GENOME PLASTICITY AND GENETIC EXCHANGE

IN *LEISHMANIA TROPICA*

Stefano Iantorno

Gonville and Caius College

This dissertation is submitted for the title of

Doctor of Philosophy

at the University of Cambridge

June 30th, 2015

Declaration

I hereby declare that this dissertation is entirely the product of my own work and contains nothing that is the product of work done in collaboration with others except when explicitly stated here and in the main text.

The sequence data that was used in this thesis was produced by the core Sequencing production teams at the Wellcome Trust Sanger Institute.

None of the work presented has been submitted for the purpose of obtaining another degree. This dissertation does not exceed 60,000 words in length, as required by the School of Biological Sciences.

Stefano Iantorno, June 2015

To my parents

ACKNOWLEDGEMENTS

Although the work presented in this dissertation is my own, this thesis is the product of concerted efforts by many different people. I thank my supervisors, David Sacks, Michael Grigg, Matt Berriman, and James Cotton for their unfailing support and critical insight in both the experimental and analytic portions of my thesis, and for always being able to provide a high level perspective to guide the direction of the research. Among those who helped carry this research project forward at NIH, I am greatly indebted to Audrey Romano and Ehud Inbar for their humour and for their assistance with a range of experimental procedures, and to Kim Beacht for assistance with the mouse work. Phillip Lawyer's wisdom provided countless opportunities to deepen my knowledge of medical entomology and his technical skills proved essential in maintaining the laboratory colonies for the sand fly feeding assays. Kashinath Ghosh also provided assistance with the sand fly colonies and sand fly midgut dissections. Alain Debrabant offered crucial assistance with transfection procedures and with genetic modification of parasite strains. At the Wellcome Trust Sanger Institute, Caroline Durrant's firm grasp on statistics was essential to some of the more complicated analyses. Adam Reid provided guidance with the RNA-seq analyses. Mandy Sanders oversaw all sequencing procedures and facilitated communications with the core sequencing teams. All of the analyses performed in this thesis would not have been possible if not for the essential work done by Alan Tracey and Karen Brooks in the pathogen genome finishing team

improving the *L. tropica* reference assembly. Lastly, I thank the NIH-OXCAM program for generous funding and for providing excellent career development opportunities throughout my graduate studies.

SUMMARY

Leishmania is a genus of unicellular eukaryotic parasites responsible for a wide range of human diseases, from cutaneous (CL) and mucocutaneous leishmaniasis (MCL) to life-threatening visceral leishmaniasis (VL). *Leishmania tropica* is responsible for significant CL in endemic areas in North and East Africa, the Middle East, and the Indian subcontinent, and has also been associated with a variant form of VL called viscerotropic leishmaniasis. Significant heterogeneity has been observed in *L. tropica* in both clinical course of disease and in response to treatment, and published data suggests there is great genetic diversity within this species. RNA-seq analysis of 12 clinical isolates of *Leishmania tropica* revealed considerable intraspecific differences in gene expression. Comparison with whole-genome sequence data generated from the same 12 isolates using a new reference genome assembly suggests that most variation in gene expression is explainable by variation in copy number at the level of individual genes, or at the level of whole chromosomes. Most field isolates appear to be near diploid, but some degree of aneuploidy is seen in all isolates. Cloning of single cells from 4 of these isolates showed variable ploidy within the same clinical isolate, a condition that in *Leishmania* has been called mosaic aneuploidy. The most significant differentially expressed genes in this set of isolates code for membrane-bound transporter proteins, which are known to be involved in uptake of nutrients and drug compounds from the extracellular environment. We identify copy number variation in these genes suggesting that a certain degree of plasticity is observed in natural

populations of *Leishmania*, creating the conditions necessary for rapid downregulation or upregulation of different transporter proteins over a limited number of mitotic generations in the presence of environmental stressors. Such an evolutionary phenomenon could be important in mediating decreased susceptibility to drug treatment in endemic areas. To further understand how such large genetic variation can be generated and the role of genetic exchange in shaping the genomic landscape in this important pathogen, we have carried out a controlled laboratory cross between one isolate collected in Israel and one collected in Lebanon. Ten hybrid lines were recovered from crosses we performed in sand flies. The present study provides the first in-depth, complete description of structural genome changes and recombination occurring during hybridization in an artificial cross of *Leishmania tropica*. The implications of this structural variation for parasite evolution in natural populations in response to drug pressure due to increased elimination efforts will be discussed.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

Amplified fragment length polymorphism	AFLP
Bayesian information criterion	BIC
Copy Number Variant	CNV
Cutaneous leishmaniasis	CL
Differentially expressed	DE
Discriminant analysis of principal component	DAPC
False discovery rate	FDR
Fluorescent *in situ* hybridization	FISH
Generalized linear model	GLM
Hardy Weinberg	HW
Human African trypanosomiasis	HAT
Human leukocyte antigen	HLA
Identical by descent	IBD
Leishmaniasis recidivans	LR
Linkage disequilibrium	LD
Lipophosphoglycan	LPG
Long runs of homozygosity	LROH
Mucocutaneous leishmanias	MCL
Multi-dimensional scaling	MDS
Multi-locus enzyme electrophoresis	MLEE
Multi-locus sequence analysis	MLSA
Multi-locus sequence typing	MLST
Neglected tropical diseases	NTD
Neighbour-Joining	NJ
Polymerase chain reaction	PCR
Post-kala-azar dermal leishmaniasis	PKDL
Principal components analysis	PCA
Promastigote secretory gel	PSG
Pulsed-field gel electrophoresis	PFGE
Quantitative trait locus	QTL
Random amplified polymorphic DNA	RAPD
Read depth	RD
Reactive oxygen species	ROS
Relative log expression	RLE
Single nucleotide polymorphism	SNP
Spliced leader	SL
T-cell receptor	TCR
Transcription start site	TSS
Trimmed mean of M-values	TMM
Untranslated region	UTR
Visceral leishmaniasis	VL
Viscerotropic leishmaniasis	VTL
Whole-genome sequencing	WGS
Identical by descent	IBD

CHAPTER 1

INTRODUCTION

1.1. Leishmaniasis, a complex parasitic disease

1.1.1 Overview

The leishmaniases are a group of vector-borne parasitic diseases that collectively affect around 12 million people around the globe, according to estimates by the World Health Organization. Approximately 310 million additional people worldwide are at risk of infection. The disease is spread by the bite of female sand flies of the genera *Lutzomyia* and *Phlebotomus*, and is caused by 20 different species of flagellated unicellular protozoa belonging to the genus *Leishmania*, although the taxonomic status of some of them is disputed. Clinical disease due to *Leishmania* presents in three major forms: cutaneous leishmaniasis, mucocutaneous leishmaniasis, and visceral leishmaniasis. Asymptomatic cases are known to exist in endemic areas, and may act as an important reservoir for re-infection (Singh, Hasker et al. 2014). The parasite may follow two different transmission cycles: a zoonotic cycle, with dogs being an especially important animal reservoir in addition to other mammals; and a strictly anthroponotic cycle, with humans as the only host, which is typically observed in densely inhabited urban areas.

Accurate epidemiological data from many endemic countries is lacking (Alvar, Velez et al. 2012), so the true burden of disease may be higher than official WHO estimates, especially if the mental health repercussions - in the form of social stigma and ostracization - associated with some clinical manifestations are appropriately factored in. In addition, HIV-*Leishmania* coinfection is a major complicating form of disease that has been systematically underreported in many endemic areas (WHO 2007), and for which poor clinical guidelines have been established (Diro, van Griensven et al. 2015). Opportunistic infection with *Leishmania* is an AIDS-defining illness in endemic settings, and the immunosuppressive effects of the parasitic infection are compounded by infection with HIV, often with irremediable consequences for the patient. *Leishmania* is widely considered to be the second biggest parasitic killer after malaria, and it is thought that global warming, anthropogenic environmental changes, and human migrations have led to an expansion of the geographic distribution of the leishmaniases (Desjeux 2004).

Visceral leishmaniasis (VL), also known as kala-azar, black fever, or Dumdum fever, is a systemic illness caused by infection of the liver, spleen, and bone marrow by the parasite. If left untreated, the disease is almost invariably lethal within two years of presentation of symptoms. According to the WHO Global Health Observatory, approximately 200 000-400 000 cases of VL occur per year, 90% of which are concentrated in 6 countries: Bangladesh, India, Ethiopia, Sudan, South Sudan, and Brazil. An estimated 20 000 to 30 000 deaths are due to VL each year, although the true number may be higher due to poor epidemiological surveillance in

many areas of active transmission. A serious complication that may occur in cases of VL following treatment is post-kala-azar dermal leishmaniasis (PKDL), a chronic syndrome characterized by the appearance, most notably on the face, of a multitude of papules, nodes, and patches, which require prolonged chemotherapy. In the Indian subcontinent, PKDL is rare, appears several years after successful therapy, and is particularly hard to treat. Conversely, in East Africa PKDL is more frequent, is noted a few months after initial therapy for VL, and often resolves spontaneously. Patients with PKDL may act as important reservoir hosts.

Mucocutaneous leishmaniasis (MCL), also known as espundia, is a severely disfiguring disease which typically occurs as a metastatic dissemination of the parasite from the site of initial infection on the skin to mucosal membranes of the body, primarily around the nose and mouth. Complete ulcerative destruction of the oro-pharyngeal tissues is often seen in the absence of timely treatment. Almost 90% of all cases of MCL occur in Bolivia, Brazil, and Peru. The disease is caused by New World species of *Leishmania* belonging to the *Viannia* subgenus, namely *L. [V.] braziliensis*, *L. [V.] panamensis*, *L. [V.] guyanensis*, and sometimes *L. amazonensis*. The first symptoms (e.g., persistent nosebleeds) may appear several months or even years after healing of the primary skin lesion. Disease pathogenesis is poorly understood, and is often associated with lack of appropriate treatment of cutaneous lesions.

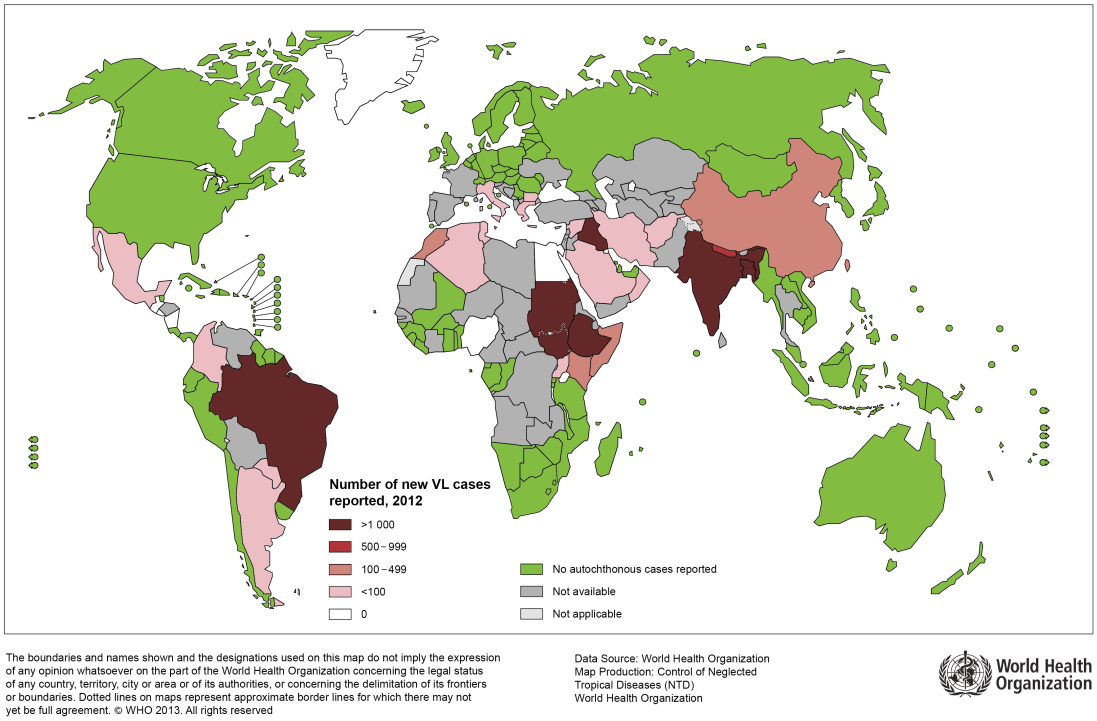


Figure 1.1: Distribution and endemicity of visceral leishmaniasis (VL) according to 2013 annual country reports. Countries in grey have no reliable epidemiological data or do not report disease incidence to the WHO Neglected Tropical Diseases (NTD) section. Countries in green had no autochthonous cases of VL reported in 2012 (source: WHO Global Health Observatory).

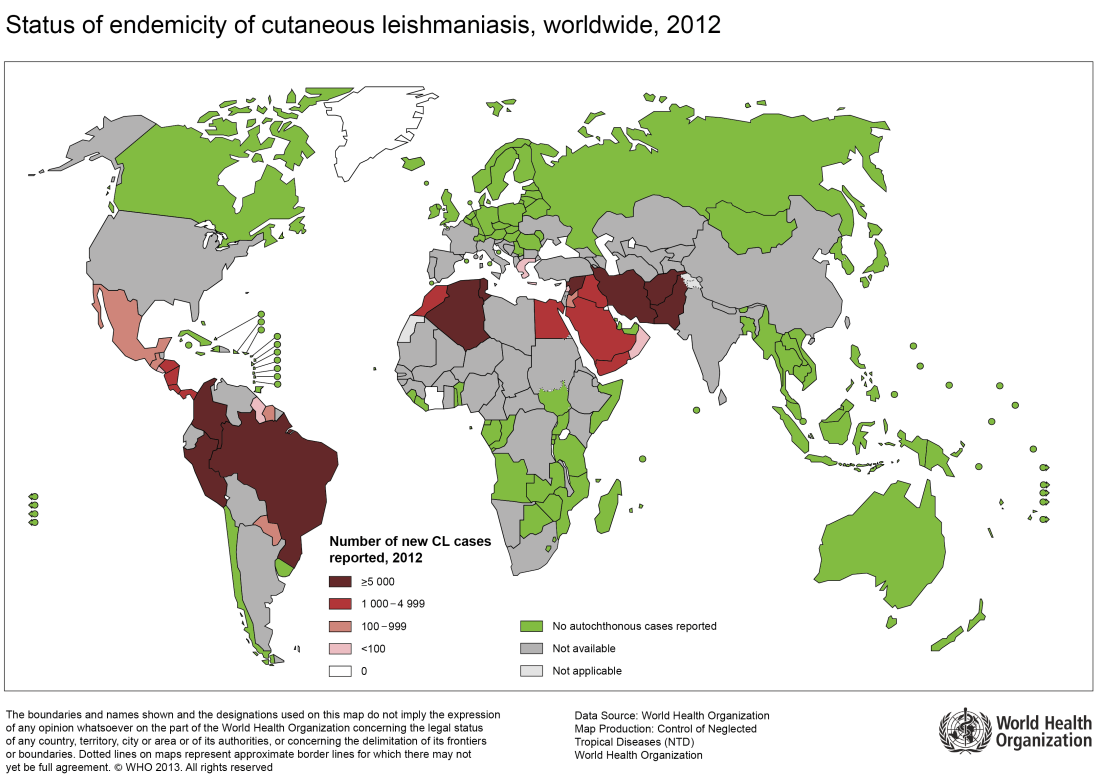


Figure 1.2: Distribution and endemicity of cutaneous leishmaniasis, (CL) according to 2013 country reports. Countries in grey have no reliable epidemiological data, or do not report disease incidence to the WHO Neglected Tropical Diseases (NTD) section. Countries in green had no autochthonous cases of CL reported in 2012 (source: WHO Global Health Observatory).

Cutaneous leishmaniasis (CL) is the most common form of disease, and is caused by both New World and Old World species of *Leishmania*. The disease manifests as ulcerative lesions that develop at the site of the sand fly bite, which can grow progressively larger and fail to heal naturally. These lesions often appear on exposed areas of the body such as the face, and are associated with significant social stigma and disability. If infected with bacteria, these lesions can be quite painful. Lymphadenopathy often precedes appearance of the lesion. Even with this localized form of disease, often multiple satellite lesions appear, and can persist for months or years. Successful treatment does not eliminate the noticeable scars that remain visible for life. Around 1.3 million new cases of CL occur annually worldwide according to the WHO Global Health Observatory.

This parasitic disease is one of the most neglected disease of the developing world, and is often entrenched in areas where poor sanitation, poor access to healthcare, and strained infrastructures due to war or social unrest lead to a combination of factors favorable to the establishment of recurrent epidemic cycles of transmission (Beyrer, Villar et al. 2007). A promising leishmaniasis elimination campaign has been championed by WHO since 2005 in Bangladesh, India, and Nepal, with the objective of reducing the incidence of VL to one case per 10 000 at the district or sub-district level by 2015. There were around 20 cases per 10 000 in the region in 2011, and the campaign has been making remarkable progress. The elimination of VL in this region is made achievable by the presence of a single sand fly vector species that is susceptible to insecticides; the distribution of cases in geographic clusters; and the fact that humans are the only reservoirs of infection.

The presence of asymptomatic carriers could complicate complete elimination in the region. Large-scale control and global elimination of all clinical forms of disease associated with *Leishmania* infection, on the other hand, is poised to be a significant challenge, given the remarkable differences observed in clinical presentation in different patient populations, the presence of zoonotic reservoirs, and the number of different parasite species causing significant disease. An improved understanding of disease pathogenesis and the transmissibility of the parasite in each clinical presentation can inform prioritization of different elimination strategies, as would the presence of an effective vaccine, the development of point-of-care diagnostics, and an affordable, easy-to-administer oral formulation for drug therapy (Matlashewski, Arana et al. 2014).

Currently, leishmaniasis is treated with a variety of remedies, ranging from first-line pentavalent antimonials, which are poorly tolerated in patients and to which many circulating parasite strains have developed resistance, to different regimens of amphotericin B, paromomycin, fluconazole, and the promising oral drug miltefosine, which recently received regulatory approval for use in India and the United States. No human vaccine is available, although there are several candidates in pre-clinical and clinical stages. The historical practice of "leishmanization", whereby live parasites are inoculated in the skin in a cosmetically acceptable part of the body, is the only fully effective way to gain immunity to CL. The fact that individuals who fully recover from VL are then resistant to re-infection suggests that immunity to symptomatic visceral disease can be achieved via vaccination (Working Group on Research Priorities for Development of Leishmaniasis, Costa et al. 2011).

1.1.2 The biology of the parasite

Leishmania belongs to a class of unicellular protists known as Kinetoplastida. The only kinetoplastid organisms known to cause disease in humans are: approximately 20 different *Leishmania* species; the two parasite species responsible for human African trypanosomiasis (HAT), or sleeping sickness, *Trypanosoma brucei brucei* and *T. brucei rhodesiense*; and *T. cruzi*, the parasite species responsible for Chagas disease in the Americas. All kinetoplastids, in addition to being flagellated for at least part of their life cycle, also share a unique DNA-containing organelle known as the kinetoplast, situated in a mitochondrian-like structure at the base of the flagellum. The kinetoplast contains multiple circular copies of kinetoplast DNA (kDNA), which serve the same function as the mitochondrial genome in more advanced eukaryotes.

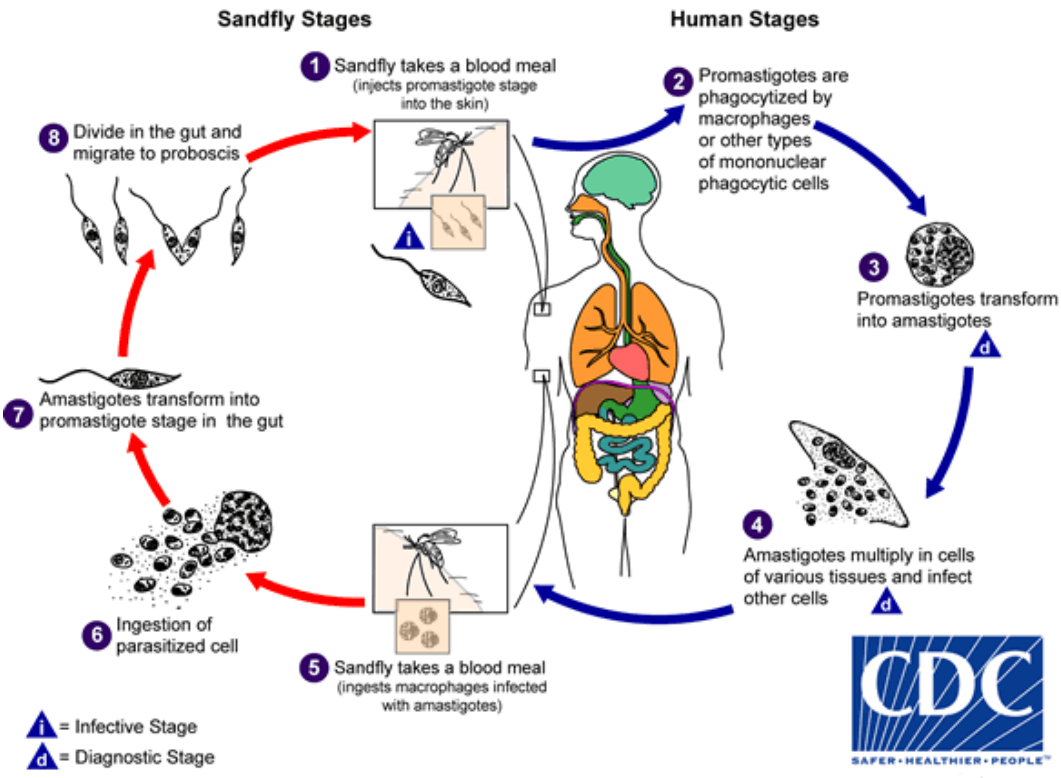


Figure 1.3. The *Leishmania* life cycle. (1) Upon bloodfeeding, the sandfly inoculates infectious metacyclic promastigotes into the host's skin; (2) once in the skin, promastigotes are ingested by phagocytic cells; (3) within the phagocytic cell the parasite differentiates into obligate intracellular amastigotes; (4) the parasite replicates intracellularly through multiple rounds of mitosis, invading neighbouring cells; amastigote-infected cells may localize to the skin lesion, or spread to other sites in the body; (5) circulating amastigote-infected macrophages are taken up in the blood meal of a sand fly; (6-7) amastigotes differentiate into extracellular promastigotes and attach to the midgut wall to survive excretion of the digested bloodmeal; (8) promastigotes migrate anteriorly and undergo a series of developmental

transitions to form infectious metacyclic promastigotes, encased within a PSG plug that blocks normal feeding of the sand fly (Source: US Center for Disease Control and Prevention, Division of Parasitic Diseases and Malaria).

Like *T. brucei* and *T. cruzi*, *Leishmania* has a digenetic life cycle, alternating between the sand fly vector and the mammalian host. When female sand flies blood feed on an appropriate host, the parasite is inoculated into the skin as metacyclic promastigotes, the infectious, extracellular, non-replicative stage (Bates 2007). These metacyclic promastigotes are lodged near the stomadeal valve in the anterior gut of the sand fly, and are encased in a gel-like "plug" created via secretion of PSG, or promastigote secretory gel. The sand fly is forced to regurgitate this plug into the skin as it takes its meal. Impaired uptake of blood leads to the sand fly attempting to feed with greater frequency, and thus increases the chance of parasite transmission (Rogers and Bates 2007).

The parasite is thus inoculated in the skin along with pro-inflammatory salivary components, where it then invades resident phagocytic cells within the skin tissues of the host. Once inside the host cell, the parasite differentiates into the obligate intracellular, non-flagellated form called the amastigote. The parasite continues to replicate by mitotic cell division, re-invading phagocytic cells as an intracellular amastigote, until it is taken up in the blood meal of the next sand fly. Both parasite and host factors are thought to be important in determining whether the infection is symptomatic and the type of pathology resulting from the infection. Tissue tropism of infecting parasites can vary, but VL is usually associated with

heavy infections of the liver, spleen, and bone marrow. Different symptomatology and distribution in the host tissues may determine differences in transmissibility of the parasite. Unusual tissue tropism has been observed in HIV-*Leishmania* coinfections, such as parasites in the gastroendothelial mucosa.

Once in the sand fly midgut, the amastigote forms differentiate into early procyclic promastigote stages, which are multiplicative and increase in numbers by cell division, while attaching to the interior wall of the sand fly midgut. Parasite attachment to the midgut wall is mediated by lipophospoglycan (LPG) covering the parasite cell surface. This molecule has a complex, branched structure composed by polysaccharides and polypeptides, and plays an important role in species-specific interactions between parasite and vector (Pimenta, Saraiva et al. 1994, Sacks 2001). By attaching to the midgut wall, the parasite survives expulsion of the digested blood meal as sand fly excrement. The parasite then differentiates into non-replicating nectomonad promastigotes, and migrates to the anterior part of the midgut where it resumes replication as leptomonad promastigotes. This stage is also responsible for production of promastigote secretory gel (PSG), and immediately precedes differentiation into mammalian-infective metacyclic promastigotes (Bates and Rogers 2004).

1.1.3 The burden of disease due to *L. tropica*

Among the approximately 20 species of *Leishmania* responsible for human disease, *L. tropica* appears to be related to *L. major*, and is considered by some to be

part of the same species complex as other CL-causing Old World species such as *L. aethiopica* and *L. killicki* (El Baidouri, Diancourt et al. 2013, Chaara, Ravel et al. 2015). *L. tropica* most often causes skin lesions similar to those caused by *L. major*, although lesions due to *L. tropica* tend to be larger, covered with scabs or crusts, and less responsive to treatment.

Skin lesions often recrudesce following apparent successful treatment, sometimes decades after the primary infection, in a syndrome unique to *L. tropica* known as leishmaniasis recidivans (LR). These encrusted or papular lesions are typically localized along the edge of the scarred tissue left by healing of the primary ulcer, and may be triggered by inflammation-activating events up to more than 40 years following healing of the primary lesion (Marovich, Lira et al. 2001).

In addition to these cutaneous manifestations, *L. tropica* has been associated with a variant form of VL known as viscerotropic leishmaniasis (VTL) that depending on both parasite and host factors may or may not resemble the classic symptomatology of VL disease, with weight loss, fever, splenohepatomegaly, general weakness and muscle pains, as well as dissemination of the parasite to internal organs and to the bone marrow, and subsequent immunosuppression and pancytopaenia in the more advanced stages of disease (Mebrahtu, Lawyer et al. 1989, Magill, Grogl et al. 1993, Sacks, Kenney et al. 1995, Alborzi, Rasouli et al. 2006, Weiss, Vogenthaler et al. 2009).

Cutaneous lesions due to *L. tropica* are treated with topical paromomycin, intralesional antimonials, and sometimes with topical miconazole. Systemic treatment is only recommended in the most serious cases or in LR, and it may

include the azoles ketoconazole and fluconazole, amphotericin B, parenteral antimonials, or a combination of several of these options (Monge-Maillo and Lopez-Velez 2013). The effectiveness of liposomal amphotericin B to treat CL due to *L. tropica* is only based on a few published case studies of patients where other treatment options had failed. No specific treatment guidelines exist for VTL due to *L. tropica*.

The *L. tropica* species complex has been found to be genetically extremely heterogeneous (Schwenkenbecher, Wirth et al. 2006), and *L. tropica* may follow different transmission cycles in different geographic areas. In densely populated urban or periurban areas of the Middle East, the parasite appears to follow an anthroponotic cycle, with no known zoonotic reservoir of infection and with human-to-human transmission contributing to the high incidence of disease. In other locations however, several animal reservoirs have been found, such as dogs in Morocco (Dereure, Rioux et al. 1991), jackals and foxes in Israel (Talmi-Frank, Kedem-Vaanunu et al. 2010), and rodents, such as rock hyraxes, belonging to the families Procaviidae and Ctenodactylidae in Tunisia, Israel, and East Africa (Sang, Njeru et al. 1994, Svobodova, Votypka et al. 2006, Talmi-Frank, Jaffe et al. 2010, Jaouadi, Haouas et al. 2011, Bousslimi, Ben-Ayed et al. 2012).

L. tropica is transmitted by several *Phlebotomus* species of sand flies. Most notably, it is transmitted by *P. sergenti* across the majority of its geographic distribution (Kamhawi, Modi et al. 2000), by *P. arabicus* in Israel (Svobodova, Volf et al. 2006), by *P. saevus* in Ethiopia (Gebre-Michael, Balkew et al. 2004), and by *P. guggisbergi* in Kenya (Lawyer, Mebrahtu et al. 1991). While *L. major* is known to be

primarily zoonotic, and the seasonal fluctuations in CL cases due to *L. major* follow seasonal changes in the population of the vector species *P. papatasi*, CL due to *L. tropica* most often occurs in urban areas, such as Aleppo in Syria and Kabul in Afghanistan, where it appears to be exclusively anthroponotic. In most areas in North Africa, however, the disease is hypoendemic, with only sporadic cases being reported in rural foci across Tunisia, Algeria, and Lybia, consistent with the presence of a zoonotic transmission cycle, although in Morocco a few large epidemics with hundreds of cases have also been reported in urban areas (Ajaoud, Es-sette et al. 2013).

Figure 1.4. Geographic distribution of Old World CL due to *L. tropica*, *L. aethiopica*, and related species (source: WHO Essential Leishmaniasis Maps).

Throughout its geographic range, cases of *L. tropica* seem to be increasing, possibly as a result of anthropogenic changes to the environment and subsequent expansion of the vector population to areas previously unaffected by the disease

(Reithinger, Dujardin et al. 2007). While *L. tropica* was rare in Kabul, Afghanistan before 1990, a marked increase in annual cases has been reported in recent years (Reithinger, Mohsen et al. 2003). In Aleppo, Syria, cases have also been steadily increasing in the last few decades, possibly reflecting changes in the vector population (Tayeh, Jalouk et al. 1997). Recent socio-political events may have precipitated these epidemics. According to the WHO, in 2013 Syria had the highest number of cases ever reported in the country, with 71,991 reported cases of CL. Afghanistan had 23,621 reported cases of CL in 2013. The majority of cases in both settings can be attributed to *L. tropica*.

New foci of infection have been reported in Northern Israel (Jacobson, Eisenberger et al. 2003), and large epidemics of *L. tropica* have been documented in refugee camps in Syria (Saroufim, Charafeddine et al. 2014) and in Pakistan (Rowland, Munir et al. 1999, Brooker, Mohammed et al. 2004). The disease is also highly endemic in the Arabian Peninsula and in Turkey, Iraq, Iran, and some areas of India. Displacement of populations due to armed conflict or civil unrest into areas with active transmission is known to be a major driver behind outbreaks of *L. tropica*. Although the majority of CL in East Africa is due to *L. major*, *L. tropica* is also present (Hotez, Savioli et al. 2012).

A reliable quantification of the burden of disease due to *L. tropica* that includes both CL and VTL manifestations has yet to be performed, although molecular typing of strains has confirmed earlier reports of *L. tropica* being a parasite species contributing to visceral disease in the region (Khanra, Datta et al. 2012, Krayter, Bumb et al. 2014).

1.1.4 Mechanisms of pathogenesis

The mechanisms by which non-healing, chronic lesions are generated upon inoculation in the skin are not fully understood. Previous exposure to uninfected sand fly bites are known to diminish the severity of disease (Kamhawi, Belkaid et al. 2000), so sand fly salivary proteins might play a role in the recruitment of phagocytic neutrophils to the site of the bite, which then act as an inflammatory "silent" route for the parasite to enter macrophages. Neutrophils are short-lived cells that undergo apoptosis, and are in turn phagocytosized by professional macrophages (van Zandbergen, Klinger et al. 2004, Peters, Egen et al. 2008). Components of the PSG plug are known to enhance infection (Rogers, Ilg et al. 2004). Although *Leishmania* can invade a variety of phagocytic and non-phagocytic mammalian cells, the parasite has a marked preference for macrophages. The metacyclic promastigote differentiates into the non-flagellated amastigote within the phagolysosome, where it carries out its entire replicative cycle. The only other microbe known to be capable of replicating within this cell compartment is the Gram-negative bacterium *Coxiella burnetti*. The LPG coat on the surface of the surface of the parasite is thought to play a role in initial uptake by neutrophils and macrophages by limiting the damage due to ROS generated during phagocytosis, although it then becomes strongly downregulated in the amastigote stages. Amastigote stages are highly opsonized by IgG antibodies, a characteristic that promotes uptake by macrophages via the Fc receptor and release of anti-inflammatory IL-10 (Kane and Mosser 2001).

Leishmania is auxotrophic for many aminoacids, which are required for protein biosynthesis, and must scavenge these from within the phagolysosome, along with carbon sources such as fatty acids and hexoses, and other essential nutrients such as heme, purines, and vitamins (McConville, de Souza et al. 2007). There is evidence that the parasite establishes a complex interplay with the host cell metabolism. Uptake of essential nutrients from the host cytosol is thought to occur via membrane proteins of the phagolysosome, and alternative activation of macrophages via Th2-associated cytokines such as IL-4 and IL-10 is associated with activation of the key host-encoded metabolic enzyme arginase-1, and subsequent increased production of amino acids essential to the parasite (Sacks and Anderson 2004).

Establishment of chronic infection has long been attributed to a Th2 cytokine profile in *Leishmania* infection (Heinzel, Sadick et al. 1991). Downregulation of effector T cells by regulatory T cells has been implicated in the establishment of chronic infection and immunity to reinfection at other sites of the body (Belkaid, Piccirillo et al. 2002). More recently, a role for NLRP3 activation and IL-18 has been found in the establishment of chronic infection (Gurung, Karki et al. 2015). In contrast to *L. major*, abrogation of IL-10 production is not sufficient to clear *L. tropica* from the site of infection (Anderson, Lira et al. 2008), suggesting important differences in pathogenesis between infection with *L. tropica* and with *L. major*.

Host genetics undoubtedly play a factor in pathogenesis. It should be noted that the majority of infections are completely asymptomatic in *L. infantum* and *L donovani* (Sakthianandeswaren, Foote et al. 2009). Recently, a large-scale case-

control study for susceptibility to VL in Brazilian and Indian populations found a strong association with the HLA locus, specifically, with the HLA class II regions HLA-DRB1 and HLA-DQA1 (LeishGEN Consortium, Wellcome Trust Case Control Consortium et al. 2013). CD4+ T cells are known to secrete interferon gamma to control infection in acute VL, and the HLA class II region may be involved in recognition of *Leishmania* antigen via classical interaction with TCRs (Kumar and Engwerda 2014). CD8+ cytotoxic T cells might also play a role (Mansueto, Vitale et al. 2007).

1.2. Alternative genetics in *Leishmania*

1.2.1 The unique genome of kinetoplastids

Leishmania parasites have a genome architecture and associated biology that are unlike those of almost any other eukaryote. Trypanosomatids as a family appear to be a very ancient clade of the eukaryotic tree of life (He, Fiz-Palacios et al. 2014), and the whole class Kinetoplastida appear to have remarkably unique features. Since publication of the *L. major* genome in 2005 (Ivens, Peacock et al. 2005), a series of discoveries have been made regarding *Leishmania* genome organization that are conserved across trypanosomatid species.

First, genes are organized head-to-tail into large clusters, often hundreds of kilobases in length, sharing the same direction of transcription. Chromosome 1, for instance, is transcribed in two large clusters, the first cluster of 29 genes on one

strand, and the remaining 50 genes on the other strand (Myler, Audleman et al. 1999). Transcription is thought to initiate and terminate in strand switch regions, with the transcribed polycistronic segments on opposite strands being either directionally divergent or convergent (Martinez-Calvillo, Yan et al. 2003, Martinez-Calvillo, Nguyen et al. 2004). Transcription is thought to primarily initiate in divergent strand switch regions, and terminate in convergent strand switch regions.

Second, in contrast to bacterial operons, polycistronic transcripts in *Leishmania* require further processing before translation. The mature mRNA transcript originates from coupled trans-splicing and polyadenylation of the initial polycistronic unit. A 39-nucleotide mini-exon sequence, called the spliced leader (SL), is trans-spliced to the 5' end of each gene (Sutton and Boothroyd 1986, Perry, Watkins et al. 1987). This process is mediated by the spliceosome complex, and is coupled with polyadenylation of the 3' end of the gene upstream of the splice site (LeBowitz, Smith et al. 1993, Matthews, Tschudi et al. 1994). The enzymatic machinery involved in mRNA maturation has not been fully characterized, although the complete set of small nuclear RNAs involved in trans-splicing has been found (Liang, Haritan et al. 2003).

Third, given the role played by polycistrons and their maturation by trans-splicing, individual *Leishmania* genes lack traditional eukaryotic promoters. Nuclear genome polycistrons appear to be constitutively transcribed by RNA polymerase II, and individual genes lack introns. The only gene known to be transcribed via eukaryotic promoter sequences in *Leishmania* is the SL RNA gene, which is well conserved in kinetoplastids (Gilinger and Bellofatto 2001). The primary transcript

of the SL RNA gene presents a distinctive three-loop secondary structure, as well as a terminal intronic sequence involved in binding with maturation factors. The intronic sequence contains two distinct domains, the -60 and -30 elements, so called based on their position with respect to the transcription start site (TSS) (Sturm, Fleischmann et al. 1998, Sturm and Campbell 1999, Sturm, Yu et al. 1999). The mature SL RNA transcript is capped at its 5' end with a modified nucleotide sequence called the ``4-cap'', due to the presence of 4 consecutive methylated bases. Trans-splicing of this capped SL RNA sequence to mature mRNAs requires an AG 3' acceptor site and U-rich polypyrimidine tract (PPT) in the target mRNA (Curotto de Lafaille, Laban et al. 1992, LeBowitz, Smith et al. 1993, Matthews, Tschudi et al. 1994). Recently, basal splicing factors were suggested to be involved in determining the exact splice site and thus the size of the 3' UTR. The size of the 3' UTR in turn may affect polyadenylation by inclusion or exclusion of a given regulatory sequence (Gupta, Carmi et al. 2013).

Fourth, kinetoplastid organisms also possess a kinetoplast-localized genome known as kDNA, which codes for mitochondrial metabolic enzymes involved in cellular respiration. The kDNA is organized into a network of interlocked minicircles and maxicircles. The minicircles, which are smaller (~1 kb) and more numerous (~100 copies), code for guide RNAs, while the maxicircles, which are larger (~25kb) and less abundant (~50 copies), code for pre-mRNAs. The pre-mRNAs encoded by the maxicircles mature by a unique process known as RNA editing, involving selective insertion and deletion of uridine residues. Ribosomal RNA and pre-mRNAs are transcribed by a phage-like polymerase in multicistronic units (Hajduk and

Ochsenreiter 2010). The guide RNAs are produced by 3' nucleolytic processing and uridylation of longer precursors encoded by the minicircles, and direct the editing reactions. Most proteins involved in the RNA editing process and all tRNAs are imported from the cytoplasm, although many of these molecules mature by additional modifications which occur in the kinetoplast (Aphasizhev and Aphasizheva 2011).

1.2.2 Karyotypic variation in *Leishmania*

The genomes of Old World species of *Leishmania* are organized into 36 heterologous chromosomes, while the New World species fall into two groups: species belonging to the *Viannia* subgenus have a 35-chromosome karyotype due to a fusion between chromosome 20 and chromosome 34, while species belonging to the *Leishmania* subgenus have a 34-chromosome karyotype, due to a fusion between chromosome 8 and 9 and between chromosome 20 and 36 (Britto, Ravel et al. 1998).

Overall, genome structure is remarkably consistent across species, with conserved synteny observed for approximately 99% of all genes. Overall, only ~200 genes are distributed in different regions of the genome in *L. major*, *L. infantum*, and *L. braziliensis*. Coding regions in *L. major* and *L. infantum* share up to 92% of the amino acid sequence, and 94% of the nucleotide sequence, while in the comparison with *L. braziliensis* these numbers drop to approximately 77% and 81%, respectively (Peacock, Seeger et al. 2007).

An important feature of *Leishmania* genomes is the extreme variation in chromosome copy number between strains of the same species, with individual chromosomes often having more than the two copies expected in diploid organisms (Rogers, Hilley et al. 2011, Sterkers, Lachaud et al. 2011, Lachaud, Bourgeois et al. 2014). Fluorescent *in situ* hybridization (FISH) experiments and whole genome sequencing (WGS) have shown that each heterologous chromosome may have a somy level different from that of the other chromosomes. Individual cells of any given lab-adapted strain have been shown to have variable patterns of aneuploidy, so that a given sample from a patient for instance, once grown in culture is rarely formed by a homogeneous population of cells with the same karyotype, a condition that has been named mosaic aneuploidy (Mannaert, Downing et al. 2012, Sterkers, Crobu et al. 2014). Monosomy of chromosomes has also been observed. For unexplained reason, chromosome 31 seems to be tetrasomic in many of the strains examined thus far.

In addition to the variation seen in chromosome number, *Leishmania* parasites are also susceptible to amplification or deletion of certain regions of the genome. Gene amplification has been documented as formation of linear or circular extrachromosomal amplicons, or as intrachromosomal gene duplication, which leads to the formation of tandem gene arrays. A conserved extrachromosomal circular amplicon was found in *L. donovani* field isolates (Downing, Imamura et al. 2011), and extrachromosomal amplicons are known to occur *in vitro* in response to drug pressure (Beverley, Coderre et al. 1984). The presence of short sequence repeats flanking individual genes was proposed as a mechanism facilitating

amplification and deletion via the molecular pathway involved in homologous recombination (Ubeda, Legare et al. 2008, Laffitte, Genois et al. 2014, Ubeda, Raymond et al. 2014).

1.2.3 Transcriptional regulation (or lack thereof)

Since *Leishmania* parasites lack eukaryotic promoters, much research has focused on how these organisms can modulate gene expression in response to environmental stimuli. The question of how parasites turn on and off genes involved in developmental transitions, such as in the differentiation from the extracellular promastigote to the intracellular amastigote stages, is an area of active inquiry. Constitutive changes in gene expression in response to selection, however, as described in Section 1.2.2, seem to arise over a limited number of generations via copy number variation at the level of individual genes, or at the level of whole chromosomes.

No control of transcriptional initiation has been so far found in *Leishmania*. Polycistrons appear to be constitutively transcribed into mRNA by RNA polymerase II. These transcriptional units resemble genomic elements lacking TATA boxes and Inr promoters found in lower eukaryotes as well as mammals (Carninci, Sandelin et al. 2006), suggesting that constitutive transcription may be the ancestral state common to all eukaryotes. In addition to protein-coding genes transcribed by RNA polymerase II, tRNA genes and rRNA genes in *Leishmania* are transcribed by RNA

polymerase III and RNA polymerase I, respectively, at defined initiator and terminator sequences (Das, Banday et al. 2008).

Transcriptional unit boundaries are enriched in histone acetylation marks: the H2A.Z and H2B.V marks have been functionally confirmed as essential, while H3.V is not required for normal transcriptional activity in both *Leishmania* (Thomas, Green et al. 2009, Anderson, Wong et al. 2013) and *Trypanosoma* (Siegel, Hekstra et al. 2009). While histone variants seem to have a conserved function in different kinetoplastids, other types of epigenetic modification appear to have a genus- or species-specific function. A hyper-modified base unique to kinetoplastid protozoa, glycosylated hydroxymethyluracil, also called base J, is enriched in telomeric regions, but is also present in strand switch regions in *Leishmania* (Genest, Ter Riet et al. 2007, van Luenen, Farris et al. 2012). This modified base is produced by a two-step hydroxylation and glycosylation reaction of thymine residues, and appears to serve a genome-wide function in *Leishmania* regulating RNA polymerase II transcription termination (van Luenen, Farris et al. 2012, Reynolds, Cliffe et al. 2014).

Gene expression studies have found a small number of genes differentially expressed between promastigote and amastigote stages. Global interspecies expression analyses have found that the majority of genes differentially regulated throughout development in one species, however, are not differentially regulated in others (Rochette, Raymond et al. 2008). In both *L. major* and *L. infantum*, only 7 to 9 percent of the genome is differentially expressed in these two life stages according to oligonucleotide microarray data (Rochette, Raymond et al. 2008). Interestingly,

up to 95 percent of the small set of genes found to be differentially expressed in *L. infantum* promastigotes are no longer differentially expressed if parasites are isolated from axenic culture as opposed to the sand fly midgut (Alcolea, Alonso et al. 2014), suggesting a significant bias introduced by *in vitro* culture. In the New World species *L. braziliensis*, only 9 percent of the genes are differentially expressed throughout the life cycle (Depledge, Evans et al. 2009). The small number of genes differentially expressed between promastigote and amastigote stages have been proposed to be a well-conserved pre-adaptation to intracellular survival.

Despite the small number of developmentally regulated genes, regulation of gene expression during development is thought to occur either post-transcriptionally, or by changes in epigenetic modifications affecting transcriptional activity. Changes in steady state transcript levels are thought to occur primarily via differences in the maturation and stability of individual mRNAs via interactions with RNA-binding proteins. Due to the small size of the sequence between the SL splice site and the start of the protein-coding sequence, *Leishmania* parasites have extremely short 5' UTRs. Regulatory sequences in the 3' UTR region of protein-coding genes have therefore been implicated in determining the stability of the trans-spliced mRNA transcript.

Translational and post-translational mechanisms determining steady state protein levels are also thought to be important in developmental regulation of *Leishmania* parasites (Bente, Harder et al. 2003). In particular, the increase in temperature associated with transmission from an insect vector to a mammalian host is thought to trigger a developmental programme via the activity of heat-shock

proteins such as HSP70, which act as chaperones and regulate production and maturation of proteins involved in the subsequent stress response.

1.3. A clonal, parasexual, or sexual organism?

1.3.1 The clonal theory

Since the 1990s, there has been ongoing debate as to the mating system and population structure of eukaryotic microbial pathogens such as *Leishmania*. These two biological aspects are closely linked to each other, and have important consequences for the epidemiology of transmissible diseases. The predominant view for many years has been that parasites such as *Leishmania* were primarily asexual, and that their population structure was essentially clonal (Tibayrenc, Kjellberg et al. 1990).

Support for the clonal theory has been based on the fact that several population genetic tests with a null hypothesis of panmixia, or random mating between individuals, have found significant deviations from Hardy-Weinberg expectations. A clonal population structure, defined by the predominance of a restricted mating system resulting in each generation being genetically identical to the generation that precedes it, is thus inferred as a result of statistical tests that reject the null hypothesis of unrestricted, random mating in the population. In this sense, asexuality and selfing via a sexual process may under most circumstances be virtually indistinguishable in population genetic data.

Several quantitative tests in *Leishmania* have yielded statistically significant results in this regard. These tests have focused on measures of linkage disequilibrium, or non-random associations between genotyped loci, and skewing of allele frequencies in the population due to prevailing homozygosity or heterozygosity (Tibayrenc and Ayala 2002). Qualitative observations have also been used in support of the clonal theory. These observations have focused on the effects one expects asexual reproduction and uniparental mating to have on the segregation and recombination of genetic markers in the population, and include considerations such as the widespread over-representation of identical genotypes, the absence of recombinant genotypes, the presence of fixed homozygosity or heterozygosity in a given population, and concordance in phylogenetic signal between independent sets of genetic markers (Ramirez and Llewellyn 2014).

In *L. tropica*, isoenzyme profiling of 27 isolates found significant heterogeneity and fixed heterozygosity (Le Blancq and Peters 1986). In addition, genotypes that should be segregating in natural panmictic populations, often cannot be found in *L. tropica* and several other Old World *Leishmania* species. The presence of ubiquitous genotypes such as the MON1 zymodeme in *L. infantum*, which is stable both across space and time (Rioux, Lanotte et al. 1990), has also been cited as evidence of predominant asexual reproduction. Species such as *L. infantum* that have a broad geographic range have been found to have extremely low genetic diversity and conserved linkage blocks by both MLEE and microsatellite MLST, with zymodemes such as MON1 being present across continents, but some sub-structuring detectable by MLST (Seridi, Amro et al. 2008, Kuhls, Alam et al. 2011). In

New World species of *Leishmania*, high concordance has been found between different typing methods, such as AFLP, MLEE, MLST and PFGE (Odiwuor, Veland et al. 2012). Both MLEE and RAPD data has been brought forward in support of linkage disequilibrium in the *Viannia* subgenus (Banuls, Jonquieres et al. 1999). Homogeneity between markers was also found in *L. donovani* in India (Alam, Kuhls et al. 2009).

1.3.2 Challenges to the clonal theory

Classic statistical tests employed in population genetics can only measure deviations from stated expectations. In the case of the clonal theory as originally proposed, quantitative and qualitative deviations from panmixia have been attributed to a predominance of self-mating (i.e. uniparental mating) and true asexual reproduction, with little differentiation between these two possible scenarios. However, in many instances rejecting the null hypothesis of panmixia doesn't in and of itself justify such broad conclusions. The original clonal theory has since been re-defined (Tibayrenc and Ayala 2012) to account for the distinction between self-mating and asexual reproduction, but some important issues remain.

Two main caveats need to be made explicit with respect to the body of evidence which has been used to argue in support of the clonal theory: these concern methodological limits due to sample size and choice of genetic markers, on one hand; and the theoretical models that may be invoked to explain trends in the population genetic data when panmixia is not observed, on the other.

Firstly, methodological issues may result in a statistically significant trend in the data, when in actuality there is none. It is a widely recognized fact in biology that inadequate sample sizes may yield artifacts due to insufficient power or to sampling bias, and it is beyond the scope of this introduction to analyse these aspects in detail. In addition, the choice of markers for MLST, AFLP, RAPD, and MLEE can significantly affect the results obtained. The fact that strains typed as belonging to the MON1 zymodeme break down into polyphyletic units when analysed with microsatellite MLST (Banuls, Hide et al. 1999) proves that high resolution markers can provide crucial additional information.

Secondly, a number of population processes may limit gene flow and recombination in sexually reproducing populations. Identifying the barriers to gene flow that may result in deviations from panmixia independently of biological features intrinsic to the parasite, however, is a challenging process. Such barriers may be posed by reproductive isolation due to geographic separation or to the presence of multiple, distinct transmission cycles, for example because of different vectors or reservoirs, even in the presence of an obligatory sexual life cycle. This has been called the Wahlund effect, and its defining feature is the detection of heterozygosity deficits as a result of substructuring of a population that is nonetheless following panmixia.

The first report of potential hybrids in natural populations concerned two inter-specific hybrids between *L. arabica* (now considered to be synonymous with *L. tropica*) and *L. major* that were isolated from zoonotic sources in Saudi Arabia, and were confirmed to be hybrids by a variety of analytical methods (Evans, Kennedy et

al. 1987, Kelly, Law et al. 1991). Putative hybrids between New World species of *Leishmania* were also found in Nicaragua (Belli, Miles et al. 1994), Ecuador (Banuls, Guerrini et al. 1997), and Venezuela (Delgado, Cupolillo et al. 1997). Natural inter-specific hybrids were later also found in Peru (Shani-Adir, Kamil et al. 2005, Nolder, Roncal et al. 2007). Hybrids between *L. major* and *L. infantum* were also reported in immunosuppressed patients in Portugal (Ravel, Cortes et al. 2006), and hybridization was found in *L. donovani* complex isolates from Turkey (Rogers, Downing et al. 2014).

The first experimental evidence of hybridization came from crosses with *L. major*, and clearly showed that *Leishmania* parasites have the capacity of undergoing sexual reproduction in the sand fly stages (Akopyants, Kimblin et al. 2009). Experimental hybrids have since also been generated in *L. donovani* (Sadlova, Yeo et al. 2011), and most recently between *L. infantum* and *L. major* (Romano, Inbar et al. 2014), indicating that cross-species hybrids may naturally occur and thus confirming previous reports of natural cross-species hybrids identified in the field. Indeed, a series of early microscope studies found evidence for cell fusion in *L. tropica* promastigotes (Lanotte and Rioux 1990). Later reports of nuclear fusion in amastigotes of several species of *Leishmania* as determined by DNA content (Kreutzer, Yemma et al. 1994) may have been detecting pre-existing aneuploids in the sampled cell population, rather than true 2*n* zygotes due to nuclear fusion events between gametes with *n* ploidy.

Several population genetics studies have since cast some doubt on the validity of the clonal theory as it was originally proposed. Using PFGE, recombinant

karyotypic variants were found in *L. infantum* (Blaineau, Bastien et al. 1992). Microsatellite MLST analyses of *L. donovani* complex populations found significant variation in inbreeding coefficients (Kuhls, Keilonat et al. 2007), and after expanding these analyses to New World *L. infantum* and *L. chagasi*, a range of inbreeding coefficients were found, reflecting possible Wahlund effects impacting the number of heterozygotes (Kuhls, Alam et al. 2011). In *L. braziliensis*, evidence for strong Wahlund effects was brought forward, with substantial heterozygote deficits and linkage disequilibrium due to sub-structuring of the population into several micro-foci of transmission (Rougeron, De Meeus et al. 2009), while in another New World species, *L. guyanensis*, very modest linkage disequilibrium was found in addition to an overrepresentation of homozygotes, suggesting substantial recombination (Rougeron, Banuls et al. 2011).

Focusing on the microgeographic scale, a mixed mode of reproduction was suggested for *L. donovani* in Sudan (Rougeron, De Meeus et al. 2011) and in Ethiopia (Gelanew, Kuhls et al. 2010). Recombinant genotypes were also found in *L. infantum* from Tunisia (Chargui, Amro et al. 2009). More evidence of a mixed mode of reproduction with an important role for recombination was offered by MLST analyses in the *Viannia* subgenus (Boite 2012, Kuhls 2013). When the local geographic scale of *Leishmania* transmission and the potential reproductive barriers present at this scale are disregarded, however, low linkage disequilibrium from MLEE or MLST markers can be misinterpreted as absence of recombination as a biological process, especially if the study is underpowered (El Baidouri, Diancourt et al. 2013).

In this dissertation, I follow the recommendations by Rougeron and colleagues (Rougeron, De Meeus et al. 2015), and advise against pooling of samples from different geographic areas and different time periods when making considerations regarding the mode of reproduction in *Leishmania*. I aimed to maintain a distinction between population genetics, which requires knowledge of the demographic units in this important human pathogen, the potential barriers to gene flow, and most importantly, access to a representative sample of individuals from each of these demographic units; and genomic analyses, which focus on genome-level processes at work in individual parasites, and allow accurate qualitative and quantitative comparative considerations to be made.

1.3.3 Models of asexuality, sexuality, and parasexuality

I chose to reject the misguiding terminology characteristic of the clonal theory, which focuses on patterns in the data rather than biological processes, and hereby describe three modes of reproduction that may be at work in *Leishmania*: asexual, sexual, and parasexual. Some overlap between these three modes is clearly possible, and as such they are not fully mutually exclusive, but for ease of reference I provide to the reader a codified description of each "model".

In the asexual model, parasites complete each transmission cycle, from human host to insect vector, without ever performing meiosis. Each cell reproduces mitotically, dividing by binary fission, and each daughter cell receives a full complement of *2n* heterologous chromosomes. Mosaic aneuploidy in this model

arises from unbalanced chromosome replication and/or unequal segregation of the chromosomes during mitotic division, whereby each daughter cell receives *Xn +/- m* chromosomes, where *X* is an integer, *n* is the full complement of heterologous chromosomes, and *m* is the absolute number of chromosomes in excess or in deficit with respect to the balanced ploidy *X*. Meiotic recombination is not observed, although homologous mitotic recombination may be possible under certain circumstances. The progeny can only inherit either the full set or a subset of the heterologous chromosomes present in the parental cell.

In the sexual model, the parasite performs meiosis as part of its transmission cycle, and generates haploid gametes that may or may not belong to separate mating types. Homologous meiotic recombination of the parental chromosomes occurs prior to segregation into the haploid gametes. Inexact cell division at any stage during meiosis may lead to gametes with chromosomes in excess or in deficit of the heterologous set *n*, thus resulting in aneuploidy of the progeny. Cell fusion thus occurs between compatible gametes, if mating types are indeed present, and gives rise to zygotes with *Xn* +/- *m* chromosomes that do not divide any further until mitosis is resumed.

In the parasexual model, there is no meiosis. Parasites perform cell fusion without prior generation of haploid gametes, and unequal segregation of the chromosomes in the daughter cells during subsequent cell division gives rise to progeny with a variable number of chromosomes *Xn* +/- *m*, that nevertheless successfully retain some combination of chromosomes from each parent. Homologous recombination may occur prior, during, or following cell fusion,

although it may affect only part of the genome. Distinct ameiotic mating types may still be present.

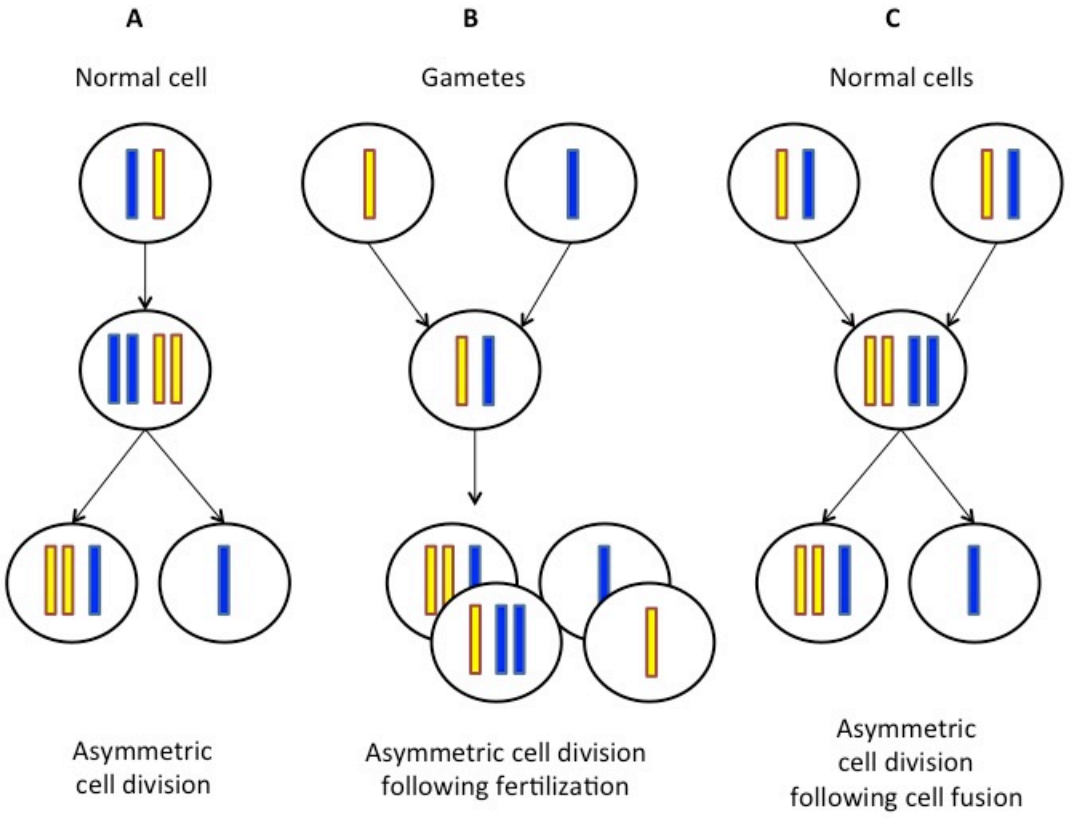


Figure 1.5. Schematic models of asexual, sexual and parasexual reproduction as referred to throughout this dissertation. (A) Asexual reproduction starts from normal somatic cells, which undergo chromosome replication and subsequent mitotic cell division. Asymmetric replication or asymmetric cell division may give rise to aneuploidy in the progeny. Homologous recombination is negligible. (B) Sexual reproduction starts with fertilization of gametes, possibly representing different mating types, which were

generated by a reductional meiotic process involving homologous recombination. Fertilization of the gametes gives rise to a zygotic form that may then generate aneuploidy daughter cells by asymmetric cell division in a mitotic process similar to that depicted in (A). (C) Parasexual reproduction starts with cell fusion of normal somatic cells, and then proceeds with reductional cell division that redistributes the chromosomes unevenly, thus generating aneuploid daughter cells. Homologous recombination may occur. (Source: adapted from Sterkers et al. 2014).

In both the sexual and parasexual models, cell divisions associated with a sexual or parasexual event may "reset" the ploidy each transmission cycle if there is balanced segregation of the chromosomes. The widespread mosaic aneuploidy reported in *Leishmania* may arise from subsequent inexact mitotic cell divisions, which may be called "somatic", rather than from the sexual or parasexual event itself.

Despite the many simplifications made in this section, I will refer back to this schematic representation of the sexual, parasexual, and asexual models throughout my dissertation as a useful way to make these discussions more accessible. In this simplified view, recombination plays a major role in shaping the *Leishmania* genome only in the parasexual or sexual models, although it may be present in all three models.

1.3.4 Aims and objectives

With this dissertation, I seek to address the open question of whether *Leishmania* parasites follow a sexual life cycle, and describe the changes that occur in their genome during hybridization. I describe how my findings fit with what we know about *Leishmania* genome biology, and with the models described in section 1.3.3. I then make observations on how genome plasticity and genetic exchange may affect parasite evolution in the field in response to selective pressures. Each chapter provides an overview of the experimental procedures and *in silico* analyses that I performed to address this central question in their respective Methods sections.

Chapter 2 describes the results from multi-locus sequence typing of 34 different isolates of *L. tropica* covering the entire geographic range of the species. This chapter then compares and contrasts the MLST data with whole-genome sequencing (WGS) data from some of these isolates, and makes some observations on how hybridization may have contributed to *L. tropica* population structure and genetic diversity.

Chapter 3 investigates in depth mechanisms of genome plasticity and their effect on gene expression via paired WGS and RNA-seq of a subset of 14 field isolates from the set of 34 discussed in Chapter 2. This set of clinical isolates was complemented with an additional 6 isogenic lines obtained by cloning of 4 of these isolates, with the aim of understanding the effects of mosaic aneuploidy on transcription.

Chapter 4 describes generation of transgenic single-drug resistant lines, sand fly feeding assays, and selection of double-drug resistant hybrids in sand fly lab-adapted colonies, and the implications of hybrid recovery rates from infected sand flies in laboratory crosses for estimating the frequency of hybridization in a natural setting.

Finally, Chapter 5 describes genetic exchange in experimental hybrids, and provides an exhaustive list of *de novo* mutations, recombination, structural rearrangements, and Mendelian violations in the inheritance of the genetic material of the two parental lines. I provide the first high-resolution, complete description of the effects of hybridization on genome sequence and structure in *L. tropica*.

In the Conclusions, I provide a summary of my findings and how they have shed new light on our current understanding of the processes generating and maintaining genetic variation in this important pathogen.

I hope that this thesis will contribute to the open debate regarding the predominant mode of reproduction in *Leishmania*. The work presented here provides for the first time experimental evidence in support of sexual or parasexual reproduction in the Old World species *L. tropica*, and on this basis I strongly reject the asexual model. Although a rigorous disqualification of either sexual or parasexual reproduction in favour of one or the other is not warranted by the present data, I provide to interested readers some indication of how future studies informed by our work can solve this debate once and for all.

CHAPTER 2

POPULATION GENETICS IN *L. TROPICA*

2.1 Introduction

As discussed in the introduction, *L. tropica* is one of approximately 21 species of *Leishmania* known to be pathogenic to humans. It is increasingly recognized as an important anthroponotic species, responsible for both cutaneous and viscerotropic disease throughout its range in Northern Africa, the Middle East, and India. The extent to which many pathogens, including kinetoplastid parasites such as *Leishmania*, undergo "sexual" as opposed to "clonal" reproduction is currently unresolved. The classic view holds the population structure of these parasites to be mostly clonal, but a growing body of evidence suggests that genetic exchanges due to intra- and inter-specific hybridization events might be an important process driving the evolution of these parasites.

Tibayrenc and colleagues (Shani-Adir, Kamil et al. 2005, WHO 2015) were the first to put forward the clonal model, suggesting that very little to no genetic exchange occurred between each clonal lineage, represented by "types" in the model parasite species *Toxoplasma gondii*, and by "species" or smaller demographic units at the subspecies or subpopulation level in *Leishmania*. The model proposes clonal evolution to be the main mode of evolutionary change in parasite populations, generating large deviations from Hardy-Weinberg equilibrium and extreme linkage

disequilibrium, as evidenced by the stable inheritance of haplotypic "blocks" or "units" of linked polymorphic markers across both space and time. Isoenzyme, RAPD, and RFLP studies of Bolivian *Trypanosoma cruzi* provide a defining paradigm for this model (Crowley, Zhabotynsky et al. 2015, Dillon, Okrah et al. 2015). Consistently, in this kinetoplastid species, diagnostic zymodemes appear to be stable across large geographical areas, and genotyping at one locus appears to be a reliable predictor of the genotype at a second linked locus. The absence of recombinant genotypes and genotypes that could result from co-segregation supports the hypothesis that populations of this parasite evolve without a consistent "shuffling" of their genetic material at each generation. Genetic exchanges, if they occur, are relatively rare events, and are not sufficient to break the predominant mode of "clonal" evolution. Importantly, this model does not preclude the presence of both natural selection and genetic drift acting on parasite allele frequencies in natural populations.

Leishmania spp. have historically been classified into distinct species based on a range of characteristics, including host preferences, transmitting vector species, and presentation of disease. Many recent studies have questioned the reliability of such taxonomy, in light of mounting evidence that both inter and intra specific hybrids do occur in nature (Belli, Miles et al. 1994, Banuls, Jonquieres et al. 1999, Ravel, Cortes et al. 2006, Nolder, Roncal et al. 2007) and that hybridization has been experimentally demonstrated in laboratory co-infections of different species of sand flies with *L. major*, *L. infantum*, and *L. donovani* (Akopyants, Kimblin et al. 2009, Sadlova, Yeo et al. 2011, Inbar, Akopyants et al. 2013, Romano, Inbar et al. 2014).

A more accurate taxonomy may be garnered by studies that consider a large number of molecular markers. Multi-locus enzyme electrophoresis (MLEE) has been for many years the single most widely used tool for strain characterization. Multi-locus sequence typing (MLST), also known as multi-locus sequence analysis (MLSA), has risen in recent years as a complementary approach to resolve species relationships within the *Leishmania* genus, especially if a large number of both markers and parasite samples is used, leading for instance to the re-definition of species that cause VL (*L. donovani*, *L. infantum*, *L. archibaldi*) and their grouping into a "*L. donovani* complex" of closely related taxa (El Baidouri, Diancourt et al. 2013, Wang, Wang et al. 2015).

Previous studies have suggested the presence of at least occasional genetic exchange in *L. tropica* as evidenced by microsatellite population structure (Schwenkenbecher, Wirth et al. 2006, Krayter, Bumb et al. 2014) and MLSA of nuclear markers (El Baidouri, Diancourt et al. 2013). These reports also indicate that extensive heterozygosity may be present in this species, possibly as a result of outcrossing. Schwenkenbecher and colleagues (Schwenkenbecher, Wirth et al. 2006) found considerable diversity in microsatellite markers across the range of the species, and suggest that a recombinant line may be propagating through Asia as a result of an hybridization event between two African strains. El-Baidouri and colleagues (El Baidouri, Diancourt et al. 2013) found evidence for intergenic recombination among housekeeping genes in *L. tropica* isolates from Morocco, Tunisia, and Kenya by linkage analysis, and by observing rearrangement of maximum likelihood tree topologies for genotypes at several different markers.

Krayter and colleagues (Krayter, Bumb et al. 2014) compare 8 Indian isolates of *L. tropica* associated with CL from human cases in Bikaner City, Northwestern India, with 156 isolates from the rest of the geographic distribution of the species, including some known human cases of VL due to *L. tropica*. They find three major populations, one in the broader Africa and Galilee region, one in Palestine and Israel, and one across most of Asia and India. High levels of heterozygosity in the latter population is consistent with outcrossing being present in the region, while the African samples appear to be associated with significant inbreeding.

In a representative set of 34 isolates covering the entire geographic range of *L. tropica*, 25 nuclear markers and 3 kinetoplast DNA markers were amplified and sequenced. MLSA was performed to sample the genetic diversity present within this species, and obtain estimates of observed heterozygosity and expected heterozygosity under the Hardy-Weinberg assumptions of random mating in the samples of this study. I then compare our MLSA results with whole-genome sequence data of 18 of these isolates.

2.2 Methods

2.2.1 Culturing and DNA extraction

A total of 14 isolates that were categorized as *L. tropica* based on zymodeme typing were selected from the collection at the US National Institutes of Health in the NIAID Laboratory of Parasitic Diseases to encompass the whole range of the

species (Table 2.1). Clinical history and other circumstantial information associated with each isolate varied based on the collector and the year of collection. Samples were from 12 different countries, and their year of collection ranged from 1958 to 2009. Although not an ideal sample for detailed population genetics analysis, this sample provides a high level perspective of overall genetic diversity in this species.

All isolates had been previously culture-adapted and preserved as axenic promastigotes in freezing solution (7.5% DMSO, 10% FBS in DMEM) at -60 °C in liquid nitrogen storage. Parasite promastigote stages were thawed in a water bath at 37 °C, washed in buffered RPMI, and resuspended in complete medium 199 (cM199) supplemented with 20% heat inactivated FCS, 100 U/mL penicillin, 100 ug/mL streptomycin, 2mM L-glutamine, 40mM Hepes, 0.1 mM adenine (in 50mM Hepes), 5 mg/mL hemin (in 50% triethanolamine), and 1 mg/mL 6-biotin. Each promastigote culture was then kept at 26 °C in a humidified incubator, and examined daily with a light microscope until parasites reached a density sufficient for DNA extraction. DNA extraction was performed using the QIAgen DNeasy Blood and Tissue kit following the manufacturer's guidelines. DNA concentration for both PCR and whole-genome sequencing was assessed using two alternative methods, by Nanodrop spectrophotometry, and confirmed by gel imaging and band intensity analysis with ImageJ image processing software. DNA samples were made available by Dr Michael Grigg for an additional 20 isolates, for which no frozen stock however could be obtained.

Isolate	Species	Zymodeme	Origin	Path.	WHO code	WGS	Stock
AM	*L. tropica*	NT	Tunisia	CL	MHOM/TN/06/AM		
CJ	*L. tropica*	NT	Tunisia	CL	MHOM/TN/06/CJ		
Killicki	*L. tropica*	MON8	Tunisia	-	MHOM/TN/80/LEM163		
Leep0920	*L. tropica*	NT	Lybia	CL	MHOM/LY/09/Leep0920		
LA28	*L. tropica*	LON16	Greece	CL	MHOM/GR/LA28		
MON497px3	*L. tropica*	MON497	Greece	-	MCAN/GR/82/MON497px3		
Tropica57	*L. tropica*	-	Palestine	-	MHOM/PS/01/ISL593		
Tropica63	*L. tropica*	-	Palestine	-	MHOM/PS/01/LRC-L838	X	
Tropica75	*L. tropica*	-	Palestine	-	MHOM/PS/02/34JnF4		
Rachnan	*L. tropica*	LON12 MON60	Israel	CL	MHOM/IL/78/Rachnan		
Gabai	*L. tropica*	LON9	Israel	-	MHOM/IL/Gabai159		
Ackerman	*L. tropica*	-	Israel	-	-	X	X
LRC-L747	*L. tropica*	-	Israel	-	MHOM/IL/02/LRC-L747	X	X
LRC-L810	*L. tropica*	-	Israel	-	MHOM/IL/00/LRC-L810	X	X
E50	*L. tropica*	-	Israel	-	-	X	X
MA-37	*L. tropica*	-	Jordan	-	MHOM/JO/94/MA37	X	X
MN-11	*L. tropica*	-	Jordan	-	-	X	X
Kubba	*L. tropica*	-	Syria	-	-	X	X
Melloy	*L. tropica*	-	Saudi Arabia	-	MHOM/SA/91/ML	X	X
Boone	*L. tropica*	-	Saudi Arabia	-	MHOM/SA/91/BN	X	X
ASinaiIII	*L. tropica*	LON11	Iraq	LR	MHOM/IQ/73/ASinaiIII		
BAG17	*L. tropica*	LON24	Iraq	CL	MHOM/IQ/73/BAG17		
BAG9	*L. tropica*	MON53	Iraq	CL	MHOM/IQ/76/BAG9		
Bum30	*L. tropica*	LON17	Iraq	VL	MHOM/IQ/73/Bumm30		
AdhanisI	*L. tropica*	MON5 LON15	Iraq	-	MRAT/IQ/73/AdhanisI	X	
L75	*L. tropica*	MON6 LON14	Iraq	CL	MHOM/IQ/65/L75	X	
SAF-K27	*L. tropica*	MON60 LON12	Ex-USSR	CL	MHOM/SU/74/SAF-K27	X	
WR683	*L. tropica*	-	Ex-USSR	-	MHOM/SU/58/WR683		
IIIKK	*L. tropica*	-	Afghanistan	-	MHOM/AF/88/KK27	X	X
Rupert	*L. tropica*	-	Afghanistan	-	MHOM/AF/87/RP	X	X
Azad	*L. tropica*	-	Afghanistan	-	MHOM/AF/82/AZ	X	X
DBKM	*L. tropica*	MON62 LON21	India	-	MCAN/IN/71/DBKM		
188	*L. tropica*	-	India	-	MHOM/IN/90/K26	X	X
311W	*L. tropica*	-	India	-	MHOM/IN/91/K112	X	X

Table 2.1. A list of the 34 isolates of *L. tropica* isolates that were used in this study. "WGS" means the strain has been whole genome sequenced by the Wellcome Trust Sanger Institute and deposited in the publicly accessible at the European Nucleotide Archives at the European Bioinformatics Institute (EBI), Hinxton (See Appendix A for all ENA accession numbers). "Stock" means that that we had access to a frozen parasite stock for that isolate. "Path." indicates pathology, when this is known.

2.2.2 Multi-locus sequence typing of field isolates

A panel of 28 different nuclear markers coding for housekeeping genes were selected to include loci on several of the 36 chromosomes observed in Old World *Leishmania* species (chromosomes 4, 5, 9, 10, 12, 14, 15, 18, 22, 24, 27, 29, 30, 31, 32, 34, 35 were covered by our panel of nuclear markers, see Table 3.2 for more information on each marker). The primer oligomer sequences were kindly provided by Dr Mourhad Barhoumi in the lab of Dr Michael Grigg at NIH. Primer pairs had been designed to amplify the same genetic locus in a number of *Leishmania* species, including *L. tropica* (Table 2.2). These primer pairs were used to genotype the 14 isolates of *L. tropica* for which a frozen stock was available. To sample more widely the genetic diversity in this species, the same primers were used to genotype 20 additional isolates for which only DNA but no frozen stock was available. In addition, 3 maxicircle kDNA markers were chosen to obtain genetic information on the parasite kinetoplast. No primers targeting the minicircle were included in this analysis.

Marker	Chromosome	Genomic location in *L. major* (Gene, start-end)	Gene product
4_80	4	LmjF.04.0070, 29378-30900	Hypothetical protein, conserved
4_360	4	LmjF.04.0380, 124100-125200	Hypothetical protein, conserved
5_180	5	LmjF.05.0180, 54000-55000	Dihydrolipoamide transacylase
5_830	5	LmjF.05.0830, 303500-304500	Methylthioadenosine phosphorylase
5_1210	5	LmjF.05.1215, 441500-442500	Surface antigen-like protein
9_740	9	LmjF.09.0740, 288100-289100	Ubiquitin ligase
10_icd	10	LmjF.10.0290, 130280-131280	Isocitrate dehydrogenase precursor
12_PGI	12	LmjF.12.0530, 293800-294700	Glucose-6-phosphate isomerase
14_NH2	14	LmjF.14.0130, 32470-33230	Inosine-guanine nucleoside hydrolase
15_810	15	LmjF.15.0770, 349700-350700	Protein kinase, putative
18_iunh	18	LmjF.18.1580, 706350-707350	Nonspecific nucleoside hydrolase
22_700	22	LmjF.22.0870, 354900-355900	Hypothetical protein, conserved
24_me	24	LmjF.24.0770, 271330-272330	Malic enzyme
27_350	27	LmjF.27.0340, 89680-90460	Editosome component MP44
27_870	27	LmjF.27.1010, 435500-436500	Hypothetical protein, conserved
27_2335	27	LmjF.27.2335, 979900-980600	Hypothetical protein, conserved
27_ITS	27	NA, 991987-992269	ITS rRNA
29_FH	29	LmjF.29.1960, 854880-855880	Fumarate hydratase
30_3800	30	LmjF.30.3740, 1396750-1397430	Ribosomal P protein AGP2 beta -1
31_AQP1	31	LmjF.31.0020, 7700-8700	Aquaglyceroporin
32_mpi	32	LmjF.32.1580, 621890-622890	Phosphomannose isomerase
34_g6pdh	34	LmjF.34.0080, 26530-27530	Glucose-6-phosphate 1-dehydrogenase
35_asat	35	LmjF.35.0820, 380109-381073	Aspartate aminotransferase
35_2160	35	LmjF.35.2125, 871300-872300	Hypothetical protein, unknown
35_GND	35	LmjF.35.3340, 1363367-1364500	6-phosphogluconate dehydrogenase
K_12sRNA	kDNA	Lt_X02354.1, 438-1610	12s rRNA
K_cytB	kDNA	Lt_X02354.1, 5403-6481	Cytochrome B
K_ND5	kDNA	Lt_X02354.1, 14924-16696	NADH dehydrogenase 5

Table 2.2. Marker panel comprising the 28 nuclear and kDNA loci that were considered for this study. Primer pair sequences were designed to amplify the same genomic locus across all *Leishmania* species, starting from the *L. major* reference genome and selecting the most highly conserved regions.

Polymerase chain reaction (PCR) was performed with Taq DNA polymerase (Sigma Aldrich D1806) in PCR buffer supplemented with 2mM dNTPs and $MgCl_2$ using a primer concentration of 25pM. The annealing temperature in the thermocycling protocol used was either 55 or 58 °C, depending on the primer pair.

The PCR product so obtained was imaged on a gel electrophoretic apparatus and free-floating nucleotides and excess primer oligomers were removed by incubation at 37 °C for 15 minutes with ExoSAP-IT nuclease (Affymetrix, product no. 78205), followed by incubation at 80 °C for 15 minutes to deactivate the enzyme. The clean PCR product was then sequenced with the Sanger dideoxy-chain termination method at the NIAID Rocky Mountain Laboratories in Hamilton, Montana.

Sequence data was obtained from all 34 isolates of *L. tropica* (See Table 2.1). The sequences thus generated were aligned marker-by-marker in LaserGene SeqMan software. All chromatographs were visually checked and sequences were manually trimmed. Biallelic heterozygous SNP calls were assigned whenever overlapping peaks of the same height were observed in the resulting chromatograph, and were labelled with the IUPAC nomenclature for ambiguous base calls (K = G or T; M = A or C; R = A or G; Y = C or T; S = C or G; W = A or T). Each sequence was run against the NCBI nucleotide database through a BLAST search to confirm the species identity as *L. tropica*, and edited sequences were then aligned using the Clustal W2 algorithm and further processed with R statistical analysis language (R Development Core Team 2009) using the *seqinr* (Charif and Lobry 2007), *ape* (Paradis 2004), and *adegenet* (Jombart 2008) package implementations.

The gametic phase (i.e. the haplotype sequence of each parental allele at biallelic loci) was estimated based on occurrence of any putative homozygous parental alleles for the marker examined in any of the other isolates in our sample set. An allelic plot was generated and color-coded based on allele sharing from this estimation. Neighbour-joining (NJ) phylogenetic trees were generated for each

unphased sequenced marker using pairwise distance matrices as input to guide assignment of putative parental alleles (Figure 2.1); concatenated sequences for all markers with complete sequence information (i.e. markers that yielded a high quality sequence for all 34 samples) were also aligned and used to generate a dendrogram, the branching order of which was determined by hierarchical clustering, and then used to organize the allelic plot information (Figure 2.3). The number of different "sequence types", i.e. matching concatenated sequences between isolates, was calculated.

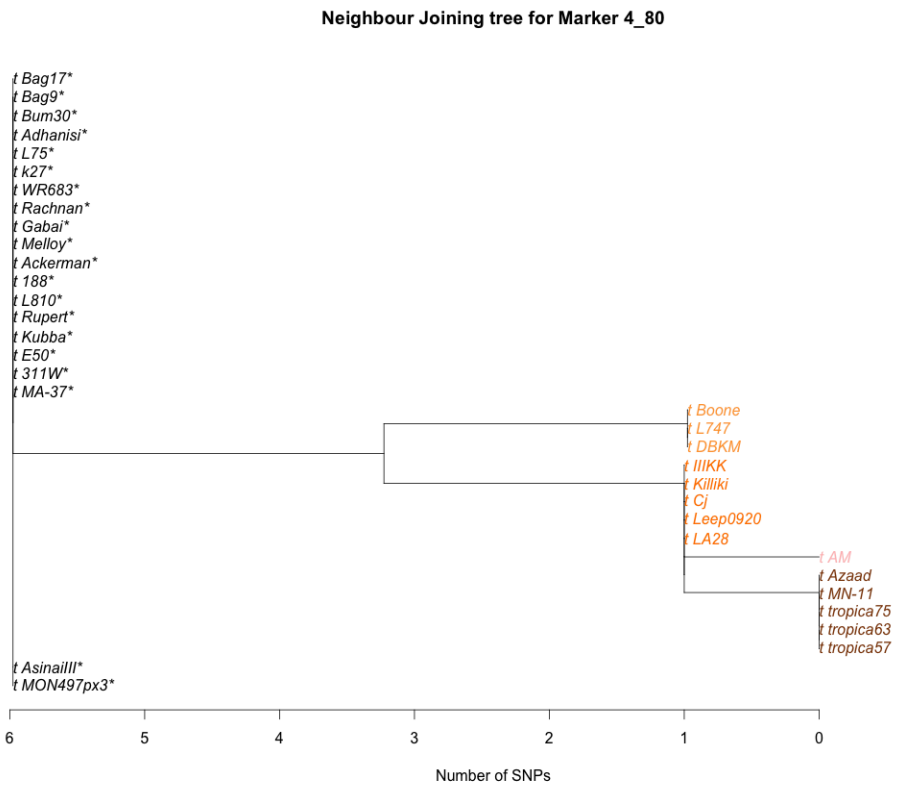


Figure 2.1. Example of a neighbour–joining (NJ) gene tree used in the assignment of individual parental alleles to heterozygous genotypes for allelic plot reconstruction. The marker shown above is 4_80 on chromosome 4. All isolates in black marked with an asterisk share the same identical heterozygous genotype (the entire sequence for the marker containing multiple SNPs was considered). The color-coded isolates at each branch tip share the same identical homozygous genotype. This marker therefore had 4 different alleles (in non-black colors in the tree), one heterozygous genotype, and no orphan alleles, since the heterozygous sequence matched two putative

parental alleles, the first seen in isolates Boone, L747, DBKM, and the second seen in isolates IIIKK, Killicki, CJ, Leep0920, and LA28. Isolates AM, Azad, MN-11, and tropica 75, 53 and 57 had homozygous genotypes not seen in heterozygous form in any of the other isolates.

2.2.3 Whole-genome sequencing of field isolates

Extraction of DNA from *in vitro* promastigote cultures of the 14 isolates with a frozen stock available was performed as described in Section 2.2.2, and the DNA amounts were measured with a Nanodrop spectrophotometer and band intensity analysis by gel imaging using a DNA ladder of known concentrations. The DNA was then quantified with an intercalating dye using the Qbit system, before being sequenced on the Illumina HiSeq 2500 platform by the sequencing operations staff at the Wellcome Trust Sanger Institute. Sequences were deposited in publicly accessible repositories at the European Nucleotide Archives (See Appendix A for ENA Accession Numbers). Each paired-end library had an average insert size of 500 base pairs, and was multiplexed over two lanes to maximize coverage, and sequenced for 100 cycles. The sequence data for 4 additional isolates (Adhanis I, L75, L838, and SAF-K27) that had previously been sequenced at the Wellcome Trust Sanger Institute (see Table 2.1) was kindly made available by Dr Gabi Schonian.

Short read sequence data was then mapped to a draft reference genome for *L. tropica* using SMALT with a sequence match threshold of 80% in parallel batches of 100000 reads, using a 13-kmer seed. The reference genome used for this study was

generated from the assembly v2.0.2 supercontigs for isolate L590, kindly provided by Dr Wes Warren and Dr Stephen Beverley at the Washington University Genome Institute. Briefly, these were scaffolded against the GeneDB release of the *L. major* Friedlin reference genome using ABACAS v2.0 (Crowley, Zhabotynsky et al. 2015), with minimum alignment length of 500bp and at least 85% identity. These parameters were empirically determined to maximise the total length of the *L. tropica* assembly that could be scaffolded. This version of the reference genome is very similar to the one currently available on TriTypDB (26 June 2015).

Variant calls for SNPs and small indels were made using the Genome Analysis Toolkit (GATK v.3.4-0) available from the Broad Institute. The variant calls were made with the UnifiedGenotyper algorithm, and any variants that were supported by reads spanning deletions or with low mapping quality were filtered out. High quality biallelic SNPs passing these strict filter settings were considered for subsequent analyses.

2.2.4 Statistical analyses

A discriminant analysis of principal components (DAPC) as implemented in the R package *adegenet* was carried out on the concatenated sequence data for markers that had complete sequence information in our sample set. Briefly, when population clusters are not known, DAPC seeks to find the number of clusters *k* that best fit the principal component-transformed data, and calculates a Bayesian Information Criterion (BIC) value for each successive *k* via the *k*-means algorithm (Jombart,

Devillard et al. 2010). The number of clusters *k* associated with the lowest BIC is the best fit for the data. This procedure maximizes between-cluster variation, while minimizing within-cluster variation.

Expected heterozygosity, defined as $2pq$ in the Hardy-Weinberg formula $p^2 + 2pq + q^2 = 1$, where p and q are the allele frequencies of a biallelic disomic locus, was calculated for each marker and averaged over all genotyped markers. This average was compared with the heterozygosity observed in the samples, which is simply the average number of markers that had heterozygous genotypes.

Heterozygosity deficiency with respect to Hardy-Weinberg expectations was quantified by estimating the inbreeding coefficient F_{IT} for each individual isolate, compared with the total population. The inbreeding coefficient F_{IT} and the frequency of a given allele p_i in the population are defined by the equation $F_{IT} + (1 - F_{IT})(\Sigma_i p_i^2)$ where *i* is the number of different loci to be considered and p_i is the allele frequency at that locus. The mean inbreeding coefficient, which is equal to the probability of an individual inheriting two identical alleles from a single ancestor, was calculated based on random sampling (N = 30) of the probability density function for that individual using the likelihood-based *inbreeding* function in the R package *adegenet* (Jombart 2008).

A similar analysis was performed starting from the high quality SNP data from the whole genome sequencing, obtained following the workflow described in Section 2.2.3, and the results were compared and contrasted. All trees were generated using the *ape* package in R (Paradis, Claude et al. 2004).

2.3 Results

2.3.1 MLST of field isolates

Sequencing of PCR amplified products from DNA of isolates with a frozen stock gave a total of 281 marker sequences, each composed by both forward and reverse strands (total n = 562). Of the 28 markers comprising our panel, 5 failed to amplify and sequence for all isolates (10_icd, 18_iunh, 34_g6pdh, 27_its, and k_ND5); 3 markers failed to amplify and sequence only for some isolates (k_12sRNA, 5_1210, 31_AQP1) and were therefore retained for subsequent analyses; 4 markers were amplified and sequenced for all isolates, but due to the highly polymorphic nature of these loci and the computational difficulty of estimating parental alleles (which increases by permutation for each heterozygous SNP following the rule $m = 2^n$, where n is the number of heterozygous SNP on a disomic chromosome, and m is the number of possible phasing solutions), these were excluded from further analysis (35_2160, 35_GND, 24_ME, 35_ASAT). A total of 18 markers were therefore considered for the subsequent analyses.

These sequences were aligned to 350 additional sequences that had been generated from the 20 isolates of *L. tropica* without a frozen stock by Dr Mourhad Barhoumi. A total of 631 sequences were then aligned and visually inspected marker-by-marker in LaserGene Seqman software to manually validate all homozygous and heterozygous SNP calls.

The marker 31_AQP1 on chromosome 31 showed signs of tetraploidy (Figure 2.2), with many peaks in the resulting chromatograph having a distinctive 3:1 height ratio. These can be attributed to heterozygous polymorphic positions, with one variant being present on three of the four chromosome copies. This hypothesis is consistent with previous findings that this chromosome is tetrasomic in *Leishmania* (Akopyants et al 2009, Rogers et al 2011). The large number of SNPs in this gene precluded estimation of the gametic phase of heterozygous genotypes, and was therefore excluded from our analyses, reducing the total number of markers to 17.

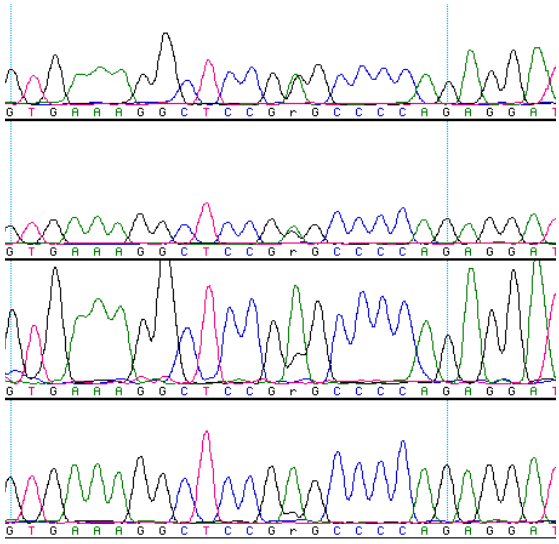


Figure 2.2. Aligned sequences for the 31_AQP1 marker for four of the isolates in our sample set. Notice the base call "r" (R in IUPAC nomenclature stands for A or G), which was used to manually annotate heterozygous positions. The presence of a green trace (nucleotide A) that is three times the size of the black trace (nucleotide G) at this position seems to suggest tetrasomy of

chromosome 31. In the first two samples, the two peaks are equal height, suggesting equal amounts of nucleotide A and G, which can be explained by a 2:2, if tetrasomic, or 1:1, if disomic, dosage of each nucleotide. In this case the heterozygous position is non-informative with respect to somy.

A total of 578 genotypes were thus obtained from 17 markers and 34 samples. Nine of the markers considered for this study lie on the same chromosome and are therefore "linked" (4_80 and 4_360, on chromosome 4; 5_830 and 5_1210 on chromosome 5, 27_350, 27_870, and 27_2335 on chromosome 27; k_cytB and k_12sRNA on maxicircle kDNA). Isolates that were heterozygous at one of these markers were also heterozygous at all linked markers on chromosome 4. Chromosomes 5 and 27 had conflicting patterns of heterozygosity at linked markers. Here, 5 isolates of *L. tropica* (Ackerman, 188, L810, Azad, MA-37) were heterozygous at one of the two markers on chromosome 5, but homozygous at the linked locus. Three more isolates (LA28, DBKM, Boone) were heterozygous at one of the three markers on chromosome 27, but homozygous at the other two linked loci, or vice versa (homozygous at one locus, but heterozygous at the remaining two). In these 8 isolates we mentioned, we therefore observed heterozygous markers in linkage with homozygous markers on at least one chromosome. All other isolates were consistently either homozygous or heterozygous at all linked markers.

A total of 101 haplotype allele sequences were identified, and 28 different heterozygous genotypes were detected. Heterozygosity per marker varied from 0 to 0.6176 in our sample set, with an average observed heterozygosity (H_0) of 0.4498. If

the population was panmictic and in Hardy-Weinberg equilibrium, the average expected heterozygosity (H_e) across all markers would be 0.6400, indicating a slight heterozygote deficiency in our sample set. Six markers had so-called "orphan alleles", defined as putative parental alleles of heterozygous genotypes for which a homozygous match could not be found in our set of isolates. The two markers with the most orphan alleles were 14_NH2 and 29_FH, with 5 orphan alleles each.

Marker	Allelic diversity, A	Heterozygous genotypes	Orphan alleles	Observed heterozygosity, H_0
4_80	4	1	0	0.5882
4_360	5	1	0	0.5882
5_830	5	1	0	0.5588
5_1210	8	4	3	0.5
9_740	6	3	2	0.4706
12_PGI	4	1	0	0.5882
14_NH2	10	4	5	0.5588
15_810	6	2	1	0.5882
22_700	6	0	0	0
27_350	5	1	0	0.5588
27_870	3	1	0	0.5588
27_2335	4	1	0	0.5588
29_FH	9	5	5	0.4117
30_3800	8	4	2	0.6176
32_mpi	6	1	0	0.5
K_cytb	5	0	0	0
K_12sRNA	7	0	0	0
Total	**101**	**28**	**18**	**Avg = 0.4498**

Table 2.3. Allelic diversity and heterozygosity of the 17 markers examined in 34 isolates of *L. tropica*. Allelic diversity (A) is defined as the number of alleles that are segregating in either homozygous or heterozygous form in our set of samples. The number of heterozygous genotypes for a marker includes genotypes with orphan alleles (see main text for definition). The observed heterozygosity (H_0) is simply the proportion of genotypes that are heterozygous for that marker. The expected heterozygosity (H_e) for diploid

organisms is defined as 2pq in the Hardy Weinberg formula $p^2 + 2pq + q^2 = 1$. The average expected heterozygosity across all markers was 0.64.

Our allelic analysis confirmed previous reports of extensive variation in patterns of heterozygosity and homozygosity in this species. The isolates seemed to fall into two categories, being either broadly homozygous, or broadly heterozygous, as exemplified by the allelic plot (Figure 2.3). Some variation at the sequence level was observed, and considerable allele sharing was evident between isolates. "Mixed" genotypes, or genotypes that matched different isolates, were observed for some kDNA markers (e.g. isolate Kubba), raising the possibility of intergenic recombination on the kDNA maxicircle. As expected, no heterozygous genotypes were observed at the kDNA markers. Isolate L810 appeared to be extremely divergent, having a unique sequence at most nuclear markers examined, although the same was not true for the kDNA markers, which were more highly conserved across isolates suggesting different evolutionary pressures acting on the nuclear and kinetoplast genomes.

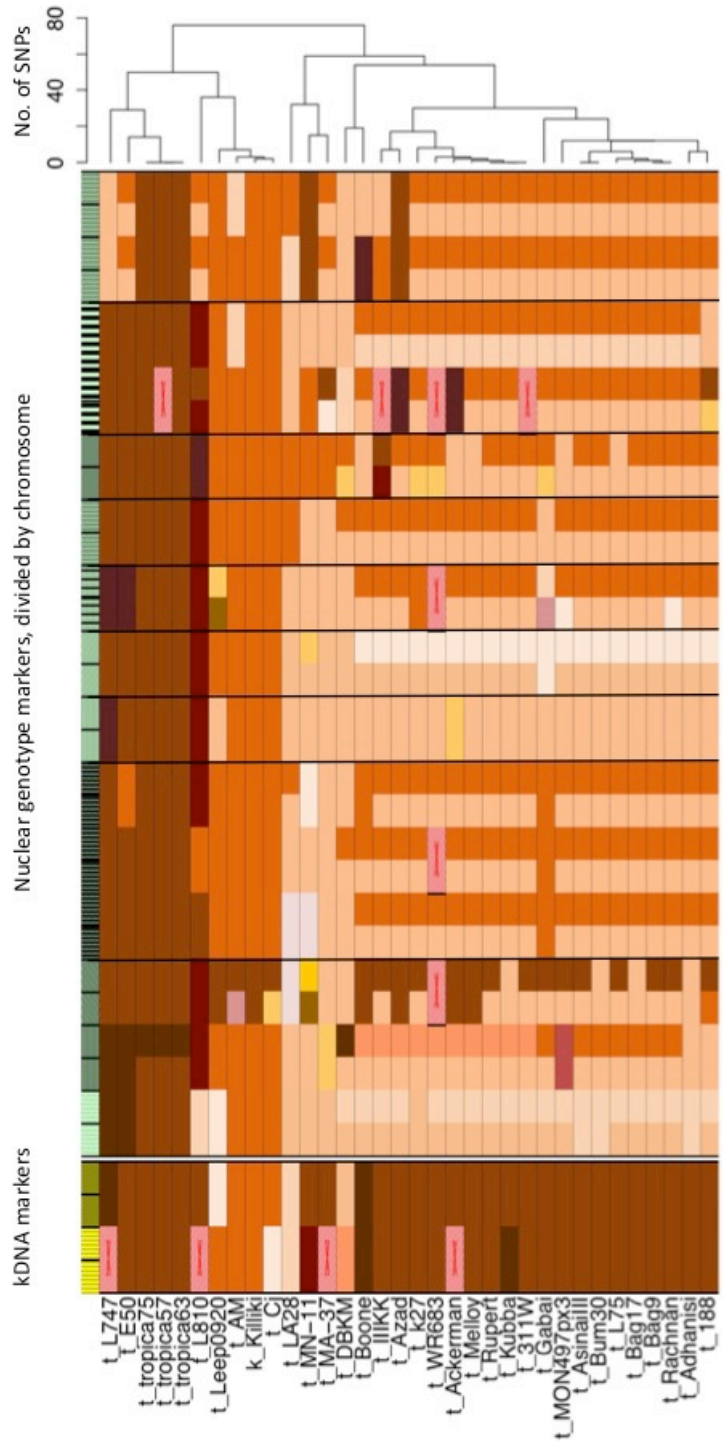

No. of SNPs
0
40
80
Nuclear genotype markers, divided by chromosome
kDNA markers
t_L747
t_E50
t_tropica75
t_tropica57
t_tropica63
t_L810
t_Leep0920
t_AM
k_Killiki
t_Ci
t_LA28
t_MN−11
t_MA−37
t_DBKM
t_Boone
t_IIIKK
t_Azad
t_k27
t_WR683
t_Ackerman
t_Melloy
t_Rupert
t_Kubba
t_311W
t_Gabai
t_MON497px3
t_Asinaiili
t_Bum30
t_L75
t_Bag17
t_Bag9
t_Rachnan
t_Adhanisi
t_188

Figure 2.3. Allelic plot of 34 isolates of *L. tropica* at 17 nuclear and kDNA markers. Each column corresponds to an isolate. Nuclear markers are marked with vertical minty green bars on the left hand side of the plot, each marker corresponding to a horizontal row, and are organized from higher to lower chromosome number (from 32_mpi on chromosome 32, in the top row, to 4_80 on chromosome 4, in the bottom row; the markers used are as in Table 2.3). Markers that lie on the kDNA maxicircle are grouped separately at the bottom of the plot, and are marked with yellow bars. Missing sequences are coloured red/pink in the plot. Heterozygous genotypes are colour-coded to show the putative parental alleles, with each half-cell per marker coloured based on the parental allelic contribution. Isolates are hierarchically clustered based on similarity in their concatenated sequences, represented by an ultrametric dendrogram matching that shown in Figure 2.4, with number of SNPs shown on the vertical axis.

Analysis of the concatenated sequence data for the markers with complete data gave 26 different "sequence types". The majority (n = 22) of the concatenated sequences were unique to each isolate. The two most common "sequence types" were characteristic of 3 isolates from Palestine (tropica75, tropica63, tropica57), and 3 isolates from Syria, Palestine, and India (Rupert, Kubba, 311W, respectively). Two more "sequence types" were represented more than once in the data, each being characteristic of two different isolates (Figure 2.4).

In order to quantify the heterozygosity deficiency of each individual compared to the total population the mean inbreeding coefficient F_{IT} was calculated for each of the isolates. The F_{IT} estimates were skewed toward either high (>0.50) or low (<0.50) values, with the largest group of isolates (n = 12) falling between 0.1 and 0.2, suggesting that despite the slight heterozygote deficiency at the level of the whole population compared to panmictic Hardy-Weinberg expectations, the majority of the isolates were heterozygous at the markers considered. The histogram of F_{IT} values appeared to be bimodally distributed (Figure 2.5), suggesting that a panmictic model may not fully explain the data and that there may be important barriers to gene flow leading to inbreeding in certain geographical locations. The F_{IT} values for all isolates are reported in Appendix B.

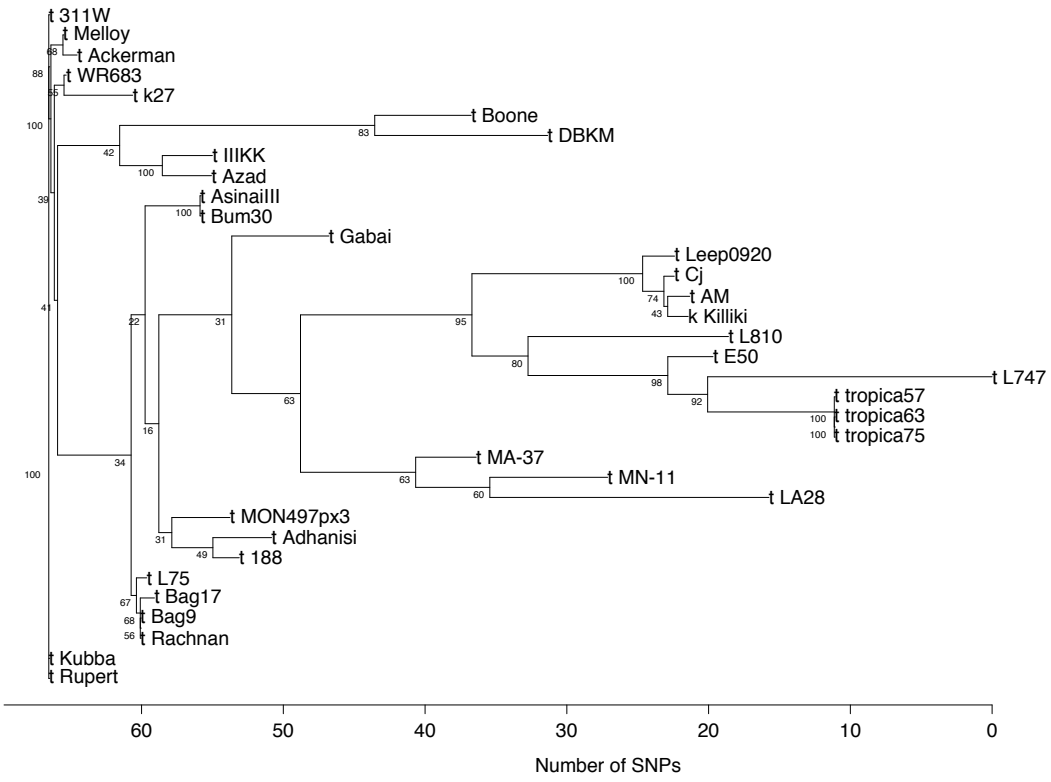


Figure 2.4. NJ tree from concatenated sequence data of markers with complete sequence information for all isolates. Note the presence of many unique "sequence types", represented by single isolates being at the tip of individual branches. Branch length represents the number of SNPs separating isolates from each other. Bootstrap confidence values are provided for each branch point (for 100 bootstrap replicates). Small clusters of identical "sequence types" are observed (e.g. tropica75, tropica63, tropica57).

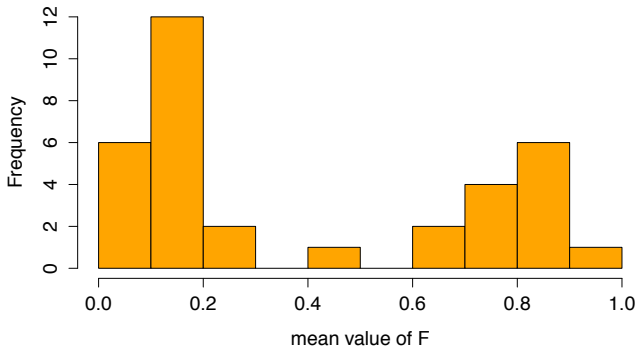


Figure 2.5. Histogram of inbreeding coefficients (F_{IT}) for the 34 isolates in this study. The data appears to be bimodally distributed, with two peaks, one centered around 0.1 and one around 0.8. See Appendix B for F_{IT} values.

DAPC-based clustering indicated an optimal number of cluster k = 3 (BIC = 86.86766) (Figure 2.6). The largest cluster encompassed 24 isolates, and two smaller clusters were found to be represented by 5 isolates each. Interestingly, the large cluster was found to contain mostly strains characterized by low F_{IT} values. This same cluster only contained Asian isolates, which originated from Greece, Israel, Syria, Saudi Arabia, Jordan, Afghanistan, Iraq, Kazakhstan, India. The only country represented by all three clusters was Israel, suggesting possible mixing of strains of different genetic backgrounds in this region, with potentially important epidemiological consequences. The other two clusters encompassed isolates from Palestine and Israel, on one hand, and North African countries such as Lybia and Tunisia, in addition to Israel, on the other. Overlaying the DAPC-based clustering on a phylogenetic tree obtained with the NJ method shows phylogenetic clades to

match these three groupings, with the exception of isolate L810, which falls in between Cluster 2 and 3 (Figure 2.7).

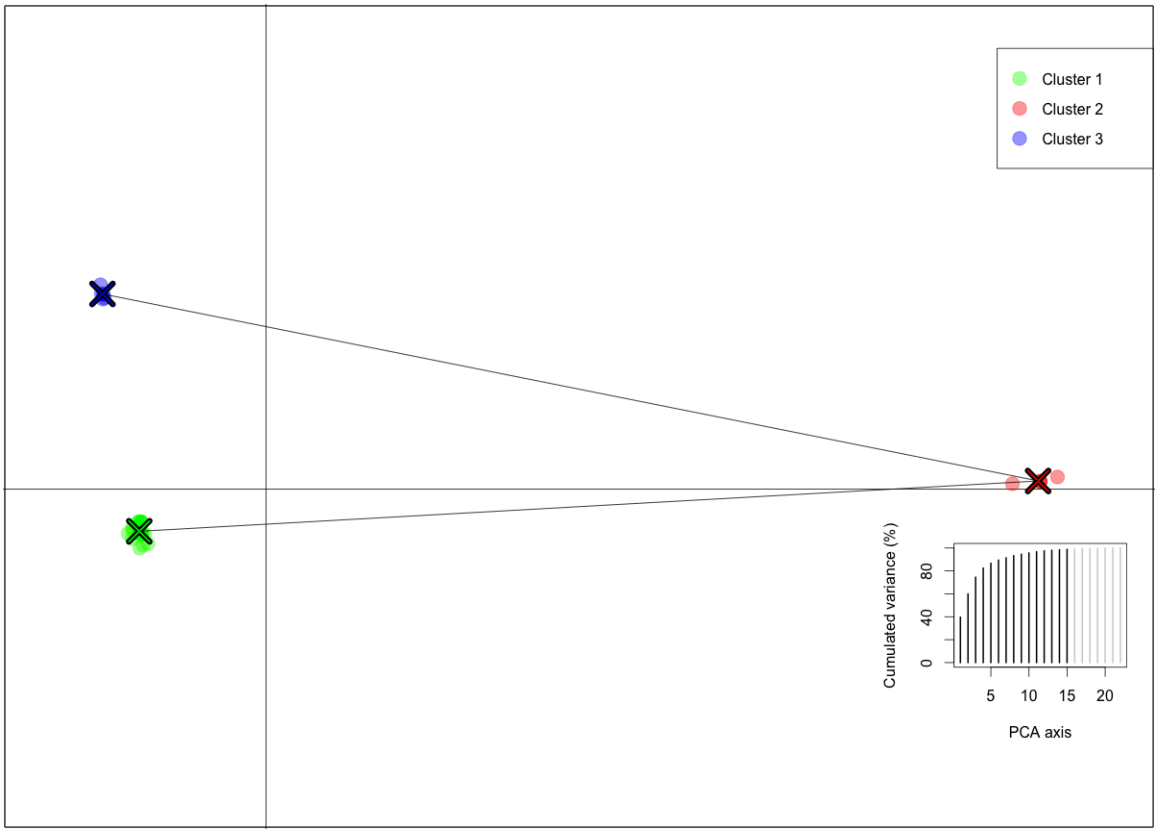


Figure 2.6. Clustering of the 34 isolates of *L. tropica* is consistent with geography and indicates possible mixing of different clusters in Israel and neighbouring countries. DAPC clustering and *k*-means analysis finds an optimal number of clusters *k* = 3. The first 15 principal component were retained in this analysis, explaining nearly 100% of the variance. The isolates within each cluster are reported in Figure 2.7.

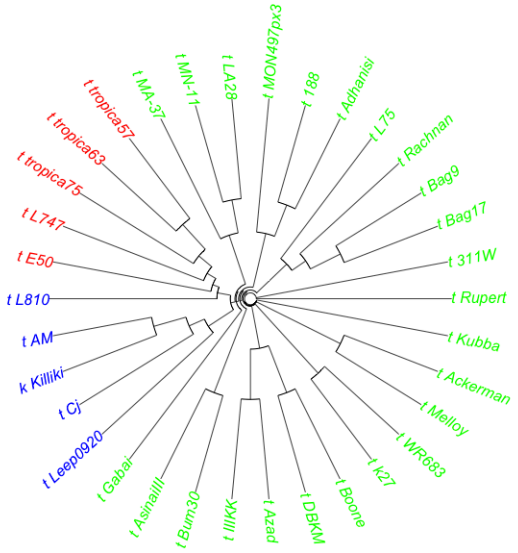


Figure 2.7. Unrooted, ultrametric NJ tree of 34 isolates of *L. tropica* based on the concatenated sequence data. Please refer to Figure 2.3 for a non-ultrametric version where branch lengths represent genetic distance. Colors are based on the DAPC clusters as in Figure 2.6. The largest cluster (green, n = 24) is composed by isolates with F_{IT} values less than 0.5, except for the clade composed by isolates LA28, MN-11, and MA-37 (with F_{IT} values 0.7996, 0.6819, and 0.5529, respectively). The isolates in the two smaller clusters (red, n = 5, and blue, n = 5) all had F_{IT} values larger than 0.5.

2.3.2 Whole-genome sequencing of field isolates

Sequencing of 18 field isolates, mapping, and variant calling predicted 732 888 variants. After filtering low quality variants, variants on unscaffolded contigs, and variants that were not SNPs or that had more than 2 alleles, we reduced our SNP set to 306 596 high quality biallelic SNPs for our population analyses.

The distribution of mean allele frequencies for each biallelic SNP showed that there were peaks corresponding to the expected frequencies for different levels of ploidy. For instance, a ploidy of 2 would predict frequencies of 0.5 for each allele in heterozygous individuals; a ploidy of 3 would be consistent with frequencies of 0.33 and 0.67; a ploidy of 4 corresponds to 0.5 and 0.5, or 0.25 and 0.75; and so on. The allele frequencies in our sample set behaved as expected, with peaks at the expected positions for disomic, trisomic, and tetrasomic heterozygous SNPs (Figure 2.8). In addition, two large peaks were observed near frequencies of 0 and 1. Note that these are allele frequencies averaged across individuals, suggesting that the individual isolates in our sample set are on average disomic, trisomic, or tetrasomic at that SNP.

To further resolve allelic variation amongst our samples, a plot representing number of alternate alleles in typed individuals (either 0 for homozygous reference individuals, 1 for heterozygotes, and 2 for homozygous alternate individuals) was generated (Figure 2.9). The plot shows a few strains to be homozygous alternate at the majority of SNP positions in the genome (E50, L747, L810, L838, and to a lesser degree, MA-37 and MN-11). No strains were uniformly homozygous reference, even

when they originated from the same geographical region as the reference isolate L590, which was isolated in Israel near Kfar Adumim in the Judean Desert (Schnur, Nasereddin et al. 2004) (Schnur 2004).

Sample	Lanes	Insert size	Cycles	Depth	Bases (pre-QC)	Bases (post-QC)
Tropica 63/ L838	1	500	100	13.27x	811.77 Mb	101.47 Mb
Ackerman	2	500	100	57.94x	1.11 Gb / 1.11 Gb	100.62 Mb / 100.80 Mb
L747	2	500	100	49.09x	986.31 Mb / 986.71 Mb	109.59 Mb / 109.63 Mb
L810	2	500	100	55.58x	1.08 Gb / 1.08 Gb	107.66 Mb / 107.77 Mb
E50	2	500	100	46.99x	932.36 Mb / 933.44 Mb	103.60 Mb / 103.72 Mb
MA-37	2	500	100	48.11x	937.02 Mb / 937.50 Mb	104.11 Mb / 104.17 Mb
MN-11	2	500	100	52.92x	1.02 Gb / 1.03 Gb	102.46 Mb / 102.52 Mb
Kubba	2	500	100	48.55x	919.97 Mb / 919.37 Gb	102.22 Mb / 102.15 Mb
Melloy	2	500	100	55.24x	1.04 Gb / 1.04 Gb	103.88 Mb / 103.94 Mb
Boone	2	500	100	39.78x	1.05 Gb / 1.05 Gb	105.15 Mb / 104.99 Mb
Adhanis I	2	500	100	26.52x	1.48 Gb	105.48 Mb
L75	1	500	100	22.15x	1.37 Gb	105.45 Mb
SAF-K27	1	500	100	32.91x	1.92 Gb	101.10 Mb
IIIKK / KK27	2	500	100	52.02x	986.09 Mb / 983.71 Mb	109.57 Mb / 109.30 Mb
Rupert	2	500	100	53.02x	1.01 Gb / 1.01 Gb	100.69 Mb / 100.91 Mb
Azad	2	500	100	50.67x	956.58 Mb / 960.51 Mb	106.29 Mb / 106.72 Mb
188 / K26	2	500	100	55.52x	1.05 Gb / 1.05 Gb	105.10 Mb / 105.33 Mb
311W / K112	2	500	100	58.42x	1.14 Gb / 1.13 Gb	103.19 Mb / 103.17 Mb

Table 2.4. Number of lanes, insert size, read length in number of cycles, depth, and total number of bases obtained from sequencing runs of the 18 isolates that were whole genome sequenced for this study. Pre- and post-QC base yield is provided by lane. See Appendix A for ENA accession numbers.

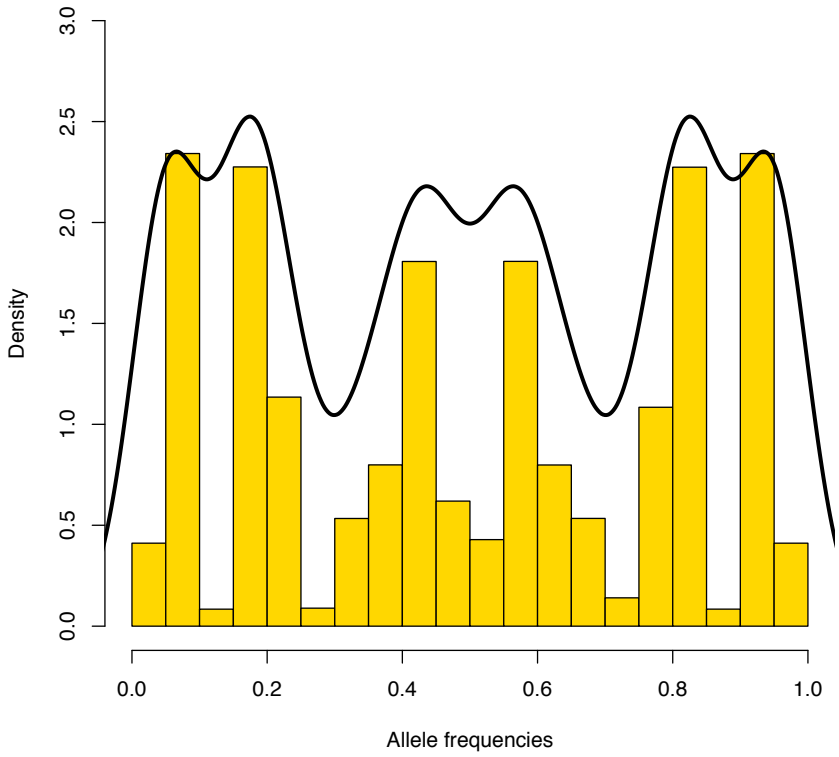


Figure 2.8. Average allele frequencies for each SNP across 18 individuals typed by WGS. Note the symmetry of the plot, expected for biallelic variants, and the peaks near 0 and 1 (homozygous), the peaks near 0.4 and 0.6 (heterozygous trisomic), and near 0.2 and 0.8 (heterozygous tetrasomic). Allele frequencies are averaged across all individuals thus slightly shifting the peaks from their expected values in the trisomic case (0.33 and 0.67). Allele frequencies shown are cumulative for all chromosomes.

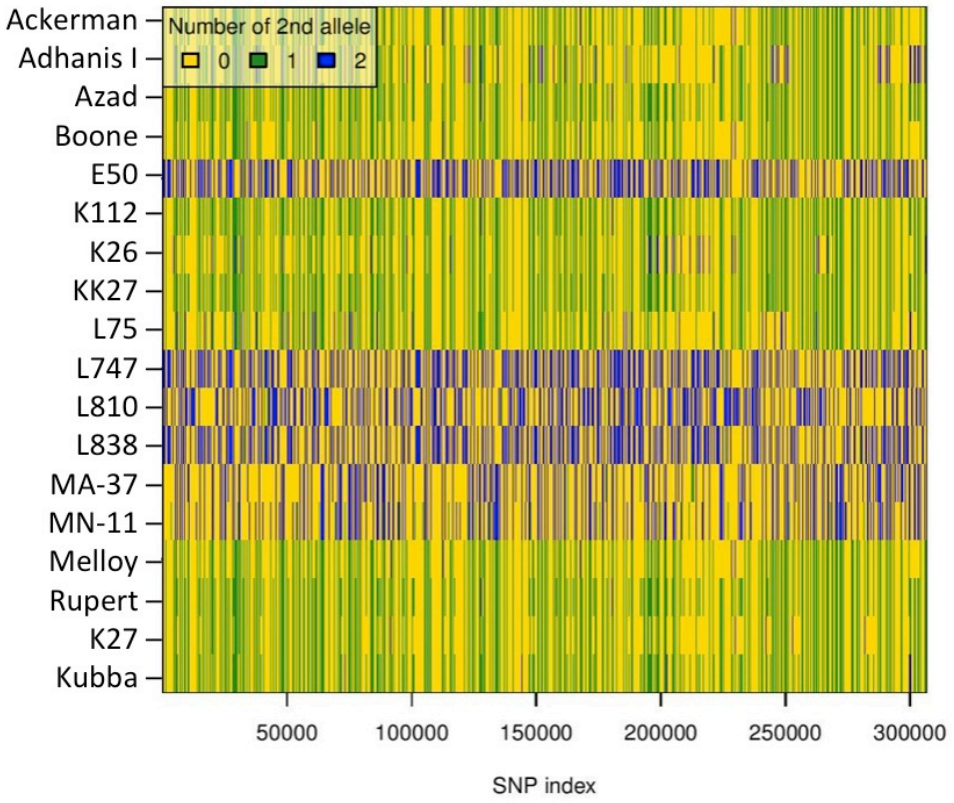


Figure 2.9. Allelic plot for all 306 596 SNPs that passed the filtering thresholds as described in the text, for all 18 strains. Heterozygous individuals have 1 copy of the 2nd allele at that typed position. Only samples that had either a frozen stock available or that had previously been sequenced were included.

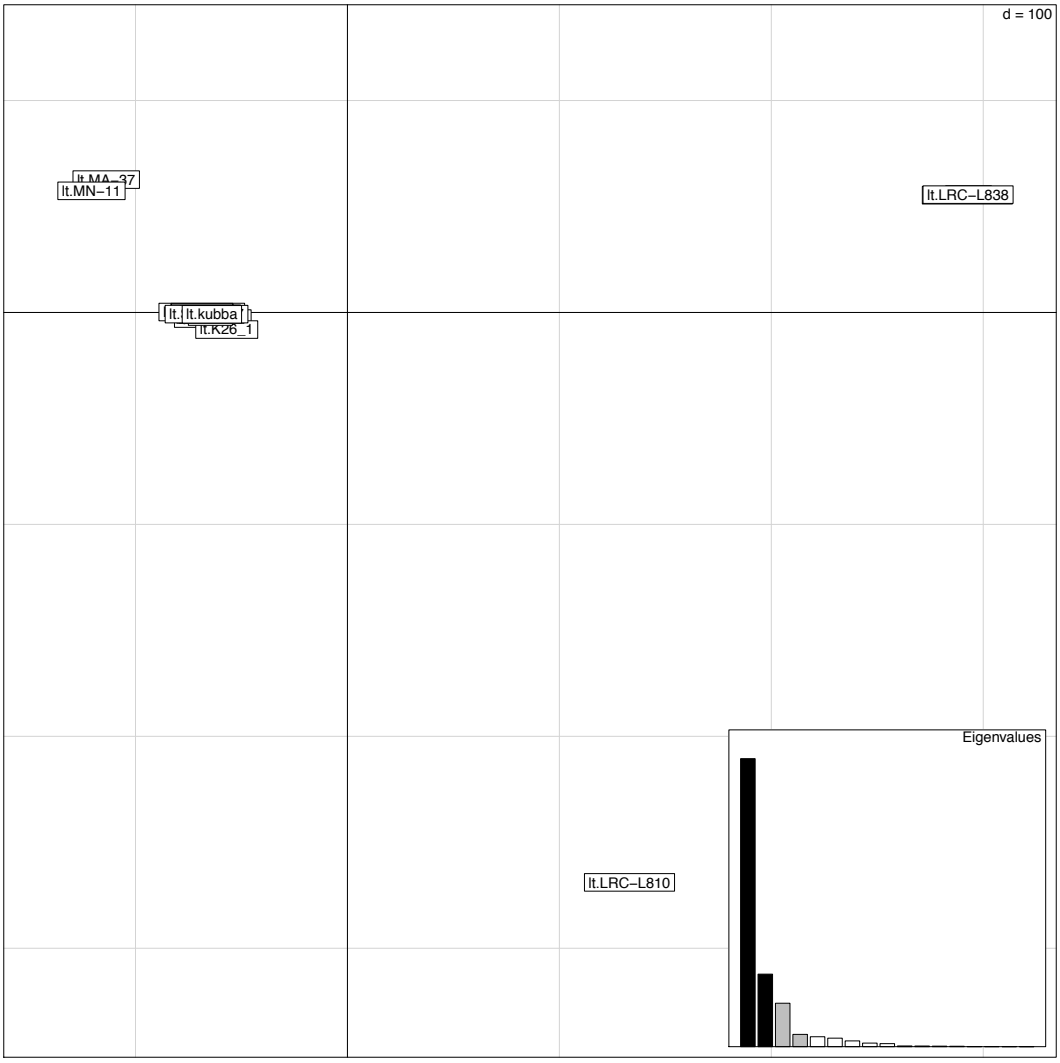


Figure 2.10. PCA plot of the 18 isolates that were typed by WGS. The first two principal components could explain almost 100% of the variance, as shown by the darker bars in the inset. Considerable distance is observed between L838 (tropica63 in the MLST analysis) and L810 and the rest of the isolates.

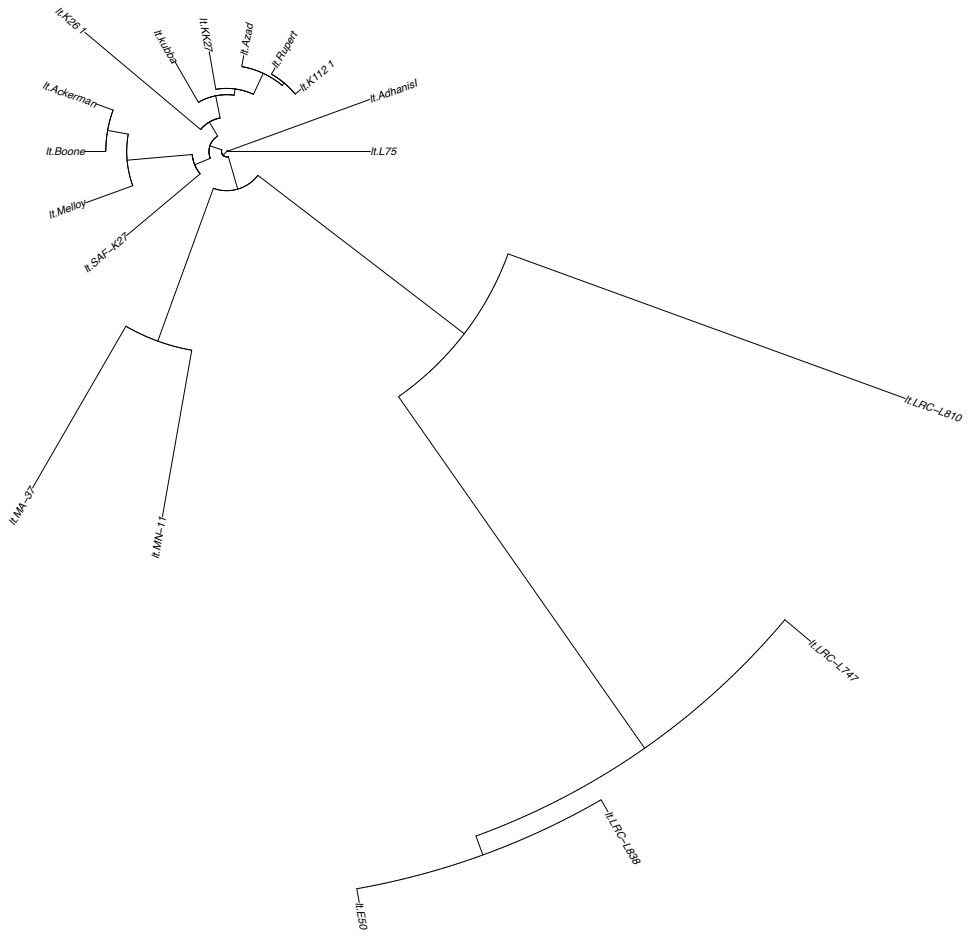


Figure 2.11. NJ tree for all 18 isolates based on the same set of 306 596 high quality biallelic SNPs as in Figure 2.9. Isolates L747, L838 (tropica63) and E50 appear to be outliers (part of Cluster 2 in Figure 2.7), while L810 appears to be quite divergent. MA-37 and MN-11 appear to be closely related, confirming the tight clustering shown in Figure 2.10.

A PCA plot was generated from the same set of genome-wide SNPs, showing close clustering of isolates MA-37 and MN-11, while the two isolates L838 (Palestine) and L810 (North Israel) appeared to be outliers (Figure 2.10). Despite the fact that L747, L838, and E50 fell in the same cluster in the DAPC plot, these do not form a cluster in the less powerful PCA plot. Both types of plot are conceptually similar in their approach, but DAPC has slightly more discriminatory power due to its emphasis on variation between groups as opposed to variation within groups, while also retaining the "blind" approach of PCA, unlike more sophisticated Bayesian methods, which instead rely on *a priori* population models (Jombart, Devillard et al. 2010). Given the small number of isolates that were typed by WGS and the limited number of principal components that explain almost 100% of the variation observed, computationally intensive DAPC was not performed on the genome-wide SNP data.

Phylogenetic analysis of the genome-wide SNP data suggests that isolates L810, L838, and E50 form a separate clade from all other isolates. MN-11 and MA-37 appear to be closely related. The relationships observed in the phylogenetic analysis based on the concatenated sequence data from a limited number of markers clash in some cases with what is seen in the genome-wide data. Specifically, K26 (188 in the MLST analysis) and Adhanis I do not appear to be as closely related in the PCA as in the NJ tree generated from the concatenated MLST data. In addition, the isolates Melloy, Boone, and Ackerman do not form a single clade in the tree generated using the MLST data, while they do so in the tree from the WGS data.

2.4 Discussion

The extent to which *Leishmania* species undergo sexual reproduction as opposed to asexual reproduction is a matter of contention. The possibility that genetic exchange is occurring in natural populations of this parasite has important consequences for disease management and control. *L. tropica* is a species responsible for both CL and VTL that has greatly increased its range in recent years, spreading to previously unaffected areas in Morocco, Kenya, Ethiopia, Israel, the Palestinian Authority, Jordan, Iraq, Afghanistan, and India.

Our MLST analysis confirms previous reports of observed heterozygosity in this species (Schwenkenbecher, Wirth et al. 2006, El Baidouri, Diancourt et al. 2013). Observed heterozygosity is slightly lower than the expected heterozygosity in our sample, under the assumption of panmixia and no population substructuring. Given the large variation in inbreeding coefficients observed in our set of isolates, there may be population substructuring due to geographical barriers or differences in transmission cycles that may be causing this heterozygous deficiency via Wahlund effects, a very well-known phenomenon in population genetics (Rougeron, De Meeus et al. 2015) which is discussed in depth in the introduction. Alternatively, heterozygous "clones" produced by ancestral meiotic hybridization events may be propagating asexually for physiological rather than epidemiological reasons, such as the absence of an obligatory meiotic stage, producing large skews in Hardy-Weinberg ratios.

The data seems to suggest the presence of genetic mechanisms that favour the appearance of heterozygous genotypes, compatible with hybridization with meiotic recombination. Allelic frequencies show a relatively high number of orphan alleles, possibly a result of the small number of isolates sampled. As previously mentioned, as the number of heterozygous SNP calls in a given genotype sequence increases, the number of possible parental haplotype alleles for that heterozygous genotype increases by 2^n (if the organism is diploid, where n = number of heterozygous SNPs), making it exceedingly difficult to find a phasing solution. The presence of orphan alleles in this data set may be associated with this difficulty, or may simply be due to limited sampling of the natural population. Given a larger number of sampled genotypes, the number of orphan alleles may decrease for the markers in this panel.

The heterozygosity that was observed in *L. tropica* may arise via mating of homozygous parental parasite lines, or through the accumulation of independent mutations in each homologous chromosome pair. The latter scenario would increase the number of orphan alleles observed, as heterozygosity in this case would evolve in a step-wise manner, each mutation effectively creating a new allele. This was not corroborated by the data. Allelic sharing between isolates was clear, with heterozygous genotypes clearly being made up by two parental alleles that were present in homozygous form in other sampled isolates. Meiotic recombination could explain why homozygous markers linked to heterozygous markers were observed on some of the chromosomes.

Alignment and hierarchical clustering of concatenated sequences identified a small number of identical genotypes. Given that this isolate set contains samples collected over a very broad span of time, this may be taken as evidence for persistence of "clonal" genotypes. However, samples that had identical sequence types tended to be from the same geographic region, and to have been collected within a narrow time span. Moreover, if we also include sequence information from markers with "missing" genotypes for some of the isolates, then all isolates have unique sequence types. The presence of identical sequence types appears to be an artefact of limited genotyping, and these disappear if we consider data from the full 34-marker genotyping panel or from WGS.

Clustering identified three main groups, which are for the most part consistently reflected in the branching of the underlying tree (colored in red, blue, and green in Figure 2.7). All African isolates appear to group together both in the clustering and phylogenetic analysis, while the rest of the Middle Eastern and Asian isolates cluster separately and also form a separate polyphyletic clade on the unrooted tree. A small group of Israeli and Palestinian samples belonging to Cluster 2 formed a separate, derived clade in the tree. The isolate L810 was problematic and provided conflicting results between the clustering and phylogenetic analyses. Although it clustered with African isolates in Cluster 3, it however fell into a separate clade in the NJ tree, and was at the base of a clade containing other isolates from Israel and Palestine that fell into Cluster 2. This fact is consistent with previously published data on this isolate, which was isolated from a *Phlebotomus arabicus* sand fly in Northern Israel. In addition to being more closely related to *L.*

major by isoenzyme analysis than other *L. tropica* isolates, this subpopulation was later shown to be transmitted via a distinct transmission cycle from the one that is more common in the wider Middle Eastern region, which involves *P. sergenti* sand flies instead of *P. arabicus* (Soares, Barron et al. 2004).

In conclusion, my analysis of heterozygosity and inbreeding coefficients suggests the presence of population substructuring. Given the bimodal distribution observed in the F statistic, it is clear that parasite isolates form two large groups with opposite patterns of heterozygosity (n=10 and n= 19), and a smaller group with intermediate levels of heterozygosity (n = 5). A more realistic modelling of the population structure in our sample set, given that the assumption of panmixia is not valid, may increase the statistical accuracy measuring gene flow between population units. Alternative F-statistic measures that are informed by population structure, such as F_{IS}, should be used when population units are known. Given that Hardy-Weinberg equilibrium is rarely observed in large natural populations, and that isolation by distance is likely occurring in *L. tropica*, an improved study design that includes sampling of a larger number of samples from well-defined population units may prove to be more informative for population genetic purposes. My aim was to gain a broad understanding of genetic variation within this species, and to prioritize a few strains for laboratory crossing experiments (see Chapter 4).

Comparisons between the MLST and WGS data suggest important differences in these two approaches. The small set of isolates that formed a single cluster (Cluster 2, in red) in the MLST analysis shows an added layer of variation that is detectable only when we expand the number of typed loci to include

thousands of SNPs across the genome. Both PCA and phylogenetics confirms two isolates, one from North Israel and one from Palestine, to be extremely divergent (L810 and L838). Their relative position with respect to other individuals in Cluster 2 however varies depending on the genotyping platform adopted. While a relationship between L810, L747, L838, and E50 is obvious in the phylogenetic analysis for both MLST and WGS, in the PCA plot the relatedness between L810 and the rest of the isolates in this clade appears to be weaker. Interestingly, the genome-wide allelic plot shows these same 4 isolates to be mostly homozygous for the alternate allele, while the majority of the other isolates are either heterozygous or homozygous reference. Moreover, the same phylogenetic analysis done on MLST or WGS data suggests different relationships between isolates Melloy, Boone, and Ackerman, hinting at the presence of fine genetic variation which may fall outside of the portion of the genome sampled by MLST.

In summary, the genetic variation observed in natural populations of *L. tropica* was grouped into three clusters. The first cluster (Cluster 1) appears to harbour considerable heterozygosity and is widespread through the Middle East and Asia. The second cluster (Cluster 2) has reduced heterozygosity, and is mostly concentrated in the Eastern Mediterranean. The third cluster (Cluster 3) also has reduced heterozygosity, and is found throughout Northern Africa. The possibility that Cluster 3 represents a "hybrid" genotype between isolates from Cluster 1 and Cluster 2 is intriguing, but this hypothesis could not be directly tested due to insufficient data. The patterns of allele sharing presented in the allele plots in Figure 2.3 and in Figure 2.9 are compatible with this scenario. Published microsatellite

evidence has also been brought forward in support of this hypothesis (Schwenkenbecher, Wirth et al. 2006, Krayter, Bumb et al. 2014). Isolate L810 appears to be problematic, falling into Cluster 3 but appearing to be closely related to other isolates in Cluster 2, both by DAPC of MLST data and PCA of WGS data. Although DAPC may artificially reduce variation within clusters and inflate variation between clusters, the relationships we have identified appear to be robust enough after validation with WGS to be employed as a useful conceptual framework for studying patterns of genetic exchange and genome plasticity in these *L. tropica* isolates in subsequent chapters.

CHAPTER 3

GENOME PLASTICITY AND GENE EXPRESSION

Publication note: *the following chapter contains excerpts of a manuscript that has been submitted for peer-review. All experimental procedures and analyses were the sole work of the first author, Stefano Iantorno, under the supervision of his PhD supervisors, Dr Matt Berriman, Dr James Cotton, Dr Michael Grigg, and Dr David Sacks, unless explicitly stated otherwise in the text. We thank Caroline Durrant for creating the EM algorithm used for estimating somy, and James Cotton, Wes Warren, and Stephen Beverley for their work generating and annotating the reference genome used in this chapter.*

3.1. Introduction

As described in the introduction, *Leishmania* parasites are characterized by a remarkable genomic plasticity. The processes by which this plasticity is translated into gene expression differences in strains circulating in the field represent an area that has been so far unexplored in *Leishmania* genetics. In this chapter, we describe the first comprehensive, high-resolution study of intraspecific differences in gene expression in an Old World species, *L. tropica*, responsible for significant CL in endemic areas in North and East Africa, the Middle East, and the Indian subcontinent.

This species is known to harbour considerable intraspecific variation compared to other *Leishmania*, as determined by microsatellite analysis (Krayter, Bumb et al. 2014), as I have described in Chapter 2. In addition to CL, some strains of *L. tropica* have also been associated with a variant form of VL known as viscerotropic leishmaniasis (Dillon, Day et al. 1995, Sacks, Kenney et al. 1995). Considerable intraspecific variation in the response of *L. tropica* to treatment has also been documented (Hadighi, Mohebali et al. 2006, Plourde, Coelho et al. 2012), with cutaneous lesions due to *L. tropica* generally being less responsive to treatment than lesions due to *L. major*.

Distinguishing characteristics of kinetoplastid organisms such *L. tropica* include a unique DNA-containing organelle called the kinetoplast, situated close to the flagellum; RNA editing of the genes encoded by kinetoplast DNA; and lack of regulation of nuclear gene expression at the level of transcription initiation, with genes being constitutively transcribed as polycistronic units and giving rise to individual protein-coding transcripts by trans-splicing.

As discussed in depth in the introduction, transcriptional units in *Leishmania* and other kinetoplastids lack traditional eukaryotic promoter and terminator elements. The majority of protein-coding genes are organized into head-to-tail polycistronic segments containing functionally unrelated genes, often hundreds of kilobases in size, which are transcribed into mRNA by RNA polymerase II. RNA pol II transcription is thought to start in divergent strand switch regions, and terminates in convergent strand switch regions. Transcriptional unit boundaries are enriched in histone acetylation marks, and in one type of hyper-modified base unique to

kinetoplastid protozoa called base J, which was recently shown to be essential in regulating RNA Pol II transcription termination (van Luenen, Farris et al. 2012, Reynolds, Cliffe et al. 2014). In addition to these epigenetic marks, divergent strand switch regions associated with transcription initiation by RNA pol II also show a positionally conserved high curvature in DNA secondary structure, which might be facilitating binding of RNA polymerase enzymes (Tosato, Ciarloni et al. 2001, Smircich, Forteza et al. 2013).

Given the absence of classically defined promoters in *Leishmania*, regulation of gene expression during parasite development is thought to occur post-transcriptionally. Changes in steady state transcript levels within the cell are primarily ascribed to differences in the maturation and stability of individual mRNAs. The nascent polycistronic mRNA transcript matures by trans-splicing of a 39-nucleotide mini-exon sequence, called the spliced leader (SL), to the 5' end of the pre-mRNA transcript (Sutton and Boothroyd 1986). This process is mediated by the spliceosome complex, and is coupled with concurrent polyadenylation of the 3' end of the upstream mature monocistron (LeBowitz, Smith et al. 1993).

Regulatory sequence elements in neighbouring untranslated regions (UTRs) determine the trans-splicing efficiency of individual protein-coding transcripts, as well as their half-life via interactions with RNA-binding proteins (De Gaudenzi, Noe et al. 2011). Most elements that have been implicated in determining the stability of trans-spliced mRNA transcripts are in the 3' UTR of protein-coding genes, due to the small size of 5’ UTRs. These regulatory sequences include AU-rich instability elements (AREs) (Milone, Wilusz et al. 2002), short interspersed degenerated

retroposons (SIDERs) (Muller, Padmanabhan et al. 2010), U-rich element (UREs) (Haile, Dupe et al. 2008), paraflagellar rod regulatory element (PREs) (Holzer, Mishra et al. 2008), as well as others (McNicoll, Muller et al. 2005).

In most eukaryotes, regulation at the level of mRNA turnover occurs by cap removal and shortening of the poly-A tail by a variety of cellular de-capping enzymes and deadenylases, followed by degradation by exonucleases with either 5' to 3' or 3' to 5' activity, as well as additional specialized pathways (Parker and Song 2004, Houseley, LaCava et al. 2006). Evidence for some of these exosome-mediated processes has been found in kinetoplastids (Milone, Wilusz et al. 2002, Haile, Estevez et al. 2003, Li, Irmer et al. 2006, Schwede, Ellis et al. 2008, Schwede, Manful et al. 2009). Alpha and beta tubulin in *Leishmania* are one of the best examples of differentially expressed multi-copy genes in promastigote and amastigote stages (Fong, Wallach et al. 1984), suggesting that the detectable differences in the length of 3' UTRs of individual copies of each gene, which are arranged in tandem arrays, may be determining transcript stability in different developmental stages (Jackson, Vaughan et al. 2006, Ramirez, Requena et al. 2013). Cis-acting regulatory elements in 3’ UTRs such as the retroposon families SIDER1 and SIDER2, which have been linked to degradation of mRNA via a endonucleolytic- and deadenylation-independent pathway (Bringaud, Muller et al. 2007, Muller, Padmanabhan et al. 2010, Muller, Padmanabhan et al. 2010), are thought to play a major role in regulating transcript levels during parasite development.

An evolutionarily conserved and potentially adaptive characteristic of *Leishmania* parasites is the ability to tolerate extensive aneuploidy, a large

proportion of chromosomes diverging from the disomic state expected in diploid eukaryotes (Rogers, Hilley et al. 2011). Variation is also seen in gene copy number, with intrachromosomal gene duplication events giving rise to tandem gene arrays. Sequencing of several *L. donovani* field isolates from Nepal found relatively low single-nucleotide polymorphism (SNP) diversity, but large differences in chromosome copy number, and the presence of a stably inherited circular episome (Downing, Imamura et al. 2011). Extrachromosomal linear and circular amplicons are known to occur in Leishmania when exposed to drug selection (Ubeda, Legare et al. 2008, Ubeda, Raymond et al. 2014).

Gene expression studies in *Leishmania* have found a small number of developmentally regulated genes between promastigote and amastigote stages, suggesting the presence of a conserved set of genes in the vector stages which confers the parasite the basic toolkit necesssary for intracellular survival with minimal changes in expression (Rochette, Raymond et al. 2009, Alcolea, Alonso et al. 2014) (Rochette 2009, Alcolea 2014). To our knowledge, the only other published study employing high-throughput RNA-seq approaches to study gene expression in *Leishmania* identified 10285 transcripts in *L. major* axenic promastigotes, 1884 of which could be considered novel compared with previously annotated genes (Rastrojo, Carrasco-Ramiro et al. 2013).

Understanding how gene expression differences arise and how they determine intraspecific phenotypic diversity in this important human pathogen may prove essential to identify parasite factors underlying tissue tropism and clinical course of disease, as well as drug susceptibility. I attempt to shed some light on

these processes by sequencing the genomes of 14 field isolates originating from different endemic regions, paired with RNA-seq of their in vitro axenic promastigote cultures. I identify the most variable expressed genes in the set of samples included in this study. In order to understand the effects of mosaic aneuploidy in these samples, I also analyze 6 clonal lines isolated from 4 of these isolates. I describe the results of these analyses in the context of observed variation in somy and gene copy number, and suggest that decreased sensitivity to antileishmanial compounds may evolve through rapid changes in gene dosage and associated gene expression.

3.2. Methods

3.2.1. Sequencing and RNA-seq of field isolates

All 14 samples were axenically cultured *in vitro*. Each isolate had previously been culture adapted and cryopreserved in DMSO under liquid nitrogen storage conditions (-60 C). Each frozen stock was thawed and cultured for 1-3 days in complete M199 promastigote medium (see Section 2.2.1 for a detailed protocol of *in vitro* culturing of promastigotes) until parasite density reached 1 x 10^6 cells per mL. Each parasite culture was then split into three separate culture flasks for biological replication, and serially passaged every 24 hours for three days to maintain log-phase growth and density, following well-established procedures (Wheeler, Gluenz et al. 2011), and to synchronize the culture in the proliferative promastigote developmental stage. After the third 24hr interval, each set of replicates was

pelleted and the RNA was extracted using the TRIzol protocol. In order to isolate the effect of aneuploidy on gene expression from genetic background, a total of 6 isogenic lines were cloned from 4 of these isolates (Kubba, MN-11, MA-37, Rupert). These clones were grown in triplicate cultures as described above, and the RNA was extracted from parasite pellets. Independent cultures of these 6 isogenic clones and all field isolates were set up for DNA purification. RNA and DNA samples were used as the starting material to prepare Illumina libraries following manufacturer's specifications. For RNA, the TruSeq stranded mRNA prep kit was used, which relies on 3' poly-A tail pull down to isolate RNA species of interest. Purified RNA species were then prepped into paired-end libraries with an average insert size of 250 bp and sequenced on the Illumina HiSeq 2500 platform for 75 cycles. DNA samples were sequenced for 100 cycles using paired-end libraries with an average insert size of 500bp, but while the cloned isogenic lines were sequenced on a single lane on the Illumina HiSeq 2000 platform, the isolates of origin were multiplexed over two lanes on the Illumina HiSeq 2500 platform to maximize coverage (Table 3.1 and Table 3.2). The same sequence data used in Chapter 2 was used for this Chapter.

3.2.2. Mapping and analysis of whole genome sequence data

Short read genomic sequence data were aligned to the reference genome using SMALT (https://smalt.sourceforge.net) (see Section 2.2.3 for complete description of genome reference). Variants and depth of coverage were called with GATK (v.3.4-0, Broad Institute), and allele frequency and read depth information was

manipulated using custom bash, Perl and R scripts to generate the desired plots. Short haplotypes to verify allele-specific expression were assembled using the physical phasing (i.e., "read-backed" phasing) procedures implemented in Freebayes (https://github.com/ekg/freebayes) (Garrison and Marth 2012) to call, filter, and phase high quality SNPs. An in-house developed expectation-maximization (EM) algorithm was used to estimate somy for each chromosome starting from the expected haploid read depth. Allele frequency and read depth plots were visually inspected and used to confirm the estimated somy for each chromosome.

The EM algorithm uses a likelihood function which models the median read depth (RD) of each chromosome from a single sample as coming from a Poisson distribution. The means of these 36 Poisson distributions are defined as the product of the somy for that chromosome (as a whole number) multiplied by the haploid RD, which is the same for all chromosomes in a particular sample. The unknown parameters are the haploid RD and the somy for each chromosome. The maximisation step uses the current estimate of the vector of somy parameters to maximise the likelihood function for the haploid RD and then in the expectation step, the maximum-likelihood estimate (MLE) of the haploid RD is used to calculate the most likely value of the somy for each chromosome based on its Poisson distribution. These steps are iterated until no change in the parameter values is observed.

3.2.3. Mapping and analysis of RNA-seq data

Despite the nearly complete absence of cis-splicing in *Leishmania*, short reads were mapped to the same reference genome that was used for the analyses in Chapter 2 using Tophat (Trapnell, Pachter et al. 2009), a splice-sensitive aligner based on the Bowtie algorithm. Since the RNAseq paired-end library preparation protocol was stranded, the option fr-firststrand was used during mapping to preserve sense/antisense directionality of sequence information. The reference genome used for this study is very similar, but not identical, to the *L. tropica* assembly version available from TriTryDB, which was produced using a similar workflow. Our assembly and annotation is available upon request.

Given the large degree of gene conservation and synteny between homologous regions in *Leishmania* species (Rogers, Hilley et al. 2011), gene annotations were transferred from the *L. major* Friedlin reference genome annotation present in GeneDB on 12/12/2013 using the version of RATT (Otto, Dillon et al. 2011) included in PAGIT v1.0 (Swain, Tsai et al. 2012), using the 'species' transfer option. *L. major* is the most closely related *Leishmania* species with a well-annotated reference genome. This resulted in a total of 7863 genes in *L. tropica*. Reads overlapping feature annotations were counted with HTSeq 0.6.1 using the htseq-count function (Anders, Pyl et al. 2014) and the "intersection nonempty" command option. Raw gene counts were imported into R statistical software and analyzed with packages edgeR (McCarthy, Chen et al. 2012) and DEseq (Anders and Huber 2010). Custom scripts were used to do the statistical analyses and to generate the

figures. Multi-dimensional scaling of the biological coefficient of variation (equal to the square root of the common dispersion calculated from each pair of libraries) was obtained from the normalized gene counts of the top 500 genes with the largest tagwise dispersion. The normalization by library size used was the weighted trimmed mean of M-values (TMM) method implemented in edgeR.

In order to confirm the clustering results from MDS analysis, a more rigorous normalization procedure was then applied to the raw read counts. The Bayes empirical dispersion for each gene was calculated using relative-log expression (RLE) normalized read counts, treating all samples as if they were replicates of the same condition. A variance stabilizing transformation was then applied to the count data as implemented in DEseq. Euclidean distances were calculated on the variance stabilized expression values for each pair of samples, and pairwise Euclidean distances were plotted as a heatmap to visualize differences in expression signatures between samples.

3.2.4. Differential gene expression analysis

Raw counts from HTseq were normalized following the standard edgeR workflow. The isolate L747 was used as the intercept for calculation of fold change relative to this baseline expression, given its similarity to most other samples by the Euclidean distance metric. A generalized linear model (GLM) for negative binomially distributed count data was built, with each set of triplicates modelled as a separate condition. The isolate L747 was chosen as the intercept due to its near diploid

karyotype and small Euclidean distance values with most other isolates. This isolate therefore provided the best baseline to measure deviations in expression due to gene dosage. Common, trended, and tagwise dispersions were calculated with the Cox-Reid estimator. Given the multifactorial model of the experiment, the negative binomial GLM was fitted with the tagwise dispersion to allow for the possibility that dispersion might vary across genes. The likelihood ratio test was then used to compare each set of triplicates to the baseline, and identify genes that were differentially expressed in any of the groups in a test analogous to a one-way ANOVA. P-values for differentially expressed genes in any of the groups were calculated using the F distribution and adjusted for multiple testing using the BH method. Pairwise exact tests were also performed between pairs of isolates (L810 vs all other isolates, and Rupert C2 vs Rupert C1) to identify and confirm genes with variable expression.

3.2.5. Copy number variation, gene dosage, and allele specific gene expression

To estimate gene dosage effects on relative expression levels, we selected the two clones originating from field isolate Rupert due to their similarity in karyotype (both were nearly diploid). Although two separate clones were also generated from the MN11 isolate, these could not be used to investigate gene dosage effects due to potential artifacts introduced by the procedure of normalizing by library size when comparing samples with different levels of nearly balanced ploidy (e.g., a nearly triploid clone vs a nearly diploid clone, as we observed in the MN11 C2 vs MN11 C1

comparison). While the MA-37 C1 and Kubba C1 isogenic clones were close to the balanced ploidy of their isolates of origin, we could not directly compare them to their isolates of origins due to the heterogeneity in karyotype known to occur in uncloned field isolates.

Average read depth was calculated in 10kb windows across the genome, and compared with the expression values of the 4634 genes DE between Rupert C2 and Rupert C1. The average read depth within a gene was plotted against the number of reads overlapping the gene per million reads (counts per million, cpm) of that gene from the expression data. The chi-square statistic was used to infer correlation between gene dosage as determined by read depth and expression levels as determined by counts per million. Copy number variants between Rupert C1 and Rupert C2 were identified and annotated using the CNV-seq pipeline (Xie and Tammi 2009). Briefly, this pipeline identifies localized regions in which read depth normalized across the length of the chromosome differs significantly between two samples.

Manhattan plots were generated by plotting p-values obtained via Fisher's exact test performed on each SNP in the genome, testing association between the alternate allele frequency at that position in the DNA sequence data and the alternate allele frequency at that position in the RNA sequence data. This set of SNPs was created by force calling variants at previously identified high quality positions shared between the samples using a somy-sensitive pipeline based on the Freebayes variant calling algorithm.

3.3. Results

3.3.1. Sequencing of 20 parasite lines

Whole-genome sequencing of 14 isolates and 6 clones generated from 4 of these isolates revealed considerable differences in the size and distribution of runs of homozygosity, as well as large variation in ploidy. *L. tropica*, like *L. major*, has a 36-chromosome karyotype. Chromosome 31 was either trisomic, tetrasomic, or hexasomic in all isolates. No chromosome was consistently disomic in all isolates. Most chromosomes varied between the disomic and trisomic state. The field isolate with the most variation in chromosome number (Azad) had 20 disomic chromosomes, 14 trisomic chromosomes, and 2 tetrasomic chromosomes. One clonal line (MN-11 C1) was nearly triploid (Figure 3.1), with a tetrasomic chromosome 7, and hexasomic chromosomes 20 and 31. Seven field isolates and one clonal line were nearly diploid, with only chromosome 31 being present in the tetrasomic state (Kubba C1, MA-37, E50, L747, L810) or in the trisomic state (K112, KK27, Kubba). The observation that clones of the same isolate differed in their karyotype confirmed that mosaic aneuploidy is an important source of genomic variation in this species.

Plotting of allele frequencies revealed many long runs of homozygosity (LROH) in chromosomal regions of several isolates, on a background of prevailing heterozygosity (Figure 3.1). Only 5 isolates (E50, MA-37, MN-11, L747, L810) out of 14 were broadly homozygous across the whole genome; the remaining isolates were

heterozygous across the majority of their genomes, except for 5 other isolates that were essentially fully heterozygous at all chromosomes examined (KK27, Rupert, Kubba, Azad, K112). Of the 6 clones, one isolate originating from a homozygous isolate was broadly heterozygous (MN-11 C2, from isolate MN-11), suggesting that considerable sequence diversity may exist within the same clinical sample.

Principal component analysis (PCA) on biallelic SNP variants called from these samples suggests overall sequence similarity within this set of isolates, both between isolates and within sets of clonal lines originating from the same isolate. Three samples, however, were very divergent from the rest and clustered in two separate groups (L747 and E50 in one group, and L810 quite separate from all other samples) (Figure 3.2). The two isogenic clones of MN-11 showed divergence in all PC comparisons performed (EV1 vs EV2, EV1 vs EV3, EV2 vs EV3, accounting for almost 100% of the variation observed), confirming clonal heterogeneity at the level of sequence diversity within the same isolate.

Sample	Lanes	Insert size	Cycles	Depth	Bases (pre-QC)	Bases (post-QC)
Rupert C1	1	500	100	89.14x	3.49 Gb	102.73 Mb
Rupert C2	1	500	100	88.18x	3.45 Gb	101.50 Mb
MN-11 C1	1	500	100	135.96x	5.39 Gb	101.75 Mb
MN-11 C2	1	500	100	81.34x	3.08 Gb	102.71 Mb
MA-37 C1	1	500	100	66.33x	2.79 Gb	103.20 Mb
Kubba C1	1	500	100	75.72x	2.89 Gb	103.38 Mb

Table 3.1. Number of lanes, insert size, read length in number of cycles, depth, and total number of bases obtained for sequencing runs of the 6 clones generated for this study. Pre- and post-QC base yield is provided by lane. Please refer to Table 2.3 for sequencing information of the 14 uncloned lines. See Appendix A for ENA accession numbers.

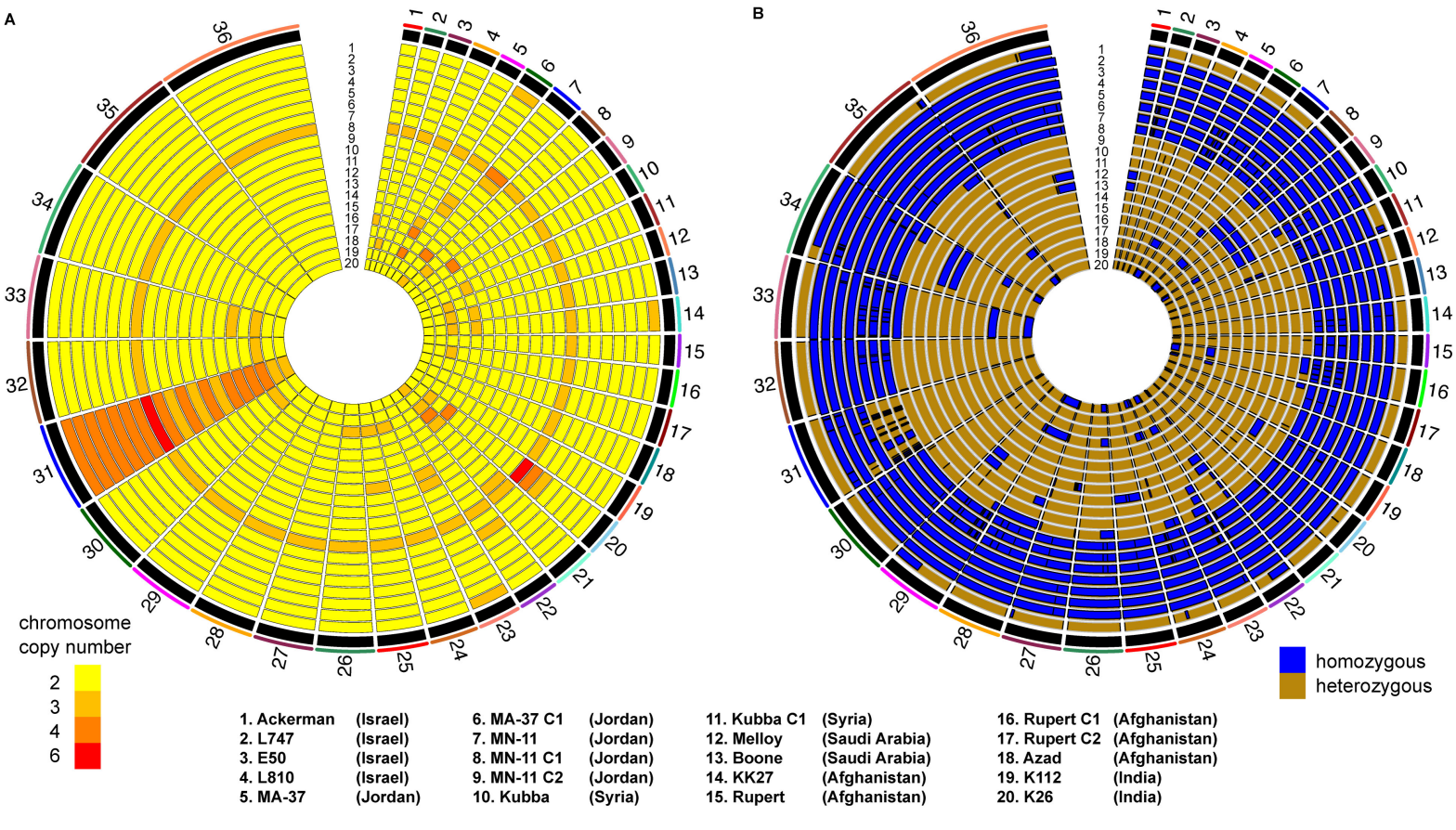
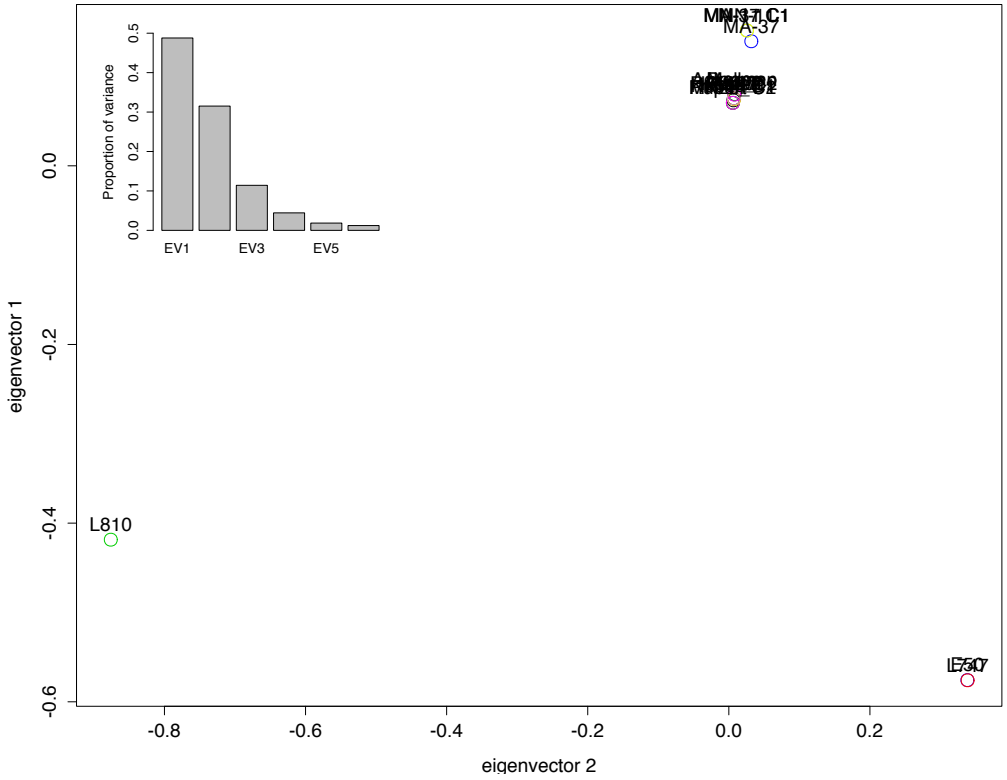
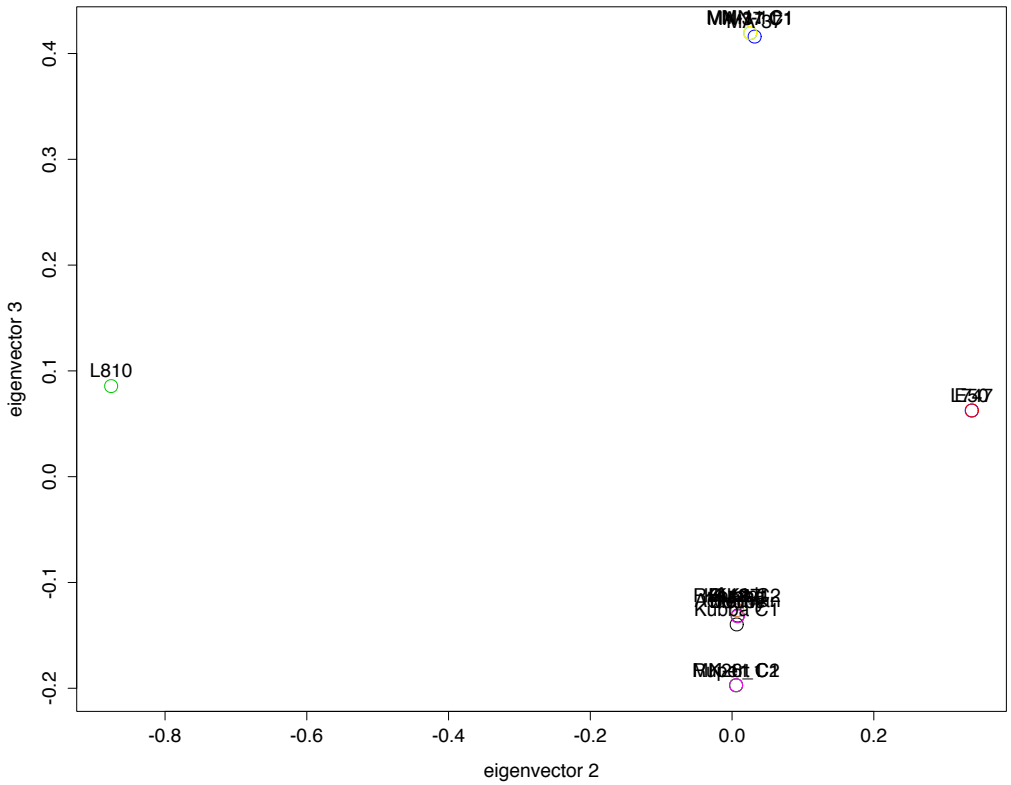

A
B
chromosome
copy number
2
3
4
6
homozygous
heterozygous
1. Ackerman (Israel)
2. L747 (Israel)
3. E50 (Israel)
4. L810 (Israel)
5. MA-37 (Jordan)
6. MA-37 C1 (Jordan)
7. MN-11 (Jordan)
8. MN-11 C1 (Jordan)
9. MN-11 C2 (Jordan)
10. Kubba (Syria)
11. Kubba C1 (Syria)
12. Melloy (Saudi Arabia)
13. Boone (Saudi Arabia)
14. KK27 (Afghanistan)
15. Rupert (Afghanistan)
16. Rupert C1 (Afghanistan)
17. Rupert C2 (Afghanistan)
18. Azad (Afghanistan)
19. K112 (India)
20. K26 (India)

Figure 3.1. Circular plots representing ploidy and long runs of homozygosity (LROH). Panel A represents somy in the panel of isolates used in this study, with increasing somy depicted with colors going from yellow to red. Panel B represents LROH, with homozygosity in blue and heterozygosity in gold.


Proportion of variance
0.0
0.1
0.2
0.3
0.4
0.5
EV1
EV3
EV5
MA-37
L810
eigenvector 1
eigenvector 2
-0.6
-0.4
-0.2
0.0
-0.8
0.2


L810
eigenvector 3
eigenvector 2
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.8
-0.6
-0.4
-0.2
0.0
0.2

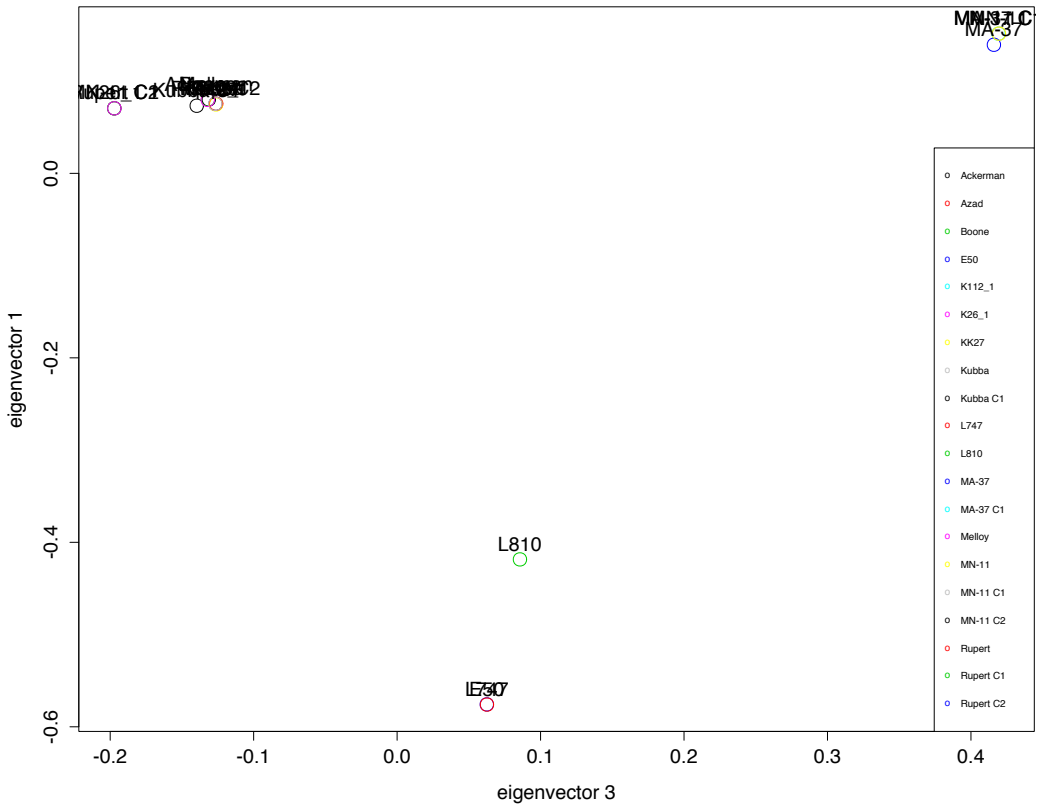


Figure 3.2. PCA plots of the 14 different isolates and 6 additional clones considered for this analysis, comparing eigenvectors 1 vs 2, 2 vs 3, and 1 vs 3. The first three eigenvectors, which represent the first three PCs, explain the majority of the variance observed, as shown by the inset in EV1 vs EV2. A color key for all 20 samples is provided in the last plot, EV 1 vs EV3.

3.3.2. RNA-seq of field isolates and clonal lines

RNA purification failed to yield enough material for RNA-seq in two samples (E50, K112), so these isolates were not submitted for sequencing due to the poor quality of the sequence data that would have been obtained. Replicates for the remaining 18 isolates, including all clonal lines, were sequenced to a median depth of between 50 and 80x. Raw read counts were normalized by library size using the weighted trimmed mean of M-values (TMM) method. Variation in the normalized data was then inspected via multi-dimensional scaling (MDS) (Figure 3.3). MDS of these 18 sets of triplicates showed that one replicate from isolate Boone clustered closely with the set of triplicates from isolate L810. Upon further inspection of the sequence data, this replicate appeared to be a cross-contamination of the culture by isolate L810. The three replicates from isolate Boone were therefore excluded from subsequent analyses. As expected, the remaining 17 sets of triplicates showed less variation within triplicates than between triplicates, with the exception of MN11 C2. One of the triplicates from the cloned line MN-11 C2 differed from the other two replicates, resembling the signature of L810, so this set of replicates was also excluded from subsequent analyses as a possible contaminant. Visual inspection of the sequence data for both MN-11 C2 and Boone suggested possible contamination that could not be rigorously disproven, and these samples were therefore excluded from further analysis.

Sample	Lanes	Insert size	Cycles	Depth	Bases (pre-QC)	Bases (post-QC)
Ackerman_1	2	250	75	35.95x	679.18 Mb / 679.96 Mb	113.20 Mb / 113.33 Mb
Ackerman_2	2	250	75	84.17x	1.70 Gb / 1.71 Gb	106.02 Mb / 100.43 Mb
Ackerman_3	2	250	75	45.98x	860.58 Mb / 860.34 Mb	107.57 Mb / 107.54 Mb
L747_1	2	250	75	57.27x	1.07 Gb / 1.08 Gb	107.26 Mb / 108.17 Mb
L747_2	2	250	75	68.27x	1.29 Gb / 1.30 Gb	107.51 Mb / 108.31 Mb
L747_3	2	250	75	67.39x	1.27 Gb / 1.27 Gb	106.12 Mb / 106.05 Mb
L810_1	2	250	75	101.29x	1.94 Gb / 1.94 Gb	102.00 Mb / 102.09 Mb
L810_2	2	250	75	62.44x	1.19 Gb / 1.19 Gb	108.52 Mb / 108.47 Mb
L810_3	2	250	75	111.86x	2.14 Gb / 2.13 Gb	102.10 Mb / 101.23 Mb
MA-37_1	2	250	75	89.46x	1.73 Gb / 1.73 Gb	101.84 Mb / 101.76 Mb
MA-37_2	2	250	75	70.59x	1.31 Gb / 1.30 Gb	100.86 Mb / 108.33 Mb
MA-37_3	2	250	75	70.74x	1.33 Gb / 1.34 Gb	102.50 Mb / 103.27 Mb
MN-11_1	2	250	75	72.45x	1.35 Gb / 1.35 Gb	104.03 Mb / 104.15 Mb
MN-11_2	2	250	75	91.26x	1.70 Gb / 1.70 Gb	100.12 Mb / 100.10 Mb
MN-11_3	2	250	75	62.79x	1.17 Gb / 1.17 Gb	106.22 Mb / 106.78 Mb
Kubba_1	2	250	75	76.00x	1.42 Gb / 1.44 Gb	101.76 Mb / 102.51 Mb
Kubba_2	2	250	75	62.01x	1.16 Gb / 1.17 Gb	105.72 Mb / 106.52 Mb
Kubba_3	2	250	75	66.46x	1.25 Gb / 1.25 Gb	103.81 Mb / 103.75 Mb
Melloy_1	2	250	75	72.51x	1.35 Gb / 1.33 Gb	103.66 Mb / 102.60 Mb
Melloy_2	2	250	75	74.65x	1.38 Gb / 1.39 Gb	106.11 Mb / 106.86 Mb
Melloy_3	2	250	75	76.13x	1.41 Gb / 1.41 Gb	100.72 Mb / 100.83 Mb
Boone_1	2	250	75	66.21x	1.23 Gb / 1.23 Gb	102.28 Mb / 102.44 Mb
Boone_2	2	250	75	67.56x	1.25 Gb / 1.26 Gb	104.24 Mb / 104.89 Mb
Boone_3	2	250	75	67.27x	1.25 Gb / 1.25 Gb	104.35 Mb / 104.26 Mb
KK27_1	2	250	75	76.37x	1.43 Gb / 1.43 Gb	101.79 Mb / 101.88 Mb
KK27_2	2	250	75	73.99x	1.39 Gb / 1.39 Gb	106.66 Mb / 106.70 Mb
KK27_3	2	250	75	70.15x	1.30 Gb / 1.31 Gb	108.26 Mb / 100.91 Mb
Rupert_1	2	250	75	69.90x	1.48 Gb / 1.49 Gb	105.49 Mb / 106.47 Mb
Rupert_2	2	250	75	60.72x	1.30 Gb / 1.30 Gb	108.12 Mb / 108.06 Mb
Rupert_3	2	250	75	51.15x	1.08 Gb / 1.09 Gb	108.35 Mb / 109.32 Mb
Azad_1	2	250	75	107.47x	1.99 Gb / 1.99 Gb	104.99 Mb / 104.92 Mb
Azad_2	2	250	75	54.80x	1.02 Gb / 1.02 Gb	101.81 Mb / 101.88 Mb
Azad_3	2	250	75	73.77x	1.37 Gb / 1.38 Gb	105.02 Mb / 105.87 Mb
K26_1	2	250	75	38.28x	731.92 Mb / 735.98 Mb	104.56 Mb / 105.14 Mb
K26_2	2	250	75	54.65x	1.03 Gb / 1.03 Gb	102.96 Mb / 102.96 Mb
K26_3	2	250	75	53.70x	1.01 Gb / 1.02 Gb	100.72 Mb / 101.66 Mb
Rupert C1_1	2	250	75	63.63x	1.20 Gb / 1.18 Gb	108.68 Mb / 107.60 Mb
Rupert C1_2	2	250	75	60.85x	1.13 Gb / 1.14 Gb	103.17 Mb / 104.08 Mb
Rupert C1_3	2	250	75	71.66x	1.34 Gb / 1.35 Gb	102.72 Mb / 103.80 Mb
Rupert C2_1	2	250	75	64.16x	1.22 Gb / 1.23 Gb	101.30 Mb / 102.41 Mb
Rupert C2_2	2	250	75	61.09x	1.16 Gb / 1.17 Gb	105.64 Mb / 106.81 Mb
Rupert C2_3	2	250	75	58.25x	1.10 Gb / 1.11 Gb	100.25 Mb / 101.22 Mb
MN-11 C1_1	2	250	75	66.27x	1.24 Gb / 1.25 Gb	102.97 Mb / 104.04 Mb
MN-11 C1_2	2	250	75	63.85x	1.19 Gb / 1.20 Gb	108.06 Mb / 100.04 Mb
MN-11 C1_3	2	250	75	64.13x	986.31 Mb / 986.71 Mb	109.59 Mb / 109.63 Mb
MN-11 C2_1	2	250	75	66.35x	1.24 Gb / 1.26 Gb	103.59 Mb / 104.59 Mb
MN-11 C2_2	2	250	75	73.83x	1.40 Gb / 1.39 Gb	100.27 Mb / 106.88 Mb
MN-11 C2_3	2	250	75	67.76x	1.29 Gb / 1.30 Gb	107.52 Mb / 100.15 Mb
MA-37 C1_1	2	250	75	57.61x	1.06 Gb / 1.07 Gb	105.97 Mb / 107.02 Mb
MA-37 C1_2	2	250	75	60.02x	1.19 Gb / 1.20 Gb	107.91 Mb / 109.00 Mb
MA-37 C1_3	2	250	75	67.76x	1.29 Gb / 1.30 Gb	107.52 Mb / 100.15 Mb
Kubba C1_1	2	250	75	49.62x	923.22 Mb / 931.89 Mb	102.58 Mb / 103.54 Mb
Kubba C1_2	2	250	75	60.80x	1.13 Gb / 1.14 Gb	102.98 Mb / 103.93 Mb
Kubba C1_3	2	250	75	55.78x	1.04 Gb / 1.05 Gb	104.02 Mb / 105.00 Mb

Table 3.2. Number of lanes, total depth, and total number of bases obtained for RNA-seq runs of each set of triplicates for the 18 cloned and uncloned lines

that were submitted for sequencing. The Array Express accession number for these samples is E-ERAD-408.

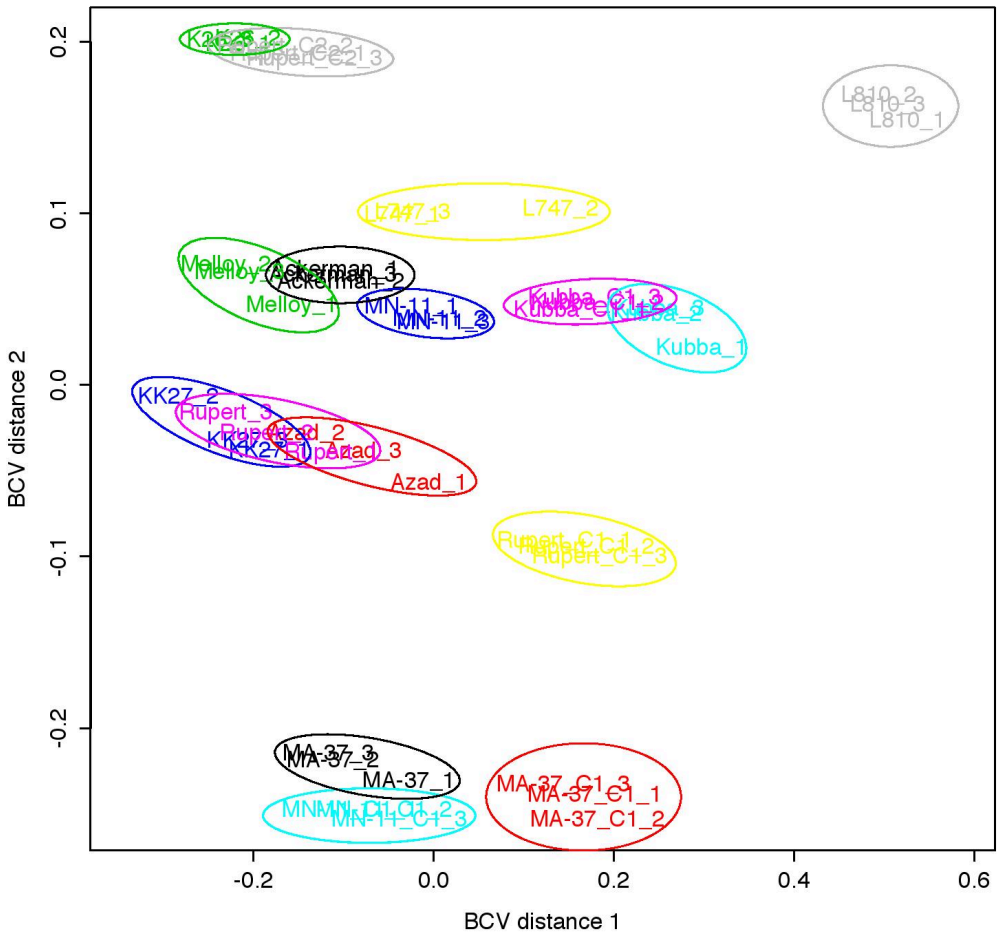


Figure 3.3. MDS plot representing expression data for all strains, except for isolates Boone, E50, and K112 (please refer to main text for description of rationale for their exclusion). Each triplicate is depicted in a different color. The replicates are numbered as per the legend included.

In order to confirm the divergent expression signature of the L810 isolate, Euclidean distances were calculated using variance-stabilized relative-log expression (RLE) normalized read counts (Figure 3.4). These distances confirmed L810 to be very divergent from the rest of the isolates in our sample (mean distance value for L810 = 50.94, SD = 13.50; mean for all other isolates = 34.97, SD = 2.984). These results were consistent with the clustering seen by MDS (Figure 3.3).

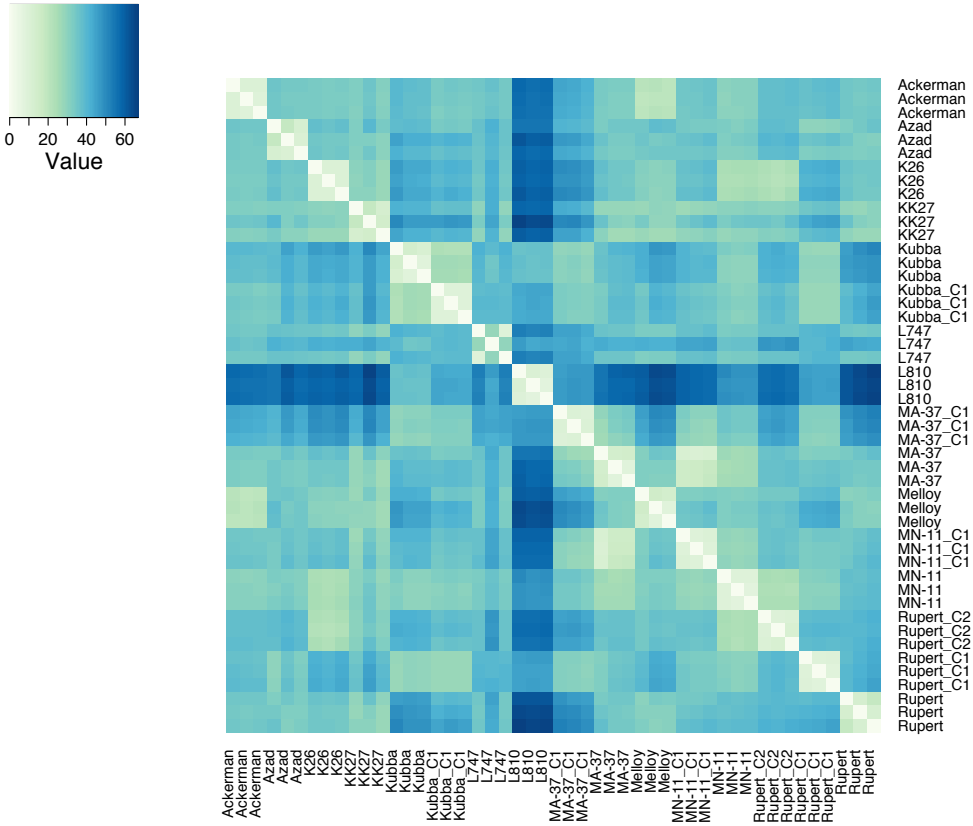


Figure 3.4. Heatmap of Euclidean distances between variance stabilized expression values for each pair of samples. MN-11 C2 and Boone showed aberrant expression signatures, and were later excluded since they were possible contaminations of the cultures.

Before identifying differentially expressed genes in this set of isolates, we searched the data to identify the most highly expressed genes in the axenic promastigote stage. The top 30 genes across all samples with the highest mean RLE-normalized read counts were identified, and the variance-stabilized expression values plotted (Figure 3.4). The most highly expressed gene in all isolates was beta-tubulin (LmjF.21.1860 in the *L. major* annotation), a well known promastigote-specific marker, known to be present in multiple orthologous copies in the genome. The top cluster of highly expressed genes included several transporter proteins, including an inosine/guanosine transporter (NT2) and two putative nucleoside transporter proteins. The next cluster of highly expressed genes included two amino-acid transporters (AAT1.4 and AAT19, named LmjF.31.0350 and LmjF.07.1160 in the *L. major* annotation), a L-lysine transporter (AAT16, or LmhF.32.2660), a glucose transporter (GT1, or LmjF.36.6300), and a putative pteridine transporter (LmjF.06.1260). Among the other highly expressed genes were many ribosomal components involved in translation of mRNA.

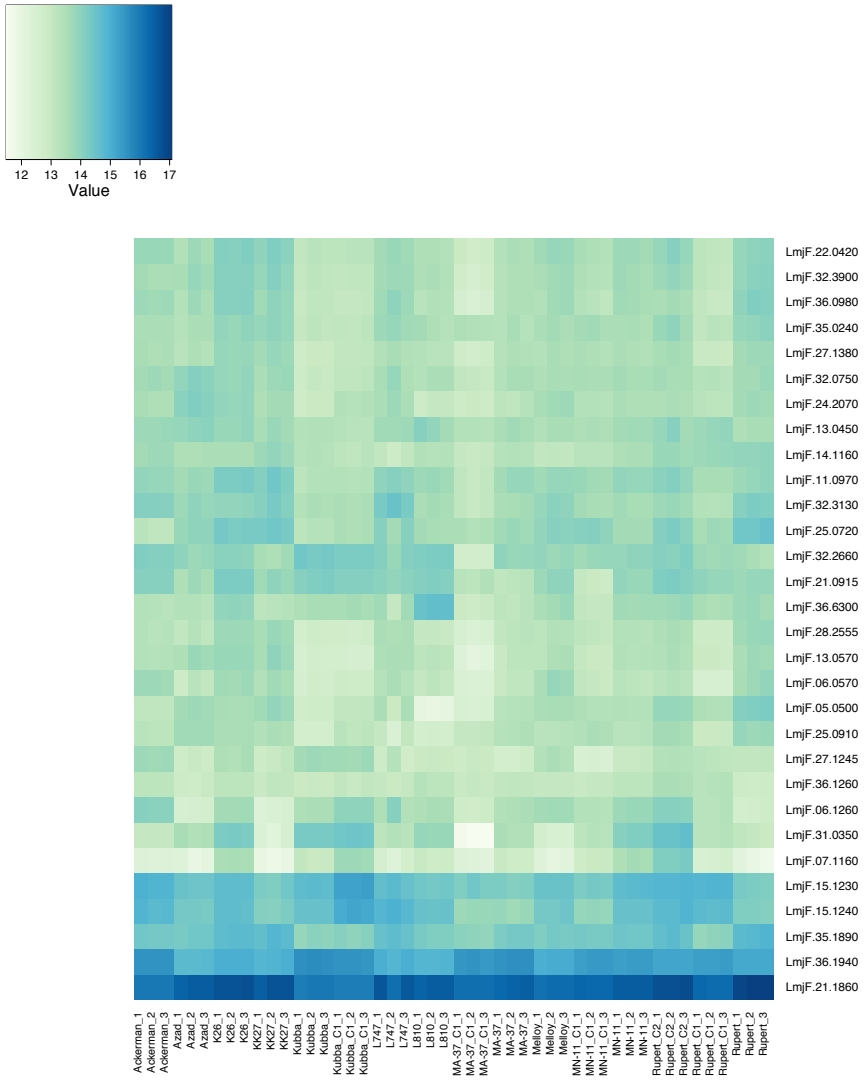


Figure 3.5. Heatmap showing the most highly expressed genes in our set of samples. The shading of each cell represents variance-stabilized expression values as implemented in DEseq. Genes were organized by hierarchical clustering, with the most highly expressed genes at the bottom of the graph. See Appendix C for list of genes.

3.3.3. Differential gene expression analysis

In order to measure differential expression between multiple sets of triplicates, isolate L747 was chosen as the baseline to which all other expression levels were compared to calculate log-fold changes (see Methods section 3.2.4 for rationale). Given the sensitivity of our study design, almost all genes were differentially expressed (DE) in at least one isolate (98.99% of all genes at FDR < 0.05, 98.26% at FDR < 0.001). The top 30 genes with the smallest p-values were selected from the F distribution (FDR < 10^{-40} for all selected genes) to generate a workable set of highly DE genes, and observe how their expression varied within our sample set.

Comparing log-fold changes in expression in each set of triplicates for these 30 genes showed that two samples (K26, Rupert C2) had dramatic downregulation of a folate transporter protein (FT1, LmjF.10.0385 in the *L. major* annotation), with concurrent upregulation of a biopterin transporter protein (BT1, LmjF.35.5150 in the *L. major* annotation). In a third sample (MN-11), the downregulation of FT1 was not as pronounced, but upregulation of BT1 was still apparent (Figure 3.5). These two transporter proteins are known to have a related function, and act in concert in pterin/folate metabolism in *Leishmania* parasites. Inspection of sequencing read depth in this region confirmed a duplication of the BT1 gene and a deletion of the FT1 gene in these two samples. A cluster of 10 genes, arranged as an array on chromosome 24 (LmjF.24.1010 to LmjF.24.1100), was comprised of highly significant DE genes (10^{-71} > FDR > 10^{-37}), and was upregulated in these same three

samples. The protein products of this DE gene cluster appeared to have unrelated functions, but included a multipass transmembrane protein (LmjF.24.1090).

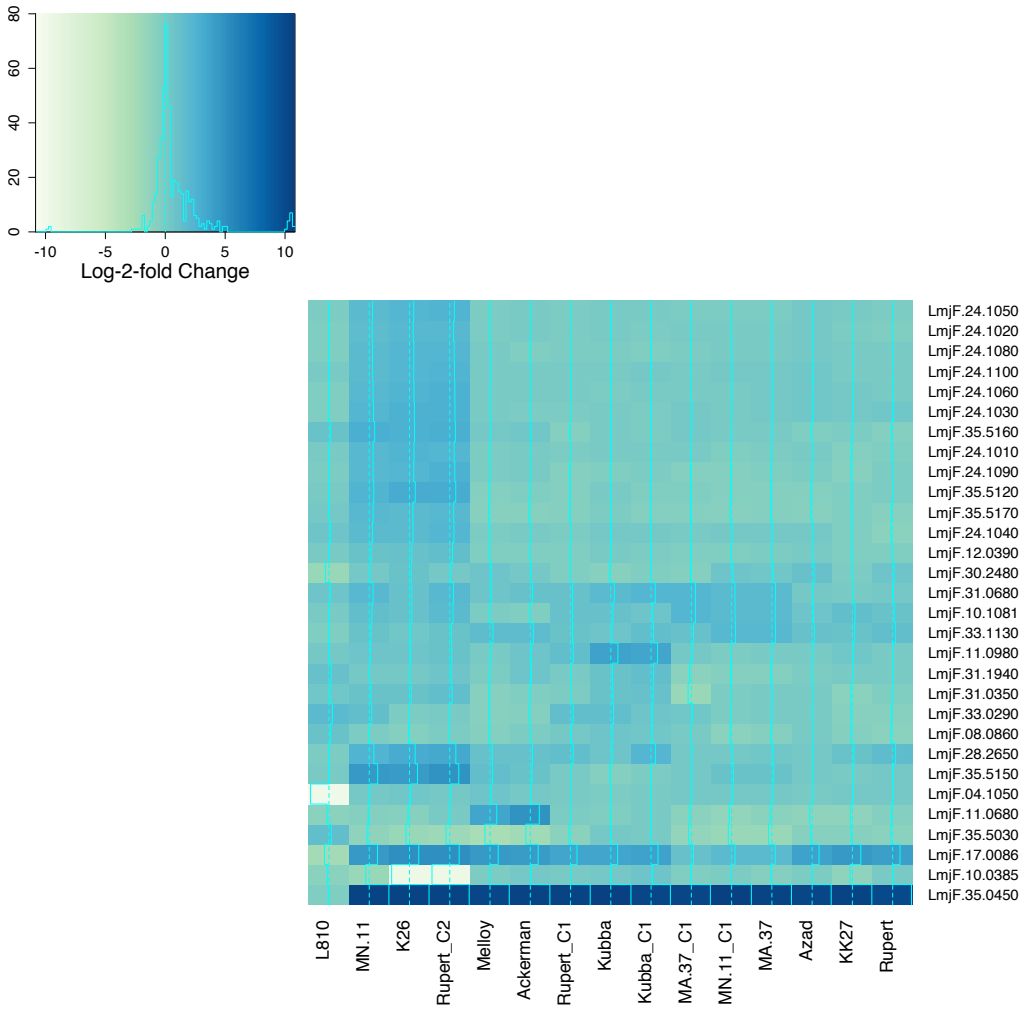


Figure 3.6. Heatmap of the top 30 most significant DE genes showing log-fold change in expression. Note the large cluster of DE genes on chromosomes 24 (gene name starting with LmjF.24) and 35 (gene name starting with LmjF.35) in 4 isolates, MN11, MN11 C1, Rupert C2, and K26. Inspection of the whole genome sequence data suggests that biopterin transporter 1 (LmjF.35.5150)

and folate transporter 1 (LmjF.10.0385) are duplicated and deleted, respectively, in these isolates. See Table 3.3 for list of genes.

Of the top 30 DE genes, 10 had transmembrane domains (Table 3.3). Of the 1546 genes containing 1 or more transmembrane domains in *L. major* as annotated in TriTypDB, 1344 are also present in the *L. tropica* annotation used for this study. By the hypergeometric distribution, the probability that at least 10 genes coding for membrane-associated protein products would be present by chance in a random sample of 30 genes is 0.008063, under the null hypothesis that there is no association between differential expression and presence of transmembrane domains in the protein product (p-value < 0.05 by Fisher's exact test), confirming a significant enrichment of transmembrane proteins among DE genes.

Gene	P-value	FDR	TM	Product
LmjF.35.5150	3.22E-78	2.53E-74	x	Biopterin transporter 1 (BT1)
LmjF.24.1090	3.98E-75	1.57E-71	x	Hypothetical predicted multipass transmembrane protein
LmjF.35.5120	2.64E-70	5.50E-67		Hypothetical protein
LmjF.11.0980	2.80E-70	5.50E-67		Hypothetical protein
LmjF.35.5160	1.28E-68	2.01E-65		P-loop containing nucleoside triphosphate hydrolase-like protein
LmjF.11.0680	6.93E-68	9.08E-65	x	MFS general substrate transporter, conserved
LmjF.24.1060	3.77E-62	4.23E-59		hypothetical protein, conserved
LmjF.24.1100	7.34E-61	7.21E-58		pre-mRNA-splicing factor ATP-dependent RNA helicase, putative
LmjF.24.1080	3.39E-60	2.83E-57		DNAJ domain protein, putative
LmjF.35.0450	3.59E-60	2.83E-57		hypothetical protein, unknown function
LmjF.35.5030	3.86E-59	2.76E-56		hypothetical protein, conserved
LmjF.24.1030	4.35E-59	2.85E-56		dynein light chain, putative
LmjF.24.1050	1.14E-56	6.91E-54		SNF2 family protein
LmjF.28.2650	8.12E-56	4.56E-53	x	membrane-bound acid phosphatase, putative
LmjF.24.1020	5.78E-55	3.03E-52		Kelch beta propeller type protein, unknown function
LmjF.24.1040	7.95E-54	3.91E-51		hypothetical protein, unknown function
LmjF.35.5170	1.03E-53	4.78E-51		P-loop containing nucleoside triphosphate hydrolase-like protein
LmjF.31.0350	1.84E-53	8.03E-51	x	amino acid transporter aATP11, putative
LmjF.10.0385	2.77E-53	1.15E-50	x	Folate transporter 1 (FT1)
LmjF.33.0290	1.29E-52	5.09E-50	x	glucose transporter/membrane transporter D2, putative
LmjF.31.0680	1.89E-52	7.09E-50		C2 calcium-dependent membrane targeting protein, conserved
LmjF.24.1010	3.62E-52	1.29E-49		Meiotic nuclear division protein 1-related protein
LmjF.33.1130	1.75E-51	5.97E-49	x	hypothetical protein, conserved
LmjF.04.1050	2.11E-51	6.91E-49	x	acyltransferase-like protein, copy 2
LmjF.17.0086	6.85E-51	2.08E-48		elongation factor 1-alpha
LmjF.08.0860	6.89E-51	2.08E-48		hypothetical protein, unknown function
LmjF.12.0390	7.30E-51	2.13E-48		hypothetical protein, conserved
LmjF.10.1081	1.11E-50	3.11E-48		hypothetical protein, conserved
LmjF.31.1940	2.92E-50	7.93E-48	x	transcription like protein nupm1, putative
LmjF.30.2480	1.55E-44	4.36E-42		heat shock 70-related protein 1, mitochondrial precursor, putative

Table 3.3. A list of the top 30 most significant differentially expressed genes from Figure 3.6. TM stands for "transmembrane" and genes marked with an "X" contain one or more transmembrane domains as predicted by TMHMM algorithm implemented in GeneDB.

Given the divergent pattern observed for L810, we performed pairwise likelihood ratio tests between L810 and the rest of the samples, considered as two separate conditions. We identified 46 gene transcripts that were significantly differentially expressed in L810 more than twofold (Figure 3.6). The two most upregulated transcripts in L810 (> 2 log-fold change) in comparison to all other samples were LmjF.17.0190, a receptor-type adenylate cyclase, and LmjF.31.2100, an unknown protein (see Appendix D for the complete list of DE genes). The majority of the downregulated transcripts in L810 belonged to unknown proteins. The two most downregulated transcripts in L810 (> 6 log-fold change) were an unknown protein and an acyltransferase-like protein (LmjF.35.0450, LmjF.04.1050).

A pairwise exact test between two separate clones of the Rupert isolate was performed to allow a more focused analysis of gene dosage effects in samples with a comparable genetic background and ploidy. There were a total of 4634 significantly DE genes found between these two samples with a FDR < 0.05 and a log-fold change greater than 1. We then proceeded to study these differences in gene expression in the context of observed copy number variation between these two clonal lines.

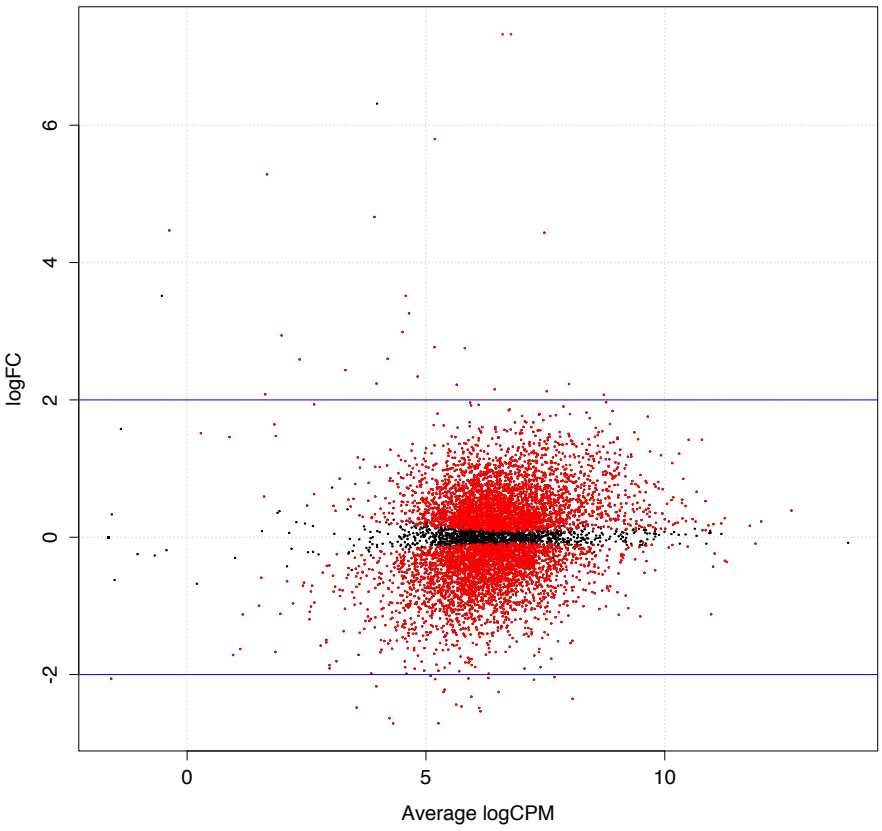


Figure 3.7. Smear plot of log-fold changes in expression versus average log-counts per million in isolate L810 compared to all other isolates. Each dot is a gene, and dots colored in red is significant at the FDR < 0.05 threshold. The blue bars represent the +2 and -2 log-fold change thresholds. There were a total of 46 genes that were significant and that fell above or below the 2 log-fold change thresholds. Negative expression values represent genes that are downregulated in all samples compared to L810 (therefore, upregulated in L810), while positive expression values represent genes that are upregulated

in all samples compared to L810 (downregulated in L810). See Appendix D for list of DE genes in L810 compared to all other isolates.

3.3.4. Copy number variation and gene dosage effects

Sequencing of two separate clonal lines from one of the field isolates (Rupert C1 and Rupert C2) showed considerable variation in karyotype and in gene copy number. Somy was found to differ between the two clones at 10 chromosomes. In each case, the log-fold change of DE genes on these chromosomes, calculated as the difference between the two clones (in this case, Rupert C2 – Rupert C1), was found to be shifted in the direction of the chromosome with the higher somy (e.g. if somy of Rupert C2 was greater than Rupert C1, then the difference in expression between Rupert C2 and Rupert C1 was positive due to higher expression of these genes in Rupert C2). The relative up- and down-regulation seemed to behave in a dose-dependent manner, with larger somy being correlated with a larger log-fold change in expression (Figure 3.7). Visual inspection of raw read counts within each of these clones confirmed that supernumerary chromosomes (i.e. chromosomes with a somy larger than 2) had higher absolute RNA-seq read counts than disomic chromosomes. A total of 156 copy number variants (CNVs) were identified between Rupert C1 and Rupert C2 (median size = 2255bp, comprising 0.9% of the genome), overlapping 64 significantly DE genes (See Appendix E for complete list of genes and their log fold changes in expression). Four large copy number variant regions on chromosome 23, 24, 27 and 35 were found spanning a total of 48 DE genes (75% of all DE genes in

CNVs). While the CNV regions on chromosome 23, 24 and 35 were up-regulated in Rupert C2, the CNV region on chromosome 27 was up-regulated in Rupert C1 (Figure 3.7). A total of 17 DE genes contained within CNVs had transmembrane domains annotated by algorithmic prediction (probability of observing this many TM genes by chance = 0.01862). Among the DE genes present in CNVs were several ABC transporters previously implicated in drug resistance, such as the multidrug resistance protein A (MRPA, LmjF.23.0250).

The relative expression differences in the majority of genes in these CNV regions were consistent with the whole-genome sequencing read depth differences between these two samples (Figure 3.8). Regions with higher relative depth in Rupert C2 showed positive log fold-changes in gene expression, whereas regions with higher relative depth in Rupert C1 showed negative log fold-changes in gene expression. A few genes, however, did not follow this general rule (two on chromosome 23, three on chromosome 24, one on chromosome 27, and one on chromosome 35), suggesting the presence of a mechanism regulating gene transcript abundance independently of gene dosage.

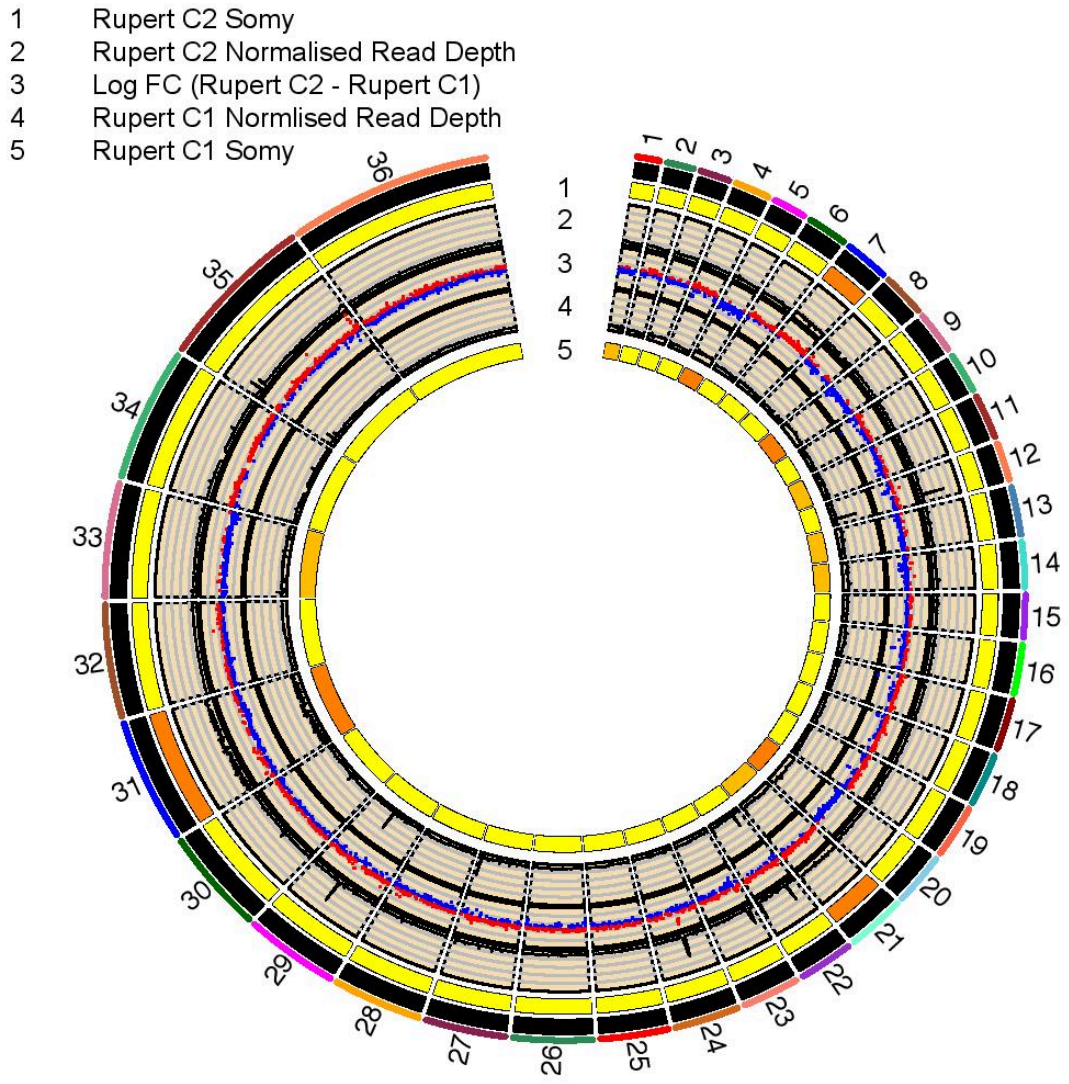


Figure 3.8. Gene dosage effects due to aneuploidy on transcription in *L. tropica* clonal lines Rupert C1 and Rupert C2. Tracks 1 and 5 represent somy as in Figure 3.1. Tracks 2 and 4 represent normalized read depth. Track 3 represents significant DE genes between these two clones. Positive log-fold change is in red, and represents overexpression in Rupert C2. Negative log-fold change is in blue, and represents overexpression in Rupert C1.

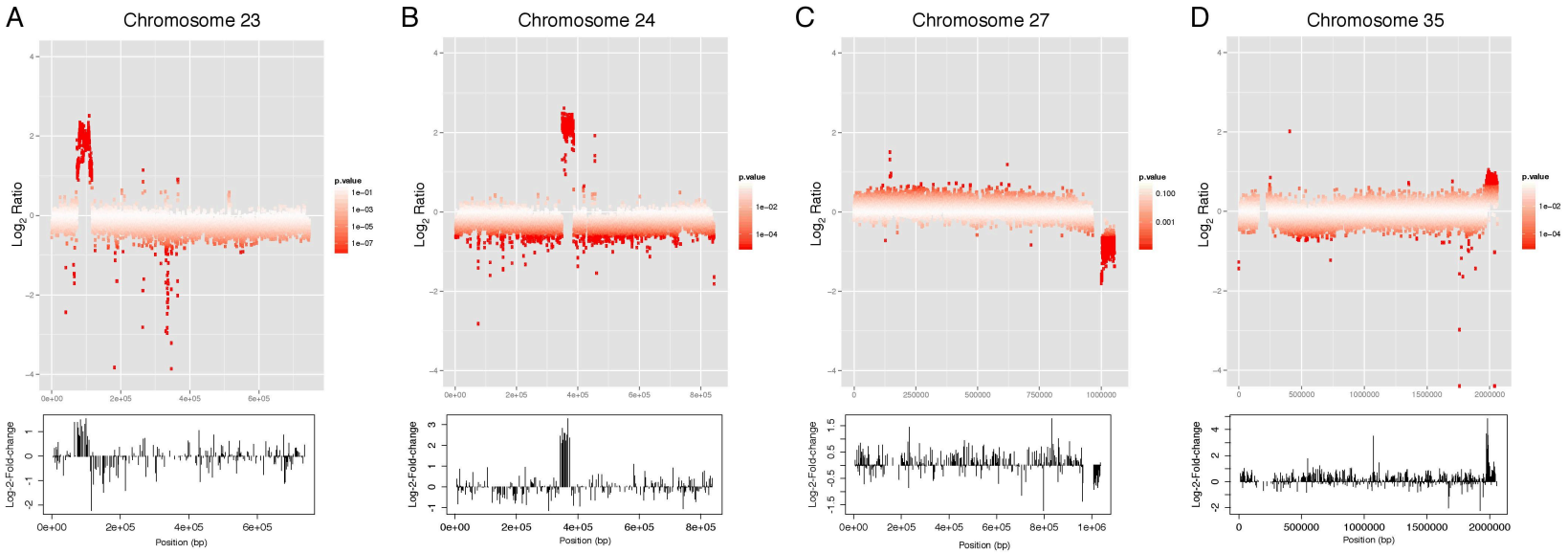

A
Chromosome 23
B
Chromosome 24
C
Chromosome 27
D
Chromosome 35
Log2 Ratio
p.value
Log-2-Fold-change
Position (bp)

Figure 3.9. Gene dosage effects due to copy number variation on transcription in *L. tropica* clonal lines Rupert C1 and Rupert C2. The top graph shows log-2 ratio of WGS read depth of Rupert C2 versus Rupert C1 across a chromosome. The bottom graph instead shows log-scale fold change in gene expression on the chromosome. Panels A through D represent the 4 largest CNVs on different chromosomes, each spanning several genes. Relative changes in expression match changes in WGS read depth in these four large CNVs.

3.3.5. Allele-specific gene expression

A genome-wide search for allele-specific gene expression was performed by comparing relative allele frequencies in the read depth information obtained from RNA and DNA sequencing of the same sample (see Methods). The two clones obtained from the Rupert isolate showed significant signals in different regions of the genome, as evidenced through Manhattan plots (Figure 3.9). One gene on chromosome 1 (LmjF.01.0830) showed highly significant p-values in both clones. A total of 641 SNPs carried significant p-values (p-value < 10^{-10}, 0.28% of all SNPs) in Rupert C2, 170 of which fell into coding regions, and 52 in putative 3' UTRs. A total of 753 SNPs (0.33% of all SNPs) were significant in Rupert C1, 187 of which were in coding regions, and 58 were in putative 3' UTRs. The majority of allele-specific expressed genes were shared between the two isolates (48 genes, 70% of genes in C1, 73% of genes in C2; see Table 3.4).

Comparison of raw sequencing read depth from RNA and DNA for one particular gene with highly significant p-values in both clones and a high number of heterozygous SNPs (LmjF.01.0830, coding for a metallo-peptidase, Clan MA(E), Family M3, with > 10 significant SNPs) confirms allele specific gene expression, with clear differences in the alternate allele frequencies in the DNA sequence data compared to the RNA data in both Rupert C1 and Rupert C2 (Figure 3.9). Reconstruction of the gene's haplotype from RNA-seq data using read-pair based phasing also found a consistent homozygous pattern across the length of the transcript. By comparison, the haplotype reconstructed from the DNA sequence data showed a consistent heterozygous pattern, suggesting that only one allelic variant is being expressed. Interestingly, the biopterin transporter 1 (BT1) gene had a very strong signal in Rupert C2, suggesting both upregulation of this gene via amplification (either extra- or intra-chromosomal) in this clone and allele-specific gene expression.

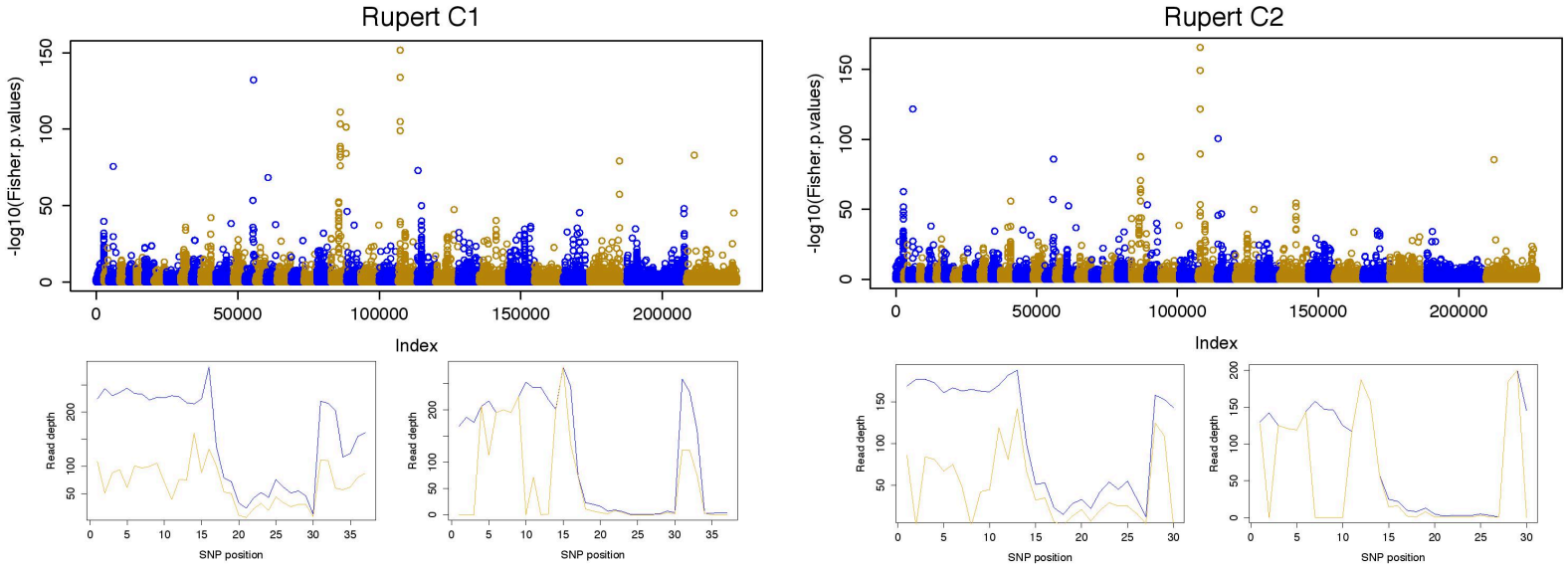

Rupert C1
-log10(Fisher.p.values)
0
50
100
150
0
50000
100000
150000
200000
Index
Read depth
SNP position
Rupert C2
-log10(Fisher.p.values)
Index
Read depth
SNP position

Figure 3.10. Evidence for allele-specific gene expression in the two clones Rupert C1 and Rupert C2, with significant hits having a p-value < 10^{-10}. The Manhattan plots for each isolate represent log-scale Fisher p-values on the y axis and SNP index on the x axis (see Methods section). The bottom two panels represent DNA read depth on the left, and RNA read depth on the right, at each SNP position in LmjF.01.0830, a metallo-peptidase that was one of the most significant hits in both clones and that had a large number of significant SNP positions (>10 SNPs). The blue line represents total read depth while the golden line represents the number of reads bearing the alternate allele at that position. The RNA read depth appears to be homozygous (the golden and blue lines overlap for SNP positions with sufficient coverage) while the DNA read depth appears to be heterozygous (the golden line is approximately half the height of the blue line).

Gene	Product	TM	Sig SNPs Rupert C1	Sig SNPs Rupert C2
LmjF.01.0540	hypothetical protein, conserved		1	1
LmjF.01.0830	metallo-peptidase, Clan MA(E), Family M3		15	13
LmjF.03.0430	60S acidic ribosomal protein P2, putative		1	1
LmjF.05.0240	viscerotropic leishmaniasis antigen, putative	x	2	5
LmjF.09.1090	translation initiation factor EIF-2B gamma subunit		1	1
LmjF.10.0360	folate/biopterin transporter, putative	x	4	5
LmjF.10.0370	folate/biopterin transporter, putative	x	5	2
LmjF.10.0380	folate/biopterin transporter, putative	x	1	2
LmjF.11.0630	metallo-peptidase, Clan MF, Family M17		1	1
LmjF.11.0680	hypothetical protein, conserved in leishmania	x	1	2
LmjF.14.0690	fatty acid elongase, putative (ELO3.2)	x	5	3
LmjF.14.1100	kinesin K39, putative		3	2
LmjF.14.1440	hypothetical protein, conserved		1	1
LmjF.15.1230	nucleoside transporter 1, putative	x	1	1
LmjF.15.1240	nucleoside transporter 1, putative	x	5	4
LmjF.15.1520	hypothetical protein, conserved		1	1
LmjF.16.1460	kinesin, putative		4	4
LmjF.17.1220	histone H2B		2	1
LmjF.18.1520	P-type H+-ATPase, putative (H1A-2)	x	2	1
LmjF.11.0680	hypothetical protein, conserved in leishmania		15	13
LmjF.19.1000	glycerol uptake protein, putative	x	1	1
LmjF.21.0240	hexokinase, putative		3	2
LmjF.21.0250	hexokinase, putative		2	1
LmjF.21.0725	hypothetical protein, conserved		1	1
LmjF.22.0850	3'a2rel-related protein	x	1	1
LmjF.25.0160	hypothetical protein, conserved		1	1
LmjF.26.1240	heat shock protein 70-related protein (HSP70.4)		12	12
LmjF.26.2170	cornifin-like protein, unknown function		2	2
LmjF.27.0240	kinetoplast-associated protein-like protein		3	3
LmjF.27.0670	Amino acid permease, putative (AAT23.1)	x	7	1
LmjF.29.1490	Asparagine synthase-related protein, conserved		5	3
LmjF.30.2550	heat shock 70-related protein 1, mitochondrial precursor		8	8
LmjF.31.0820	hypothetical protein, conserved		3	1
LmjF.31.2330	3,2-trans-enoyl-CoA isomerase, mitochondrial precursor		2	2
LmjF.32.2270	membrane associated protein-like protein		1	2
LmjF.33.2270	hypothetical protein, conserved		1	1
LmjF.33.2300	udp-glc 4'-epimerase, putative		1	1
LmjF.34.0690	flagellar attachment zone protein, putative (FAZ1)	x	1	1
LmjF.34.2560	hypothetical protein, conserved		2	1
LmjF.34.2800	tuzin, putative	x	4	4
LmjF.35.0010	phosphoglycan beta 1,3 galactosyltransferase 7 (SCG7)	x	4	5
LmjF.35.0500	proteophosphoglycan ppg3, putative		1	2
LmjF.35.0550	proteophosphoglycan ppg1	x	1	1
LmjF.35.1310	histone H4		1	1
LmjF.35.1410	threonyl-tRNA synthetase, putative		1	1
LmjF.35.1670	60S ribosomal protein L26, putative		1	1
LmjF.35.4420	mitochondrial phosphate transporter		4	2
LmjF.36.6300	glucose transporter 1 (GT1)	x	5	8
LmjF.36.6480	histidine secretory acid phosphatase, putative		1	1

Table 3.4. A list of the genes with evidence of allele specific gene expression in both clones originating from the Rupert isolate.

3.4. Discussion

Overall, the present study confirms previous reports of significant intra-specific heterogeneity within *L. tropica,* and confirms that structural genomic variation provides an added layer of complexity in addition to simple sequence variation. The variation observed in genome structure at the level of somy and copy number variation underlie most differences in gene expression within our set of isolates via gene dosage effects, suggesting that this variation may be functional and serve an evolutionarily adaptive purpose. We also confirm previous reports that clinical isolates of *Leishmania* are mosaics of closely related parasites, differing particularly in chromosome copy number and smaller structural variants (CNVs), by comparison of different lines generated via multiple independent cloning of the same isolate. This mosaicism may maximize the presence of standing variation within a population of parasites in a single host, and in this way help the parasite cope with certain selective pressures, such as drug selection or nutrient deficiency. The mechanisms giving rise to this mosaicism remain unknown, although unequal chromosome replication during mitotic cell division appears to be the most likely immediate cause of mosaic aneuploidy (Sterkers, Lachaud et al. 2011).

Differentially expressed genes in these 20 isolates (i.e., genes that varied more *across* triplicates than *within* triplicates) appear to be significantly enriched in transmembrane proteins, especially transporter proteins. These surface proteins are known to play an important role in uptake of essential nutrients from the external environment, as well as import of drug compounds into the cell (Marquis,

Gourbal et al. 2005, Leprohon, Legare et al. 2006, Mandal, Mandal et al. 2015). The BT1 and FT1 transporters are among the most well studied examples associated with *in vitro* resistance to the antifolate drug methotrexate (Cunningham and Beverley 2001, Ouameur, Girard et al. 2008). The redundancy observed in the function of many of these transporters and the relative ease with which their transcript levels can be regulated via amplification or deletion of the corresponding protein-coding gene suggest that the natural variation observed in our study may be an important pre-adaptation for survival in nutrient-poor environments or in environments that are otherwise hostile to the parasite.

RNA-seq analysis of this set of isolates found one isolate (LRC-L810) to have a very different expression signature to all other isolates. Interestingly, this sample was isolated from an infected sand fly in Northern Israel, and additional studies found this strain to be preferentially transmitted by a different vector species than other *L. tropica* isolates originating from the same region (Soares, Barron et al. 2004). One of the two most highly upregulated genes in this isolate compared to all other isolates was LmjF.17.0190, a receptor-type adenylate cyclase. This type of protein has been linked in African trypanosomes to differentiation of epimastigote into trypomastigote forms (Fraidenraich, Pena et al. 1993), inhibition of the host immune response (Salmon, Vanwalleghem et al. 2012), and coordinated social motility in the insect stages (Lopez, Saada et al. 2015). Most other differentially expressed genes had unknown functions (see Appendix D). Further functional studies are needed to shed additional light on the role of these strain-specific differences in gene expression.

Focusing our analysis of gene expression on the comparison between two different clonal lines obtained from the same field isolate (Rupert C1 and Rupert C2, from isolate Rupert) suggests that transcript levels in each of these clones correlate with chromosome number. Our initial hypothesis was that in order to minimize negative repercussions on the overall cellular homeostasis of the parasite, genes on extranumerary chromosomes had to be downregulated so that overall steady state transcript levels matched those found in diploid parasites. However, this was not the case in our data: higher somy is consistently associated with both higher relative and absolute expression values of the genes on these chromosomes in the two aneuploid clones that we considered. Comparing raw read counts within each of these two genomes showed that supernumerary chromosomes, such as chromosome 31, had higher expression than disomic chromosomes. Most genome-wide differences in gene expression between these two clones can thus be attributed to differences in somy, with a clear proportional effect seen as somy increases.

In addition to large variation in chromosome number between chromosomes, we also observed local variation in copy number within chromosomes. We identified 156 copy number variants (CNVs) between these two clones, which could be associated with differences in expression of 64 genes. Although these comprise a minority of DE genes, CNVs may be an evolutionarily important mechanism for parasites to rapidly upregulate (or downregulate) individual genes or gene clusters involved in drug resistance, as has been observed for ABC transporters in antimony-resistant *L. infantum* (Leprohon, Legare et al. 2009) and possibly relapsing *L.*

braziliensis and *L. panamensis* parasites following treatment with miltefosine (Obonaga, Fernandez et al. 2014). Certain regions of the genome are known to be more prone to deletion or amplification by virtue of being flanked with repetitive sequences that facilitate formation of both linear and circular amplicons via RAD51 recombinase-dependent and independent mechanisms (Ubeda, Raymond et al. 2014).

Given the well-recognized role played by RNA-binding proteins in regulation of gene expression during kinetoplastid parasite development (Kramer and Carrington 2011), we postulate that these might also be playing a role in determining expression of only one of the two alleles present at a given heterozygous coding region. However, given the remarkable genomic plasticity thus observed in *Leishmania*, we cannot rigorously exclude the possibility that some form of mitotic gene conversion could have occurred in some of these genes in the time spent in culture between DNA extraction and RNA extraction. The similar genomic positions of peaks with significant p-values observed in two independent clones, however, means that the majority of genes showing evidence of allele-specific gene expression in one clone also show preferential expression of one allele in the other clone, thus making the occurrence of the same gene conversion events in two independently *in vitro* cultured clonal lines highly unlikely. In our view, this fits better with a model predicting the presence of conserved regulatory elements in the 3'UTR of these genes which may be determining increased stability of one allelic transcript over the other at some heterozygous loci. Enrichment of SNPs with significant p-values in 3' UTRs would support this model.

In conclusion, we identify multiple layers of genomic variation that are controlling steady state mRNA transcript levels in *Leishmania* parasites. We observe considerable variation in genome structure in our set of *L. tropica* isolates, which represents most of the geographic distribution of this Old World species, and in the expression of genes associated with structural variants. The extent of genome plasticity observed in *L. tropica* may very well be larger than in other species, given the greater heterogeneity that has been documented at the sequence level in this species. Explicitly comparing populations of different *Leishmania* species could prove illuminating to identify the extent of species-specific genome plasticity. Interestingly, from a clinical perspective cutaneous lesions due to *L. tropica* generally show a poorer response to treatment than lesions due to *L. major* (Hadighi, Mohebali et al. 2006). A greater plasticity in terms of both gene regulation and copy number variation and a larger pool of standing intra-specific structural variation may facilitate and hasten the evolution of reduced sensitivity to drugs in this species.

In addition to gene dosage effects and the important role of trans-regulators in determining transcript stability, it is crucial to note that transcript abundance in *Leishmania* generally correlate poorly with cellular protein levels. Additional downstream processes may be shaping the proteomic landscape at the level of protein translation from this initial pool of mRNA transcripts.

The circumstances giving rise to both aneuploidy and gene copy number variation in *Leishmania* remain poorly understood. Specifically, there is a need to quantify the activity of these processes in both sexual and asexual stages of the life

cycle, and measure the relative contribution of each in generating genetic diversity. Such an understanding would provide valuable information to build a formally explicit mathematical model to understand and possibly predict how populations of this widespread human pathogen will change over time in an epidemiological context, especially in response to increased elimination efforts.

CHAPTER 4

EXPERIMENTAL CROSSES OF *L. TROPICA*

4.1. Introduction

Protozoan parasites of the genus *Leishmania* are thought to be facultative sexual organisms, capable of - at least occasionally - generating "hybrid" strains, bearing genetic markers of different parental lines of the same species (Akopyants, Kimblin et al. 2009, Sadlova, Yeo et al. 2011, Inbar, Akopyants et al. 2013, Calvo-Alvarez, Alvarez-Velilla et al. 2014, Rogers, Downing et al. 2014) or even different species (Romano, Inbar et al. 2014). Despite the fact that the frequency of these hybridization events in the wild is currently disputed, laboratory experiments have demonstrated that several species, including *L. major*, *L. infantum*, and *L. donovani*, can undergo genetic exchange during their extracellular growth and development inside the sand fly vector, resulting in hybrids bearing a full complement of genetic markers from each parent. These unicellular organisms therefore retain the molecular machinery necessary for carrying out cellular fusion and segregation of the parental genetic material to the progeny, and possibly also perform homologous recombination.

The first experimental cross in *Leishmania* was carried out in *Phlebotomus duboscqi* sand flies between *L. major* Friedlin and LV39 strains (Akopyants, Kimblin et al. 2009). Using clonal lines of each strain in which two different drug resistance

markers were introduced, hybrids that were resistant to both hygromycin B and nourseothricin were recovered from sand fly midguts. Attempts to recover double-drug resistant hybrid lines from *in vitro* co-culture or from *in vivo* co-infections of mice were unsuccessful. Out of 18 hybrid lines recovered, 7 showed 3n DNA content by propidium iodide staining, while the rest showed 2n DNA content. Intermediate levels of DNA content were not observed. SNP-CAPS analysis confirmed triploid and diploid patterns in the sequencing traces of these hybrids. At all markers where the parental lines were homozygous, the parental lines were heterozygous for each of the parental alleles. The hybrids were fully viable, and were capable of reinfecting mice and undergoing metacyclogenesis in the sand fly. Markers on the kDNA maxicircle appeared to follow uniparental inheritance rules.

Several other crosses have been performed in *L. major* with strains originating from different geographic regions. Inbar and colleagues (Inbar, Akopyants et al. 2013) recovered 9 additional double drug resistant lines from the same lines that were used in the first experimental cross, Friedlin SAT, resistant to nourseothricin, and LV39 HYG, resistant to hygromycin B. Pairing *L. majo*r Friedlin SAT with a different drug resistant strain, Sd HYG, gave yield to 15 additional hybrids. Lastly, crosses between the LV39 HYG strain and two other strains, Sd BSD and Ryan SAT, gave yield to 4 and 5 more hybrids, respectively. In addition to these crosses, which were performed in *P. duboscqi*, crosses were also performed in *Lutzomyia longipalpis*, a non-natural but permissive vector of *L. major*, between strains Friedlin SAT and LV39 HYG. A total of 61 hybrids were generated from this last cross.

These experiments suggest that *Leishmania* lacks distinct "mating types", and that there are no barriers to hybridization between strains originating from different geographic regions. Hybrids were recovered at different stages of metacyclogenesis, with no strict association between timing of hybrid recovery and developmental stages present in the sand fly midgut. The developmental stage most closely associated with hybrid recovery appeared to be the nectomonad stage, although more mature forms, such as haptomonads and metacyclics, may retain mating competency. Early replicating procyclic forms are unlikely to be mating competent, as *in vitro* promastigotes do not advance past these stages, and no hybrids have ever been recovered from selection of *in vitro* promastigote cultures. Mating competent developmental forms must therefore arise from late promastigote forms that only appear *in vivo* in the sand fly.

At the time this dissertation was written, a cross between *L. infantum* and *L. major* had also been successfully performed (Romano, Inbar et al. 2014). *L. major* Friedlin strain resistant to nourseothricin was crossed with *L. infantum* strain LLM-320 resistant to hygromycin B in *L. longipalpis* sand flies, and 11 double drug resistant hybrids were recovered. The hybrids lines differed in their phenotypes in mouse models of visceral and cutaneous disease, suggesting differential inheritance of genes involved in tissue tropism and pathogenesis.

The process through which parasite lines can undergo hybridization and the characteristics of this genetic exchange remain largely undescribed. In the first cross in *L. major*, the majority of hybrids were diploid, although triploid hybrids were also observed. In subsequent crosses, again triploid and even tetraploid hybrids were

seen. A tetraploid *L. major* hybrid line reverted to diploid and lost resistance to the two antibiotics following passage through the mouse, while all other lines were stably diploid or triploid both *in vitro* and *in vivo*. While kDNA markers were inherited uniparentally in all progeny clones, most nuclear markers followed biparental inheritance rules and were for the most part heterozygous for each parental allele. In a few cases, however, loss of heterozygosity and reversion to one of the two parental alleles was seen for a number of loci (< 3% of all genotypes markers). Whether this is due to aneuploidy, a common phenomenon in *Leishmania* (see Chapter 3), or to a more specific gene conversion mechanism remains unknown.

In the inter-specific cross between *L. major* and *L. infantum*, again diploid and triploid hybrid clones were recovered, in addition to a single tetraploid hybrid clone. The tetraploidy remained present following repeated *in vitro* passage and recovery of tissue amastigotes from infected mice, suggesting that polyploidy is not intrinsically unstable. Triploid - or near-triploid - hybrids showed intermediate tissue tropism phenotypes in a dose-dependent manner depending on the parental origin of the supernumerary chromosomes. As discussed in Chapter 3, gene dosage seems to have a major effect on transcription and may subsequently affect infection phenotypes, as seen for these inter-specific cross progeny in a mouse model of visceral and cutaneous disease. Interestingly, some of the *L. major* x *L. infantum* hybrid lines established better infections in a vector species, *P. duboscqi*, which is normally refractory to *L. infantum*, with the 3n hybrids showing a dosage-dependent

trend in their ability to infect this vector species based on how many of the supernumerary chromosomes came from the *L. major* parental line.

Following passage through the mouse, one of the inter-specific hybrids reverted to the *L. major* allelic variant at one of the heterozygous nuclear markers on chromosome 29, resulting in a homozygous genotype. The same nuclear marker was also stably homozygous in one other hybrid following maintenance of the line under *in vitro* and *in vivo* conditions, despite being heterozygous at all other nuclear markers, indicating the possible presence of a mechanism such as gene conversion reverting some heterozygous loci to one or the other of the parental alleles, as documented in previous crosses of *L. major*. Again, kDNA maxicircle inheritance appeared to be strictly uniparental. It is important to note that in *T. brucei*, a related kinetoplastid parasite, hybrids recovered from earlier time points showed that progeny inherited both parental copies of kDNA maxicircles, and later lost one of the two allelic variants due to unbalanced segregation in subsequent mitotic divisions (Gibson and Garside 1990).

These experiments confirm that although rare, genetic exchange does occur in *Leishmania* at low frequencies, with approximately 10^{-4} to 10^{-5} or less meiotic events per mitotic division, after correcting for recovery of only double drug resistant hybrids. Considerable work has been done on studying cellular processes underlying reproduction in the related kinetoplastid species *T. brucei*, a pathogenic organism that shares many life cycle similarities with *Leishmania*. Genomic hybrids were recovered from tse-tse fly salivary glands (Jenni, Marti et al. 1986), suggesting that hybridization happens in these stages, as later confirmed using fluorescently

tagged parasites (Gibson, Peacock et al. 2008). Many features of parasite development, such as epithelial attachment through the flagellum, are present in both *Leishmania* and *T. brucei*. A similar approach screening for double drug resistant hybrids in *T. cruzi*, responsible for American trypanosomiasis, recovered double drug resistant hybrids in mammalian host stages (Gaunt, Yeo et al. 2003), suggesting there may be important differences between trypanosome species.

As mentioned already, *T. brucei* hybrids have biparental inheritance of kDNA maxicircles and minicircles, but one of the two parental copies is subsequently lost, while the mature hybrid parasite retains minicircle kDNA from both parents. Similarly, nuclear DNA is also inherited biparentally in a Mendelian fashion, and homologous non-sister chromosomes undergo meiotic recombination. Unlike in *Giardia*, a distantly related binucleated excavate that performs genetic exchange without meiosis (Poxleitner, Carpenter et al. 2008), expression of meiosis-specific genes that are widely conserved in eukaryotes and that act at different stages during homologous recombination seems to be closely linked to a meiotic process in *T. brucei* salivary gland stages (Peacock, Ferris et al. 2011). Tracking the expression of fluorescently tagged proteins that function during prophase of meiosis I (SPO11, MND1, HOP1, and DMC1, a homologue of RAD51) in *T. brucei* epimastigotes, a distinct meiotic-competent stage was identified in which replication of the flagellum and kinetoplast precedes cellular division into daughter cells containing 2n nuclear DNA. These cells then undergo meiosis II without further DNA replication to produce haploid gametes, although the exact process – which may or may not involve nuclear fission without cell division - remains uncharacterized. Haploid

promastigote-like cells are the final product of meiosis II, making *T. brucei* an essentially meiotic organism (Peacock, Bailey et al. 2014).

There is no reason to believe that similar processes may not also underlie genetic exchanges in *Leishmania*. In addition to the crosses mentioned above, fluorescently tagged clonal lines have been used to generate crosses in *L. infantum* and in *L. donovani*, although the number of hybrid lines recovered was much smaller in these experiments (Sadlova, Yeo et al. 2011, Calvo-Alvarez, Alvarez-Velilla et al. 2014). Natural hybrids between different species of *Leishmania*, including *L. tropica*, have also been reported from many different countries (Chapter 2).

In conclusion, *Leishmania* species appear able to interbreed if presented the opportunity, both within their species and across species designations. *L. tropica* isolates often bear within their genome signatures of past occurrences of this genetic exchange (Chapter 2 and 3). Cell fusion between promastigote-like stages was documented in *L. tropica* (ISER/MA/89/LEM/1685) originating from a sand fly captured in Morocco (Lanotte and Rioux 1990) (video recording available online at: http://goo.gl/X6xs1L). It is therefore expected that *L. tropica* lines should be able to give rise to hybrid lines in a laboratory setting. This chapter describes experiments performed to generate experimental hybrids.

4.2. Methods

4.2.1. Laboratory infections of *L. longipalpis* and *P. arabicus* sand flies

Sixteen parasite lines were selected at the time of this study, fourteen of which were confirmed to be *L. tropica* by MLST (See Chapter 2). These isolates were selected from the collections deposited at NIH in the Laboratory of Parasitic Diseases, National Institute of Allergy and Infectious Diseases. Isolates whose sequences fully aligned with over 90% agreement to *L. tropica* sequences after a BLAST search in GeneBank were prioritized for sand fly feeding assays.

Batches of approximately 50 individuals of 3-5 days old female *L. longipalpis* and *P. arabicus* sand flies from laboratory colonies that were initiated from specimens collected in Jacobina, Brazil (LLJB strain) and in North Israel (PAIS strain), respectively, were used in the feeding assays. The colonies were permanently housed at the Walter Reed Army Institute of Research, in Rockville, Maryland. Sand flies were placed in feeding cups, starved overnight, and fed through a chick skin membrane with heparinized mouse blood kept at 37 °C in glass feeders, containing 4 x 10^6 parasites per mL of log-phase growth promastigotes that had been previously harvested from *in vitro* culture and washed in PBS (See Section 2.2.1 for detailed description of promastigote *in vitro* culture). Laboratory feeding of promastigotes has been previously shown to be the best model to study sexual processes, even if amastigotes are normally taken up by sand fly bites *in vivo* (Inbar, Akopyants et al. 2013).

Sand fly midguts were dissected at two time points: when the blood meal was still present within the midgut, at day 2 post-infection (n = 5 sand flies); and after the blood meal had passed, at day 8 post-infection (n = 10 sand flies). Individual midguts were dissected with sterile needles, and homogenized with a pestel in 50 µL of complete M199 medium (see Chapter 2) in a 1.5 mL eppendorf tube. The proportion of procyclic, nectomonad, haptomonad, and metacyclic forms was determined by counting the number of each form present in the midgut homogenate using a haemocytometer and a light microscope. Some *ad hoc* exploratory feedings were also performed with dissections every 2-3 days for 14 days, to further assess fluctuations of parasite development in the laboratory colonies available to us for each species of sand fly. Promastigote stages were identified based on morphological criteria previously described in the literature (Saraiva, Pimenta et al. 1995).

4.2.2. Generation of drug resistant parasite lines

Plasmid constructs bearing markers conferring resistance to nourseothricin (NTC/SAT), hygromycin B (HYG), and neomycin/G418 (NEO) were kindly provided by Dr Alain Debrabant, Division of Emerging and Transfusion Transmitted Diseases, U.S. Food and Drug Administration. Each construct integrated a cassette, containing a drug resistance marker and a fluorescence marker, into one of approximately 20 different tandem copies of the small subunit RNA domains on chromosome 27. Previous studies have shown that disruption of this redundant locus does not result

in impaired parasite growth *in vitro* or in *in vivo*. Plasmid DNA was transformed into *E. coli* and grown at 37 °C under standard procedures. Plasmid DNA was then extracted and purified using the Qiagen MiniPrep Kit following manufacturer's specifications.

Plasmid DNA was cut with SwaI restriction enzyme at 30 °C for 15 minutes, precipitated in 100% ethanol, and resuspended in 20 µL of TE buffer. Approximately 2 x 10^8 promastigotes were washed in cold PBS and resuspended in electroporation buffer (21mM Hepes, pH = 7.0, 137mM NaCl, 5mM KCl, 0.7mM Na_2HPO_4, 6mM glucose). Parasites were then electroporated on a BioRad BTX ECM 630, with settings at 450 V, resistance at 15 Ω, and capacitance at 500 µF. The transfected parasites were incubated at 26 °C overnight prior to drug selection with the chosen antibiotic (25 µg/mL hygromycin, 100 µg/mL nourseothricin, 50 µg/mL G418). After 1-2 weeks of drug selection, cultures with drug resistant parasites were diluted and plated on cM199 agar plates. Single parasite colonies were picked with a pipette tip, and individual clones were then grown briefly in culture for genotyping. Clones positive for integration were stored in freezing solution under liquid nitrogen conditions. Successful genomic integration of the resistance cassette was verified by PCR using previously designed primers that span the 5' and 3' SwaI restriction sites (Figure 4.1), and by expression of the expected fluorescent protein in drug-resistant transgenic parasites as verified by fluorescent microscopy and flow cytometry.

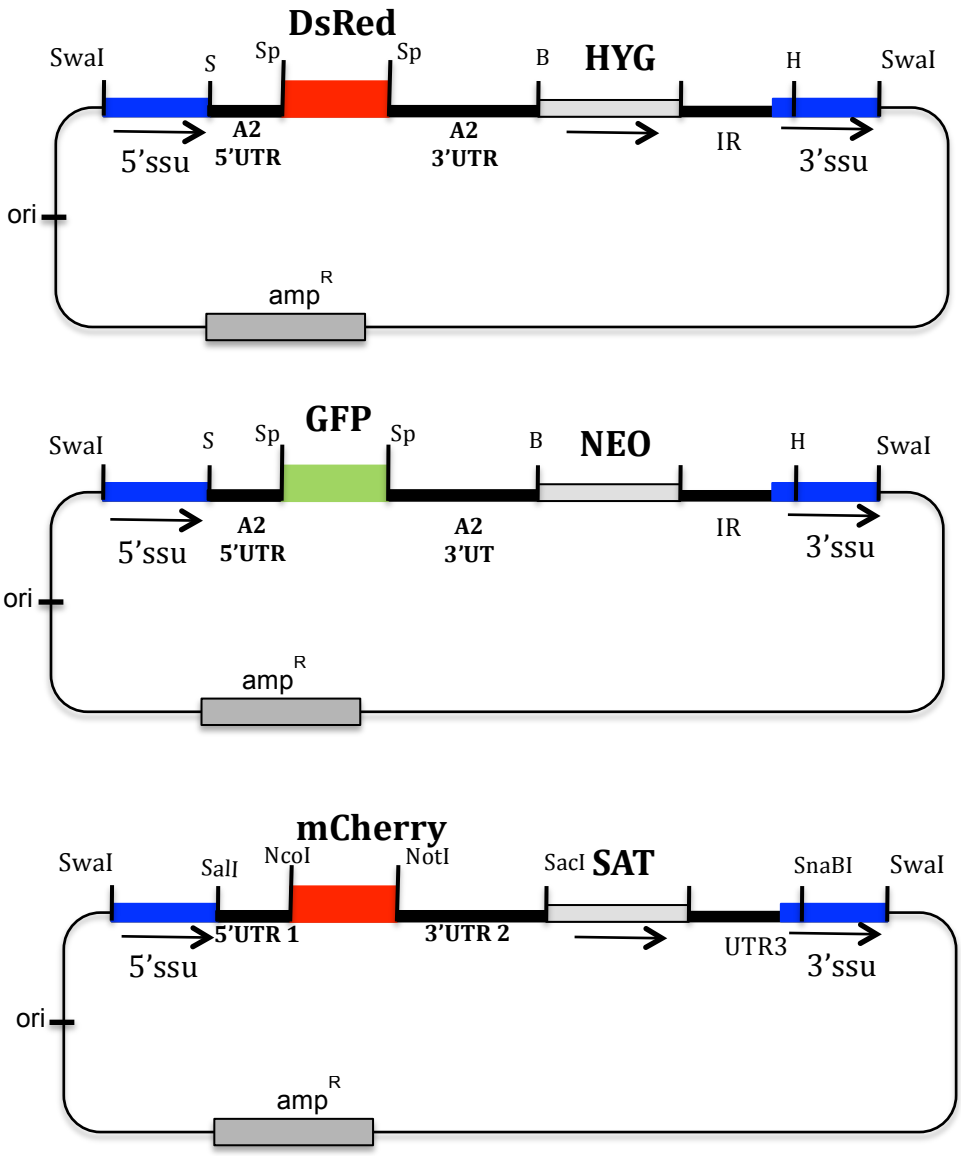


Figure 4.1. Targeting vectors used in this study for integration of different drug resistance markers in the *L. tropica* genome. From top to bottom, the plasmids are pA2-RFP-HYG, pLEXSY-cherry-SAT2, and pA2-GFP-NEO. In blue are the homologous arms for integration into the small subunit RNA locus. The HYG marker codes for resistance to hygromycin B (Hyg B), SAT codes for

resistance to nourseothricin (NTC), NEO codes for resistance to neomycin and geneticin (G418).

4.2.3. Screening for double drug resistant hybrids

A number of different crosses were attempted using drug resistant lines that were generated as described in Section 4.2.2. In each attempted cross, we seeded heparinized mouse blood with a mixture of 4 x 10^6 parasites of two different lines of drug-resistant *L. tropica*, each line being resistant to a different antibiotic. Cross-resistance to both antibiotics by a single drug resistance marker was excluded by *in vitro* susceptibility assays for all drug resistance markers used (data not shown). *L. longipalpis* (LLJB strain) and *P. arabicus* (PAIS strain) sand flies were fed on this mixture of parasites as described in Section 4.2.1. Blood-fed sand flies were separated from unfed females the day following the infectious feed. At different time points after the infectious feed, ranging from 7 to 13 days post-infection, a variable number of blood-fed sand flies were anesthetized with CO_2, and their midgut were dissected, homogenized in 100 µL of complete M199 medium, and placed in a single well of a 96-well flat bottom plate. Plates were incubated at 26 °C overnight, and the appropriate combination of antibiotics for selection of double-drug resistant parasites was added the following day (100ug/mL NTC, 50 µg/mL G418, 25 µg/mL Hyg B). Double-drug resistant parasites were cloned by limiting dilution, or by plating the diluted culture on cM199 agar plates, as described in Section 4.2.2. The

presence of both drug resistance markers in the putative hybrid lines was confirmed by PCR. A summary of the screening procedures is shown in Figure 4.2.

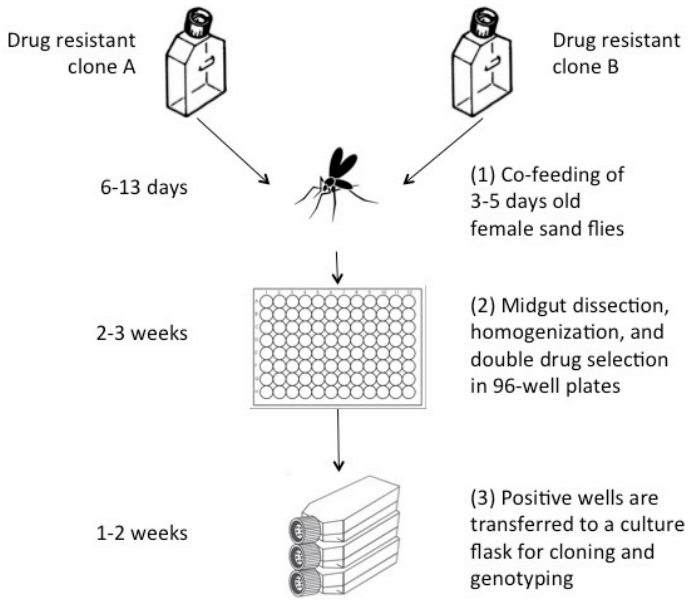


Figure 4.2. Schematic representation of the screening procedures for recovery of double-drug resistant hybrids in sand fly co-infections. Clones A and B have been engineered to be resistant to two different antibiotics. The two lines are mixed in heparinized mouse blood and co-fed to female sand flies (1). At different points during metacyclogenesis, individual sand fly midguts are dissected and placed in a 96 well plate following homogenization, and both antibiotics are added to the culture medium (2). Following 2-3 weeks or longer of double drug selection, any positive wells are transferred to a larger

culture flask for cloning and genotyping by PCR to verify the presence of both drug resistance markers (3).

4.3. Results

4.3.1. Laboratory infections of *L. longipalpis* and *P. arabicus* sand flies

All 14 isolates confirmed to be *L. tropica* by MLST were able to infect the non-natural vector *L. longipalpis* LLJB strain, a well-established permissive laboratory model to study parasite-vector interactions (Figure 4.3). Five parasite isolates (Melloy, Ackerman, 188, E50, Azad) had a lower parasite load present in the midgut on day 8 post-infection compared to the parasite load on day 2, when the blood meal was still present. All other parasite isolates had a higher parasite load on day 8 post-infection, suggesting successful attachment to the inner surface of the midgut during bloodmeal discharge and subsequent parasite replication. Examination of the stomadeal valve at day 13 post-infection showed promastigote secretory gel plug formation, suggesting that *L. tropica* can complete its life cycle within this insect vector and may be infectious to mammalian hosts.

Four isolates (Rupert, Kubba, E50, L747) were also tested in the natural vector *P. arabicus* PAIS strain. All isolates successfully developed within the sand fly midgut, showing an increase in parasite load and progression from procyclic to nectomonad, haptomonad, and metacyclic forms on day 8 (Figure 4.4). Comparison with *L. longipalpis* infections shows similar parasite loads and comparable

proportions of each developmental stage, except for one isolate, E50, which seems to infect *L. longipalpis* LLJB sand flies with lower effectiveness than *P. arabicus* PAIS. The isolate L747 showed appreciable infection at day 8 post-infection, but due to the limited size of the laboratory colony and the small proportion of sand flies that completed blood feeding in our feeding assays no sand flies could be dissected at day 2 (data not shown).

Two exploratory feedings were also performed in *P. sergenti* PSSS strain, another natural vector of *L. tropica* throughout most of its range in the Middle East. Due to the instability of the laboratory sand fly colony and the low infection levels in the small number of sand flies that successfully blood fed in laboratory conditions (data not shown), this vector species was abandoned in subsequent experiments.

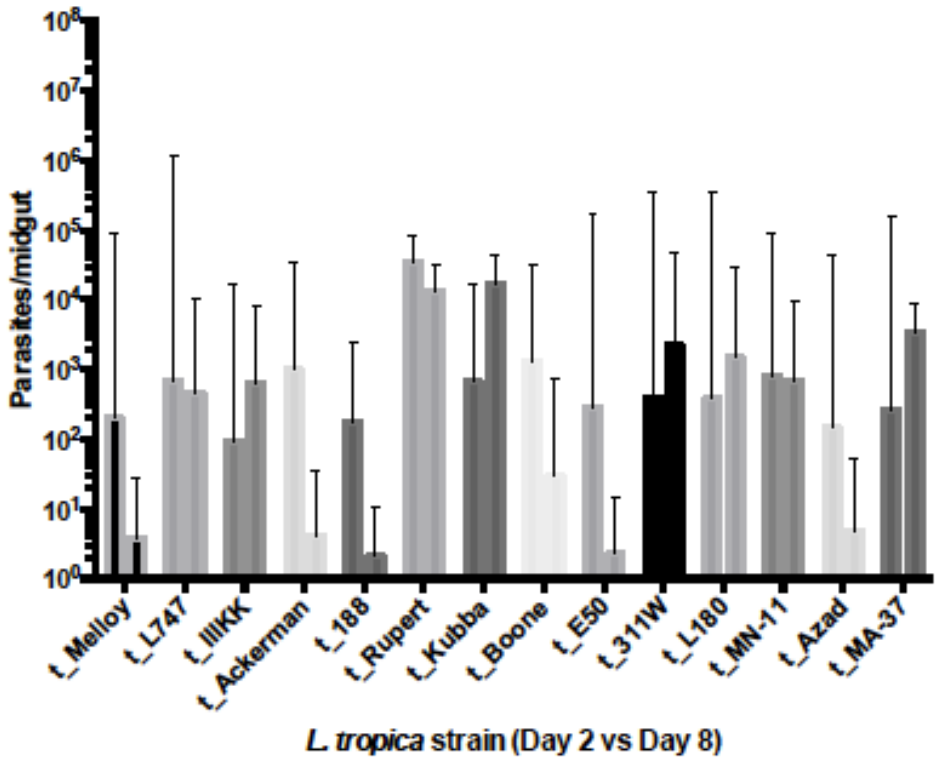


Figure 4.3. Infection loads of 14 *L. tropica* isolates in *L. longipalpis* LLJB sand flies, at day 2 and day 8 post-infection. Each pair of bars represents the geometric mean number of parasites (all developmental stages) in 5 infected midguts at day 2 (left bar) and in 10 infected midguts at day 8 (right bar). The 95% confidence interval around the geometric mean is shown with error bars.

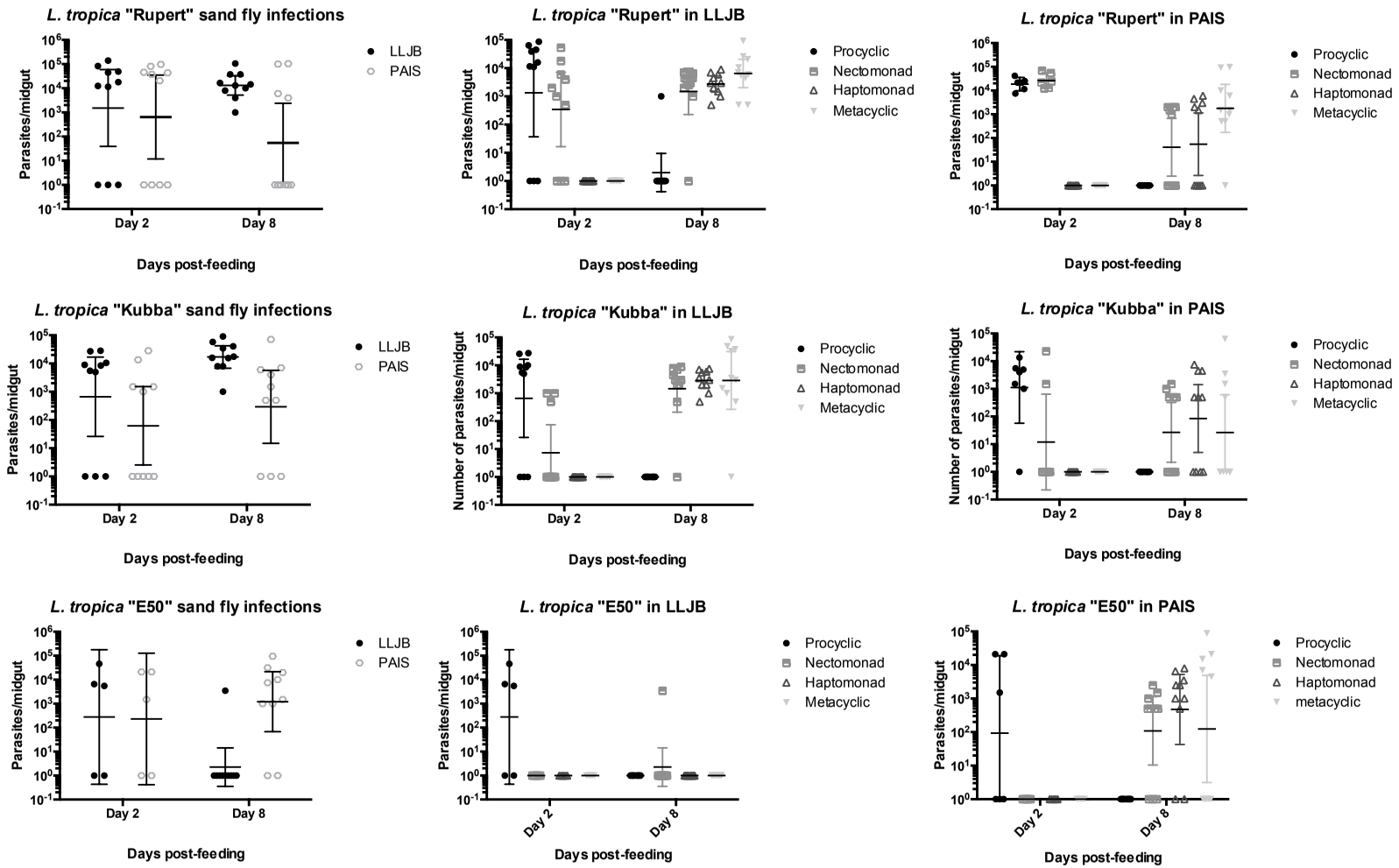

L. tropica "Rupert" sand fly infections
Parasites/midgut
Day 2
Day 8
Days post-feeding
LLJB
PAIS
L. tropica "Rupert" in LLJB
Procyclic
Nectomonad
Haptomonad
Metacyclic
L. tropica "Rupert" in PAIS
L. tropica "Kubba" sand fly infections
L. tropica "Kubba" in LLJB
Number of parasites/midgut
L. tropica "Kubba" in PAIS
L. tropica "E50" sand fly infections
L. tropica "E50" in LLJB
L. tropica "E50" in PAIS
metacyclic

Figure 4.4. Infection loads and developmental stages of *L. tropica* isolates Rupert, Kubba, and E50 in *L. longipalpis* LLJB and *P. arabicus* PAIS sand flies. The geometric mean and 5% confidence interval of each experiment is shown. The number of infected midguts dissected at day 2 and day 8 were 5 and 10, respectively.

4.3.2. Generation of drug resistant parasite lines

Five strains were selected based on country of origin, robust growth phenotypes throughout *in vitro* culture and *in vivo* sand fly feedings, and variation in genotype (Chapter 2). One of three possible constructs containing a fluorescent marker and a drug resistant marker was integrated into the genome of isolates Kubba, Rupert, MN-11, MA-37, and L747, in one of approximately 20 small subunit RNA domains on chromosome 27 (Table 4.1). Constructs stably integrated in the genome as determined by PCR (data not shown). Fitness of the transgenic line was confirmed by passage through the sand fly (Figure 4.5). Drug resistance and fluorescence was confirmed during repeated *in vitro* culture and stable inheritance of the integrated construct was confirmed by flow cytometry (Figures 4.6, Kubba SAT and MN-11 NEO shown as examples).

Line	Origin	Fluorescent marker	Drug resistant marker	Plasmid
Kubba SAT	Syria	mCherry	Nourseothricin	pLEXSY-cherry-SAT2
Rupert HYG	Afghanistan	dsRed	Hygromycin B	pA2-RFP-HYG
Rupert NEO	Afghanistan	GFP	Neomicin/G418	pA2-GFP-NEO
MN-11 HYG	Jordan	dsRed	Hygromycin B	pA2-RFP-HYG
MN-11 NEO	Jordan	GFP	Neomicin/G418	pA2-GFP-NEO
MA-37 NEO	Jordan	GFP	Neomicin/G418	pA2-GFP-NEO
L747 HYG	Israel	dsRed	Hygromycin B	pA2-RFP-HYG

Table 4.1. Transgenic drug resistant lines generated for crossing experiments with drug resistance markers NEO, HYG, and SAT.

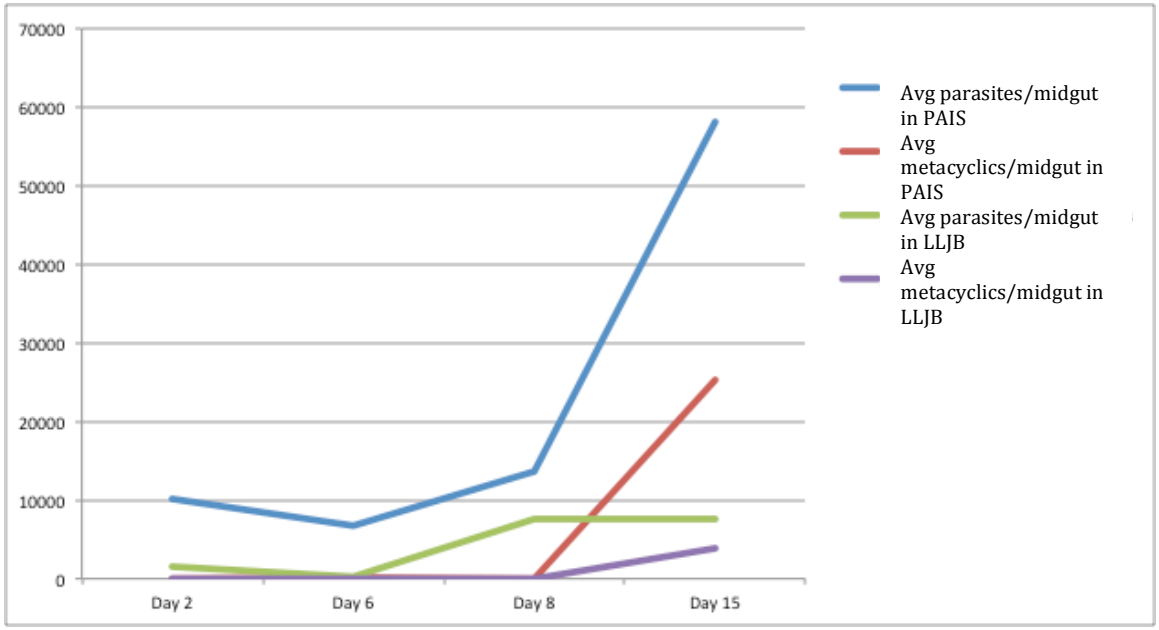


Figure 4.5. Growth phenotype of the transgenic line Kubba SAT in *P. arabicus* PAIS and *L. longipalpis* LLJB sand flies. For both PAIS and LLJB, the first line represents the average number of parasites per midgut at days 2, 6, 8, and 15 post-infection. The second line represents the average number of infectious

metacyclic promastigotes per midgut, and is therefore a fraction of the total number of parasites per midgut. Notice the increase in metacyclic forms at days 8 and 15 post-infection, meaning successful establishment of infection and progression to infectious mature stages.

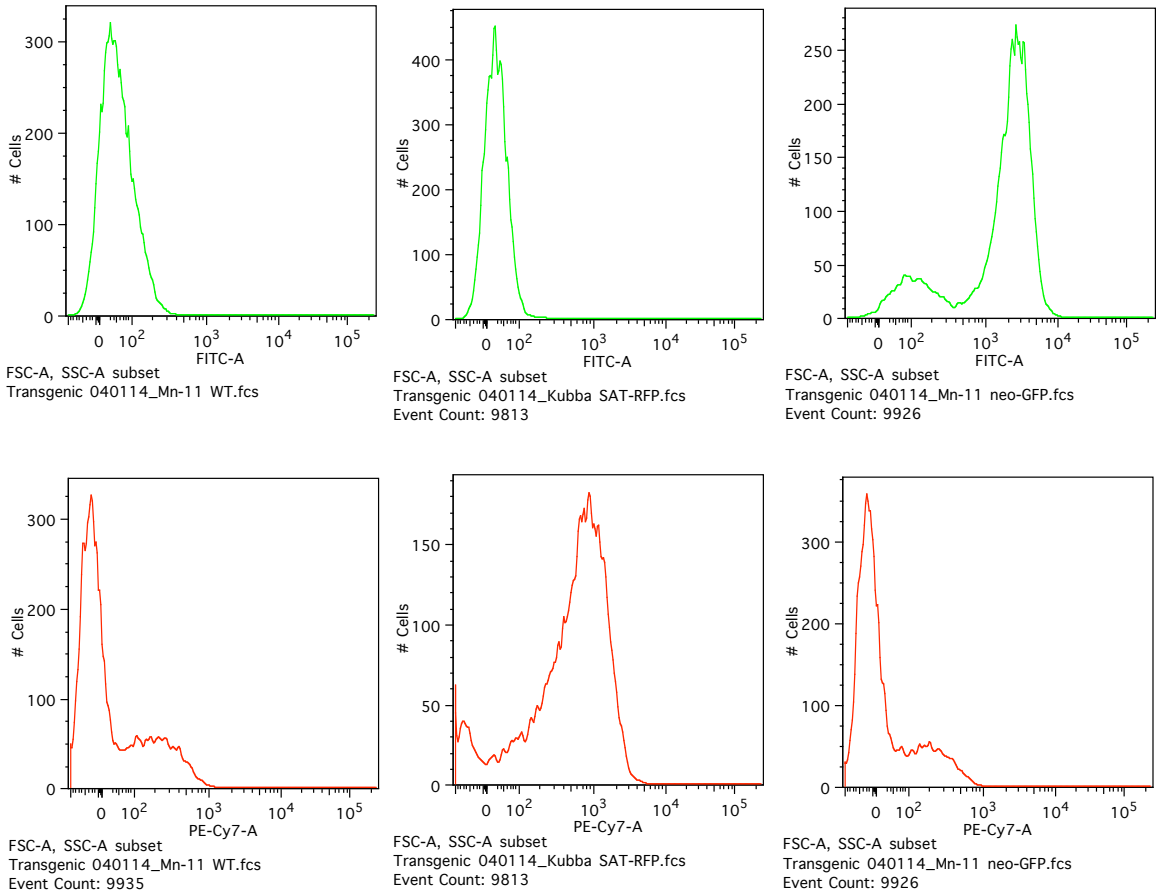


Figure 4.6. Confirmation of expression of the fluorescent markers in the transgenic drug-resistant lines Kubba SAT and MN-11 NEO. From left to right, each panel shows fluorescence in channels FITC-A (in green, top row) and PE-Cy7-A (in red, bottom row) for lines MN-11 wild type, Kubba SAT (expressing

mCherry fluorescent protein), and MN-11 NEO (expressing GFP). Kubba wild type had a similar fluorescence profile as MN-11 wild type (negative in both FITC-A and PE-Cy7-A, not shown).

4.3.3. Screening for double drug resistant hybrids

A total of 7 crosses were attempted using different combinations of drug resistant lines in both *P. arabicus* PAIS and *L. longipalpis* LLJB sand flies (Table 4.2). Recurrent health problems with the laboratory colony of *P. arabicus* PAIS limited the number of females available for feeding assays. A total of 931 sand flies were dissected, and 178 wells (~19%) were lost to fungal or bacterial contamination despite our attempts at maintaining sterile conditions. Sand flies cannot be kept under aseptic laboratory conditions and harbour many bacteria within their gut. A double-drug resistant recovery rate of 2.26 % was seen in the MN-11 HYG x Kubba SAT cross (2.23 % in LLJB, 2.13 % in PAIS). Parasites from 6 wells were positive for double drug resistance (resistant to both NTC and Hyg B) in this cross, however they failed to grow after transfer to a new culture flask with fresh antibiotics, and were therefore lost. A double-drug resistant recovery rate of 31.5 % was seen in the MA-37 NEO x L747 HYG cross in LLJB sand flies, with 3 double-drug resistant parasites recovered from 15 clean wells at day 7 post-infection, 27 recovered from 82 clean wells at day 8, and 15 out of 46 clean wells at day 11. A total of 5 positive wells showed strong double drug resistance even following passage of the parasites into a new culture flask with the same concentration of fresh antibiotics (50 µg/mL

G418, 25 μg/mL Hyg B). These represent 5 independent possible mating events (Table 4.2). Two clonal lines were obtained from each of these double drug resistant parasite cultures. Each pair of clonal lines can be considered "sibling" hybrid lines, which may or may not have originated from the same mating event. The drug resistance profile of each clone was ascertained (Table 4.3) and the Inheritance of each drug resistance marker, NEO and HYG, in each of the hybrid lines was confirmed by PCR (Figure 4.7). The uncloned, double drug resistant parasite population from each positive well was also preserved as a frozen stock for subsequent analyses.

Cross	Sand fly	No. dissected	No. contaminated	No. double-drug resistant	No. hybrids recovered
MN-11 NEO x Kubba SAT	PAIS	46	26	0	0
MN-11 NEO x Kubba SAT	LLJB	138	2	0	0
MN-11 HYG x Kubba SAT	PAIS	48	7	1	0
MN-11 HYG x Kubba SAT	LLJB	282	58	5	0
Rupert HYG x Kubba SAT	PAIS	50	18	0	0
Rupert HYG x Kubba SAT	LLJB	217	42	0	0
MA-37 NEO x L747 HYG	LLJB	150	7	45	5

Table 4.2. Number of sand fly dissections, wells lost to bacterial or fungal contamination, and positive wells with double-drug resistant parasites for each of the attempted crosses. Variation in the number of sand flies dissected is due to fluctuations in colony size and in the number of sand flies that successfully fed in each experiment.

	Clone	Positive well	NEO	HYG
1	H1a	H1	+	+
2	H1b	H1	+	+
3	H2a	H2	+	+
4	H2b	H2	+	+
5	H3a	H3	+	+
6	H3b	H3	+	+
7	H4a	H4	+	+
8	H4b	H4	+	+
9	H5a	H5	+	+
10	H5b	H5	+	+
11	L747 HYG	NA	-	+
12	MA-37 NEO	NA	+	-

Table 4.3. Summary of the 10 hybrid lines originated from the MA-37 NEO x L747 HYG cross, and PCR positivity for presence of each drug resistance marker. Hybrid lines that were cloned from the same positive well may or may have arisen from independent mating events.

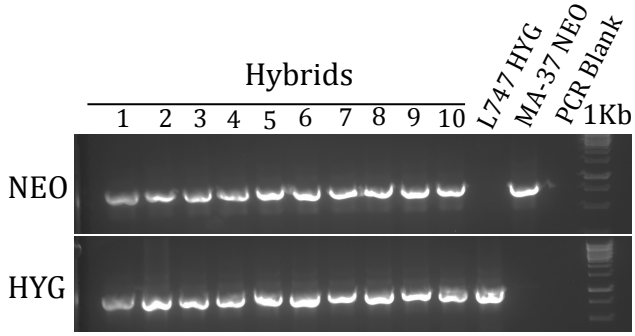


Figure 4.7. Confirmation by PCR of inheritance of the drug resistance markers NEO and HYG in the 10 putative hybrid lines generated in the MA-37 NEO x L747 HYG cross. Hybrid lines are numbered as in Table 4.3.

4.4. Discussion

The experiments summarized in this chapter demonstrate that *L. tropica*, like other *Leishmania* species, is able to perform genetic exchange in the sand fly stages. A total of nearly 1000 sand flies were dissected, and only about 50 wells showed strong growth in the presence of both antibiotics. Low double-drug resistant recovery rates were seen in the MN-11 HYG x Kubba SAT crosses (~2%), but since these subsequently failed to grow *in vitro* under drug selection, they cannot be considered real hybrids. Possible rationales for this phenomenon include loss of either drug resistance marker, poor initial exposure to drug in culture due to high parasite density, or generation of hybrids with reduced fitness. An additional area that needs further investigation is the possibility that some hybrids may be arising after midgut dissection in extremely low numbers, although extensive previous experimentation has effectively ruled out this possibility as exceedingly rare (Akopyants, Kimblin et al. 2009).

A relatively high double-drug resistant recovery rate was seen in the L747 HYG x MA-37 NEO crosses (~30%), with a true hybrid recovery rate of 3.5%. No double drug resistance was observed in any of the other cross combinations. These differences may indicate that there are mating incompatibilities between strains of different genetic backgrounds. As described in Chapters 2 and 3, some isolates appear to be homozygous at the majority of typed marker (Chapter 2) and have long stretches of homozygosity throughout their genomes (Chapter 3), while others harbour considerable heterozygosity. An intriguing interpretation of these findings

is that as heterozygosity arises through outcrossing events between homozygous strains, the mating potential of these outbred hybrids is reduced, both in crosses with either homozygous parent, or with other heterozygous hybrids. In yeast interspecific crosses, F1 heterozygous hybrids produce only 1% viable gametes, while parental lines produce 90-100% viable gametes (Greig, Louis et al. 2002, Xu and He 2011). Different mechanisms could be mediating this phenomenon, from deleterious epistatic interactions to dominant genic incompatibility. A similar process may reduce the mating potential of heterozygous hybrids in *L. tropica*: isolates Rupert and Kubba appear to be heterozygous throughout the majority of their genomes both by MLST and WGS, and crosses involving either as one of the parental lines yielded no or very few positive wells (6 out of a total of 781 dissected midguts). Post-zygotic barriers to hybrid fertility might reduce interbreeding between the hybrid and parental populations and favour speciation in *Leishmania* as in yeast (Greig 2007, Schumer, Cui et al. 2015). Interestingly, in *T. brucei*, F1 hybrid progeny show a range of mating compatibilities that do not readily fit a simple two mating type system (Peacock, Ferris et al. 2014).

The vector species used for the crosses may also play a role in hybridization. In previous crosses in *L. major*, the same *L. longipalpis* LLJB strain used in this study gave much higher hybrid recovery rates than the natural vector, and at earlier time points of parasite development within the sand fly. Reports of natural interspecific hybrids are quite common in South America, where the *Lutzomyia* genus is endemic. Parasite development within *L. longipalpis* has been shown to occur faster

compared to Old World vector species, with nectomonads being the dominant life cycle stage by day 3 post-infection (Walters, Irons et al. 1993).

The 5 positive wells found in this study were recovered at days 7, 8, and 11 post-infection, confirming an association between the nectomonad and later stages and the presence of hybrids. The infection assays I performed in *L. longipalpis* show a large number of mature forms at day 8, with a large number of elongated nectomonad forms and fully mature metacyclics being present. The relatively smaller parasite load seen in *P. arabicus* may reflect intrinsic physiological differences in the amount of blood consumed per blood meal by the sand fly, rather than parasite fitness within the midgut environment. It is important to note that the minimum infectious dose necessary to ensure transmission of the parasite to the host is very small, and previous crosses have recovered hybrids from midguts with very low parasite infections, making a heavy parasite load in the midgut a condition which may not necessarily improve parasite survival (Kimblin, Peters et al. 2008, Stamper, Patrick et al. 2011).

As reviewed in Chapter 2, the heterozygosity observed in natural populations of *L. tropica* is larger than that seen in other species. This and additional lines of evidence suggest that *L. tropica* is capable of Mendelian genetic exchange. I have here reported the first successful cross between two isolates of *L. tropica*, L747 from Israel and MA-37 from Jordan, with different genetic backgrounds as elucidated by MLST and WGS, and demonstrated biparental inheritance of the two drug resistance markers used for selection of double drug resistant hybrids. In the next chapter, I introduce the results of WGS data analyses from these 10 hybrid clones to elucidate

genome-wide patterns of inheritance and genetic exchange that arise through hybridization.

CHAPTER 5

GENETIC EXCHANGE IN EXPERIMENTAL HYBRIDS

5.1. Introduction

As described in Chapter 4, several crosses of *Leishmania* species have now been performed. Specifically, combinations of *L. major* isolates with different geographical origins and an interspecific cross between *L. major* and *L. infantum* have successfully given rise to experimental hybrids. To this list, we must now add *L. tropica* as the most recent species for which a sizable number of hybrids have been recovered in laboratory crosses.

Crosses between individuals that differ in a phenotypic trait of interest have been used in the mapping of genes involved in determining that trait in many different organisms. This type of analysis needs a sizable number of hybrid progeny and relies on recombination as the mechanism by which genes are shuffled in the progeny, and is best done on biparental hybrid populations such as near-isogenic lines, resulting from a backcross between F1 hybrids and one of the parental lines; recombinant inbred lines, resulting from selfing of F1 hybrids over multiple generations; or advanced intercross lines, resulting from multiple rounds of intercrossing, or sib pair mating, between F1 hybrids (Jamann, Balint-Kurti et al. 2015).

This type of analysis, known as quantitative trait locus (QTL) mapping, has been used in plants and other organisms since the 1980's. Although high-throughput genotyping has increased the number of molecular markers that can be captured at once by at least three orders of magnitude, this by itself does not increase the resolution of the genetic mapping, since hybrid populations that have undergone multiple meiotic events are required in order to increase the chances of capturing recombination between closely linked markers.

The process of genetic exchange in *Leishmania* F1 hybrids analysed so far seems to follow classic Mendelian inheritance rules, with each hybrid clone being heterozygous at genotyped markers where each of the parental lines is homozygous. The cellular processes that underlie this genetic exchange are poorly characterized, although the nectomonad stage in the sand fly appears to be the developmental stage most closely associated with genetic exchange, and cell fusion between promastigote-like cells has been observed (see Chapter 4). Most hybrids recovered so far are either diploid or triploid, suggesting a reduction of ploidy at some stage prior to or following cell fusion. The presence of a meiosis-derived haploid or near-haploid gamete stage is possible. Cell fusion between near-haploid and near-diploid gametes may explain the recovery of hybrids with ploidy greater than 2, such as near-triploid hybrids.

The ability of *Leishmania* parasites of tolerating extensive aneuploidy is poorly understood (and may be connected to their unique transcriptional properties, see Chapter 3), but seems to be constitutive to the genus and may further complicate patterns of homozygosity or heterozygosity seen in the hybrids. The loss of

heterozygosity seen in a minority of genotyped markers is compatible with gene conversion, or with reversion to the haploid condition at these loci. Currently, it is unclear whether aneuploidy arises during mitotic division, or whether it is linked to cell fusion between meiotic products with unbalanced ploidy. Consistently, however, *in vitro* cultured isolates have been found to have a heterogeneous cell population, each parasite potentially having a different total number of chromosomes (Sterkers, Crobu et al. 2014)(Chapter 3).

Recombination events have been identified in several natural populations of *Leishmania*. Recently, a vector-isolated hybrid population was identified in Turkey, with phylogenetically distinct *L. donovani* complex strains as putative parents (Rogers 2014). In addition, metrics measuring linkage disequilibrium (LD) have been used to identify populations in which recombination has led to a reshuffling of the genotyped markers on each chromosome, with few markers being in LD with each other in the genotyped population. However, no analysis specifically measuring recombination in experimental hybrids has so far been performed.

The presence of widespread recombination would strongly suggest the presence of a distinct meiotic stage. In most eukaryotes, although mitotic recombination may be present, meiotic recombination is the main driver behind genome-wide recombination of parental markers. These are typically clustered around so-called "recombination hotspots", and occur to a first approximation at a fixed rate (Paigen 2015).

Haplotypes of linked markers and recombination between markers that are linked in the parents can be identified in F1 progeny and subsequent generations

through a process called phasing. Phasing involves determining the gametic phase of each set of markers – in diploid organisms, this means attributing the origin of each heterozygous or homozygous genotype seen in the progeny to either parent, so that haplotypes, i.e. a series of linked markers which are passed on from parent to offspring, can be identified. For homozygous genotypes, the solution is trivial, since each parent will have contributed the same allele to the offspring. For heterozygous genotypes, it becomes necessary to consider the genotype of the parents. Heterozygous genotypes in the offspring for which at least one parent is homozygous can be non-ambiguously phased. Heterozygous genotypes in the offspring for which both parental genotypes are heterozygous, however, have no non-ambiguous phasing solution, and these must therefore be called non-informative markers (see Table 5.1).

	AA	**AT**	**TT**
GG	AG AG	AG TG	TG TG
GC	AG AC	AG TC or AC TG	TG TC
CC	AC AC	AC TC	TC TC

Table 5.1. An example illustrating the phasing problem for two linked biallelic, disomic loci. Cells represent the possible phasing solutions in the genotyped progeny. Note that when both parental genotypes are heterozygous there is no non-ambiguous phasing solution (AG on one chromosome and TC on the other chromosome, or AC on one chromosome and TG on the other, are both acceptable phasing solutions for two linked heterozygous loci AT and GC).

Once phase has been determined at all informative markers, in order to identify recombination events pairs of linked markers that are heterozygous in one parent need to have changed positions in the offspring – i.e., for biallelic linked markers, an AG TG individual, where each pair of letters represents a marker, and the order represents the chromosome each allele is found on, generates a recombinant haplotype A-G instead of A-T. For this purpose, orthogonal approaches are necessary. One approach involves physical phasing, which in genotyping by WGS consists of identifying which variants are represented within the same read. Given the length of the read, in Illumina sequencing rarely exceeding 100 base pairs, only variants that are close to each other on the chromosome and for which adequate sequencing coverage has been achieved can be phased in this way. The result is short haplotype blocks, which are broken up by intervening homozygous stretches along the length of the chromosome, although read pair information, by making use of the longer insert size, can help resolve phase across some of these homozygous blocks. However, for longer stretches, no solution can be found by using this method only.

An alternative solution is population-based phasing, that involves identifying related individuals within a population that carry stretches of linked heterozygous markers in homozygous form. These can be thought of as ancestral haplotypes that have given rise to heterozygous haplotypes by hybridization. The main limitation of this approach is the number of individuals that need to be genotyped, although this

number is greatly reduced in highly inbred populations where there are long stretches of chromosomal regions that are identical-by-descent (IBD).

When information about the parents is available, phasing allows one to identify regions where there are Mendelian violations. These could be due to point mutations reverting a variant allele in one of the chromosomes to the other parental allele, to loss of one of the two parental haplotypes, or to gene conversion.

Although F1 hybrids from previous crosses of *Leishmania* have been sequenced, at present published evidence suggesting genome-wide biparental inheritance of genomic material is limited, and mostly relies on a limited number of markers typed by PCR (Inbar 2014). No information on *de novo* mutation rates associated with hybridization has been published. In this chapter, we present our findings confirming that biparental inheritance of genetic material is the rule in *Leishmania* and that balanced segregation of the chromosomes following a meiotic process is very likely, with aneuploidy arising following subsequent mitotic divisions.

5.2. Methods

5.2.1. Culturing of parasites, DNA extraction, and whole genome sequencing

Please refer to Section 2.2.1 for a detailed description of *Leishmania* culture procedures. Parasite promastigote cultures were pelleted and DNA was extracted using the QIAgen DNeasy Blood and Tissue extraction kit following manufacturer's

procedures. The DNA was quantified using a Nanodrop spectrophotometer prior to library preparation for sequencing on the Illumina platform. The samples were sequenced on two lanes of the Illumina HiSeq 2500 platform, with an average insert size of 500 and a read length of 100 bp.

5.2.2. Mapping, variant calling, and quality control of called variants

A new reference genome assembly was created by the Parasite Genomics finishing team at the Wellcome Trust Sanger Institute. This assembly was generated by *de novo* assembly of PacBio sequence data using the HGAP assembly pipeline, and then using contigs from the assembly used in previous chapters. Contigs from this assembly were scaffolded based on optical mapping of the *L. tropica* L590 reference strain. Gaps in this assembly were filled using a combination of Illumina reads (with IMAGE and GapFiller software) and PacBio reads using PBJelly. This was followed by manual examination of both Illumina reads, and PacBio filtered, corrected sub-reads from HGAP in Gap5 (Bonfield and Whitwham 2010). This last step adds further corrections to the data. REAPR was used to evaluate the quality of the resulting assembly, and its completeness was confirmed by CEGMA and by comparison to other *Leishmania* finished reference assemblies. Such an approach is independent of the *L. major* reference genome, and ensures that scaffolds are not incorrectly combined based on homology to *L. major*, but rather based on an experimentally determined optical map. Gaps in this assembly were filled using a

combination of PacBio reads and illumina reads in Gap4 (Bonfield 1995). Reads for each sample were mapped using smalt to this reference assembly (Ltro_freeze_v.2).

Raw Fastq reads were mapped to the Ltro_freeze_v.2 reference using SMALT. The resulting mapped reads were then processed with the Genome Analysis Tool Kit (GATK v.3.4-0, Broad Institute) to call SNPs and small insertion-deletions (indels). The HaplotypeCaller algorithm was used with genotyping mode set to discovery, which makes use of haplotype information and is generally more accurate than the UnifiedGenotyper algorithm that is part of the same program. Variants were then quality filtered to exclude any genotyping errors: briefly, this involved screening for variants that fell into clusters of 3 or more per 10 bp window, variants with a mapping quality less than 50, variants with a Fisher strand bias greater than 20, variants with reads spanning deletions, and variants with a base quality, mapping quality, or read position rank sum that fell outside empirically determined limits to their normal distribution (a base quality rank sum greater than 3.1 or smaller than -3.1, a mapping quality rank sum greater than 4 or smaller than -4, and a read position rank sum greater than 3.1 or less than -3.1).

Variants on disomic chromosomes (see Section 5.2.3) were then further processed in PLINK (Purcell, Neale et al. 2007) for input into phasing programs and R statistical analysis software. Only biallelic variants on disomic chromosomes that passed quality filters were considered for subsequent analyses.

5.2.3. Estimation of chromosome number in parental and hybrid lines

The mapped reads for each sample were processed with GATK DepthOfCoverage to obtain total read depth, reference, and alternate allele frequencies at each base pair of the reference genome assembly. Multi-allelic positions were included in this analysis. These raw estimates were then processed with custom Perl, Unix, and R scripts to generate plots of allele frequencies for each chromosome. A custom-built expectation-maximization (EM) algorithm (see Chapter 3) was built to model the haploid read depth on each chromosome with respect to the rest, and an estimated somy number was obtained for each chromosome in each sample. The estimated somy obtained with the EM algorithm was manually validated by inspection of allele frequency plots.

5.2.4. Phasing of high quality variants and recombination

Biallelic variants on disomic chromosomes were phased using SHAPEIT (O'Connell, Gurdasani et al. 2014) with the –duohmm option for complex pedigrees. All variants were validated for missingness and consistency within the pedigree formed by two parents and 10 siblings in PLINK. These were then input into SHAPEIT for phasing and identification of Mendelian errors. First, variants were phased ignoring pedigree relationships using the –noped option. Then the duoHMM algorithm, which uses a Hidden Markov Model to identify haplotypes in related trios and duos, was run over 10 possible phasing solutions generated from the diploid graph obtained with SHAPEIT to find probable recombination crossover points between the parents and the progeny.

5.2.5. Identifying *de novo* SNPs, indels, and structural variants

SNPs and indels that passed quality filters were then further analysed with vcftools (Danecek, Auton et al. 2011) to identify variants that were private to the hybrid lines, i.e. present in the hybrid progeny lines but absent in the parental lines. Structural variants were detected by using DELLY (Rausch, Zichner et al. 2012). DELLY utilizes read pair information to identify regions where paired mates align improperly, or in a manner that is incompatible with the expected insert size. These can be indicative of duplications, inversions, translocations, or large deletions. A similar variant calling procedure was performed for sequencing reads obtained from the *L. tropica* reference strain L590 that were used to generate the reference genome assembly. Structural rearrangements that are present in both the reference strain and the hybrids are likely due to sequencing reads being mapped to a misassembled region in the reference assembly, and should therefore be excluded as structural artifacts. Called variants were further processed with vcftools to identify variants that were private to the hybrid progeny.

5.3. Results

5.3.1. Whole-genome sequencing, mapping and variant calling

Whole-genome sequencing gave yield to a minimum of 35.42x and a maximum of 67.12x coverage per sample (see Table 5.2). After mapping to the reference assembly, a total of 1533533 raw variant calls were made using GATK HaplotypeCaller. These were quality filtered as described in the Methods section to reduce the list to 499769 high quality SNPs and indels for the 12 samples considered in this study.

Sample	Lanes	Insert sizes	Cycles	Depth	Bases (pre-QC)	Bases (post-QC)
L747_HYG	2	500	100	55.87x	1.10 Gb / 1.14 Gb	100.11 Mb / 103.23 Mb
MA-37_NEO	2	500	100	35.42x	1.18 Gb / 1.21 Gb	107.28 Mb / 101.10 Mb
H1a	2	500	100	50.83x	1.41Gb / 1.46 Gb	100.80 Mb / 103.95 Mb
H1b	2	500	100	43.25x	1.09 Gb / 1.12 Gb	108.79 Mb / 101.75 Mb
H2a	2	500	100	49.39x	1.19 Gb / 1.16 Gb	108.46 Mb / 105.66 Mb
H2b	2	500	100	40.96x	1.03 Gb / 1.05 Gb	102.69 Mb / 105.27 Mb
H3a	2	500	100	36.05x	872.68 Mb / 899.10 Mb	109.08 Mb / 112.39 Mb
H3b	2	500	100	52.67x	1.21 Gb / 1.24 Gb	100.68 Mb / 103.48 Mb
H4a	2	500	100	51.73x	1.32 Gb / 1.35 Gb	101.24 Mb / 103.89 Mb
H4b	2	500	100	39.78x	991.87 Mb / 1.02 Gb	110.21 Mb / 102.05 Mb
H5a	2	500	100	51.50x	1.23 Gb / 1.27 Gb	102.54 Mb / 105.64 Mb
H5b	2	500	100	67.12x	1.56 Gb / 1.61 Gb	104.21 Mb / 100.58 Mb

Table 5.2. Number of lanes, insert size, read length in number of cycles, depth, and total number of bases obtained from sequencing runs of the 10 hybrid lines and the 2 parental lines used in this study. Pre- and post-QC base yield is provided by lane. See Appendix A for ENA accession numbers.

5.3.2. Estimation of chromosome number in parental and hybrid lines

All hybrid offspring and parental lines appeared to be near-diploid (see Figure 5.1). Chromosome 31 was tetrasomic in all samples, as seen in many *Leishmania* species studied to date. Chromosome 23 was trisomic in the two parental lines and in all hybrid lines, with the exception of hybrids H3a and H3b. These two hybrid lines were disomic at chromosome 23, suggesting the presence of a step reducing chromosome number from the level seen in the parents. Chromosome 4 was also trisomic in hybrids H4a and H4b, while it was disomic in all other parental and hybrid lines.

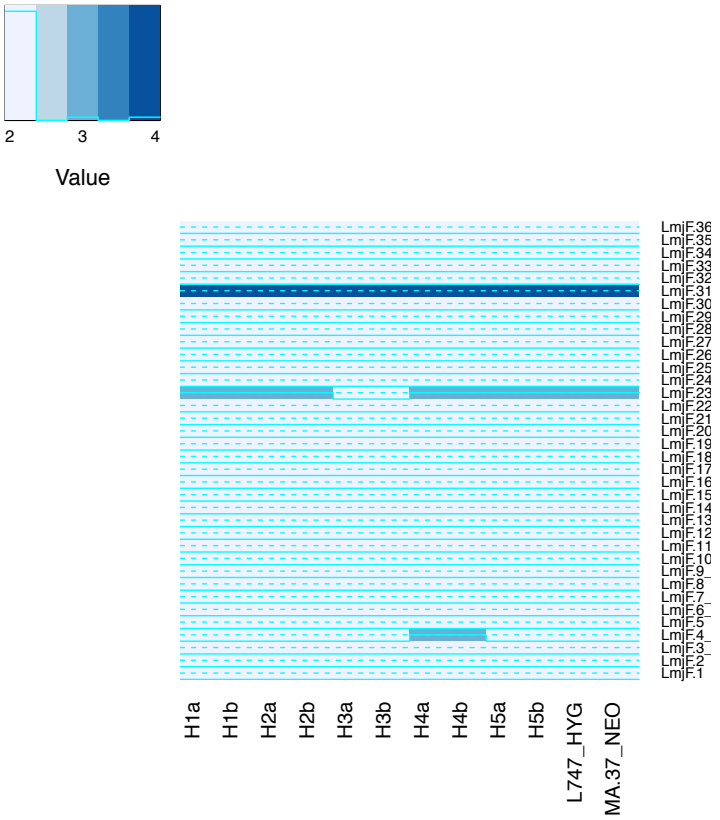
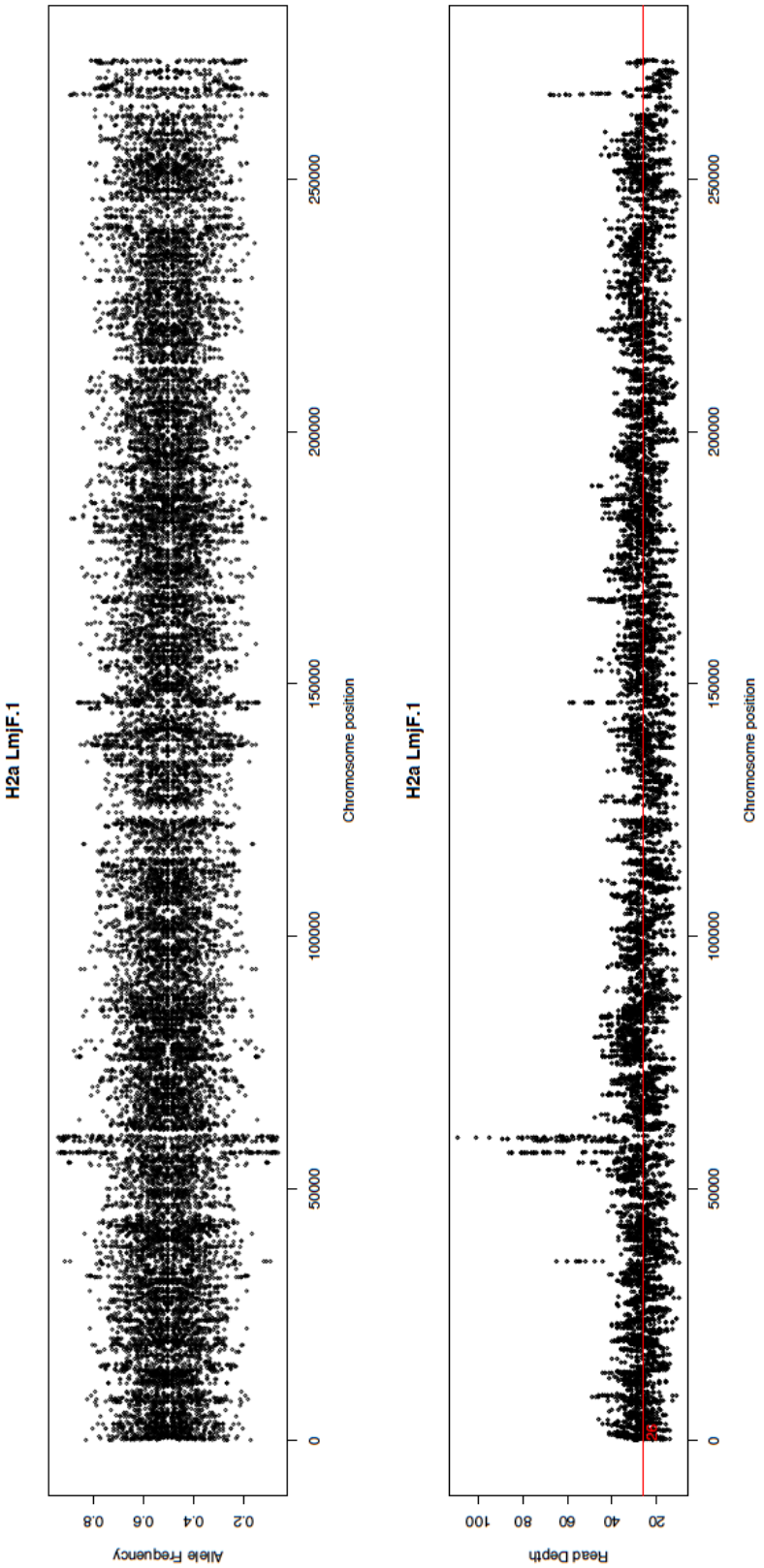


Figure 5.1. Somy estimates from WGS data for all hybrid and parental lines represented as a heatmap. Hybrids are indicated by the letter H. Darker shades of blue are associated with larger somy number, from a minimum somy of 2 to a maximum somy of 4. Note that chromosome 23 (LmjF.23) is trisomic in the parental lines (L747_HYG and MA-37_NEO) and in the majority of hybrid lines, while chromosome 31 (LmjF.31) is tetrasomic in all lines.

Interestingly, inspection of allele frequencies per chromosome revealed that in all individual samples chromosome 31 was heterozygous for each of the parental alleles in a 1:2, or 0.5, ratio, meaning that two of the chromosomes had one allelic variant and two of the chromosomes had the other variant. While this is expected for heterozygous disomic chromosomes, and was indeed observed in our sample, on tetrasomic chromosomes allele frequencies can vary from the 1:2 ratio to 1:4 and 3:4 ratios (0.25 and 0.75), if only one chromosome has a different allele from the other three. This however was not observed in our study, suggesting that each of the parents contributed two copies of chromosome 31. All trisomic chromosomes showed 1:3 and 2:3 ratios (approximately 0.33 and 0.67) in their allele frequencies, due to the fact that for heterozygous positions these are the only allele frequencies possible, while disomic chromosomes all appeared to be heterozygous with variant allele frequencies of 0.5 (Figure 5.2, 5.3 and 5.4). Inspection of read depth across the chromosome confirms that read coverage is even across the chromosome, with median read depth increasing in a dose-dependent manner with somy (Figure 5.2 and 5.3).


H2a LmjF.1
Allele Frequency
0.2
0.4
0.6
0.8
0
50000
100000
150000
200000
250000
Chromosome position
H2a LmjF.1
Read Depth
20
40
60
80
100
0
50000
100000
150000
200000
250000
Chromosome position

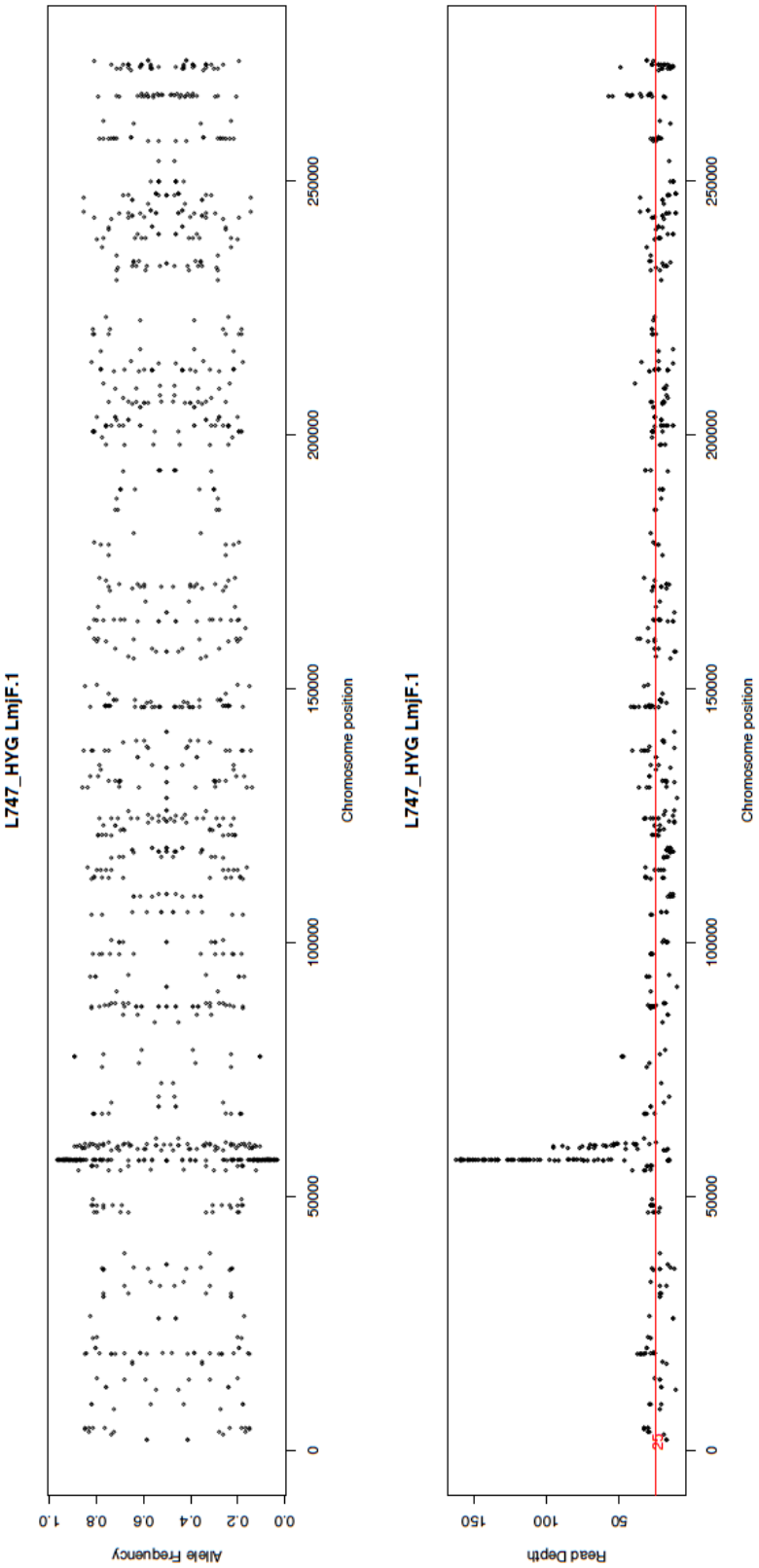
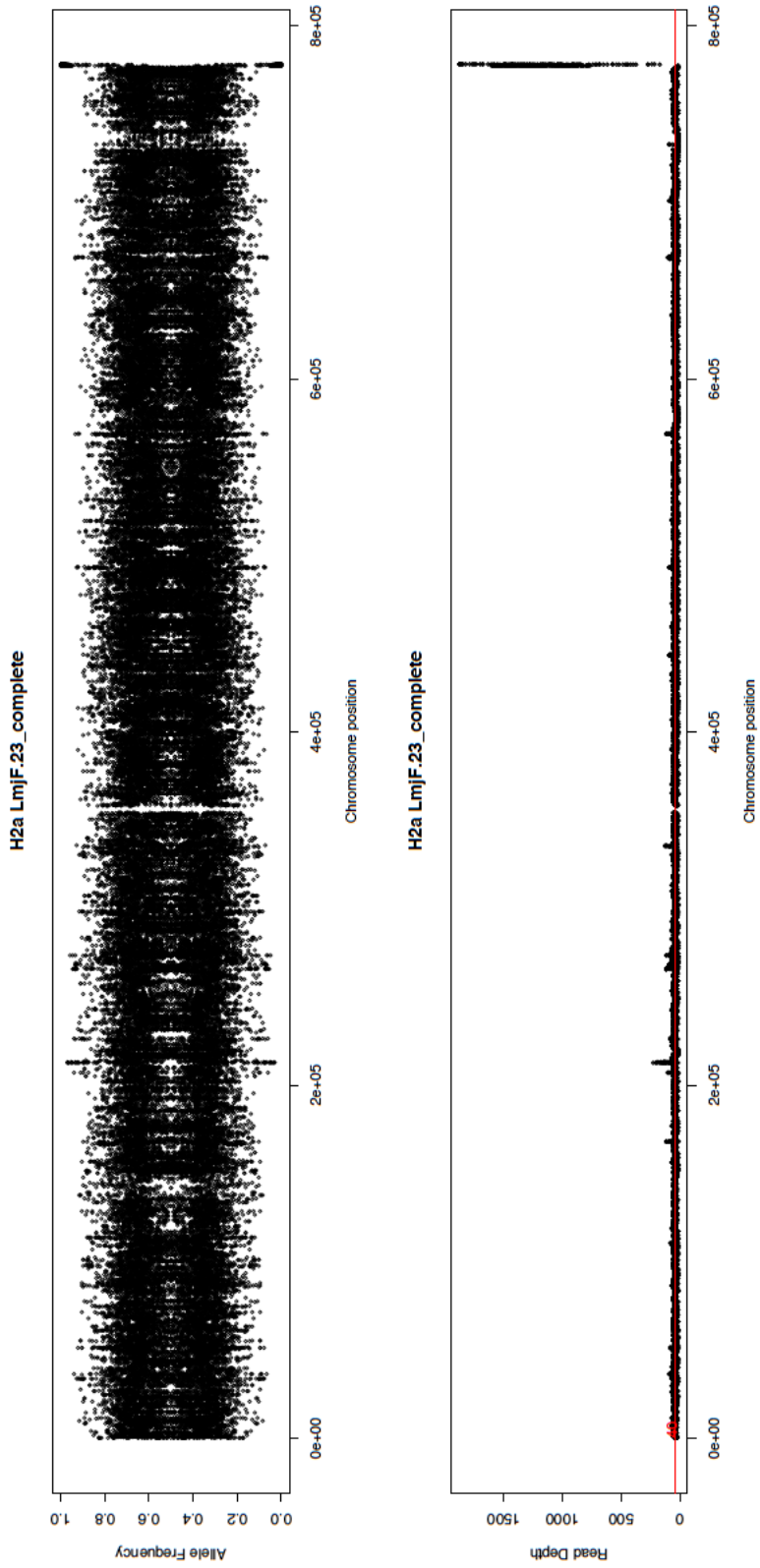

L747_HYG LmjF.1
Allele Frequency
0.0
0.2
0.4
0.6
0.8
1.0
0
50000
100000
150000
200000
250000
Chromosome position
L747_HYG LmjF.1
Read Depth
50
100
150
25
0
50000
100000
150000
200000
250000
Chromosome position

Figure 5.2. Heterozygous allele frequencies for chromosome 1 (LmjF.1) in one of the hybrids (H2a) and in one of the two parental lines (L747 HYG), depicted in the upper panel, with read depth across the chromosome depicted in the lower panel. Median read depth is marked by a red line, and is 26 and 20, respectively. Both of these chromosomes were identified as disomic by the EM algorithm. Note the paucity of data points in the parental line, due to most of the variants called being homozygous and therefore either 1 or 0 (very few data points are seen in the 0.1 to 0.9 y-axis range, whereas there is an overabundance of data points around 0.5 in the hybrid line, indicating heterozygous variants present in about half of the reads). These plots are illustrative of all disomic chromosomes in the hybrids and the parents, respectively.


H2a LmjF.23_complete
Allele Frequency
0.0
0.2
0.4
0.6
0.8
1.0
0e+00
2e+05
4e+05
6e+05
8e+05
Chromosome position
H2a LmjF.23_complete
Read Depth
0
500
1000
1500
0e+00
2e+05
4e+05
6e+05
8e+05
Chromosome position

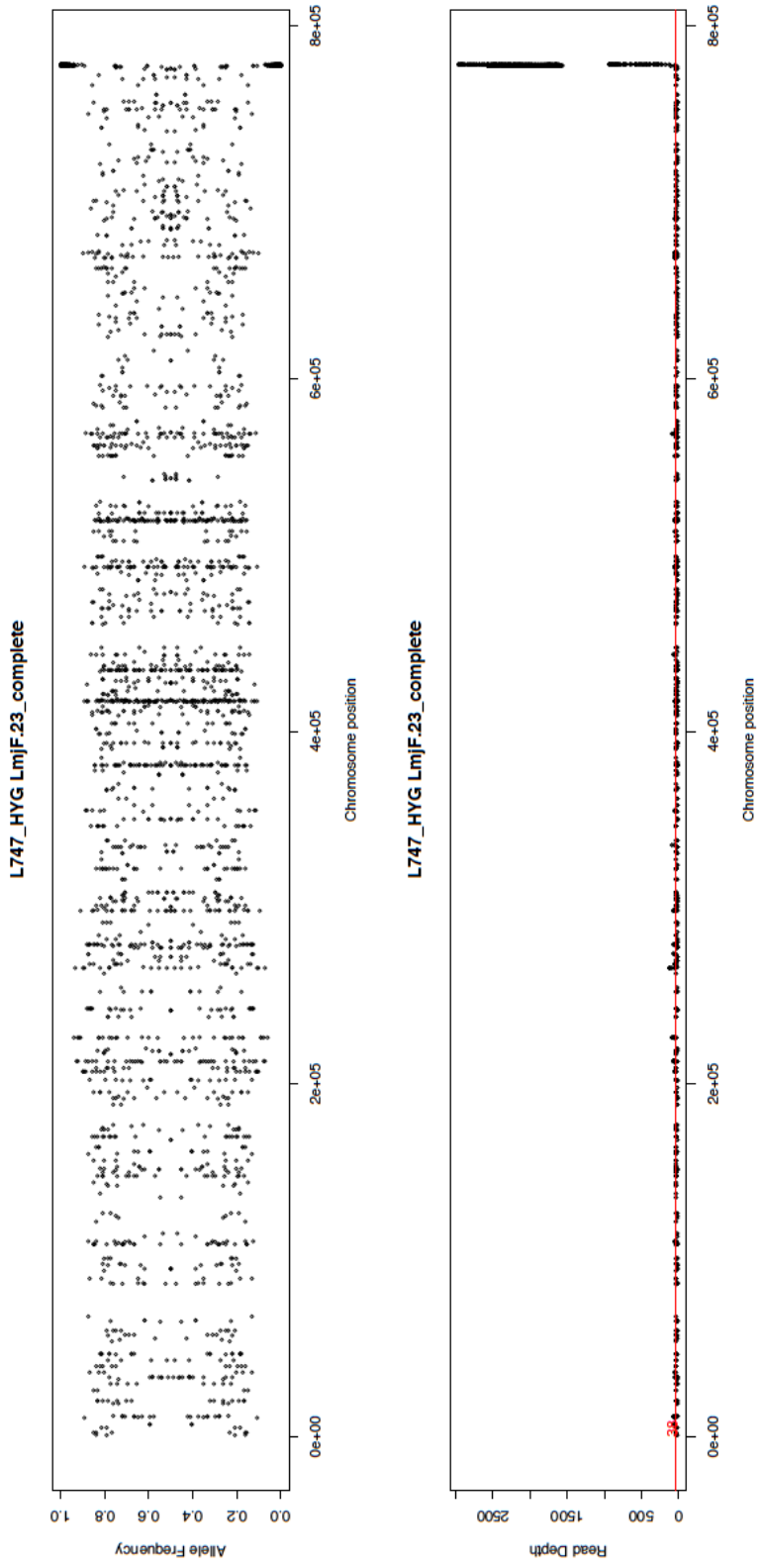

L747_HYG LmjF.23_complete
Allele Frequency
0.0
0.2
0.4
0.6
0.8
1.0
0e+00
2e+05
4e+05
6e+05
8e+05
Chromosome position
L747_HYG LmjF.23_complete
Read Depth
0
500
1500
2500
30
0e+00
2e+05
4e+05
6e+05
8e+05
Chromosome position

Figure 5.3. Heterozygous allele frequencies and read depth for chromosome 23 (LmjF.23) in the same two lines as in Figure 5.2 (hybrid H2a and parental line L747 HYG). Note how the density of the data points increases as chromosome size increases (chromosomes are inversely numbered with size in *Leishmania*, differently from humans and other mammals). Data points are concentrated around 0.33 and 0.67 in H2a compared to the parental line, which have sparse variant calls with no clear distribution in the allele frequency spectrum (homozygous frequencies of 1 or 0 were excluded)., confirming trisomy of chromosome 23 as estimated by the EM algorithm. Median read depth is 40 and 38, respectively. The isolated high read depth at the end of the chromosome visible in both lines is likely due to a misassembly of the reference.

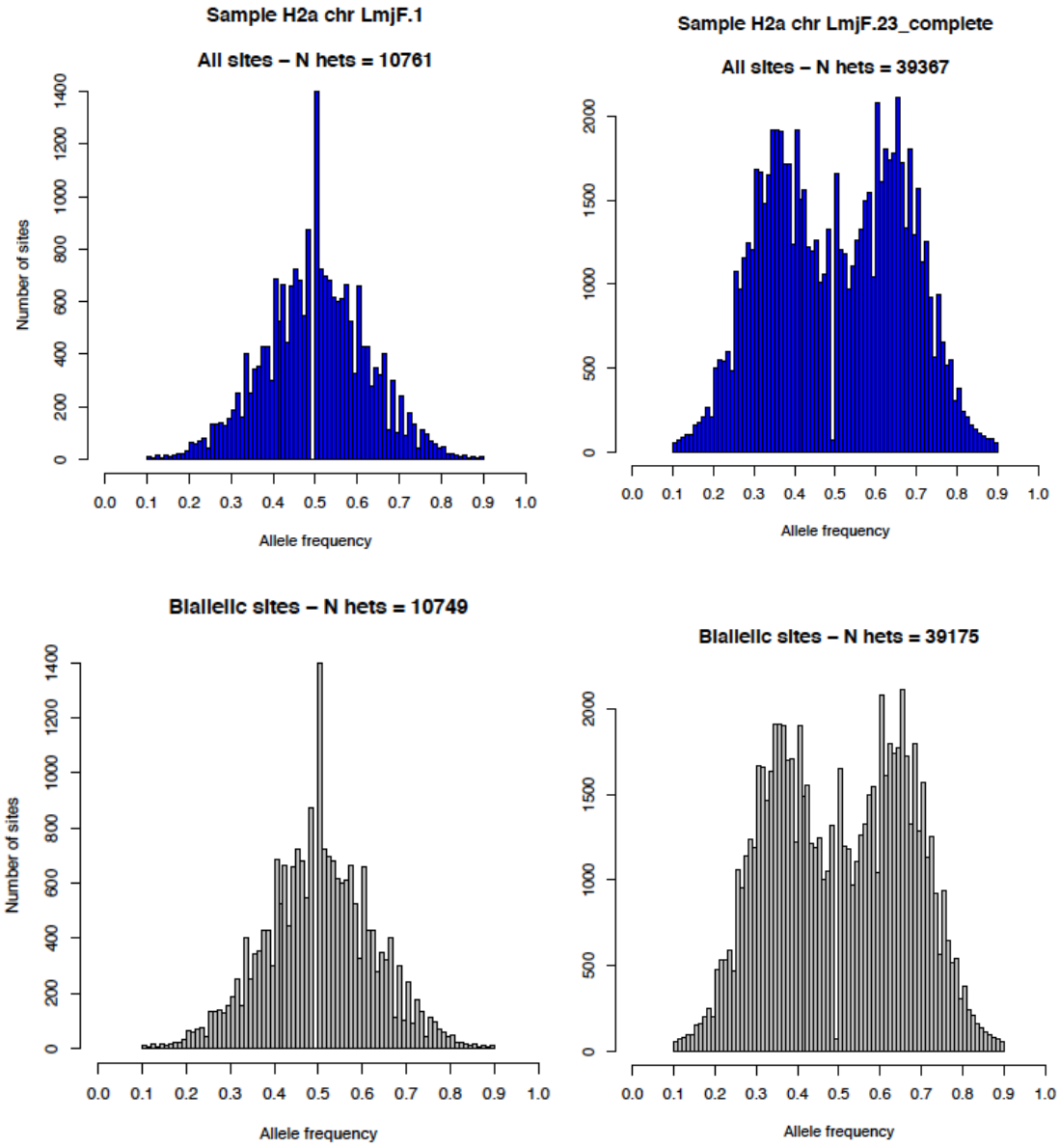


Figure 5.4. Allele frequency histograms for all heterozygous sites and for heterozygous biallelic sites only on chromosomes 1 (LmjF.1) and 23 (LmjF.23) in hybrid H2a, with the total number of sites indicated as "N hets". These distributions are illustrative of disomic and trisomic chromosomes, respectively, in all hybrids. Note the peaks at 0.33 and 0.67 on chromosome

23, indicating 1/3 and 2/3 possible ratios for heterozygous trisomic sites, and at 0.5 on chromosome 1, indicating the only possible ratio of ½ for heterozygous disomic sites.

5.3.3. Phasing of high quality variants and recombination

Biallelic variants on disomic chromosomes were phased using SHAPEIT software. A total of 2035 loci (0.4% of all biallelic sites) had evidence of Mendelian violations (Table 5.3), and were therefore excluded from phasing attempts. Due to the fact that the new reference genome assembly for *L. tropica* has not been annotated as of yet, the location of these Mendelian violations with respect to coding regions could not be investigated further. These were distributed genome-wide, with no clear continuous stretches of variants violating Mendelian expectations.

All biallelic variants on disomic chromosomes with no Mendelian errors were phased with SHAPEIT and duoHMM. All hybrid lines were heterozygous at positions where each of the parents was homozygous for a different allele. Positions in which one of the two parents was heterozygous were either homozygous for one allele, homozygous for the other allele, or heterozygous for both as expected by Mendelian inheritance rules. A graphic representation of the inheritance of alleles from parents to offspring is shown in Figure 5.5.

No recombination was detected by the SHAPEIT ouput. Due to the complex pedigree of the experimental set up with 10 sibling hybrid lines, no significant recombination crossovers were detected after averaging haplotype sets over 10

simulations, likely due to the small number of offspring analysed and to the high homozygosity of the two parental lines involved in the cross. SHAPEIT is optimized for human genetics, and is used to analyze hundreds of individuals with different degrees of relatedness. More sensitive *ad hoc* approaches will have to be developed to detect recombination crossovers breaking haplotype patterns in this and similar datasets generated from experimental crosses.

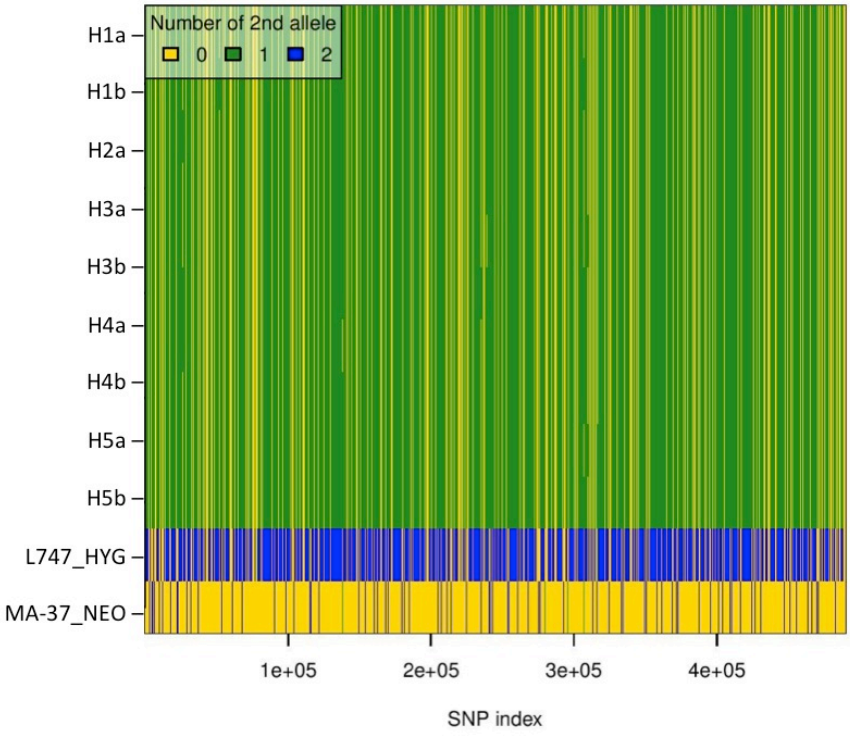


Figure 5.5. Allelic plot showing biparental inheritance of alleles from the two parental lines (bottom two rows, individual indices 11 and 12) to the offspring (indices 1 through 10). A blue cell represent a position that is homozygous for the first allele, a yellow cell is homozygous for the second allele, and a green cell is a heterozygous position of a blue plus a yellow allele.

Individual	No. of Mendelian violations
L747_HYG	413
MA-37_NEO	1780
H1a	885
H1b	333
H2a	183
H2b	173
H3a	155
H3b	81
H4a	53
H4b	88
H5a	54
H5b	30

Table 5.3. Number of Mendelian violations distributed by individual. A total of 2035 SNPs had Mendelian violations from a total of 489822 SNP calls (0.4 %). A Mendelian violation needs to be shared between at least one parent and one offspring, by definition.

5.3.4. Identifying de novo SNPs, indels, and structural variants

A total of 6543 structural variants were identified using DELLY after applying the built-in quality filters. The majority of these were deletions (4529), followed by translocations (1511), inversions (257) and tandem duplications (246). Due to the properties of read pair mapping and how these are modelled by the program, DELLY is unable to identify extrachromosomal duplications, so these remain unidentified. The largest group of structural variants that appear to have occurred *de novo* in the progeny are translocations, with 17.09% of called variants being present in the progeny but absent from the two parental lines, followed by inversions, duplications, deletions, and lastly by SNPs and small indels.

	Total called passing QCs	Total non-reference	Total private to hybrids	Percent *de novo*
SNPs and small indels	499769	NA	342	0.068 %
Deletions	4529	4474	61	1.36 %
Duplications	246	194	9	4.64 %
Translocations	1511	1363	233	17.09%
Inversions	257	212	13	6.13 %

Table 5.4. Number of *de novo* variants private to the hybrid offspring, shown as a fraction of the total number of variants identified. Non-reference variants are variants that are present in the study samples but not in the reference strain used to generate the assembly, and are thus true structural variants and not misassemblies. SNPs and small indels are called following mapping to the reference genome and are therefore non-reference by definition. Note that these are structural variants that are smaller than the large CNVs discussed in Chapter 3 and are therefore not comparable with that analysis, which employed different analytic methods.

5.4. Discussion

Leishmania experimental crosses are starting to provide a glimpse into the genetic processes that shape genetic variation in this important pathogen. Despite decades of research and debate on the topic, the presence of an obligatory sexual stage in the life cycle of *Leishmania* remains contested. The ability of *Leishmania* species to hybridize both within and between species has now however been rigorously demonstrated (Romano 2014, Inbar 2014, Akopyants 2009). The possibility of performing crosses between strains that differ in phenotypic traits

linked to pathogenesis promises to be a powerful tool for the dissection of the genetic basis of these traits with forward genetic approaches.

In Chapter 4, we have demonstrated that *L. tropica* isolates are capable of hybridizing through a process compatible with sexual reproduction. In this chapter, we have demonstrated that the resulting hybrids are full genomic hybrids, heterozygous throughout most of the genome for markers inherited from each parent. We identify ploidy changes that are consistent with the presence of a haploid stage, but however fail to identify evidence for meiotic recombination. This failure can be attributed to the extensive homozygosity of the parental lines that could be masking recombination crossovers between the two homologous chromosomes, which are for the most part identical. Given that recombination occurs through chiasmata between homologous non-sister chromatids during meiosis, if the two chromosomal arms are identical such as expected in the homozygous condition, the resulting recombination would exchange identical stretches of DNA and therefore go undetected. Any heterozygous positions in the parental genomes are separated by long runs of homozygosity that make finding a non-ambigouous phasing solution across these haplotypes difficult without pedigree-based phasing approaches. More sensitive statistical approaches will have to be developed to identify associations between single heterozygous SNPs suggestive of recombination, despite the long stretches of homozygosity seen in these F1 hybrids.

All hybrid lines were near-diploid, providing convincing evidence for evolutionary constraints on the cellular processes associated with hybridization that make balanced segregation of the chromosomes the rule rather than the

exception. The fact that chromosome 23, which was trisomic in both parents, is also trisomic in the majority of hybrid lines, but heterozygous rather than homozygous suggests that a reductional step similar to meiosis has occurred (Figures 5.2 and 5.3; heterozygous allele frequencies plots similar to these were generated for all chromosomes in all hybrid and parental lines but are not included in this work for brevity; they are available upon request from iantornostefano@yahoo.com). In the two hybrid lines where chromosome 23 is disomic instead of trisomic, the same pattern of heterozygosity was observed, suggesting that the parental lines provided only one copy each of chromosome 23. Chromosome 31 was tetraploid in all samples examined, but again this chromosome showed allele frequency patterns consistent with heterozygosity in the hybrids, while only a limited number of heterozygous positions were found in the parental lines, suggesting that the parents contributed two copies each of this chromosome. In summary, trisomic chromosomes in the parents were passed on to the progeny only in single or double copy, never in their original trisomic state, while tetrasomic chromosomes were always passed on in double copy and not in quadruple copy. It is important to note that chromosome 4 showed evidence of unbalanced segregation during meiosis, with hybrids H4a and H4b being trisomic, although this could also have arisen during subsequent *in vitro* culture through mitotic divisions from an initial disomic state.

For disomic chromosomes, a phasing solution could be found, showing overwhelming evidence for the parental origin of each of the chromosomes in homologous pairs. A minority of sites (0.4%) were found to be in violation of

Mendelian inheritance rules. These could arise through genotyping errors, which are unlikely given the strict quality control measures adopted for variant calls in this study, or to genetic processes such as gene conversion or overlapping point mutations switching genotypes to a new variant allele. The lack of annotation of the most recent reference genome means that the genomic position of *L. tropica* genes is not known with any confidence, preventing further characterization of these Mendelian violations with respect to coding regions. A brief survey of their genomic location suggests that they are distributed across the genome and that they are not clustered in continuous stretches.

While the majority of structural variants were not private to the hybrids and thus were not acquired *de novo*, a large proportion (17.06 %) of translocations were present only in the hybrid lines. This would confirm a remarkable plasticity in the genome of *Leishmania*, which in addition to aneuploidy could also tolerate a non-negligible amount of translocation activity between non-homologous chromosomes, possibly in well-defined regions such as subtelomeric regions. This possibility will require further investigation and validation once an annotated reference genome is made available. Only 0.068% of the point mutations detected (including both SNPs and small indels) were private to the hybrids, situating the point mutation rate at around 2×10^{-6} per meiotic cycle if we assume a total genome size of approximately 35Mb (36561031bp in the latest reference genome used in this study), which is three orders of magnitude higher than the rate of germline mutations seen in humans at approximately 1×10^{-8} for single nucleotide variants (Campbell and Eichler 2013). The general principle found to be valid across most eukaryotic

organisms is that mutation rates scales directly with genome size (Lynch 2010). Rogers and colleagues (2014) have estimated mutation rates per generation in *Leishmania* to be one order of magnitude less than in humans, conforming to this general rule. This could be due to the filters selected in this study being too lax, to the fact that hybrid parasites were grown briefly in culture, or to a combination of these and other factors. In order to experimentally determine the mutation rate per generation in *L. tropica*, mutation accumulation experiments will have to be designed.

In conclusion, we have found strong evidence for Mendelian segregation of the parental genetic material in the experimental hybrids, both in terms of parental alleles and in terms of chromosome numbers. No hybrids with ploidy larger than near-2n were seen, suggesting that cell fusion between cells containing more than n chromosomes is relatively rare, if it does occur as seen in other *Leishmania* species. These results provide convincing evidence that classic forward genetic approaches for mapping of traits of interest are feasible in *L. tropica*, and that further attempts at crossing F1 hybrids with either parental line could generate backcross F2 near-isogenic lines that would prove extremely informative for mapping recombination events in this species.

CHAPTER 6

CONCLUSIONS

The aim of this thesis was to investigate genome plasticity and the effects of hybridization on *L. tropica* genome structure and function. I have demonstrated that like *L. infantum* and *L. major*, *L. tropica* is also capable of performing genetic exchange. In the Introduction, I have delineated three possible models of reproduction that this genetic exchange might follow: asexual, sexual, and parasexual. I hereby discuss each in light of my findings, and explore to what extent my objective of describing genetic exchange in this species has been achieved.

6.1. Population genetics in *L. tropica*

6.1.1. Heterozygosity and reproduction

As discussed in Chapter 2, a characteristic feature of the *L. tropica* species complex is the great genetic heterogeneity encountered. As I have presented in that chapter, both MLST and WGS data show that isolates covering the entire geographic distribution of the species often bear genome-wide differences in patterns of homozygosity and heterozygosity that can be attributed to differences in the frequency of intercrossing between individual "clones" or "clonal lineages". Previous studies that have analysed in detail microsatellite data from a much larger sample

set than that considered in this thesis (Schwenkenbecher, Wirth et al. 2006, Krayter, Bumb et al. 2014) have found three broad clusters that broadly speaking overlap geographically with those identified in the analyses performed in this thesis: one cluster for Israel/Palestine, one for Africa and the Galilee region, and one for Asia/India. The genetic distance between isolates within each of these clusters was often quite large in the studies mentioned above.

It is important to bear in mind that the concept of "clones" in *Leishmania* species comes with some caveats given the large variation in ploidy seen within individual culture-adapted isolates (Sterkers, Crobu et al. 2014). Changing genome structure appears to confer an added level of genetic variability that may serve an adaptive purpose in the downregulation or upregulation of specific genes. The functional effects of this plasticity in genome structure were presented in Chapter 3 and are summarized in Section 6.2.1.

From my analysis of observed allele frequencies with respect to Hardy-Weinberg equilibrium expectations, it became evident that individuals carrying higher heterozygosity than expected, and therefore with low inbreeding coefficients (F_{IT}), are widespread throughout Asia and the Indian subcontinent. On the other hand, the majority of individuals with lower heterozygosity than expected were typically found in Northern and Eastern Africa, and a restricted area of the Middle East region, mainly Israel and Palestine. Denser sampling from neighbouring countries could further elucidate patterns of genetic exchange in these populations, given the apparent overlap between different parasite clusters that could be associated with interbreeding.

Previous studies have suggested the possibility that during the evolutionary history of *L. tropica* a hybridization event associated with an Out-of-Africa origin for this species gave rise to the most common circulating strains found in Asia and India, which have therefore lower inbreeding coefficients. Based on the evidence I presented, I do not exclude this possibility, but warn against the practice of drawing conclusions from the results of inbreeding analyses that mainly simply establish that panmixia is not observed in this species, and can only hint at a directional skew towards inbreeding or outbreeding. The exact quantification of gene flow between population units will require the application of more advanced techniques such as coalescent-based approaches and a more detailed knowledge of the epidemiologically relevant demographic units in this pathogen species.

6.1.2. Wahlund effects and reproduction

As discussed in the Introduction, Wahlund effects have often been invoked to explain patterns of genetic variation in *Leishmania* species. Wahlund effects are defined as the apparent lack of heterozygosity seen in populations where there is population substructuring. The individual subpopulations themselves may be in Hardy-Weinberg equilibrium, and therefore follow panmictic assumptions, but sampling irrespective of these subpopulation boundaries will artificially inflate the expected heterozygosity due to different equilibrium allele frequencies being present in each subpopulation, thus increasing the difference between observed and expected heterozygosity.

Wahlund effects can therefore explain low heterozygosity even when there is significant interbreeding. Genotyping of the limited number of isolates considered in this study shows that homozygous markers are often quite diverse in terms of the number of different allelic sequence variants (Table 2.3), which would confirm the presence of significant substructuring. Given that the inbreeding coefficient calculation did not take into consideration allele sharing between individuals, a DAPC, which is informed by the number of shared alleles between individuals, was performed to gather information on genetic distances between isolates and the number of distinct clusters. Most of the isolates fell into three distinct clusters, which were validated by phylogenetic analysis of the concatenated sequence data.

One of the isolates, L810 from Northern Israel, proved problematic, with conflicting results between the phylogenetic analysis and DAPC. This isolate can be taken as a case in point to further present the unique epidemiological scenarios exemplified by a vector-borne disease such as *Leishmania*, and specifically a species such as *L. tropica* that has been shown to have both a urban/periurban anthroponotic cycle, and a rural zoonotic cycle, depending on the geographical region considered (see Section 1.1.3). This isolate was isolated from a *P. arabicus* sand fly in Northern Israel (Jacobson, Eisenberger et al. 2003, Schnur, Nasereddin et al. 2004), in a region were transmission of *L. tropica* was previously thought to be uniquely due to *P. sergenti*. These and following studies demonstrating its preference for a different vector species and the presence of two distinct transmission cycles in a restricted geographical area (Svobodova, Votypka et al. 2006) exemplify the complexity of the epidemiological setting in endemic areas, and

how parameters such as the presence of a sylvatic reservoir or the biting habits of different vector species, which may be more or less prone to bite again following a first infectious feed, may increase or reduce the chance of co-infection of different strains of *L. tropica* and thus the possibility of mating in the midgut of the insect vector.

A deeper understanding of gene flow between demographic units of *L. tropica* may be offered by studies that tackle population genetics within an epidemiologically relevant setting. Although the clonal theory has focused on the concept of evolutionarily stable "clones" which are transmitted from person to person within epidemic foci, this theory fails to provide useful information on the frequency of hybridization events in natural population of *Leishmania* species, which are now known to occur. Specifically, no distinction is made between lack of hybridization due to physiological boundaries in the capacity of different strains to perform genetic exchange, on one hand, and lack of hybridization due to epidemiological reasons on the other, such as presence of transmission cycles by different vector species, geographic barriers, or even transmission intensity, which for instance may reduce the chances of two different parasite strains meeting within the vector if transmission is low, such as in those areas associated with a sylvatic zoonotic cycle where CL due to *L. tropica* is known to be very sporadic.

6.1.3. Population genetics and models of reproduction

In summary, the population genetics evidence reviewed in this and preceding chapters, and the data I presented in Chapter 2 suggest that genetic exchange consistent with Mendelian inheritance may be present in *L. tropica*, given the pattern of allele sharing observed in isolates from different geographical clusters, and the presence of abundant heterozygosity in a subset of these samples. The presence of meiotic homologous recombination was suggested by previous studies (Krayter, Alam et al. 2014), and the observation of heterozygous markers in linkage with homozygous markers in the samples analyzed in this thesis would seem to confirm this suspicion. In a different species of *Leishmania*, recombination breakpoints were directly detected by WGS in field samples isolated from sand flies (Rogers, Downing et al. 2014). These parasite lines were likely produced by an outcrossing event between two divergent lines with subsequent inbreeding within and between hybrid lines, both of which would produce the genomic pattern of patchy heterozygosity that was observed. However, processes associated with meiotic homologous recombination other than crossing over, such as gene conversion, are suspected to be present in *Leishmania*, although the process lacks molecular characterization, and may give rise to stretches of homozygous regions where heterozygous regions are expected (Akopyants, Kimblin et al. 2009). In human genomes, for which considerable more evidence is available, the rate of gene conversion seems to be higher than the rate of crossovers (Jeffreys and May 2004, Gay, Myers et al. 2007).

The presence of meiotic homologous recombination, in the form of either gene conversion or chromosomal crossing over, would be in line with either the parasexual or sexual model of reproduction as put forward in the Introduction. Although our evidence for recombination is limited, the presence of both homozygous and heterozygous regions is apparent from the WGS, with predominance of either allele 1 or allele 2 in homozygous form in a few isolates and the same alleles present in short stretches in heterozygous form in other isolates (Figure 2.9). This genomic survey of genetic variability in these *L. tropica* isolates provided a useful conceptual framework to further prioritize different strains for experimental crosses, and revealed important differences in the resolving power of WGS technologies when compared to traditional MLST, which only captures variation in a small fraction of the genome (only approximately 15 kb were amplified by the PCR probes used in Chapter 2), and is prone to bias depending on which coding or non-coding regions are selected for genotyping.

6.2. Genome plasticity in *L. tropica*

6.2.1. Variation in somy and effects on transcription

As I presented in Chapter 3, extensive variation in chromosome number is seen in *Leishmania* species, and the *L. tropica* isolates considered in this study were no exception. Cloning of individual cells in a subset of isolates and their paired WGS and RNA-seq analysis revealed mosaicism in the number of chromosomes within a

single *in vitro* adapted parasite population. Whether this mosaicism is also seen to the same extent in natural parasite populations isolated from either individual hosts or individual sand flies is unknown, as all previous studies concern *in vitro* adapted parasites (Sterkers, Crobu et al. 2014). Mitotic divisions associated with parasite proliferation *in vitro* may be intrinsically less likely to have balanced segregation of the chromosomes: i*n vitro* promastigote forms are known to be pleomorphic and display a range of shapes and sizes, are developmentally arrested, and do not proceed to the infectious metacyclic stage (Sacks 1989, Schuster and Sullivan 2002).

Regardless of whether these findings are consistent in field isolates that have never been adapted to growth in culture medium, the results I discuss in Chapter 3 demonstrate that *L. tropica* is in principle capable of tolerating extensive aneuploidy, confirming previous reports of aneuploidy in this and other *Leishmania* species (Bastien, Blaineau et al. 1992, Sterkers, Lachaud et al. 2011). These reports have also been validated by WGS and other experimental approaches (Cruz, Titus et al. 1993, Rogers, Hilley et al. 2011). In previous fluorescence *in situ* hybridization (FISH) studies, asymmetric nuclear chromosome allocations were seen in mitotically dividing cells (Sterkers, Lachaud et al. 2011). The mechanism by which aneuploidy arises was thus postulated to be a defect in chromosomal replication, which explained the odd number of chromosomes seen in all asymmetrically dividing cells. Cloning of individual cell lines from the original population showed that any combination of somy was observed to occur, each chromosome being present in at least two of the following conditions: monosomy, disomy, and trisomy.

In the set of isolates considered in this study, I found most cloned and uncloned lines to be near-diploid, although variation at the level of individual chromosome number was common. One clonal line appeared to be near-triploid both by modelling of the expected haploid read depth and by inspection of allele frequencies on individual chromosomes. The fact that a near triploid clone was generated from a near diploid isolate suggests that considerable standing cellular variation in karyotype exists within culture-adapted isolates.

Aneuploidy is generally deleterious in most eukaryotic species. In cancer cells, for example, karyotype alterations are associated with oncogenesis(Giam and Rancati 2015), while if they occur in human germline cells, they are associated with developmental defects that are often lethal (Hassold and Hunt 2001). The hypothesis I aimed to test in Chapter 3 involved determining whether supernumerary chromosomes are effectively downregulated at the level of steady state transcript levels to compensate for the increased gene dosage. The transcriptional machinery in *Leishmania* lacks the ability to regulate transcriptional initiation, and all genes are constitutively expressed. However, several RNA-binding motifs in the 3' UTR of mRNA transcripts have been found to determine the half-life of specific transcripts via RNA-binding proteins (See Section 3.1). A possible mechanism for reduced steady state transcript levels of supernumerary chromosomes can be in theory attributed to cis-acting RNA-binding proteins, which are also more highly expressed by virtue of being on the same chromosome, and therefore may function in a "balancing act" to proportionally reduce steady state transcript levels to those seen in the disomic state. If the majority of RNA-binding

proteins however function in trans, steady state transcript levels of genes on supernumerary chromosomes would still be higher than expected in the disomic condition.

The results I present in Chapter 3 suggest that gene expression on supernumerary chromosomes behaves in a dose dependent manner, with higher somy associated with higher steady state mRNA levels. Gene dosage effects entirely explain the patterns observed at the chromosome level in two clonal lines generated from the same field isolate, where somy differed at a subset of chromosomes. This leaves open the question as to how *L. tropica* can cope with the increased expression of a large number of genes to maintain cellular homeostasis. Further regulation at the level of translation into protein may explain this ability, possibly via the activity of chaperones such as seen in the heat shock response (Spath, Drini et al. 2015) or via RNA-binding proteins that regulate translation activity via sequestration or modification of the mRNA transcript, while leaving the total level of mRNA transcript (which is what is captured by RNA-seq) present in the cell unchanged.

6.2.2. Variation in gene copy number and effects on transcription

In addition to describing changes in expression due to aneuploidy, one of the aims of Chapter 3 was also to identify which genes were most differentially expressed (DE) between the isolates considered. Remarkably, the most highly represented category in the DE gene set included transmembrane proteins with a transporter function. These are known to play an important role in shuttling

nutrients and other compounds in and out of the cell, thus affecting parasite fitness in unfavourable environmental conditions, such as when a drug is present in the extracellular environment. Many structural variants that we found to be differentially expressed span loci that have been previously implicated in drug resistance, for instance at the FT1 transporter locus in antifolate drug resistance (Ouameur, Girard et al. 2008). In order to further dissect gene dosage effects at the sub-chromosomal levels, two clonal lines of the same isolate were again compared to identify copy number variants (CNVs) that differed between the two lines. Many of the top 30 DE genes identified from the entire sample set of isolates fell into genomic regions that appeared to be large CNVs in these two clones, spanning dozens of genes.

Relative shifts in the read depth ratio between the two clonal lines showed that genes within these CNVs were consistently upregulated in the clone in which the CNV had higher relative read depth. Gene dosage effects thus again largely accounted for differences in expression of genes within CNVs; however, at a minority of genes the trend was opposite to that expected from differences in read depth. In order to address the presence of regulatory mechanisms determining transcript abundance independently of the copy number in the genome context, a screen identifying significant differences between DNA and RNA read depth was performed at a genome-wide SNP level, looking for allele-specific gene expression. The presence of only one allelic variant in RNA transcripts at heterozygous loci would by definition be a violation of gene dosage rules, which otherwise seem to reliably predict in most cases the amount of transcription seen for a given gene. A

large number of SNPs in coding regions gave highly significant p-values, and inspection of the DNA and RNA sequencing reads overlapping those positions indeed showed that only one allele was present in the RNA data, whereas the DNA data showed equal proportions of two different alleles.

Allele-specific gene expression is a previously uncharacterized phenomenon in *Leishmania* and represents an important exception to the predictive power of gene dosage as an explanation for RNA steady state levels. While gene dosage generally can reliably predict transcript levels at a given locus, there are additional layers of regulation that can, in select cases, dictate which allelic variant is actually expressed. Given the overlap between the most significant hits from the allele-specific expression analysis of the two clones, it is highly unlikely that the same gene conversion event has independently occurred in two different cultured parasite populations at the same exact genomic locus, and must instead reflect the presence of variation in DNA sequence affecting transcript abundance. Our data, however, cannot distinguish between sequence variation affecting RNA stability, translational efficiency, or variation in conserved binding motifs mediating interactions between specific gene transcripts and RNA-binding proteins.

In conclusion, I have provided an in depth analysis of patterns of gene expression in *L. tropica*, and confirmed that this species can tolerate significant plasticity in both chromosome number and intrachromosomal structural variants. CNVs appear to be an important mechanism for parasites to upregulate or downregulate, via either amplifications or deletions, the expression of specific genes.

6.2.3. Genome structure and models of reproduction

While the focus of Chapter 3 was on the effects of genome structure on transcription, some conclusions can be drawn on reproductive strategies in *L. tropica*. In particular, the observation that considerable variation in genome structure is observed within individual isolates after being grown *in vitro*, which can therefore be considered true "mosaics", complicates the concept of "clonal lineages" which is central to the clonal theory. Indeed, if the parasite population within a given host or vector is composed by an assemblage of individual parasites with different patterns of aneuploidy, and this variation in genome structure is indeed functional, as suggested by the fact that expression of the genes on these chromosomes increases in a dose-dependent manner, then this added layer of genetic variation must be carefully taken into consideration when determining relationships between different isolates.

Moreover, long runs of homozygosity (LROH) were seen in both cloned and uncloned isolates, resembling the patterns of patchy heterozygosity mentioned in Section 6.1.3 that have been associated with hybridization in *L. donovani* complex isolates from Turkey (Rogers, Downing et al. 2014). Such blocks of homozygosity alternating with heterozygous tracts can be explained by the presence of homologous recombination in natural populations of *L. tropica*. The two models, parasexual and sexual, are therefore supported by the variation in genome structure presented in Chapter 3, with aneuploidy expected to appear through mitotic divisions following cell fusion. The difference between these two models lie in the

type of cell that can perform cell fusion: if the cell is a haploid or near-haploid gamete, then the reproductive strategy would effectively be sexual and involve a meiotic step, during which homologous recombination occurs; if the cell is a somatic cell with a ploidy greater than *n* (where *n* is the set of haploid chromosomes, equal to 36 in *L. tropica*), then no meiosis has occurred and the reproductive strategy would be considered parasexual, although recombination may still be present. The evidence presented here is unfortunately insufficient to draw conclusions on whether a sexual or parasexual cycle is more likely.

6.3. Genetic exchange in *L. tropica*

6.3.1. Sand fly infections and hybridization

In Chapter 4 and Chapter 5 I describe the experimental crosses that were performed in *L. tropica* using combinations of different drug resistant parasite lines, and the results of WGS of 10 different hybrid lines. I hereby provide a short summary of the hybrid recovery rate seen in these crosses and the implications this has for hybridization in natural populations of *L. tropica*.

Out of 7 different cross combinations, only one gave hybrid lines that retained double drug resistance following passage into a new culture flask. These results suggest that only certain combinations of parasite lines have mating competency, or that certain pairings may be more fertile than others: an intriguing possibility that the data seems to suggest is that lines with abundant heterozygosity may have

decreased fertility compared to lines in which homozygosity is more abundant across the whole genome. Heterozygosity and homozygosity can in this case be considered as a proxy measure for outcrossing and inbreeding. None of the crosses in which at least one of the two parental lines was largely heterozygous (see Chapter 2) produced any viable hybrid lines. If isolates that fell into Cluster 3 can be considered "outbred" due to their elevated heterozygosity, as previous analyses seem to suggest, then they must have recently arisen trough an hybridization event between two different "inbred" strains. "Inbred" strains may have greater reproductive potential than "outbred" strains: such a situation is the opposite of that seen in heterosis, or hybrid vigour, where hybrids have greater fitness than either parental line. A condition in which hybrids have a reduced capability to undergo sexual or parasexual reproduction is consistent with the establishment of post-zygotic barriers to interbreeding between parental lines and the new hybrid lines, which may therefore facilitate speciation.

6.3.2. Genomic consequences of hybridization

The main aim of Chapter 5 was to describe the changes observed during hybridization at a genome level. One of the most important objectives was to establish patterns of inheritance of the genetic material as it was passed on from the parental lines to the offspring. Previous crosses in *L. major* and in *L. infantum* found biparental inheritance at a number of genomic markers. We expand these analyses to include inheritance of all biallelic loci, resolution of the gametic phase in all

hybrid lines, somy estimation for all chromosomes, and identification of *de novo* variants associated with hybridization.

All hybrid lines were near-diploid, as were the two parental lines. The only chromosomes showing evidence of aneuploidy were chromosomes 4, 23, and 31. While chromosome 4 was disomic in both parental lines, chromosomes 23 and 31 were trisomic and tetrasomic, respectively. Interestingly, all chromosomes were homozygous in the parents, while all chromosomes in the hybrid lines were heterozygous. Phasing of disomic chromosomes confirms biparental inheritance, and while multisomic chromosomes cannot be phased with traditional approaches, the same pattern of heterozygosity was seen in all hybrid lines, suggesting that meiotic processes may consistently reduce the number of multisomic chromosomes and in a sense "reset" aneuploidy.

As evidenced from allelic plots, the inheritance of markers on disomic chromosomes is largely Mendelian. A minority of SNP positions appeared to be violating Mendelian inheritance rules. A more accurate analysis of the distribution of these SNPs in the genome could not be perfomed due to the lack of annotation of the reference genome used. An automated annotation pipeline has been developed for *L. tropica* and will be deployed for future studies. Many *de novo* structural variants were also detected in the hybrids. Herein I have limited myself to providing a list of the different types of structural variants detected and have postponed a more detailed analysis of the distribution of these variants with respect to coding regions to subsequent studies that will be informed by an accurate genome annotation.

6.3.3. Hybridization and models of reproduction

Tracking the inheritance of whole chromosomes can shed some light on the processes governing *L. tropica* biparental inheritance following cell fusion. All trisomic and tetrasomic chromosomes in the hybrid lines were heterozygous (with allele frequency peaks at 0.33 and 0.67 if trisomic, and 0.5 if tetrasomic). This consistency across several independent mating events suggests that non-random segregation of the chromosomes occurs during hybridization. The presence of a parasexual cycle would involve cell fusion followed by random loss of certain chromosomes during mitotic division. This was not observed in the data presented in Chapter 5. Consistently, each hybrid line inherits at least one chromosome from each parent, suggesting a more complicated mechanism being at work than concerted loss of chromosomes during mitotic divisions following a cell fusion event between near-diploid cells, as predicted by the parasexual model. By the combinatorial formula, the probability that the 33 disomic chromosomes that were phased in the hybrids each inherited a single chromosome from each parent if they originated from a cell fusion between the two diploid parents and subsequently lost two of the four chromosomes at random is approximately $1x10^{-6}$. The probability decreases if we account for potential aneuploidy that arises in that step, providing convincing evidence that a meiotic haploid or near haploid stage is very likely.

These results are indeed in line with what is expected from a meiotic stage being present in *L. tropica*. The obligatory generation of haploid gametes would explain why all hybrid lines inherit at least one chromosome from each parent.

Chromosomes that are trisomic in the parental lines cannot be split evenly in the resulting gametes, thus resulting in two possible near haploid gametes with either a monosomic or a disomic version of these chromosomes. Both of these possibilities are in fact seen in the hybrid lines for chromosome 23, with some hybrids being disomic (1 + 1) and others being trisomic (2 + 1). Chromosomes that are tetrasomic in the parental lines can be split evenly in the resulting gametes, and seem to be preferentially passed on as disomic chromosomes in the gametes (2 + 2). This would explain the maintenance of tetrasomy at chromosome 31 in most *L. tropica* isolates analysed in this thesis, and the fact that all hybrids had a single peak at 0.5 in the allele frequencies for this chromosome.

The exception to this remarkable consistency is given by chromosome 4, which is trisomic in two of the hybrid lines. Such isolated cases of aneuploidy in the offspring when the parental lines have balanced ploidy could arise during the *in vitro* growth phase rather than as a result of chromosomal segregation during meiosis. Indeed, over prolonged *in vitro* culture, the consistency in chromosome number seen across independent mating events could be lost due to asymmetric cell divisions (Section 6.2.1).

6.4. Future directions

Additional crosses of *L. tropica*, especially backcrosses involving the F1 hybrids, will help generate enough recombination events in the progeny to break associations between even closely linked markers. As described in Section 5.1, near

isogenic lines produced by a backcross of the F1 hybrids to one of the parental lines provide some of the best tools for forward genetic mapping of traits of interest in several model organisms. *L. tropica* has very few *in vitro* or *in vivo* models of pathogenesis, although infection in hamsters has been described as a possible model of viscerotropic disease. The establishment of simple, quantitative, biologically relevant phenotype readouts will help screen a large number of hybrid clones for presence or absence of the trait of interest. A high throughput, accurate, and reliable phenotyping platform will be necessary to obtain informative results.

In the short term, more sensitive approaches to detect recombination events on disomic chromosomes in the presence of extensive homozygosity will be developed. These statistical methods need to take into consideration the unique genomic features of *Leishmania* and other kinetoplastid parasites, and will focus on detecting recombinant haplotype blocks between heterozygous SNPs in different F1 hybrids. The exact location of the crossover won't be known if it falls within a stretch of homozygous SNPs, but evidence that a recombination has occurred between the two heterozygous SNPs can be gathered.

The development of *ad hoc* phasing approaches for multisomic chromosomes will prove to be difficult given the large number of possible phasing solutions. Phasing and detection of recombination are both areas of active research in bioinformatics, and in humans these benefit from the presence of large projects such as the HapMap or the 1000 Genomes Project that provide a steady influx of relevant information for this type of analyses.

Crosses between parasite lines engineered with different fluorescent markers will be used to identify the meiosis competent stage in *L. tropica*. A cross between green and red fluorescent lines can point to the location within the midgut where hybrids, which will fluoresce yellow, make their first appearance. Morphological observations on the type of cells expressing red, green, or both fluorescent markers can narrow down identification of the promastigote stage undergoing meiosis, as has been done for *T. brucei* crosses.

The reference genome that was generated utilizing optically mapped data and assembled independently of synteny with other *Leishmania* spp. will be annotated using a combination of coding region prediction tools and tools that identify homology with other species such as *L. major*. Such an annotated genome will be used to further dissect structural changes and the location of new single nucleotide variants that occur during hybridization. The presence of abundant RNA-seq data for this species will allow discovery and annotation of transcription start and end sites.

In terms of genome plasticity and the effects of structural changes on transcription, spike-in RNA-seq of lines with polyploid genomes will allow absolute quantitation of RNA levels in diploid, triploid, and tetraploid lines. By comparing the relative gene expression observed within experimental samples to a known amount of a reference RNA transcript spiked into the preparation, these type of studies will be able to better inform our understanding of how *Leishmania* is able to cope with polyploidy and aneuploidy.

In plants, aneuploidy seems to be associated with dampening of transcription to avoid deleterious effects on cellular homeostasis. The effects of ploidy on transcription can be better understood in terms of absolute, rather than relative, mRNA levels. The exciting prospect of performing single-cell WGS and RNA-seq, which has now been achieved in several unicellular and multicellular eukaryotes, will allow resolution of the dynamics associated with mosaic aneuploidy. The possibility of performing single-cell WGS on parasites from a clinical or field isolate without prior culturing can reveal the extent of mosaic aneuploidy *in natura*.

Lastly, large-scale population genetics studies of *L. tropica* and other *Leishmania* species are currently underway. These will provide crucial information for arriving at a more accurate and effective systematics in *Leishmania*. Evolutionary relationships between different species or subspecies can be explored with these large datasets. However, to reach a more detailed understanding of the population dynamics in active foci of transmission, dense sampling from a localized area in conjunction with adequate epidemiological surveys will need to be carried out in order to assess the relative frequency of hybridization in different endemic settings.

REFERENCES

. "Genetic Testing Registry, National Institutes of Health." from http://www.ncbi.nlm.nih.gov/gtr/.

Ajaoud, M., N. Es-sette, S. Hamdi, A. L. El-Idrissi, M. Riyad and M. Lemrani (2013). "Detection and molecular typing of *Leishmania tropica* from *Phlebotomus sergenti* and lesions of cutaneous leishmaniasis in an emerging focus of Morocco." Parasit Vectors **6**: 217.

Akopyants, N. S., N. Kimblin, N. Secundino, R. Patrick, N. Peters, P. Lawyer, D. E. Dobson, S. M. Beverley and D. L. Sacks (2009). "Demonstration of genetic exchange during cyclical development of *Leishmania* in the sand fly vector." Science **324**(5924): 265-268.

Alam, M. Z., K. Kuhls, C. Schweynoch, S. Sundar, S. Rijal, A. K. Shamsuzzaman, B. V. Raju, P. Salotra, J. C. Dujardin and G. Schonian (2009). "Multilocus microsatellite typing (MLMT) reveals genetic homogeneity of *Leishmania donovani* strains in the Indian subcontinent." Infect Genet Evol **9**(1): 24-31.

Alborzi, A., M. Rasouli and A. Shamsizadeh (2006). "*Leishmania* tropica-isolated patient with visceral leishmaniasis in southern Iran." Am J Trop Med Hyg **74**(2): 306-307.

Alcolea, P. J., A. Alonso, M. J. Gomez, M. Postigo, R. Molina, M. Jimenez and V. Larraga (2014). "Stage-specific differential gene expression in *Leishmania infantum*: from the foregut of *Phlebotomus perniciosus* to the human phagocyte." BMC Genomics **15**: 849.

Alvar, J., I. D. Velez, C. Bern, M. Herrero, P. Desjeux, J. Cano, J. Jannin, M. den Boer and WHO Leishmaniasis Control Team (2012). "Leishmaniasis worldwide and global estimates of its incidence." PLoS One **7**(5): e35671.

Anders, S. and W. Huber (2010). "Differential expression analysis for sequence count data." Genome Biol **11**(10): R106.

Anders, S., P. T. Pyl and W. Huber (2014). "HTSeq-a Python framework to work with high-throughput sequencing data." Bioinformatics.

Anderson, B. A., I. L. Wong, L. Baugh, G. Ramasamy, P. J. Myler and S. M. Beverley (2013). "Kinetoplastid-specific histone variant functions are conserved in *Leishmania major*." Mol Biochem Parasitol **191**(2): 53-57.

Anderson, C. F., R. Lira, S. Kamhawi, Y. Belkaid, T. A. Wynn and D. Sacks (2008). "IL-10 and TGF-beta control the establishment of persistent and transmissible infections produced by *Leishmania tropica* in C57BL/6 mice." J Immunol **180**(6): 4090-4097.

Aphasizhev, R. and I. Aphasizheva (2011). "Mitochondrial RNA processing in trypanosomes." Res Microbiol **162**(7): 655-663.

Banuls, A. L., F. Guerrini, F. Le Pont, C. Barrera, I. Espinel, R. Guderian, R. Echeverria and M. Tibayrenc (1997). "Evidence for hybridization by multilocus enzyme electrophoresis and random amplified polymorphic DNA between *Leishmania braziliensis* and *Leishmania panamensis/guyanensis* in Ecuador." J Eukaryot Microbiol **44**(5): 408-411.

Banuls, A. L., M. Hide and M. Tibayrenc (1999). "Molecular epidemiology and evolutionary genetics of *Leishmania* parasites." Int J Parasitol **29**(8): 1137-1147.
Banuls, A. L., R. Jonquieres, F. Guerrini, F. Le Pont, C. Barrera, I. Espinel, R. Guderian, R. Echeverria and M. Tibayrenc (1999). "Genetic analysis of *Leishmania* parasites in Ecuador: are *Leishmania (Viannia) panamensis* and *Leishmania (V.) guyanensis* distinct taxa?" Am J Trop Med Hyg **61**(5): 838-845.
Bastien, P., C. Blaineau and M. Pages (1992). "Molecular karyotype analysis in *Leishmania*." Subcell Biochem **18**: 131-187.
Bates, P. A. (2007). "Transmission of *Leishmania* metacyclic promastigotes by phlebotomine sand flies." Int J Parasitol **37**(10): 1097-1106.
Bates, P. A. and M. E. Rogers (2004). "New insights into the developmental biology and transmission mechanisms of *Leishmania*." Curr Mol Med **4**(6): 601-609.
Belkaid, Y., C. A. Piccirillo, S. Mendez, E. M. Shevach and D. L. Sacks (2002). "CD4+CD25+ regulatory T cells control *Leishmania major* persistence and immunity." Nature **420**(6915): 502-507.
Belli, A. A., M. A. Miles and J. M. Kelly (1994). "A putative *Leishmania panamensis/Leishmania braziliensis* hybrid is a causative agent of human cutaneous leishmaniasis in Nicaragua." Parasitology **109 (Pt 4)**: 435-442.
Bente, M., S. Harder, M. Wiesgigl, J. Heukeshoven, C. Gelhaus, E. Krause, J. Clos and I. Bruchhaus (2003). "Developmentally induced changes of the proteome in the protozoan parasite *Leishmania donovani*." Proteomics **3**(9): 1811-1829.
Beverley, S. M., J. A. Coderre, D. V. Santi and R. T. Schimke (1984). "Unstable DNA amplifications in methotrexate-resistant *Leishmania* consist of extrachromosomal circles which relocalize during stabilization." Cell **38**(2): 431-439.
Beyrer, C., J. C. Villar, V. Suwanvanichkij, S. Singh, S. D. Baral and E. J. Mills (2007). "Neglected diseases, civil conflicts, and the right to health." Lancet **370**(9587): 619-627.
Blaineau, C., P. Bastien and M. Pages (1992). "Multiple forms of chromosome I, II and V in a restricted population of *Leishmania infantum* contrasting with monomorphism in individual strains suggest haploidy or automixy." Mol Biochem Parasitol **50**(2): 197-204.
Bonfield, J. K. and A. Whitwham (2010). "Gap5--editing the billion fragment sequence assembly." Bioinformatics **26**(14): 1699-1703.
Bousslimi, N., S. Ben-Ayed, I. Ben-Abda, K. Aoun and A. Bouratbine (2012). "Natural infection of North African gundi (*Ctenodactylus gundi*) by *Leishmania tropica* in the focus of cutaneous leishmaniasis, Southeast Tunisia." Am J Trop Med Hyg **86**(6): 962-965.
Bringaud, F., M. Muller, G. C. Cerqueira, M. Smith, A. Rochette, N. M. El-Sayed, B. Papadopoulou and E. Ghedin (2007). "Members of a large retroposon family are determinants of post-transcriptional gene expression in *Leishmania*." PLoS Pathog **3**(9): 1291-1307.
Britto, C., C. Ravel, P. Bastien, C. Blaineau, M. Pages, J. P. Dedet and P. Wincker (1998). "Conserved linkage groups associated with large-scale chromosomal rearrangements between Old World and New World *Leishmania* genomes." Gene **222**(1): 107-117.

Brooker, S., N. Mohammed, K. Adil, S. Agha, R. Reithinger, M. Rowland, I. Ali and J. Kolaczinski (2004). "Leishmaniasis in refugee and local Pakistani populations." Emerg Infect Dis **10**(9): 1681-1684.
Calvo-Alvarez, E., R. Alvarez-Velilla, M. Jimenez, R. Molina, Y. Perez-Pertejo, R. Balana-Fouce and R. M. Reguera (2014). "First evidence of intraclonal genetic exchange in trypanosomatids using two *Leishmania infantum* fluorescent transgenic clones." PLoS Negl Trop Dis **8**(9): e3075.
Campbell, C. D. and E. E. Eichler (2013). "Properties and rates of germline mutations in humans." Trends Genet **29**(10): 575-584.
Carninci, P., A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C. A. Semple, M. S. Taylor, P. G. Engstrom, M. C. Frith, A. R. Forrest, W. B. Alkema, S. L. Tan, C. Plessy, R. Kodzius, T. Ravasi, T. Kasukawa, S. Fukuda, M. Kanamori-Katayama, Y. Kitazume, H. Kawaji, C. Kai, M. Nakamura, H. Konno, K. Nakano, S. Mottagui-Tabar, P. Arner, A. Chesi, S. Gustincich, F. Persichetti, H. Suzuki, S. M. Grimmond, C. A. Wells, V. Orlando, C. Wahlestedt, E. T. Liu, M. Harbers, J. Kawai, V. B. Bajic, D. A. Hume and Y. Hayashizaki (2006). "Genome-wide analysis of mammalian promoter architecture and evolution." Nat Genet **38**(6): 626-635.
Chaara, D., C. Ravel, A. Banuls, N. Haouas, P. Lami, L. Talignani, F. El Baidouri, K. Jaouadi, Z. Harrat, J. P. Dedet, H. Babba and F. Pratlong (2015). "Evolutionary history of *Leishmania killicki* (synonymous *Leishmania tropica*) and taxonomic implications." Parasit Vectors **8**: 198.
Chargui, N., A. Amro, N. Haouas, G. Schonian, H. Babba, S. Schmidt, C. Ravel, M. Lefebvre, P. Bastien, E. Chaker, K. Aoun, M. Zribi and K. Kuhls (2009). "Population structure of Tunisian *Leishmania infantum* and evidence for the existence of hybrids and gene flow between genetically different populations." Int J Parasitol **39**(7): 801-811.
Crowley, J. J., V. Zhabotynsky, W. Sun, S. Huang, I. K. Pakatci, Y. Kim, J. R. Wang, A. P. Morgan, J. D. Calaway, D. L. Aylor, Z. Yun, T. A. Bell, R. J. Buus, M. E. Calaway, J. P. Didion, T. J. Gooch, S. D. Hansen, N. N. Robinson, G. D. Shaw, J. S. Spence, C. R. Quackenbush, C. J. Barrick, R. J. Nonneman, K. Kim, J. Xenakis, Y. Xie, W. Valdar, A. B. Lenarcic, W. Wang, C. E. Welsh, C. P. Fu, Z. Zhang, J. Holt, Z. Guo, D. W. Threadgill, L. M. Tarantino, D. R. Miller, F. Zou, L. McMillan, P. F. Sullivan and F. P. de Villena (2015). "Corrigendum: analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance." Nat Genet **47**(6): 690.
Crowley, J. J., V. Zhabotynsky, W. Sun, S. Huang, I. K. Pakatci, Y. Kim, J. R. Wang, A. P. Morgan, J. D. Calaway, D. L. Aylor, Z. Yun, T. A. Bell, R. J. Buus, M. E. Calaway, J. P. Didion, T. J. Gooch, S. D. Hansen, N. N. Robinson, G. D. Shaw, J. S. Spence, C. R. Quackenbush, C. J. Barrick, R. J. Nonneman, K. Kim, J. Xenakis, Y. Xie, W. Valdar, A. B. Lenarcic, W. Wang, C. E. Welsh, C. P. Fu, Z. Zhang, J. Holt, Z. Guo, D. W. Threadgill, L. M. Tarantino, D. R. Miller, F. Zou, L. McMillan, P. F. Sullivan and F. Pardo-Manuel de Villena (2015). "Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance." Nat Genet **47**(4): 353-360.
Cruz, A. K., R. Titus and S. M. Beverley (1993). "Plasticity in chromosome number and testing of essential genes in *Leishmania* by targeting." Proc Natl Acad Sci U S A **90**(4): 1599-1603.

Cunningham, M. L. and S. M. Beverley (2001). "Pteridine salvage throughout the *Leishmania* infectious cycle: implications for antifolate chemotherapy." Mol Biochem Parasitol **113**(2): 199-213.
Curotto de Lafaille, M. A., A. Laban and D. F. Wirth (1992). "Gene expression in *Leishmania*: analysis of essential 5' DNA sequences." Proc Natl Acad Sci U S A **89**(7): 2703-2707.
Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin and G. Genomes Project Analysis (2011). "The variant call format and VCFtools." Bioinformatics **27**(15): 2156-2158.
Das, A., M. Banday and V. Bellofatto (2008). "RNA polymerase transcription machinery in trypanosomes." Eukaryot Cell **7**(3): 429-434.
De Gaudenzi, J. G., G. Noe, V. A. Campo, A. C. Frasch and A. Cassola (2011). "Gene expression regulation in trypanosomatids." Essays Biochem **51**: 31-46.
Delgado, O., E. Cupolillo, R. Bonfante-Garrido, S. Silva, E. Belfort, G. Grimaldi Junior and H. Momen (1997). "Cutaneous leishmaniasis in Venezuela caused by infection with a new hybrid between *Leishmania (Viannia) braziliensis* and *L. (V.) guyanensis*." Mem Inst Oswaldo Cruz **92**(5): 581-582.
Depledge, D. P., K. J. Evans, A. C. Ivens, N. Aziz, A. Maroof, P. M. Kaye and D. F. Smith (2009). "Comparative expression profiling of *Leishmania*: modulation in gene expression between species and in different host genetic backgrounds." PLoS Negl Trop Dis **3**(7): e476.
Dereure, J., J. A. Rioux, M. Gallego, J. Perieres, F. Pratlong, J. Mahjour and H. Saddiki (1991). "*Leishmania tropica* in Morocco: infection in dogs." Trans R Soc Trop Med Hyg **85**(5): 595.
Desjeux, P. (2004). "Leishmaniasis: current situation and new perspectives." Comp Immunol Microbiol Infect Dis **27**(5): 305-318.
Dillon, D. C., C. H. Day, J. A. Whittle, A. J. Magill and S. G. Reed (1995). "Characterization of a *Leishmania tropica* antigen that detects immune responses in Desert Storm viscerotropic leishmaniasis patients." Proc Natl Acad Sci U S A **92**(17): 7981-7985.
Dillon, L. A., K. Okrah, V. K. Hughitt, R. Suresh, Y. Li, M. C. Fernandes, A. T. Belew, H. Corrada Bravo, D. M. Mosser and N. M. El-Sayed (2015). "Transcriptomic profiling of gene expression and RNA processing during *Leishmania major* differentiation." Nucleic Acids Res **43**(14): 6799-6813.
Diro, E., J. van Griensven, R. Mohammed, R. Colebunders, M. Asefa, A. Hailu and L. Lynen (2015). "Atypical manifestations of visceral leishmaniasis in patients with HIV in north Ethiopia: a gap in guidelines for the management of opportunistic infections in resource poor settings." Lancet Infect Dis **15**(1): 122-129.
Downing, T., H. Imamura, S. Decuypere, T. G. Clark, G. H. Coombs, J. A. Cotton, J. D. Hilley, S. de Doncker, I. Maes, J. C. Mottram, M. A. Quail, S. Rijal, M. Sanders, G. Schonian, O. Stark, S. Sundar, M. Vanaerschot, C. Hertz-Fowler, J. C. Dujardin and M. Berriman (2011). "Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance." Genome Res **21**(12): 2143-2156.

El Baidouri, F., L. Diancourt, V. Berry, F. Chevenet, F. Pratlong, P. Marty and C. Ravel (2013). "Genetic structure and evolution of the *Leishmania* genus in Africa and Eurasia: what does MLSA tell us." PLoS Negl Trop Dis **7**(6): e2255.
Evans, D. A., W. P. Kennedy, S. Elbihari, C. J. Chapman, V. Smith and W. Peters (1987). "Hybrid formation within the genus *Leishmania*?" Parassitologia **29**(2-3): 165-173.
Fong, D., M. Wallach, J. Keithly, P. W. Melera and K. P. Chang (1984). "Differential expression of mRNAs for alpha- and beta-tubulin during differentiation of the parasitic protozoan *Leishmania mexicana*." Proc Natl Acad Sci U S A **81**(18): 5782-5786.
Fraidenraich, D., C. Pena, E. L. Isola, E. M. Lammel, O. Coso, A. D. Anel, S. Pongor, F. Baralle, H. N. Torres and M. M. Flawia (1993). "Stimulation of *Trypanosoma cruzi* adenylyl cyclase by an alpha D-globin fragment from *Triatoma* hindgut: effect on differentiation of epimastigote to trypomastigote forms." Proc Natl Acad Sci U S A **90**(21): 10140-10144.
Garrison, E. and G. Marth (2012) "Haplotype-based variant detection from short-read sequencing." arXiv preprint **arXiv:1207.3907 [q-bio.GN]**.
Gaunt, M. W., M. Yeo, I. A. Frame, J. R. Stothard, H. J. Carrasco, M. C. Taylor, S. S. Mena, P. Veazey, G. A. Miles, N. Acosta, A. R. de Arias and M. A. Miles (2003). "Mechanism of genetic exchange in American trypanosomes." Nature **421**(6926): 936-939.
Gay, J., S. Myers and G. McVean (2007). "Estimating meiotic gene conversion rates from population genetic data." Genetics **177**(2): 881-894.
Gebre-Michael, T., M. Balkew, A. Ali, A. Ludovisi and M. Gramiccia (2004). "The isolation of *Leishmania tropica* and *L. aethiopica* from *Phlebotomus* (Paraphlebotomus) species (Diptera: Psychodidae) in the Awash Valley, northeastern Ethiopia." Trans R Soc Trop Med Hyg **98**(1): 64-70.
Gelanew, T., K. Kuhls, Z. Hurissa, T. Weldegebreal, W. Hailu, A. Kassahun, T. Abebe, A. Hailu and G. Schonian (2010). "Inference of population structure of *Leishmania donovani* strains isolated from different Ethiopian visceral leishmaniasis endemic areas." PLoS Negl Trop Dis **4**(11): e889.
Genest, P. A., B. Ter Riet, T. Cijsouw, H. G. van Luenen and P. Borst (2007). "Telomeric localization of the modified DNA base J in the genome of the protozoan parasite *Leishmania*." Nucleic Acids Res **35**(7): 2116-2124.
Giam, M. and G. Rancati (2015). "Aneuploidy and chromosomal instability in cancer: a jackpot to chaos." Cell Div **10**: 3.
Gibson, W. and L. Garside (1990). "Kinetoplast DNA minicircles are inherited from both parents in genetic hybrids of *Trypanosoma brucei*." Mol Biochem Parasitol **42**(1): 45-53.
Gibson, W., L. Peacock, V. Ferris, K. Williams and M. Bailey (2008). "The use of yellow fluorescent hybrids to indicate mating in *Trypanosoma brucei*." Parasit Vectors **1**(1): 4.
Gilinger, G. and V. Bellofatto (2001). "Trypanosome spliced leader RNA genes contain the first identified RNA polymerase II gene promoter in these organisms." Nucleic Acids Res **29**(7): 1556-1564.
Greig, D. (2007). "A screen for recessive speciation genes expressed in the gametes of F1 hybrid yeast." PLoS Genet **3**(2): e21.

Greig, D., E. J. Louis, R. H. Borts and M. Travisano (2002). "Hybrid speciation in experimental populations of yeast." Science **298**(5599): 1773-1775.
Gupta, S. K., S. Carmi, H. Waldman Ben-Asher, I. D. Tkacz, I. Naboishchikov and S. Michaeli (2013). "Basal splicing factors regulate the stability of mature mRNAs in trypanosomes." J Biol Chem **288**(7): 4991-5006.
Gurung, P., R. Karki, P. Vogel, M. Watanabe, M. Bix, M. Lamkanfi and T. D. Kanneganti (2015). "An NLRP3 inflammasome-triggered Th2-biased adaptive immune response promotes leishmaniasis." J Clin Invest **125**(3): 1329-1338.
Hadighi, R., M. Mohebali, P. Boucher, H. Hajjaran, A. Khamesipour and M. Ouellette (2006). "Unresponsiveness to Glucantime treatment in Iranian cutaneous leishmaniasis due to drug-resistant *Leishmania tropica* parasites." PLoS Med **3**(5): e162.
Haile, S., A. Dupe and B. Papadopoulou (2008). "Deadenylation-independent stage-specific mRNA degradation in *Leishmania*." Nucleic Acids Res **36**(5): 1634-1644.
Haile, S., A. M. Estevez and C. Clayton (2003). "A role for the exosome in the in vivo degradation of unstable mRNAs." RNA **9**(12): 1491-1501.
Hajduk, S. and T. Ochsenreiter (2010). "RNA editing in kinetoplastids." RNA Biol **7**(2): 229-236.
Hassold, T. and P. Hunt (2001). "To err (meiotically) is human: the genesis of human aneuploidy." Nat Rev Genet **2**(4): 280-291.
He, D., O. Fiz-Palacios, C. J. Fu, J. Fehling, C. C. Tsai and S. L. Baldauf (2014). "An alternative root for the eukaryote tree of life." Curr Biol **24**(4): 465-470.
Heinzel, F. P., M. D. Sadick, S. S. Mutha and R. M. Locksley (1991). "Production of interferon gamma, interleukin 2, interleukin 4, and interleukin 10 by CD4+ lymphocytes in vivo during healing and progressive murine leishmaniasis." Proc Natl Acad Sci U S A **88**(16): 7011-7015.
Holzer, T. R., K. K. Mishra, J. H. LeBowitz and J. D. Forney (2008). "Coordinate regulation of a family of promastigote-enriched mRNAs by the 3'UTR PRE element in *Leishmania mexicana*." Mol Biochem Parasitol **157**(1): 54-64.
Hotez, P. J., L. Savioli and A. Fenwick (2012). "Neglected tropical diseases of the Middle East and North Africa: review of their prevalence, distribution, and opportunities for control." PLoS Negl Trop Dis **6**(2): e1475.
Houseley, J., J. LaCava and D. Tollervey (2006). "RNA-quality control by the exosome." Nat Rev Mol Cell Biol **7**(7): 529-539.
Inbar, E., N. S. Akopyants, M. Charmoy, A. Romano, P. Lawyer, D. E. Elnaiem, F. Kauffmann, M. Barhoumi, M. Grigg, K. Owens, M. Fay, D. E. Dobson, J. Shaik, S. M. Beverley and D. Sacks (2013). "The mating competence of geographically diverse *Leishmania major* strains in their natural and unnatural sand fly vectors." PLoS Genet **9**(7): e1003672.
Ivens, A. C., C. S. Peacock, E. A. Worthey, L. Murphy, G. Aggarwal, M. Berriman, E. Sisk, M. A. Rajandream, E. Adlem, R. Aert, A. Anupama, Z. Apostolou, P. Attipoe, N. Bason, C. Bauser, A. Beck, S. M. Beverley, G. Bianchettin, K. Borzym, G. Bothe, C. V. Bruschi, M. Collins, E. Cadag, L. Ciarloni, C. Clayton, R. M. Coulson, A. Cronin, A. K. Cruz, R. M. Davies, J. De Gaudenzi, D. E. Dobson, A. Duesterhoeft, G. Fazelina, N. Fosker, A. C. Frasch, A. Fraser, M. Fuchs, C. Gabel, A. Goble, A. Goffeau, D. Harris, C. Hertz-Fowler, H. Hilbert, D. Horn, Y. Huang, S. Klages, A. Knights, M. Kube, N. Larke, L. Litvin, A.

Lord, T. Louie, M. Marra, D. Masuy, K. Matthews, S. Michaeli, J. C. Mottram, S. Muller-Auer, H. Munden, S. Nelson, H. Norbertczak, K. Oliver, S. O'Neil, M. Pentony, T. M. Pohl, C. Price, B. Purnelle, M. A. Quail, E. Rabbinowitsch, R. Reinhardt, M. Rieger, J. Rinta, J. Robben, L. Robertson, J. C. Ruiz, S. Rutter, D. Saunders, M. Schafer, J. Schein, D. C. Schwartz, K. Seeger, A. Seyler, S. Sharp, H. Shin, D. Sivam, R. Squares, S. Squares, V. Tosato, C. Vogt, G. Volckaert, R. Wambutt, T. Warren, H. Wedler, J. Woodward, S. Zhou, W. Zimmermann, D. F. Smith, J. M. Blackwell, K. D. Stuart, B. Barrell and P. J. Myler (2005). "The genome of the kinetoplastid parasite, *Leishmania major*." Science **309**(5733): 436-442.

Jackson, A. P., S. Vaughan and K. Gull (2006). "Comparative genomics and concerted evolution of beta-tubulin paralogs in *Leishmania* spp." BMC Genomics **7**: 137.

Jacobson, R. L., C. L. Eisenberger, M. Svobodova, G. Baneth, J. Sztern, J. Carvalho, A. Nasereddin, M. El Fari, U. Shalom, P. Volf, J. Votypka, J. P. Dedet, F. Pratlong, G. Schonian, L. F. Schnur, C. L. Jaffe and A. Warburg (2003). "Outbreak of cutaneous leishmaniasis in northern Israel." J Infect Dis **188**(7): 1065-1073.

Jamann, T. M., P. J. Balint-Kurti and J. B. Holland (2015). "QTL mapping using high-throughput sequencing." Methods Mol Biol **1284**: 257-285.

Jaouadi, K., N. Haouas, D. Chaara, M. Gorcii, N. Chargui, D. Augot, F. Pratlong, J. P. Dedet, S. Ettlijani, H. Mezhoud and H. Babba (2011). "First detection of *Leishmania killicki* (Kinetoplastida, Trypanosomatidae) in *Ctenodactylus gundi* (Rodentia, Ctenodactylidae), a possible reservoir of human cutaneous leishmaniasis in Tunisia." Parasit Vectors **4**: 159.

Jeffreys, A. J. and C. A. May (2004). "Intense and highly localized gene conversion activity in human meiotic crossover hot spots." Nat Genet **36**(2): 151-156.

Jenni, L., S. Marti, J. Schweizer, B. Betschart, R. W. Le Page, J. M. Wells, A. Tait, P. Paindavoine, E. Pays and M. Steinert (1986). "Hybrid formation between African trypanosomes during cyclical transmission." Nature **322**(6075): 173-175.

Jombart, T. (2008). "adegenet: a R package for the multivariate analysis of genetic markers." Bioinformatics **24**(11): 1403-1405.

Jombart, T., S. Devillard and F. Balloux (2010). "Discriminant analysis of principal components: a new method for the analysis of genetically structured populations." BMC Genet **11**: 94.

Kamhawi, S., Y. Belkaid, G. Modi, E. Rowton and D. Sacks (2000). "Protection against cutaneous leishmaniasis resulting from bites of uninfected sand flies." Science **290**(5495): 1351-1354.

Kamhawi, S., G. B. Modi, P. F. Pimenta, E. Rowton and D. L. Sacks (2000). "The vectorial competence of *Phlebotomus sergenti* is specific for *Leishmania tropica* and is controlled by species-specific, lipophosphoglycan-mediated midgut attachment." Parasitology **121 (Pt 1)**: 25-33.

Kane, M. M. and D. M. Mosser (2001). "The role of IL-10 in promoting disease progression in leishmaniasis." J Immunol **166**(2): 1141-1147.

Kelly, J. M., J. M. Law, C. J. Chapman, G. J. Van Eys and D. A. Evans (1991). "Evidence of genetic recombination in *Leishmania*." Mol Biochem Parasitol **46**(2): 253-263.

Khanra, S., S. Datta, D. Mondal, P. Saha, S. K. Bandopadhyay, S. Roy and M. Manna (2012). "RFLPs of ITS, ITS1 and hsp70 amplicons and sequencing of ITS1 of recent

clinical isolates of Kala-azar from India and Bangladesh confirms the association of *L. tropica* with the disease." Acta Trop **124**(3): 229-234.
Kimblin, N., N. Peters, A. Debrabant, N. Secundino, J. Egen, P. Lawyer, M. P. Fay, S. Kamhawi and D. Sacks (2008). "Quantification of the infectious dose of *Leishmania major* transmitted to the skin by single sand flies." Proc Natl Acad Sci U S A **105**(29): 10125-10130.
Kramer, S. and M. Carrington (2011). "Trans-acting proteins regulating mRNA maturation, stability and translation in trypanosomatids." Trends Parasitol **27**(1): 23-30.
Krayter, L., M. Z. Alam, M. Rhajaoui, L. F. Schnur and G. Schonian (2014). "Multilocus Microsatellite Typing reveals intra-focal genetic diversity among strains of Leishmania tropica in Chichaoua Province, Morocco." Infect Genet Evol **28**: 233-239.
Krayter, L., R. A. Bumb, K. Azmi, J. Wuttke, M. D. Malik, L. F. Schnur, P. Salotra and G. Schonian (2014). "Multilocus microsatellite typing reveals a genetic relationship but, also, genetic differences between Indian strains of *Leishmania tropica* causing cutaneous leishmaniasis and those causing visceral leishmaniasis." Parasit Vectors **7**: 123.
Kreutzer, R. D., J. J. Yemma, M. Grogl, R. B. Tesh and T. I. Martin (1994). "Evidence of sexual reproduction in the protozoan parasite *Leishmania* (Kinetoplastida: Trypanosomatidae)." Am J Trop Med Hyg **51**(3): 301-307.
Kuhls, K., M. Z. Alam, E. Cupolillo, G. E. Ferreira, I. L. Mauricio, R. Oddone, M. D. Feliciangeli, T. Wirth, M. A. Miles and G. Schonian (2011). "Comparative microsatellite typing of new world *Leishmania infantum* reveals low heterogeneity among populations and its recent old world origin." PLoS Negl Trop Dis **5**(6): e1155.
Kuhls, K., L. Keilonat, S. Ochsenreither, M. Schaar, C. Schweynoch, W. Presber and G. Schonian (2007). "Multilocus microsatellite typing (MLMT) reveals genetically isolated populations between and within the main endemic regions of visceral leishmaniasis." Microbes Infect **9**(3): 334-343.
Kumar, R. and C. Engwerda (2014). "Vaccines to prevent leishmaniasis." Clin Transl Immunology **3**(3): e13.
Lachaud, L., N. Bourgeois, N. Kuk, C. Morelle, L. Crobu, G. Merlin, P. Bastien, M. Pages and Y. Sterkers (2014). "Constitutive mosaic aneuploidy is a unique genetic feature widespread in the *Leishmania* genus." Microbes Infect **16**(1): 61-66.
Laffitte, M. C., M. M. Genois, A. Mukherjee, D. Legare, J. Y. Masson and M. Ouellette (2014). "Formation of linear amplicons with inverted duplications in *Leishmania* requires the MRE11 nuclease." PLoS Genet **10**(12): e1004805.
Lanotte, G. and J. A. Rioux (1990). "[Cell fusion in *Leishmania* (Kinetoplastida, Trypanosomatidae)]." C R Acad Sci III **310**(7): 285-288.
Lawyer, P. G., Y. B. Mebrahtu, P. M. Ngumbi, P. Mwanyumba, J. Mbugua, G. Kiilu, D. Kipkoech, J. Nzovu and C. O. Anjili (1991). "*Phlebotomus guggisbergi* (Diptera: Psychodidae), a vector of *Leishmania tropica* in Kenya." Am J Trop Med Hyg **44**(3): 290-298.
Le Blancq, S. M. and W. Peters (1986). "*Leishmania* in the Old World: 2. Heterogeneity among *L. tropica* zymodemes." Trans R Soc Trop Med Hyg **80**(1): 113-119.

LeBowitz, J. H., H. Q. Smith, L. Rusche and S. M. Beverley (1993). "Coupling of poly(A) site selection and trans-splicing in *Leishmania*." Genes Dev **7**(6): 996-1007.
LeishGEN Consortium, Wellcome Trust Case Control Consortium, M. Fakiola, A. Strange, H. J. Cordell, E. N. Miller, M. Pirinen, Z. Su, A. Mishra, S. Mehrotra, G. R. Monteiro, G. Band, C. Bellenguez, S. Dronov, S. Edkins, C. Freeman, E. Giannoulatou, E. Gray, S. E. Hunt, H. G. Lacerda, C. Langford, R. Pearson, N. N. Pontes, M. Rai, S. P. Singh, L. Smith, O. Sousa, D. Vukcevic, E. Bramon, M. A. Brown, J. P. Casas, A. Corvin, A. Duncanson, J. Jankowski, H. S. Markus, C. G. Mathew, C. N. Palmer, R. Plomin, A. Rautanen, S. J. Sawcer, R. C. Trembath, A. C. Viswanathan, N. W. Wood, M. E. Wilson, P. Deloukas, L. Peltonen, F. Christiansen, C. Witt, S. M. Jeronimo, S. Sundar, C. C. Spencer, J. M. Blackwell and P. Donnelly (2013). "Common variants in the HLA-DRB1-HLA-DQA1 HLA class II region are associated with susceptibility to visceral leishmaniasis." Nat Genet **45**(2): 208-213.
Leprohon, P., D. Legare, I. Girard, B. Papadopoulou and M. Ouellette (2006). "Modulation of *Leishmania* ABC protein gene expression through life stages and among drug-resistant parasites." Eukaryot Cell **5**(10): 1713-1725.
Leprohon, P., D. Legare, F. Raymond, E. Madore, G. Hardiman, J. Corbeil and M. Ouellette (2009). "Gene expression modulation is associated with gene amplification, supernumerary chromosomes and chromosome loss in antimony-resistant *Leishmania infantum*." Nucleic Acids Res **37**(5): 1387-1399.
Li, C. H., H. Irmer, D. Gudjonsdottir-Planck, S. Freese, H. Salm, S. Haile, A. M. Estevez and C. Clayton (2006). "Roles of a *Trypanosoma brucei* 5'->3' exoribonuclease homolog in mRNA degradation." RNA **12**(12): 2171-2186.
Liang, X. H., A. Haritan, S. Uliel and S. Michaeli (2003). "trans and cis splicing in trypanosomatids: mechanism, factors, and regulation." Eukaryot Cell **2**(5): 830-840.
Lopez, M. A., E. A. Saada and K. L. Hill (2015). "Insect stage-specific adenylate cyclases regulate social motility in African trypanosomes." Eukaryot Cell **14**(1): 104-112.
Lynch, M. (2010). "Evolution of the mutation rate." Trends Genet **26**(8): 345-352.
Magill, A. J., M. Grogl, R. A. Gasser, Jr., W. Sun and C. N. Oster (1993). "Visceral infection caused by *Leishmania tropica* in veterans of Operation Desert Storm." N Engl J Med **328**(19): 1383-1387.
Mandal, G., S. Mandal, M. Sharma, K. S. Charret, B. Papadopoulou, H. Bhattacharjee and R. Mukhopadhyay (2015). "Species-specific antimonial sensitivity in *Leishmania* is driven by post-transcriptional regulation of AQP1." PLoS Negl Trop Dis **9**(2): e0003500.
Mannaert, A., T. Downing, H. Imamura and J. C. Dujardin (2012). "Adaptive mechanisms in pathogens: universal aneuploidy in *Leishmania*." Trends Parasitol **28**(9): 370-376.
Mansueto, P., G. Vitale, G. Di Lorenzo, G. B. Rini, S. Mansueto and E. Cillari (2007). "Immunopathology of leishmaniasis: an update." Int J Immunopathol Pharmacol **20**(3): 435-445.
Marovich, M. A., R. Lira, M. Shepard, G. H. Fuchs, R. Kruetzer, T. B. Nutman and F. A. Neva (2001). "Leishmaniasis recidivans recurrence after 43 years: a clinical and immunologic report after successful treatment." Clin Infect Dis **33**(7): 1076-1079.

Marquis, N., B. Gourbal, B. P. Rosen, R. Mukhopadhyay and M. Ouellette (2005). "Modulation in aquaglyceroporin AQP1 gene transcript levels in drug-resistant *Leishmania*." Mol Microbiol **57**(6): 1690-1699.
Martinez-Calvillo, S., D. Nguyen, K. Stuart and P. J. Myler (2004). "Transcription initiation and termination on *Leishmania major* chromosome 3." Eukaryot Cell **3**(2): 506-517.
Martinez-Calvillo, S., S. Yan, D. Nguyen, M. Fox, K. Stuart and P. J. Myler (2003). "Transcription of *Leishmania major* Friedlin chromosome 1 initiates in both directions within a single region." Mol Cell **11**(5): 1291-1299.
Matlashewski, G., B. Arana, A. Kroeger, A. Be-Nazir, D. Mondal, S. G. Nabi, M. R. Banjara, M. L. Das, B. Marasini, P. Das, G. Medley, A. Satoskar, H. Nakhasi, D. Argaw, J. Reeder and P. Olliaro (2014). "Research priorities for elimination of visceral leishmaniasis." Lancet Glob Health **2**(12): e683-684.
Matthews, K. R., C. Tschudi and E. Ullu (1994). "A common pyrimidine-rich motif governs trans-splicing and polyadenylation of tubulin polycistronic pre-mRNA in trypanosomes." Genes Dev **8**(4): 491-501.
McCarthy, D. J., Y. Chen and G. K. Smyth (2012). "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation." Nucleic Acids Res **40**(10): 4288-4297.
McConville, M. J., D. de Souza, E. Saunders, V. A. Likic and T. Naderer (2007). "Living in a phagolysosome; metabolism of *Leishmania* amastigotes." Trends Parasitol **23**(8): 368-375.
McNicoll, F., M. Muller, S. Cloutier, N. Boilard, A. Rochette, M. Dube and B. Papadopoulou (2005). "Distinct 3'-untranslated region elements regulate stage-specific mRNA accumulation and translation in *Leishmania*." J Biol Chem **280**(42): 35238-35246.
Mebrahtu, Y., P. Lawyer, J. Githure, J. B. Were, R. Muigai, L. Hendricks, J. Leeuwenburg, D. Koech and C. Roberts (1989). "Visceral leishmaniasis unresponsive to pentostam caused by *Leishmania tropica* in Kenya." Am J Trop Med Hyg **41**(3): 289-294.
Milone, J., J. Wilusz and V. Bellofatto (2002). "Identification of mRNA decapping activities and an ARE-regulated 3' to 5' exonuclease activity in trypanosome extracts." Nucleic Acids Res **30**(18): 4040-4050.
Monge-Maillo, B. and R. Lopez-Velez (2013). "Therapeutic options for old world cutaneous leishmaniasis and new world cutaneous and mucocutaneous leishmaniasis." Drugs **73**(17): 1889-1920.
Muller, M., P. K. Padmanabhan and B. Papadopoulou (2010). "Selective inactivation of SIDER2 retroposon-mediated mRNA decay contributes to stage- and species-specific gene expression in *Leishmania*." Mol Microbiol **77**(2): 471-491.
Muller, M., P. K. Padmanabhan, A. Rochette, D. Mukherjee, M. Smith, C. Dumas and B. Papadopoulou (2010). "Rapid decay of unstable *Leishmania* mRNAs bearing a conserved retroposon signature 3'-UTR motif is initiated by a site-specific endonucleolytic cleavage without prior deadenylation." Nucleic Acids Res **38**(17): 5867-5883.
Myler, P. J., L. Audleman, T. deVos, G. Hixson, P. Kiser, C. Lemley, C. Magness, E. Rickel, E. Sisk, S. Sunkin, S. Swartzell, T. Westlake, P. Bastien, G. Fu, A. Ivens and K.

Stuart (1999). "Leishmania major Friedlin chromosome 1 has an unusual distribution of protein-coding genes." Proc Natl Acad Sci U S A **96**(6): 2902-2906.
Nolder, D., N. Roncal, C. R. Davies, A. Llanos-Cuentas and M. A. Miles (2007). "Multiple hybrid genotypes of *Leishmania (Viannia)* in a focus of mucocutaneous leishmaniasis." Am J Trop Med Hyg **76**(3): 573-578.
O'Connell, J., D. Gurdasani, O. Delaneau, N. Pirastu, S. Ulivi, M. Cocca, M. Traglia, J. Huang, J. E. Huffman, I. Rudan, R. McQuillan, R. M. Fraser, H. Campbell, O. Polasek, G. Asiki, K. Ekoru, C. Hayward, A. F. Wright, V. Vitart, P. Navarro, J. F. Zagury, J. F. Wilson, D. Toniolo, P. Gasparini, N. Soranzo, M. S. Sandhu and J. Marchini (2014). "A general approach for haplotype phasing across the full spectrum of relatedness." PLoS Genet **10**(4): e1004234.
Obonaga, R., O. L. Fernandez, L. Valderrama, L. C. Rubiano, M. Castro Mdel, M. C. Barrera, M. A. Gomez and N. Gore Saravia (2014). "Treatment failure and miltefosine susceptibility in dermal leishmaniasis caused by *Leishmania* subgenus *Viannia* species." Antimicrob Agents Chemother **58**(1): 144-152.
Odiwuor, S., N. Veland, I. Maes, J. Arevalo, J. C. Dujardin and G. Van der Auwera (2012). "Evolution of the *Leishmania braziliensis* species complex from amplified fragment length polymorphisms, and clinical implications." Infect Genet Evol **12**(8): 1994-2002.
Otto, T. D., G. P. Dillon, W. S. Degrave and M. Berriman (2011). "RATT: Rapid Annotation Transfer Tool." Nucleic Acids Res **39**(9): e57.
Ouameur, A. A., I. Girard, D. Legare and M. Ouellette (2008). "Functional analysis and complex gene rearrangements of the folate/biopterin transporter (FBT) gene family in the protozoan parasite *Leishmania*." Mol Biochem Parasitol **162**(2): 155-164.
Paradis, E., J. Claude and K. Strimmer (2004). "APE: Analyses of Phylogenetics and Evolution in R language." Bioinformatics **20**(2): 289-290.
Parker, R. and H. Song (2004). "The enzymes and control of eukaryotic mRNA turnover." Nat Struct Mol Biol **11**(2): 121-127.
Peacock, C. S., K. Seeger, D. Harris, L. Murphy, J. C. Ruiz, M. A. Quail, N. Peters, E. Adlem, A. Tivey, M. Aslett, A. Kerhornou, A. Ivens, A. Fraser, M. A. Rajandream, T. Carver, H. Norbertczak, T. Chillingworth, Z. Hance, K. Jagels, S. Moule, D. Ormond, S. Rutter, R. Squares, S. Whitehead, E. Rabbinowitsch, C. Arrowsmith, B. White, S. Thurston, F. Bringaud, S. L. Baldauf, A. Faulconbridge, D. Jeffares, D. P. Depledge, S. O. Oyola, J. D. Hilley, L. O. Brito, L. R. Tosi, B. Barrell, A. K. Cruz, J. C. Mottram, D. F. Smith and M. Berriman (2007). "Comparative genomic analysis of three *Leishmania* species that cause diverse human disease." Nat Genet **39**(7): 839-847.
Peacock, L., M. Bailey, M. Carrington and W. Gibson (2014). "Meiosis and haploid gametes in the pathogen Trypanosoma brucei." Curr Biol **24**(2): 181-186.
Peacock, L., V. Ferris, M. Bailey and W. Gibson (2014). "Mating compatibility in the parasitic protist *Trypanosoma brucei*." Parasit Vectors **7**: 78.
Peacock, L., V. Ferris, R. Sharma, J. Sunter, M. Bailey, M. Carrington and W. Gibson (2011). "Identification of the meiotic life cycle stage of *Trypanosoma brucei* in the tsetse fly." Proc Natl Acad Sci U S A **108**(9): 3671-3676.
Perry, K. L., K. P. Watkins and N. Agabian (1987). "Trypanosome mRNAs have unusual "cap 4" structures acquired by addition of a spliced leader." Proc Natl Acad Sci U S A **84**(23): 8190-8194.

Peters, N. C., J. G. Egen, N. Secundino, A. Debrabant, N. Kimblin, S. Kamhawi, P. Lawyer, M. P. Fay, R. N. Germain and D. Sacks (2008). "In vivo imaging reveals an essential role for neutrophils in leishmaniasis transmitted by sand flies." Science **321**(5891): 970-974.
Pimenta, P. F., E. M. Saraiva, E. Rowton, G. B. Modi, L. A. Garraway, S. M. Beverley, S. J. Turco and D. L. Sacks (1994). "Evidence that the vectorial competence of phlebotomine sand flies for different species of *Leishmania* is controlled by structural polymorphisms in the surface lipophosphoglycan." Proc Natl Acad Sci U S A **91**(19): 9155-9159.
Plourde, M., A. Coelho, Y. Keynan, O. E. Larios, M. Ndao, A. Ruest, G. Roy, E. Rubinstein and M. Ouellette (2012). "Genetic polymorphisms and drug susceptibility in four isolates of *Leishmania tropica* obtained from Canadian soldiers returning from Afghanistan." PLoS Negl Trop Dis **6**(1): e1463.
Poxleitner, M. K., M. L. Carpenter, J. J. Mancuso, C. J. Wang, S. C. Dawson and W. Z. Cande (2008). "Evidence for karyogamy and exchange of genetic material in the binucleate intestinal parasite *Giardia intestinalis*." Science **319**(5869): 1530-1533.
Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly and P. C. Sham (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." Am J Hum Genet **81**(3): 559-575.
Ramirez, C. A., J. M. Requena and C. J. Puerta (2013). "Alpha tubulin genes from Leishmania braziliensis: genomic organization, gene structure and insights on their expression." BMC Genomics **14**: 454.
Ramirez, J. D. and M. S. Llewellyn (2014). "Reproductive clonality in protozoan pathogens--truth or artefact?" Mol Ecol **23**(17): 4195-4202.
Rastrojo, A., F. Carrasco-Ramiro, D. Martin, A. Crespillo, R. M. Reguera, B. Aguado and J. M. Requena (2013). "The transcriptome of *Leishmania major* in the axenic promastigote stage: transcript annotation and relative expression levels by RNA-seq." BMC Genomics **14**: 223.
Rausch, T., T. Zichner, A. Schlattl, A. M. Stutz, V. Benes and J. O. Korbel (2012). "DELLY: structural variant discovery by integrated paired-end and split-read analysis." Bioinformatics **28**(18): i333-i339.
Ravel, C., S. Cortes, F. Pratlong, F. Morio, J. P. Dedet and L. Campino (2006). "First report of genetic hybrids between two very divergent *Leishmania* species: *Leishmania infantum* and *Leishmania major*." Int J Parasitol **36**(13): 1383-1388.
Reithinger, R., J. C. Dujardin, H. Louzir, C. Pirmez, B. Alexander and S. Brooker (2007). "Cutaneous leishmaniasis." Lancet Infect Dis **7**(9): 581-596.
Reithinger, R., M. Mohsen, K. Aadil, M. Sidiqi, P. Erasmus and P. G. Coleman (2003). "Anthroponotic cutaneous leishmaniasis, Kabul, Afghanistan." Emerg Infect Dis **9**(6): 727-729.
Reynolds, D., L. Cliffe, K. U. Forstner, C. C. Hon, T. N. Siegel and R. Sabatini (2014). "Regulation of transcription termination by glucosylated hydroxymethyluracil, base J, in *Leishmania major* and *Trypanosoma brucei*." Nucleic Acids Res **42**(15): 9717-9729.

Rioux, J. A., G. Lanotte, E. Serres, F. Pratlong, P. Bastien and J. Perieres (1990). "Taxonomy of *Leishmania*. Use of isoenzymes. Suggestions for a new classification." Ann Parasitol Hum Comp **65**(3): 111-125.
Rochette, A., F. Raymond, J. Corbeil, M. Ouellette and B. Papadopoulou (2009). "Whole-genome comparative RNA expression profiling of axenic and intracellular amastigote forms of *Leishmania infantum*." Mol Biochem Parasitol **165**(1): 32-47.
Rochette, A., F. Raymond, J. M. Ubeda, M. Smith, N. Messier, S. Boisvert, P. Rigault, J. Corbeil, M. Ouellette and B. Papadopoulou (2008). "Genome-wide gene expression profiling analysis of *Leishmania major* and *Leishmania infantum* developmental stages reveals substantial differences between the two species." BMC Genomics **9**: 255.
Rogers, M. B., T. Downing, B. A. Smith, H. Imamura, M. Sanders, M. Svobodova, P. Volf, M. Berriman, J. A. Cotton and D. F. Smith (2014). "Genomic confirmation of hybridisation and recent inbreeding in a vector-isolated *Leishmania* population." PLoS Genet **10**(1): e1004092.
Rogers, M. B., J. D. Hilley, N. J. Dickens, J. Wilkes, P. A. Bates, D. P. Depledge, D. Harris, Y. Her, P. Herzyk, H. Imamura, T. D. Otto, M. Sanders, K. Seeger, J. C. Dujardin, M. Berriman, D. F. Smith, C. Hertz-Fowler and J. C. Mottram (2011). "Chromosome and gene copy number variation allow major structural change between species and strains of *Leishmania*." Genome Res **21**(12): 2129-2142.
Rogers, M. E. and P. A. Bates (2007). "*Leishmania* manipulation of sand fly feeding behavior results in enhanced transmission." PLoS Pathog **3**(6): e91.
Rogers, M. E., T. Ilg, A. V. Nikolaev, M. A. Ferguson and P. A. Bates (2004). "Transmission of cutaneous leishmaniasis by sand flies is enhanced by regurgitation of fPPG." Nature **430**(6998): 463-467.
Romano, A., E. Inbar, A. Debrabant, M. Charmoy, P. Lawyer, F. Ribeiro-Gomes, M. Barhoumi, M. Grigg, J. Shaik, D. Dobson, S. M. Beverley and D. L. Sacks (2014). "Cross-species genetic exchange between visceral and cutaneous strains of *Leishmania* in the sand fly vector." Proc Natl Acad Sci U S A **111**(47): 16808-16813.
Rougeron, V., A. L. Banuls, B. Carme, S. Simon, P. Couppie, M. Nacher, M. Hide and T. De Meeus (2011). "Reproductive strategies and population structure in *Leishmania*: substantial amount of sex in *Leishmania Viannia guyanensis*." Mol Ecol **20**(15): 3116-3127.
Rougeron, V., T. De Meeus and A. L. Banuls (2015). "A primer for *Leishmania* population genetic studies." Trends Parasitol **31**(2): 52-59.
Rougeron, V., T. De Meeus, M. Hide, G. Le Falher, B. Bucheton, J. Dereure, S. H. El-Safi, A. Dessein and A. L. Banuls (2011). "Multifaceted population structure and reproductive strategy in *Leishmania donovani* complex in one Sudanese village." PLoS Negl Trop Dis **5**(12): e1448.
Rougeron, V., T. De Meeus, M. Hide, E. Waleckx, H. Bermudez, J. Arevalo, A. Llanos-Cuentas, J. C. Dujardin, S. De Doncker, D. Le Ray, F. J. Ayala and A. L. Banuls (2009). "Extreme inbreeding in *Leishmania braziliensis*." Proc Natl Acad Sci U S A **106**(25): 10224-10229.
Rowland, M., A. Munir, N. Durrani, H. Noyes and H. Reyburn (1999). "An outbreak of cutaneous leishmaniasis in an Afghan refugee settlement in north-west Pakistan." Trans R Soc Trop Med Hyg **93**(2): 133-136.

Sacks, D. and C. Anderson (2004). "Re-examination of the immunosuppressive mechanisms mediating non-cure of *Leishmania* infection in mice." Immunol Rev **201**: 225-238.
Sacks, D. L. (1989). "Metacyclogenesis in *Leishmania* promastigotes." Exp Parasitol **69**(1): 100-103.
Sacks, D. L. (2001). "*Leishmania*-sand fly interactions controlling species-specific vector competence." Cell Microbiol **3**(4): 189-196.
Sacks, D. L., R. T. Kenney, R. D. Kreutzer, C. L. Jaffe, A. K. Gupta, M. C. Sharma, S. P. Sinha, F. A. Neva and R. Saran (1995). "Indian kala-azar caused by *Leishmania tropica*." Lancet **345**(8955): 959-961.
Sadlova, J., M. Yeo, V. Seblova, M. D. Lewis, I. Mauricio, P. Volf and M. A. Miles (2011). "Visualisation of *Leishmania donovani* fluorescent hybrids during early stage development in the sand fly vector." PLoS One **6**(5): e19851.
Sadlova, J., M. Yeo, V. Seblova, M. D. Lewis, I. Mauricio, P. Volf and M. A. Miles (2011). "Visualisation of *Leishmania donovani* fluorescent hybrids during early stage development in the sand fly vector." PLoS ONE **6**(5).
Sakthianandeswaren, A., S. J. Foote and E. Handman (2009). "The role of host genetics in leishmaniasis." Trends Parasitol **25**(8): 383-391.
Salmon, D., G. Vanwalleghem, Y. Morias, J. Denoeud, C. Krumbholz, F. Lhomme, S. Bachmaier, M. Kador, J. Gossmann, F. B. Dias, G. De Muylder, P. Uzureau, S. Magez, M. Moser, P. De Baetselier, J. Van Den Abbeele, A. Beschin, M. Boshart and E. Pays (2012). "Adenylate cyclases of *Trypanosoma brucei* inhibit the innate immune response of the host." Science **337**(6093): 463-466.
Sang, D. K., W. K. Njeru and R. W. Ashford (1994). "A zoonotic focus of cutaneous leishmaniasis due to *Leishmania tropica* at Utut, Rift Valley Province, Kenya." Trans R Soc Trop Med Hyg **88**(1): 35-37.
Saraiva, E. M., P. F. Pimenta, T. N. Brodin, E. Rowton, G. B. Modi and D. L. Sacks (1995). "Changes in lipophosphoglycan and gene expression associated with the development of *Leishmania major* in *Phlebotomus papatasi*." Parasitology **111 (Pt 3)**: 275-287.
Saroufim, M., K. Charafeddine, G. Issa, H. Khalifeh, R. H. Habib, A. Berry, N. Ghosn, A. Rady and I. Khalifeh (2014). "Ongoing epidemic of cutaneous leishmaniasis among Syrian refugees, Lebanon." Emerg Infect Dis **20**(10): 1712-1715.
Schnur, L. F., A. Nasereddin, C. L. Eisenberger, C. L. Jaffe, M. El Fari, K. Azmi, G. Anders, M. Killick-Kendrick, R. Killick-Kendrick, J. P. Dedet, F. Pratlong, M. Kanaan, T. Grossman, R. L. Jacobson, G. Schonian and A. Warburg (2004). "Multifarious characterization of *Leishmania tropica* from a Judean desert focus, exposing intraspecific diversity and incriminating phlebotomus sergenti as its vector." Am J Trop Med Hyg **70**(4): 364-372.
Schumer, M., R. Cui, G. G. Rosenthal and P. Andolfatto (2015). "Reproductive isolation of hybrid populations driven by genetic incompatibilities." PLoS Genet **11**(3): e1005041.
Schuster, F. L. and J. J. Sullivan (2002). "Cultivation of clinically significant hemoflagellates." Clin Microbiol Rev **15**(3): 374-389.

Schwede, A., L. Ellis, J. Luther, M. Carrington, G. Stoecklin and C. Clayton (2008). "A role for Caf1 in mRNA deadenylation and decay in trypanosomes and human cells." Nucleic Acids Res **36**(10): 3374-3388.
Schwede, A., T. Manful, B. A. Jha, C. Helbig, N. Bercovich, M. Stewart and C. Clayton (2009). "The role of deadenylation in the degradation of unstable mRNAs in trypanosomes." Nucleic Acids Res **37**(16): 5511-5528.
Schwenkenbecher, J. M., T. Wirth, L. F. Schnur, C. L. Jaffe, H. Schallig, A. Al-Jawabreh, O. Hamarsheh, K. Azmi, F. Pratlong and G. Schonian (2006). "Microsatellite analysis reveals genetic structure of *Leishmania tropica.*" Int J Parasitol **36**(2): 237-246.
Seridi, N., A. Amro, K. Kuhls, M. Belkaid, C. Zidane, A. Al-Jawabreh and G. Schonian (2008). "Genetic polymorphism of Algerian *Leishmania infantum* strains revealed by multilocus microsatellite analysis." Microbes Infect **10**(12-13): 1309-1315.
Shani-Adir, A., S. Kamil, D. Rozenman, E. Schwartz, M. Ramon, L. Zalman, A. Nasereddin, C. L. Jaffe and M. Ephros (2005). "*Leishmania tropica* in northern Israel: a clinical overview of an emerging focus." J Am Acad Dermatol **53**(5): 810-815.
Siegel, T. N., D. R. Hekstra, L. E. Kemp, L. M. Figueiredo, J. E. Lowell, D. Fenyo, X. Wang, S. Dewell and G. A. Cross (2009). "Four histone variants mark the boundaries of polycistronic transcription units in *Trypanosoma brucei*." Genes Dev **23**(9): 1063-1076.
Singh, O. P., E. Hasker, D. Sacks, M. Boelaert and S. Sundar (2014). "Asymptomatic Leishmania infection: a new challenge for *Leishmania* control." Clin Infect Dis **58**(10): 1424-1429.
Smircich, P., D. Forteza, N. M. El-Sayed and B. Garat (2013). "Genomic analysis of sequence-dependent DNA curvature in *Leishmania*." PLoS One **8**(4): e63068.
Soares, R. P., T. Barron, K. McCoy-Simandle, M. Svobodova, A. Warburg and S. J. Turco (2004). "*Leishmania tropica*: intraspecific polymorphisms in lipophosphoglycan correlate with transmission by different Phlebotomus species." Exp Parasitol **107**(1-2): 105-114.
Spath, G. F., S. Drini and N. Rachidi (2015). "A touch of Zen: post-translational regulation of the *Leishmania* stress response." Cell Microbiol **17**(5): 632-638.
Stamper, L. W., R. L. Patrick, M. P. Fay, P. G. Lawyer, D. E. Elnaiem, N. Secundino, A. Debrabant, D. L. Sacks and N. C. Peters (2011). "Infection parameters in the sand fly vector that predict transmission of *Leishmania major*." PLoS Negl Trop Dis **5**(8): e1288.
Sterkers, Y., L. Crobu, L. Lachaud, M. Pages and P. Bastien (2014). "Parasexuality and mosaic aneuploidy in *Leishmania*: alternative genetics." Trends Parasitol **30**(9): 429-435.
Sterkers, Y., L. Lachaud, L. Crobu, P. Bastien and M. Pages (2011). "FISH analysis reveals aneuploidy and continual generation of chromosomal mosaicism in *Leishmania major*." Cell Microbiol **13**(2): 274-283.
Sturm, N. R. and D. A. Campbell (1999). "The role of intron structures in trans-splicing and cap 4 formation for the *Leishmania* spliced leader RNA." J Biol Chem **274**(27): 19361-19367.
Sturm, N. R., J. Fleischmann and D. A. Campbell (1998). "Efficient trans-splicing of mutated spliced leader exons in *Leishmania tarentolae*." J Biol Chem **273**(30): 18689-18692.

Sturm, N. R., M. C. Yu and D. A. Campbell (1999). "Transcription termination and 3'-End processing of the spliced leader RNA in kinetoplastids." Mol Cell Biol **19**(2): 1595-1604.
Sutton, R. E. and J. C. Boothroyd (1986). "Evidence for trans splicing in trypanosomes." Cell **47**(4): 527-535.
Svobodova, M., P. Volf and J. Votypka (2006). "Experimental transmission of *Leishmania tropica* to hyraxes (*Procavia capensis*) by the bite of *Phlebotomus arabicus*." Microbes Infect **8**(7): 1691-1694.
Svobodova, M., J. Votypka, J. Peckova, V. Dvorak, A. Nasereddin, G. Baneth, J. Sztern, V. Kravchenko, A. Orr, D. Meir, L. F. Schnur, P. Volf and A. Warburg (2006). "Distinct transmission cycles of *Leishmania tropica* in 2 adjacent foci, Northern Israel." Emerg Infect Dis **12**(12): 1860-1868.
Swain, M. T., I. J. Tsai, S. A. Assefa, C. Newbold, M. Berriman and T. D. Otto (2012). "A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs." Nat Protoc **7**(7): 1260-1284.
Talmi-Frank, D., C. L. Jaffe, A. Nasereddin, A. Warburg, R. King, M. Svobodova, O. Peleg and G. Baneth (2010). "*Leishmania tropica* in rock hyraxes (*Procavia capensis*) in a focus of human cutaneous leishmaniasis." Am J Trop Med Hyg **82**(5): 814-818.
Talmi-Frank, D., N. Kedem-Vaanunu, R. King, G. K. Bar-Gal, N. Edery, C. L. Jaffe and G. Baneth (2010). "*Leishmania tropica* infection in golden jackals and red foxes, Israel." Emerg Infect Dis **16**(12): 1973-1975.
Tayeh, A., L. Jalouk and S. Cairncross (1997). "Twenty years of cutaneous leishmaniasis in Aleppo, Syria." Trans R Soc Trop Med Hyg **91**(6): 657-659.
Thomas, S., A. Green, N. R. Sturm, D. A. Campbell and P. J. Myler (2009). "Histone acetylations mark origins of polycistronic transcription in *Leishmania major*." BMC Genomics **10**: 152.
Tibayrenc, M. and F. J. Ayala (2002). "The clonal theory of parasitic protozoa: 12 years on." Trends Parasitol **18**(9): 405-410.
Tibayrenc, M. and F. J. Ayala (2012). "Reproductive clonality of pathogens: a perspective on pathogenic viruses, bacteria, fungi, and parasitic protozoa." Proc Natl Acad Sci U S A **109**(48): E3305-3313.
Tibayrenc, M., F. Kjellberg and F. J. Ayala (1990). "A clonal theory of parasitic protozoa: the population structures of *Entamoeba*, *Giardia*, *Leishmania*, *Naegleria*, *Plasmodium*, *Trichomonas*, and *Trypanosoma* and their medical and taxonomical consequences." Proc Natl Acad Sci U S A **87**(7): 2414-2418.
Tosato, V., L. Ciarloni, A. C. Ivens, M. A. Rajandream, B. G. Barrell and C. V. Bruschi (2001). "Secondary DNA structure analysis of the coding strand switch regions of five *Leishmania major* Friedlin chromosomes." Curr Genet **40**(3): 186-194.
Trapnell, C., L. Pachter and S. L. Salzberg (2009). "TopHat: discovering splice junctions with RNA-Seq." Bioinformatics **25**(9): 1105-1111.
Ubeda, J. M., D. Legare, F. Raymond, A. A. Ouameur, S. Boisvert, P. Rigault, J. Corbeil, M. J. Tremblay, M. Olivier, B. Papadopoulou and M. Ouellette (2008). "Modulation of gene expression in drug resistant *Leishmania* is associated with gene amplification, gene deletion and chromosome aneuploidy." Genome Biol **9**(7): R115.
Ubeda, J. M., F. Raymond, A. Mukherjee, M. Plourde, H. Gingras, G. Roy, A. Lapointe, P. Leprohon, B. Papadopoulou, J. Corbeil and M. Ouellette (2014). "Genome-wide

stochastic adaptive DNA amplification at direct and inverted DNA repeats in the parasite *Leishmania*." PLoS Biol **12**(5): e1001868.
van Luenen, H. G., C. Farris, S. Jan, P. A. Genest, P. Tripathi, A. Velds, R. M. Kerkhoven, M. Nieuwland, A. Haydock, G. Ramasamy, S. Vainio, T. Heidebrecht, A. Perrakis, L. Pagie, B. van Steensel, P. J. Myler and P. Borst (2012). "Glucosylated hydroxymethyluracil, DNA base J, prevents transcriptional readthrough in *Leishmania*." Cell **150**(5): 909-921.
van Zandbergen, G., M. Klinger, A. Mueller, S. Dannenberg, A. Gebert, W. Solbach and T. Laskay (2004). "Cutting edge: neutrophil granulocyte serves as a vector for *Leishmania* entry into macrophages." J Immunol **173**(11): 6521-6525.
Walters, L. L., K. P. Irons, G. Chaplin and R. B. Tesh (1993). "Life cycle of *Leishmania major* (Kinetoplastida: Trypanosomatidae) in the neotropical sand fly *Lutzomyia longipalpis* (Diptera: Psychodidae)." J Med Entomol **30**(4): 699-718.
Wang, W., W. Wang, W. Sun, J. J. Crowley and J. P. Szatkiewicz (2015). "Allele-specific copy-number discovery from whole-genome and whole-exome sequencing." Nucleic Acids Res **43**(14): e90.
Weiss, F., N. Vogenthaler, C. Franco-Paredes and S. R. Parker (2009). "*Leishmania tropica*-induced cutaneous and presumptive concomitant viscerotropic leishmaniasis with prolonged incubation." Arch Dermatol **145**(9): 1023-1026.
Wheeler, R. J., E. Gluenz and K. Gull (2011). "The cell cycle of *Leishmania*: morphogenetic events and their implications for parasite biology." Mol Microbiol **79**(3): 647-662.
WHO (2007). Report of the Fifth Consultative Meeting on *Leishmania*/HIV coinfection. WHO Reports. Addis Ababa, Ethiopia, World Health Organization.
WHO (2015). WHO Global Health Observatory.
Working Group on Research Priorities for Development of Leishmaniasis, V., C. H. Costa, N. C. Peters, S. R. Maruyama, E. C. de Brito, Jr. and I. K. Santos (2011). "Vaccines for the leishmaniases: proposals for a research agenda." PLoS Negl Trop Dis **5**(3): e943.
Xie, C. and M. T. Tammi (2009). "CNV-seq, a new method to detect copy number variation using high-throughput sequencing." BMC Bioinformatics **10**: 80.
Xu, M. and X. He (2011). "Genetic incompatibility dampens hybrid fertility more than hybrid viability: yeast as a case study." PLoS One **6**(4): e18341.

APPENDICES

Appendix A. European Nucleotide Archive (ENA) accession numbers for all samples sequenced for this dissertation. For RNA-seq samples, the Array Express accession number is E-ERAD-408. The ENA accession numbers for all RNA-seq triplicate samples were from ERS763603 to ERS763656.

Sample	ENA Accession Number
Melloy	ERS364123
LRC-L747	ERS364124
KK27	ERS364125
Ackerman	ERS364126
K26_1	ERS364127
LRC-L810	ERS364128
MN-11	ERS364129
Rupert	ERS364130
R55	ERS364131
Kubba	ERS364132
Boone	ERS364133
Azad	ERS364134
E50	ERS364135
MA-37	ERS364136
K112_1	ERS364137
Sultan	ERS364138
Melloy	ERS364123
LRC-L747	ERS364124
KK27	ERS364125
Ackerman	ERS364126
K26_1	ERS364127
LRC-L810	ERS364128
SAF-K27	ERS198200
LRC-L838	ERS198235
Adhanis I	ERS214826
L75	ERS198220
SAF-K27	ERS198200
LRC-L838	ERS198235
Rupert C1	ERS471347
Rupert C2	ERS471346
MA-37 C1	ERS763658
MN-11 C1	ERS471344
MN-11 C2	ERS471343
Kubba C1	ERS471342
L590	ERS218438
L747 HYG	ERS763657
MA-37 NEO	ERS763658
H1a	ERS763659
H1b	ERS763664
H2a	ERS763665
H2b	ERS763660
H3a	ERS763666
H3b	ERS763667
H4a	ERS763661
H4b	ERS763668
H5a	ERS763669
H5b	ERS763662

Appendix B. Inbreeding coefficients (FIT) and clustering results for the 34 isolates analysed in Chapter 2. Isolates are ranked from highest to lowest inbreeding coefficient. The inbreeding coefficient goes from 0 to 1 and increases with decreasing heterozygosity compared to HW expectations.

Sample	F_{IT}	Cluster
LA28	0.7996	Cluster 1
L747	0.7361	Cluster 2
Killicki	0.7268	Cluster 3
MN-11	0.6819	Cluster 1
L810	0.6071	Cluster 3
Tropica63	0.5603	Cluster 2
MA-37	0.5529	Cluster 1
Leep0920	0.5424	Cluster 3
Tropica75	0.5420	Cluster 2
Tropica57	0.5365	Cluster 2
AM	0.5031	Cluster 3
DBKM	0.4571	Cluster 1
CJ	0.3629	Cluster 3
E50	0.3096	Cluster 2
Azad	0.2406	Cluster 1
Boone	0.1625	Cluster 1
Gabai	0.1439	Cluster 1
Melloy	0.1340	Cluster 1
IIIKK	0.1094	Cluster 1
MON497px3	0.1078	Cluster 1
Bum30	0.1064	Cluster 1
WR683	0.1063	Cluster 1
K27	0.1040	Cluster 1
Ackerman	0.1010	Cluster 1
Adhanis I	0.1005	Cluster 1
Kubba	0.09649	Cluster 1
188	0.09235	Cluster 1
Asinai III	0.09151	Cluster 1
Rachnan	0.09119	Cluster 1
L75	0.08906	Cluster 1
311W	0.08197	Cluster 1
Bag17	0.07922	Cluster 1
Rupert	0.07415	Cluster 1
Bag9	0.06428	Cluster 1

Appendix C. Most highly expressed genes in the promastigote stage from RNA-seq of the isolates analysed in Chapter 3. Mean VSD is the mean variance-stabilized dispersion across the set of isolates. Genes are ranked from most to least expressed.

Gene	Product	Mean VSD
LmjF.21.1860	Beta tubulin	16.41689
LmjF.36.1940	Inosine/guanosine transporter (NT2)	15.23829
LmjF.15.1230	Nucleoside transporter 1, putative	14.59714
LmjF.15.1240	Nucleoside transporter 1, putative	14.45777
LmjF.35.1890	60S ribosomal protein L5	14.36469
LmjF.32.2660	L-lysine transport protein (AAT16)	13.86768
LmjF.21.0915	Histone H2A	13.82777
LmjF.25.0720	Eukaryotic initiation factor 5a (EIF5A1)	13.82076
LmjF.11.0970	40S ribosomal protein S5	13.76264
LmjF.32.3130	Ribosomal protein L3	13.7539
LmjF.13.0450	AlbA-like nucleic acid binding protein	13.64462
LmjF.31.0350	Aminoacid transporter (AAT1.4)	13.42285
LmjF.36.6300	Glucose transporter 1 (GT1)	13.52832
LmjF.22.0420	40S ribosomal protein S15	13.58769
LmjF.32.3900	60S ribosomal protein L2/L8	13.52428
LmjF.35.0240	60S ribosomal protein L30	13.53991
LmjF.32.0750	RNP1-like RNA binding protein	13.52044
LmjF.06.1260	Pteridine transporter, putative	13.40097
LmjF.36.0980	40S ribosomal protein S10	13.46978
LmjF.14.1160	Enolase	13.44339
LmjF.24.2070	40S ribosomal protein S8	13.42711
LmjF.27.1380	60S acidic ribosomal protein P0	13.37493
LmjF.05.0500	ATPase alpha subunit	13.30601
LmjF.25.0910	Cyclophilin a (CYPA)	13.24228
LmjF.13.0570	40S ribosomal protein S12	13.21298
LmjF.28.2555	40S ribosomal protein S17	13.21329
LmjF.06.0570	60S ribosomal protein L23a	13.18204
LmjF.27.1245	Carboxypeptidase-like protein	13.15227
LmjF.36.1260	Fructose-1,6-biphosphate aldolase (ALD)	13.09008
LmjF.07.1160	Aminoacid transporter (AAT19)	12.75687

Appendix D. List of DE genes in L810 compared to all other isolates (excluding MN-11 C2 and Boone). Only genes that had significant p-values and that had a log-fold change greater than 2 or less than -2 were considered.

Gene	Log-FC	P-value	Product
LmjF.02.0400	2.081753	2.88E-10	Transmembrane hypothetical protein, unknown function
LmjF.04.0310	4.434332	0.00E+00	beta-fructofuranosidase, putative
LmjF.04.0320	2.754353	2.36E-175	beta-fructofuranosidase, putative
LmjF.04.1050	10.20729	0.00E+00	acyltransferase-like protein, copy 2
LmjF.09.0003	6.314717	1.00E+00	hypothetical protein, conserved
LmjF.09.0690	2.939361	3.84E-20	hypothetical protein, conserved
LmjF.17.0086	5.799281	7.53E-186	elongation factor 1-alpha
LmjF.17.0730	3.260571	3.68E-87	transmembrane hypothetical protein, conserved
LmjF.17.0733	4.468265	2.24E-05	hypothetical protein, conserved
LmjF.17.0740	2.220649	3.42E-129	Transmembrane hypothetical protein, conserve
LmjF.20.0830	2.340252	3.81E-115	phosphopantetheinyl transferase-like protein
LmjF.20.1175	5.285097	1.77E-21	hypothetical protein
LmjF.21.0015	2.434106	9.56E-35	histone H4
LmjF.26.0640	2.126129	1.79E-109	10 kDa heat shock protein, putative
LmjF.28.0350	2.237566	8.52E-47	Transmembrane hypothetical protein, unknown function
LmjF.30.2460	2.768505	7.25E-132	heat shock 70-related protein 1, mitochondrial precursor, putative
LmjF.30.2480	2.23119	8.076193e-316	heat shock 70-related protein 1, mitochondrial precursor, putative
LmjF.30.2490	3.514708	7.40E-121	heat shock 70-related protein 1, mitochondrial precursor, putative
LmjF.31.0420	2.987786	8.51E-99	Transmembrane hypothetical protein, unknown function
LmjF.31.0470	2.598772	6.07E-72	hypothetical protein, conserved
LmjF.31.2680	3.515456	1.00E+00	hypothetical protein, unknown function
LmjF.32.1380	4.66376	5.53E-111	hypothetical protein, conserved
LmjF.32.1390	2.155002	4.45E-274	hypothetical protein, conserved
LmjF.33.1630	2.074137	3.27E-211	cyclophilin 4, putative
LmjF.35.0450	10.170533	3.96E-14	hypothetical protein, unknown function
LmjF.35.2600	2.589783	5.11E-14	hypothetical protein, unknown function
LmjF.03.0010	-2.06585	8.30E-208	hypothetical protein
LmjF.04.1210	-2.220235	9.88E-114	casein kinase I, putative
LmjF.05.0360	-2.059694	4.23E-03	ATP-dependent RNA helicase, putative
LmjF.13.0100	-2.635757	1.43E-220	excreted hypothetical protein, unknown function
LmjF.16.0480	-2.353516	0.00E+00	fucose kinase, putative
LmjF.17.0190	-2.709674	6.31E-226	receptor-type adenylate cyclase, putative
LmjF.22.0750	-2.252003	1.26E-98	CCCH-type zinc finger hypothetical protein, conserved
LmjF.23.0730	-2.481566	1.17E-68	RNA-binding protein, putative
LmjF.23.1665	-2.321696	0.00E+00	Transmembrane hypothetical phosphatidic acid phosphatase type-2
LmjF.25.1730	-2.018702	0.00E+00	signal peptide-containing hypothetical protein, conserved
LmjF.26.2680	-2.075803	5.60E-182	Excreted hypothetical protein, unknown function
LmjF.28.1570	-2.253956	0.00E+00	hydrolase, alpha/beta fold family, putative
LmjF.28.2660	-2.437439	5.97E-245	hypothetical methionine sulphoxide reductase B protein, conserved
LmjF.31.2100	-2.711269	7.20E-223	hypothetical protein, unknown function
LmjF.31.2460	-2.050621	1.09E-116	lipase, putative
LmjF.33.0520	-2.486323	0.00E+00	d-xylulose reductase, putative
LmjF.35.0710	-2.056016	2.44E-186	Nuclease-like hypothetical protein, conserved
LmjF.35.2810	-2.036244	5.89E-298	MFS-1 transporter, putative
LmjF.35.4240	-2.17067	2.01E-76	hypothetical protein, conserved
LmjF.35.5030	-2.534736	0.00E+00	hypothetical protein, conserved
LmjF.36.2760	-2.464428	2.39E-168	MFS general substrate transporter protein

Appendix E. List of DE genes between the two clones of isolate Rupert that were compared in Chapter 3 that also overlap with CNVs identified in this study. LogFC is the log-fold change in expression between the two clones, representing the difference between Rupert C2 and Rupert C1. Genes with positive logFC are more highly expressed in Rupert C2 than Rupert C1. Log-2-ratio is the ratio between the average read depth for that gene in Rupert C2 and Rupert C1. Positive Log-2-ratio means that a CNV has either amplified the gene in Rupert C2 or deleted a copy of the gene in Rupert C1.

Gene	Product	Log-FC	CNV region	Log-2-ratio
LmjF.06.0080	ATP-binding cassette protein subfamily G, member 1	0.5524548	CNVR_113	-0.920944
LmjF.06.0090	ATP-binding cassette protein subfamily G, member 2	0.2920734	CNVR_114	-0.9580283
LmjF.14.1100	kinesin K39	-0.2676972	CNVR_110	0.7792663
LmjF.15.1480	cAMP specific phosphodiesterase	0.4970286	CNVR_37	-0.9574448
LmjF.17.0190	receptor-type adenylate cyclase, putative	0.6477952	CNVR_8	-2.746126
LmjF.22.0600	acetyltransferase (GNAT) domain containing protein	-0.63869	CNVR_61	2.383092
LmjF.22.1390	hypothetical protein, conserved	0.3337642	CNVR_71	-0.774964
LmjF.23.0210	ATP-binding cassette protein subfamily C, member 1	1.3873056	CNVR_53	1.786927
LmjF.23.0223	hypothetical protein, conserved	1.1552017	CNVR_53	1.786927
LmjF.23.0225	hypothetical protein, conserved	0.8470822	CNVR_53	1.786927
LmjF.23.0230	hypothetical protein, conserved	0.8052322	CNVR_53	1.786927
LmjF.23.0240	terbinafine resistance locus protein (yip1)	1.486045	CNVR_53	1.786927
LmjF.23.0250	multidrug resistance protein A	1.2144793	CNVR_53	1.786927
LmjF.23.0260	argininosuccinate synthase, putative	1.0082369	CNVR_53	1.786927
LmjF.23.0270	pteridine reductase 1	1.313278	CNVR_53	1.786927
LmjF.23.0280	zinc finger, C3HC4 type	1.539302	CNVR_53	1.786927
LmjF.23.0290	checkpoint protein HUS1	0.451299	CNVR_53	1.786927
LmjF.23.0300	tryptophanyl-tRNA synthetase	0.8019783	CNVR_53	1.786927
LmjF.23.0310	hypothetical protein, conserved	0.6507995	CNVR_53	1.786927
LmjF.23.0320	hypothetical protein, conserved	-0.5984561	CNVR_53	1.786927
LmjF.23.0340	(H+)-ATPase G subunit, putative	-2.2208945	CNVR_53	1.786927
LmjF.23.0220	ATP-binding cassette protein subfamily C, member 2	1.3804016	CNVR_54	-1.347129
LmjF.24.1010	Mnd1 family, putative	2.2562565	CNVR_10	2.126852
LmjF.24.1020	F-box domain/Galactose oxidase containing protein	2.5872305	CNVR_10	2.126852
LmjF.24.1030	dynein light chain	2.581097	CNVR_10	2.126852
LmjF.24.1040	hypothetical protein, unknown function	2.2095486	CNVR_10	2.126852
LmjF.24.1050	SNF2 family protein	2.4974836	CNVR_10	2.126852
LmjF.24.1060	AKAP7 2'5' RNA ligase-like domain containing protein	3.2916332	CNVR_10	2.126852
LmjF.24.1070	hypothetical protein, conserved	2.3595209	CNVR_10	2.126852
LmjF.24.1080	DNAJ domain protein, putative	-0.3976447	CNVR_10	2.126852
LmjF.24.1100	pre-mRNA-splicing factor ATP-dependent RNA helicase	-0.8934357	CNVR_10	2.126852
LmjF.24.1125	hypothetical protein, conserved	-0.2420312	CNVR_14	-0.9032868
LmjF.25.1443	hypothetical protein, conserved	0.264172	CNVR_44	-1.198483
LmjF.27.2340	sucrose hydrolase-like protein	-0.7033248	CNVR_16	-1.525739
LmjF.27.2350	vesicle-associated membrane protein	-0.9150209	CNVR_16	-1.525739
LmjF.27.2360	hypothetical protein, conserved	-0.6365974	CNVR_16	-1.525739
LmjF.27.2370	hypothetical protein, conserved	-0.4573555	CNVR_16	-1.525739

LmjF.27.2380	hypothetical protein, unknown function	-0.6501815	CNVR_16	-1.525739
LmjF.27.2390	TPR-repeat protein, putative	-0.5023864	CNVR_16	-1.525739
LmjF.27.2400	heat shock protein DNAJ, putative	-0.7452128	CNVR_16	-1.525739
LmjF.27.2410	aldo-keto reductase-like protein	-0.3671546	CNVR_17	-0.9651632
LmjF.27.2420	hypothetical protein, conserved	-0.2953249	CNVR_17	-0.9651632
LmjF.27.2430	hypothetical protein, conserved	-0.7725951	CNVR_17	-0.9651632
LmjF.27.2440	3-oxoacyl-ACP reductase, putative	-0.856198	CNVR_19	-0.9582698
LmjF.27.2450	hypothetical protein, conserved	-0.3068082	CNVR_19	-0.9582698
LmjF.27.2460	hypothetical protein, conserved	-0.6411425	CNVR_20	-0.8594242
LmjF.27.2590	dynein heavy chain	0.3560877	CNVR_26	-0.8456079
LmjF.28.1540	hypothetical protein, unknown function	0.3502462	CNVR_141	-1.515351
LmjF.33.2020	hypothetical protein, unknown function	-0.3044996	CNVR_127	-1.244153
LmjF.33.2760	hypothetical protein, conserved	-0.4947443	CNVR_133	-0.7255231
LmjF.33.2955	hypothetical protein	-0.4253012	CNVR_134	-0.8459617
LmjF.34.1520	p25-alpha, putative	-0.3007447	CNVR_1	1.194699
LmjF.34.1530	p25-alpha, putative	0.7736714	CNVR_1	1.194699
LmjF.35.5170	hypothetical protein, conserved	2.7907753	CNVR_85	0.7553396
LmjF.35.5180	hypothetical protein, conserved	1.0829859	CNVR_86	0.641145
LmjF.35.5190	NIMA-related kinase	1.4470836	CNVR_87	0.7353458
LmjF.35.5240	hypothetical protein, unknown function	-0.2359127	CNVR_93	0.7487856
LmjF.35.5290	Lsm12 protein, putative	0.6631023	CNVR_96	0.7330686
LmjF.35.5300	hypothetical protein, conserved	0.8586335	CNVR_96	0.7330686
LmjF.35.5320	SAC3/GANP/Nin1/mts3/eIF-3 p25 family protein, putative	0.4844933	CNVR_97	0.6892603
LmjF.35.5330	isopentenyl-diphosphate delta-isomerase (type II)	0.7756338	CNVR_98	0.6620098
LmjF.35.5350	AAT27.1, amino acid permease	0.9528822	CNVR_99	0.6878376
LmjF.36.1520	NIMA-related protein kinase	-0.7818998	CNVR_46	-0.9723839