CHAPTER 3

GENOME PLASTICITY AND GENE EXPRESSION

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

3.1. Introduction

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.2. Methods

3.2.1. Sequencing and RNA-seq of field isolates

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

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

3.2.2. Mapping and analysis of whole genome sequence data

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

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

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

3.2.3. Mapping and analysis of RNA-seq data

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

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

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

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

3.2.4. Differential gene expression analysis

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

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

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

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

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

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

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

3.3. Results

3.3.1. Sequencing of 20 parasite lines

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

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

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

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

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

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

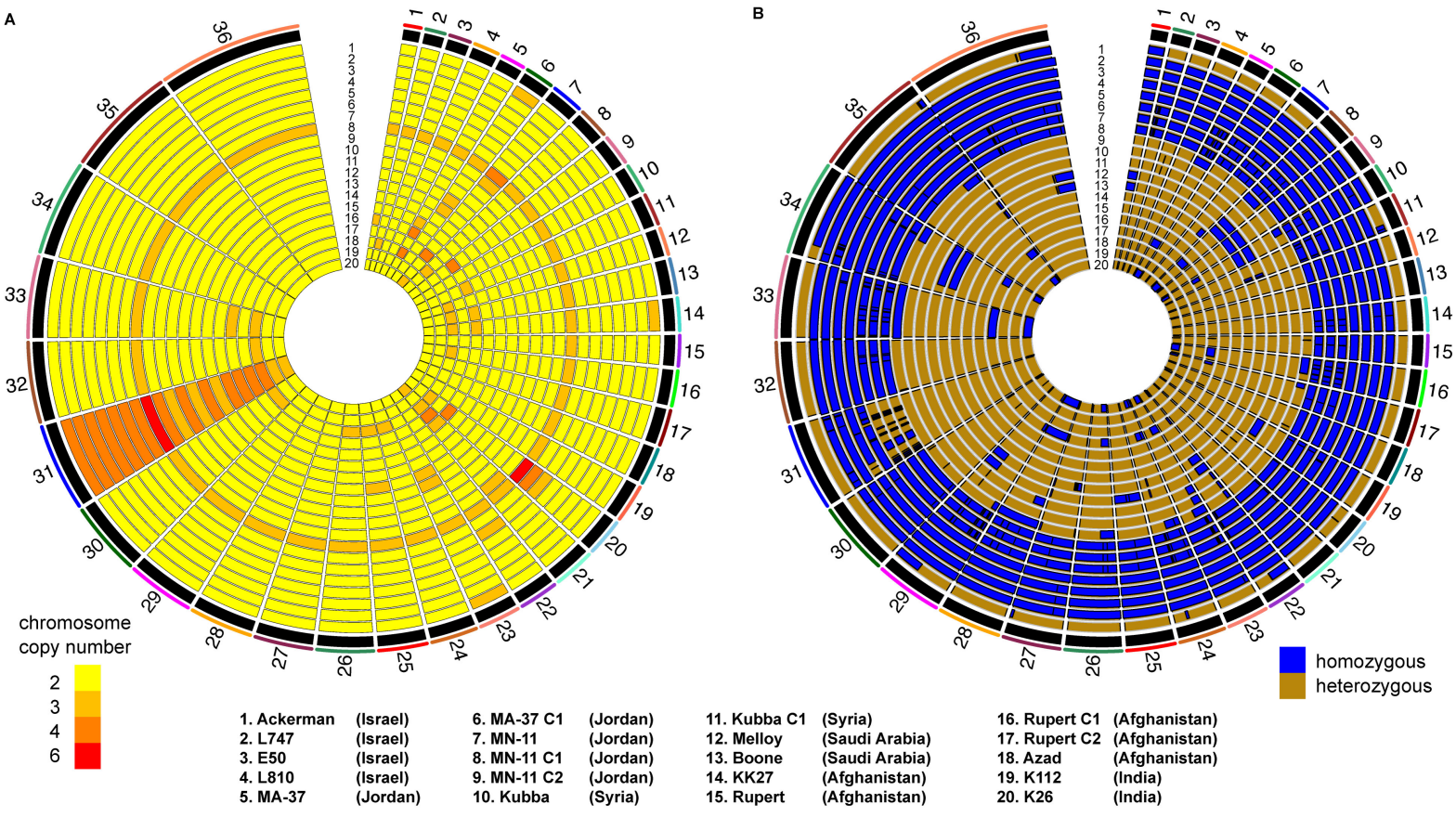
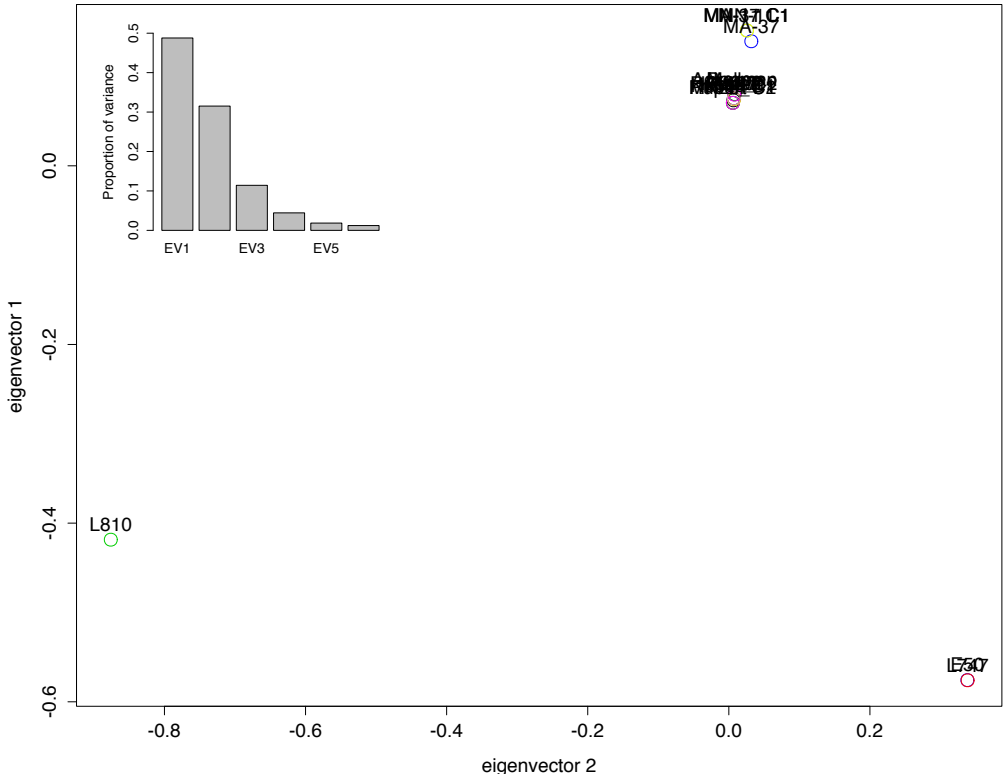
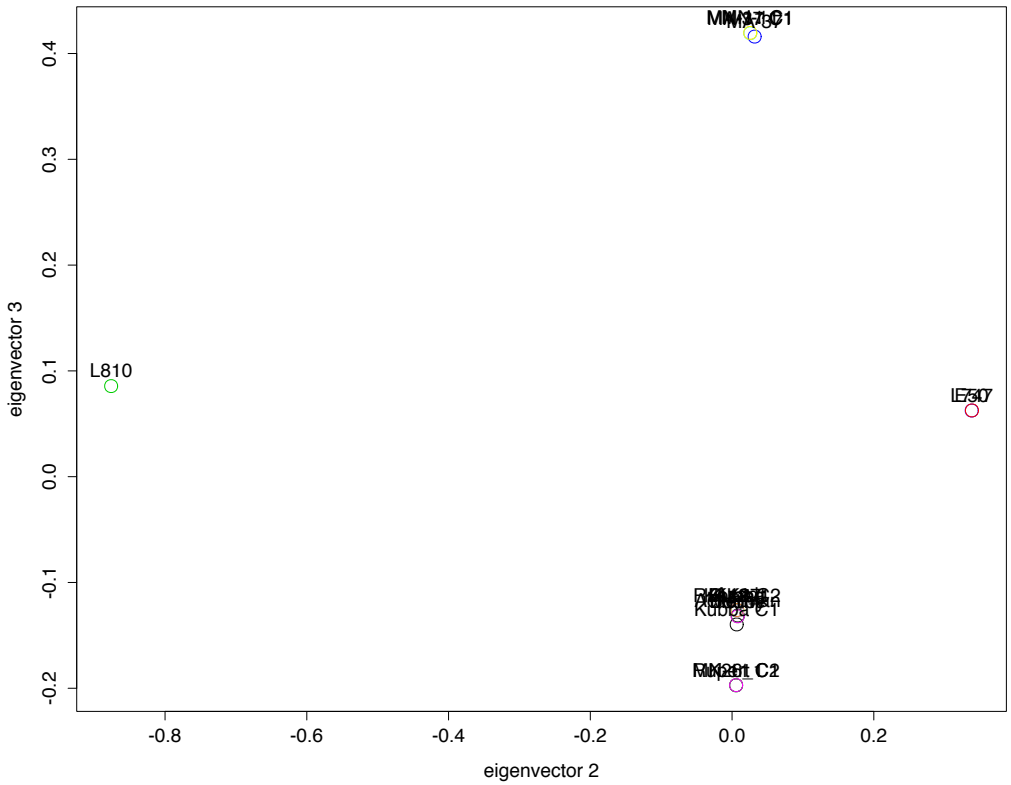

A
B
chromosome
copy number
2
3
4
6
homozygous
heterozygous
1. Ackerman (Israel)
2. L747 (Israel)
3. E50 (Israel)
4. L810 (Israel)
5. MA-37 (Jordan)
6. MA-37 C1 (Jordan)
7. MN-11 (Jordan)
8. MN-11 C1 (Jordan)
9. MN-11 C2 (Jordan)
10. Kubba (Syria)
11. Kubba C1 (Syria)
12. Melloy (Saudi Arabia)
13. Boone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

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


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


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

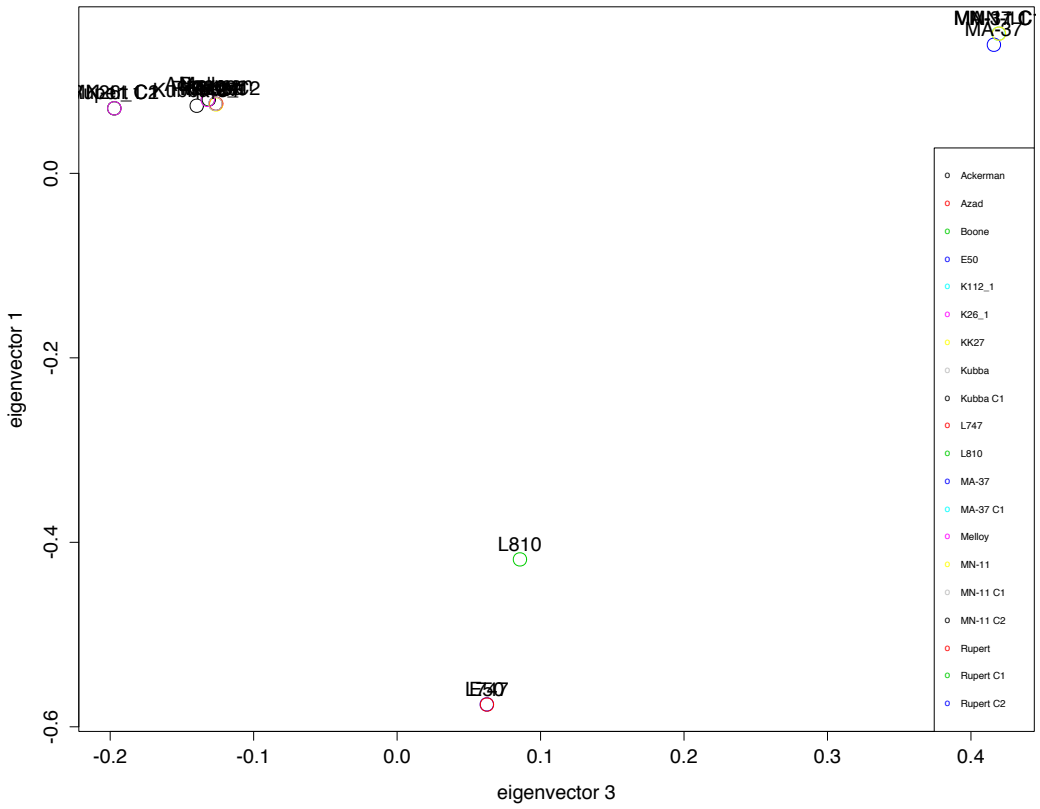


Figure 3.2. PCA plots of the 14 different isolates and 6 additional clones considered for this analysis, comparing eigenvectors 1 vs 2, 2 vs 3, and 1 vs 3. The first three eigenvectors, which represent the first three PCs, explain the majority of the variance observed, as shown by the inset in EV1 vs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

3.3.2. RNA-seq of field isolates and clonal lines

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

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

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

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

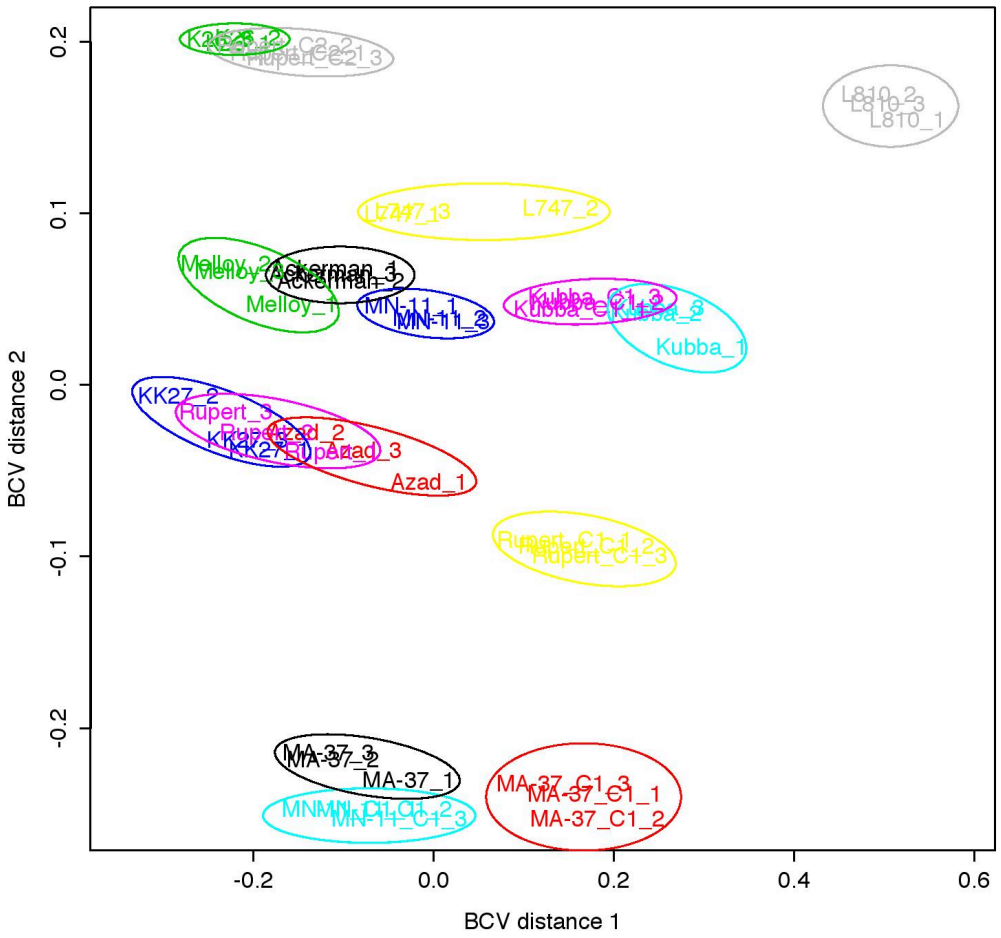


Figure 3.3. MDS plot representing expression data for all strains, except for isolates Boone, E50, and K112 (please refer to main text for description of rationale for their exclusion). Each triplicate is depicted in a different color. The replicates are numbered as per the legend included.

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

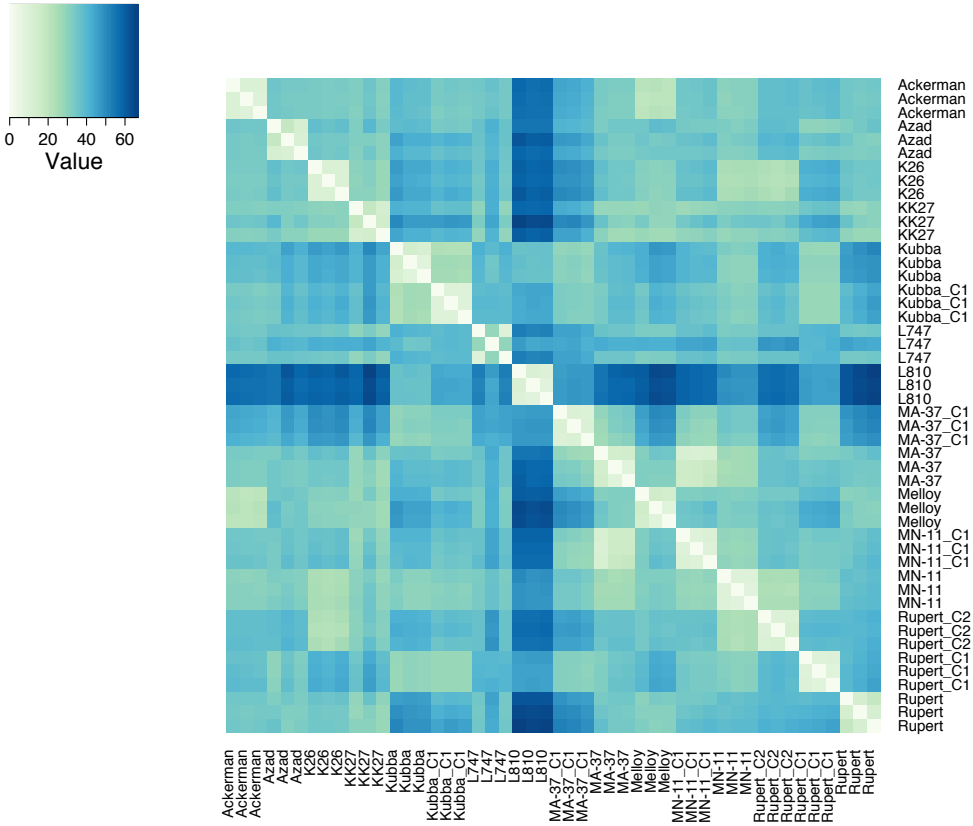


Figure 3.4. Heatmap of Euclidean distances between variance stabilized expression values for each pair of samples. MN-11 C2 and Boone showed aberrant expression signatures, and were later excluded since they were possible contaminations of the cultures.

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

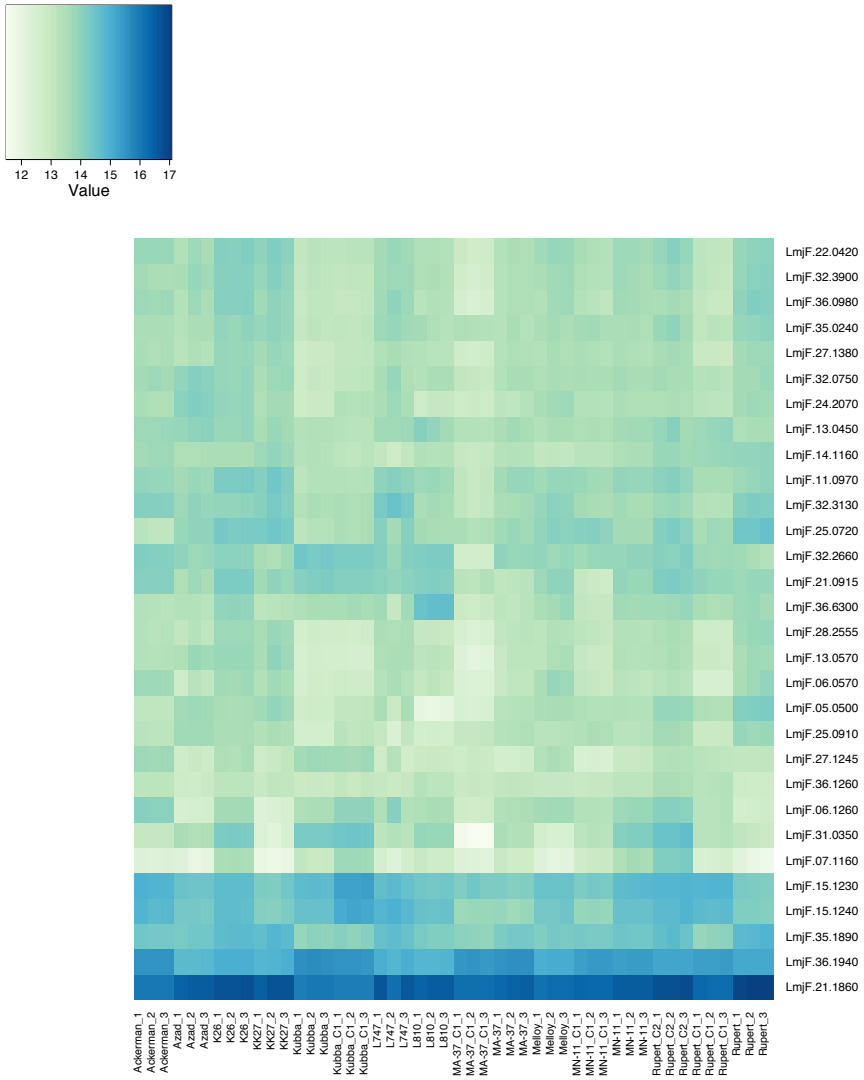


Figure 3.5. Heatmap showing the most highly expressed genes in our set of samples. The shading of each cell represents variance-stabilized expression values as implemented in DEseq. Genes were organized by hierarchical clustering, with the most highly expressed genes at the bottom of the graph. See Appendix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

3.3.3. Differential gene expression analysis

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

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

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

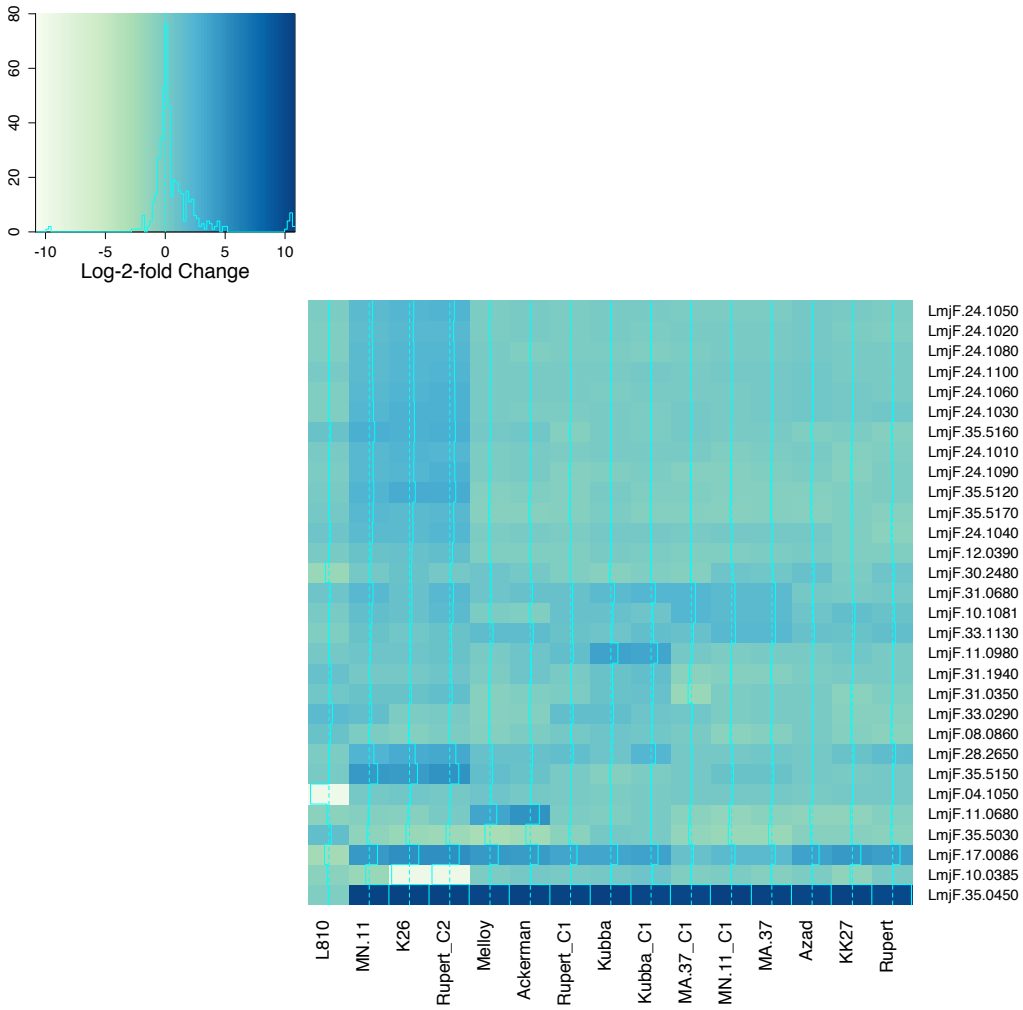


Figure 3.6. Heatmap of the top 30 most significant DE genes showing log-fold change in expression. Note the large cluster of DE genes on chromosomes 24 (gene name starting with LmjF.24) and 35 (gene name starting with LmjF.35) in 4 isolates, MN11, MN11 C1, Rupert C2, and K26. Inspection of the whole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

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

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

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

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

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

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

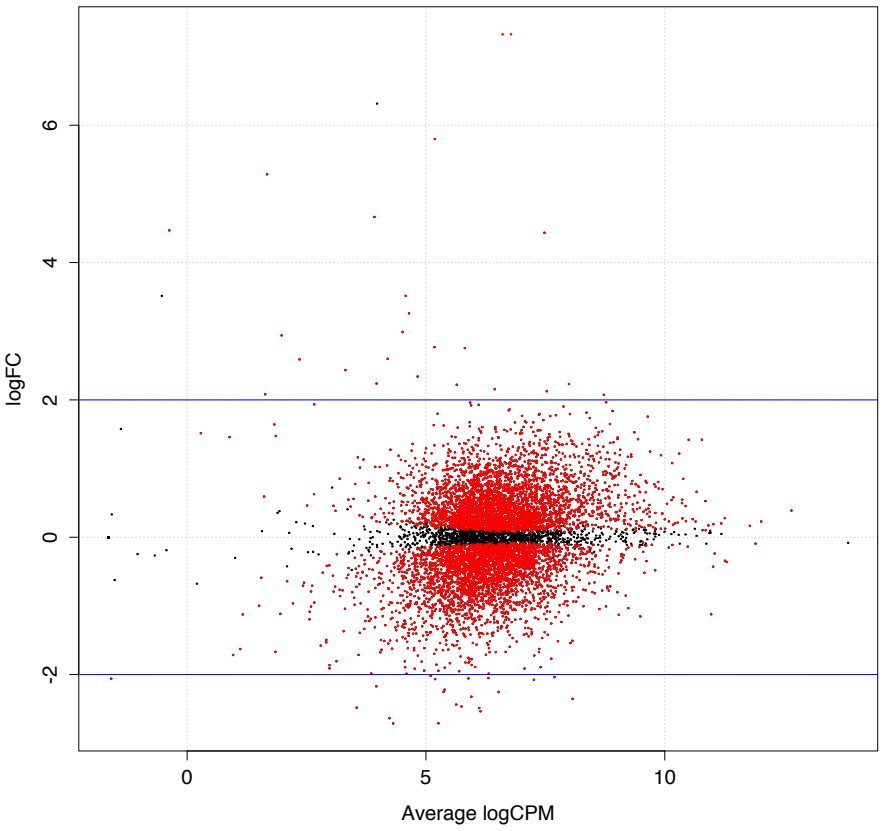


Figure 3.7. Smear plot of log-fold changes in expression versus average log-counts per million in isolate L810 compared to all other isolates. Each dot is a gene, and dots colored in red is significant at the FDR < 0.05 threshold. The blue bars represent the +2 and -2 log-fold change thresholds. There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

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

3.3.4. Copy number variation and gene dosage effects

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

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

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

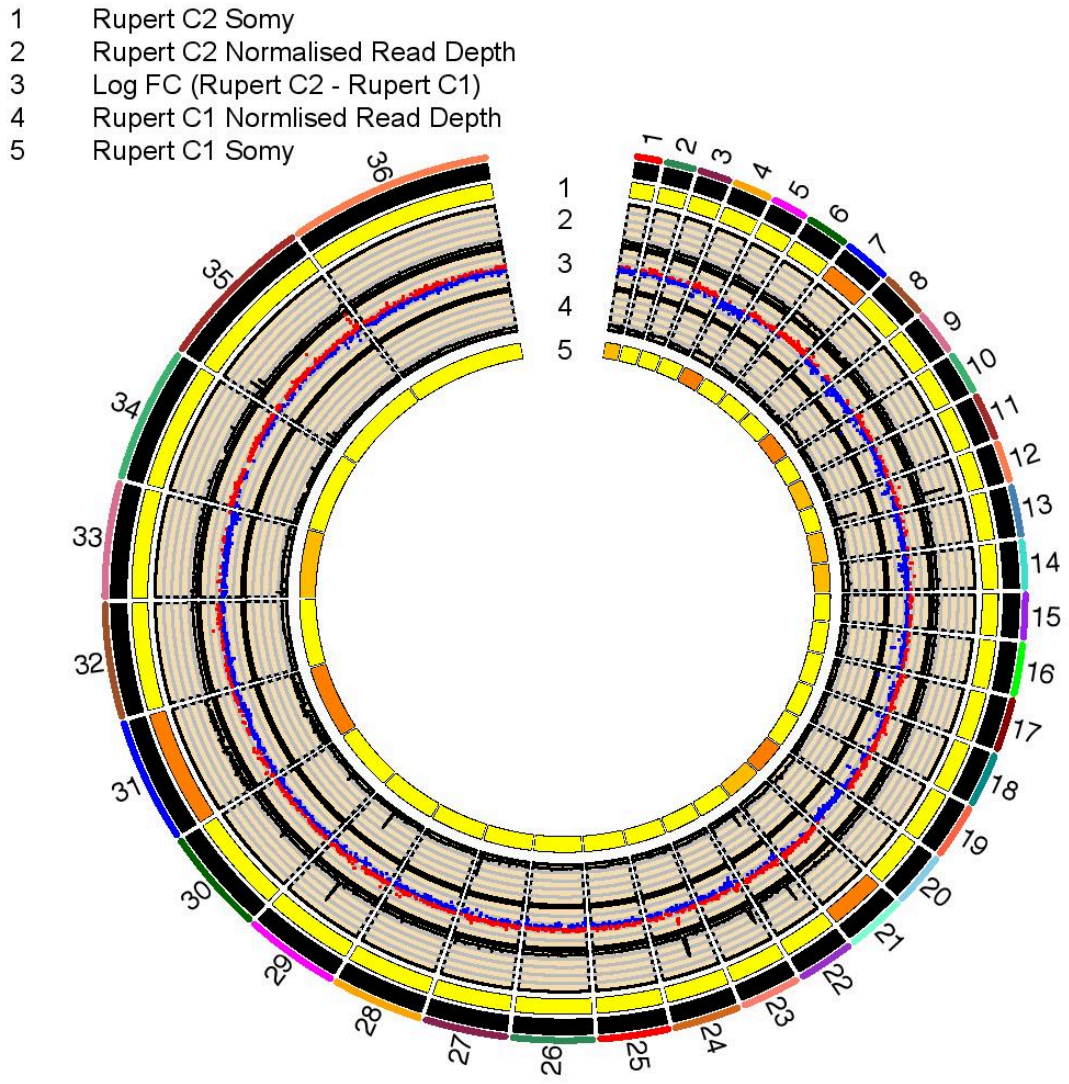


Figure 3.8. Gene dosage effects due to aneuploidy on transcription in *L. tropica* clonal lines Rupert C1 and Rupert C2. Tracks 1 and 5 represent somy as in Figure 3.1. Tracks 2 and 4 represent normalized read depth. Track 3 represents significant DE genes between these two clones. Positive log-fold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

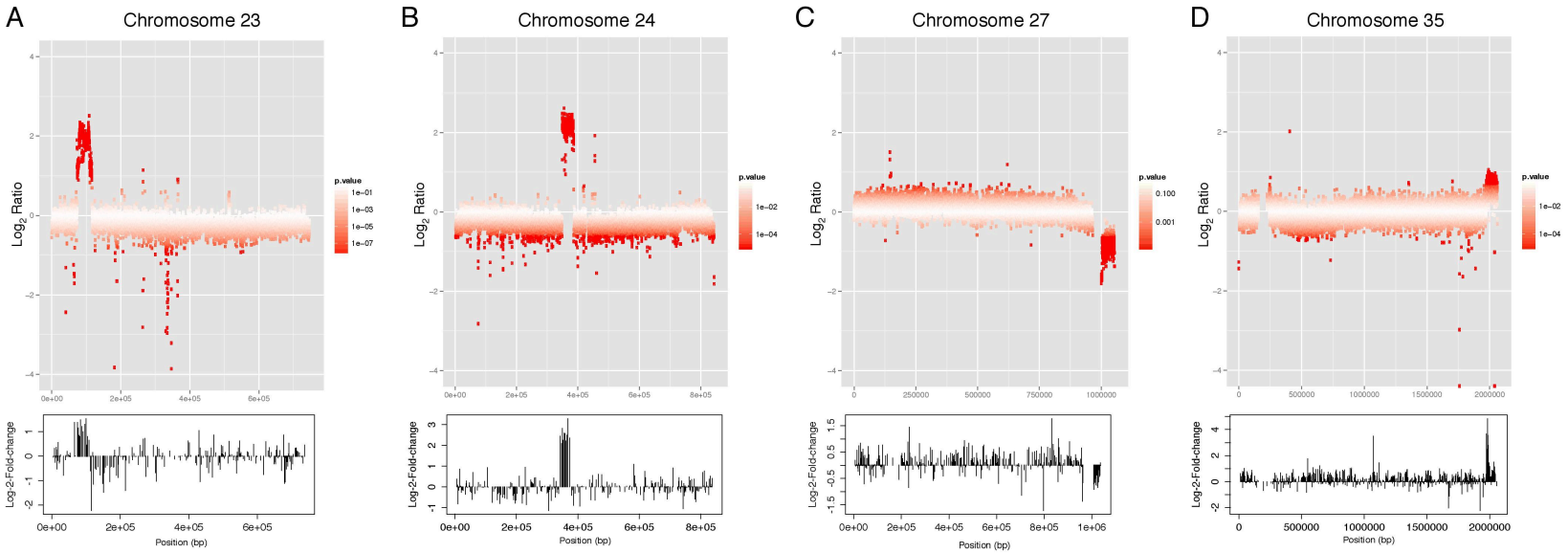

A
Chromosome 23
B
Chromosome 24
C
Chromosome 27
D
Chromosome 35
Log2 Ratio
p.value
Log-2-Fold-change
Position (bp)

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

3.3.5. Allele-specific gene expression

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

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

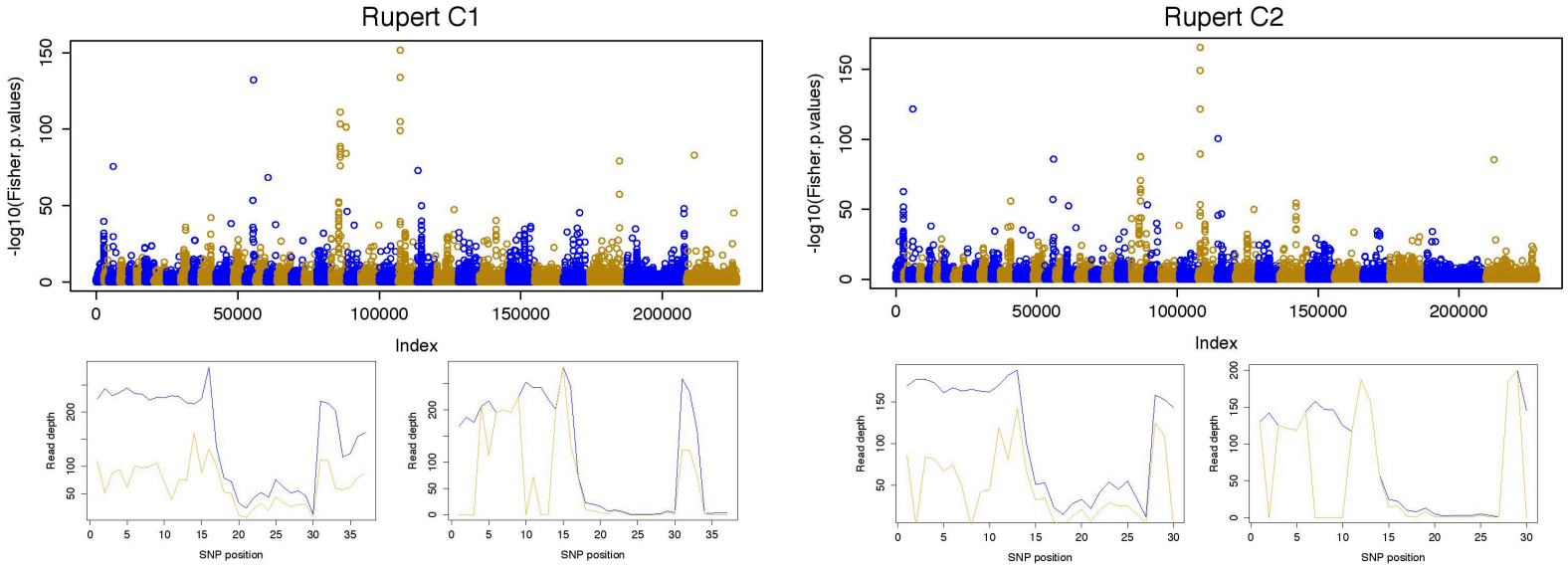

Rupert C1
-log10(Fisher.p.values)
0
50
100
150
0
50000
100000
150000
200000
Index
Read depth
SNP position
Rupert C2
-log10(Fisher.p.values)
Index
Read depth
SNP position

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

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

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

3.4. Discussion

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

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

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

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

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

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

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

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

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

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

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

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