheets/detail/leishmaniasis (Accessed: 1 October 2023).

Xia, X. (2012) ‘DNA replication and strand asymmetry in prokaryotic and mitochondrial genomes’, *Current Genomics*, 13(1), pp. 16–27. Available at: https://doi.org/10.2174/138920212799034776.

Xu, W. *et al*. (2017) ‘The R-loop is a common chromatin feature of the Arabidopsis genome’, *Nature Plants*, 3(9), pp. 704–714. Available at: https://doi.org/10.1038/s41477-017-0004-x.

Zhang, T. *et al*. (2024) ‘Nanopore sequencing: flourishing in its teenage years’, *Journal of Genetics and Genomics*, 51(12), pp. 1361–1374. Available at: https://doi.org/10.1016/j.jgg.2024.09.007.

Zomerdijk, J.C. *et al*. (1990) ‘The promoter for a variant surface glycoprotein gene expression site in Trypanosoma brucei.’, *The EMBO Journal*, 9(9), pp. 2791–2801.

Zomerdijk, J.C. *et al*. (1991) ‘Antigenic variation in Trypanosoma brucei: a telomeric expression site for variant-specific surface glycoprotein genes with novel features.’, *Nucleic Acids Research*, 19(7), pp. 1359–1368.

Zomerdijk, J.C., Kieft, R. and Borst, P. (1992) ‘A ribosomal RNA gene promoter at the telomere of a mini-chromosome in Trypanosoma brucei’, *Nucleic Acids Research*, 20(11), pp. 2725–2734. Available at: https://doi.org/10.1093/nar/20.11.2725.