

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

Acknowledgements	p. 4
Summary	p. 7
Table of Contents	p. 9
List of Figures	p. 12
List of Tables	p. 16
Abbreviations	p. 18
Chapter 1: Introduction	
1.1. Leishmaniasis, a complex parasitic disease	p. 19
1.1.1. Overview	p. 19
1.1.2. The biology of the parasite	p. 26
1.1.3. The burden of disease due to <i>L. tropica</i>	p. 29
1.1.4. Mechanisms of pathogenesis	p. 34
1.2. Alternative genetics in <i>Leishmania</i>	
1.2.1. The unique genome of kinetoplastids	p. 36
1.2.2. Karyotypic variation in <i>Leishmania</i>	p. 39
1.2.3. Transcriptional regulation (or lack thereof)	p. 41
1.3. A clonal, sexual, or parasexual organism?	
1.3.1. The clonal theory	p. 44
1.3.2. Challenges to the clonal theory	p. 46
1.3.3. Models of asexuality, sexuality, and parasexuality	p. 50
1.4. Aims and Objectives	p. 54

Chapter 2: Population genetics in *L. tropica*

- 2.1. Introduction p. 56
- 2.2. Methods p. 59
- 2.3. Results p. 70
- 2.4. Discussion p. 89

Chapter 3: Genome plasticity and gene expression

- 3.1. Introduction p. 95
- 3.2. Methods p. 100
- 3.3. Results p. 107
- 3.4. Discussion p. 132

Chapter 4: Experimental crosses of *L. tropica*

- 4.1. Introduction p. 140
- 4.2. Methods p. 147
- 4.3. Results p. 153
- 4.4. Discussion p. 163

Chapter 5: Genetic exchange in experimental hybrids

- 5.1. Introduction p. 167
- 5.2. Methods p. 172
- 5.3. Results p. 177
- 5.4. Discussion p. 191

Chapter 6: Conclusions

- 6.1. Population genetics in *L. tropica* p. 196
 - 6.1.1. Heterozygosity and reproduction p. 196

6.1.2.	Wahlund effects and reproduction	p. 198
6.1.3.	Population genetics and models of reproduction	p. 201
6.2.	Genome plasticity in <i>L. tropica</i>	p. 202
6.2.1.	Variation in copy number and effects on transcription	p. 202
6.2.2.	Variation in gene copy number and effects on transcription	p. 205
6.2.3.	Genome structure and models of reproduction	p. 208
6.3.	Genetic exchange in <i>L. tropica</i>	p. 209
6.3.1.	Sand fly infections and hybridization	p. 209
6.3.2.	Genomic consequences of hybridization	p. 210
6.3.3.	Hybridization and models of reproduction	p. 212
6.4.	Future directions	p. 213
	Appendices	p. 217
	References	p. 224

LIST OF FIGURES

Figure 1.1. Distribution and endemicity of visceral leishmaniasis (VL) according to 2013 annual country reports (source: WHO Global Health Observatory). (p.22)

Figure 1.2. Distribution and endemicity of cutaneous leishmaniasis (CL) according to 2013 annual country reports (source: WHO Global Health Observatory). (p. 23)

Figure 1.3. The *Leishmania* life cycle (source: CDC). (p.27)

Figure 1.4. Geographic distribution of Old World CL due to *L. tropica*, *L. aethiopica*, and related species. (p. 32)

Figure 1.5. Schematic models of asexual, sexual, and parasexual reproduction in *Leishmania* as referred to throughout this dissertation. (p. 52)

Figure 2.1. Example of a NJ gene tree used in the assignment of individual parental alleles to heterozygous genotypes for allelic plot reconstruction. (p. 66)

Figure 2.2. Aligned sequences for the 31_AQP1 marker for four of the isolates in the sample set used in this study. (p. 71)

Figure 2.3. Allelic plot of 34 isolates of *L. tropica* at 17 nuclear and kDNA markers. (p. 75)

Figure 2.4. NJ tree from concatenated sequence data of markers with complete sequence information for all isolates. (p. 78)

Figure 2.5. Histogram of inbreeding coefficients (F_{IT}) for the 34 isolates in this study. (p. 79)

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. (p. 80)

Figure 2.7. Unrooted, ultrametric NJ tree of 34 isolates of *L. tropica* based on the concatenated sequence data. (p. 81)

Figure 2.8. Average allele frequencies for each SNP across 18 individuals typed by WGS. (p. 84)

Figure 2.9. Allelic plot for all 306 596 SNPs that passed the filtering thresholds as described in the text, for all 18 strains. (p. 85)

Figure 2.10. PCA plot of the 18 isolates that were typed by WGS. (p. 86)

Figure 2.11. NJ tree for all 18 isolates based on the same set of high quality biallelic SNPs as in Figure 2.9. (p. 87)

Figure 3.1. Circular plots representing ploidy and long runs of homzygosity in 20 uncloned and cloned parasite lines. (p. 109)

Figure 3.2. PCA plots of the 14 isolates and 6 clones considered in this analysis. (p. 111)

Figure 3.3. MDS plot representing expression data for all strains, except for isolates MN-11_C2, Boone, E50, and K112. (p. 115)

Figure 3.4. Heatmap of Euclidean distances between variance stabilized expression values for each pair of samples. (p. 116)

Figure 3.5. Heatmap showing the most highly expressed genes in the set of isolates considered in this study. (p. 118)

Figure 3.6. Heatmap of log-fold changes in expression for the top 30 most significant DE genes. (p. 120)

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. (p. 124)

Figure 3.8. Gene dosage effects due to aneuploidy on transcription in *L. tropica* clonal lines Rupert C1 and Rupert C2. (p. 127)

Figure 3.9. Gene dosage effects due to copy number variation on transcription in *L. tropica* clonal lines Rupert C1 and Rupert C2. (p. 128)

Figure 3.10. Evidence for allele-specific gene expression in the two clones Rupert C1 and Rupert C2. (p. 131)

Figure 4.1. Targeting vectors used in this study for integration of different drug resistance markers in the *L. tropica* genome. (p. 150)

Figure 4.2. Schematic representation of the screening procedures for recovery of double-drug resistant hybrids in sand fly co-infections. (p. 152)

Figure 4.3. Infection loads of 14 *L. tropica* isolates in *L. longipalpis* LLJB sand flies at day 2 and day 8 post-infection. (p. 155)

Figure 4.4. Infection loads and development stages of *L. tropica* isolates Rupert, Kubba, and E50 in *L. longipalpis* LLJB and *P. arabicus* PAIS sand flies. (p. 156)

Figure 4.5. Growth phenotype of the transgenic line Kubba SAT in *P. arabicus* PAIS and *L. longipalpis* LLJB sand flies. (p. 158)

Figure 4.6. Confirmation of expression of fluorescent markers mCherry and GFP in transgenic drug-resistant lines Kubba SAT and MN-11 NEO. (p. 159)

Figure 4.7. Confirmation by PCR of inheritance of the drug resistance markers NEO and HYG un the 10 putative hybrid lines generated in the MA-37 NEO x L747 HYG cross. (p. 162)

Figure 5.1. Somy estimates for all hybrid and parental lines. (p. 179)

Figure 5.2. Heterozygous allele frequencies on chromosome 1 for one of the hybrids (H2a) and for one of the two parental lines (L747 HYG). (p. 181)

Figure 5.3. Heterozygous allele frequencies for chromosome 23 in the same two isolates as in Figure 5.2. (p.186)

Figure 5.4. Allele frequency histograms and read depth for all sites and biallelic sites only on chromosomes 1 and 23 in hybrid H2a. (p. 187)

Figure 5.5. Allelic plot showing biparental inheritance of alleles from the two parental lines to the offspring. (p. 189)

LIST OF TABLES

Table 2.1. A list of the 34 isolates of *L. tropica* that were used in this study. (p. 61)

Table 2.2. Marker panel comprising the 28 nuclear and kDNA loci that were considered for this study. (p. 63)

Table 2.3. Allelic diversity and heterozygosity of the 17 markers examined in 34 isolates of *L. tropica*. (p. 73)

Table 2.4. Number of lanes, total depth, and total number of bases obtained from sequencing runs of the 18 isolates that were whole genome sequenced for this study. (p. 83)

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. (p. 108)

Table 3.2. Number of lanes, insert size, read length in number of cycles, 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. (p. 114)

Table 3.3. A list of the top 30 most significant differentially expressed genes as depicted in Figure 3.5. (p. 122)

Table 3.4. A list of the genes with evidence of allele specific gene expression in both clones originating from the Rupert isolate. (p. 133)

Table 4.1. Transgenic drug resistant lines generated for crossing experiments with drug resistance markers NEO, HYG, and SAT. (p. 158)

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. (p. 161)

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. (p. 162)

Table 5.1. An example illustrating the phasing problem for two linked, biallelic disomic loci. (p. 170)

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. (p. 177)

Table 5.3. Number of Mendelian violations in the L747 HYG x MA-37 NEO cross of *L. tropica*, grouped by individual. (p. 190)

Table 5.4. Number of *de novo* variants private to the hybrid offspring, shown as a fraction of the total number of variants identified. (p. 191)

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 <i>in situ</i> 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 leishmaniasis	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