Understanding the Genomic Relationship Between Nuclear DNA Replication and Genome Plasticity in Kinetoplastid Genomes

Samantha Jane Campbell

BSc, MSc

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

Wellcome Centre for Molecular Parasitology

Institute of Infection, Immunity and Inflammation

College of Medical, Veterinary and Life Sciences

University of Glasgow

January 2019

Abstract

DNA replication is an essential process in all eukaryotes initiated from sites termed origins of replication. Recent studies in the kinetoplastid species *Leishmania* and *Trypanosoma brucei* have revealed striking differences in the process of DNA replication between the largely syntenic genomes. *T. brucei* replication origins are generally consistent with previous eukaryotic models while *Leishmania* chromosomes appears to contain a single major origin, as is observed in bacteria, although how the parasites can complete replication in this manner remains unknown. Sites of DNA replication co-localise to strand switch regions where transcription initiation and termination also occur. However, not every strand switch region contains an origin of replication and differences between those containing an origin and those without have not been identified. The use of a variety of computational approaches, including machine learning, allow the investigation of origins of replication in both *Leishmania* and *Trypanosoma brucei* at the DNA sequence level and within the structure of the surrounding genomic context and further characterization of the different classes of strand switch region.

A significant feature of the *Leishmania* genome is its ability to adapt in response to environmental pressures through copy number variation of genes and chromosomes and the formation of episomes, allowing the parasites to evade the host immune system and rapidly develop drug resistance through modulation of gene expression. Analysis of the sequence and structure of the *Leishmania mexicana* genome in serial passage conditions provides insight into the mechanisms underlying genome plasticity and presents a novel hypothesis explaining the potential relationship with DNA replication.

Table of Contents

List of Tables

List of Figures

List of Accompanying Material

Github repositories containing Java and Python scripts can be found at:

https://github.com/CampbellSam

Acknowledgements

I would like to thank my supervisors, Dr Nicholas Dickens, Dr Richard McCulloch and Dr Richard Burchmore, for accepting me as a PhD student and giving me the opportunity to work on an interesting project in a fascinating area of research. And to Dr Kathryn Crouch, for providing technical bioinformatics support and temporary day to day supervision.

I would like to acknowledge my assessors, Dr Michael Barrett and Dr William Weir, who provided critical and essential annual feedback throughout my four years of study and especially through difficult transition periods.

And to Dr Catarina Marques who laid the foundations for this project and continually inspires with her enthusiasm and passion for research.

I would also like to thank the Doctoral Training Centre in Cell and Proteomic Technologies for funding this project through the EPSRC and BBSRC research councils, and the college of Medical, Veterinary and Life Sciences at the University of Glasgow for funding my final thesis-pending year.

Also, Suzann Rundell, who frequently helped me navigate the administrative systems and Sue Barnett and the PGR team at the Institute of Infection, Immunity and Inflammation for answering my endless questions.

To everyone that performed the molecular biology work and is mentioned throughout including Dr Samuel Duncan and Dr Jennifer Ann Stortz for patiently explaining how the data is generated and allowing me to analyse it.

I would like to thank all the current and previous members of the Wellcome Centre for Molecular Parasitology Bioinformatics team and the colleagues, friends and officemates who created a welcoming and pleasant work environment.

Finally, I would like to thank my family and friends for their constant and continued support.

Author's Declaration

I here declare that this thesis and the results herein presented are a result of my own work, except where otherwise stated and acknowledged. None of the results herein presented have been used previously to obtain a degree at any university.

Samantha Jane Campbell

List of Abbreviations

ARS	autonomous replication sequence
ATP	adenosine triphosphate
BES	bloodstream expression site
bp	base pair
BSF	bloodstream form
ChIP	chromatin immunoprecipitation
CNV	copy number variation
DGC	directional gene cluster
DNA	deoxyribonucleic acid
DSB	double strand break
dNTP	deoxyribonucleotide triphosphate
ESAG	expression site-associated gene
FACS	fluorescence activated cell sorting
G4	G-quadruplex
gDNA	genomic DNA
HAT	human African trypanosomiasis
HR	homologous recombination
HU	hydroxyurea
Kbp	kilobase pairs
kDNA	kinetoplast DNA
Mbp	megabase pairs
MFA-seq	marker frequency analysis by deep sequencing
MMEJ	microhomology-mediated end-joining
mRNA	messenger RNA
NGS	next generation sequencing
NHEJ	non-homologous end joining
OGRE	origin G-rich repeated elements
ORB	origin recognition box
ORC	origin recognition complex
PCF	procyclic form
PCR	polymerase chain reaction
PKDL	post kala-azar dermal leishmaniasis
pre-RC	pre-replication complex
qPCR	quantitative real-time PCR
RNA	ribonucleic acid

RNA pol I	RNA polymerase I
RNA pol II	RNA polymerase II
RNA pol III	RNA polymerase III
rRNA	ribosomal RNA
SSR	strand switch region
SVM	support vector machine
tRNA	transfer RNA
TSS	transcription start site
UTR	untranslated region
VSG	variant surface glycoprotein

1 General Introduction

1.1 Biology of Kinetoplastid genomes

1.1.1 *Trypanosoma brucei* and *Leishmania spp.*

Leishmania and *Trypanosoma* are closely related protozoan parasites belonging to the order Trypanosomatida and class Kinetoplastea. These organisms are characterised by the presence of a kinetoplast, a cellular structure containing the cell's highly unusual mitochondrial DNA (kDNA) (Adl et al., 2012; Liu et al., 2005).

The *L. major* and *T. brucei* genomes display a high level of synteny, with ~70% of genes found in similar genomic context (El-Sayed et al, 2005): an overview of this is shown in Figure 1.1. Despite this high level of synteny, the clinical disease manifestations, infection and transmission approaches, and strategies for immune evasion differ greatly between the two genera (El-Sayed et al., 2005a;

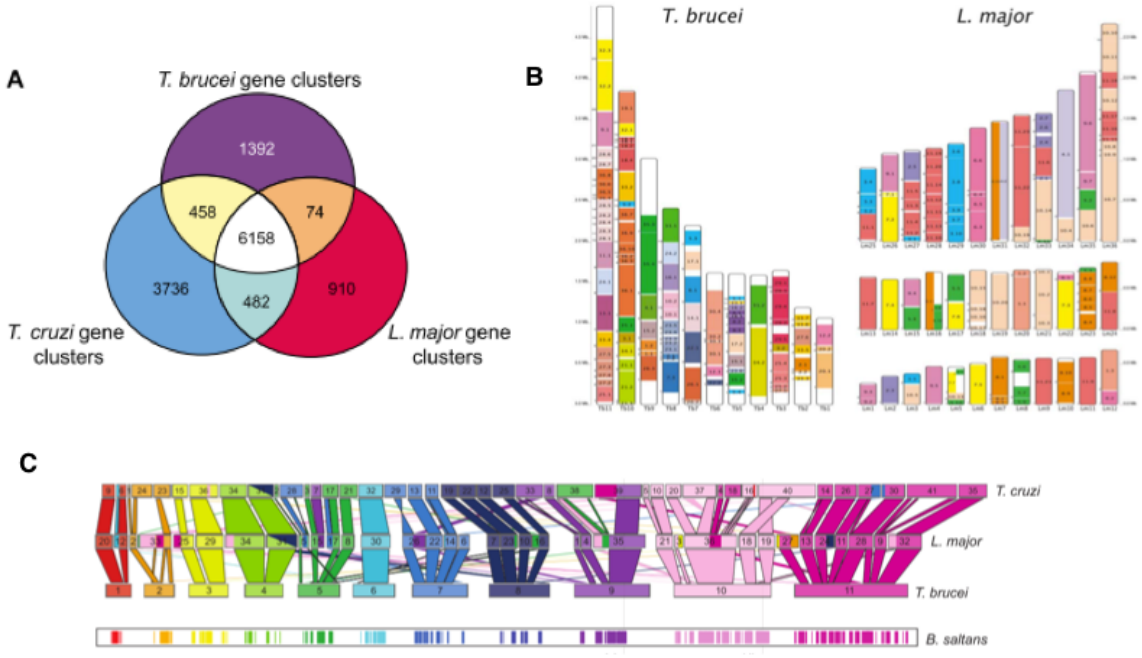


Figure 1.1 Conservation of genomic material between *T. brucei, T. cruzi* and *L. major.*

A. Gene clusters shared between and unique to *T. brucei*, *T. cruzi* and *L. major*. Adapted from El-Sayed et al, 2005**.** Currently requesting permission for reuse. **B.** Diagram representing the location of synteny between *T. brucei* and *L. major*. The left panel contains the 11 *T. brucei* chromosomes colour coded in 36 colours which represent the chromosomes of *L. major*. The right panel shows the 36 chromosomes of *L. major* with 11 colours depicting the corresponding *T. brucei* chromosomes. Adapted from El-Sayed et al, 2005**. C.** Coloured blocks depicting homologous regions across the genomes of *T. brucei, T. cruzi* and *L. major*. Syntenic regions between *T. brucei* and *B. saltans*, a non-parasitic kinetoplastid, are also included. Adapted with permission from (Jackson et al., 2015).

Jackson et al., 2015). Comparative genomics between these genera has revealed very few species-specific genes and the cause of the varying clinical manifestations of *Leishmania* species is currently unclear (Rogers et al., 2011).

1.1.2 Life cycle, vector and transmission of *Trypanosoma brucei*

Trypanosomiasis affects poor rural populations in sub-Saharan Africa, Asia and South America making development of these areas more difficult. *T. brucei* infects a host via an infected tsetse fly vector. Sleeping sickness disease caused by *Trypanosoma brucei* species affects both humans (human African trypanosomiasis (HAT)) and animals and contributes a significant medical and economic burden to the developing countries affected. *T. brucei brucei* infects cattle and domestic animals, while *T. brucei gambiense* and *T. brucei rhodesiense* cause disease in humans in West and East Sub-Saharan Africa, respectively (Franco et al., 2014). Animal trypanosomiasis, also known as nagana, is also caused by the species *Trypanosoma vivax* and *Trypanosoma congolense* in Africa. *Trypanosoma cruzi* causes Chagas disease across large parts of the Americas.

An overview of the lifecycle of the *T. brucei* parasite can be found in figure 1.2. The non-replicating metacyclic form of the parasite is found in the tsetse fly salivary gland and is injected into the bloodstream during a blood meal. The cells then differentiate into the slender replicative bloodstream form (BSF) and establish an infection in the host. BSF *T. brucei* differentiates in the host to a non-replicative short stumpy form which can be taken up by the tsetse fly during a blood meal. In the midgut of the tsetse fly, the short stumpy form then differentiates into the replicative procyclic form (PCF).

BSF *Trypanosoma brucei* evades the host immune system through antigenic variation employed by the periodic switching of variant surface glycoproteins (VSGs), which form a dense coat. The VSG genes are organized into arrays in the sub-telomeric and telomeric regions of the chromosomes , although VSGs are also present among the intermediate chromosomes and mini-chromosomes (Glover et al., 2013). Only one VSG is expressed at a time, with all others suppressed. The C-terminal domain is conserved while the N-terminal is exposed

to the immune system and hypervariable among different VSGs. The basic organisation of a VSG includes one or more 70 bp repeats at the 5' end and sequence homology at the 3' end. The bloodstream expression sites (BESs) are located adjacent to the telomeres and include five to ten ESAGs that are co-transcribed with the active VSG, with all BES to containing ESAG 6 and ESAG7 (Becker et al., 2004). The expressed VSG is switched into the active BES where it is transcribed by RNA pol I. This switching process is known to occur through gene duplication and recombination of VSG cassettes and more recently by RAD51-independent pathways such as microhomology-mediated end-joining (MMEJ) that repairs double strand breaks (DSBs) using 5-25bp of imperfectly matched sequence (Glover et al., 2011). The mechanism of VSG switching creates diversity within the population, which allows the parasite to maintain infection within the host and continue to evade the immune response. Maintaining VSG transcriptional control is essential for parasite survival and coordination between this process and DNA replication has been suggested by studies based on *T. brucei* ORC1/CDC6 (a factor of the origin recognition complex related to Orc1). Following knockdown of ORC1/CDC6, metacyclic VSGS are derepressed in PCF cells and BESs are derepressed in BSF cells and PCF cells, although to a lesser extent in the insect-stage cells (Benmerzouga et al., 2013; Tiengwe et al., 2012). It is becoming clear that antigenic variation and DNA replication are linked but there is currently no evidence that DNA replication drives antigenic variation and these processes may be linked at a regulatory level to achieve optimal efficiency.

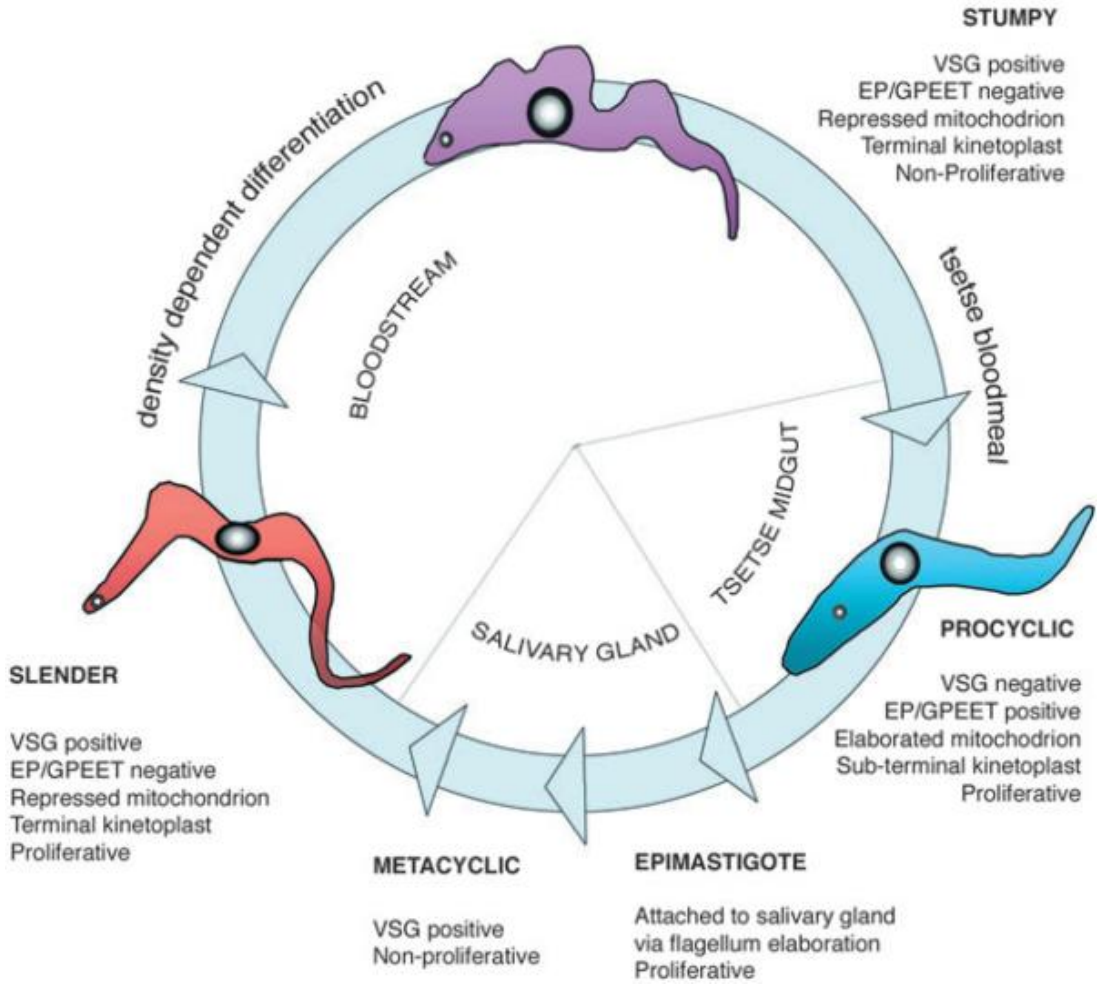


Figure 1.2. Lifecycle of *Trypanosoma brucei*.

An overview of the *T. brucei* lifecycle between the mammalian host and tsetse fly vector. Slender BSF trypanosomes proliferate in the host bloodstream, expressing a VSG coat to evade the host immune system. Cells begin to differentiate to stumpy form as parasite numbers increase. The non-replicating stumpy form is then taken up by the tsetse fly during a bloodmeal. Parasites then differentiate into procyclic form in the fly midgut. These cells no longer have a VSG coat but instead express EP and GPEET procyclins. Procyclic form cells migrate to the salivary gland and attach as epimastigotes, which then generate non-dividing metacyclic cells that acquire the VSG coat and are transmitted to the next mammalian host during the tsetse fly bloodmeal. Reproduced with permission from Gull, 2009 [license number: 4597121021546].

1.1.3 Life cycle, vector and transmission of *Leishmania major*

Several *Leishmania* species cause cutaneous and visceral leishmaniasis in humans with varying severity. Leishmaniasis is transmitted to human hosts by phlebotomine sand flies in the Old World and Lutzomyia sand flies in the New World. The parasite then invades and alters the function of macrophage cells in order to avoid the host immune response and the infection results in leishmaniasis (El-Sayed et al., 2005b). The pathology of the disease depends on the *Leishmania* species and also appears to be affected by the host immune response (Kaye & Scott, 2011). There are roughly 20 species of *Leishmania* responsible for varying species-dependent clinical manifestations of leishmaniasis (Alvar et al., 2012). *L. major* and *L. mexicana* cause a cutaneous form of the disease while *L. donovani* and *L. infantum* infect the liver and spleen causing a visceral form of leishmaniasis that can be fatal if untreated. A muco-cutaneous form of the disease is also caused by *L. braziliensis*. In some areas, treated cases of visceral leishmaniasis infection caused by *L. donovani* can reoccur some time after recovery as a cutaneous form known as post kala-azar dermal leishmaniasis (PKDL) (Zijlstra, Musa, Khalil, & Hassan, 2003).

In the host, *Leishmania* parasites exist in an intracellular amastigote form and as motile promastigotes in the sand fly vector. Several forms of promastigote occur in the sand fly vector and this stage of the life cycle has been discovered to be increasingly complex, in great contrast to the amastigote form in the mammalian host (Gossage et al., 2003). An overview of the *Leishmania* lifecycle, where the promastigote stages have been simplified, is shown in figure 1.3. Infective metacyclic promastigotes are transmitted to the host during the blood meal by the sand fly and phagocytosed by a host phagocyte cell, where the parasite then develops into the amastigote form and divides. Infected host phagocytes then lyse and release amastigotes, which then infect further phagocyte cells. The amastigotes taken up by the fly during a blood meal proliferate in the fly midgut and develop into procyclic promastigotes which then undergo several morphological changes to become non-dividing metacylic promastigotes that can infect a new mammalian host when transmitted during a blood meal (Kaye and Scott, 2011).

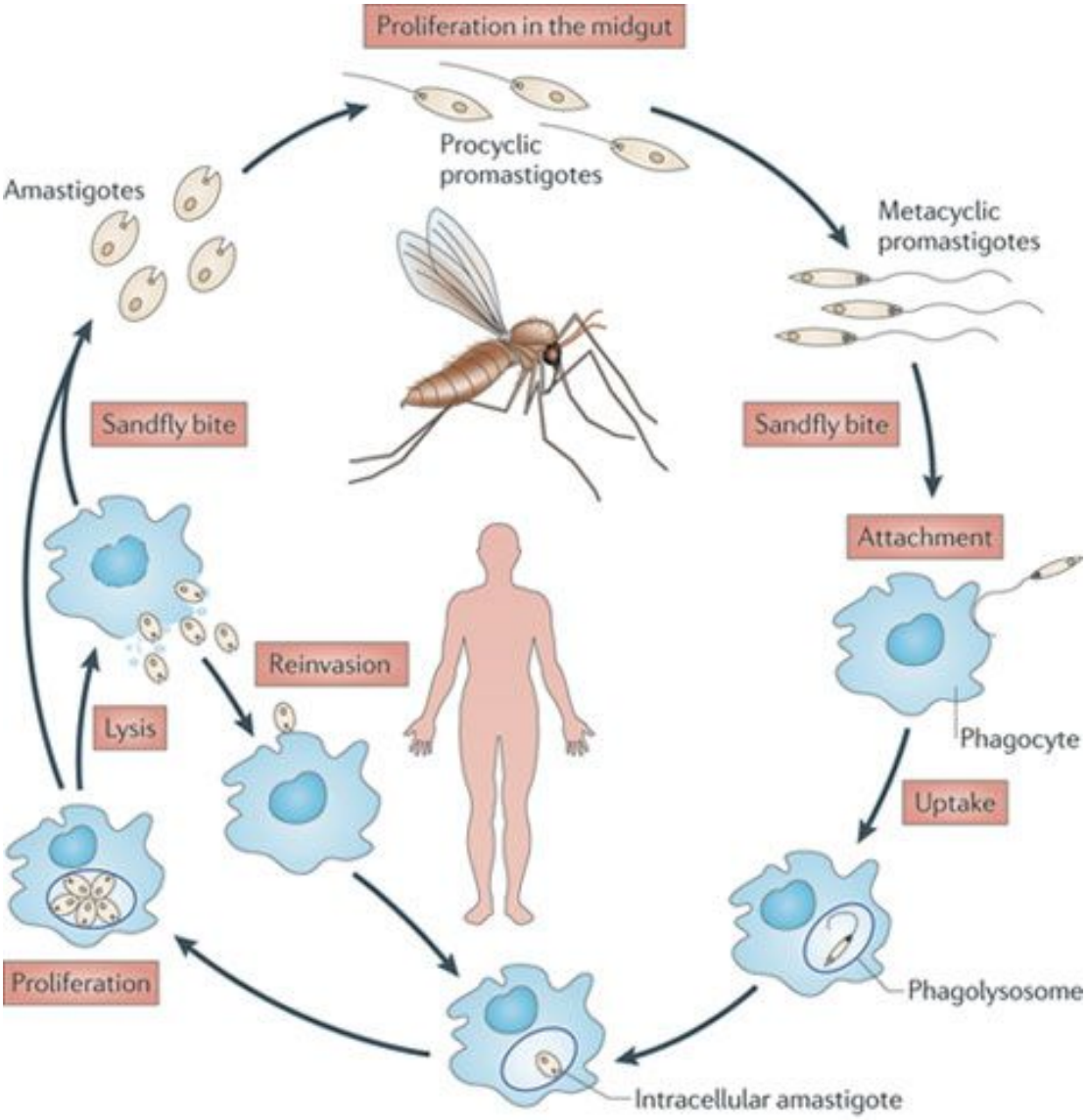


Figure 1.3. The lifecycle of *Leishmania*.

Diagram of the general lifecycle of *Leishmania* parasites. Sandfly vector stage promastigote forms have been simplified. Amastigotes are taken up by phagocyte cells in the mammalian host where they then proliferate and infect further phagocyte cells. Amastigote parasites taken up by the sandfly during a blood meal proliferate in the fly midgut and differentiate into procyclic promastigotes and then metacyclic promastigotes which infect a new mammalian host. Reproduced with permission from Kaye and Scott., 2011 [license number: 4597121427977].

1.1.4 Unusual genome organisation and transcription

Kinetoplastid nuclear genomes are unusual in that virtually all genes are grouped into arrays known as directional gene clusters (DGCs) that are transcribed polycistronically from a single promoter. DGCs can be divergent, convergent or head-to-tail (H-T), depending on the orientation of the genes. The region in between the DGCs is termed a strand switch region (SSR), which are also known as divergent, convergent or H-T based on the surrounding DGCs. An extreme example of this arrangement is chromosome 1 in *L. major*, a small chromosome which contains only two DGCs on opposite strands with a 1.6kb SSR in between (Dubessay, 2002). Polycistronic transcription is also observed in bacterial genomes, but kinetoplastids are distinct as the genes within a DGC do not appear to be grouped by function, and functionally related genes are found in different arrays although there are some notable exceptions including the *T. brucei* tubulin gene array on chromosome 1 (Kelly et al, 2012). The resulting transcripts are *trans*-spliced and polyadenylated: the 5' end of each individual transcript is capped with a 39nt splice leader RNA sequence and the 3' end is polyadenylated ready for translation (Liang et al, 2003). Bidirectional RNA pol II transcription initiation sites are marked by epigenetic factors, primarily modification of histones as well as the presence of histone variants. Transcription initiation sites in *Leishmania* are marked by acetylated H3, and in *T. brucei* display increased levels of acetylated K10 of H4 (H4K10Ac) and variants of H3 and H4 (Thomas et al. 2009; Siegel et al, 2005).

Little is known concerning the termination of RNA Pol II transcription at the end of a DGC. However, it is significant to note that investigation of the hypermodified base J in *Leishmania and T. brucei* has found that this base localizes to the boundaries of transcription units (Cliffe et al, 2010) and the lack of this base causes transcriptional readthrough at termination sites, causing improper transcripts and resulting in death of the parasite (Reynolds et al, 2014; Van Luenen et al, 2012). Base J may therefore be associated with the correct termination of transcription.

A high level of copy number variation is observed in kinetoplastids as a result of polycistronic transcription. A gene with a high copy number will generate a high level of mRNA transcripts in contrast to single copy genes. The position of a gene

within a transcription unit has also been associated with the level of resulting transcripts (Kelly et al, 2012). It was observed that rapidly downregulated genes were more often located close to the transcription initiation site while upregulated genes tended to be distal to the initiation site. GO-term analysis of proximal and distal genes found enrichment of genes associated with translation, the cell cycle and the cytoskeleton proximal to initiation sites while genes involved in transcription and RNA processing were enriched in distal regions (Kelly et al., 2012). This organisation may play a role in the regulation of transcription while allowing variation in gene copy number to generate further diversity between species.

1.1.5 Genome plasticity in *Leishmania spp.*

Mosaic aneuploidy, where cells can sustain uneven numbers of chromosomes in states other than diploid, is a common feature across *Leishmania* species and chromosome number therefore varies between species, strains and cell to cell. This means chromosomes in each cell may be present in more than two ploidy states, ranging from haploid to tetraploid depending on the chromosome, highlighting the level of plasticity in the *Leishmania* genomes (Rogers et al., 2011; Sterkers et al, 2012). *L. major* chromosome 31 is consistently present in a state higher than diploid and this is observed across all studied *Leishmania* species. The consequences of abnormal ploidy are detrimental in most eukaryotes, due to the effects of gene dosage, however *Leishmania* are able to tolerate these changes, perhaps as a result of the unusual transcriptional control in kinetoplastids. Mosaic aneuploidy is used as a mechanism for generating genetic diversity and is thought to be involved in the parasite's ability to quickly adapt to environmental changes, such as drug treatment, and evade the host immune system (Lachaud et al., 2014; Leprohon et al., 2009; Mukherjee et al., 2013).

Copy number variation (CNV) of specific genes can occur as well as whole chromosome CNVs that lead to aneuploidy. Homologous direct or inverted repeats in the DNA sequence have been associated with amplification of specific regions in the genome and the formation of extrachromosomal elements in the form of linear or circular episomes (J. M. Ubeda et al., 2014). The involvement of MRE11 in the formation of linear amplicons and RAD51 recombinase in the

formation of circular amplicons has been demonstrated although the mechanism underlying amplicon formation remains unclear (Laffitte et al., 2014) .

A known source of genomic rearrangement in eukaryotes occurs from errors in DNA replication which can lead to collapsed forks and result in DNA double strand breaks (DSBs) (Gent et al., 2001). DSBs are predominantly resolved by homologous recombination (HR) or non-homologous end joining (NHEJ), although few proteins associated with NHEJ have been identified in *Leishmania* and it is not clear if the process occurs in this parasite genome (Genois et al., 2014; Symington & Gautier, 2011). MRE11(Meiotic Recombination 11) and RAD50 proteins are involved in DNA repair and protection of chromosome ends. The absence of these proteins in *L. infantum* leads to a decrease in the rate of DNA repair by HR and observed chromosomal translocations are associated with a microhomology-mediated end-joining mechanism (MMEJ) that has previously been observed in *T. brucei* (Burton et al., 2007; Glover et al., 2011).

1.2 Initiation of nuclear DNA Replication

DNA replication is an essential process in all organisms that is required for the inheritance of genetic material and organism survival. Strict regulation of the cell cycle and coordination of several factors are required to ensure that complete and faithful duplication of DNA molecules occurs only once during S phase. Initiation of DNA replication involves the recruitment of conserved initiation factors, a replicative helicase and DNA polymerase(s) to begin DNA synthesis (reviewed by Masai et al., 2010) .

The region of DNA sequence where replication initiates is termed the origin of replication. Origins of replication are not well characterized across eukaryotes; with the exception of autonomous replicating sequences (ARS) in budding yeast, *Saccharomyces cerevisiae* and close relatives, there is no known motif or consensus sequence for replication origins in eukaryotic genomes (Nicholas P Robinson & Bell, 2005). Errors in the replication process can have significant consequences that may lead to deleterious effects within the cell and it is

therefore important that each full chromosome replicates correctly, most commonly from multiple origins.

1.2.1 DNA origins of replication in bacteria

A conserved origin sequence has been established in bacterial genomes, a 9 bp sequence called the DnaA box that determines specific sites of replication initiation (Robinson et al, 2005). Most bacterial genomes are circular and also have a single termination region. Bacterial origins of replication are termed *oriC* and contain several DnaA boxes. In *E. coli, oriC* contains five DnaA boxes (Messer, 2002). The DnaA protein, comprised of 4 functional domains, binds to the DnaA boxes in *oriC* and initiates replication when complexed with ATP (Sekimizu et al., 1987). ATP-DnaA then binds additional ATP-DnaA boxes composed of the 5' 6-mer AGATCT which flank the *oriC* site (Schaper & Messer, 1995).

1.2.2 Archaeal origins of replication

In archaea, replication of the circular chromosomes can occur from a single origin or multiple origins (reviewed in (Ausiannikava & Allers, 2017)). The first replication origin in archaea was identified in *Pyrococcus*, which was shown to have a single origin per chromosome (Baldwin et al., 2000). Since then, studies in *Sulfolobus islandicus* and *Haloferax volcanii* have identified three replication origins per chromosome and four in *Pyrobaculum calidifontis* (Lundgren et al., 2004; Norais et al., 2007; Pelve et al., 2013). A conserved motif, termed origin recognition box (ORB), reminiscent of the DnaA box in bacteria, is associated with DNA replication initiation in archaea and is a demonstrated binding site for Orc1/Cdc6 proteins. Although the relatively simple genome is evocative of the organisation observed in bacteria, the replication machinery is distinct and more closely resembles the Orc1 and Cdc6 model associated with eukaryotes providing an interesting combination of features associated with DNA replication (Nicholas P Robinson & Bell, 2005). Archaeal genomes with multiple origins of replication, also contain multiple Orc1/Cdc6 homologs. For example, in *Sulfolobus solfataricus*, there are three origins and three Orc1/Cdc6 homologs, which demonstrate different binding affinities at each origin site where as this does not appear to be the case in the closely related *S. islandicus* where two of the

origin sites are bound exclusively by a specific Orc1/Cdc6 complex (N. P. Robinson & Bell, 2007). Study of distantly related archaeal species demonstrated that the presence of multiple origins is likely to have occurred due to horizontal gene transfer via integration of extrachromosomal elements (Robinson and Bell., 2007).

1.2.3 Origins of DNA replication in eukaryotes

Based on the study of eukaryotic model organisms, it is known that replication is initiated through recruitment and binding of a specific complex of proteins to a region of DNA termed the origin at multiple sites on each chromosome (reviewed by (Burgers & Kunkel, 2017)). This process is tightly temporally and spatially regulated to ensure it only occurs once in each cell every cell cycle. An overview of the proteins involved, and complexes formed during eukaryotic replication initiation is shown in figure 1.4. All potential origins are recognised by the origin recognition complex but only a subset of these are activated in each cell during S phase. The first step of activation involves tightly regulated assembly of the pre-replication complex and licensing of the origin, and occurs during G1 phase. The origin recognition complex (ORC) is composed of six subunits, Orc1-6 (Bell & Stillman, 1992), and ORC-like initiation complexes are conserved in all eukaryotes although binding and recruitment can vary and no specific binding sequence has been identified (Zellner et al, 2007). ORC binds the origin sequence forming a ring structure. Cdc6, bound by ATP, binds to this complex, stabilising the interaction between ORC and the DNA. Binding of Cdc6 changes the complex conformation, allowing the recruitment of a replicative helicase, the minichromosome maintenance (MCM) complex, (MCM2-7) which is loaded onto the complex by Cdt1, to form the pre-RC (Bochman & Schwacha, 2009). The correct formation of the pre-RC (ORC-Cdc6-Cdt1-MCM), signals that the origin is licensed and is ready for replication to occur (Sun et al, 2013). Activation of licensed origins occurs during S phase and is dependent on the levels of two protein kinases, cyclin-dependent kinase (CDK) and Dbf4-dependent kinase (DDK). Assembly of the pre-initiation complex (pre-IC) involves the loading of additional proteins and complexes (general overview in figure 1.4; reviewed by Burgers and Kunkel., 2017). Dpb11, Sld2, Sld3, Sld7, Cdc45 and the GINS complex are recruited along with Pol ε to form the pre-IC. Helicase activation is then achieved by Mcm10 and RPA. Once active, the helicase can

unwind the origin DNA, allowing formation of the replication fork and the initiation of synthesis.

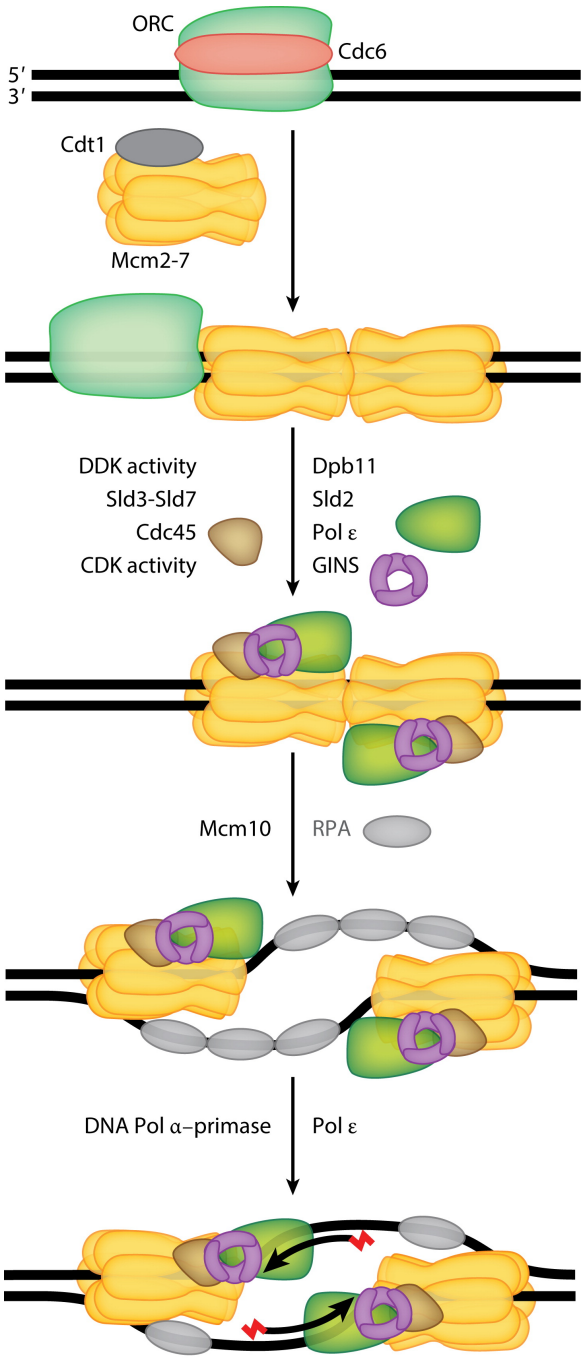


Figure 1.4. Initiation of eukaryotic DNA replication.

General overview of proteins and complexes involved in formation of the pre-replication complex (pre-RC) and pre-initiation complex (pre-IC). Binding of the ORC complex to the DNA is initially stabilised by the binding of ATP-bound Cdc6. This causes a conformational change which allows loading of the helicase complex, Mcm2-7, by Cdt1, forming the pre-RC. Recruitment of additional factors Dpb11, Sld2, Sld3, Sld7, Cdc45 and the GINS complex forms the pre-IC which is ready for helicase activation. Activation of the helicase is performed by Mcm10 and RPA, allowing unwinding of the origin DNA. Reproduced with permission from Burgers and Kunkel, 2017 [license number: 4597760295630].

Despite the lack of a consensus sequence in most well studied eukaryotic genomes, recent studies demonstrate correlation between the presence of G-rich sequences known as origin G-rich repeated elements (OGREs) and the majority of active replication origins (Cayrou et al, 2012; Besnard et al, 2012). The presence of OGREs can predict the formation of G-quadruplex (G4) structures, which potentially have a role in replication initiation and origin activation (Valton et al, 2014). The motif associated with these regions that have the potential to form the quadruplex is currently thought to be $G_{\geq 3}N_xG_{\geq 3}N_xG_{\geq 3}N_xG_{\geq 3}$ (Maizels & Gray, 2013). The quadruplex is formed through the bonding of four guanine nucleotides into a cyclic hydrogen bond arrangement where neighbouring guanines each share a hydrogen bond (Maizels, 2006). Enrichment of the G4 motif is observed in telomeres and origin regions and is thought to be associated with the regulation of replication as well as transcription and translation (Valton et al, 2014). Centromeric regions are known to be very early–replicating in many eukaryotic genomes and generally contain highly repetitive sequences. In many species, centromeres contain 171bp repeat arrays known as alpha satellite DNA (reviewed by (Murphy & Karpen, 1998). Repetitive sequences are often associated with the impairment of DNA replication as these sequences can cause conformational changes and lead to genome instability.

1.2.4 Origins of replication in *Trypanosoma brucei* and *Leishmania spp.*

Recent studies in *T. brucei* have demonstrated that much of the replication machinery is retained in kinetoplastids although divergence is observed between the early-acting proteins involved in initiation (Tiengwe et al, 2014). Identification of kinetoplastid initiation factors orthologous to the ORC subunits and Cdc6 has been difficult, with only one initially determined through homology searches, TbORC1/CDC6 (Godoy et al., 2009). It is now possible to perform sequence homology studies for these proteins as ORC1-ORC5 and CDC6 contain AAA+ domains and are members of the ATPase family (Iyer et al, 2004). Several putative orthologs have been identified but only TbORC1/CDC6 contains this domain along with TbORC1B (Tiengwe et al, 2014). It has therefore been difficult to confirm further orthologous subunits using sequence comparison methods. Although it was previously suggested that the initiator role is filled by

the single factor TbORC1/CDC6, as this is observed in archaea, additional factors of the ORC complex have since been identified (TbORC4, Tb3120 and TbORC1B) (De Melo Godoy et al., 2009; Marques et al., 2016). Therefore, conservation of some proteins involved in the main replication initiation roles have been uncovered in *T. brucei* although some divergence is observed and the presence of additional factors associated with the process of initiation remains unclear (Tiengwe et al, 2014).

ChIP-chip mapping of TbORC1/CDC6 confirmed that this protein is involved in DNA replication initiation as binding is observed at transcription unit boundaries, within the core *T. brucei* genome (Tiengwe et al, 2012). Origins were also mapped using marker frequency analysis (MFAseq) and co-localise with a fraction of the ORC1/CDC6 binding sites, suggesting only some of the sites are activated as origins. Further, more limited work, suggested that HU treatment could lead to activation of at least one further origin in chromosome 1 (Calderano et al., 2015). Despite all origins mapping to SSRs, no common sequence has been observed in these regions, nor is it clear how the origins are distinct from the non-origin SSRs, which also bind ORC1/CDC6. An abundance of TbORC1/CDC6 binding is also observed in the transcriptionally silent arrays of VSGs located at the telomeric and sub-telomeric regions of chromosomes (Tiengwe et al, 2012). Active replication origins have not yet been detected in these regions due to limitations of the MFAseq mapping technique, and the role of TbORC1/CDC6 binding remains unknown. It is plausible that the binding of this protein indicates currently undefined replication or transcriptional activity in VSG arrays (Benmerzouga et al, 2013). The cues responsible for recruitment of TbORC1/CDC6 to the origin are currently unclear, though it may be of interest to study epigenetic features during the initiation of the replication process as this is known to be a common mechanism of regulation in higher eukaryotes.

The *T. brucei* nuclear genome is comprised of 11 Mb chromosomes present in a diploid state but also contains aneuploid intermediate chromosomes and several minichromosomes (Berriman et al., 2005). Multiple early-firing (first half of S phase) origins have been mapped on each of the core 11 chromosomes of *T. brucei* through the integration of TbORC1/CDC6 ChIP-seq data and MFAseq (Tiengwe et al, 2012). MFAseq predicts origin location based on the ratio of reads that align to the genome in different cell cycle stages. This approach

identified 42 origins, mapping to the boundaries of transcription units in the SSR sections of the chromosome. The spacing between each origin is much higher than other eukaryotes, highlighting a low density of origins in the core *T. brucei* genome. No origins are observed in the sub-telomeric VSG arrays despite the high abundance of TbORC1/CDC6 binding. It is plausible that origins have not been detected in these regions because studies have not yet mapped late-firing origins.

At the outset of this project, origins of replication had not yet been mapped in *Leishmania* and therefore no comparisons of origins between *T. brucei* and *Leishmania* and could yet be made. However, MFAseq analysis was applied to *L. major* and *L. mexicana* close to the start of the project, suggesting similarities and differences when comparing the two species (Marques et al., 2015). This is discussed in further detail in chapter 3.

1.2.5 Replication and transcription in kinetoplastid genomes

In both kinetoplastid genomes, the origins of replication map to regions on the chromosome where polycistronic gene arrays diverge or converge, called strand switch regions (SSRs) (Myler et al., 1999). There is a high level of activity at SSRs as initiation and/or termination of transcription is also known to take place at SSRs and therefore RNA pol II binding sites are located here. There is also a strong G/C bias in the *Leishmania* genome that may obscure a motif specific to origin sequences (Martínez-Calvillo, Vizuet-De-Rueda, Florencio-Martínez, Manning-Cela, & Figueroa-Angulo, 2010) . Characterization of origins at the sequence level is therefore confounded by the presence of several other elements involved in a number of processes, making it difficult to identify the origin DNA sequence.

In *T. brucei*, the replication initiator protein TbORC1/CDC6 binds to replication origins at SSRs, sites also marked by RNA pol II promoters, indicating that sites of replication and transcription initiation appear to co-localise. RNAi against TbORC1/CDC6 results in an increase in transcript levels at the boundaries of gene arrays and also a loss of VSG silencing and therefore increased levels of VSG mRNA (Tiengwe et al, 2012; Benmerzouga et al, 2013). These changes may suggest a functional role for TbORC1/CDC6 in transcription as well as

replication, or at least an overlap between the machineries. Although this relationship remains unclear, it provides evidence for the co-ordination of replication and transcription in kinetoplastid genomes, which may involve interaction of the required machinery. This correlation in both position and function of replication and transcriptional initiation may be possible due to the post-transcriptional regulation of gene expression.

There is a high level of constitutive transcriptional activity that traverses most of the genome in kinetoplastids and collisions between replication and transcription are likely occur more often than in other eukaryotes. This is frequently observed in *T. brucei* although a possible explanation for the localization of the replication and transcription initiation sites is to limit these events from happening excessively (Tiengwe et al, 2012).

1.3 Aims and Objectives

The current mapping of DNA origins of replication in kinetoplastid genomes indicates that the organisation of replication initiation sites differs between the closely related genomes of *Trypanosoma brucei* and *Leishmania major*. Due to the unusual structure of kinetoplastid genomes and the location of predicted origins, there is potential to identify the features that comprise an origin at the DNA sequence level. Additionally, the observed differences between these species are particularly interesting as genome plasticity, including mosaic aneuploidy, is a feature of the *Leishmania* genome that is not present in *Trypanosoma brucei*.

The overall goal of this investigation was to further characterise DNA origins of replication in kinetoplastid genomes and understand the potential relationship between the process of DNA replication and the plasticity observed in *Leishmania* genomes. This was tackled as follows:

1.3.1 Objectives and Aims

Objective 1: Characterise origins of replication in *T. brucei* and *Leishmania* at the DNA sequence level

The first objective was to improve the current mapping of origin-containing sites by DNAseq-based marker-frequency analysis (MFAseq) and assess the sensitivity of this approach to confirm the observation of a single dominant origin per chromosome in *Leishmania major* and *L. mexicana*. It was also of interest to compare the MFAseq results with alternative peak-calling software. This was achieved with the following specific aims.

Aim 1.1 Mapping replication origins in kinetoplastid genomes using MFAseq

The existing MFAseq pipeline used to predict sites of DNA replication initiation in *Trypanosoma brucei* and *Leishmania major* and *L. mexicana* was built in Perl and required improvement for speed and simplification. The algorithm was re-written in Java and used to reproduce the previous MFAseq results in *Leishmania* and *T. brucei* for comparison.

Aim 1.2 Assessing the sensitivity of the MFAseq approach in *Leishmania major*

Several simulations were performed in order to assess the reliability of MFAseq output in *Leishmania major*. These included the modelling of an origin-containing region and the sampling of its presence in non-origin containing SSRs belonging to the same chromosome. The presence of individual dominant origins was modelled as well as multiple origins at a lower usage within the population.

Aim 1.3 Refinement of peak calling

Sites of DNA replication initiation predicted by MFAseq are currently determined by eye and there is currently no statistical method implemented to computationally define peaks within this output. A variety of peak calling software designed for analysis of ChIP-seq software is available and a selection

was chosen, based on the current literature, to attempt to solve this problem. To refine the current method of identifying peaks, the ChIP-seq peak calling software was applied to the MFAseq input data, the results were compared and improvements were evaluated.

Aim 1.4 Conservation between species

The current mapping of origins in *T. brucei* and *Leishmania* indicates differences in the genomic organisation of replication initiation sites. A comparison of the similarities and conserved features between the kinetoplastid genomes is discussed. Additional analysis of mapped origins in *T. brucei* is also of interest as little is currently known about replication of BESs, which are now assigned as contigs in the *T. brucei* Lister 427 reference genome. Statistical analysis was performed to validate the observation of a potential DNA replication origin at the BES expressed in this cell line.

Objective 2: Investigate similarities and differences in replication origins across kinetoplastid genomes

As the regions currently predicted to contain an origin in *Leishmania* are large, one of the objectives of this project was to investigate methods of refining the SSRs and the coordinates we consider to contain an origin of replication. Due to the unusual genome structure shared by kinetoplastids it was possible to focus on refining specific regions and compare the origin and non-origin containing SSRs at the DNA sequence level. This was achieved with the following specific aims.

Aim 2.1 Identifying DNA sequence features of origin-containing regions

It was important to identify existing annotated features of SSRs, including histone markers, and determine potential sequence features characteristic of origins e.g. the G4 motif previously associated with sites of DNA replication initiation. Multiple sequence alignment of origin-containing SSRs was performed and motif searching was applied to the SSR sequences. Genomic features,

including SNP rate, were also compared across different SSR types, in the context of origin presence and SSR orientation as well as transcriptional activity, which is predicted based on mapping of histone markers.

Aim 2.2 Refining the SSR coordinates

To update the SSR coordinates in *Leishmania*, existing and updated gene models were examined using RNA-seq datasets. Use of RNA-seq data allowed the *de novo* prediction of gene models which extends the annotations of 3' and 5' UTRs. An alternative method, mapping the coverage of existing RNA-seq data, was also investigated to establish a read depth cut-off threshold that would denote the boundaries of a SSR.

Aim 2.3 Predicting origins across species

The applications of machine learning algorithms were investigated in an effort to determine DNA sequence features characteristic of a DNA replication origin. A support vector machine (SVM) algorithm was applied to the DNA sequence read data across SSRs to classify and predict origin-containing regions. Analysis of the features used by the algorithm makes it possible to discern sequence features that characterise DNA origins of replication.

Objective 3: Use of genomic techniques to investigate the relationship between mosaic aneuploidy and DNA replication in *Leishmania*

The final aim was to understand the potential relationship between DNA replication and the processes underlying genome plasticity in *Leishmania*. It is reasonable to hypothesise that these processes may be linked, particularly as replication may predominantly occur from a single origin on each chromosome in *Leishmania*. This association was investigated with the following specific aims.

Aim 3.1 Identifying a relationship between plasticity and replication

To establish whether it was possible to detect a relationship between the process of DNA replication and genome plasticity, analysis of genomic changes in serially passaged *Leishmania mexicana* promastigotes at the sequence and

structural level was performed. Genomic analyses, including SNP calling and investigation of gene and chromosome copy number variation, was carried out.

Aim 3.2 Plasticity and the cell cycle

A second dataset, in which replication is impeded, was generated to validate the observed relationship between DNA replication from a single dominant origin and chromosomal aneuploidy. The addition of hydroxyurea (HU) is expected to emphasize any changes in chromosome ploidy which occur during DNA replication.

Aim 3.3 Comparison of DNA replication and aneuploidy in *Leishmania*

Based on the results of the previous analyses, it was important to further analyse the relationship between chromosome size and ploidy change throughout passage as the observed changes in chromosome ploidy may be connected to replication from a single origin as the limits of effective replication are reached. This relationship and a potential explanation of the link between replication and aneuploidy are discussed.

2 General Methods

2.1 Next-generation sequencing

Next generation sequencing of all *T. brucei* and *Leishmania* DNA libraries was performed on Illumina sequencing systems at Glasgow Polyomics.

2.1.1 DNA sequencing in *T. brucei* TREU 927

T. brucei TREU 927 DNA libraries were prepared by Dr Calvin Tiengwe and sequenced on an Illumina platform as described in Tiengwe et al., 2012. Reads were trimmed and pre-processed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and aligned to the *T. brucei* TREU 927 reference genome available at TriTrypDB (genome version 26) using Bowtie2 (Langmead & Salzberg, 2012).

2.1.2 Generation of *L. major* Friedlin next-generation sequencing data

DNA libraries were prepared and sequenced on an Illumina MiSeq platform as described in Marques et al., 2015. This work was carried out by Dr Catarina Marques and Craig Lapsley at the Wellcome Centre for Molecular Parasitology. The DNA sequence reads were trimmed using FastQC and aligned to the *L. major* reference genome (obtained from TriTrypDB version 26) using Bowtie2.

2.1.3 DNA sequencing in *L. mexicana* M379

2.1.3.1 Serial Passage of *L. mexicana* M379 promastigotes

$2x10^6$ stationary phase *L. mexicana* M379 promastigotes were injected into the right footpads of Balb/c mice and a lesion was allowed to form. Amastigotes were purified from the footpad lesions then inoculated into promastigotes media and cultivated as promastigotes to stationary phase. The stationary phase culture was split into 3 flasks by diluting 200µL of the original culture into 10 mL new media for each flask. After 7 days of growth to stationary phase, the cells were passaged by adding 100µL of each culture to 5mL new media. If gDNA was extracted, 200µL was added to 10mL of new media. This was continued until passage 29. At passage 16, Balb/c mice were inoculated once more with $2x10^6$ stationary phase *L. mexicana* M379 promastigotes. Genomic DNA was extracted

at passage 0, 5, 10. 16. 20 and 29. This work was performed by Dr Samuel Duncan at the Wellcome Centre for Molecular Parasitology.

DNA library preparation and next generation sequencing was performed by Glasgow Polyomics for all samples. Samples from passage 0 to 10 were sequenced on the Illumina Miseq platform and samples from passage 16 to 29 were sequenced using an Illumina NextSeq 500 system. DNA sequence reads were trimmed and pre-processed using TrimGalore which incorporates FastQC and Cutadapt (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Reads were then aligned to the *L. mexicana* U1103 reference genome (obtained from TriTrypDB version 26) using Bowtie2.

2.1.3.2 *L. mexicana* M379 with HU treatment

L. mexicana M379 promastigotes were obtained from the lab group of Dr Richard Burchmore and set up in cultures of $1x10^6$ cells/ml. This culture underwent 7 passages, maintained below $1x10^7$ cell/ml by passage every 48 hours, before being split into 3 replicates. Each replicate was then split into three once more and each culture exposed to 0mM, 0.1mM or 0.2mM hydroxyurea (HU). Cells underwent 7 further passages with HU added each time. At each time point gDNA was collected along with RNA, cells for DAPI staining and imaging and FACS to monitor cell cycle progression. This work was performed by Dr Jennifer-Ann Stortz.

Genomic DNA libraries from the first triplicate passage and passage 7 under HU treatment were prepared by Craig Lapsley at the Wellcome Centre for Molecular Parasitology and sequenced on the Illumina NextSeq 500 platform by Glasgow Polyomics. DNA sequence reads were trimmed and pre-processed using TrimGalore and aligned to the *L. mexicana* U1103 reference genome (obtained from TriTrypDB version 26) using Bowtie2.

2.1.4 Access of RNA sequencing datasets in *L. major*

20 *L. major* promastigote RNA-seq datasets were downloaded from the sequence read archive (http://www.ncbi.nlm.nih.gov/sra), full details of the data is

provided in Table 2.1. The data was aligned using Hisat2 and the aligned RNA-seq reads were merged into one file using samtools (Kim et al., 2015; Li et al., 2009).

2.2 Java programs

2.2.1 Writing the MFAseq pipeline

The Java program takes the raw genome DNA sequence alignment files for parasite populations sorted for early S phase and G2 phase cell cycle progression. This program is available at (https://github.com/CampbellSam/MFAseq). The plotted ratio is calculated by the number of reads aligning to the region during early S phase of replication and the number of reads aligning during G2 phase, scaled for the total read count in each phase. The output is a list of ratios for the for every 2.5kb window of each chromosome written to a wiggle file. This can then be uploaded as a custom track and viewed on TriTrypDB.org. The original pipeline was rewritten in Java utilizing API (application program interface) functionality from picard provided by samtools (http://www.htslib.org). A Python script was then adapted from this Java program by Dr Nicholas Dickens.

2.2.2 Assessing the sensitivity of the MFAseq analysis in *Leishmania*

The Python script calculates the distribution of reads across the real chromosome 36 origin and simulates this at a chosen region by allocating reads across 10kb regions based on the probability of a read starting within this region. A modified alignment file can then be written containing the real and simulated origins. To simulate the presence of a second origin that is used by 80% the population; the simulated alignment for this region was merged with an unmodified alignment file at a ratio of 5:1. This was repeated for 66% (2:1), 50% (1:1), 33% (1:2) and 17% (1:5). All of the modified and merged alignment files are from early S phase and therefore MFAseq could then be performed against a normal unmodified alignment file from G2 phase to observe the ‘fake origin’ peak sizes using TriTrypDB.org. Peaks considered an origin are determined by eye and there is currently no metric for exclusively determining an origin region, only the SSR it falls within. The SSR coordinates are determined from the

adjacent gene start/end locations. This script is available at (https://github.com/CampbellSam/MFAseq-sensitivity).

Run	Acc. No.	Platform	Read Length	Selection	Layout	Biosample	Description	Published
SRR2136722	SRX1126065	Illumina	101	unspecified	paired	SAMN03945365	Leishmania major, Friedlin strain, metacyclic promastigote, PNA isolation	25/06/2015
SRR2136721	SRX1126064	Illumina Hi-seq	101	unspecified	paired	SAMN03945364	Leishmania major, Friedlin strain, metacyclic promastigote, PNA isolation	25/06/2015
SRR2171252	SRX1126064	Illumina Hi-seq	101	unspecified	paired	SAMN03945364	Leishmania major, Friedlin strain, metacyclic promastigote, PNA isolation	25/06/2015
SRR2136720	SRX1126062	Illumina Hi-seq 1500	101	unspecified	paired	SAMN03945363	Leishmania major, Friedlin strain, metacyclic promastigote, PNA isolation	25/06/2015
SRR2136708	SRX1126060	Illumina Hi-seq 1500	101	unspecified	paired	SAMN03945361	RNA-seq of Leishmania major, Friedlin strain, metacyclic promastigote, PNA isolation	25/06/2015
SRR2136703	SRX1126059	Illumina Hi-seq 1500	101	unspecified	paired	SAMN03945361	Leishmania major, Friedlin strain, metacyclic promastigote, PNA isolation	25/06/2015
SRR2136702	SRX1126058	Illumina Hi-seq 1500	101	unspecified	paired	SAMN03945360	Leishmania major, Friedlin strain, metacyclic promastigote, Ficoll isolation	25/06/2015
SRR1460775	SRX625729	Illumina Hi-seq 1000	101	polyA selectio	paired	SAMN02870610	Leishmania major, Friedlin strain, metacyclic promastigote, Ficoll isolation	25/06/2015
SRR1460774	SRX625728	Illumina Hi-seq 1000	101	polyA selectio	paired	SAMN02870609	Leishmania major, Friedlin strain, metacyclic promastigote, PNA isolation	25/06/2015
SRR1460773	SRX625727	Illumina Hi-seq 1000	101	polyA selectio	paired	SAMN02870608	Leishmania major, Friedlin strain, procyclic promastigote	25/06/2015
SRR1460772	SRX625726	Illumina Hi-seq 1000	101	polyA selectio	paired	SAMN02870607	Leishmania major, Friedlin strain, metacyclic promastigote, Ficoll isolation	25/06/2015
SRR1460771	SRX625725	Illumina Hi-seq 1000	101	polyA selectio	paired	SAMN02870606	Leishmania major, Friedlin strain, metacyclic promastigote, PNA isolation	25/06/2015
SRR1460770	SRX625724	Illumina Hi-seq 1000	101	polyA selectio	paired	SAMN02870605	Leishmania major, Friedlin strain, procyclic promastigote	25/06/2015
SRR1460769	SRX625723	Illumina Hi-seq 1000	101	polyA selectio	paired	SAMN02870604	Leishmania major, Friedlin strain, metacyclic promastigote, Ficoll isolation	25/06/2015
SRR1460768	SRX625722	Illumina Hi-seq 1000	101	polyA selectio	paired	SAMN02870603	Leishmania major, Friedlin strain, procyclic promastigote	25/06/2015
SRR1460767	SRX625721	Illumina Hi-seq 1000	101	polyA selectio	paired	SAMN02870602	Leishmania major, Friedlin strain, metacyclic promastigote, Ficoll isolation	25/06/2015
SRR1460766	SRX625720	Illumina Hi-seq 1000	101	polyA selectio	paired	SAMN02870601	Leishmania major, Friedlin strain, procyclic promastigote	25/06/2015
SRR1460765	SRX625719	Illumina Hi-seq 1000	101	polyA selectio	paired	SAMN02870600	Leishmania major, Friedlin strain, metacyclic promastigote, Ficoll isolation	25/06/2015
SRR1460764	SRX625718	Illumina Hi-seq 1000	101	polyA selectio	paired	SAMN02870599	Leishmania major, Friedlin strain, metacyclic promastigote, PNA isolation	25/06/2015
SRR1460763	SRX625717	Illumina Hi-seq 1000	101	polyA selectio	paired	SAMN02870598	Leishmania major, Friedlin strain, procyclic promastigote	25/06/2015

Table 2.1 Details of *L. major* Friedlin RNA-seq datasets accessed from the sequence read archive.

2.2.3 *T. brucei* VSGs Monte Carlo Simulation

An in-house Java script was written to perform a monte-carlo simulation against MFAseq profiles in BESs compared to random genomic regions (https://github.com/CampbellSam/MFAseq-validation). The script requires aligned reads from early S and G2 phase cells in 2 samples and a predefined comparison region. For this region, the program will calculate a list of ratios representing the number of reads aligned in each eS and G2 for the procyclic form and again for the blood stream form. These lists are then compared using a mann-whitney u test which determines the likelihood that two distributions are the same. The resulting p-value (stored as x) represents the likelihood that the lists of numbers are the same. The program then chooses a random segment of the same length as the comparison region (in testing - 60kb) for a set number of iterations (user input, default=1000). The p-value (stored as y) from the random segment is then compared to that from the comparison region. If this value is less than the test value (y <= x), a counter is incremented and the chromosomal location and the value are added to an output file. When the set number of iterations are complete, the counter is divided by the number of iterations to obtain a final result which is an indication of how often you would observe the difference in ratios at the region of interest by chance.

2.2.4 Motif searching

2.2.4.1 Motif searching in *T. brucei* TREU927

Bias caused by repeated DNA sequences within the *T. brucei* genome was removed using the tool RepeatMasker (http://www.repeatmasker.org). The output from this software is a file containing annotated coordinates of repeat regions and an alignment file with these regions removed. This version of the genome was then searched for sequence motifs using Trawler (Ettwiller et al., 2007). An in-house Java script was written to parse the T. brucei genome for the G4 motif ($G_{\geq 3}N_xG_{\geq 3}N_xG_{\geq 3}N_xG_{\geq 3}$) using BioJava and regular expressions text analysis (Yates et al., 2012). The returned output from this script is a bed file containing the coordinates of the region matching this motif and the sequence that was a ‘hit’. The Java code can be found at https://github.com/CampbellSam/motif-search.

2.2.4.2 Motif searching in *L. major* Friedlin

RepeatMasker was applied to the *L. major* Friedlin genome to reduce bias caused by repeated DNA sequences (http://www.repeatmasker.org). Sequence at origin-containing SSRs was then extracted from the masked genome file and used as input to Trawler. A Java script was then generated to detect the frequency of each base in origin and non-origin sequence using regular expressions. This was then extended to double base counts (e.g CC or AA) in a second script. The Java program used to perform this analysis can be found at https://github.com/CampbellSam/base-counts.

2.3 Refining MFAseq output using peak calling software

2.3.1 Preparation of input data

The data used for input to each of the algorithms consisted of reads from early S *L. major* cells as the treated sample containing enrichment and G2 phase cells as the control sample with even coverage. As this data is contained in BAM files, Pyicos (https://bitbucket.org/regulatorygenomicsupf/pyicoteo) was used to convert the data to ELAND format and separate each chromosome into separate files which is required for input to FindPeaks.

2.3.2 Command line ChIP-seq peak-calling software

FindPeaks, MACS and a spatial clustering approach for the identification of ChIP-enriched regions (SICER) were used to detect peaks of DNA sequence enrichment in our data. SICER has been developed to improve the detection of signals from enrichment over a broad region of the genome (Zang et al., 2009). FindPeaks has been designed to detect enrichment in short-read sequencing data (Fejes et al., 2008). A working protocol for MACS is also available (Feng et al., 2012). Each piece of ChIP-seq analysis software was run using the command line in a Unix environment. An overview of the specific steps involved when using ChIP-seq peak-calling workflows is outline in figure 2.1.

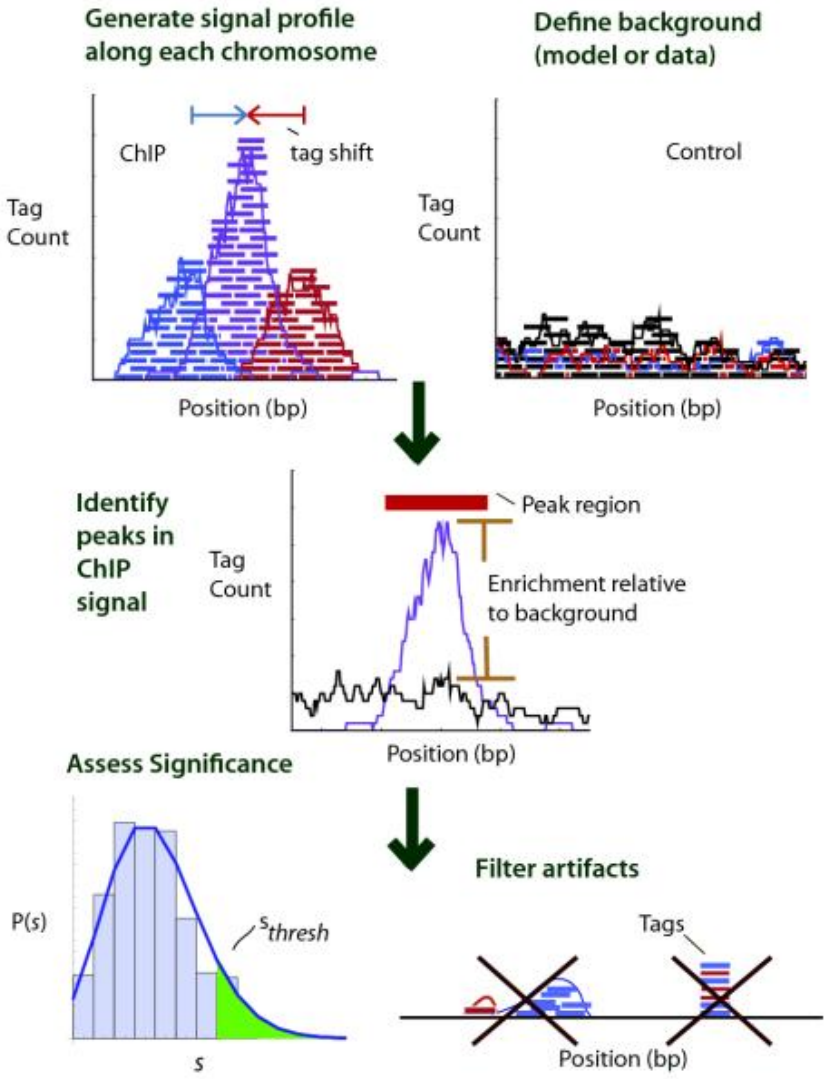


Figure 2.5. ChIP-seq peak calling workflow.

Diagram of the subtasks involved in the workflow of a typical ChIP-seq peak calling algorithm. Signal profiles are defined for both the ChIP sample and the background control. Peaks are then identified based on relative enrichment between the test sample and the background. A significance threshold is defined and used to refine the called peaks before filtering to give the final list of peaks signifying enriched regions. Figure adapted from (Pepke, Wold, & Mortazavi, 2009), reproduced with permission (license number: 4499441401470).

2.3.3 Visualisation using Circos

The output from each algorithm was converted to bed format and visualised using Circos (Krzywinski et al., 2009). Circos is visualisation software that accesses appropriately formatted data from locations specified in a configuration file and plots them as circular tracks. Thorough documentation is available at http://circos.ca/.

2.4 Re-annotation of strand switch regions

2.4.1 Re-annotating SSR coordinates in *T. brucei* Lister 427

SSRs containing protein-coding genes were identified using the custom search feature available at TriTrypDB and coordinates were updated using the Genome Browser tool and the gene annotations available for *T. brucei* TREU 927 and *T. brucei* Lister 427.

2.4.2 Generation of *de novo* gene annotations in *L. major*

L. major Friedlin RNA-seq reads were accessed from the sra and aligned into transcripts *de novo* using Trinity (https://github.com/trinityrnaseq/trinityrnaseq/wiki) (Grabherr et al., 2013). This was performed on an AWS EC2 instance as this process requires a large amount of disk space. The resulting transcripts were then used to update existing gene models using the PASA pipeline (http://pasapipeline.github.io/) (Haas et al., 2003). Updated annotations were viewed against the original annotations using IGV (https://www.broadinstitute.org/igv/) and regions of interest were identified using Bedtools intersect (http://bedtools.readthedocs.io/en/latest/) (Quinlan & Hall, 2010; J. T. Robinson et al., 2011).

2.4.3 Determining an RNA-seq coverage threshold

RNA-seq coverage data based on overlapping read depth counts was generated using HTSeq implemented on the Linux command line (Anders, Pyl, & Huber, 2015). A Python script to parse the read depth across the whole genome and SSRs was generated.

2.4.4 Analysis of spurious alignments across SSRs

Analysis of RNA-seq read coverage at SSRs was performed using a Jupyter IPython notebook (https://ipython.readthedocs.io/en/stable/). The total read count across SSRs containing an origin and those without was calculated and visualised as a boxplot using ggplot (http://ggplot.yhathq.com/).

2.5 Machine learning

Before applying the classifier the data must first be (1) kmerised, (2) vectorized and (3) transformed. An outline of this process is shown in figure 2.2. (1) The reads are split into kmers, which are then considered as the features i.e. the kmers belonging to each read are the features of each sample. This provides a large sample size compared to using the SSRs alone. (2) The kmers are then vectorized. (3) The final step is to transform the data set so that the least common features have the highest weighting and highly common features have a low weighting.

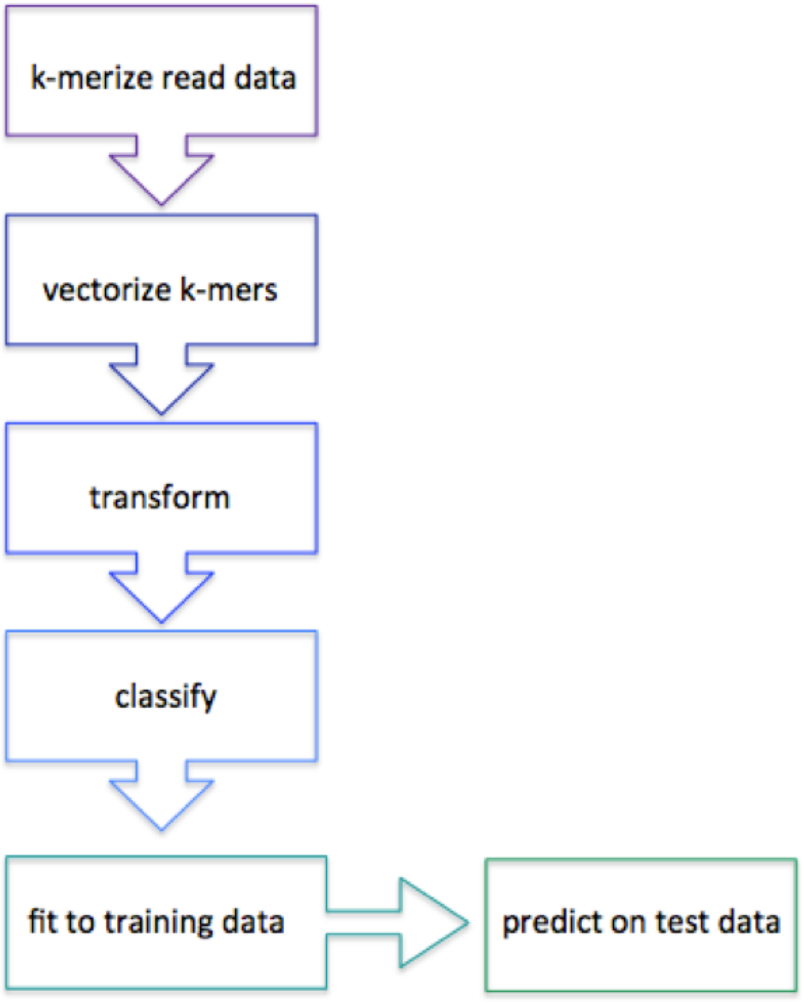


Figure 6.2 Outline of machine learning workflow.

DNA sequence reads are broken down into *k*-mers of a specified length. The *k*-mer features are then vectorised, transformed and used to perform the classification. This is done in two stages, first the *k*-mer data is used to train the SVM classifier before testing and predicting in an uncharacterised dataset.

2.5.1 Preparing *k*-mers from DNA sequence read data

2.5.1.1 Manual region extraction and exporting prepared *k*-mers

The *k*-mer generation step was initially written in a stand-alone Python script. Prior to running the script, DNA sequence reads aligning to the SSRs of the *L. major* Friedlin genome were extracted manually to a small alignment file using SAMtools and provided as input to the Python script. Tiled 10-mers were then generated for all reads and assigned class labels of 'origin' or 'non-origin' depending on the SSR they originate from. These *k*-mers were temporarily stored as individual lists for each SSR before being exported and saved in file to be used as input for the Python script that would perform the classification based on statistical learning and SVM implementation.

2.5.1.2 Optimsation of *k*-mer generation

This process was optimised by Dr Nicholas Dickens and the updated version is implemented in the current pipeline.The updated *k*-mer generation step does not require prior manual read extraction, only a list of regions of interest in BED format. The generated *k*-mers are no longer hard-saved to output files but stored within the Python script and formatted appropriately for input to the SVM.

2.5.2 Support vector machine implementation using Scikit-learn

The SVM was implemented in Python using Scikit-learn (Chang & Lin, 2013; Pedregosa et al., 2011). The Python program was co-written by Dr Nicholas Dickens. The resulting script is available at the following repository (https://github.com/CampbellSam/Origin-classifier). The following input parameters are required to train a SVM: indexed BAM file of aligned reads for training sample and 2 bed files containing the coordinates of origin and non-origin regions in the sample species. The following optional parameters can also be specified: output file name, base directory, cpu, random seed value and the option print the top 10 features used by the algorithm to perform the classification. The trained machine can also be saved and used as input to the

script. Running the same script with and specifying --machineIn will use the existing classifier to predict on supplied test data. The following input is required to test the algorithm on a new sample: base name of trained machine, indexed BAM file of aligned reads for training sample and a bed file of the test region coordinates.

2.5.3 Implementation of Gkm-SVM

The gapped *k*-mer SVM classifier, gkm-SVM is available to download at http://www.beerlab.org/gkmsvm/. The software requires sequence data in fasta format for positive and negative data sets. The sequence for SSRs containing origins was provided as positive data and those with no detectable origin as negative data. The software follows a general SVM outline and runs in three steps: (1) generate the kernel based on input data (2) train the kernel on training data (3) test the classifier by predicting on new test data. In the output each tested region will have a positive or negative value indicating whether it is predicted as origin or non-origin, respectively.

2.6 *L. mexicana* Serial Passage Analysis

2.6.1 Similarity of aligned DNA sequence samples

Deeptools (https://github.com/fidelram/deepTools) was used to assess similarity between aligned samples. Specifically, the multiBamSummary command which can be used to produce heatmaps and correlation plots as well as a PCA. Each of these commands can use either a Pearson or Spearman correlation for the analysis. The data shown was produced using a Pearson correlation. CummeRbund (http://compbio.mit.edu/cummeRbund/manual_2_0.html), a set of tools intended for differential expression analysis of RNA-Seq datasets was also applied to investigate the average gene coverage across all samples. This requires Cufflinks FPKM (fragments per kilobase of transcript per million mapped reads) estimates from all samples as input.

2.6.2 Analysing the dataset using the PReP pipeline for genomic analysis

Triplicate DNA sequence datasets from *L. mexicana* M379 in serial passage were provided from regular intervals sp0, 10, 16, 20 and 29, with the exception of serial passage 0 (sp0, passaged once) which is a single sample. Sample sp0 was extracted from a mouse infection and then passaged once to increase sample size. At sp16, the replicate samples were used to inoculate three mice and allowed to form a lesion over 10 weeks before a sample was extracted from the lymph node of each mouse. These samples again underwent a single passage to increase available DNA content and sequences were provided from these three replicate samples, allowing a comparison of post-mouse samples to be performed. The following steps were performed using the in-house Perl pipeline PreP (https://bitbucket.org/ndickens/prep).The reads were pre-processed including trimming and quality scoring using TrimGalore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) before alignment using Bowtie2 and validation by samtools flagstats. Pre-processing and alignment of reads was performed as a batch job on an Amazon web services CfnCluster (https://aws.amazon.com/hpc/cfncluster/). Gene CNV data was generated across all samples using Cufflinks tools, including CuffQuant and CuffDiff (https://github.com/cole-trapnell-lab/cufflinks). While chromosome CNV was estimated using in-house scripts.

2.6.3 Generation of SNP data

Using the output from PReP, the Bayesian variant caller FreeBayes was used to detect SNP and variant data in each sample which was then visualised with Circos (http://circos.ca/).

2.6.3.1 Analysis of SNP rate at SSRs

Python notebooks using Jupyter (http://jupyter.org/) were used to write analysis and visualization of SNP rate across SSRs. In the initial analysis of SNP rate at origin vs non-origin SSRs, SNP rate was calculated per kb for each region using vcf file output from FreeBayes (https://github.com/ekg/freebayes). When comparing average SNP rate across all regions, SSRs were broken down into ten

bins and the average SNP rate was generated per kb for each bin to allow comparison of all regions accounting for differences in size.

2.6.4 Investigating correlation between chromosome size and chromosome fold change

In house Python scripts were written to investigate the relationship between chromosome length and fold change throughout passage. Python tools were also used in this script to perform a linear regression analysis for chromosome foldchange between sp0 and sp10 and sp0 and sp16.

2.6.5 Protein domain search

Protein domain annotations, based on the Pfam database and functional protein classification by InterPro, were accessed from TriTrypDB (version 26) for the *L. mexicana* genome (El-gebali et al., 2019; Mitchell et al., 2019). A Python script was written to parse the annotation file and focus on Pfam annotations only. Counts for each Pfam ID were generated for the whole genome and individual chromosomes and then calculated for the chosen samples. Enrichment analysis was then performed by hypergeometric distribution.

Gene ontology (GO) term enrichment analysis was performed using TopGO, implemented in a pre-existing R markdown script written by Dr Kathryn Crouch at the Wellcome Centre for Molecular Parasitology (A & J, 2018; Ashburner et al., 2000).

Coordinates of genes containing the enriched amastin domains and GP63-associated genes were obtained from TriTrypDB as tab-delimited text files and their locations in the genome visualised using Circos.

2.7 Investigation of sequence features within multi-copy genes in *L. major*

2.7.1 Calculating a haploid threshold to determine gene copy number

The PReP pipeline was applied to aligned DNA sequence reads from each genome: *L. mexicana* U1103, *L. mexicana* M379 and *L. major* Friedlin. This provided haploid counts for each set of genes grouped by ortholog group which could be plotted to generate a single copy haploid count threshold. This visualisation was performed in R using ggplot2 (H, 2016). These plots were similar across the three samples and so a consensus of thresholds were chosen: 0.75, 1.8 and 10 as the minimum, single copy and maximum thresholds respectively. A list of gene IDs were generated for each genome using Python scripts and the coordinates for these were downloaded from TriTrypDB. Many of the Python scripts used in this analysis exist within Jupyter notebooks. After a small amount of formatting on these files they could be used as bed format input for the machine learning Python script (https://github.com/CampbellSam/Origin-classifier).

2.7.2 Application of machine learning

The SVM classifier scripts are written in Python and makes use of machine learning packages from Scikit-learn as described previously (http://scikit-learn.org/stable/). A machine with a linear kernel was trained on DNA sequence reads aligning to single/multi copy genes in *L. Mexicana* U1103 or *L. major* Friedlin and each has been tested on the other dataset. In the case of the *L. mexicana* U1103 trained classifier, this was also tested on *L. mexicana* M379. Input data must be pre-processed before input to the SVM by the following steps: (1) k-merisation. Reads are broken down into tiled *k*-mers of 7 nucleotides (previously 10) which are then considered as features of the sample data set. This step increases the sample size although many specific *k*-mers will only be present once, generating a sparse feature matrix. (2) Vectorisation: each *k*-mer feature is turned into a numerical vector which allows the machine learning algorithm to implicitly plot features in high-dimensional space in order to perform the classification. (3) Transformation: The entire feature set is transformed to give rare *k*-mers a higher weighting while features that are

common will have a low weighting. The classifier can then be trained and tested on this data. Accuracy is currently evaluated by testing back on the training data and also a new data set to determine the number of correct predictions. Selection of significant features used by the classifier is performed within the Python script and included as final output.

2.8 Data storage and resources

2.8.1 Use of AWS S3 bucket storage

Large datasets including processed DNA sequence reads, intermediate data and additional files were deposited in online cloud storage containers. Amazon Web Services Simple Storage Service (AWS S3) provides secure cloud storage that is easy to upload to, store and retrieve data from using the management console user interface on a web browser or through command line tools (https://docs.aws.amazon.com/s3).

2.8.2 Remote virtual machines in the form of AWS EC2 Instances

Some analyses required a larger volume of disk space and computational power and therefore could not be performed locally. In this case, analysis was completed on an Amazon Web Services EC2 instance, which provides efficient cloud computing resources (https://aws.amazon.com/ec2/).

2.8.3 Genome sequence retrieval

Genome sequences in FASTA format were accessed from TriTrypDB, the online database resource (http://tritrypdb.org/tritrypdb), for *T. brucei* TREU 927, *T. brucei* Lister 427, *L. major* Friedlin and *L. mexicana* U1103 (Aslett et al., 2010). Genome version 36 was used consistently across all analysis. Additional datasets including gene and protein functional annotations and histone data tracks in varying formats, such as BED and tab-delimited Excel, were also accessed using this resource.

3 Characterising origins of replication in *T. brucei* and *Leishmania* spp.

3.1 Introduction

Faithful replication of DNA molecules is a complex and tightly regulated process that generally occurs once during each cell cycle and is essential for the inheritance of genetic material. The initiation of DNA replication is spatially and temporally controlled through a variety of factors to minimise clashes with other cellular machineries and minimise the occurrence of errors during duplication. A variety of mechanisms involved in the regulation of DNA replication initiation have been identified in bacteria, archaea, fission yeast and higher eukaryotes (Nicholas P Robinson & Bell, 2005).

Regions of DNA replication initiation were first mapped across the 11 megabase chromosomes in the *Trypanosoma brucei* genome using chromatin immunoprecipitation coupled with microarray hybridisation (ChIP-chip) and high-throughput sequencing, followed by MFAseq analysis which later included the BESs mapped to contigs (Devlin et al., 2016; Tiengwe et al., 2012). A similar analysis to the MFAseq approach was applied to yeast around the same time (Muller et al., 2014). Marker Frequency Analysis by deep sequencing (MFA-seq), is a technique similar to repli-seq used to temporally map replicons in human cells (Hansen et al., 2010). Given the success of this MFAseq mapping in *T. brucei*, the same approach was applied to *Leishmania* at the outset of this PhD, revealing a potentially striking difference from *T. brucei* (Marques et al., 2015). Given the novelty of only identifying a single MFAseq peak in each *Leishmania* chromosome, suggesting only a single origin, it was considered important to test, as far as possible, the limits and effectiveness of the MFAseq approach. This chapter describes a number of tests and modifications of the MFAseq analysis in *Leishmania*.

The MFAseq pipeline implemented in this analysis compares read depth across each chromosome in early S and G2/M phase cells. Cells are sorted by FACS on DNA content, with G2/M cells having twice the DNA content. The output provides a list of ratios for 2.5 kb regions of each chromosome in the sample genome that can be easily visualised. It is expected that there will be an enrichment in read coverage in regions of the genome where replication has initiated in early S phase cells relative to the non-replicating G2 cells and therefore the ratios at these sites will be higher. Sites of replication initiation

appear as peaks relative to the background and represent potential origins of replication. The coordinates for origins of replication in kinetoplastid genomes mapped using this approach are currently based on the annotations of the surrounding genes.

The results from these different studies indicate that the *Leishmania* genome is complex and confounding factors such as the presence of mosaic aneuploidy may hinder current interpretation. Attempts to locate DNA origins of replication have produced disparate results and there is currently no consensus on an accurate model. In this chapter the sensitivity of the MFAseq technique in *Leishmania* is assessed and potential optimisation is discussed. Results from several peak-calling algorithms applied to optimise this approach are compared and the problem of computationally identifying peaks is addressed. Further analysis is also employed to confirm unusual observations based on MFAseq data of the VSG-containing BESs in *T. brucei*

3.2 Optimisation of the MFAseq pipeline

A variation of the MFAseq method previously used to effectively predict the location of replication origins in *T. brucei*, was now applied to *L. major and L. mexicana* (Marques et al., 2015; Tiengwe et al., 2012). In the case of *T. brucei*, the predicted origin regions were verified through integration with ChIP-chip data of TbORC1/CDC6, a protein involved in the pre-replication complex (Tiengwe et al, 2012), but no such replication factor mapping is available in *Leishmania*. The MFAseq method requires DNA sequence from cells in S and G2 phases and compares the ratio of aligned reads between the replicating and non-replicating cells. Peaks in the ratio indicate regions of the chromosome which are enriched in S phase relative to G2 phase and are therefore presumed to represent a site of replication initiation. Results from this analysis in *Leishmania* led to the identification of a single peak per chromosome (Figure 3.1), suggesting a single early-replicating site. To address the reliability of this observation and generate comparable results in *T. brucei*, the MFAseq pipeline was optimised to be faster and better suited to this analysis and the sensitivity of this approach in *Leishmania* was assessed.

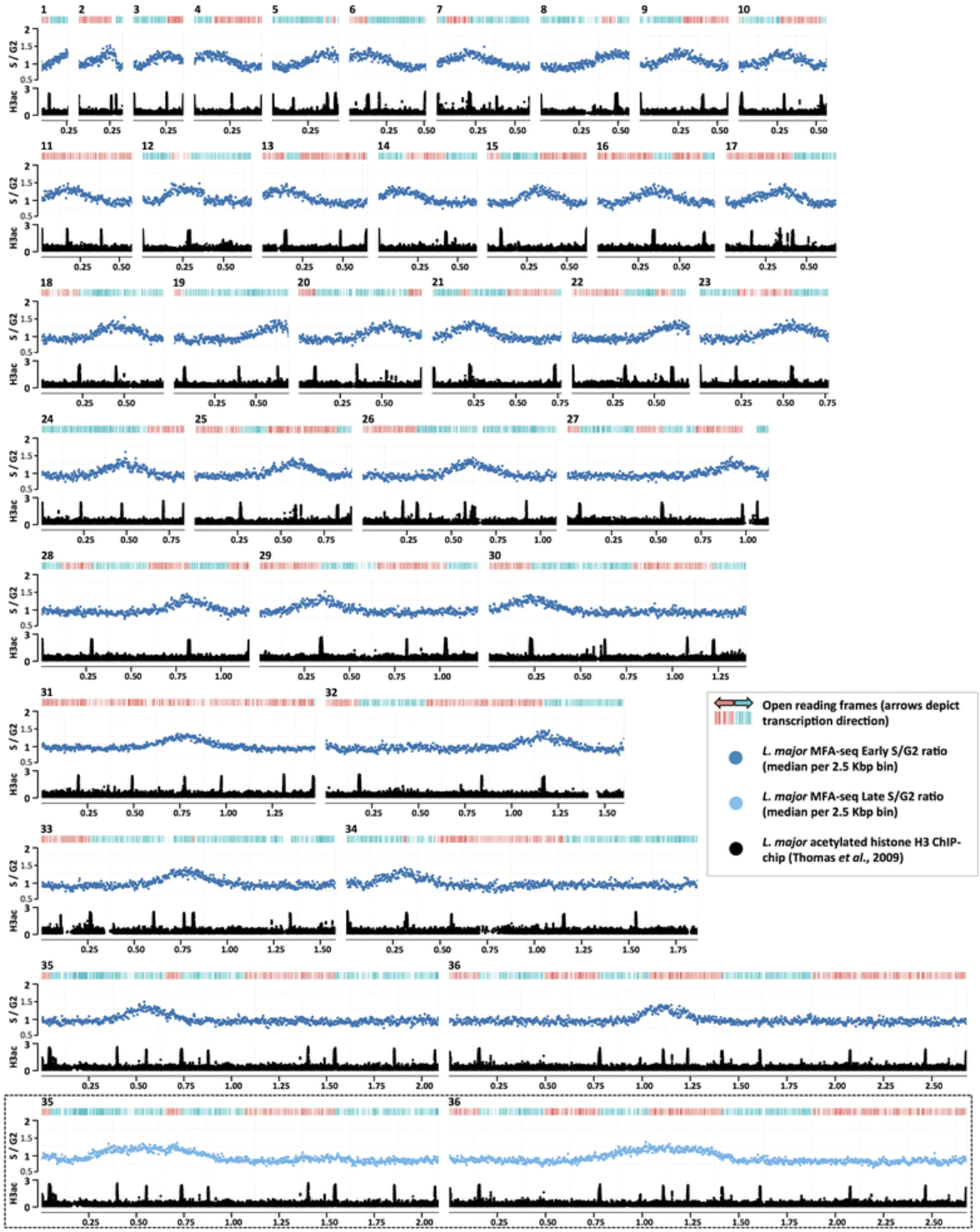


Figure 7.1. MFAseq mapping of replication origins in *L. major.*

Replication origins mapped by MFAseq on the 36 chromosomes of the *L. major* genome. The top track shows gene coding sequences with transcriptional direction indicated by colour, right to left in red and left to right in blue. The read depth ratio between S phase and G2 phase cells is plotted below in blue. Histone H3 acetylation is included at the bottom to indicate transcriptional start sites. The final box contains the plotted ratios between late S phase and G2 cells for chromosomes 35 and 36. This figure has been reproduced with permission from Marques et al, 2015.

3.2.1 Re-writing the MFAseq pipeline in Java

As the MFAseq approach has been shown to be effective in predicting sites of replication initiation, the first step of the project was to rebuild the analysis pipeline in order to generate a robust and efficient program with simple data input and output. The existing pipeline consisted of several Perl and Bash scripts that required pre-requisite formatting and intermediate steps that were slow and inefficient. This has been re-written into a quicker, more user-friendly Java program that can be run from the command line in a single step, requiring less prior bioinformatics training for use (https://github.com/CampbellSam/MFAseq). The updated MFAseq program requires just two indexed BAM files containing aligned reads as input and will output MFAseq ratios in wiggle format to the current directory or a specified Amazon Web Services bucket. The program counts the total number of reads in each sample and then counts the number of reads in segments along each chromosome (the default window size is 2.5kb but this can be changed using an additional input parameter). The read count of each segment in each sample is then used to calculate the ratio of aligned reads to give the MFAseq output. The new program was first tested on the *T. brucei* 927 genome to ask if the previously predicted origins were still detected. Following this trial, the new program was then applied to the DNA sequence data from *L. major* and *L. mexicana,* revealing that only one peak is detected for each chromosome, in agreement with previous analysis. This new MFAseq approach and resulting data is used for the following analyses in *Leishmania*.

MFAseq analysis indicates that megabase chromosomes in *T. brucei* contain multiple sites of replication initiation with varying firing times but these regions are present in a lower density than expected compared to what is currently known in eukaryotes (Tiengwe et al., 2012). Performing the same analysis in both *L. major* and *L. mexicana* identifies only a single origin per chromosome (Marques et al, 2015). The MFAseq peaks in *Leishmania* are also very broad and of consistent amplitude relative to *T. brucei*. When the same analysis is performed with late S phase cells against G2 cells in *Leishmania*, the observed

peaks broaden as replication progresses and there are no late-firing origins detectable with this method (Marques et al., 2015).

Further study of origins has been performed using alternative techniques to map these regions in *Leishmania*. DNA molecular combing allows the analysis of single DNA molecules and was used to analyse replication parameters in *T. brucei, L. major, L. mexicana* and *L. donovani* (Stanojcic et al., 2016). This analysis found that most DNA fibres contained more than one active origin, suggesting multiple origins per chromosome in *Leishmania,* -although still at a lower density than other well-studied eukaryotes. However, the lack of sufficient mapping data in the single molecule DNA combing analysis does not allow the resolution of sites of replication initiation with specific chromosomes.

Attempting to map origins in *Leishmania major* based on next-generation sequencing of purified small leading nascent strands (SNS-seq) also suggests multiple sites of replication initiation per chromosome, although at a much higher density (Lombraña et al., 2016). A region of high density initiation sites on each chromosome approximately correlates with the broad peaks observed by MFAseq. Mapping nucleosome occupancy using MNase-seq indicates that DNA replication is more likely to initiate at sites of RNA polymerase pausing and termination. The data from this high-resolution analysis indicates that the temporal and spatial initiation of DNA replication is potentially flexible across the chromosome and greatly influenced by transcriptional activity, highlighting the extent of the connection between these processes in kinetoplastids.

3.2.2 Simulation of multiple origins

Given the unexpected detection of only a single MFAseq peak per chromosome in *Leishmania*, the limits of this mapping approach were now tested. Indeed, subsequent to our MFAseq analysis, mapping of nuclear DNA replication initiation sites in *Leishmania* by methods other than MFAseq suggests that each chromosome contains multiple sites in which replication can initiate. Origin prediction by MFAseq is based on DNA sequence from a population of organisms and it was therefore important to assess the sensitivity of the approach in order

to determine a usage threshold at which active origins can be observed. It is then possible to verify the single origin observation in *L. major* by determining the minimum usage frequency within a population where an origin can be detected using MFAseq analysis. The scripts written to perform this analysis can be found at (https://github.com/CampbellSam/MFAseq-sensitivity). These simulations were performed using the raw alignment data for *L. major Friedlin* chromosome 36 (LmjF.36), since this is the largest chromosome and contains at least 7 annotated SSRs. The simulated data was initially visualized using the Genome Browser resource on the online database TriTrypDB.org. We simulated the presence of an additional origin used by 50% of the population at each of the remaining 6 SSRs on LmjF.36 that are not predicted by MFAseq to contain an origin (figure 3.2A). Although each of the simulated origins was easily detected by eye in this instance, when the data was merged it strongly resembled the real output and it is therefore possible that if several other origins were used by half of the population, we would be unable to detect them. As the entire population may not use all origins equally, I then performed 5 simulations beginning at 80% of the population using a second origin and decreasing to 66%, 50%, 33% and 17% usage, focusing on a single LmjF.36 SSR (figure 3.2B). The simulated origin was detectable by eye at 33% but no longer at 17%. Therefore, at <17% this region would not have been considered an origin. Plotting the median value for each of the simulated peaks against the usage frequency within the population generated a linear correlation and indicates that an origin would need to be used by at least 25% of the population to be detectable by this method (Figure 3.2C).

L. major chromosome 36 was chosen due to its size and the presence of multiple SSRs where replication could potentially initiate. Performing further simulation analyses with varying sizes of chromosomes and fewer potential SSRs could provide more detailed insight into the sensitivity of MFAseq in a plastic genome. Application to the whole genome may not be beneficial as it has been difficult to define replication origins on the smaller chromosomes and for others, the predicted peak encompasses the greater part of the chromosome (Figure 3.1).

This analysis indicates that we are able to detect an active origin in the *Leishmania* genome using the MFAseq approach if it is constitutively used by ~25% of the population or higher. This method cannot detect origins that are

fired variably at low frequencies across the population. Compared to the narrow, defined peaks detected by MFAseq in *T. brucei*, the peaks in *Leishmania* are comparatively broad and encompass large regions. This may indicate that the distribution which we are modelling potentially represents an early firing cluster of origins as opposed to a single origin. If this is the case, it would correlate with the dense regions of initiation sites observed on each chromosome by SNS-seq (Lombrana et al., 2016). However, it should be noted that the MFA-seq predicted origins in both parasites colocalise with SSRs, where ORC binds in *T. brucei*. Nonetheless, every relatively broad MFAseq peak mapped in *Leishmania* may also co-localise with recent mapping of potential centromeres in *L. major* by localisation and ChIP-seq analysis of LmKKT1 (Sollelis et al., 2017), unlike in *T. brucei*, where only some of the origins colocalise with centromeres. The centromeric regions are predicted to be 2-10kb in length and possibly represent a region where sites of DNA replication initiation are enriched. No consistently conserved motifs or repetitive sequences characteristic of centromeres in eukaryotic genomes have currently been identified in *Leishmania* (Sollelis et al., 2017).

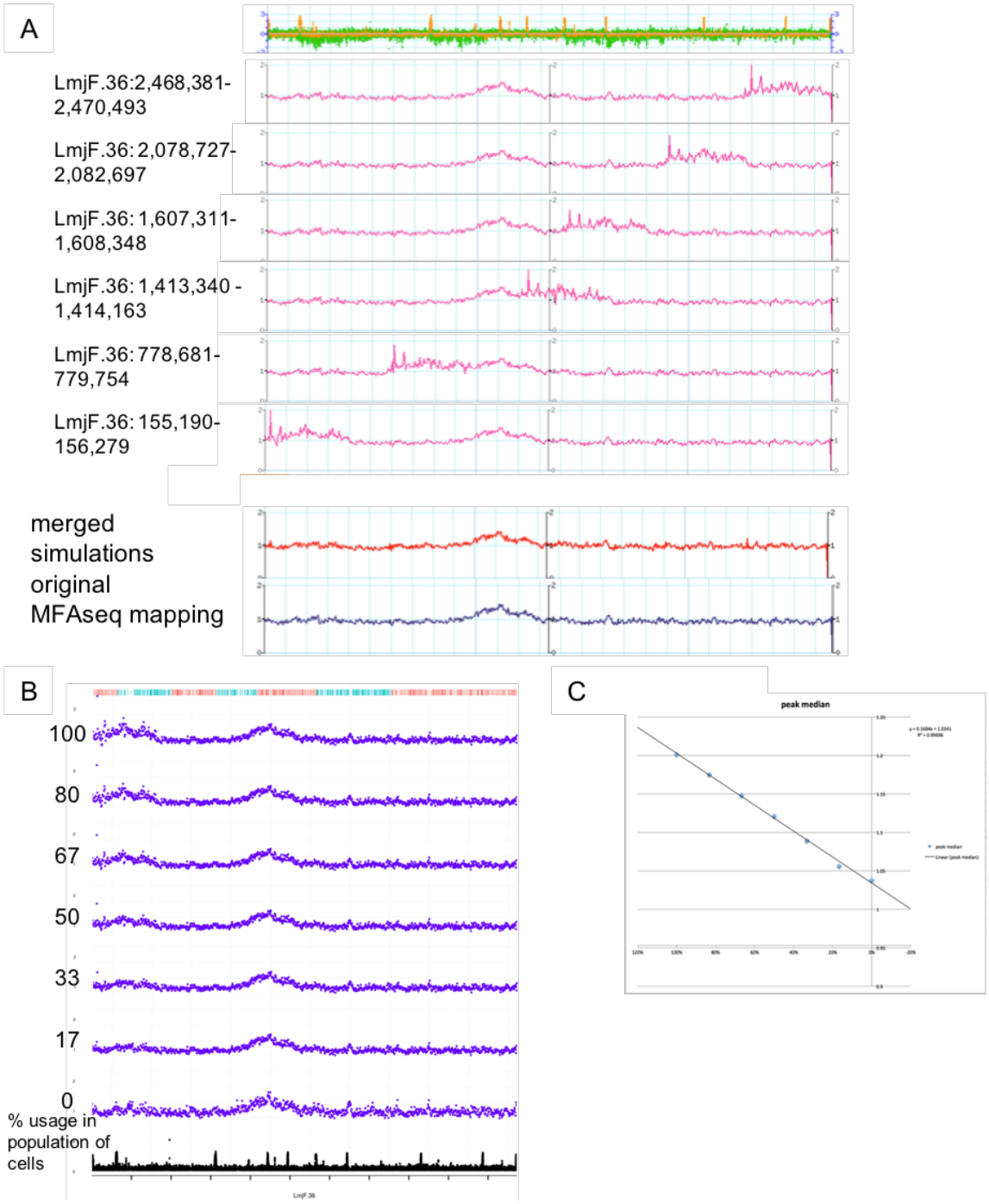


Figure 3.2. Assessing MFAseq sensitivity in the genome of *Leishmania major*.

A. H3Ac mapping on Lmj.36 in *L. major* promastigotes shown in green to highlight predicted SSRs. MFAseq mapping across simulations of an additional origin of the same usage frequency at six alternative SSRs shown in pink. The distribution of reads at the natural origin was determined and this read distribution was added to each SSR on Lmj.36. In red, these simulations have been merged and the dark blue track represents the original MFAseq mapping. This data was visualised using the Genome Browser toll at TriTrypDB.org. **B.** Visualisation of MFAseq data from Lmj.36 containing a simulated additional origin at varying usage frequencies in the population (80%, 67%, 50%, 33% and 17%). **C**. linear regression of median peak amplitude and frequency of usage within the population.

3.2.3 Attempt to refine MFAseq peak calling using ChIP-seq peak-calling software

The MFAseq peak data cannot give exact coordinates of replication initiation, since the data is mapped as broad peaks. Furthermore, the current MFAseq annotation of replication origins in *Leishmania* relies on rough co-localisation of origins with SSRs, but these regions have not been well characterised, with their boundaries merely based on the surrounding gene annotations. We wanted to define a more specific set of coordinates for each peak and, potentially, identify additional regions that may have been missed when the previous peaks were identified by eye. To achieve this, we asked if the MFAseq peak data could be refined using pre-existing software designed for ChIP-seq analysis. Based on literature reviews and initial testing using S phase and G2 DNA sequence read data from *L. major*, the peak calling software MACS, SICER and FindPeaks were chosen for the analysis.

ChIP-seq peak calling software is used to identify enrichment between samples, traditionally comparing a treated ChIP-seq sample to background data. Each program utilises a slightly different method to calculate enrichment and identify candidate peaks (see below). However, each is designed for analysis such as mapping transcription factor binding sites and identifying chromatin-binding factors and histone modifications, and therefore the target fragment size can vary widely in length and sequence composition. It is important to choose an appropriate algorithm and normalisation method for the data to be analysed as results and performance can vary depending on the dataset and chosen software. For example, some algorithms are designed to detect narrow peaks and would therefore not work well in an analysis trying to detect broad peaks. FindPeaks identifies candidate sites on the basis of peak height and overlapping fragments, while MACS (model-based analysis of ChIP-seq) models a local background distribution to reduce local bias in the genome (Fejes et al., 2008; Feng et al., 2012) . SICER (spatial clustering approach for the identification of ChIP-enriched regions) (Zang et al., 2009) was designed to identify clusters of histone modifications and therefore performs better when the target fragments form broad peaks, such as we observe in the MFAseq output from *Leishmania*. SICER partitions the genome into windows and clusters candidate windows into islands (Zang et al, 2009). Candidate islands are then assessed for significance

based on a calculated threshold. Normally a control sample will be included as well as a treated sample where specific proteins have been targeted and cross-linked to DNA leading to enrichment of certain sequences.

We supplied aligned reads from *Leishmania major* Friedlin G2 cells as the control, and early S reads as the sample. A step included in each approach is correcting for shift as the algorithms expect the target sequence to be present on both strands and therefore estimate the distribution of the fragment size to suit the bimodial enrichment pattern (Laajala et al., 2009). This step involves correcting the distance between the centre of the true binding site and the position of observed tags and is specific to ChIP-seq analysis. It was therefore considered not to be a necessary step in the case of our data.

The visualisation tool Circos (Krzywinski et al., 2009) was used to plot the output from each algorithm, as shown in figure 3.3. Of the software utilised, FindPeaks performed the worst, as MFAseq origin peaks were only called correctly in 8 of the 36 *L. major* chromosomes and no peaks were detected for several chromosomes. Almost all peaks defined using FindPeaks were called towards the end of chromosomes, with the exception of chromosome 36, indicating a potential increased sensitivity to enrichment in these regions. SICER called a higher number of peaks than FindPeaks but these only co-localised with our current MFAseq origin predictions on 5 chromosomes. This result was unexpected, given that SICER is based on detection of broad peaks.

MACS is a frequently used ChIP-seq algorithm that detects genome-wide binding sites by calculating peak enrichment using model based local background normalisation (Zhang et al, 2008). This algorithm performed the best in broad peak mode, as a single peak window was identified on almost all of the 36 chromosomes and the output was consistent with the previously generated early S and late S MFAseq plots. On 10 of the chromosomes a precise peak was mapped although the predicted regions were very large, possibly caused by the merging of several small neighbouring windows. Due to this, the predicted peak window was almost the same as the chromosome length for a minimum of 5 of the smaller chromosomes. We were therefore unable to refine the broad peaks mapped by MFAseq using this method as no specific spikes of enrichment were detected in any output.

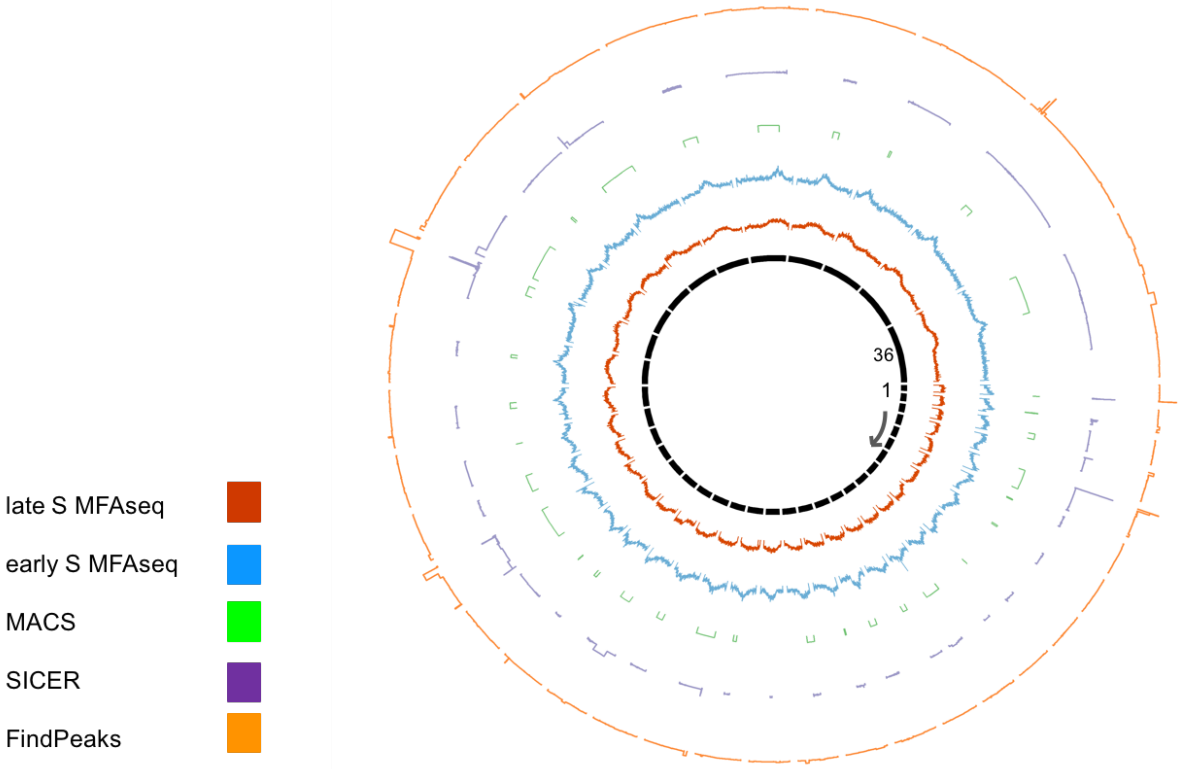


Figure 3.3. Peak calling profiles from ChIP-seq software

FindPeaks (orange), SICER (purple) and MACS (green) visualised using Circos. L. major chromsomes 1-36 are represented by black segments in the centre of the plot. MFAseq profiles for each chromosome are also included from early S (blue) and late S (dark red) samples

This investigation suggests that the pre-existing peak-callers designed for ChIP-seq analysis may not be appropriate for our data and the corresponding enrichment distribution pattern. Initial analysis with USeq, another peak caller designed for ChIP-seq data, which uses a control sample to reduce false positives, could not detect distinct peaks in our data (Nix et al., 2008). It may be of interest to investigate shape-based peak callers, which are available to use as command line software and/or packages in R such as PICS and PolyaPeak (Wu, 2012; Zhang et al., 2011). This approach might improve our current data and

refine the coordinates of the region defined as a peak down to as small as a few kilobases.

Alternatively, this method could be improved by the implementation of an in-house peak calling script which would define peaks computationally in a similar method to the ChIP-seq software. However, it is difficult to define a peak computationally from the MFAseq profile as a significance threshold must be pre-defined. We cannot currently determine the ratio level representative of a detectable active origin and therefore, addressing this issue is beyond the scope of the current project.

3.3 Identifying conserved features of DNA replication origins in kinetoplastid genomes

Although the genomes of *T. brucei* and *L. major* are highly syntenic, the organisation of currently mapped origins differs between *L. major* and *T. brucei*, with the observation of a single origin also seen in *L. mexicana*. In contrast to the largely stable diploid genome of *T. brucei*, the *Leishmania* genome is highly flexible and therefore able to rapidly adapt to external conditions. It is of interest to investigate the similarities and conserved features of DNA replication origins between these species and assess the possible divergence of the mechanisms underlying the DNA replication process in *Leishmania*.

A common feature of many of the mapped replication origins in *Leishmania* and *T. brucei* is the co-localisation of these regions with SSRs. In *Leishmania*, origin-containing SSRs are consistently larger than SSRs that do not contain an origin (Marques et al., 2015). *T. brucei* and *Leishmania spp*. also employ different methods of surviving immune attack. This likely has an effect on chromosome structure and content, as genes associated with immune evasion will be strongly selected for. A VSG coat is expressed by *T. brucei* parasites to evade the immune system and the genome contains several sub-telomeric VSG arrays, which are not present in *Leishmania*. The sequence of BESs, where the VSGs are expressed, had not yet been mapped to core genome chromosomes during this analysis, making it difficult to investigate replication dynamics at these regions.

The following focuses on validation of a potentially early-replicating BES, highlighted by MFAseq analysis.

3.3.1 VSG origin prediction

MFAseq analysis was performed in bloodstream form (BSF) and procyclic form (PCF) of *T. brucei* strain Lister 427 to investigate replication dynamics and antigenic variation (Devlin et al., 2016). The telomeric BESs have been sequenced and characterised in *T. brucei* Lister 427, allowing a comparison of these regions in different lifecycle stages but they have not been mapped to specific chromosomes and the 14 BESs are mapped across 16 contigs (Hertz-Fowler et al., 2008). A difference in MFAseq mapping between PCF and BSF is only observed for one of the regions, termed BES1 and encoding VSG221, which is the actively transcribed BES in the BSF cells (figure 3.4). It was predicted that this region is early-replicating in BSF only, while all other BESs are late-replicating. As these regions are small, between 10 kbp – 50 kbp, it is not possible to detect peaks at this resolution. We therefore performed a series of simulations using the Monte Carlo method to test if the discrete MFAseq ratios for the BES1 region, relative to all other BESs, are different between BSF and PCF.

The simulation analysis was written in Java and initially generates a Mann-Whitney *U* test p-value for the BSF and PCF MFAseq ratios for a specified region. It then randomly samples regions of the same length from the genome and performs the same test (https://github.com/CampbellSam/MFAseq-validation). Counting how many times the Mann-Whitney *U* test p-value from the random samples is lower than that for BSF and PCF gives an indication of how often we would expect to see this difference occur at random. Repeated simulations gave results between 17-28 out of 1000 random samples, giving us an output value of 0.017-0.028. The resulting p-value of 2.3×10^{-8} indicates that the BSF and PCF ratios are significantly different and from the output of the Monte Carlo simulation we can verify that it is unlikely this difference would occur at random. Although we can therefore predict that this region is early-replicating, this approach does not confirm the presence of an active origin.

As the SSR coordinates for *T. brucei* Lister427 have now been generated, it would be possible to extend this analysis by comparing the sequence features of the BESs with that of the origin and non-origin containing SSRs. For example, PCA may be useful to identify any conserved similarities and determine whether the BES1 region contains any resemblance to an origin-containing site.

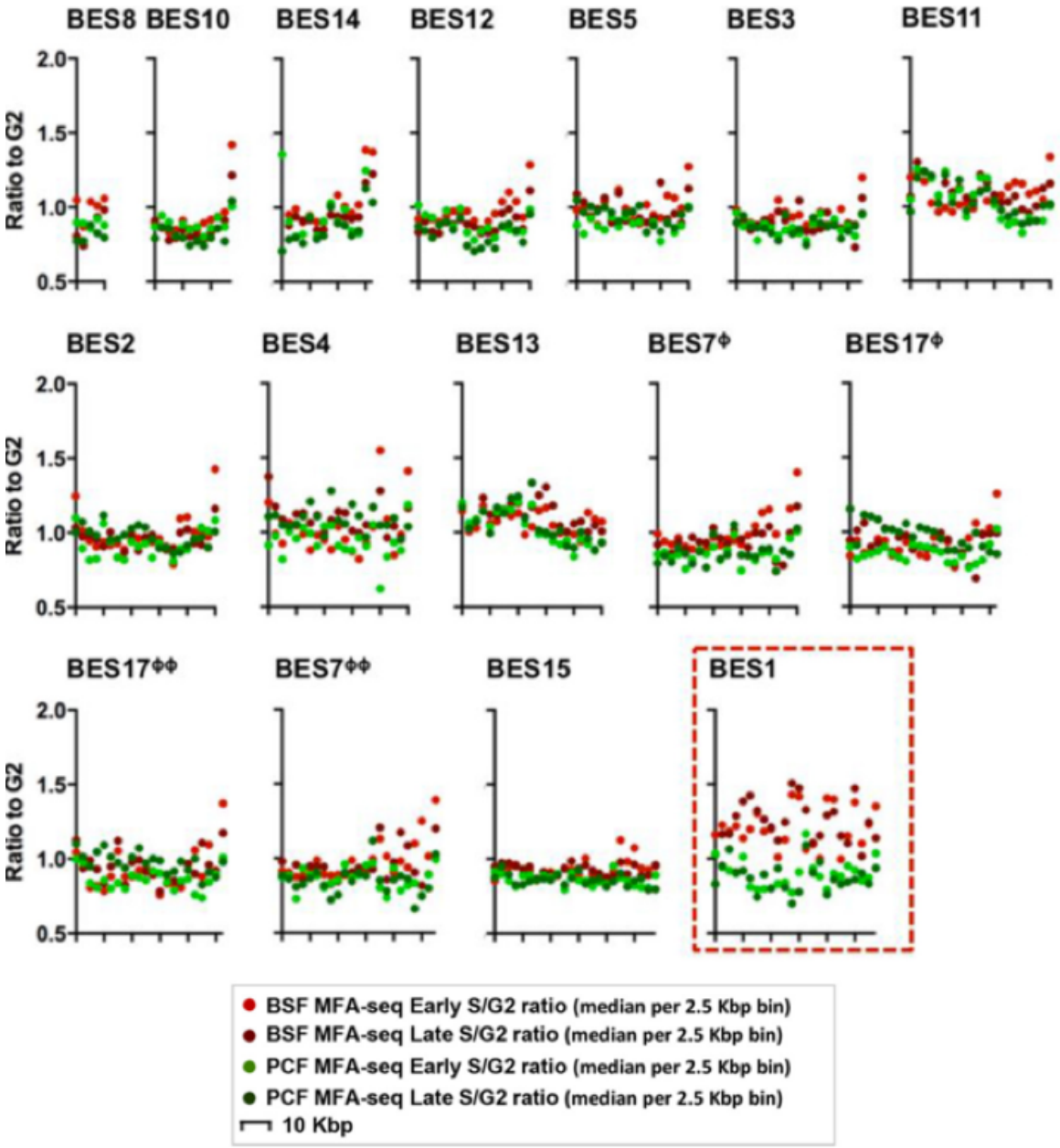


Figure 3.4. MFAseq mapping of BESs in BSF and PCF.

MFAseq profiles of the contigs containing BESs in T. brucei Lister427 in BSF (red) and PCF (green) during both early (light) and late (dark) S phase. Adapted with permission from Devlin et al, 2016.

3.4 Conclusions

The analysis in this chapter was aimed primarily at determining the effectiveness of MFAseq in a genome where mosaic aneuploidy is a feature, based on the different results observed between *L. major* and *T. brucei*. In contrast to the largely stable diploid genome of *T.* brucei, the *Leishmania* genome can tolerate variations in chromosome copy number and sustain aneuploid chromosomes from haploid to more than tetraploid. The presence of aneuploidy in *Leishmania* varies between species, as does the number of chromosomes which may have occurred due to fission or fusion events (Rogers et al, 2011). The presence of aneuploidy in this genome may obfuscate the flexible and dormant replication origins that are used at a much lower frequency in the population. This may account for the differences observed between the MFAseq profiles of *Leishmania* and *T. brucei*. The newly built MFAseq pipeline was used to predict origins in *L. major* and investigating the addition of a simulated origin allowed the estimation of a minimum usage frequency within the population for an origin to be detectable by eye from this data. From this analysis, we conclude that an origin used by >25% of the population would be detectable by the current MFAseq method. This approach is still limited as peaks are detected by eye and an algorithm to computationally predict origins from the MFAseq ratio data has not yet been developed. Use of ChIP-seq peak callers to identify peaks from the early S and G2 phase read data did not perform as well as expected. However, the predictions made by MACS were close to our current mapped peaks and use of this software could be further optimised in regards to our data. It may be useful to also assess the performance of peak callers on early S and late S phase reads from the *T. brucei* genome, where ChIP-chip of TbORC1/CDC6 confirms the binding of the ORC complex at MFAseq predicted origin sites.

The peaks predicted by MFAseq in *Leishmania* are broad and span a large region. The apex of these broad peaks co-localise with SSRs and current annotations of a single dominant replication origin are based on the surrounding gene boundaries of the corresponding SSR. It is however possible, that these peaks denote a region enriched for replication initiation sites that is spatially conserved and used by all cells. The broad peak we observe could therefore represent a cluster of replication origins as opposed to a single origin. This suggestion may explain the detection of multiple origins on a single chromosome by DNA combing and

SNS-seq approaches. Recent predictions of centromeric regions in *L. major* by ChIP-seq assay of LmKKT1, identifies a major peak on each chromosome spanning roughly 4kb which co-localises with the peaks mapped by MFAseq (Sollelis et al., 2017). Additionally, motif searching analysis highlighted two conserved motifs at these regions in 14 of the 36 chromosomes and a high density of retroposons was also noted in the majority of predicted centromeric regions (Sollelis et al., 2017).

Although the MFAseq Monte Carlo simulation analysis does not confirm the presence of an active origin at an early replicating BES in BSF *T. brucei*, it does indicate that this region is likely to be proximal to a replication origin in the subtelomere of a chromosome. This is a reasonable conclusion as this BES contains the VSG which is actively expressed in this cell line. Since this analysis was performed, a new assembly of the *T. brucei* 427 Lister genome has been generated using PacBio single-molecule real-time (SMRT) sequencing technology followed by genome-wide chromosome conformation capture (Hi-C) which provides high resolution information about spatial organisation of chromosomes (Muller et al., 2018). With the use of this data, the BESs were mapped to subtelomeric regions of the megabase chromosomes. This potentially allows further investigation of the early replicating BES and the proximity of this site to a predicted origin of replication could be determined.

Recent research highlights the differences in the organisation and characteristics of DNA origins of replication between *L. major* and *T. brucei*. Identification and study of DNA machinery and centromeric proteins in *Leishmania* would provide further insight into the similarities and differences between these species. Due to their different survival strategies, it is reasonable to hypothesise that features specific to *Leishmania* may have emerged as they were beneficial to the maintenance of a plastic genome and parasite survival and characterisation of DNA replication in *Leishmania* may elucidate the mechanisms underlying genome plasticity.

4 Investigation of similarities and differences of replication origins across kinetoplastid genomes

4.1 Introduction

Kinetoplastid genomes are characterized by their unusual structure in which genes are organized into polycistronic arrays, each traversed by RNA pol that is loaded at a single transcription initiation site for all genes in an array. Putative multigene transcripts are then converted to mature mRNAs by trans-splicing and linked polyadenylation. Gene expression is therefore regulated post-transcriptionally and, in order to modulate the expression of individual genes, the copy number of a gene is sometimes increased or decreased. Neighbouring polycistronic gene arrays are frequently found on different strands and the regions in between are termed strand-switch regions (SSRs). Initiation of transcription occurs between the first genes bordering divergent SSRs, whereas termination is thought to occur at SSRs where the gene arrays converge (Nguyenet al,. 2004). SSRs also exist between gene arrays of the same orientation and are referred to as head-to-tail (H-T) SSRs. In all of the kinetoplastid genomes analysed thus far, all of the predicted origins of replication mapped by MFAseq localise to SSRs but not all SSRs contain an active origin of replication (Marques et al., 2015; Tiengwe et al., 2012). In the case of *T. brucei,* of the 42 replication origins identified, 19 of the MFAseq peaks co-localised with divergent SSRs, 3 with convergent SSRs and 20 with H-T SSRs (Tiengwe et al., 2012). In *L. major,* all 36 mapped origins are thought to span SSRs, although the origin on chromosome 1 appears at the end of the chromosome, at the end of a DGC adjacent to the telomere and therefore the replication signal here could be due to telomere-directed replication initiation (Marques et al., 2015).

Several essential processes including transcription initiation and termination, also occur at SSRs and putative features relevant to DNA replication initiation (is separate from these reactions)could be easily obscured as the relationship between replication and transcriptional processes is not well understood (Reynolds et al., 2016; Siegel et al., 2005; Thomas et al., 2009). RNA pol II binding motifs have not been identified in either *T. brucei* or *L. major* and sites of transcription initiation have been identified by mapping histone markers. In *L. major*, ChIP-chip was used to map TATA-binding protein, $SNAP_{50}$ and modified H3 histones as acetylation of histone H3 has been associated with transcriptional activity and is found at transcriptional start sites (TSSs) (Liang et al., 2004;

Thomas et al., 2009). The majority of H3ac sites in *L. major* map to divergent SSRs, with some occurring within gene clusters and near chromosome ends, indicating that divergent SSRs are a preferred site for transcription initiation (Thomas et al., 2009). Previous analysis of a divergent SSR in *L. major* revealed a high AT composition and DNA curvature that also suggested a potential role in transcription (Tosato et al., 2001) In the case of *T. brucei,* four histone markers have been identified that denote predicted TSSs. H4K10ac and the histone variants H2AZ and H2BV are enriched at TSSs, as is the bromodomain factor BDF3 (Siegel et al., 2005). Additionally, mapping histone variants also suggests that H3V and H4V are enriched at sites of transcriptional termination (Siegel et al., 2005). Sites of RNA pol II termination in *T. brucei* are predominantly characterised by this observed enrichment of H3V and also an increased amount of base J (David Reynolds et al., 2016). The hyper modified base J (β-D-glucosyl-hydroxymethyluracil) is found in all kinetoplastid flagellates and has also been associated with transcriptional termination in Leishmania (Van Luenen et al., 2012). Although >99% of base J has been mapped to telomeric regions, it has also been mapped to convergent SSRs where termination of transcription occurs (Genest et al,. 2007). Additionally, loss of base J demonstrates massive transcriptional readthrough (Van Luenen et al, 2012).

Current knowledge indicates that chromatin conformation may restrict the locations of transcription and replication initiation sites to a small number of regions in the genome, predominantly in between polycistronic gene clusters at SSRs. Collisions between replication and transcriptional machinery can be a source of genomic instability if not resolved efficiently. Head on collisions between these processes can cause an accumulation of RNA-DNA hybrids termed R-loops (Pollock et al., 2017). Recent mapping of R-loops in *T. brucei* highlights conserved localisation of R-loops at centromeres and sites of RNA pol II initiation, with most R-loops occurring at intergenic regions but also localised at sites of transcription initiation (Briggs et al., 2018). No enrichment of R-loops was detected at origin-containing SSRs compared to non-origin SSRs and any functional interaction between the replication and transcription machineries remains unclear (Briggs et al., 2018).

As all the MFAseq mapped replication origins co-localise with SSRs, the coordinates for the origins are based on those of the corresponding SSR and the SSR coordinates are generated from the neighbouring gene annotations, encompassing a large region between the polycistron arrays. Characterization of origins at the sequence level is therefore confounded by the presence of several other elements involved in a number of processes, making it difficult to identify the origin DNA sequence which may be associated with several functions. To better characterise replication origins, it is necessary to better define the SSRs, as these regions are complex and can be very large. Due to the kinetoplastid genome structure, many processes occur here and updating the current gene models could help to reduce noise and provide an interesting model to study eukaryotic origins of replication. In this chapter, the SSRs and DNA origins of replication are further characterised in *T. brucei* and *Leishmania* and compared between these species.

Traditional sequence comparison methods and motif identification tools have so far been unable to determine a consensus sequence at virtually all eukaryotic origins highlighting the need for a different approach in solving this problem. Consistent with previous eukaryotic studies, with the exception of *S. cerevisiae* and relatives, a consensus sequence has not been identified at kinetoplastid replication origins (S.Newlon & F.Theis, 1993). Machine learning algorithms have been successfully applied across many diverse disciplines to learn from large datasets and used in genomics to determine features specific to different categories of data, such as cell types, DNA binding domains and proteins, and to make functional predictions in novel data based on these features (Čuklina et al., 2016; Ding et al., 2016; Kumar, Gromiha, & Raghava, 2007). The quality of data from existing genomics approaches and technologies can also be improved through the use of machine learning techniques (DePristo et al., 2011). Support vector machines (SVMs) are based on statistical learning theory and use a learning algorithm specified by a kernel to predict the optimal features representative of different classes of complex data (Vapnik, 1999). A previous study of replication origins in *Drosophila melanogaster* effectively implemented a SVM to discriminate between ORC-associated and ORC-free sequences (MacAlpine et al., 2010). The application of machine learning in the context of

origin and non-origin DNA sequences in *Leishmania* is also described in this chapter.

4.2 Identifying sequence features of DNA replication origins in *T. brucei* and *L. major*

4.2.1 Identifying motifs related to replication origins in the *T. brucei* genome

Initial approaches to characterizing origins of replication at the DNA sequence level utilized online resources and software available for motif identification although pattern searching in DNA sequence remains a challenge due to the presence of confounding factors such as mutations. The aim of using these tools was to identify enriched sequences within origin sequences that can be used to predict novel origins in unannotated kinetoplastid species. The *T. brucei* DNA genome sequence was masked for repeats and assessed for motifs using the software RepeatMasker and Trawler (http://www.repeatmasker.org; Ettwiller et al. 2007). The output from Trawler showed enrichment of several homopolymers within the *T. brucei* genome (data not shown). Further investigation of this to look at the relative bias and tract length of each nucleotide revealed that this was not restricted to homopolymer tracts of G, which are involved in the formation of G-quadruplex (G4) structures, but observed for all nucleotides.

The presence of the G4 motif ($G_{\geq 3}N_xG_{\geq 3}N_xG_{\geq 3}N_xG_{\geq 3}$) has been identified in several eukaryotic genomes and is known to be associated with the formation of quadruplex structures that are thought to be involved in origin activation and replication initiation (Maizels, 2006; Valton et al., 2014). As a result, a script was written to perform a basic text search of the G4 motif against the DNA sequence of the *T. brucei* genome. The motif appeared frequently throughout the genome and although our analysis did not search for specific regions of enrichment, it appears difficult to determine a statistically significant correlation between the G4 motif and replication origins in *T. brucei*. The results from this brief analysis are therefore so far inconclusive, and are not shown.

4.2.2 Identifying motifs related to replication origins in the *Leishmania* genome

As in the previous analysis of sequence features in the *T. brucei* genome, RepeatMasker was used to mask repetitive sequences in *L. major*. Applying the Trawler software to the DNA sequence of the *L. major* origins of replication did not identify any conserved motifs or significant features. This analysis was run with a minimum occurrence of 18 (half the total number of chromosomes). A second motif identification tool, MEME, was also applied to the sequence predicted to contain DNA replication origins in *L. major* but, again, this did not yield significant results and the enriched sequences identified by both of the motif searching software were predominantly homopolymer tracts e.g CCCc or AAAaA (Bailey et al., 2009).

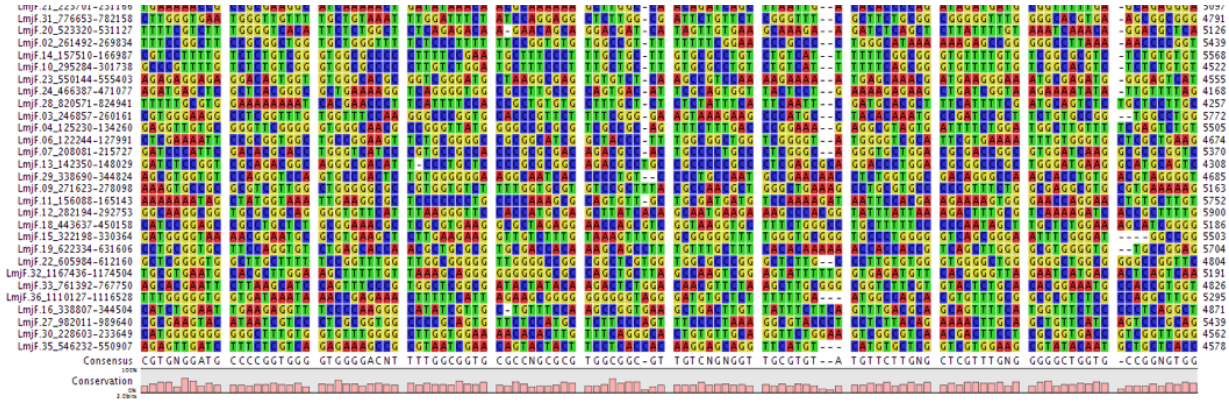
A program was then written in Java to initially compare counts of individual bases within origin and non-origin SSRs in *L. major* (https://github.com/CampbellSam/base-counts). This analysis found that the GC-/AT-content of the origin and non-origin SSRs is roughly the same. The base composition of origin-containing SSRs (A: 21.8%, C: 28.16%, G: 28.47% , T: 21.57%) was very similar to that of non-origin SSRs (A: 22.86%, C: 27.04%, G: 27.22%, T: 22.87%) Enrichment of long homopolymer tracts have previously been observed in the *L. major* genome (Zhou, Bizzaro, & Marx, 2004). The original script designed to analyse base frequency, was extended to include larger tracts, incrementally from 3 bases up to 6, and found that non-specific homopolymer runs tend to appear more often at origin-containing sequences than those without, although this requires further statistical analysis as the observed differences are small and no strong conclusions can be drawn from the current data. It is possible that these sites and other repetitive regions are masked in some motif searches which include a repeat-masking step. It is also likely that the tracts are not spatially conserved in the large regions predicted to contain origins of replication in *Leishmania*, making it difficult to detect conservation.

Figure 4.1 shows a multiple sequence alignment of the DNA sequence from SSRs in *L. major* that contain mapped origins, demonstrating that there is no

conserved sequence similarity across these regions, which is consistent with current knowledge of eukaryotic replication origins.

Figure 4.1 Sample extract of multiple sequence alignment highlights lack of conserved sequence.

Multiple sequence alignment between *Leishmania major* SSRs predicted to contain an origin of replication by MFAseq



4.3 Defining SSRs in *Leishmania* & *T. brucei*

SSRs have been identified as important sites in the genomes of *T. brucei* and *Leishmania* and current research suggests that the regulation of essential processes is more likely to occur in these regions. The current annotations for SSRs could be greatly improved as they are based on the coordinates of the surrounding gene boundaries. The coordinates of untranslated regions (UTRs) were not available in *Leishmania* outside of *L. major*. Gene coordinate annotations are based on the CDS and the true gene boundaries are currently not known. Due to this lack of reliable annotations plus the variation in the size and structure of these regions, the lack of any consensus sequences and the absence of epigenetic studies in Leishmania, it has so far been difficult to characterise SSRs.

4.3.1 Updating the SSR coordinates in *T. brucei* Lister427

SSR coordinates determined by histone markers and surrounding gene boundaries were available for the *T. brucei* TREU927 genome but previous attempts, by the bioinformatics team at the Wellcome Centre for Molecular Parasitology, to transfer these to the *T. brucei* Lister427 genome using orthology had been unsuccessful. The annotations were transferred using RATT (Rapid Annotation Transfer Tool) (Otto et al., 2011) but several problems occurred due to the inversion of some gene arrays in *T. brucei* Lister 427 relative to *T. brucei* TREU 927. Several predicted SSRs in *T. brucei* Lister 427 were very large, with one region spanning almost 1Mb, and contained gene arrays.

To begin the process of manually updating the coordinates for these regions, I used the online resource TriTrypdb.org to filter the annotations generated for the *T. brucei* Lister 427 genome by RATT and compile a list of those which contained protein-coding genes and therefore required investigation. Of the 91 SSRs annotated, 47 of these contained genes, of which 41 of these were predicted to contain protein-coding genes. I used the gene annotations from *T. brucei* TREU 927 to manually correct the start and end of each SSR using the orthologous coordinates in *T. brucei* Lister 427. In most cases this was straightforward, as the SSR was clear from the presence and direction of entire gene arrays, however I observed 5 cases where several small consecutive SSRs

had been merged into one large region, possibly due to subtle genomic rearrangements between the two strains. This process yielded a corrected list of coordinates for SSRs in *T. brucei* Lister427.

4.3.2 Different methods to refine SSR coordinates

As mentioned previously, the current SSR coordinates are based on the boundaries of the surrounding genes and it is likely that these could be improved by updating the gene annotations. This would greatly aid the machine learning analyses as DNA sequence reads aligning across SSRs comprise the input for machine learning, including predicting whether or not an SSR contains an origin based on thousands of sequence features. In order to improve the output of this algorithm, it is essential to first refine the initial input data. We approached this problem in two ways, the first using *de novo* gene prediction to update the existing gene models and UTRs, and in the second by looking at gene expression coverage data to determine an expression cut off threshold at regions of low signal between gene arrays.

4.3.2.1 Use of RNA-seq to perform *de novo* gene annotation in *L. major*

This analysis sought to update existing gene models by extending the annotated coordinates at 5' and 3' untranslated regions (UTRs) and potentially map spliced leader and polyA sites through the integration of RNA-seq datasets in an effort to reduce the length of sequence considered to be a SSR. SRA-tools was used to access and download 20 *L. major* Friedlin RNA-seq datasets available in the sequence read archive (http://www.ncbi.nlm.nih.gov/sra) (Table 2.1). The reads from all samples were aligned using Hisat2 (Kim et al,. 2015). The resulting RNA-seq alignment files were then merged into a single alignment file and *de novo* transcript reconstruction was performed using Trinity. Gene structure annotations were then generated using the complementary tool PASA (Program to Assemble Spliced Alignments) (Grabherr et al., 2013; Haas et al., 2003). The updated annotations were viewed against the original gene models using Integrative Genomics Viewer (IGV). I used bedtools intersect on both datasets to locate updated regions across the genome and found the majority to be located in UTRs. Extended UTRs were observed in 7 chromosomes and on 2 of

these chromosomes, these modifications were proximal to SSRs containing origins which would potentially allow gene coordinates to be updated. A SSR on chromosome 14 which does not contain an origin could also be updated. An example of the output from this analysis can be viewed in figure 4.2.

The number of origin-containing SSRs that could be updated by this process was only 2 of the total 36 and several optimisations would be required to make this a viable approach. Several issues arose during the analysis of the large merged RNA-seq dataset, which hindered progress of this approach. The total size of temporary files generated by the software outlined above was extremely large (in excess of 300GB) due to the size of the input dataset, and lack of disk space meant that, despite repeated attempts, the pipeline could not be run locally. The use of a large EC2 instance was therefore required to complete the analysis shown, which builds additional costs. This cost may need to be factored in when planning to repeat this process. However, it would be necessary to first determine the computational power required to run the samples individually and the number that could run in parallel on a local machine.

The output from this analysis was extremely noisy and it is likely that merging the datasets together has masked the subtleties of potentially interesting updated annotations. Due to this, I think it would be beneficial to repeat the same process with each of the datasets individually and merge the final updated annotations at the end. If this analysis was repeated it could also be improved by the generation of a script to compare the new and original gene annotations.

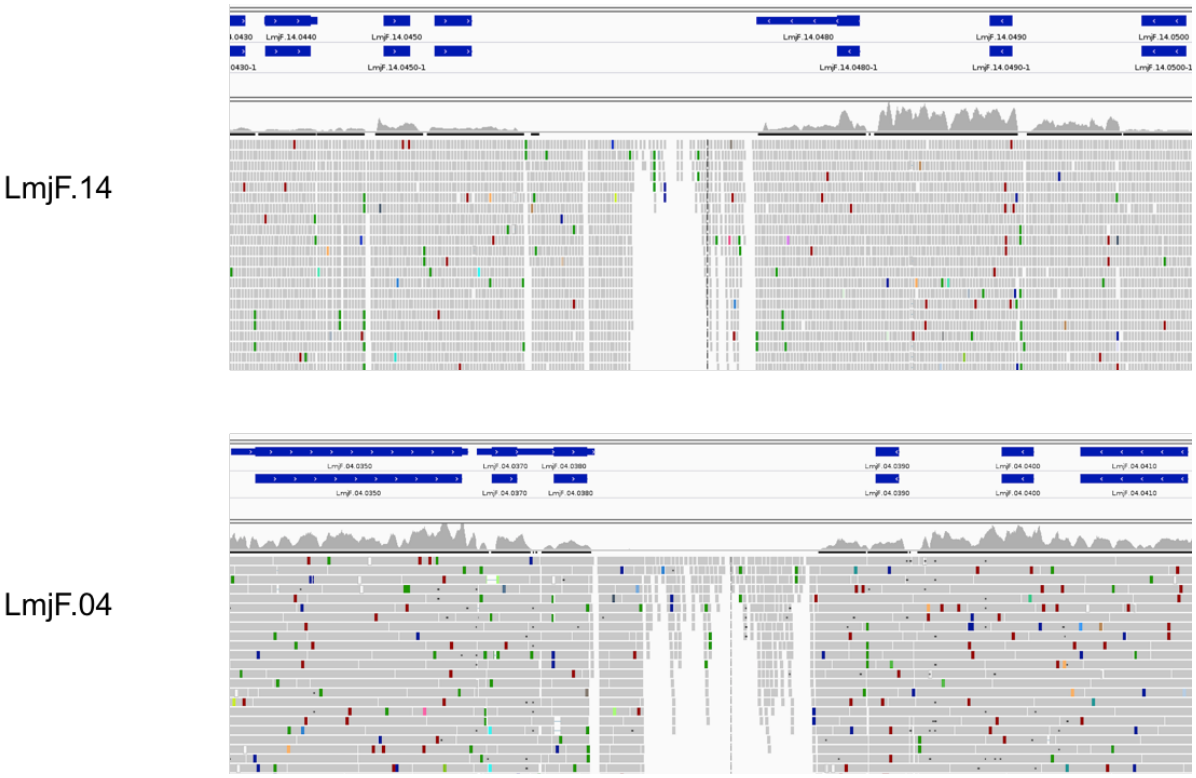


Figure 4.2. Updated UTRs at SSRs based on *de novo* gene annotation in *L. major* chromosomes 4 and 14.

Visualisation of the current gene annotations and the *de novo* annotations, assembled using Trinity and PASA, was performed using IGV (J. T. Robinson et al., 2011). Both sets of annotations are shown here across a SSR in *L. major* chromosome 14 and chromosome 4. In both chromosome panels, the top track shown is the updated gene annotations with the pre-existing annotations shown below, each gene is displayed as an individual box with white arrows to indicate direction (both shown in dark blue). Individual mapped reads supporting the *de novo* assembly are shown in grey and coverage is represented above in a grey histogram plot. The SSR in each panel is marked by the reduced mapping coverage of RNA-seq reads. Coloured highlights in mapped reads provide information regarding bases that do not match the reference genome and insertions. Deletions are denoted by a gapped read joined by a black bar to show the deletion relative to the reference..

4.3.2.2 Determining a read coverage threshold of RNA-seq data in *L. major*

As an alternative to above approach, a script to analyse the read depth coverage of the RNA-seq data was established with the intention of generating a read depth cut-off threshold to determine regions of low gene expression and refine the gene coordinates this way. SSRs generally do not contain pol II transcribed, protein-coding genes, although small arrays of tRNAs associated with transcription initiation or termination, and some rRNA genes, transcribed by RNA

pol III and RNA pol I respectively, are occasionally present within some SSRs. Based on this, few RNA-seq reads would be expected to map across SSRs and a coverage threshold could be used to refine the coordinates that denote the boundaries of SSRs.

Read coverage data was generated across each chromosome using htseq-count, a command line tool made available by HTSeq (Anders et al., 2015). A Python script was then generated to calculate average read depth across the chromosome and assess relative coverage across SSRs. This analysis requires further refinement in assessing a cut-off threshold and is currently incomplete, therefore data is not shown.

4.3.2.3 Spurious alignment of RNA-seq reads at SSRs

During the analysis of RNA-seq read data across SSRs, a correlation between the number of reads with potentially spurious alignments aligning to non-origin SSRs was observed relative to origin-containing SSRs in *L. major* (Figure 4.3). Non-specific alignments were consistently observed in non-origin SSRs at a higher frequency than in origin-containing SSRs. In *Leishmania,* SSRs containing an origin are much larger than those which do not but, despite this observation, we detected an increase in spurious RNA-seq reads aligning across non-origin SSRs compared with SSRs that contain an origin (Marques et al., 2015). 30 of the 36 predicted origins in *L. major* map to sites of transcription initiation and very few are at sites of transcription termination where transcription may run-on. This may account for the decreased read depth observed across SSRs containing an origin of replication. To investigate this further, it would be interesting to extend this analysis of RNA-seq read depth at SSRs to *T. brucei* where SSR length does not differ significantly between those containing origins and those without.

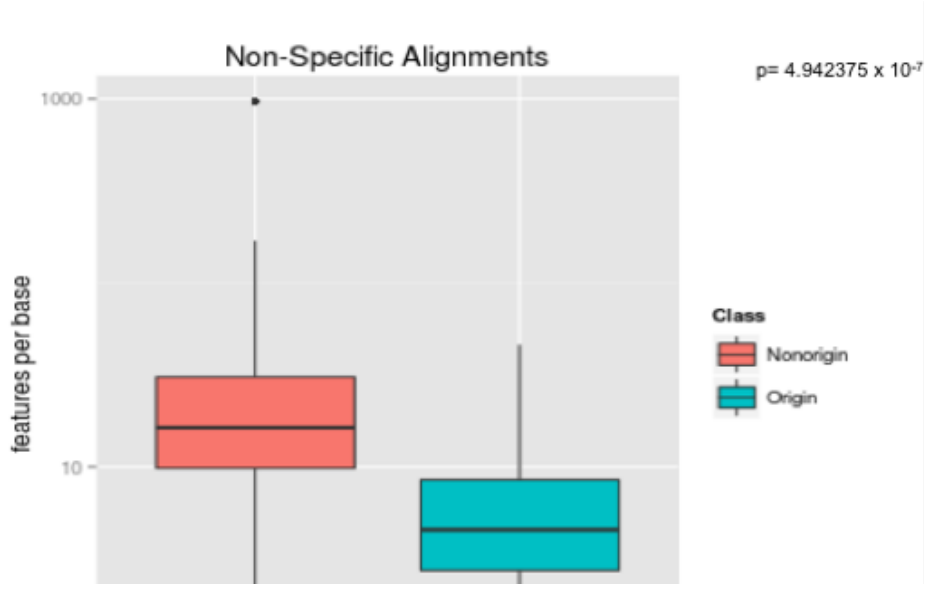


Figure 4.3. Low number of spurious RNA-seq read alignments in origin-containing SSRs in *L. major* relative to non-origin SSRs.

Boxplot visualisation of *L. major* RNA-seq aligned reads across SSRs that are predicted to contain a DNA origin of replication mapped by MFAseq (blue) relative to SSRs that are not predicted to contain an origin.

4.4 Investigation of machine learning and characterisation of DNA sequence features at SSRs

This analysis investigates the application of machine learning algorithms, focusing on support vector machines (SVMs) and their implementation in binary classification of DNA sequence, as an alternative approach to attempt to identify sequence features at kinetoplastid origins of replication. As not all SSRs are predicted to contain an origin of replication, it is reasonable to hypothesise that there are differences in the sequence features associated with the processes occurring at origin-active and inactive SSRs, and it may be possible to use these to distinguish SSRs that contain an origin from those without. Support vector machines are suited to this problem as the algorithms learn features of different classes within a training data set and use this information to accurately predict these classes in an uncharacterized set of test data. Using this approach, it may therefore be possible to characterise the sequence features associated with origins of replication and predict potential origins in less studied, related parasite species.

4.4.1 Generating *k*-mers from DNA sequence data

The input required to train the SVM, for this analysis, consists of small sequence features of *k* length, termed *k*-mers. In *L. major* Friedlin, DNA sequence reads that align across the SSR coordinates were extracted from the full alignment file and treated as samples belonging to the origin-containing or non-origin classes. This provides a large sample size compared to using the single sequence for each SSR form the reference genome alone. Each of the mapped reads was then broken up into *k*-mers to form the training data for the SVM.

The *k*-mer generation step was written in Python and performed numerous times as improvements were required. The initial attempt used a sliding window across each SSR, from size 6bp up to 50% of the read size but the time consumed by this process was extensive and generated too much redundancy. An alternative approach was therefore adapted to generate tiled *k*-mers instead, which reduces the size of the input dataset but is more efficient. The first set of *k*-mers generated in this manner were 10bp in length, although further investigation of varying lengths may improve the training of the SVM. Using a larger *k*-mer size,

the chance of detecting associated spatial features increases but the frequency of observing a specific *k*-mer in each sample diminishes resulting in a sparse matrix that predominantly consists of single counts.

Several issues occur when generating sparse data,, and efficient data storage became a key factor in this analysis. In the preliminary stages of this process, *k*-mers were stored as lists with origin and non-origin labels and output to several files within a hierarchy of directories. This became a time-consuming step that was later adapted, avoiding the file output step. The optimisation of the *k*-mer generation step requires a strong knowledge of computer science and appropriate methods of efficient data storage and access and was therefore beyond the scope of the current project. The process of generating *k*-mers therefore remained a time-consuming step for a significant duration of this analysis but was optimised by Dr Nicholas Dickens through storage within the pipeline and this step has now been integrated as part of the SVM pipeline described below and is no longer an individual pre-requisite script.

4.4.2 Building the SVM pipeline

The SVM pipeline was built in Python and the machine learning algorithm is implemented through scikit-learn, a Python package of machine learning classification and clustering tools (Pedregosa et al., 2011).

Support vector machines classify data with a large number of features by implicitly plotting them as vectors in high-dimensional space. The kernel algorithm that performs in high-dimensional space can be changed by the user depending on the type of classification problem. Changing the kernel affects the decision boundaries that will split the different classes. There are three commonly used kernel types: linear, radial basis function (rbf) and polynomial. Linear classifiers are efficient when applied to text classification and were used in this analysis. In the case of a linear classifier, the decision boundary is a hyperplane that linearly separates the two data classes. In our analysis, the two classes are SSRs containing an origin and SSRs not known to contain an origin.

The first step is to vectorise the input *k*-mers so that the classifier is able to implicitly plot them. The weighting of the features is transformed in the dataset

so that more common features have a low weighting while rare *k*-mers have a higher weighting, as they are more likely to be indicative of the associated class. The classifier is then trained on the prepared data where it learns the groups of features that are the best representatives of each class.

Multiple parameters must be defined when training the classifier and steps were implemented to optimise these. The performance of the linear classifier was also compared with that of a linear model stochastic gradient descent (SGD) classifier with and without optimised parameters. The output from each of these is shown in Table 4.1.

The current SVM pipeline was co-written with Dr Nicholas Dickens and can be found at https://github.com/CampbellSam/Origin-classifier.

	Linear	**Linear SGD**	**Linear SGD optimized parameters**
LmjF eS (training)	99.98	98.10	99.99
LmjF G2	97.97	94.72	97.67
LmjF unsorted	92.12	81.12	91.26
Lmj WT unsorted	93.33	82.70	92.59
Lmx G2	78.50	72.50	77.11

Table 4.1 The accuracy of the original linear kernel compared with two optimised alternatives, a linear model SGD and a linear model SGD with optimised parameters. The performance was assessed in the available datasets: early S phase *L. major* Friedlin data set used for training the classifier, G2 phase *L. major* Friedlin, unsorted *L. major* Friedlin, unsorted *L. major* wild-type strain and a G2 phase sample from *L. mexicana.*

4.4.3 Testing the SVM classifier in *Leishmania*

The linear classifier was trained on DNA sequence reads aligning to SSRs in *L. major* Friedlin cells from early S phase labelled origin/non-origin containing regions. Testing the classifier back on the training data achieved an accuracy of 99.98% in *L. major*. Initial testing of the trained classifier on the training data is expected to give a high accuracy and it is important to be careful at this stage not to over-fit the classifier to the training data. This could be avoided by performing a n-fold cross-validation on the training set although this is a more significant problem in multi-class problems and was not currently implemented here.

The trained classifier was then used to predict which SSRs contain origins in unclassified test data sets from *L. major* and *L. mexicana*, since origins had been mapped by MFAseq in the latter (Marques et al., 2015). Results from all of the test datasets are included in table 4.2. When testing the classifier on the G2 data from *L. major* Friedlin to predict SSRs containing origins, 97.97% accuracy was achieved. Although this approach requires further optimisation, this preliminary result indicates that there are *k*-mers within the samples that can be used as features to predict replication origins. High accuracy was maintained when testing the classifier on a different strain of *L. major*, indicating that the same features are conserved within *L. major* strains. The classifier was also applied to two unsorted samples to identify and account for any bias from a larger number of reads aligning to origins during early S phase. It was then important to establish whether the classifier could use the features identified from the *L. major* training set to predict origins in *L. mexicana*. When taking the reads as samples, as with the training set and all other test sets, a drop in accuracy was observed, although 78.6% of reads aligning to SSRs were still correctly classified as origin-containing/non-origin. When removing the read level from the analysis and allocating all *k*-mers only to the SSR they belong to, a steep increase in accuracy occurred. The sample size was much smaller as only 124 SSRs are input for classification; however, 96.77% of these regions were classified correctly. When the classifier was applied to *L. brazilensis*, a distinct drop in performance was observed that may highlight an overfitting issue.

Test data	Accuracy score	No. of samples
L. major Friedlin G2	97.97%	50,815
L. major Friedlin unsorted	92.12%	1,553,127
L. major WT unsorted	93.33%	135,652
L. mexicana G2 (reads)	78.59%	53,352
L. mexicana G2 (SSRs)	96.77%	124

Table 4.2. SVM performance when predicting the presence of DNA replication origins in SSRs in *Leishmania*

Output of testing the classifier on each of the data sets when trained on reads aligning to SSRs in early S phase *L. major Friedlin.* The accuracy is the percentage of reads that were correctly classified as belonging to origin or non-origin containing SSRs, except in the case of the final data set where SSRs were treated as samples instead of the reads. The total number of samples is also included.

Although we observed a high success rate with this approach, many of the steps require refinement and the assessment of accuracy must be greatly improved. There are several modifications to the approach which can be made to overcome the current issues, most importantly the addition of a cross-validation training step to avoid over-fitting and inclusion of a more robust assessment of classifier performance (for instance, the use of ROC curves to evaluate accuracy). Better understanding of the SSRs, thus providing robust input data would dramatically reduce the bias introduced through inclusion of sequence present in SSRs which is not involved in DNA replication and therefore improve the significance of extracted features.

4.4.4 Alternative method using gkm-SVM

The SVM pipeline implemented above samples variation in the genome through the use of reads as input data as opposed to raw sequence data. An alternative approach is to allow for gaps in the sequence features used in the classification. As described previously, increasing the *k*-mer size increases the identification of spatially-associated features but also introduces noise and leads to low *k*-mer counts. The use of gapped *k*-mers could also improve the identification of associated DNA sequence features (Ghandi et al., 2014).

To assess the different approaches, we applied the gapped k-mer classifier gkm-SVM, generated by Ghandi et al, to our data (Ghandi et al., 2014). A linear kernel was first trained on the sequence data from SSR regions containing origins (positive dataset) and those which do not contain origins (negative dataset). The trained classifier is then tested on unlabelled sequence data from SSRs where we know the correct labels. The classifier predicts whether the sequences from each input region belong to origin or non-origin containing SSRs and accuracy is determined based on the number of correct predictions. Gkm-SVM was able to correctly predict 25 of the 36 origins in *L. major*. Of the 33 origins mapped to 36 *L. mexicana* chromosomes, 23 were predicted correctly by Gkm-SVM. The performance of the algorithm appears highly consistent across both samples and all non-origin SSRs were classified correctly, indicating a very low false discovery rate (FDR). It may be beneficial to apply this classifier to the *T. brucei* genome in future studies to further assess performance and compare this method with our own approach.

4.4.5 Predicting the presence of DNA replication origins in *T. brucei* 427 BESs

Before the generation of the Lister 427 SSR coordinates, we were unable to fully apply the machine learning approach to the characterisation of origins in *T. brucei* as we only had a single dataset comprised of the regions in 927. However, we were able to investigate BES1, the previously investigated region in *T. brucei* Lister427 bloodstream form (BSF), which appeared 'origin-like' in output from MFAseq analysis (Devlin et al,. 2016).

To further validate this observation, we trained a second support vector machine learning algorithm on the reads aligning to origin and non-origin containing SSRs in *T. brucei* TREU927. We then used this algorithm to predict the presence of active origins in the sequence of the BES contigs in *T. brucei* Lister427. Control regions were added to ensure the data was classified correctly: sequence from a region in *T. brucei* 927 known to contain an origin was added to the orthologous site in Lister427 test file as a positive control; the negative control was a 7kb SSR from TREU927 which is not predicted to contain a replication origin.

Despite the difference in aligned reads at BES1 between eS and G2, an active origin was not predicted for this site, or any of the other expression sites. The analysis was repeated with smaller tiled windows across each region, and similarly, all of the small regions were predicted as non-origin sequence. These data may indicate that any origin in the BES is distinct in sequence from core origins. An alternative explanation for the early replicating behaviour observed at this expression site is that the actively transcribed BES1 is located proximal to an active origin on a megabase chromosome, but this hypothesis could not been tested due to lack of linkage of all BES to the core chromosome genome annotations and a lack of complete annotation of the subtelomeres.

4.5 Conclusions

The analysis presented here firstly investigated sequence features at SSRs in *T. brucei* and *L. major*. Motif searching analysis highlights the enrichment of homopolymer tracts at origins of replication in *L. major*, while specific searches

for the G4 motif in *T. brucei* revealed its presence in abundance, although no current association with replication origins has been made in this genome.

Although attempts to refine the SSR coordinates proved difficult, the approach of utilising RNA-seq data to better define these regions could still be successful. The analysis outlined here has several inefficiencies but could be improved through the use of smaller datasets and refinement of the comparison between the updated and pre-existing gene annotations. Investment of time in the generation of *de novo* gene annotations and a novel script to update the SSRs, combined with the secondary approach of investigating read depth cut-off, would provide a method to update these coordinates that could be applied across the kinetoplastid genomes.

These results also demonstrate that the application of machine learning is a promising approach, as the current SVM algorithm is able to correctly predict SSRs containing origins between strains of *L. major* and also across species, at least to *L. mexicana*. Application of gkm-SVM to our data also showed promising results; classification of non-origin SSRs by this software was comparable with the results from the in-house sci-kit learn SVM implementation although a high false-negative rate was observed when classifying origin-containing SSRs. This indicates that although previous methods have failed to identify sequence motifs enriched within origins, there is conserved sequence information that can be identified between species to predict origins of replication.

5 Use of genomic techniques to investigate the relationship between mosaic aneuploidy and DNA replication in *Leishmania*

5.1 Introduction

A feature within the *Leishmania* genome is its ability to tolerate pervasive aneuploidy, a property thought to allow the parasite to efficiently adapt to changing environmental conditions. The presence of aneuploid chromosomes is usually detrimental in other organisms and is frequently associated with disease, such as cancer and trisomy 21 in humans (Pfau & Amon, 2012). However, the budding yeast *Saccharomyces cerevisiae* is able to tolerate chromosomal aneuploidy and a model system has been established to study the causes and consequences of aneuploidy in this model organism (Mulla et al., 2013; Parry & Cox, 1970).

In response to changing environmental pressures, *Leishmania* parasites undergo gene and whole chromosome copy number variation, often to regulate the expression of genes that may be drug targets and efficiently develop drug resistance (Sterkers et al., 2011; Ubeda et al., 2008). Constitutive mosaic aneuploidy has been observed across several *Leishmania* species and variations in chromosome ploidy are observable between species and even between cells in a population (Lachaud et al., 2014; Rogers et al., 2011).

The mechanisms underlying chromosomal amplification and the tolerance of any gene dosage effects are not currently understood. However, recent studies have highlighted the large number of repeated DNA sequences throughout the *Leishmania* genome and found that these sequences contribute to the recombination-driven rearrangement of genetic material and the formation of extrachromosomal elements, termed episomes, that can also be sustained within the genome (J. M. Ubeda et al., 2014). The relationship between these processes and DNA replication is currently unclear, but is of interest to further investigate potential links as the process of DNA replication may be unusual and occur from a single dominant initiation site on each chromosome, which is not the case in the closely related *T. brucei*.

To further investigate genome plasticity in *Leishmania*, DNA sequence data from *L. mexicana* M379 promastigotes in serial passage was analysed using genomic techniques. This investigation is also interesting as parasites are often cultured for long periods of time in a variety of investigations, including drug resistance

studies but it is not well understood how the parasite genome adapts to these conditions. This study could therefore also highlight regions that are prone to genomic changes in *in vitro* passage conditions and help to eliminate a bias in regions of the genome which are not necessarily correlated with the focus of the study, and may be misinterpreted as key areas of investigation in biological investigations such as drug resistance studies.

5.2 Genomic changes during serial passage of *L. mexicana* promastigotes

To investigate the adaptation of the *Leishmania* genome to culture conditions, post-mouse infection DNA sequence data generated from *Leishmania mexicana* M379 promastigotes extracted at regular intervals throughout serial passage was analysed. An overview of the experimental set up is shown in Figure 5.1. *L. mexicana* parasites from a mouse footpad lesion were passaged in order to generate an adequate amount of DNA and then sequenced. This sample was then split into 3 and the rest of the experiment was carried out in triplicate. DNA samples were then extracted at passages 5, 10, 16, 20 & 29. At serial passage 16 (sp16), a sample from each replicate was used to inoculate mice in the footpad and allowed to form a lesion. Parasites were then extracted from the lymph node of each mouse and passaged before sequencing. These data allow us to compare the parasite genomes after a mouse infection and during growth in culture, and also ask whether any putative adaptive changes are 'reset' after a host infection. The parasite culture, passaging and DNA extraction were performed by Dr Samuel Duncan, a former PhD student in the research group of Dr Jeremy Mottram.

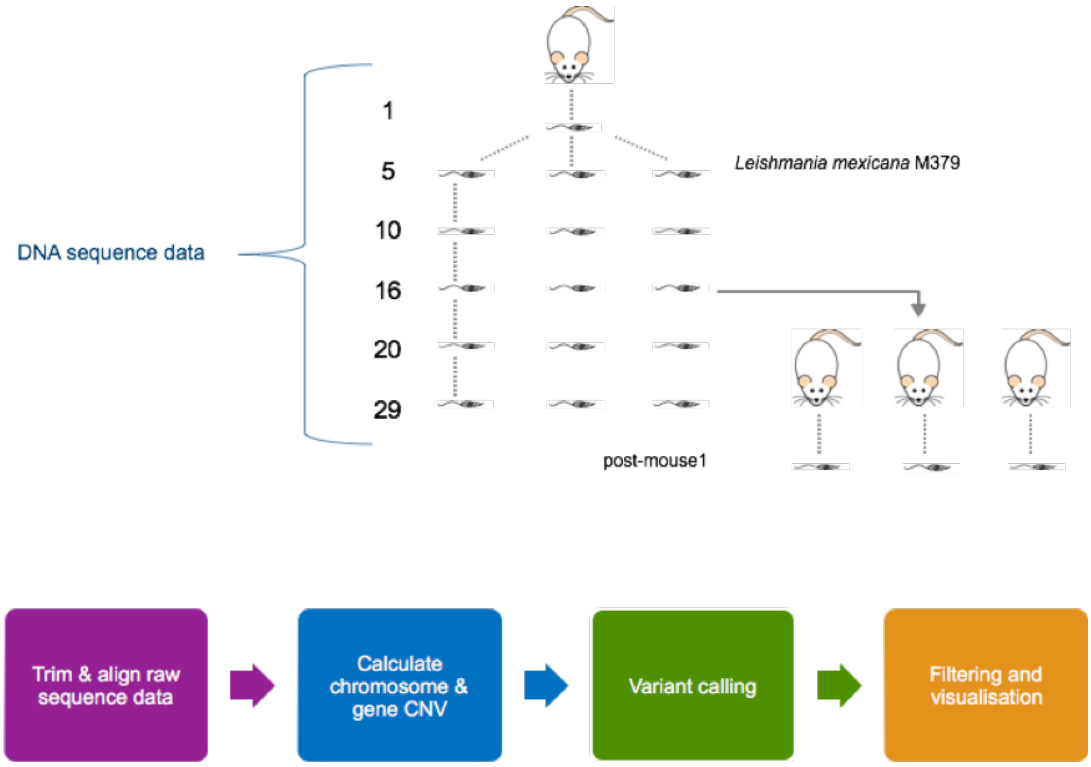


Figure 5.1 Overview of *L. mexicana* serial passage experimental set up.

Outline of experimental set up and serial passage samples used to provide DNA sequence data for analysis at regular intervals (above). Data processing pipeline from raw data to aligned reads which were then used for further analysis and visualization including chromosome and gene CNV data, variant calls through to filtering and visualisation. (below).

5.2.1 Genome analysis using the PReP pipeline

5.2.1.1 Preparing the read data and alignment similarity

The DNA sequence data for each sample was pre-processed using TrimGalore and aligned to the *L. mexicana* U1103 genome using Bowtie2 (Langmead & Salzberg, 2012). The alignments were validated using samtools to ensure high quality data. At the time that the wet lab work was performed a change in sequencing platform at Glasgow Polyomics sequencing facility took place, between samples sp10 and sp16. Samples sp0–sp10 were sequenced using the Illumina miSeq platform, while sp16–sp29 were sequenced on an Illumina NextSeq 500 machine. As there are differences in these platforms, I have used deepTools (https://github.com/fidelram/deepTools) to assess the similarity of the aligned DNA sequence samples to ensure the trends we observe (described below) are not due to a bias caused by this change in sequencing platform. As we normalize our samples before comparison, we should not expect to see any changes in ploidy which is explained by different sequencing platform coverage bias. Figure 5.2 shows the output from several approaches used to assess the similarity of each aligned sample. We did not observe distinct clusters of samples from each platform when plotting by PCA (figure 5.2A), although a degree of clustering occurred between the samples when assessed by Pearson correlation (figure 5.2B). To ensure any clustering observed in the above approaches was not due to a bias caused by sequencing platform, CummeRbund, available as an R package as part of the Cufflinks suite, was also used to visualize the effect of different sequencing platforms on the estimation of gene coverage (http://compbio.mit.edu/cummeRbund/manual_2_0.html). The gene abundance estimates, FPKMs (fragments per kilobase of transcript per million mapped reads), generated by Cufflinks are used to calculate the chromosome ploidy estimates and therefore any variation between samples here would have a significant impact on the analysis. However, we did not observe any significant inconsistencies when comparing median FPKMs across samples (figure 5.2C).

From this analysis, we confirm that any observations made in this analysis are not the consequence of bias caused by different sequencing platforms.

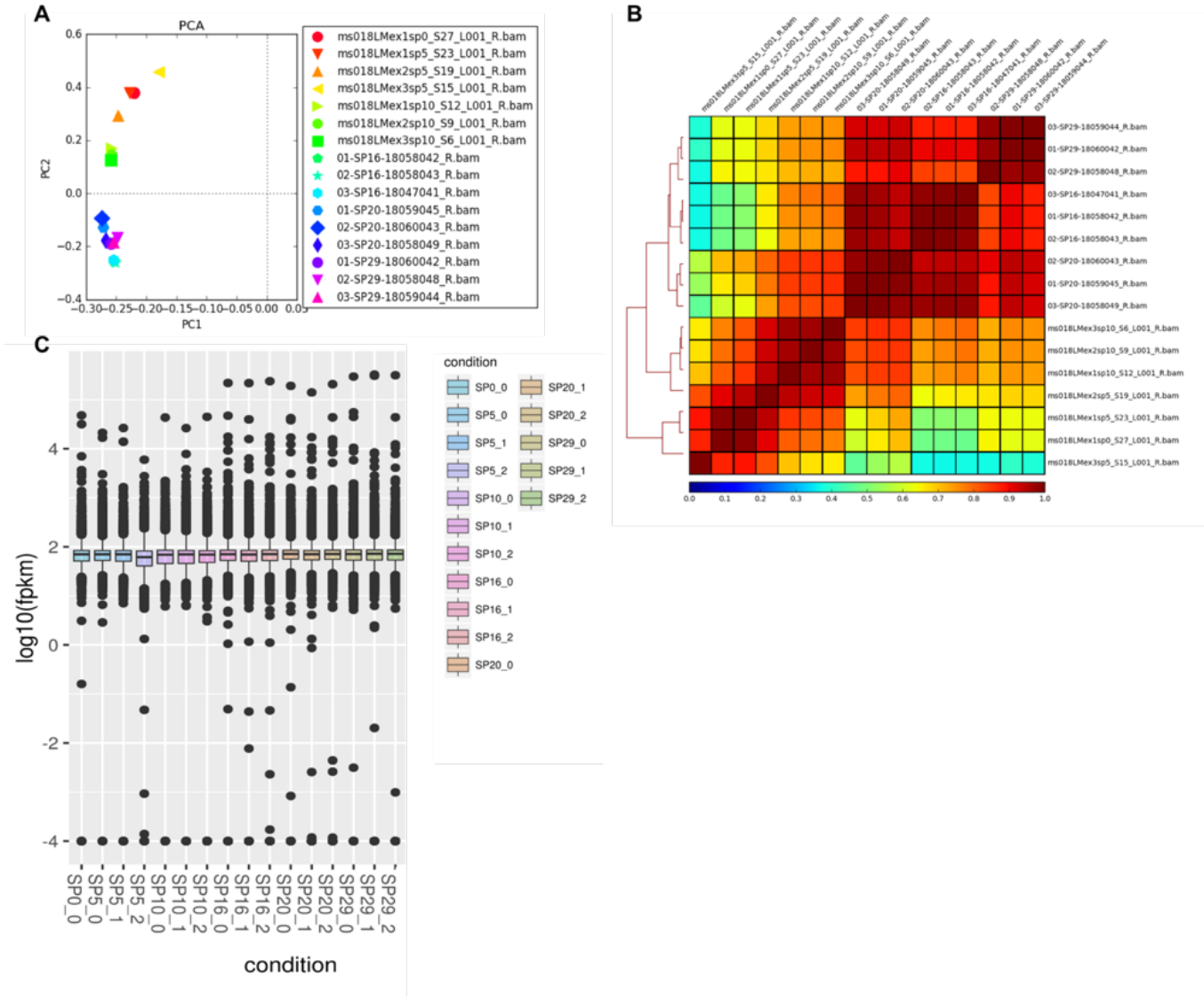


Figure 5.2. DNA sequence alignment similarity.

A. PCA plot representing variability between aligned read files for all *L. mexicana* serial passage samples. B. Pearson correlation of aligned sample similarity represented by heatmap. C. Boxplot, generated by the R package CummeRbund, representing gene FPKMs across sp0 and the triplicate samples from sp5 to sp19, here labelled as conditions. Samples sp0–sp10 were sequenced using the Illumina miSeq platform and samples sp16–sp29 were sequenced on an Illumina NextSeq 500 machine.

5.2.1.2 Estimating gene copy number and chromosome ploidy

The in-house pipeline (known as PReP) was used to perform genomic sequence analysis based on parameters within a specified configuration file. The pipeline is written in Perl and, given a configuration file, access to the required input files and software directories, uses the information contained in the configuration file to appropriately format data and make software calls to perform genome alignments, generate gene copy number variation information and perform SNP calling. Although the reference genome alignments were performed separately, PReP was used to generate gene (CNV) data that was then used to generate chromosome ploidy estimates for each sample.

Gene CNV data was generated using tools from the Cufflinks package, CuffQuant and CuffDiff (Trapnell et al., 2010). CuffQuant is used to generate abundance estimates for each gene and CuffDiff can be used to perform a series of analyses, including direct sample comparisons and time-series analysis. CuffDiff results were generated using the whole dataset and normalised copy number estimates were obtained for genes in each sample. In-house Python scripts were written to generate chromosome ploidy data for each sample and further scripts were used to generate CNV and fold change estimates across samples. The output from this analysis can be seen as heatmaps in figure 5.3. Chromosomes 16 and 30 were present in consistently higher copy number while chromosome 23 exhibited a decrease in copy number in two of the replicates across most samples (fig 5.3A). A clear trend emerged when fold change in copy number of each chromosome was evaluated over the time of passage relative to sp01 (fig 5.3B & C). Over several passages, predominantly between sp10 and sp16, small chromosomes showed an overall increase in copy number whereas larger chromosomes showed an overall slight decrease in copy number. Further support of this can be seen by examining the putative ‘fusion’ chromosomes present in the *L. mexicana* genome (L.mx08 and L.mx20) which are much longer in length (1.7 Mb and 3.3 Mb) and show a decrease in copies consistent with the observed trend.

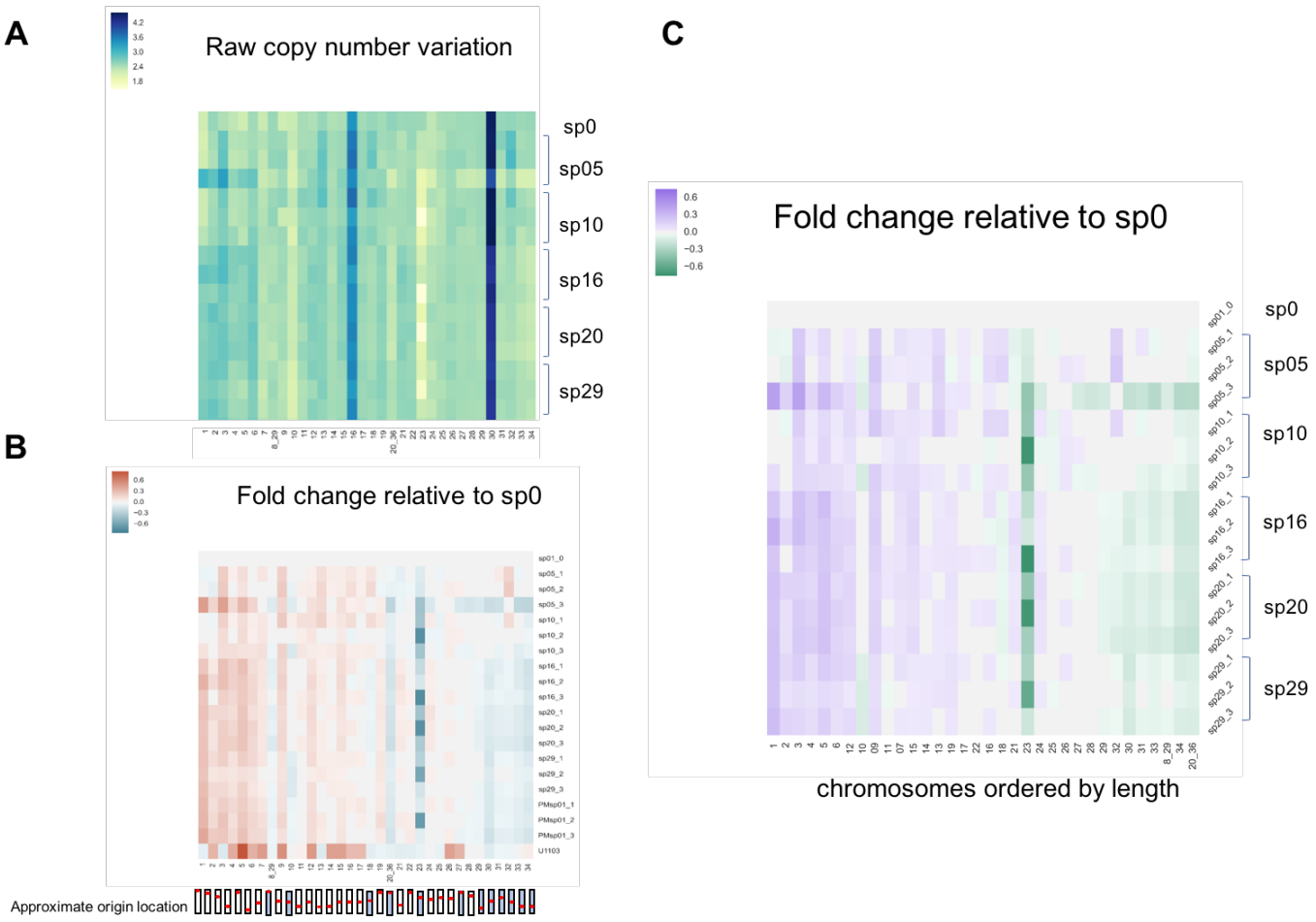


Figure 5.3. Chromosome copy number and relative fold change in *L. mexicana* serial passage

A. Raw copy number estimates for *L. mexicana* chromosomes (x-axis) in each serial passage sample (y-axis). Chromosomes are ordered numerically. B. Fold change of chromosome copy number in each sample relative to the sp0 sample. A model of approximate origin location is included below. Chromosomes are ordered numerically. C. Fold change of chromosome copy number in each sample relative to sp0, with chromosomes ordered by length.

5.2.2 SNP analysis of *L. mexicana* serial passage samples

5.2.2.1 SNP rate across passage samples and visualisation with Circos

FreeBayes variant calling software was applied to each sample to generate SNP data (Garrison & Marth, 2012). FreeBayes was chosen to perform SNP calling as it is haplotype-based and does not assume a diploid genome. The average SNP rate for each chromosome across each sample is shown in figure 5.4A. Variants/100 kb and variants/1 kb were calculated using a custom script (https://github.com/CampbellSam/calculateSNPrate) across every chromosome for each sample and visualised using Circos (figure 5.4B). When the SNP rate was plotted alongside the MFAseq, a reduction in SNP rate was observed on some chromosomes in the region surrounding the SSR which contains the origin of replication. Further statistical analysis will be required to verify the potential significance of this observation. At the level of SNPs/kb, it will be possible to further mine this data for regions with an enriched SNP rate that are prone to mutagenesis and also regions with a particularly low SNP rate which may contain highly conserved regions. Integration of SNP data with gene CNV and MFAseq data could provide insight into the structure of SSRs and the process of DNA replication and even highlight potential regions of interest.

5.2.2.2 Investigation of SNP rate at SSRs

Due to the predicted association between changes in chromosome ploidy and DNA replication, it was of interest to investigate this dataset for other potential relationships with origins of replication. To investigate the potential relationship between the frequency of SNP occurrence and sites of DNA replication initiation in *Leishmania*, the SNP rate was plotted with distance relative to the origin of replication for all chromosomes in triplicate samples from passage 29. As predicted replication origins co-localise with SSRs, the centre of each origin-containing SSR was taken as the location of the origin to maintain consistency, despite the fact that this is supposition. An increase in SNP rate is observed proximal to origins in all replicate samples, as shown in figure 5.5. This is in contrast to the previous observation made by eye from the data showin in figure 5.4B. Analysis of SNP rate within origin-containing SSRs compared to non-origin SSRs revealed a small but significant increase in SNP rate at the centre of origin-

containing SSRs. The different types of SSR were also considered (divergent, convergent and head-to-tail) in the context of origin presence. There was an increased SNP rate at divergent SSRs containing an origin compared to those without a predicted origin, although the significance of this in a biological context has not yet been recognized. The function of increased sequence variability at origin containing SSRs is currently unclear.

Analysis of large structural variations in these samples to detect any chromosomes with enrichment of break end, inversion, deletion or duplication events may reveal correlations with the previously observed changes in chromosome ploidy.

5.2.2.3 Generation of chromosome ploidy estimates based on allele frequency

Allele frequencies were calculated based on the SNP data generated using FreeBayes, which allowed us to confirm the previously predicted chromosome ploidies. This script can be found at https://github.com/CampbellSam/calculateAlleleFrequency. These data were visualised briefly using ggplot in Python but requires further refinement and is therefore not shown. The script used to perform this analysis could be extended to also test for loss-of-heterozygosity (LOH) in the *L. mexicana* dataset.

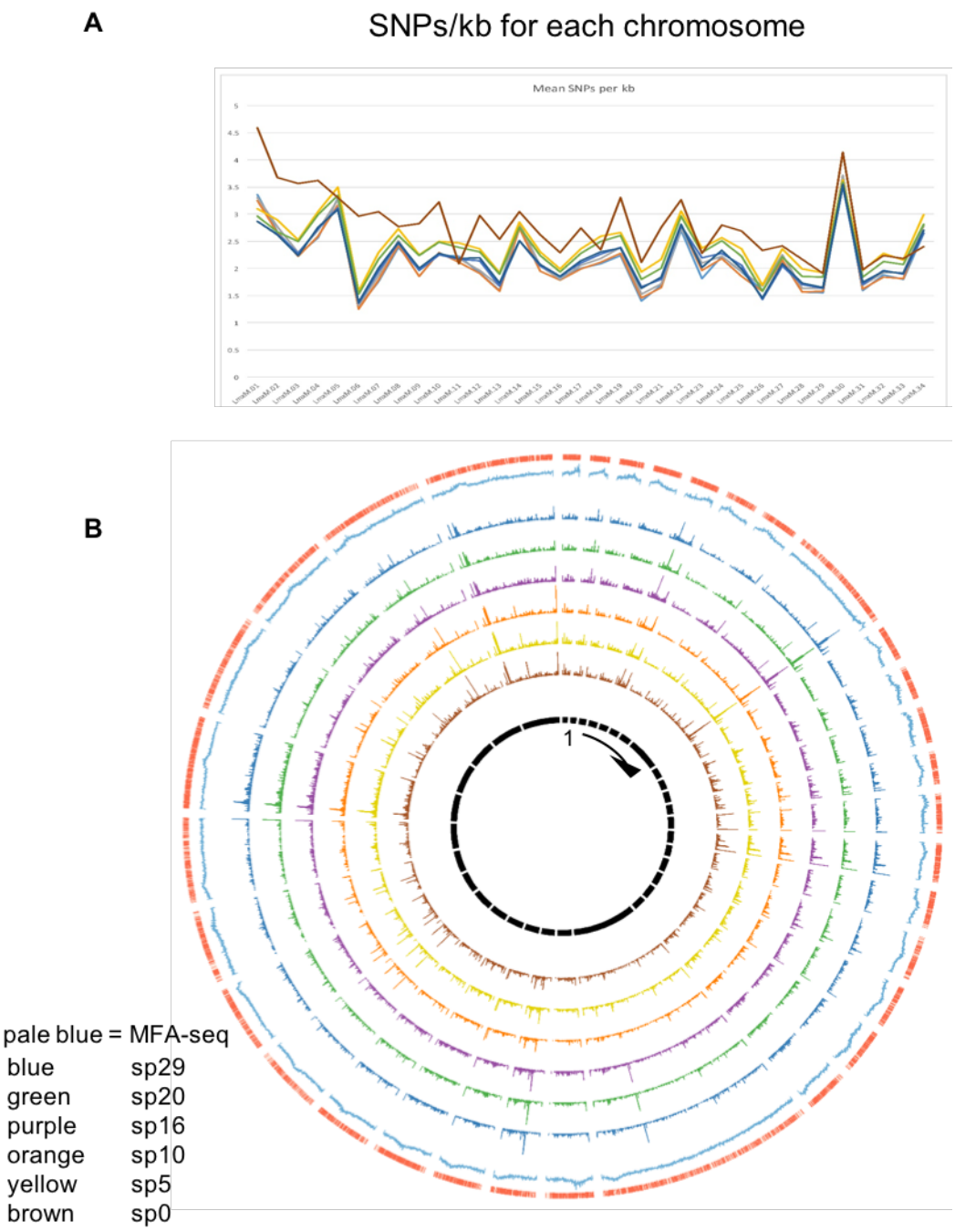


Figure 5.4 Overview of SNP rate per chromosome and SNP rate per kb across serial passage samples.

A. Mean SNPs/kb (y-axis) plotted for all *L. mexicana* chromosomes, ordered numerically (x-axis) for all samples (stacked coloured lines) Mean SNPs/kb included in dark red represents an independent *L. mexicana* U1103 strain provided for comparison (generated by Dr Nicholas Dickens). **B.** Circos visualisation of SNPs/kb across each chromosome in L. mexicana samples. From inside brown = sp01, yellow = sp5, orange = sp10, purple = sp16, green = sp20, blue = sp29 and the early S MFAseq profile is included in light blue with an added gene track highlighting enriched regions.

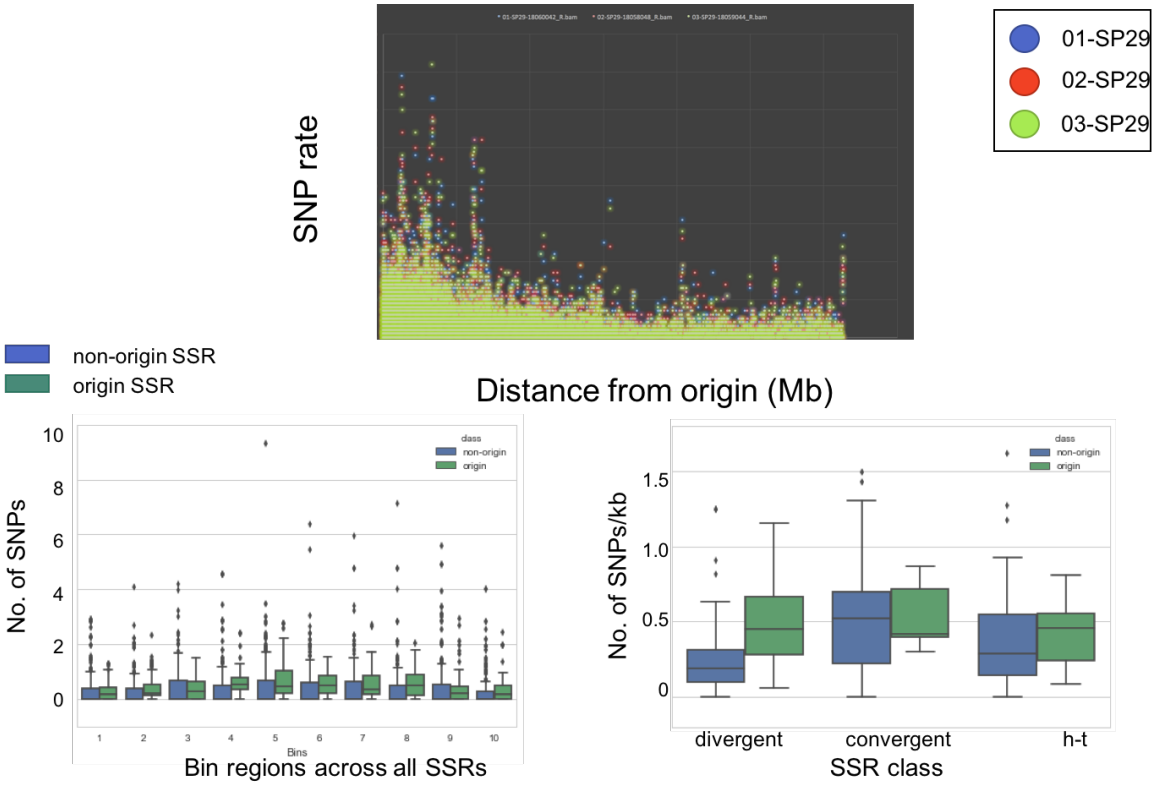


Figure 5.5. SNP rate and origins of replication.

Average SNPs/kb in triplicate samples from sp29 plotted for all chromosomes against distance from origin of replication (above). SNP rate plotted against location of DNA origins of replication visualisation generated by Dr Nicholas Dickens. Boxplot of SNPs across all SSRs broken into 10 bins to account for length. SSRs containing an origin are shown in green and non-origin SSRs are shown in blue (left). Number of SNPS/kb for SSR types - divergent, convergent and head-to-tail (right). Origin/non-origin SSRs shown in green and blue respectively.

5.2.3 Chromosome copy number variation vs chromosome size

Based on the previous observation of chromosome fold change during parasite growth in serial passage, the correlation between chromosome fold change and chromosome length was assessed. DNA read coverage was plotted using Fluff for three chromosomes of varying length to ask if any loss of coverage occurred as a result of incomplete replication (Georgiou & van Heeringen, 2016). Read depth was normalised and plotted for chromosomes 3 (~300kb in length), 20 (~3.3Mb in length) and 27 (~1.1MB in length) at passage 0 and passage 16. The shape of the overall read distribution was maintained in both the small and large chromosomes, although the largest chromosome, chromosome 20 (a 'fusion' chromosome), exhibited an overall decrease in coverage from sp0 to sp16, while

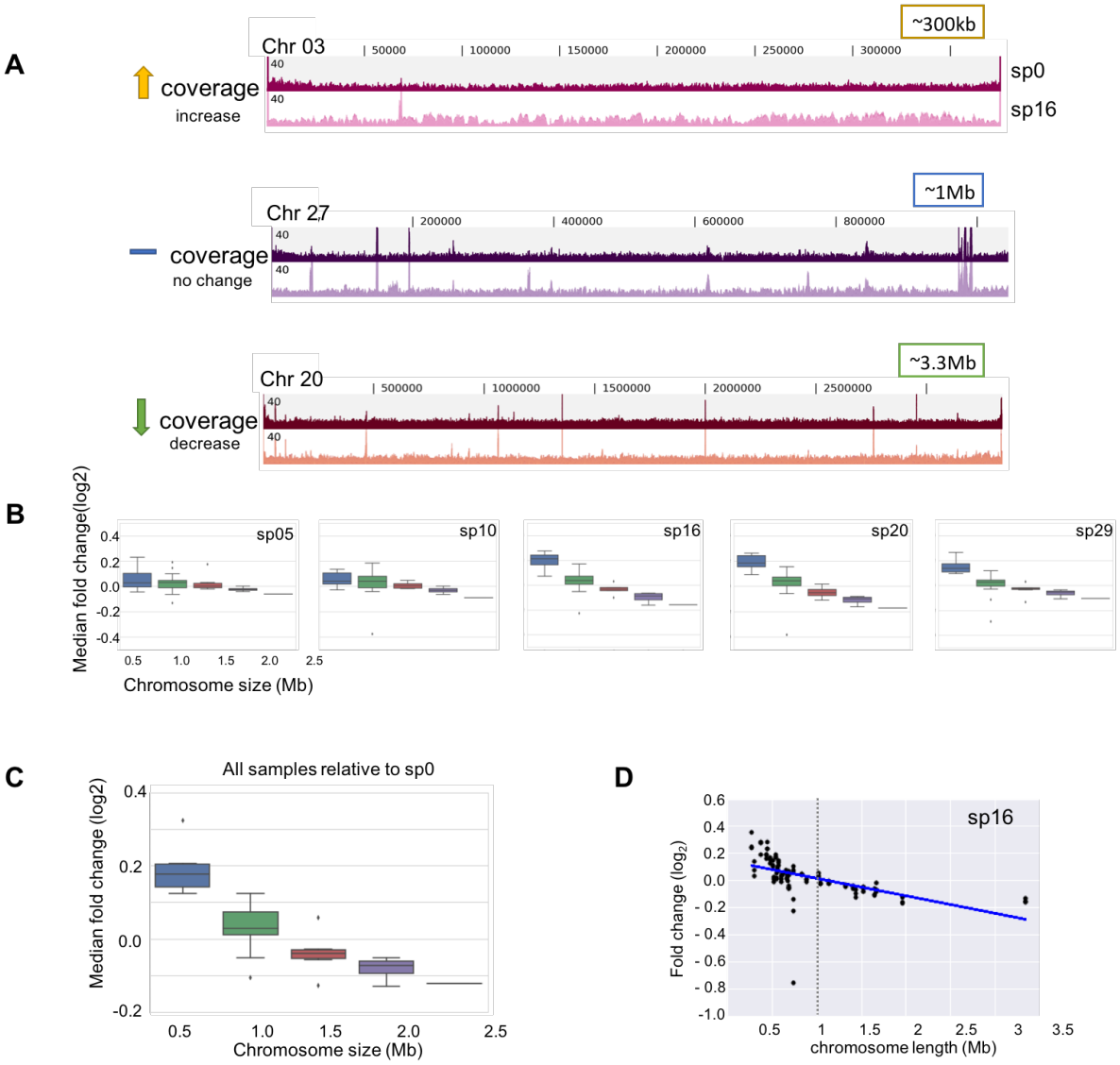


Figure 5.6. Changes in whole chromosome coverage and the relationship between relative fold change and chromosome size.

A. Normalised read coverage for chromosomes 3, 27 & 20 between sp0 and sp16. B. Chromosome length plotted against fold change shown for every interval of serial passage, from the left: sp5, sp10, sp16, sp20 & sp29. Chromosomes are binned by length, increasing in 500 kb increments. Blue = 0-500 kb, green = 500 kb- 1 Mb, red = 1-1.5 Mb, purple = 1.5 Mb and the remaining bin represents > 2 Mb. C. Chromosome fold change plotted against chromosome length for all samples relative to sp0. D. Linear regression of the relationship between chromosome length and relative fold change for samples at passage 16, r= -1.18966x10^{-7}.

the smallest chromosome (3) showed a slight overall increase in coverage. Chromosome 27 did not display a significant change in read depth. These findings are consistent with the previously observed fold change, and rule out that copy number variation is due to localised loss or gain of sequence within each chromosome.

Investigation of the median chromosome fold change relative to chromosome size in passage 5 through to 29 revealed that at passage 16, a trend begins to emerge where the small chromosomes - up to ~1Mb show a slight fold change increase, while the larger chromosomes decrease. This trend continues at passage 20 and 29. We predict a chromosome of length 1.1 Mb will exhibit zero fold-change, as shown in the linear regression in figure 5.6D. This suggests that chromosomes smaller than 1 Mb may be prone to re-replication while those larger than 1 Mb may be under replicated.

5.2.4 Investigation of advantageous genes on small and large chromosomes

Although genes are organised into arrays in kinetoplastid genomes, there is currently no evidence that this structure is based on a functional association. The previous analysis suggests that the copy number of small and large chromosomes is affected differently throughout promastigote growth in culture. It is therefore of interest to ask if genes of associated function may be enriched on large or small chromosomes. In order to test for the presence of genes with a potentially advantageous function, Interpro and Pfam protein domain annotations for *L. mexicana* U1103 were obtained from TriTrypDB version 26 and assessed for enrichment in a group of small chromosomes (LmxM.01, LmxM.02, LmxM.03, LmxM.04, LmxM.05, LmxM.06 and LmxM.12) and in a group of large chromosomes (LmxM.08, LmxM.20, LmxM.30, LmxM.31, LmxM.32, LmxM.33 and LmxM.34) relative to the annotations for the whole genome. Hypergeometric distribution analysis was performed for each domain annotation ID in each sample group to determine if any domains appear to be enriched in the small chromosome group or in the large chromosome group. The output from this is shown in table 5.1. A domain associated with motility, PF07004, is enriched in the small chromosome group ($q=3.11x10^{-19}$) and contains a PGP motif that is repeated in eukaryotic sperm tail proteins. A second domain, PF00501, encoding an AMP-binding enzyme is also slightly enriched (q=0.03766) in this group suggesting a potential role in metabolic or catalytic activity. Of the 77 amastin surface glycoprotein domains (PF07344) annotated in the genome, 55 are found in the large chromosome group ($q=2.16x10^{-12}$) and 15 of these are located on LmxM.30. A domain of unknown function, PF09149, is also enriched in the large chromosome group ($q=9.9x10^{-8}$), while a zinc knuckle-zinc binding motif,

PF00098, involved in eukaryotic gene regulation is significantly underrepresented (q=0.032997). Testing for enrichment on LmxM.30 alone highlights two further domains, PF01490 ($q=4.53x10^{-7}$) and PF00664 (q=0.00348), both associated with transmembrane transport. This result may suggest that changes in copy number of this chromosome are associated with the regulation of genes involved in modulating drug resistance, although follow-up experimental study and validation of these observations would be required.

A similar analysis was performed using GO terms to test further the protein domain analysis and assess overlap of significantly enriched functions in each approach (Table 5.1). In this analysis, small chromosomes displayed an enrichment of FMN- and copper-ion binding proteins, which was not identified when utilising Pfam domain annotations. In contrast to this, the functions enriched on large chromosomes were highly similar to those from the previous approach, with transmembrane transport being the most significant. CRAL/TRIO lipid-binding domain containing proteins also appeared as enriched on large chromosomes in both approaches and this may be of interest as this structural domain was found to be elevated in the stumpy form of *T. brucei* by Capewell et al., 2013 (Capewell et al., 2013).Enriched GO-terms on LmxM.30 were associated with membrane transport and multi-drug resistance genes.

An unresolved concern with this current analysis of enriched Pfam domains is that one gene may contain multiple copies of the same domain and this will lead to bias in the results. To resolve this issue, the analysis would need to be repeated, minimising the unique domain counts to 1 annotation per gene.

5.2.4.1 Visualisation of amastin surface glycoprotein domain distribution in *L. mexicana* U1103

Amastin surface glycoproteins are present in large gene arrays in *L. major* and *L. mexicana* and are involved in immune evasion. This analysis highlights an enrichment of these arrays on large chromosomes and therefore it was of interest to visualise the distribution of these domains throughout the *L. mexicana* genome to ask if this observation is a result of the previously described potential bias in this approach. We wanted to investigate the distribution of these genes to see whether they were present in large clusters,

evenly distributed across large chromosomes or mainly located on only one or two large chromosomes. The regions containing amastin domains were visualised using Circos and GP63 domains were also included as these proteins are also associated with immune evasion. The visualisation of the distribution of genes containing these domains is shown in figure 5.7. From this Circos plot, we can see that clusters of amastins are predominantly found on large chromosomes but amastin domain containing genes are also present on chromosomes 10, 16 and 24 which are intermediate in size and not included in either the small or large chromosome groups. Chromosome 33 contains a distribution of amastin domain containing genes on one half of the chromosome where as a concentrated cluster of domains is observed in chromosome 8. The cluster observed on chromosome 8 may be due to multiple copies of the domain in a single gene as described above in 5.2.4. Based on this analysis, it appears that genes containing GP63 and amastin domains are more likely to be found on large chromosomes and are dispersed across several chromosomes.

	Pfam/GO-term ID	Function	q-value	Fisher's Exact p-value
Chromosomes 1-6 + 12	PF07004	sperm-tail PG rich repeat	3.11E-19	-
	GO:0010181	FMN binding	-	0.00034
	GO:0005507	copper ion binding	-	0.0096
Chromosomes 8, 20, 30-34	PF07344	amastin surface glycoprotein	2.16E-12	-
	PF09149	unknown function	9.90E-08	-
	GO:0008408	3'-5' exonuclease activity	-	0.0018
	GO:0004576	oligosaccharyl transferase activity	-	0.0064
	GO:0042626	ATPase activity, coupled to transmembrane	-	0.0064
	GO:0004784	superoxide dismutase activity	-	0.0083
Chromosome 30	PF00648	calpain family cysteine protease	2.57E-07	-
	PF01490	transmembrane amino acid transporter	4.53E-07	-
	PF07344	amastin surface glycoprotein	0.0002857	-
	PF09149	unknown function	0.00148	-
	PF00664	ABC transporter transmembrane region	0.00348	-

Table 5.1. Enriched Pfam and GO term IDs in samples of chromosome subsets in *L. mexicana* M379 identified based on length and level of fold change during serial passage.

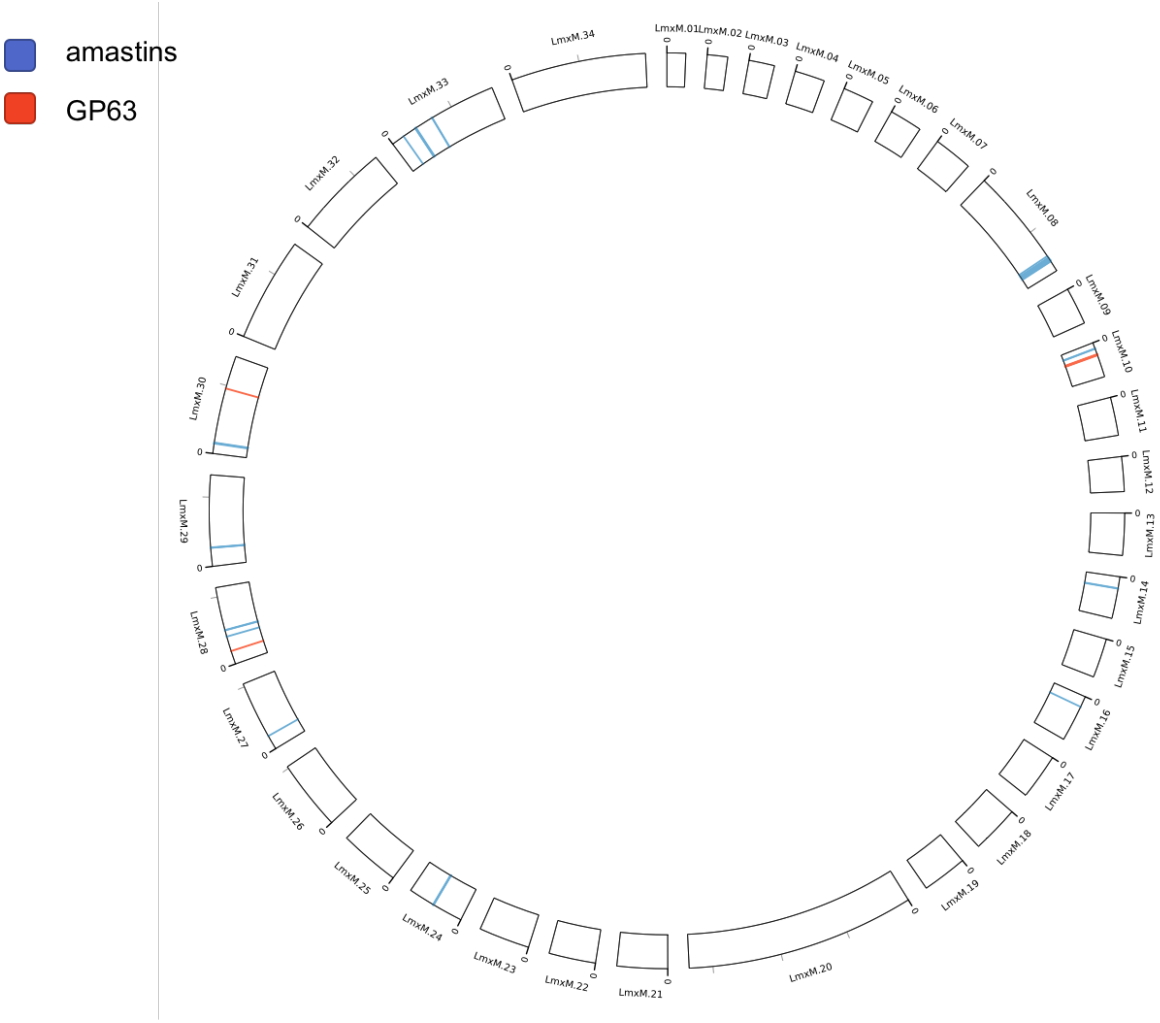


Figure 5.7. Amastin and GP63 gene distribution visualised using Circos.

Locations of amastin associated genes are highlighted in blue while GP63 genes are shown in red.

5.2.5 Analysis of changes in chromosome ploidy in an additional dataset

The observation of size-dependent changes in chromosome ploidy is based on the analysis of a single dataset and it was therefore of value to test the validity of this observation in an independent set of data before planning further lab experiments. Aspects of this analysis were performed by Jade Bartolo as part of a bioinformatics MSc Medical Genetics project. DNA sequence data is available for the progenitor sample for the *L. mexicana* M379 line cultured by the research group of Professor Michael Barrett in the Wellcome Centre for Molecular Parasitology, and passage of parasites from this sample have been performed by Dr Andrew Pountain, a former PhD student in this group. DNA sequence data was obtained from a *L. mexicana* sample that had undergone ~10 passages as a control sample in a drug resistance study. Repeating the previous genomic analysis using the PReP pipeline allowed a similar comparison of chromosome fold change between the progenitor sample, treated as a zero passage sample,

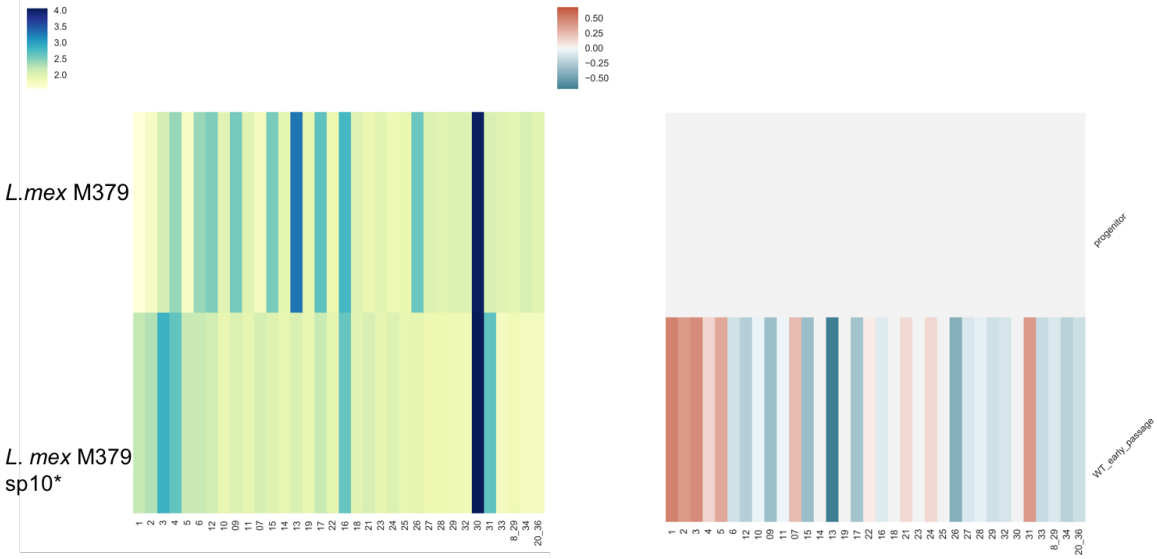


Figure 5.8. Chromosome copy number and relative fold change in a second independent *L. mexicana* dataset.

Left: Raw chromosome copy number estimates in a *L. mexicana* M379 progenitor sample (above) and a sample from the same line that has undergone a short passage of roughly 10 (below). Right: Fold change of chromosome copy number relative to the progenitor sample. In both heat maps, chromosomes (x-axis) are ordered by size.

and the passage 10 sample. The results from this brief comparison demonstrate a similar pattern of size-dependent changes in chromosome ploidy, although to a lesser extent with the trend being visible only in the smallest and largest chromosomes (figure 5.8).

5.3 Assessing the relationship between DNA replication and changes in chromosome ploidy

To further investigate the potential relationship between changes in chromosome ploidy and DNA replication, a second experiment was designed involving the addition of hydroxyurea (HU). The introduction of HU depletes the dNTP pool and therefore impedes the rate of DNA replication (Bianchis et al., 1986). Stalled replication forks can also occur ,which can collapse and cause DNA breaks. It was hypothesised that inhibition of DNA replication with HU would enhance the previously observed changes in chromosome ploidy, allowing the changes to be detectable over a shorter period of passaging, if replication limitations are the basis for ploidy variation.

5.3.1 Addition of HU during passage of *L. mexicana* promastigotes

The culture and passage of *L. mexicana* M379 promastigotes was performed by Dr Jennifer Stortz, a post-doctoral researcher in Dr Richard McCulloch's group. The experimental set up is outlined in figure 5.9. An initial step was identifying a concentration of hydroxyurea (HU) that could be used to treat the cells and deplete dNTPs to a level which impedes DNA replication but allowed the cells to continue growing. Growth curve analysis indicated that concentration above 0.2mM caused cells to die rapidly. Cells were therefore treated with 0.1 mM or 0.2 mM HU during each passage, and a third control sample was included with no HU treatment. The experiment was carried out in triplicate and samples underwent 7 passages in total. DNA was extracted at passage 1 and passage 7. FACS data was also generated at passage 1, 6 and 7 for the 0 mM, 0.1mM and 0.2mM samples (figure 5.10). The FACs analysis highlighted an increase in S phase cells in the passage 6 and 7 samples with 0.2mM HU treatment relative to the 0 mM control samples (figure 5.10). This is likely due to cells not completing

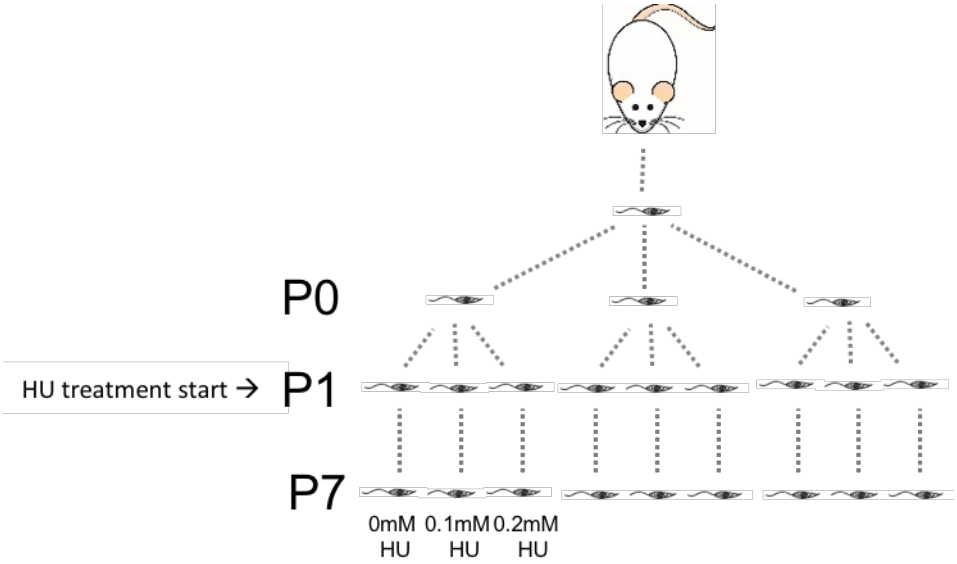


Figure 5.9. Overview of *L. mexicana* serial passage with HU treatment experimental set up.

Outline of experimental set up and triplicate serial passage samples that provide DNA sequence data for analysis at passage 1 and passage 7. Within each replicate, one sample has been treated with 0.1 mM HU, a second with 0.2 mM HU and a third control sample with no HU added.

or taking longer to complete DNA replication and therefore spending longer in S phase.

DNA sequence data was prepared as previously outlined in 5.2.1, and genomic analysis was performed by running PReP. Visualisation of raw chromosome copy number did not reveal any significant changes with HU treatment (figure 5.11A), although chromosomes 3, 4, 16 and 31 were consistently triploid in this dataset and chromosome 30 was at least tetraploid. When fold change of the passage 7 samples was visualised relative to the initial passage 1 sample, a size-dependent trend in ploidy variation was observed in cells treated with 0.2 mM HU. This pattern was stronger still when comparing the chromosome copy number fold change in passage 7 0.2 mM HU treated cells relative to the passage 7 0 mM HU control sample(figure 5.11B (left)). Although the relative fold change is limited to a small scale, plotting the median fold change against chromosome size as in the previous analysis, revealed a significant difference between the small and

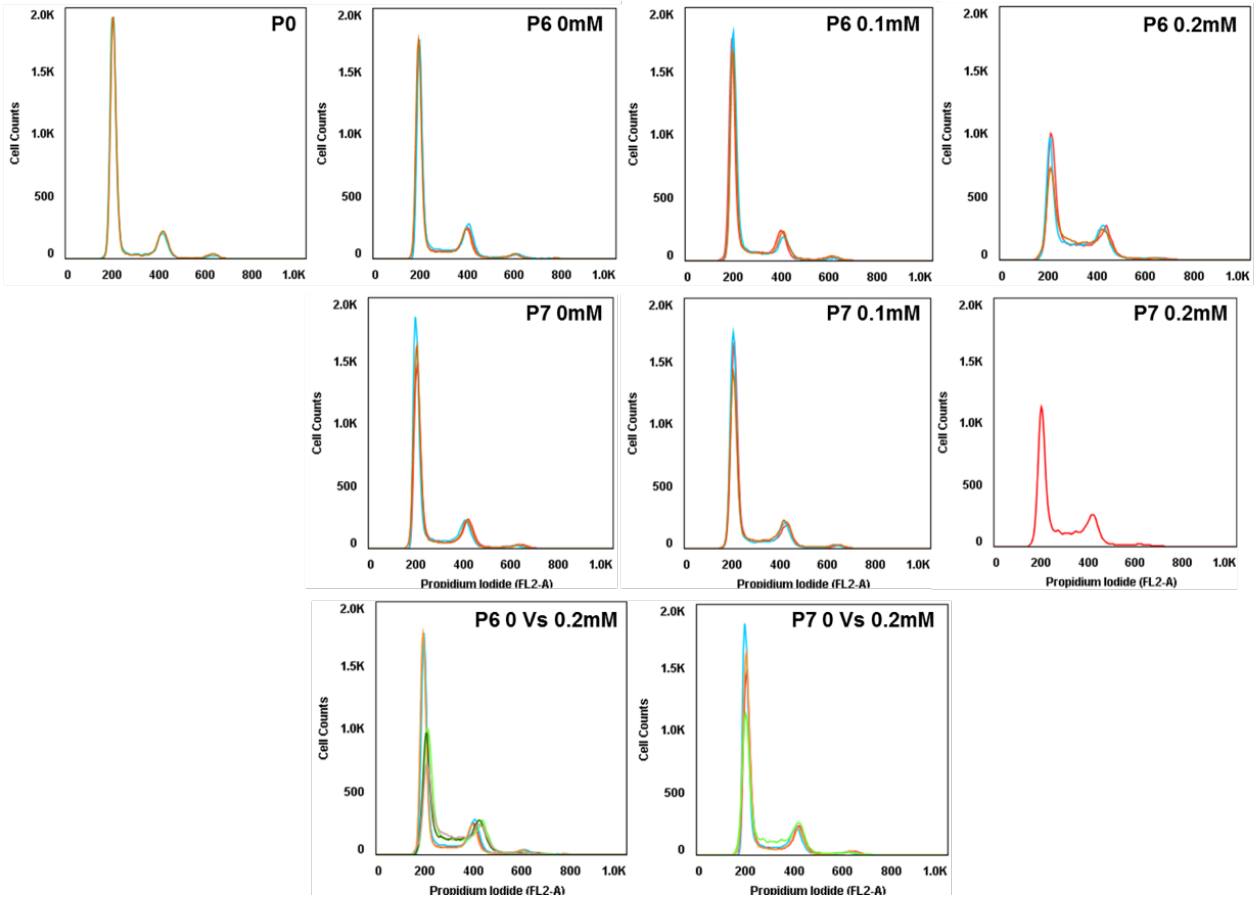


Figure 8.10. FACS profiles for triplicate untreated and 0.1mM and 0.2mM HU treated samples at passage 0, passage 6 and passage 7.

Colours represent overlaid replicate samples. Plots of 0mM vs 0.2mM HU treated samples at passage 6 and passage 7 are also included in the lower panels (0mM and 0.2mM profiles overlaid).

large chromosome groups (figure 5.11B(right)). This observation is consistent in that small chromosomes increase in copy number and large chromosomes decrease in copy number, this time under a much shorter serial passage experiment (7 passages relative to 29). These results support our previous observations in that the changes in chromosome copy number fold change are associated with the process of DNA replication.

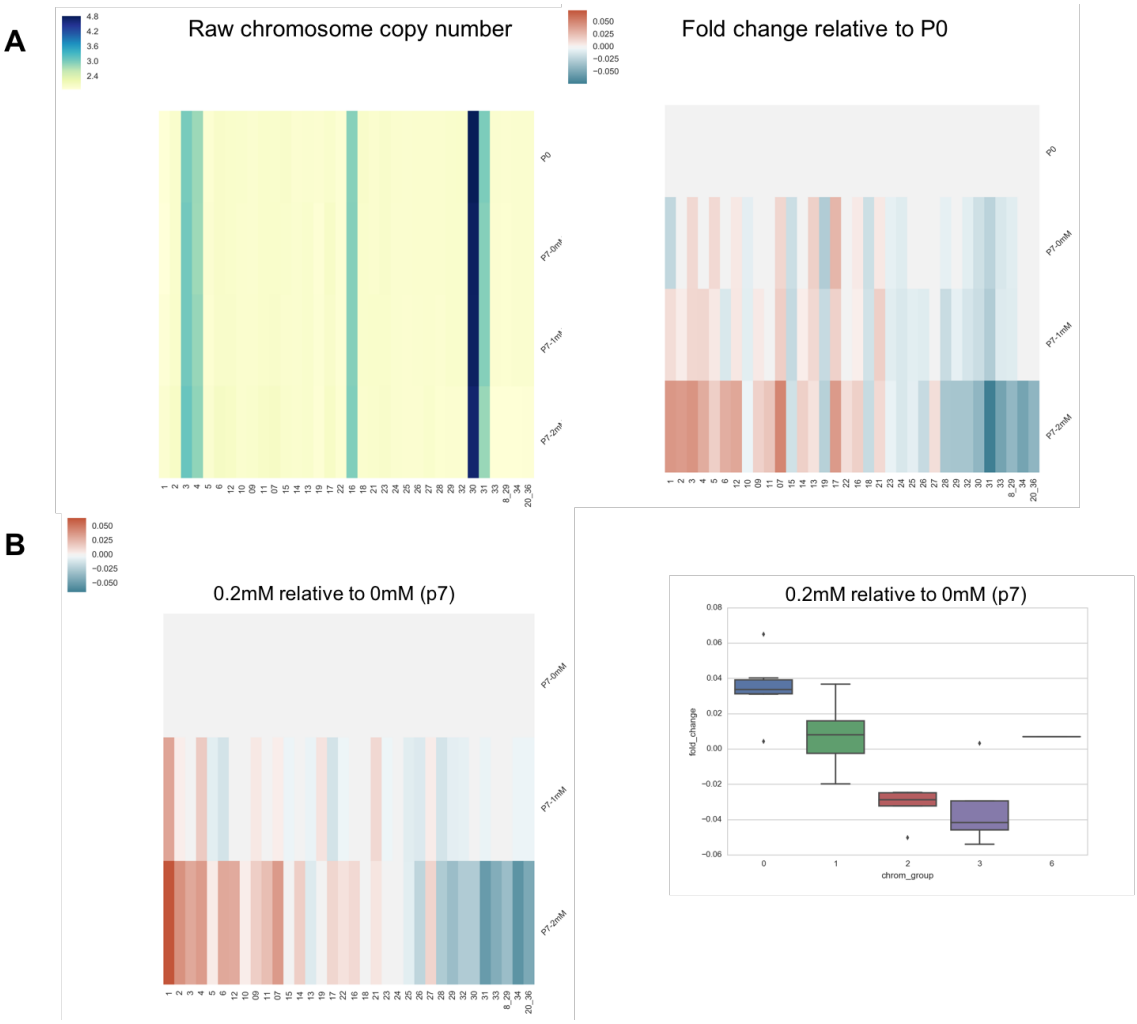


Figure 5.11. Chromosome copy number and relative fold change in a *L. mexicana* short passage dataset with HU treatment

A. Raw chromosome copy number estimates in a *L. mexicana* M379 dataset of 7 serial passages with HU treatment (left). Fold change of chromosome copy number relative to the untreated P0r sample (right). In both heat maps, chromosomes (x-axis) are ordered by length B. Relative fold change of passage 7 0.2 mM HU treated sample to the passage 7 0 mM HU sample (left). Chromosomes are ordered by length. Boxplot visualisation of the relative changes in chromosome ploidy against chromosome size (right).

5.4 Comparison of DNA replication and changes in chromosome ploidy

It is possible that the serial passage of *L. mexicana* M379 promastigotes artificially selects for fast growing cells, and therefore potentially exposes the limitations of replication from a single origin. Treating the samples with HU enhances this limitation and it is therefore possible to observe the same changes in chromosome ploidy over a much short timescale. The decrease in copy number observed for large chromosomes correlates with this hypothesis, while the possible re-replication of small chromosomes may indicate that *Leishmania* cells are less efficient at detecting if an origin has fired. *Leishmania* have a 2.9 hour S phase as measured by (Wheeler et al., 2011), and a replication fork rate of 2.4-2.6 kb/minute as measured by (Stanojcic et al., 2016). It is estimated that within this S phase, it would be possible to replicate ~870 kbp from a bi-directional single origin. This is sufficient to replicate many of the chromosomes in *L. major* and *L. mexicana* (24 out of 36 and 23 out of 34, respectively). However, in the case of larger chromosomes that are >1 Mb in length, a single origin would not be sufficient to replicate the entire chromosome. For example, the longest chromosome in the *L. mexicana* genome, LmxM.20, which is ~3.3Mb in length, would require ~8 hours to complete replication at this rate from a single origin. As the doubling time of *Leishmania* is around 6 hours, the replication of this chromosome at the measured rates is unfeasible. A model of this is illustrated in figure 5.12.

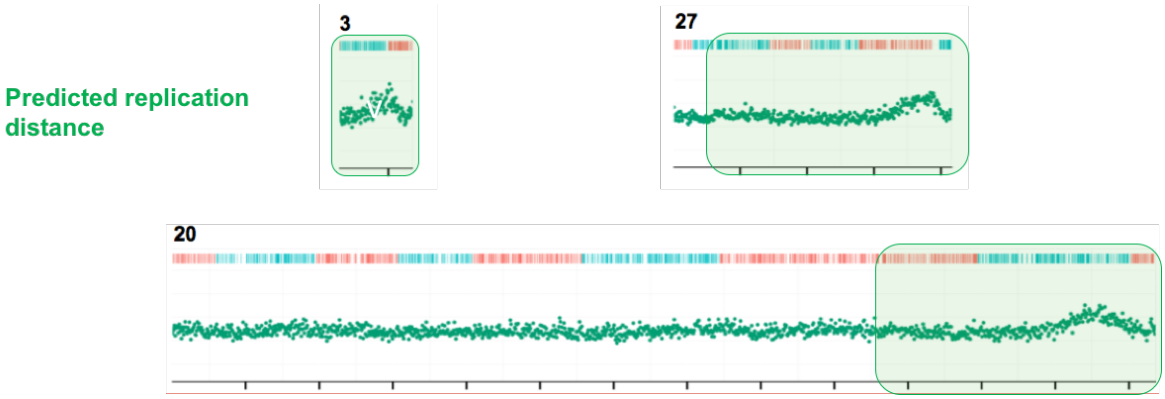


Figure 5.12. Model of predicted DNA replication distance from a single dominant origin in *L. mexicana*.

Predicted replication distance from a single origin is illustrated for chromosome 3 (~300kb), chromosome 27 (~1 Mb) and chromosome 20 (~3.3 Mb) in *L. mexicana*.

The above prediction indicates that replication in *Leishmania* must be initiated at multiple sites on each chromosome during S phase to complete replication of large chromosomes within the allocated time. However, it is plausible, based on this analysis, that the observed changes in chromosome ploidy are the consequence of replication from a single dominant origin per chromosome. In this case, a secondary mechanism could be present to complete replication of large chromosomes, which may be ORC-independent and potentially relies on homologous recombination and proteins associated with DNA repair pathways. This is a novel hypothesis that proposes an underlying mechanism for the relationship between DNA replication and mosaic aneuploidy in *Leishmania,* which is currently not well understood.

Alternatively, it is possible that across a population of *Leishmania* cells, several minor origins are used at a low frequency to complete replication of each chromosome, which cannot be detected due to limitations of the MFAseq approach. In *T. brucei* the major early-firing origin on each chromosome co-localises with the centromere, therefore the major origin we observe in *Leishmania* could potentially co-localise with the chromosome centromere which have recently been predicted through ChIP-seq enrichment analysis of LmKKT1 (Sollelis et al., 2017; Tiengwe et al., 2012).

The size-dependent changes in chromosome ploidy assessed in this analysis are observed at the level of a whole population of promastigote cells and the changes detected at this level are small. More sensitive investigation techniques would be required to detect these changes at the level of individual cells and provide insight into the proportion of cells undergoing changes in chromosome copy number.

5.5 Characterisation of sequence features present in multi-copy genes in *L. major* and *L. mexicana*

As previously described, genomic adaptation through variation in gene and whole chromosome copy number is well documented in *Leishmania* (Downing et al., 2011; Rogers et al., 2011). The mechanisms underlying this process are not yet

clear and little is known about any differences that may exist between single and multi-copy genes at the DNA sequence level. The above investigation focused on understanding the potentially conserved regulatory sequences at SSRs, and below the approach of binary sequence classification is extended to the classification of multi copy genes. This may allow the identification of sequence features enriched within multi copy genes that may have functional significance, or allow future prediction of multi copy genes.

5.5.1 Defining single and multi-copy gene datasets

The available DNA sequence datasets for the genomes of *L. major* Friedlin, *L. mexicana* U1103 and *L. mexicana* M379 were used to investigate characteristic sequence features in multi-copy genes compared to single copy genes using a SVM approach similar to the one used previously to characterise origins of replication in *Leishmania*. Classes of single and multi copy genes were defined in

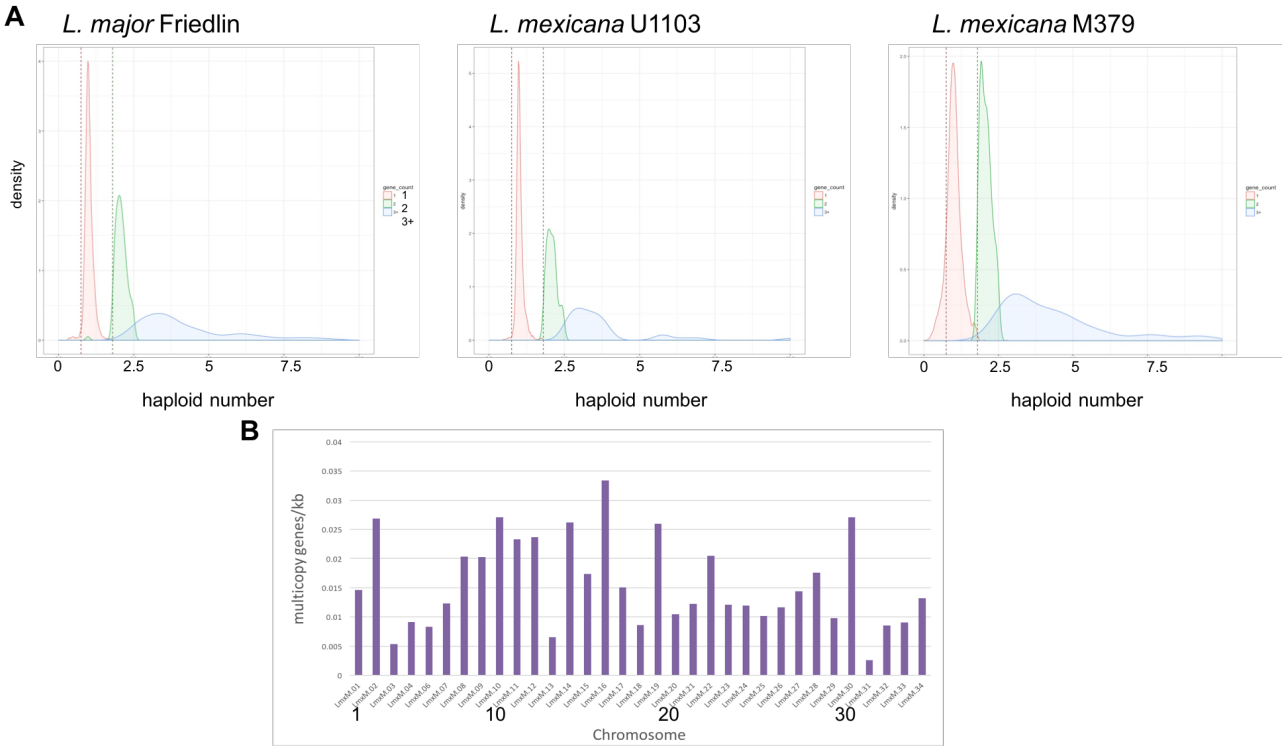


Figure 5.13. Visualisation of estimated gene haploid frequencies.

A. Gene haploid count frequency plots for L. major, L. mexicana U1103 and M379. Genes are grouped by orthomcl ID where a single gene is shown in red, 2 genes in green and 3+ in blue. Thresholds are represented by dotted lines at 0.75 and 1.8. B. Number of genes determined as multi-copy per kb in each chromosome of *L. mexicana* U1103 (plot shown in B. generated by Dr Nicholas Dickens).

the genomes of *L. mexicana* U1103, *L. mexicana* M379 and *L. major* Friedlin based on a gene haploid count threshold of 1.8 for single copy genes and a minimum and maximum cut-off of 0.75 and 10 respectively (figure 5.13A). This analysis allowed the generation of a formatted file containing gene IDs, haploid number and labelled as single or multi-copy, providing the labels required to train and test the SVM classifier. Additionally, plotting the number of multi copy genes per chromosome in *L mexicana* U1103 reveals that chromosome 16 has the highest number of multi copy genes in this genome per kb (figure 5.13B).

5.5.2 Application of machine learning to characterise multi copy genes

The DNA sequence reads that map to the coordinates of the single and multi-copy gene lists were again taken as samples and broken up into *k*-mers, this time with a length of 7 bases. A SVM was trained using a linear kernel on the *k*-mer features from *L. mexicana* U1103 and tested in *L. mexicana* M379 and *L. major* Friedlin. The classifier had high self-accuracy but this is expected, as a cross validation step has not yet been implemented during training. High prediction accuracy was maintained when testing in a different strain, *L. mexicana* M379, and across species to *L. major* Friedlin (Table 5.2). A second classifier was then trained on *L. major* Friedlin *k*-mer features and tested on *L. mexicana* U1103 to perform the reverse analysis and investigate whether similar sequence features were used by the machine in each genome to make the classifications. The second classifier also performed well and therefore provides confidence that this approach accurately distinguishes single and multi-copy genes based on DNA sequence features.

As part of this initial analysis, feature selection was included to detail the top 10 *k*-mer features used by the SVM to classify the single/multi-copy gene regions. This step was generated by Dr Nicholas Dickens. The feature lists for the *L. mexicana* U1103 and *L. major* Friedlin-trained classifiers are shown in table 5.3. Visualisation of this data on an individual chromosome was not effective and further refinement of this step would be required to reduce noise and ensure clarity of the results.

	Trained on *L. mexicana* U1103	Trained on *L. major* Friedlin
L. mexicana M379	*L. major* Friedlin	*L. mexicana* U1103
98.97%	94.28%	91.52%

Table 5.2. SVM classifier accuracy in *L. major* and *L.mexicana*.

Although, there is still much work to be done in improving the machine learning approach to ensure accurate results that reflect the datasets, the high classifier accuracy obtained during initial attempts to define single and multi copy genes indicates the presence of sequence features specific to multi-copy genes. Given additional time, it would be important to optimise and repeat this process to ensure the appropriate features are being detected and extracted before further analysis. Improvement of the general SVM approach and the feature selection process would allow correlation between the appearance of relevant sequence features with the location of annotated multi-copy genes. Implementation of a cross-validation step during training of the classifier would be a priority, as this

L. mexicana U1103	*L. major* Friedlin
gtgaccg	cggccac
gaatgtg	tgttctc
agaacat	cctcttc
ggtaccg	gacatga
gcagtct	agtttgt
cggcagc	aagctca
cggtacc	gtacttg
ccttgga	gacgctc
gtcgcgg	agaggta
cttggag	ccgccgc

Table 5.3. Extracted list of top 10 *k*-mer features used by the SVM algorithm to perform classification

can limit potential overfitting of the model to the training data and therefore improve overall accuracy. The datasets used in this analysis are relatively small and although this is alleviated to some degree by the use of DNA sequence reads as samples, it would likely only be possible to perform a 5-fold cross-validation before the sample sizes become too small. The consideration of alternative kernels may also be beneficial, primarily the non-linear RBF kernel, which is frequently used with SVMs. Also, the current assessment of classifier accuracy could be improved as the current method is not sufficient. Common accuracy detection methods that consider false positive and true positive rate would be implemented to improve this method, such as assessing the area under the curve (AUC) of the receiver operating curve (AUC-ROC) and potentially the precision-recall curve (PR-ROC).

5.6 Conclusions

Exploration of the changes that occur in the sequence and structure of the *L. mexicana* genome in response to serial passage conditions has revealed changes in chromosome ploidy that are size-dependent and emphasised by continued parasite growth. Little variation is observed at the sequence level over this time and further investigation of structural variation throughout passage could identify break point regions, particularly on the potentially under-replicated large chromosomes, that may be prone to breakage or re-arrangement. The observations of genome plasticity in this analysis may be a consequence of DNA replication from a single dominant origin and selection of fast-growing cells. In this case, replication may be completed by flexible or dormant origins, which we have not been able to detect by MFAseq, or by currently unknown recombination based methods, which initiates and completes replication independent of origins and ORC initiation factors.

Attempts to detect changes in copy number of small and large chromosomes by qPCR were unsuccessful. This is likely due to the small level of fold change detectable in the overall population and a more sensitive molecular technique would be required to validate these. Addition of HU to cells over a second shorter passage produced similar changes in chromosome ploidy, although to a

lesser extent. The trend was most pronounced in passage 7 cells treated with 0.2mM HU, in which replication was most impeded by the depletion of dNTPs. FACS analysis indicates an increase of S phase cells in 0.2mM HU treated samples at passage 6 and passage 7 relative to untreated cells at the same passage, providing further support that the observed changes in chromosome copy number are associated with DNA replication.

The enrichment analysis performed in this study found that small chromosomes contained a larger number of genes associated with motility, catalytic and metabolic activity and an increase in DNA read coverage was observed across these chromosomes during growth. However, this analysis does not show that this increased coverage translates to increased gene expression. Large chromosomes that displayed an overall decrease in DNA read coverage with growth displayed an enrichment of amastin surface glycoprotein genes involved in host immune evasion and also multi-drug resistance genes. The presence of drug-resistance genes was markedly enriched on chromosome 30, a large chromosome which is consistently present in a ploidy state greater than diploid across all species of *Leishmania* characterised thus far. Although RNA was only extracted from the shorter passage HU-treated cells and not from the initial serial passage samples, RNA-seq analysis would allow correlation between changes in gene expression and the observed changes in chromosome ploidy. This would allow the potential correlation of expression levels of the aforementioned genes of specific function on the chromosomes displaying overall changes in ploidy.

Brief investigation of DNA sequence at genes prone to amplification and copy number variation revealed the presence of sequence features specific to multi-copy genes. Application of machine learning successfully classifies multi-copy genes in *L. major* and *L. mexicana*, although optimisation of this approach is required for further elucidation of the relevant sequence features.

6 Concluding Remarks and Future Perspectives

The use of whole genome computational methods, still novel to kinetoplastid biology, reveal striking differences between the genomes of *Leishmania* and *Trypanosoma* despite their evolutionary relatedness. Further analysis of kinetoplastid genomes using molecular and computational techniques, highlights divergence between the *L. major* and *T. brucei* genomes and differences in essential processes such as DNA replication become apparent.

The replication origins in *L. major* have been mapped using MFAseq, and subsequent analysis of a second species, *L. mexicana,* supports the hypothesis of DNA replication predominantly occurring from a single origin on each chromosome (Marques et al, 2015). This mechanism of replication is also seen in bacterial and archaeal genomes (Robinson & Bell, 2005). The mapped origins demonstrate significant conservation in location with those mapped in the *T. brucei* genome, although no sequence homology has been observed (Marques et al, 2015). The *L. mexicana* genome is made up of 34 chromosomes compared to 36 in the *L. major* genome. This is due to the presence of two putative chromosome fusions occurring between chromosomes 8 and 29 and chromosomes 20 and 36 (Rogers et al, 2011). Origin loss occurs as a consequence of this process and the 'fused' chromosomes appear to replicate from a single origin. In contrast with the *T. brucei* genome, the origins in *Leishmania* are used with equal efficiency. Having multiple origins is thought to allow redundancy, a mechanism that allows the recovery of the bi-directional replication process if a problem occurs and replication cannot continue. It is unknown how species of *Leishmania* recover DNA replication in these circumstances, although recombination may be a plausible mechanism.

The MFAseq pipeline could be further improved by the inclusion of a purpose-built script that would analyse the raw MFAseq ratio data, generate a significance threshold and computationally define peak regions in a similar way to existing software designed for ChIP-seq analyses. The investigation of algorithms designed for analysis of ChIP-seq data found that they are not appropriate in application to the MFAseq data. A limitation of the MFAseq approach is that it relies on sorting cells in S and G2 phases, meaning any unscheduled replication, such as in G1 or M phase, would not be seen. Indeed, it is possible that such unscheduled replication may explain the dichotomy

between MFAseq replication mapping and the DNA combing and SNSseq approaches described later (Lombrana et al., 2016; Stanojcic et al., 2016)

A support vector machine algorithm was applied to the SSRs in *L. major* and *L. mexicana* to better characterize origins of replication at the sequence level. Although this approach was highly successful in origin/non-origin sequence classification, optimization of the approach is required. Working to resolve issues such as updating poor reference genome annotations and reduction of computational time involved in data pre-processing steps will strongly improve the speed and accuracy of the classifier and allow extraction of meaningful features that can be used to characterize replication origins and start to generate a more robust model of replication initiation, which includes sequence level detail. Further optimization of the support vector machine parameters would allow classification at a greater accuracy with identified false discovery rates and optimized parameter and kernel choices. It would be of interest to apply this machine learning approach to the *T. brucei* genome and further investigate the conservation of replication origins between kinetoplastid species by also applying the *Leishmania* trained classifier to *T. brucei*. To date, origins have not been mapped in other kinetoplastid genomes, but if the approach can be shown to traverse the evolutionary gap between *Leishmania* and *T. brucei*, it may be a way to identify origins in further species, without laborious wet lab experiments. If the classifier is able to predict origins across species, it would likely provide insight into the level of conservation of sequence features associated with DNA replication. Alternatively, if the classifier performs well in *Leishmania* but is in fact characterising centromeric sequences as opposed to replication origins, then a difference in performance may be observed in *T. brucei,* which contains multiple origins per chromosome, not all of which colocalise with centromeres.

The results described in this analysis indicate that features exist within these regions that can be used to accurately separate the two different types of SSR (origin-active and inactive) and potentially characterise DNA replication origins at the sequence level. A limitation of the machine learning technique is that the quality of the predictions made by the algorithm are dependent on the quality of the input data and how well defined the region of interest is. This analysis could therefore be improved by providing more precise region coordinates to reduce

the amount of sequence used to perform the classification. Further investigation will provide insight into the sequence and structure of SSRs, the initiation of replication at the DNA sequence level, and potentially the relationship between replication and transcription in kinetoplastid genomes.

Integration of several datasets from *L. mexicana* in serial passage has revealed the presence of size-dependent changes in chromosome ploidy that provide insight into the potential relationship between genome plasticity and DNA replication. The Hi-C approach which was recently applied to *T. brucei,* is also used to identify chromosomal rearrangements in tumour cells and assess copy number variation (Harewood et al, 2017; Siegel et al, 2018). Potential further experiments in *Leishmania* could include this type of analysis which would give insight into general chromosome organisation and also identify potential structural rearrangements that occur during serial passage. In addition to this, protocols to assess nucleosome occupancy have been effective in *Leishmania,* such as MNase-seq which could be repeated at a high resolution Lombrana et al, 2016). This analysis could be extended to techniques such as ATAC-seq, which assesses genome-wide chromatin accessibility at a high resolution and can be used to infer regions of increased accessibility (Buenrostro et al., 2015). If possible, it would also be of interest to repeat the analysis in the sandfly vector to ask if the observations are caused by serial passage in culture or reflect true *in vivo* phenomena. Further investigation, both computationally and in the wet lab, may increase our understanding of aneuploidy and the relationship with replication as we work to understand if it is possible for these single cell parasites to replicate each chromosome from a single origin.

List of References

A, A., & J, R. (2018). *topGO: Enrichment Analysis for Gene Ontology*. *R package version 2.34.0*.

ADL, S. M., SIMPSON, A. G. B., LANE, C. E., LUKES, J., BASS, D., BOWSER, S. S., … SPIEGEL, F. W. (2012). The revised classificatin of eukaryotes. *The Journal of Eukaryotic Microbiology*, *59*(5), 429–493. https://doi.org/10.1111/j.1550-7408.2012.00644.x

Alvar, J., Vélez, I. D., Bern, C., Herrero, M., Desjeux, P., Cano, J., … de Boer, M. (2012). Leishmaniasis worldwide and global estimates of its incidence. *PLoS ONE*, Vol. 7. https://doi.org/10.1371/journal.pone.0035671

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166–169. https://doi.org/10.1093/bioinformatics/btu638

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2000). Gene Ontology : tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556.Gene

Aslett, M., Aurrecoechea, C., Berriman, M., Brestelli, J., Brunk, B. P., Carrington, M., … Wang, H. (2010). *TriTrypDB : a functional genomic resource for the Trypanosomatidae*. *38*(October 2009), 457–462. https://doi.org/10.1093/nar/gkp851

Ausiannikava, D., & Allers, T. (2017). Diversity of DNA Replication in the Archaea. *Genes*, *8*(2), 56. https://doi.org/10.3390/genes8020056

Bailey, T. L., Boden, M., Buske, F. a., Frith, M., Grant, C. E., Clementi, L., … Noble, W. S. (2009). MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Research*, *37*(SUPPL. 2), 202–208. https://doi.org/10.1093/nar/gkp335

Baldwin, J. M., Schertler, G. F. X., Unger, V. M., Mol, J., Myllykallio, H., Lopez, P., … Forterre, P. (2000). *Bacterial Mode of Replication with Eukaryotic-Like Machinery in a Hyperthermophilic Archaeon*. *288*(June), 2212–2216.

Becker, M., Aitcheson, N., Byles, E., Wickstead, B., Louis, E., & Rudenko, G. (2004).

Isolation of the repertoire of VSG expression site containing telomeres of Trypanosoma brucei 427 using transformation-associated recombination in yeast. *Genome Research*, *14*(11), 2319–2329. https://doi.org/10.1101/gr.2955304

Bell, S. P., & Stillman, B. (1992). ATP-dependent recognition of eukaryotic origins of DNA replication by a multiprotein complex. *Nature*, *357*, 128–134. https://doi.org/10.1038/357128a0

Benmerzouga, I., Concepción-Acevedo, J., Kim, H. S., Vandoros, A. V., Cross, G. A. M., Klingbeil, M. M., & Li, B. (2013). Trypanosoma brucei Orc1 is essential for nuclear DNA replication and affects both VSG silencing and VSG switching. *Molecular Microbiology*, *87*(1), 196–210. https://doi.org/10.1111/mmi.12093

Berriman, M., Ghedin, E., & Hertz-fowler, C. (2005). The genome of the African trypanosome, Trypanosoma brucei. *Science*, *309*, 416–422. https://doi.org/10.1126/science.1112642

Besnard, E., Babled, A., Lapasset, L., Milhavet, O., Parrinello, H., Dantec, C., … Lemaitre, J.-M. (2012). Unraveling cell type–specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. *Nature Structural & Molecular Biology*, *19*(8), 837–844. https://doi.org/10.1038/nsmb.2339

Bianchis, V., Pontis, E., & Reichard, P. (1986). Changes of Deoxyribonucleoside Triphosphate Pools Induced by Hydroxyurea and Their Relation to DNA Synthesis*. *The Journal of Biological Chemistry*, *261*(34), 16037–16042.

Bochman, M. L., & Schwacha, A. (2009). The Mcm complex: unwinding the mechanism of a replicative helicase. *Microbiology and Molecular Biology Reviews : MMBR*, *73*(4), 652–683. https://doi.org/10.1128/MMBR.00019-09

Briggs, E., Hamilton, G., Crouch, K., Lapsley, C., & Mcculloch, R. (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), 11789–11805. https://doi.org/10.1093/nar/gky928

Burgers, P. M. J., & Kunkel, T. A. (2017). Eukaryotic DNA Replication Fork. *Annual Review of Biochemistry*, *86*, 417–438. https://doi.org/10.1146/annurev-biochem-061516-044709

Burton, P., Mcbride, D. J., Wilkes, J. M., Barry, J. D., & Mcculloch, R. (2007). Ku Heterodimer-Independent End Joining in Trypanosoma brucei Cell Extracts Relies upon Sequence Microhomology □. *Eukaryotic Cell*, *6*(10), 1773–1781. https://doi.org/10.1128/EC.00212-07

Calderano, S. G., Drosopoulos, W. C., Quaresma, M. M., Marques, C. A., Kosiyatrakul, S., McCulloch, R., … Elias, M. C. (2015). Single molecule analysis of Trypanosoma brucei DNA replication dynamics. *Nucleic Acids Research*, *43*(5), 2655–2665. https://doi.org/10.1093/nar/gku1389

Capewell, P., Monk, S., Ivens, A., Macgregor, P., Fenn, K., Walrad, P., … Matthews, K. R. (2013). Regulation of Trypanosoma brucei Total and Polysomal mRNA during Development within Its Mammalian Host. *Plos One*, *8*(6). https://doi.org/10.1371/journal.pone.0067069

Cayrou, C., Coulombe, P., Puy, A., Rialle, S., Kaplan, N., Segal, E., & Méchali, M. (2012). New insights into replication origin characteristics in metazoans. *Cell Cycle*, *11*(February 2015), 658–667. https://doi.org/10.4161/cc.11.4.19097

Chang, C., & Lin, C. (2013). *LIBSVM : A Library for Support Vector Machines*. 1–39.

Cliffe, L. J., Siegel, T. N., Marshall, M., Cross, G. A. M., & Sabatini, R. (2010). Two thymidine hydroxylases differentially regulate the formation of glucosylated DNA at regions flanking polymerase II polycistronic transcription units throughout the genome of Trypanosoma brucei. *Nucleic Acids Research*, *38*(12), 3923–3935. https://doi.org/10.1093/nar/gkq146

Čuklina, J., Hahn, J., Imakaev, M., Omasits, U., Förstner, K. U., Ljubimov, N., … Evguenieva-Hackenberg, E. (2016). Genome-wide transcription start site mapping of Bradyrhizobium japonicum grown free-living or in symbiosis – a rich resource to identify new transcripts, proteins and to study gene regulation. *BMC Genomics*, *17*(1), 302. https://doi.org/10.1186/s12864-016-2602-9

De Melo Godoy, P. D., Nogueira-Junior, L. A., Paes, L. S., Cornejo, A., Martins, R. M., Silber, A. M., … Elias, M. C. (2009). Trypanosome prereplication machinery contains a single functional Orc1/Cdc6 protein, which is typical of archaea. *Eukaryotic Cell*, *8*(10), 1592–1603. https://doi.org/10.1128/EC.00161-09

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., … Daly,

M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. https://doi.org/10.1038/ng.806

Devlin, R., Marques, C. A., Paape, D., Prorocic, M., & Zurita-leal, A. C. (n.d.). *Mapping replication dynamics in Trypanosoma brucei reveals a link with telomere transcription and antigenic variation Samantha J . Campbell , Craig Lapsley , Nicholas Dickens and Richard McCulloch The Wellcome Trust Centre for Molecular Parasitology , Ins*. https://doi.org/10.7554/eLife.12765

Devlin, R., Marques, C. A., Paape, D., Prorocic, M., Zurita-leal, A. C., Campbell, S. J., … Mcculloch, R. (2016). Mapping replication dynamics in Trypanosoma brucei reveals a link with telomere transcription and antigenic variation. *ELife*, *5*(e12765), 1–30. https://doi.org/10.7554/eLife.12765

Ding, H., Liang, Z.-Y., Guo, F.-B., Huang, J., Chen, W., & Lin, H. (2016). Predicting bacteriophage proteins located in host cell with feature selection technique. *Computers in Biology and Medicine*, *71*, 156–161. https://doi.org/10.1016/j.compbiomed.2016.02.012

Downing, T., Imamura, H., Decuypere, S., Clark, T. G., Coombs, G. H., Cotton, J. a., … Berriman, M. (2011). *Whole genome sequencing of multiple* Leishmania donovani *clinical isolates provides insights into population structure and mechanisms of drug resistance*. 2143–2156. https://doi.org/10.1101/gr.123430.111

Dubessay, P. (2002). The switch region on Leishmania major chromosome 1 is not required for mitotic stability or gene expression, but appears to be essential. *Nucleic Acids Research*, *30*(17), 3692–3697. https://doi.org/10.1093/nar/gkf510

El-gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Potter, S. C., Qureshi, M., … Finn, R. D. (2019). *The Pfam protein families database in 2019*. (8), 1–6. https://doi.org/10.1093/nar/gky995

El-Sayed, N. M., Myler, P. J., Blandin, G., Berriman, M., Crabtree, J., Aggarwal, G., … Hall, N. (2005a). Comparative genomics of trypanosomatid parasitic protozoa. *Science (New York, N.Y.)*, *309*(5733), 404–409. https://doi.org/10.1126/science.1112181

El-Sayed, N. M., Myler, P. J., Blandin, G., Berriman, M., Crabtree, J., Aggarwal, G., …

Hall, N. (2005b). Comparative genomics of trypanosomatid parasitic protozoa. *Science (New York, N.Y.)*, *309*(2005), 404–409. https://doi.org/10.1126/science.1112181

Ettwiller, L., Paten, B., Ramialison, M., Birney, E., & Wittbrodt, J. (2007). Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. *Nature Methods*, *4*(7), 563–565. https://doi.org/10.1038/nmeth1061

Fejes, A. P., Robertson, G., Bilenky, M., Varhol, R., Bainbridge, M., & Jones, S. J. M. (2008). FindPeaks 3.1: A tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, *24*(15), 1729–1730. https://doi.org/10.1093/bioinformatics/btn305

Feng, J., Liu, T., Qin, B., Zhang, Y., & Liu, X. S. (2012). Identifying ChIP-seq enrichment using MACS. *Nature Protocols*, *7*(9), 1728–1740. https://doi.org/10.1038/nprot.2012.101

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *ArXiv Preprint ArXiv:1207.3907*, 9. https://doi.org/arXiv:1207.3907 [q-bio.GN]

Genest Paul-Andre, Bas ter Riet, Tony Cijsouw, H. G. A. M. van L. and P. B. (2007). Telomeric localization of the modified DNA base J in the genome of the protozoan parasite Leishmania. *Nucleic Acids Research*, *35*(7), 2116–2124. https://doi.org/10.1093/nar/gkm050

Genois, M.-M., Paquet, E. R., Laffitte, M.-C. N., Maity, R., Rodrigue, A., Ouellette, M., & Masson, J.-Y. (2014). DNA repair pathways in trypanosomatids: from DNA repair to drug resistance. *Microbiology and Molecular Biology Reviews : MMBR*, *78*, 40–73. https://doi.org/10.1128/MMBR.00045-13

Gent, D. C. Van, Hoeijmakers, J. H. J., & Kanaar, R. (2001). *CHROMOSOMAL STABILITY AND THE DNA DOUBLE-STRANDED BREAK CONNECTION*. *2*(March), 196–206.

Georgiou, G., & van Heeringen, S. J. (2016). Fluff: Exploratory Analysis and Visualization of High-Throughput Sequencing Data. *BioRxiv*, 045526. https://doi.org/10.1101/045526

Ghandi, M., Lee, D., Mohammad-Noori, M., & Beer, M. a. (2014). Enhanced Regulatory

Sequence Prediction Using Gapped k-mer Features. *PLoS Computational Biology*, *10*(7), e1003711. https://doi.org/10.1371/journal.pcbi.1003711

Ghandi, M., Mohammad-Noori, M., & Beer, M. a. (2014). Robust k -mer frequency estimation using gapped k -mers. In *Journal of Mathematical Biology* (Vol. 69). https://doi.org/10.1007/s00285-013-0705-3

Glover, L., Hutchinson, S., Alsford, S., Mcculloch, R., Field, M. C., & Horn, D. (2013). Antigenic variation in African trypanosomes: The importance of chromosomal and nuclear context in VSG expression control. *Cellular Microbiology*, *15*(12), 1984–1993. https://doi.org/10.1111/cmi.12215

Glover, L., Jun, J., & Horn, D. (2011). *Microhomology-mediated deletion and gene conversion in African trypanosomes*. *39*(4), 1372–1380. https://doi.org/10.1093/nar/gkq981

Gossage, S. M., Rogers, M. E., & Bates, P. A. (2003). Two separate growth phases during the development of Leishmania in sand flies : implications for understanding the life cycle. *International Journal for Parasitology*, *33*, 1027–1034. https://doi.org/10.1016/S0020-7519(03)00142-5

Grabherr Manfred G., Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W., N., & Friedman, and A. R. (2013). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883.Trinity

H, W. (2016). ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York*. https://doi.org/ISBN 978-3-319-24277-4

Haas, B. J., Delcher, A. L., Mount S.M., S. M., Wortman, J. R., Smith, R. K., Hannick, L. I., … White, O. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, *31*(19), 5654–5666. https://doi.org/10.1093/nar/gkg770

Hansen, R. S., Thomas, S., Sandstrom, R., Canfield, T. K., Thurman, R. E., Weaver, M., … Stamatoyannopoulos, J. A. (2010). Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National*

Academy of Sciences of the United States of America, *107*(1), 139–144. https://doi.org/10.1073/pnas.0912402107

Hertz-Fowler, C., Figueiredo, L. M., Quail, M. A., Becker, M., Jackson, A., Bason, N., … Berriman, M. (2008). Telomeric expression sites are highly conserved in Trypanosoma brucei. *PLoS ONE*, *3*(10). https://doi.org/10.1371/journal.pone.0003527

Iyer, L. M., Leipe, D. D., Koonin, E. V., & Aravind, L. (2004). Evolutionary history and higher order classification of AAA+ ATPases. *Journal of Structural Biology*, *146*, 11–31. https://doi.org/10.1016/j.jsb.2003.10.010

Jackson, A. P., Otto, T. D., Aslett, M., Armstrong, S. D., Bringaud, F., Schlacht, A., … Berriman, M. (2015). Kinetoplastid Phylogenomics Reveals the Evolutionary Innovations Associated with the Origins of Parasitism. *Current Biology*, (1). https://doi.org/10.1016/j.cub.2015.11.055

Kaye, P., & Scott, P. (2011). Leishmaniasis: Complexity at the host-pathogen interface. *Nature Reviews Microbiology*, *9*(8), 604–615. https://doi.org/10.1038/nrmicro2608

Kelly, S., Kramer, S., Schwede, A., Maini, P. K., Gull, K., & Carrington, M. (2012). Genome organization is a major component of gene expression control in response to stress and during the cell division cycle in trypanosomes. *Open Biology*, *2*(4), 120033. https://doi.org/10.1098/rsob.120033

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357–360. https://doi.org/10.1038/nmeth.3317

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Jones, S. J., & Marra, M. A. (2009). Circos : an Information Aesthetic for Comparative Genomics. *Genome Research*, *19*, 1639–1645.

Kumar, M., Gromiha, M. M., & Raghava, G. P. S. (2007). Identification of DNA-binding proteins using support vector machines and evolutionary profiles. *BMC Bioinformatics*, *8*, 463. https://doi.org/10.1186/1471-2105-8-463

Laajala, T. D., Raghav, S., Tuomela, S., Lahesmaa, R., Aittokallio, T., & Elo, L. L. (2009). A practical comparison of methods for detecting transcription factor

binding sites in ChIP-seq experiments. *BMC Genomics*, *10*, 618. https://doi.org/10.1186/1471-2164-10-618

Lachaud, L., Bourgeois, N., Kuk, N., Morelle, C., Crobu, L., Merlin, G., … Sterkers, Y. (2014). Constitutive mosaic aneuploidy is a unique genetic feature widespread in the Leishmania genus. *Microbes and Infection*, *16*, 61–66. https://doi.org/10.1016/j.micinf.2013.09.005

Laffitte, M.-C. N., Genois, M.-M., Mukherjee, A., Légaré, D., Masson, J.-Y., & Ouellette, M. (2014). Formation of Linear Amplicons with Inverted Duplications in Leishmania Requires the MRE11 Nuclease. *PLoS Genetics*, *10*(12), e1004805. https://doi.org/10.1371/journal.pgen.1004805

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Leprohon, P., Légaré, D., Raymond, F., Madore, É.´, Hardiman, G., Corbeil, J., & Ouellette, M. (2009). Gene expression modulation is associated with gene amplification, supernumerary chromosomes and chromosome loss in antimony-resistant Leishmania infantum. *Nucleic Acids Research*, *37*(5), 1387–1399. https://doi.org/10.1093/nar/gkn1069

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Liang, G., Lin, J. C. Y., Wei, V., Yoo, C., Cheng, J. C., Nguyen, C. T., … Jones, P. A. (2004). *Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome*. *4*.

Liang, X. H., Haritan, A., Uliel, S., & Michaeli, S. (2003). trans and cis splicing in trypanosomatids: Mechanism, factors, and regulation. *Eukaryotic Cell*, *2*(5), 830–840. https://doi.org/10.1128/EC.2.5.830-840.2003

Liu, B., Liu, Y., Motyka, S. A., Agbo, E. E. C., & Englund, P. T. (2005). Fellowship of the rings: the replication of kinetoplast DNA. *Trends in Parasitology*, *21*(8), 363–369. https://doi.org/https://doi.org/10.1016/j.pt.2005.06.008

Lombraña, R., Álvarez, A., Fernández-Justel, J. M., Almeida, R., Poza-Carrión, C.,

Gomes, F., … Gómez, M. (2016). Transcriptionally Driven DNA Replication Program of the Human Parasite Leishmania major. *Cell Reports*, 1774–1786. https://doi.org/10.1016/j.celrep.2016.07.007

Lundgren, M., Andersson, A., Chen, L., Nilsson, P., & Bernander, R. (2004). Three replication origins in Sulfolobus species: Synchronous initiation of chromosome replication and asynchronous termination. *Proceedings of the National Academy of Sciences*, *101*(18), 7046–7051. https://doi.org/10.1073/pnas.0400656101

MacAlpine, H. K., Gordân, R., Powell, S. K., Hartemink, A. J., & MacAlpine, D. M. (2010). Drosophila ORC localizes to open chromatin and marks sites of cohesin complex loading. *Genome Research*, *20*(2), 201–211. https://doi.org/10.1101/gr.097873.109

Maizels, N. (2006). Dynamic roles for G4 DNA in the biology of eukaryotic cells. *Nature Structural & Molecular Biology*, *13*(12), 1055–1059. https://doi.org/10.1038/nsmb1171

Maizels, N., & Gray, L. T. (2013). The G4 Genome. *PLoS Genetics*, *9*(4). https://doi.org/10.1371/journal.pgen.1003468

Marques, C. A., Dickens, N. J., Paape, D., Campbell, S. J., & McCulloch, R. (2015a). Genome-wide mapping reveals single-origin chromosome replication in Leishmania, a eukaryotic microbe. Supplementary methods. *Genome Biology*, *16*(1), 230. https://doi.org/10.1186/s13059-015-0788-9

Marques, C. A., Dickens, N. J., Paape, D., Campbell, S. J., & McCulloch, R. (2015b). Genome-wide mapping reveals single-origin chromosome replication in Leishmania, a eukaryotic microbe. *Genome Biology*, *16*(1), 230. https://doi.org/10.1186/s13059-015-0788-9

Martínez-Calvillo, S., Vizuet-De-Rueda, J. C., Florencio-Martínez, L. E., Manning-Cela, R. G., & Figueroa-Angulo, E. E. (2010). Gene expression in trypanosomatid parasites. *Journal of Biomedicine and Biotechnology*, *2010*. https://doi.org/10.1155/2010/525241

Masai, H., Matsumoto, S., & You, Z. (2010). *Eukaryotic Chromosome DNA Replication : Where , When , and How ?* https://doi.org/10.1146/annurev.biochem.052308.103205

Messer, W. (2002). *The bacterial replication initiator DnaA . DnaA and oriC , the bacterial mode to initiate DNA replication*. *26*.

Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., Bridge, A., … Finn, R. D. (2019). *InterPro in 2019 : improving coverage , classification and access to protein sequence annotations*. 1–10. https://doi.org/10.1093/nar/gky1100

Mukherjee, A., Boisvert, S., Monte-neto, R. L., Coelho, A. C., Raymond, F., Mukhopadhyay, R., … Ouellette, M. (2013). *Telomeric gene deletion and intrachromosomal amplification in antimony-resistant Leishmania*. *88*(March), 189–202. https://doi.org/10.1111/mmi.12178

Mulla, W., Zhu, J., & Li, R. (2013). Yeast: a simple model system to study complex phenomena of aneuploidy. *FEMS Microbiology Reviews*. https://doi.org/10.1111/1574-6976.12048

Muller, C. A., Hawkins, M., Retkute, R., Malla, S., Wilson, R., Blythe, M. J., … Nieduszynski, C. A. (2014). The dynamics of genome replication using deep sequencing. *Nucleic Acids Research*, *42*(1), 1–11. https://doi.org/10.1093/nar/gkt878

Murphy, T. D., & Karpen, G. H. (1998). Centromeres take flight: Alpha satellite and the quest for the human centromere. *Cell*, *93*(3), 317–320. https://doi.org/10.1016/S0092-8674(00)81158-7

Myler, P. J., Audleman, L., deVos, T., Hixson, G., Kiser, P., Lemley, C., … Stuart, K. (1999). Leishmania major Friedlin chromosome 1 has an unusual distribution of protein-coding genes. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(6), 2902–2906. https://doi.org/10.1073/pnas.96.6.2902

Nguyen, D., Stuart, K., & Myler, P. J. (2004). Transcription Initiation and Termination on Leishmania major Chromosome 3. *Society*, *3*(2), 506–517. https://doi.org/10.1128/EC.3.2.506

Nix, D. a, Courdy, S. J., & Boucher, K. M. (2008). Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. *BMC Bioinformatics*, *9*, 523. https://doi.org/10.1186/1471-2105-9-523

Norais, C., Hawkins, M., Hartman, A. L., Eisen, J. A., Myllykallio, H., & Allers, T. (2007). Genetic and physical mapping of DNA replication origins in Haloferax volcanii. *PLoS Genetics*, *3*(5), 729–743. https://doi.org/10.1371/journal.pgen.0030077

Otto, T. D., Dillon, G. P., Degrave, W. S., & Berriman, M. (2011). *RATT : Rapid Annotation Transfer Tool*. *44*(0), 1–7. https://doi.org/10.1093/nar/gkq1268

Parry, B. Y. E. M., & Cox, B. S. (1970). The tolerance of aneuploidy in yeast. *Genetic Research*, (16), 333–340.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Pelve, E. A., Martens-Habbena, W., Stahl, D. A., & Bernander, R. (2013). Mapping of active replication origins in vivo in thaum- and euryarchaeal replicons. *Molecular Microbiology*, *90*(3), 538–550. https://doi.org/10.1111/mmi.12382

Pepke, S., Wold, B., & Mortazavi, A. (2009). Computation for ChIP-seq and RNA-seq studies. *Nature Methods*, *6*(11 Suppl), S22–S32. https://doi.org/10.1038/nmeth.1371

Pfau, S. J., & Amon, A. (2012). Chromosomal instability and aneuploidy in cancer: from yeast to man. *EMBO Reports*, *13*(6), 515–527. https://doi.org/10.1038/embor.2012.65

Pollock, A. J., Woodward, J. J., Dreifus, J. E., Merrikh, H., Lang, K. S., Hall, A. N., … Tabakh, H. (2017). Replication-Transcription Conflicts Generate R-Loops that Orchestrate Bacterial Stress Survival and Pathogenesis. *Cell*, *170*(4), 787–799. https://doi.org/10.1016/j.cell.2017.07.044.Replication-Transcription

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

Reynolds, D., Cliffe, L., Forstner, K. U., Hon, C.-C., Siegel, T. N., & Sabatini, R. (2014). Regulation of transcription termination by glucosylated hydroxymethyluracil, base J, in Leishmania major and Trypanosoma brucei. *Nucleic Acids Research*, 1–13. https://doi.org/10.1093/nar/gku714

Reynolds, David, Hofmeister, B. T., Cliffe, L., Alabady, M., Siegel, T. N., Schmitz, R. J., & Sabatini, R. (2016). Histone H3 Variant Regulates RNA Polymerase II Transcription Termination and Dual Strand Transcription of siRNA Loci in Trypanosoma brucei. *PLOS Genetics*, *12*(1), e1005758. https://doi.org/10.1371/journal.pgen.1005758

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative Genomics Viewer. *Nature Biotechnology*, *29*(1), 24–26. https://doi.org/10.1038/nbt.1754.Integrative

Robinson, N. P., & Bell, S. D. (2007). Extrachromosomal element capture and the evolution of multiple replication origins in archaeal chromosomes. *Proceedings of the National Academy of Sciences*, *104*(14), 5806–5811. https://doi.org/10.1073/pnas.0700206104

Robinson, Nicholas P, & Bell, S. D. (2005). Origins of DNA replication in the three domains of life. *The FEBS Journal*, *272*(15), 3757–3766. https://doi.org/10.1111/j.1742-4658.2005.04768.x

Rogers, M. B., Hilley, J. D., Dickens, N. J., Wilkes, J., Bates, P. a., Depledge, D. P., … Mottram, J. C. (2011). Chromosome and gene copy number variation allow major structural change between species and strains of Leishmania. *Genome Research*, *21*, 2129–2142. https://doi.org/10.1101/gr.122945.111

S.Newlon, C., & F.Theis, J. (1993). The structure and function of yeast ARS elements. *Current Opinion in Genetics & Development*, *3*(5), 752–758. Retrieved from https://doi.org/10.1016/S0959-437X(05)80094-2

Schaper, S., & Messer, W. (1995). Interaction of the initiator protein DnaA of Escherichia coli with its DNA target. *The Journal of Biological Chemistry*, *270*(29), 17622–17626.

Sekimizu, K., Bramhill, D., & Kornberg, A. (1987). ATP Activates dnaA Protein in Initiating Replication of Plasmids Bearing the Origin of the E . coli Chromosome. *Cell*, *50*, 259–265.

Siegel, T. N., Hekstra, D. R., Kemp, L. E., Figueiredo, L. M., Lowell, J. E., Fenyo, D., … Cross, G. a M. (2005). Four histone variants mark the boundaries of polycistronic transcription units in Trypanosoma brucei. *Genes & Development*, 1063–1076. https://doi.org/10.1101/gad.1790409.7

Sollelis, L., Macpherson, C. R., Stanojcic, S., Kuk, N., Crobu, L., Bringaud, F., … Scherf, A. (2017). *Identification of the centromeres of Leishmania major : revealing the hidden pieces*. 1–10. https://doi.org/10.15252/embr.201744216

Stanojcic, S., Sollelis, L., Kuk, N., Crobu, L., Balard, Y., Schwob, E., … Sterkers, Y. (2016). Single-molecule analysis of DNA replication reveals novel features in the divergent eukaryotes Leishmania and Trypanosoma brucei versus mammalian cells. *Scientific Reports*, *6*(October 2015), 23142. https://doi.org/10.1038/srep23142

Sterkers, Y., Lachaud, L., Bourgeois, N., Crobu, L., Bastien, P., & Pagès, M. (2012). Novel insights into genome plasticity in Eukaryotes: Mosaic aneuploidy in Leishmania. *Molecular Microbiology*, *86*(August), 15–23. https://doi.org/10.1111/j.1365-2958.2012.08185.x

Sterkers, Y., Lachaud, L., Crobu, L., Bastien, P., & Pagès, M. (2011). FISH analysis reveals aneuploidy and continual generation of chromosomal mosaicism in Leishmania major. *Cellular Microbiology*, *13*(2), 274–283. https://doi.org/10.1111/j.1462-5822.2010.01534.x

Subramanian, A., & Sarkar, R. R. (2015). Comparison of codon usage bias across Leishmania and Trypanosomatids to understand mRNA secondary structure, relative protein abundance and pathway functions. *Genomics*. https://doi.org/10.1016/j.ygeno.2015.05.009

Sun, J., Evrin, C., Samel, S., Fernandez-Cid, A., Riera, A., Kawakami, H., … Li, H. (2013). Cryo-EM structure of a helicase loading intermediate containing ORC-Cdc6-Cdt1-MCM2-7 bound to DNA. *Nature Structural & Molecular Biology*, *20*(8), 944–951. https://doi.org/10.1038/nsmb.2629

Symington, L. S., & Gautier, J. (2011). Double-Strand Break End Resection and Repair Pathway Choice. *Annual Reviews of Genetics*, *45*, 247–273. https://doi.org/10.1146/annurev-genet-110410-132435

Thomas, S., Green, A., Sturm, N. R., Campbell, D. a, & Myler, P. J. (2009). Histone acetylations mark origins of polycistronic transcription in Leishmania major. *BMC Genomics*, *10*, 152. https://doi.org/10.1186/1471-2164-10-152

Tiengwe, C., Marcello, L., Farr, H., Dickens, N., Kelly, S., Swiderski, M., … McCulloch, R. (2012). Genome-wide analysis reveals extensive functional interaction between

DNA replication initiation and transcription in the genome of trypanosoma brucei. *Cell Reports*, *2*(1), 185–197. https://doi.org/10.1016/j.celrep.2012.06.007

Tiengwe, C., Marques, C. a., & McCulloch, R. (2014). Nuclear DNA replication initiation in kinetoplastid parasites: New insights into an ancient process. *Trends in Parasitology*, *30*, 27–36. https://doi.org/10.1016/j.pt.2013.10.009

Tosato, V., Ivens, A. C., & Rajandream, Á. (2001). *Secondary DNA structure analysis of the coding strand switch regions of ® ve Leishmania major Friedlin chromosomes*. 186–194. https://doi.org/10.1007/s002940100246

Trapnell, C., Williams, B. a, Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., … Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*(5), 511–515. https://doi.org/10.1038/nbt.1621

Ubeda, J.-M., Légaré, D., Raymond, F., Ouameur, A. A., Boisvert, S., Rigault, P., … Ouellette, M. (2008). Modulation of gene expression in drug resistant Leishmania is associated with gene amplification, gene deletion and chromosome aneuploidy. *Genome Biology*, *9*(7), R115. https://doi.org/10.1186/gb-2008-9-7-r115

Ubeda, J. M., Raymond, F., Mukherjee, A., Plourde, M., Gingras, H., Roy, G., … Ouellette, M. (2014). Genome-Wide Stochastic Adaptive DNA Amplification at Direct and Inverted DNA Repeats in the Parasite Leishmania. *PLoS Biology*, *12*(5). https://doi.org/10.1371/journal.pbio.1001868

Valton, A. L., Hassan-Zadeh, V., Lema, I., Boggetto, N., Alberti, P., Saintomé, C., … Prioleau, M. N. (2014). G4 motifs affect origin positioning and efficiency in two vertebrate replicators. *EMBO Journal*, *33*(7), 732–746. https://doi.org/10.1002/embj.201387506

Van Luenen, H. G. a M., Farris, C., Jan, S., Genest, P. A., Tripathi, P., Velds, A., … Borst, P. (2012). Glucosylated hydroxymethyluracil, DNA base J, prevents transcriptional readthrough in Leishmania. *Cell*, *150*, 909–921. https://doi.org/10.1016/j.cell.2012.07.030

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks / a Publication of the IEEE Neural Networks Council*, *10*(5), 988–999. https://doi.org/10.1109/72.788640

Wheeler, R. J., Gluenz, E., & Gull, K. (2011). *The cell cycle of Leishmania : morphogenetic events and their implications for parasite biology*. *79*(December 2010), 647–662. https://doi.org/10.1111/j.1365-2958.2010.07479.x

Wu, H. (2012). *polyaPeak : a tool for reranking ChIP-seq peaks with peak shape information*. 1–2.

Yates, A., Bliven, S. E., Rose, P. W., Jacobsen, J., Troshin, P. V, Chapman, M., … Rims, G. (2012). BioJava : an open-source framework for bioinformatics in 2012. *Bioinformatics*, *28*(20), 2693–2695. https://doi.org/10.1093/bioinformatics/bts494

Zang, C., Schones, D. E., Zeng, C., Cui, K., Zhao, K., & Peng, W. (2009). *A clustering approach for identification of enriched domains from histone modification ChIP-Seq data*. *25*(15), 1952–1958. https://doi.org/10.1093/bioinformatics/btp340

Zellner, E., Herrmann, T., Schulz, C., & Grummt, F. (2007). Site-specific interaction of the murine pre-replicative complex with origin DNA: Assembly and disassembly during cell cycle transit and differentiation. *Nucleic Acids Research*, *35*(20), 6701–6713. https://doi.org/10.1093/nar/gkm555

Zhang, X., Robertson, G., Krzywinski, M., Ning, K., Droit, A., Jones, S., & Gottardo, R. (2011). PICS: Probabilistic Inference for ChIP-seq. *Biometrics*, *67*(1), 151–163. https://doi.org/10.1111/j.1541-0420.2010.01441.x

Zhou, Y., Bizzaro, J. W., & Marx, K. a. (2004). Homopolymer tract length dependent enrichments in functional regions of 27 eukaryotes and their novel dependence on the organism DNA (G+C)% composition. *BMC Genomics*, *5*, 95. https://doi.org/10.1186/1471-2164-5-95

Zijlstra, E. E., Musa, A. M., Khalil, E. A. G., & Hassan, I. M. El. (2003). *Review Post-kala-azar dermal leishmaniasis*. *3*(February), 87–98.