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.

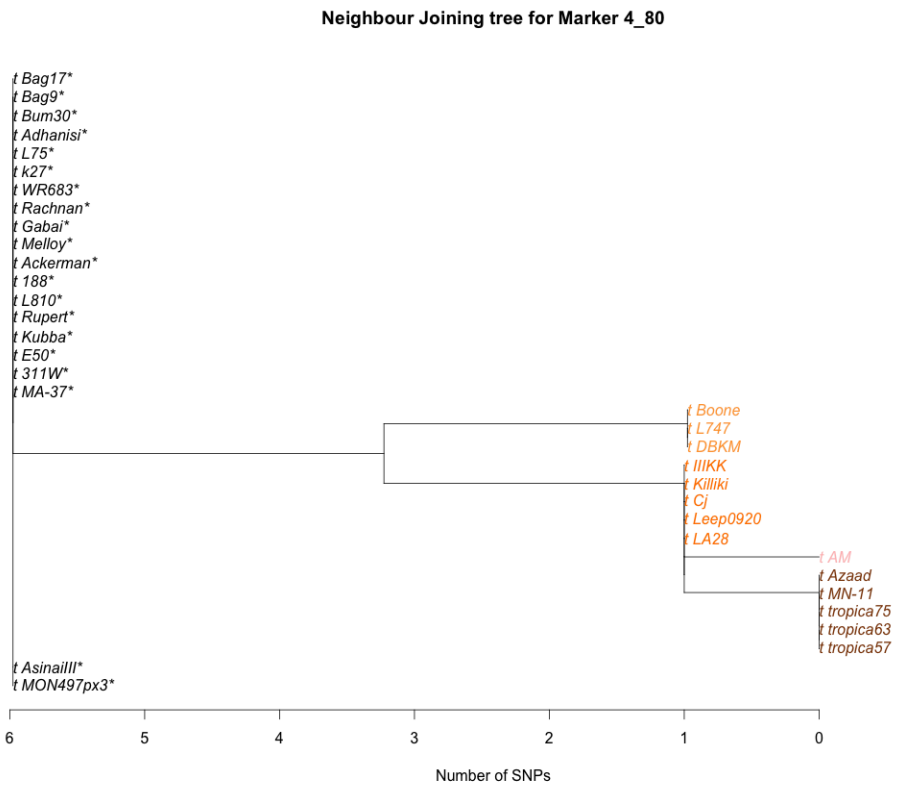

Figure 2.1. Example of a neighbour–joining (NJ) gene tree used in the assignment of individual parental alleles to heterozygous genotypes for allelic plot reconstruction. The marker shown above is 4_80 on chromosome 4. All isolates in black marked with an asterisk share the same identical heterozygous 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.

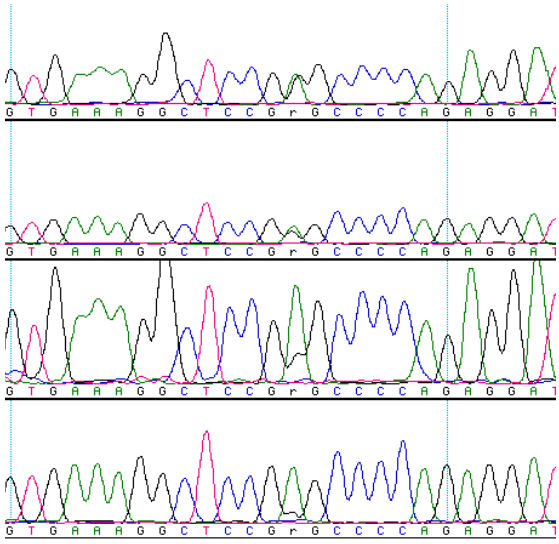

Figure 2.2. Aligned sequences for the 31_AQP1 marker for four of the isolates in our sample set. Notice the base call "r" (R in IUPAC nomenclature stands for A or G), which was used to manually annotate heterozygous positions. The presence of a green trace (nucleotide A) that is three times the size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

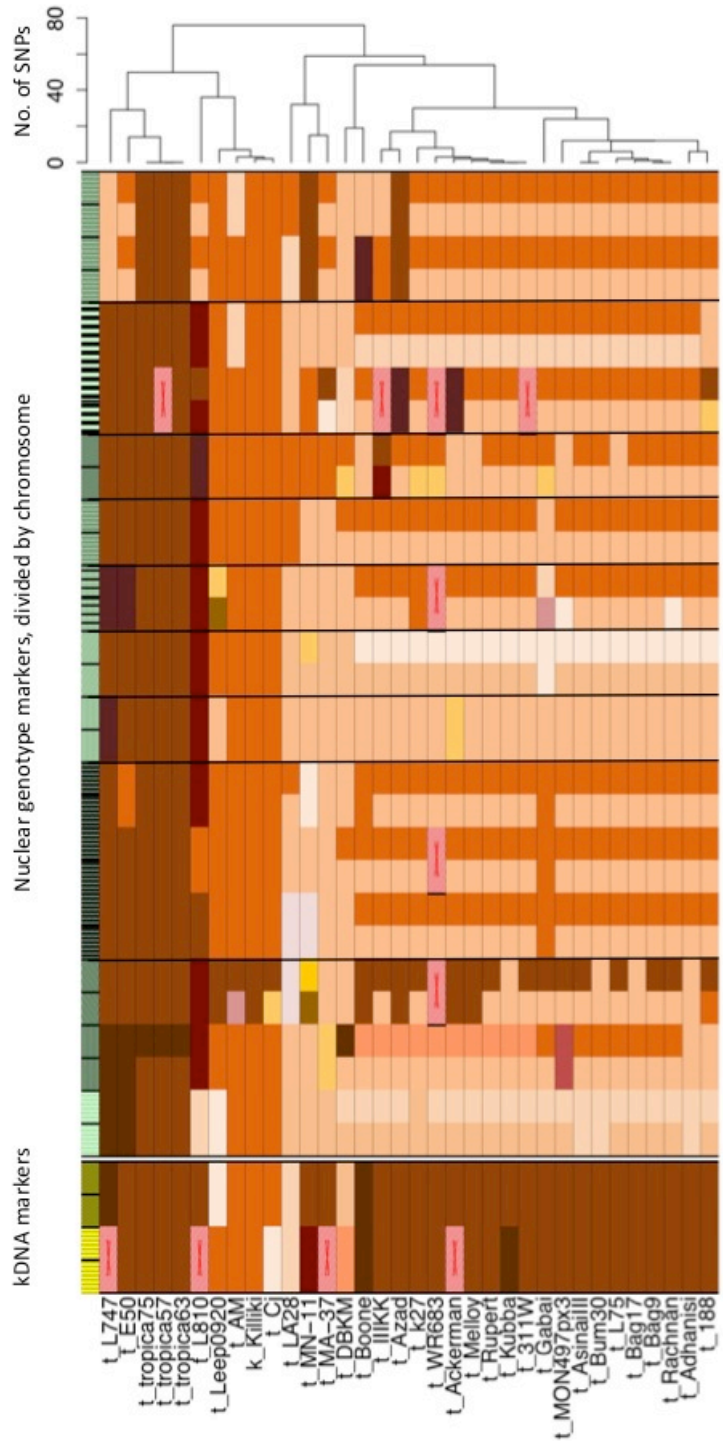
No. of SNPs
0
40
80
Nuclear genotype markers, divided by chromosome
kDNA markers
t_L747
t_E50
t_tropica75
t_tropica57
t_tropica63
t_L810
t_Leep0920
t_AM
k_Killiki
t_Ci
t_LA28
t_MN−11
t_MA−37
t_DBKM
t_Boone
t_IIIKK
t_Azad
t_k27
t_WR683
t_Ackerman
t_Melloy
t_Rupert
t_Kubba
t_311W
t_Gabai
t_MON497px3
t_Asinaiili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

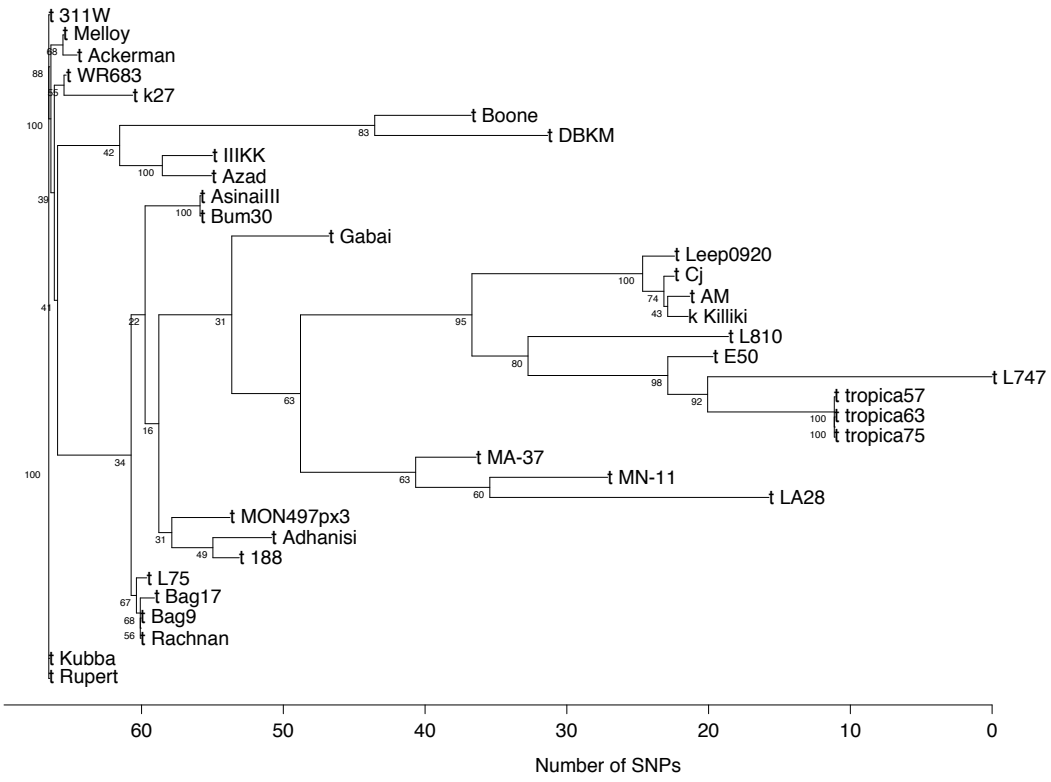

Figure 2.4. NJ tree from concatenated sequence data of markers with complete sequence information for all isolates. Note the presence of many unique "sequence types", represented by single isolates being at the tip of individual branches. Branch length represents the number of SNPs separating isolates 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).

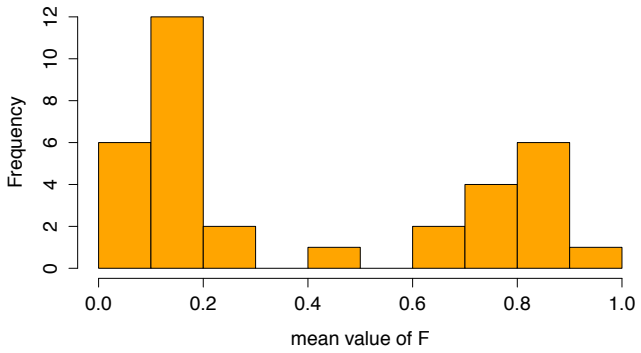

Figure 2.5. Histogram of inbreeding coefficients (F_{IT}) for the 34 isolates in this study. The data appears to be bimodally distributed, with two peaks, one centered around 0.1 and one around 0.8. See Appendix B for F_{IT} values.

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).

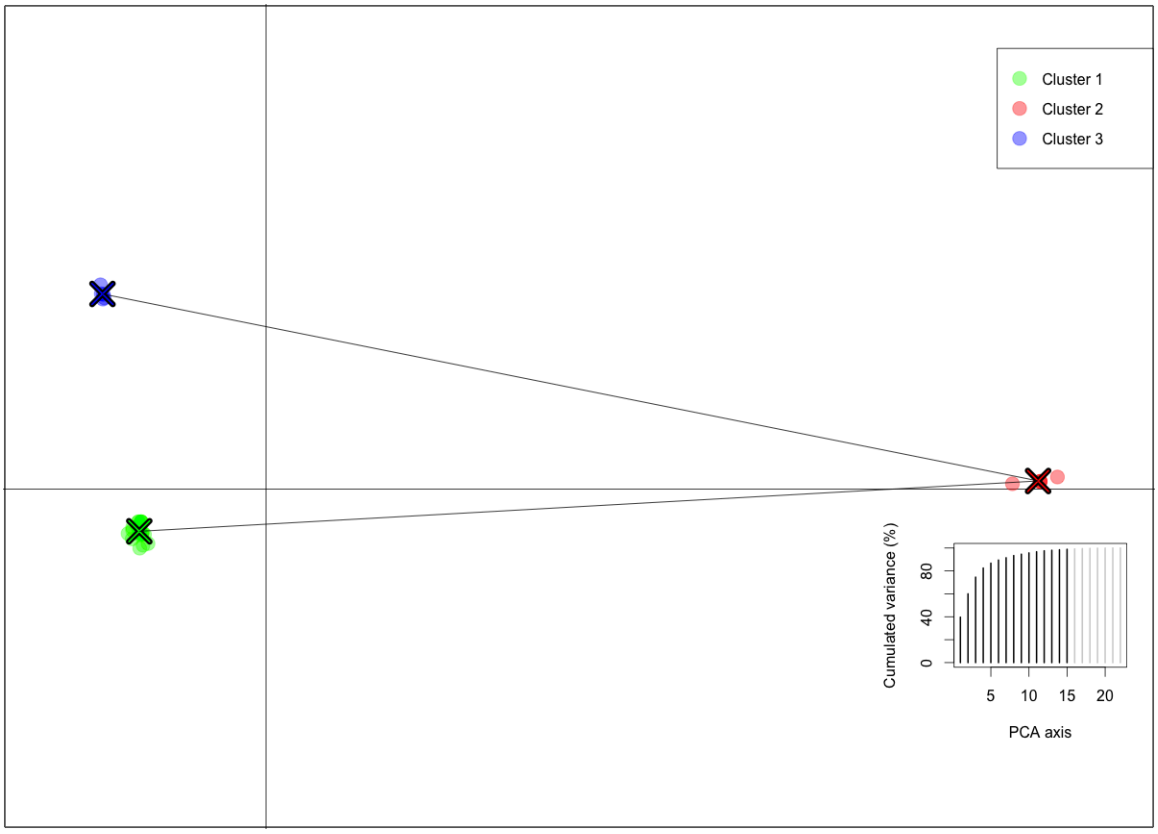

Figure 2.6. Clustering of the 34 isolates of *L. tropica* is consistent with geography and indicates possible mixing of different clusters in Israel and neighbouring countries. DAPC clustering and *k*-means analysis finds an optimal number of clusters *k* = 3. The first 15 principal component were retained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

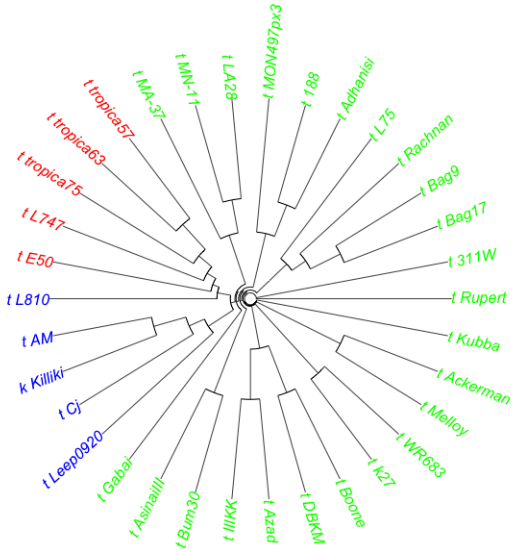

Figure 2.7. Unrooted, ultrametric NJ tree of 34 isolates of *L. tropica* based on the concatenated sequence data. Please refer to Figure 2.3 for a non-ultrametric version where branch lengths represent genetic distance. Colors are based on the DAPC clusters as in Figure 2.6. The largest cluster (green, 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.

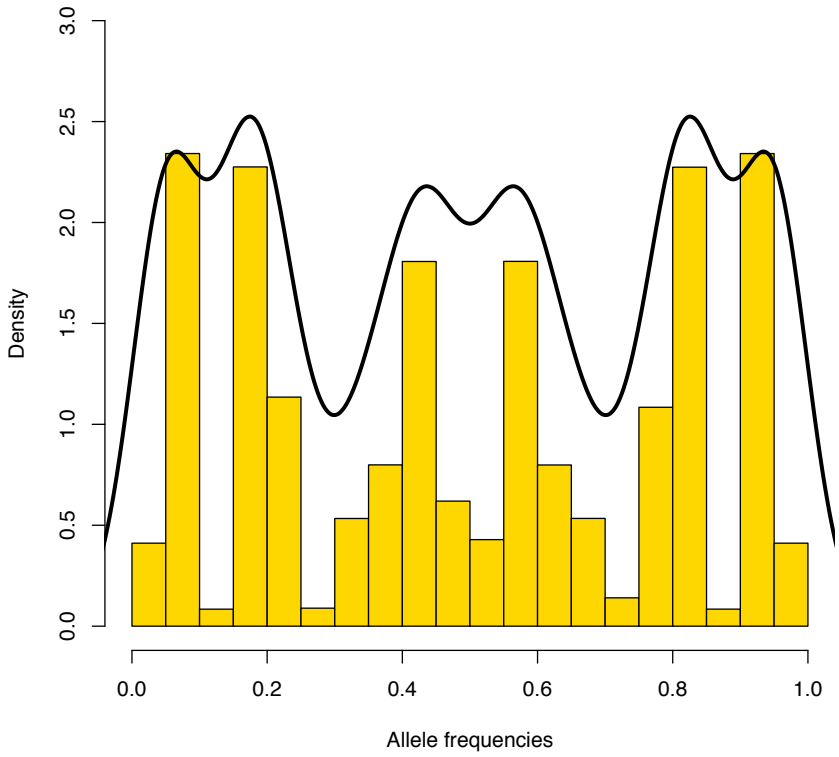

Figure 2.8. Average allele frequencies for each SNP across 18 individuals typed by WGS. Note the symmetry of the plot, expected for biallelic variants, and the peaks near 0 and 1 (homozygous), the peaks near 0.4 and 0.6 (heterozygous trisomic), and near 0.2 and 0.8 (heterozygous tetrasomic). Allele frequencies 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.

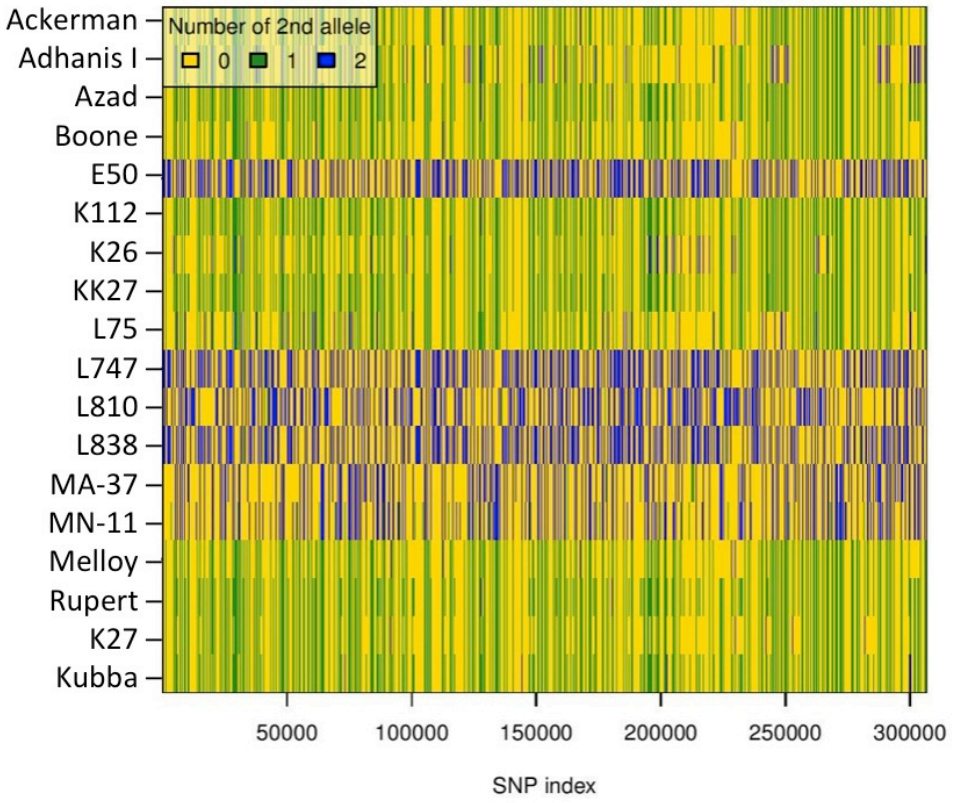

Figure 2.9. Allelic plot for all 306 596 SNPs that passed the filtering thresholds as described in the text, for all 18 strains. Heterozygous individuals have 1 copy of the 2nd allele at that typed position. Only samples that had either a frozen stock available or that had previously been sequenced were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

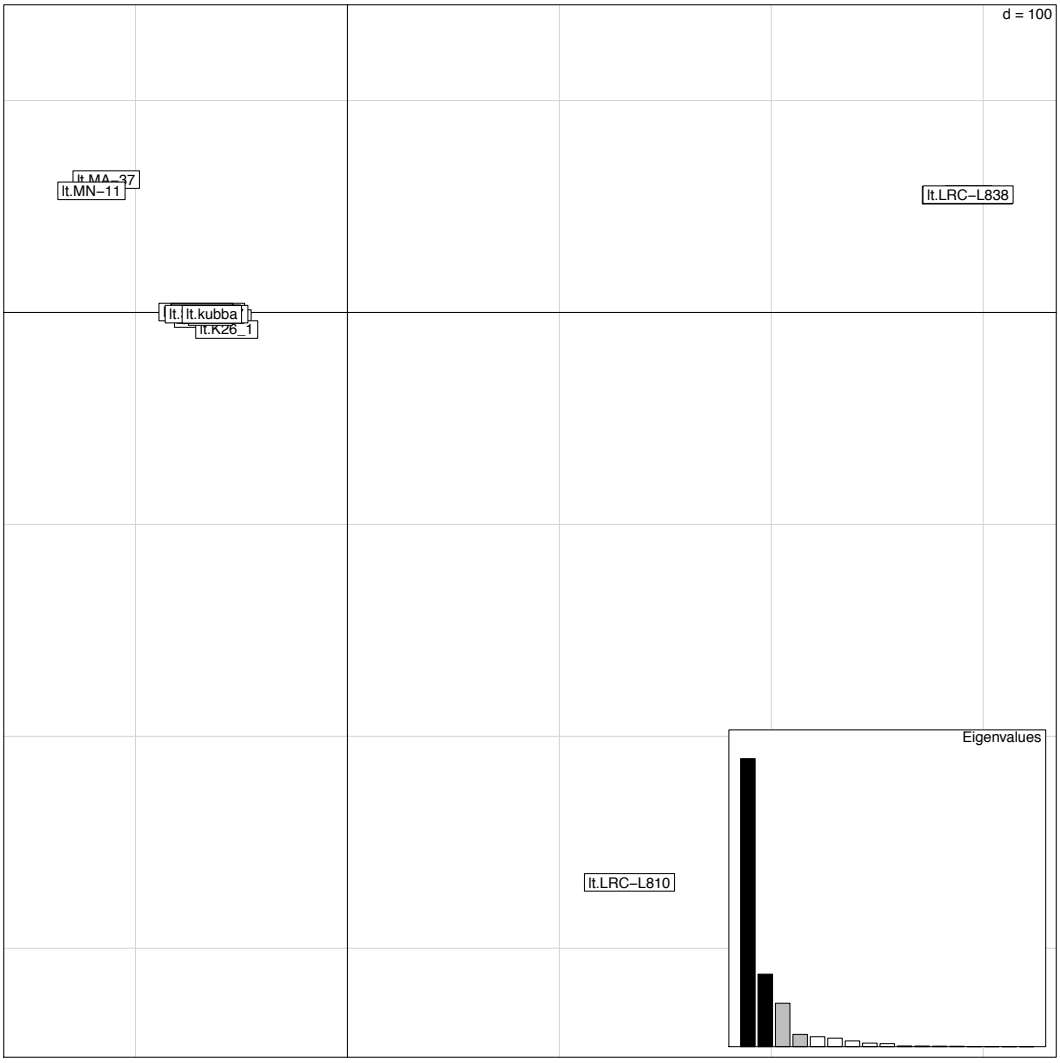

Figure 2.10. PCA plot of the 18 isolates that were typed by WGS. The first two principal components could explain almost 100% of the variance, as shown by the darker bars in the inset. Considerable distance is observed between L838 (tropica63 in the MLST analysis) and L810 and the rest of the isolates.

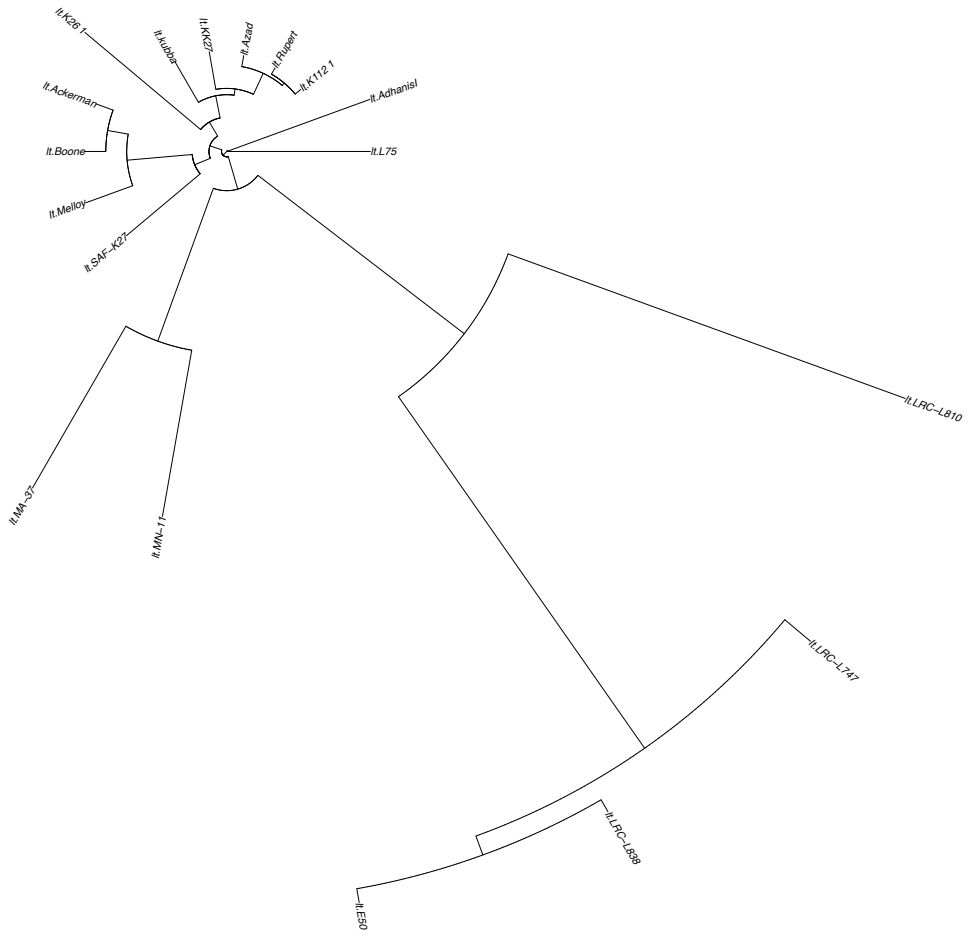

Figure 2.11. NJ tree for all 18 isolates based on the same set of 306 596 high quality biallelic SNPs as in Figure 2.9. Isolates L747, L838 (tropica63) and E50 appear to be outliers (part of Cluster 2 in Figure 2.7), while L810 appears to be quite divergent. MA-37 and MN-11 appear to be closely related,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