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.

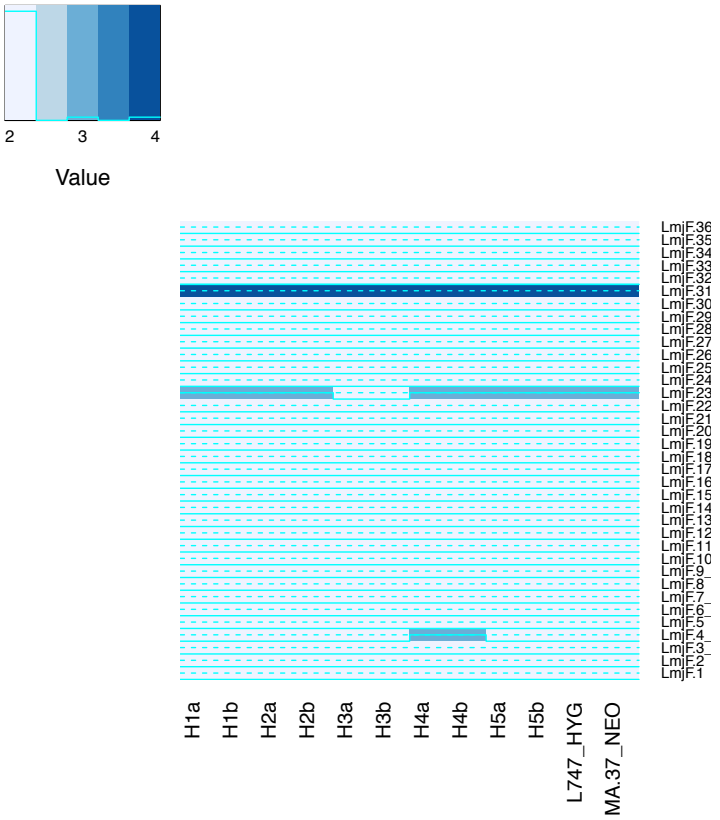
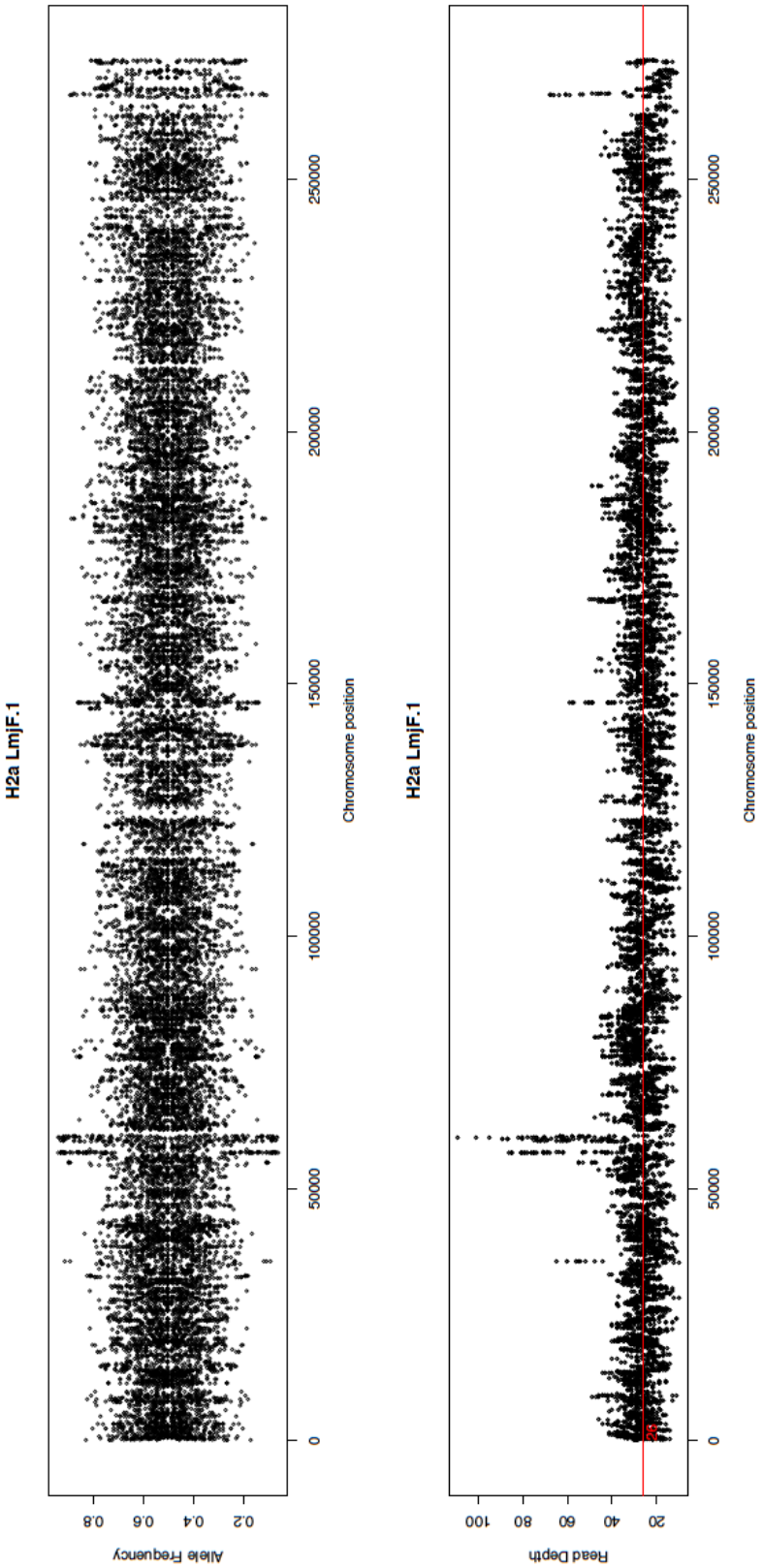


Figure 5.1. Somy estimates from WGS data for all hybrid and parental lines represented as a heatmap. Hybrids are indicated by the letter H. Darker shades of blue are associated with larger somy number, from a minimum somy of 2 to a maximum somy of 4. Note that chromosome 23 (LmjF.23) is trisomic in the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

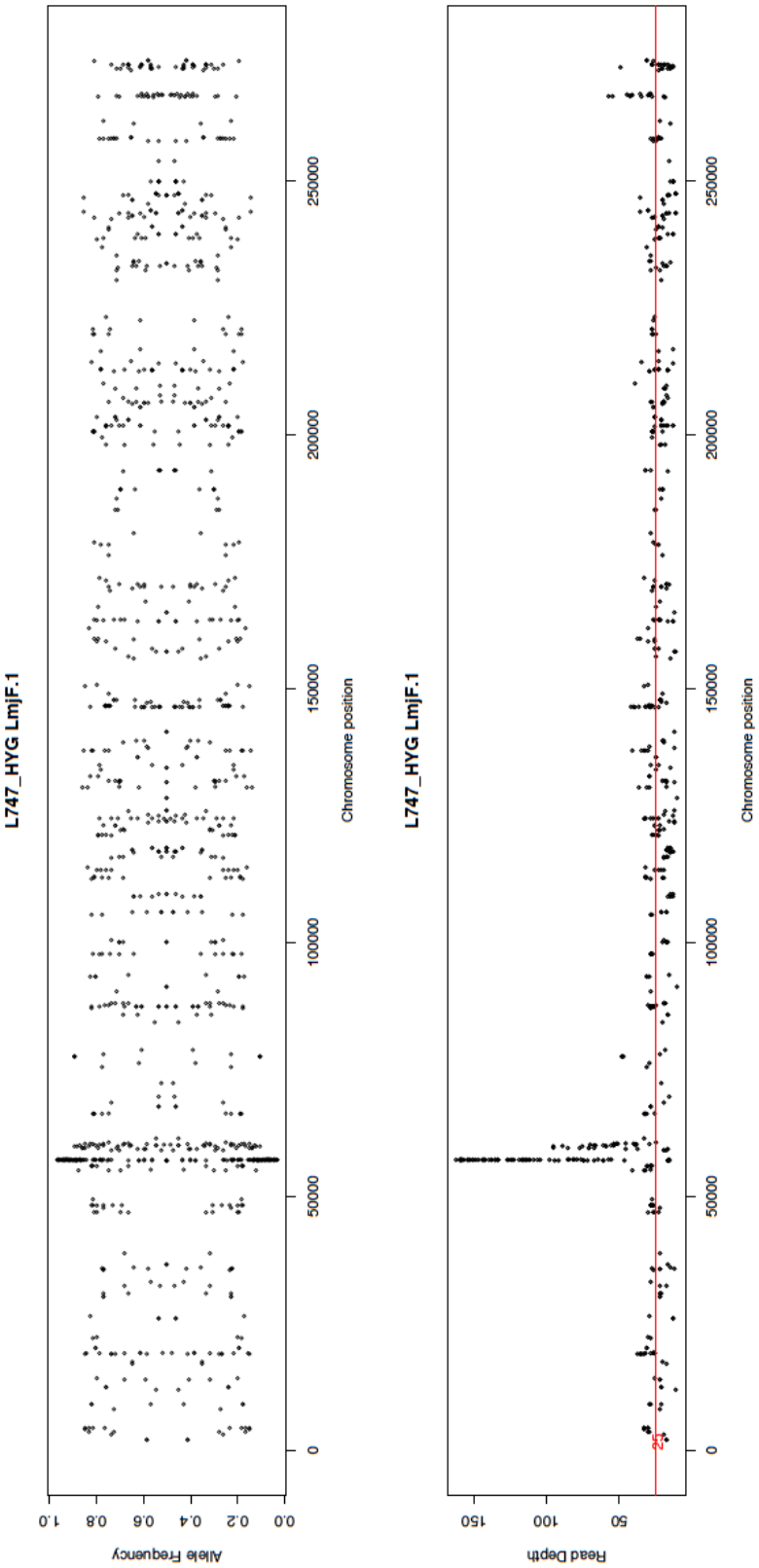
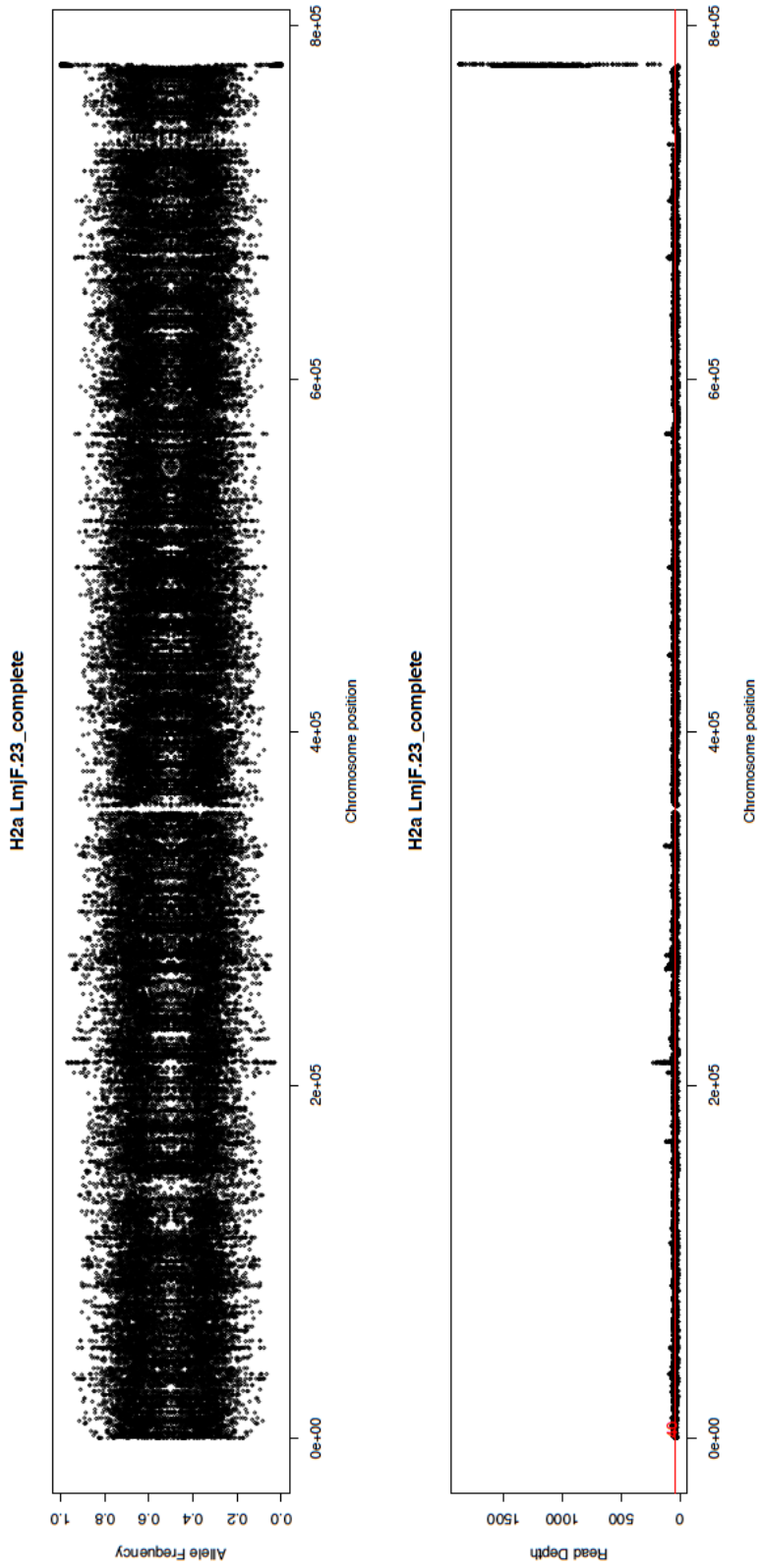

L747_HYG LmjF.1
Allele Frequency
0.0
0.2
0.4
0.6
0.8
1.0
0
50000
100000
150000
200000
250000
Chromosome position
L747_HYG LmjF.1
Read Depth
50
100
150
25
0
50000
100000
150000
200000
250000
Chromosome position

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

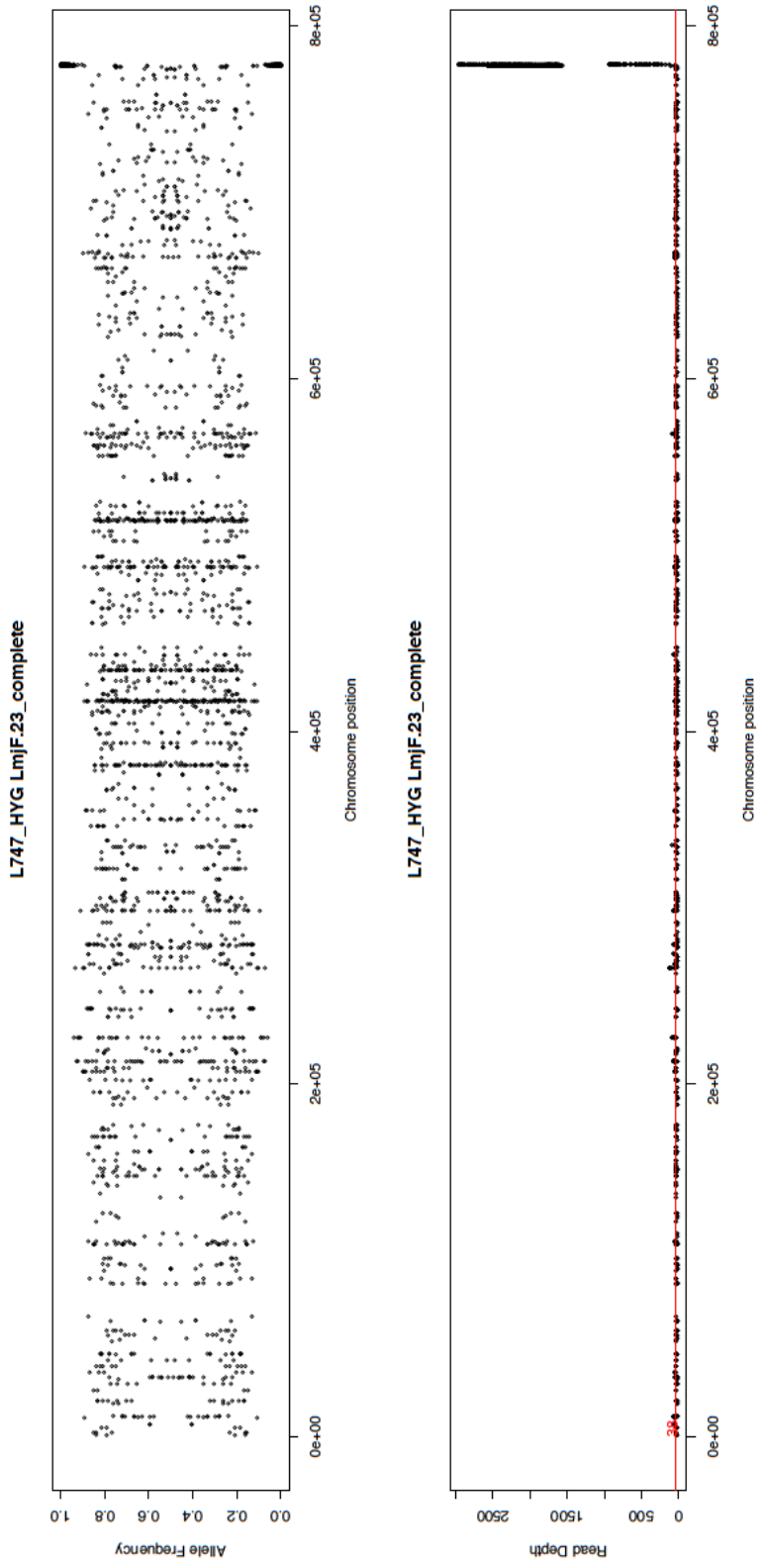

L747_HYG LmjF.23_complete
Allele Frequency
0.0
0.2
0.4
0.6
0.8
1.0
0e+00
2e+05
4e+05
6e+05
8e+05
Chromosome position
L747_HYG LmjF.23_complete
Read Depth
0
500
1500
2500
30
0e+00
2e+05
4e+05
6e+05
8e+05
Chromosome position

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.

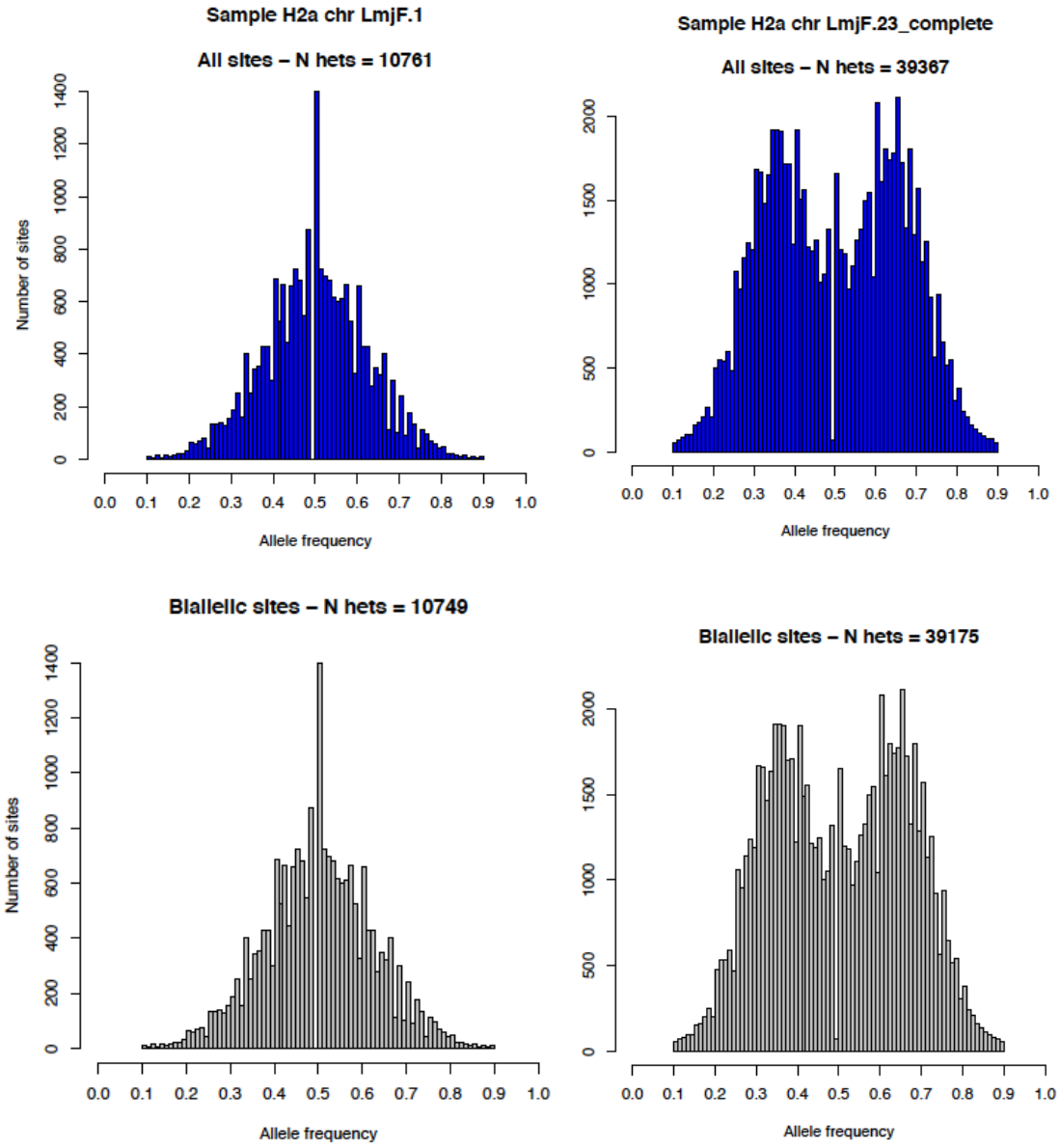


Figure 5.4. Allele frequency histograms for all heterozygous sites and for heterozygous biallelic sites only on chromosomes 1 (LmjF.1) and 23 (LmjF.23) in hybrid H2a, with the total number of sites indicated as "N hets". These distributions are illustrative of disomic and trisomic chromosomes, respectively,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

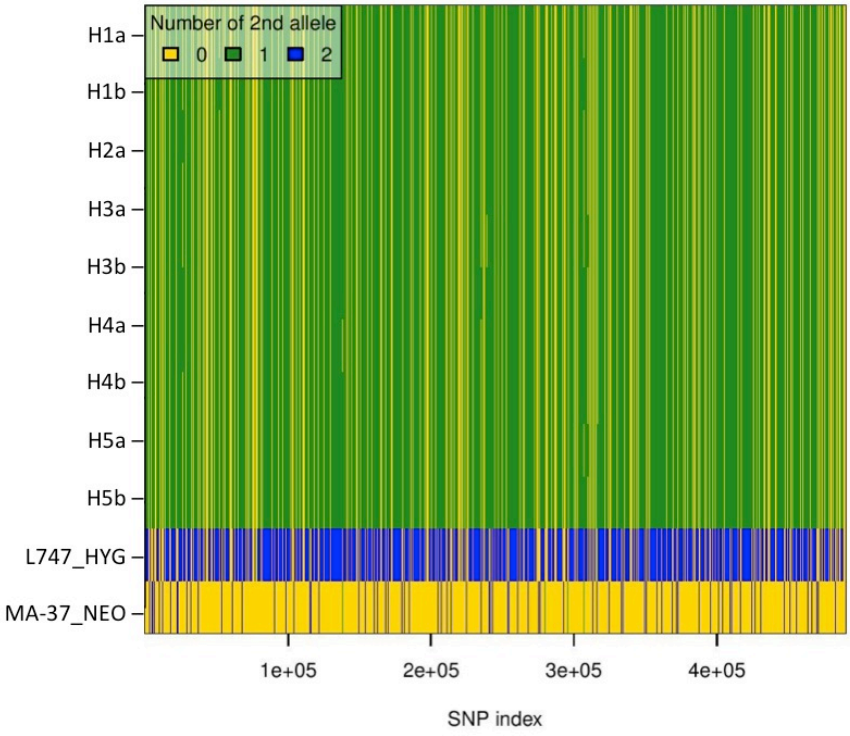


Figure 5.5. Allelic plot showing biparental inheritance of alleles from the two parental lines (bottom two rows, individual indices 11 and 12) to the offspring (indices 1 through 10). A blue cell represent a position that is homozygous for the first allele, a yellow cell is homozygous for the second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