Chromosome and gene copy number variation allow major structural change between species and strains of *Leishmania*

Matthew B. Rogers,[1,2,9] James D. Hilley,[3,9] Nicholas J. Dickens,[3,9] Jon Wilkes,[3] Paul A. Bates,[4] Daniel P. Depledge,[1,2] David Harris,[1] Yerim Her,[2] Pawel Herzyk,[5] Hideo Imamura,[1,6] Thomas D. Otto,[1] Mandy Sanders,[1] Kathy Seeger,[1] Jean-Claude Dujardin,[6,7] Matthew Berriman,[1] Deborah F. Smith,[2] Christiane Hertz-Fowler,[1,8,10] and Jeremy C. Mottram[3,10]

[1]*Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SA, United Kingdom;* [2]*Department of Biology, University of York, Heslington, York YO10 5DD, United Kingdom;* [3]*Wellcome Trust Centre for Molecular Parasitology, Institute of Infection, Immunity and Inflammation, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8TA, United Kingdom;* [4]*Division of Biomedical and Life Sciences, School of Health and Medicine, Lancaster University, Lancaster LA1 4YQ, United Kingdom;* [5]*Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8TA, United Kingdom;* [6]*Unit of Molecular Parasitology, Department of Parasitology, Institute of Tropical Medicine, B-2000 Antwerp, Belgium;* [7]*Department of Biomedical Sciences, Faculty of Pharmaceutical, Biomedical and Veterinary Sciences, University of Antwerp, B-2020 Antwerp, Belgium*

Leishmania parasites cause a spectrum of clinical pathology in humans ranging from disfiguring cutaneous lesions to fatal visceral leishmaniasis. We have generated a reference genome for *Leishmania mexicana* and refined the reference genomes for *Leishmania major*, *Leishmania infantum*, and *Leishmania braziliensis*. This has allowed the identification of a remarkably low number of genes or paralog groups (2, 14, 19, and 67, respectively) unique to one species. These were found to be conserved in additional isolates of the same species. We have predicted allelic variation and find that in these isolates, *L. major* and *L. infantum* have a surprisingly low number of predicted heterozygous SNPs compared with *L. braziliensis* and *L. mexicana*. We used short read coverage to infer ploidy and gene copy numbers, identifying large copy number variations between species, with 200 tandem gene arrays in *L. major* and 132 in *L. mexicana*. Chromosome copy number also varied significantly between species, with nine supernumerary chromosomes in *L. infantum*, four in *L. mexicana*, two in *L. braziliensis*, and one in *L. major*. A significant bias against gene arrays on supernumerary chromosomes was shown to exist, indicating that duplication events occur more frequently on disomic chromosomes. Taken together, our data demonstrate that there is little variation in unique gene content across *Leishmania* species, but large-scale genetic heterogeneity can result through gene amplification on disomic chromosomes and variation in chromosome number. Increased gene copy number due to chromosome amplification may contribute to alterations in gene expression in response to environmental conditions in the host, providing a genetic basis for disease tropism.

[Supplemental material is available for this article.]

The leishmaniases are a complex of diseases caused by species of the protozoan parasite *Leishmania*. Parasites are transmitted to mammalian hosts via the bite of female phlebotomine sand flies, with the geographical range of sand fly species capable of transmitting parasites being the main limiting factor for disease prevalence in the Americas, Asia, Africa, and Europe. This distribution puts an estimated 350 million people at risk of infection across 88 countries, and up to 2 million people become infected annually, 0.5 million with the most severe visceral form of the disease (Murray et al. 2005).

The three main forms of leishmaniases are usually associated with particular species of *Leishmania*, although the host immune response to infection is also a critical determinant in development of disease. At least 20 species of *Leishmania* cause disease in man, with the most severe visceral leishmaniasis (VL) disseminating to visceral organs such as the liver and spleen. The severe form is caused by species of the *Leishmania donovani* complex, including *L. donovani* and *Leishmania infantum*, that disseminate to visceral organs such as the liver and spleen. Cutaneous leishmaniasis (CL) is caused by species such as the Old World *Leishmania major* and the New World *Leishmania mexicana* and is generally confined to tissues immediately surrounding the sand fly bite site. These spe-

[8]Present address: Centre for Genomic Research, Institute of Integrative Biology, Biosciences Building, University of Liverpool, Liverpool L69 7ZB, UK.
[9]These authors contributed equally to this work.
[10]Corresponding authors.
E-mail C.Hertz-Fowler@liverpool.ac.uk.
E-mail jeremy.mottram@glasgow.ac.uk.
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.122945.111. Freely available online through the *Genome Research* Open Access option.

cies may also disseminate to other areas of the skin, often following drug treatment, giving rise to diffuse CL (DCL). Although the disease caused by *Leishmania braziliensis* and other species of the *Viannia* subgenera normally presents in a cutaneous form, the parasite can also migrate to the nasopharyngeal tissues in a small proportion of cases, resulting in highly disfiguring mucocutaneous leishmaniasis (MCL).

Leishmania genomes, like those of other kinetoplastids, are characterized by a high degree of synteny, genes organized into polycistronic transcription units, and rare spliceosomal introns. *Leishmania* genomes lack large subtelomeric regions, which in the closely related African trypanosomes encode species-specific genes (Berriman et al. 2005; Peacock et al. 2007). The genomes of *L. infantum*, *L. donovani*, and *L. major* each consist of 36 chromosomes (Wincker et al. 1998), whereas the more distantly related *L. braziliensis* genome has only 35 chromosomes as a result of a fusion event involving chromosomes 20 and 34 (Britto et al. 1998; Peacock et al. 2007). The genome of *L. mexicana* consists of 34 chromosomes, with two unique fusion events having occurred between chromosomes 8 and 29, and between chromosomes 20 and 36 (Britto et al. 1998).

In common with other trypanosomatids, *Leishmania* has an unusual mechanism of transcriptional control (Clayton and Shapira 2007). Protein-coding genes in polycistronic transcription units are cotranscribed by RNA polymerase II, and precursor mRNA is subsequently *trans*-spliced and polyadenylated (Martinez-Calvillo et al. 2003). Messenger RNA levels are regulated by RNA stability, rather than the activity of promoters, precluding the up-regulation of gene expression through increased RNA polymerase II activity. One feature of *Leishmania* genome structure is the presence of tandem arrays of duplicated genes (Ivens et al. 2005; Peacock et al. 2007). These are predicted to allow increased gene expression in the absence of the regulated transcriptional control that is found in other eukaryotes.

The genome sequences of three *Leishmania* species have been reported to date, *L. major* Friedlin (Ivens et al. 2005), *L. infantum* JPCM5, and *L. braziliensis* M2904 (Peacock et al. 2007). Comparative genomic analysis of these species revealed highly conserved gene synteny, despite an estimated divergence of 46 million years (Lukes et al. 2007), as well as an unexpectedly small number of species-specific genes (Peacock et al. 2007). Most of these encode predicted proteins currently of no known function, and these have been proposed to contribute to the parasite tropism and pathology associated with the different forms of leishmaniasis (Peacock et al. 2007; Smith et al. 2007). Several *L. donovani*–specific genes have been expressed in *L. major* and shown to significantly increase parasite survival in visceral organs in mice, indicating that individual genes can contribute to parasite tropism in the host (Zhang et al. 2008; Zhang and Matlashewski 2010). Here, we present a reference genome for *L. mexicana* U1103 together with refined analyses of the published *L. major* Friedlin, *L. infantum* JPCM5, and *L. braziliensis* M2904 genomes. In addition, Illumina high-throughput resequencing of additional *Leishmania* strains or species, including two *L. donovani* strains (BPK206/0 and LV9), *L. major* LV39, and *L. mexicana* M379, has revealed the presence of species-specific genes in other isolates of the same species or complex. The present study has also identified gene and chromosome copy number differences between species and strains of *Leishmania* as a major source of genomic variation. These observations have important implications for the understanding of parasite variation and will guide further investigations into the genetic basis of disease tropism.

Results and Discussion

A *L. mexicana* reference genome, and updated versions of the *L. major*, *L. infantum*, and *L. braziliensis* genomes

L. mexicana U1103 is an isolate taken from the ear lesion of a 30-yr-old male patient in Guatemala. The U1103 genome was assembled de novo using capillary sequencing reads generated from a whole-genome shotgun library. The assembly was subsequently improved by scaffolding against three other *Leishmania* species (Ivens et al. 2005; Peacock et al. 2007) and using deep coverage Illumina sequencing reads for correction of sequencing errors (Otto et al. 2010). The resulting improved high-quality draft assembly (as defined by Chain et al. 2009) consists of 929 contigs, totaling 32 Mb of data, of which 375 contigs are ordered and scaffolded as 34 pseudo-chromosomes. The junctions of the fusion events between chromosomes 8 and 29 and between chromosomes 20 and 36 (Britto et al. 1998) were mapped (Fig. 1A). The former are fused at their 5′ ends (based on the homologous *L. major* chromosomes), while the latter are fused at the 3′ end of chromosome 36 and the 5′ end of chromosome 20. The gene models and functional annotation, based on predicted orthology, were transferred from *L. major* to the *L. mexicana* U1103 genome using the Rapid Annotation Transfer Tool (Otto et al. 2011), and thereafter manually annotated using codon bias and BLAST searches against the NCBI nr database as a guide for gene prediction.

The three other reference *Leishmania* genomes—*L. major* Friedlin, *L. infantum* JPCM5, and *L. braziliensis* M2904 (Table 1)—were resequenced on the Illumina Genome Analyser platform, and the iterative mapping algorithm iCORN (Otto et al. 2010) was used to correct single-base sequencing errors as well as small insertions or deletions, which have resulted in several hundred corrections (see Supplemental Tables S1, S2).

Gene content comparison between the four *Leishmania* species

Generation of a reference genome for *L. mexicana*, together with updated and refined *L. major* Friedlin, *L. braziliensis* M2904, and *L. infantum* JPCM5 reference genomes, has allowed a reevaluation of the number of genes that are differentially distributed between these *Leishmania* species. Using a combination of OrthoMCL (Li et al. 2003) and ACT (Carver et al. 2008) alignments, we have assembled an updated set of ortholog predictions for the four *Leishmania* reference genomes (Supplemental Table S3). We found only two unique genes (LmxM.14.0870, LmxM.31.2501) present in the *L. mexicana* U1103 genome, both of which encode predicted proteins of unknown function. LmxM.31.2501 contains a predicted kelch actin binding domain (PFAM:PF01344). LmxM.14.0870 is predicted to be a pseudogene in the other reference *Leishmania* species and is orthologous to an intact copy in *Trypanosoma brucei* (Tb927.7.4050). Genes with clear orthologs but containing internal stop codons or insertions or deletions resulting in frameshifts were annotated as pseudogenes. We now predict 19 *L. infantum*–specific genes, or paralogous groups, of which 15 encode proteins of unknown function, 14 are *L. major*–specific genes or paralogous groups, of which 13 encode proteins of unknown function, and one encodes a predicted PfpI peptidase (Eschenlauer et al. 2006). Sixty-seven are *L. braziliensis*–specific genes or paralogous groups, of which 54 encode proteins of unknown function (Fig. 1B). *L. braziliensis* M2904 also has the highest number of gene loss or pseudogene-formation events compared with the other *Leishmania* species in terms of orthologous groups that are absent in *L. braziliensis* or in which the *L. braziliensis* orthologs appear to be pseudogenes. In contrast, *L. mexicana* has the most unique losses and interestingly

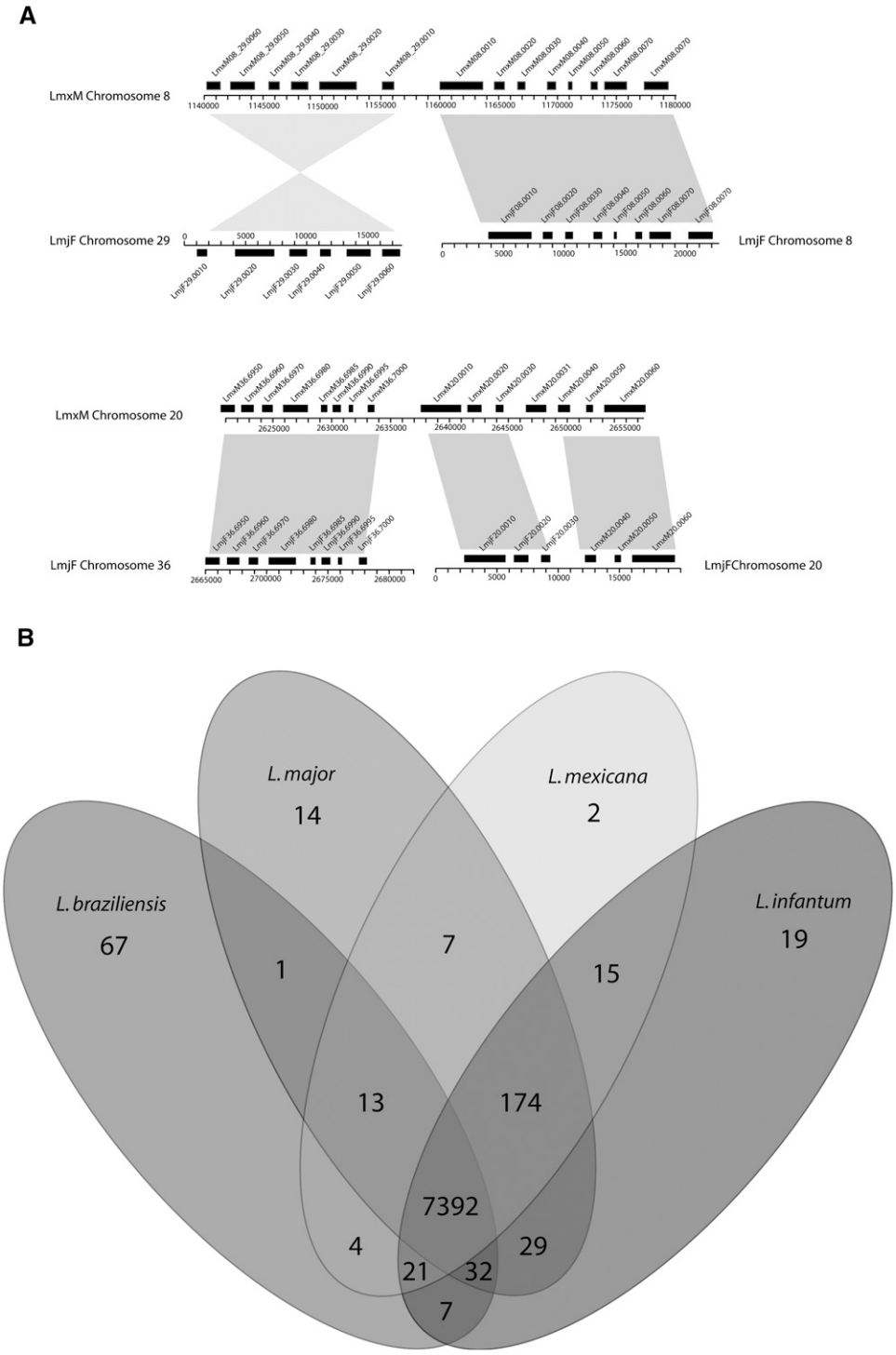

Figure 1. (*A*) Chromosome fusion events in *L. mexicana*. (*B*) Venn diagram showing number of conserved genes in *L. mexicana* U1103, *L. major* Friedlin, *L. infantum* JPCM5, and *L. braziliensis* M2904.

has more orthologous groups in common with *L. infantum* than any other *Leishmania* genome (see Supplemental Table S3). Twelve of these genes were previously predicted to be unique to *L. infantum* (Peacock et al. 2007; Zhang et al. 2008). Overall, the number of unique genes in each species is remarkably small in relation to the coding capacity of the genomes.

SNP analysis of *Leishmania* genomes

Illumina reads were used to call heterozygous SNPs in the reference genomes of *L. major* Friedlin (297 SNPs), *L. infantum* (629 SNPs), *L. braziliensis* (44,588 SNPs), and *L. mexicana* (12,531 SNPs) (Supplemental Table S2). The remarkably low level of heterozygosity

in *L. major* and *L. infantum* compared with *L. mexicana* and *L. braziliensis* could result from a higher degree of inbreeding within these two species relative to *L. braziliensis* and *L. mexicana*. Conversely, in *L. braziliensis* M2904, the high level of heterozygosity could in part be due to its triploidy (see below) and the supernumerary nature of many of its chromosomes, which, if stable, may have resulted in a redundancy of essential genes and allowed for a higher rate of neutral mutation than that of the typically diploid *L. major* Friedlin or *L. infantum* JPCM5 genomes. Recent population studies of *Leishmania* species have proposed inbreeding, natural selection, the Wahlund effect, or conversion on a genome-wide scale as possible causes of deficiency of heterozygous sites in strains of *L. braziliensis* (Rougeron et al. 2009) and *L. donovani* (Gelanew et al. 2010).

Unique genes are conserved in strains of the same species or species complex

To investigate whether the species-specific genes are present in different isolates of the same species or different species within a complex, Illumina sequencing was performed on a second *L. major* isolate (LV39) and a second *L. mexicana* isolate (M379). In addition, *L. donovani* BPK206/0, a field isolate from Nepal (Downing et al. 2011), and *L. donovani* LV9, a human strain isolated in Ethiopia in 1967 and subsequently maintained by animal passage (Bradley and Kirkley 1977; Stager et al. 2000), were also sequenced. The 19 *L. infantum*–specific genes appear conserved in *L. donovani* LV9 and are not clearly absent in *L. donovani* BPK282/0/c14 (Downing et al. 2011), despite an estimated divergence of 1 million years between *L. infantum* and *L. donovani*. One gene (*LinJ.19.1170*) was found to contain an internal stop codon in *L. donovani* BPK206/0. The two *L. major* strains, LV39 and Friedlin, cause similar disease progression in susceptible mouse strains but distinct differences when analyzed in an IL4-receptor-deficient host genetic background (Noben-Trauth et al. 2003). These strains were used recently to show that genetic exchange can occur in the sand fly vector, *Phlebotomus papatasi* (Akopyants et al. 2009). Thirteen of the 14 *L. major*–specific genes identified in the orthoMCL analysis are intact in LV39, the one exception being hypothetical protein LmjF.32.2470, which contains a premature stop codon. *L. mexicana* M379 contained intact copies of both unique genes found in *L. mexicana* U1103, neither of which contains any predicted SNPs.

Hence, species-specific genes in *Leishmania* are conserved in isolates from within the same complex (*L. donovani* vs. *L. infantum*) and between strains of the same species isolated from different geographical locations (e.g., *L. donovani* BPK206/0 from Nepal and *L. donovani* LV9 from Ethiopia) or similar geographical locations (e.g., *L. mexicana* U1103 from Guatemala and *L. mexicana* M379 from Belize). The majority of the *L. donovani* complex species-specific genes are uncharacterized, and for most, their transgenic expression in *L. major* failed to identify a phenotype correlating with visceral infection in mice (Zhang et al. 2008; Zhang and Matlashewski 2010). An exception was *LinJ.28.0340*, a *L. infantum*–specific gene apparently important for survival of axenic amastigotes (Zhang and Matlashewski 2010). The finding that unique genes are present in both *L. infantum* and *L. donovani* suggests that these sequences play an important role in the potential for visceralization associated with the *L. donovani* complex. However, other structural and functional components of the genome, such as gene copy number (Fig. 2) and differential gene expression (Depledge et al. 2009), are also likely to be important.

Gene copy number variation in *Leishmania*

The highly repetitive nature of tandem arrays is problematic for de novo genome assembly, leading to "collapsed" arrays of unknown length (Ivens et al. 2005), as verified with the HASP genes and their related copies in *L. braziliensis* (Depledge et al. 2009, 2010). We have used read depth coverage to estimate protein-coding gene copy numbers in the four reference genomes (Supplemental Table S4). Multicopy genes (tandem arrays) in this analysis are defined as genes of more than one copy that have the same OrthoMCL group identifier and are encoded on the same chromosome. To provide a confidence limit of >95% for the analysis, the most complete and refined of the *Leishmania* genomes, *L. major* Friedlin, was used to

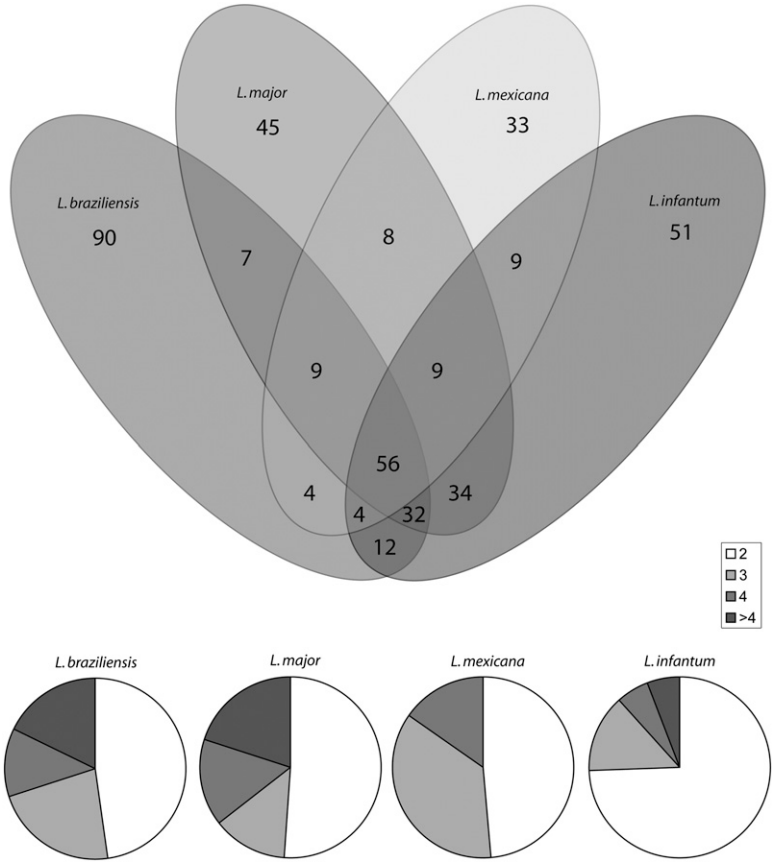

Figure 2. Variation in multicopy arrays in the *Leishmania* reference genomes. Venn diagram that shows the overlap of arrays present in the different species, by unique ortholog group identifier for ease of comparison. Pie charts indicate the proportion of unique arrays grouped by the number of genes.

Table 1. Summary of *Leishmania* genomes

	L. major	*L. infantum*	*L. braziliensis*	*L. mexicana*
Size	32,855,089	32,101,728	31,997,773	32,108,741
Chromosomes	36	36	35	34
Fold coverage (capillary)	N/A	5.1	7.0	6.7
Fold coverage (Illumina)	56	79	105	107
GC%	59.7	59.6	57.8	59.7
Predicted protein-coding genes	8412	8241	8357	8250

benchmark single- and dual-copy genes. In the *L. major* Friedlin genome (TriTrypDB version 2.4), our gene copy number analysis was in agreement for 99% of the single-copy and 81% of the dual-copy genes annotated in the genome (Supplemental Methods). The total number of arrays varies between species: 200 in *L. major* Friedlin, 207 in *L. infantum* JPCM5, 214 in *L. braziliensis* M2904, and 132 in *L. mexicana* U1103. The number of gene copies within each array also varies between species (Fig. 2).

Only 56 protein-coding genes were found to be multicopy in all four species, although the number of copies varied considerably (Table 2; Supplemental Table S5). These include well-characterized surface proteins (PSA2/GP46) (Devault and Banuls 2008), amastin (Jackson 2010), GP63/leishmanolysin (Gomez et al. 2009; Halle et al. 2009), structural proteins (alpha- and beta-tubulin) (Jackson et al. 2006), 40S ribosomal protein (EF1A), and chaperones (HSP70 and HSP83). In *L. major* Friedlin, the tandem array containing the largest number of gene copies was a class I nuclease-like protein, with sequence similarity to externally orientated surface membrane enzymes with 3′-nucleotidase activity (Yamage et al. 2000). This particular tandem array was one of 22% of tandem arrays that were found to be unique to *L. major* Friedlin. Twenty-five percent of tandem arrays were unique for *L. mexicana*, 25% for *L. infantum*, and 42% for *L. braziliensis*. The increase in gene dosage that arises from tandem duplication may allow higher transcript levels for multicopy, possibly stage-regulated genes. Among the largest species-specific tandem arrays are genes encoding proteins of unknown function (see Table 3; Supplemental Table S6) (OG4_18490, OG4_17568 in *L. major*; OG4_21149 in *L. mexicana*; OG4_15862 in *L. infantum*; OG4_28794 in *L. braziliensis*). Most of the tandem arrays that are unique to each species are hypothetical proteins, so further investigation is warranted to examine these particular hypothetical genes for their role in the biology of *Leishmania*.

Chromosome copy number variation in *Leishmania*

One feature of the massively parallel sequencing achieved with the Illumina analyzer is that it provides unprecedented read depth coverage across all the *Leishmania* chromosomes. Analysis of median read depth for the *L. major* Friedlin genome (Fig. 3A) shows that all chromosomes except one (chromosome 31) have an even read depth, indicating that the chromosomes within the population of cells are disomic. Chromosome 31 has a read depth of greater than twofold, indicating that this chromosome is at least tetrasomic, as has been reported previously (Akopyants et al. 2009). The increase in read depth is not due to amplification of one specific area of the chromosome, because the read depth coverage is even along the whole of both disomic and tetrasomic chromosomes (Supplemental Fig. S1).

Chromosome copy number analysis revealed large differences for eight strains and species of *Leishmania* (Fig. 3; Table 4). Nine chromosomes in *L. infantum* JPCM5, three in *L. mexicana* U1103, and one in *L. mexicana* M379 have trisomic read depth. Two chromosomes are supernumerary in *L. braziliensis* M2904, while 10 chromosomes in *L. mexicana* U1103 and two in *L. mexicana* M379 appear to be of intermediate read depth, being neither disomic nor trisomic. These chromosomes with "intermediate" read depth may be a mixture of individual cells within a population with monosomic, disomic, and trisomic chromosomes (Sterkers et al. 2011), resulting in chromosome copy number mosaicism in the sampled population. This effect appears to be more prevalent among the smaller chromosomes (Fig. 3). This is unlikely to be due to the loss of partial fragments from duplicated chromosomes, because read depth does not indicate any deletion or amplification of large regions across disomic and supernumerary chromosomes (Supplemental Fig. S2). *L. mexicana* U1103 chromosomes 1 and 3 have a read depth coverage that indicates that the population as a whole is less than disomic, potentially including some monosomic cells. The effect is also seen in *L. braziliensis* M2904 with two chromosomes that are neither trisomic nor tetrasomic. The two *L. major* strains have a single supernumerary chromosome, 31, the only chromosome that is supernumerary in all strains and species (Fig. 3A,E; Supplemental Table S7). Only 12 chromosomes are disomic in all strains and species analyzed—chromosomes 3, 10, 11, 18, 19, 21, 24, 28, 30, 32, 34, and 36 (*L. major* chromosome numbering).

To provide further evidence for mixed chromosomal ploidy, approximate distributions of base frequencies for each predicted heterozygous SNP in *L. mexicana* U1103 and *L. braziliensis* were plotted (Fig. 4; Supplemental Fig. S3). For the purpose of cross-comparison, chromosomes were grouped according to their allele frequency profiles, and allele frequency counts were normalized against the total allele counts for the chromosome. In *L. mexicana*, the distribution of base frequencies for predicted heterozygous sites, plotted for all chromosomes, gives four distinct distributions. Twenty-one chromosomes show a distinct peak at a frequency of 0.5, indicative of disomic chromosomes (Fig. 4A,C). Eleven chromosomes show distinct peaks surrounding frequencies of 0.33 and 0.66, suggestive of trisomic chromosomes, while chromosome 30 shows peaks at 0.25, 0.5, and 0.75, indicative of a tetrasomic chromosome. For 32 of the 34 chromosomes, these data match those predicted from read depth analysis (Fig. 3). The two exceptions are chromosomes 5 and 16, which are predicted to be disomic by allele frequency distribution (Fig. 4B), but supernumerary by read depth analysis (Fig. 3B). We hypothesize that 21 of the 34 chromosomes in *L. mexicana* are stable within the population, so that they accumulate SNPs on disomic alleles. In contrast, supernumerary alleles of chromosomes 5 and 16 are likely to be recent events, because they have not accumulated mutations at frequencies that would suggest that these occurred on one of the tetrasomic chromosomes. For the 11 trisomic chromosomes, it is not possible to determine whether these duplications are recent or ancestral, because even a recent duplication of a single chromosome leading to trisomy would result in previously disomic frequencies appearing trisomic. Nevertheless, the trisomic plots and the clear absence of frequencies resembling disomy or tetrasomy demonstrate that all of these chromosomes are trisomic in this sample rather than a combination of disomic and tetrasomic states. This method, in combination with read depth coverage, can be used to discriminate between recent and established aneuploidy in the case of tetrasomic chromosomes, or other chromosomes that exist as multiples of 2*n*. The same analysis was performed on

Table 2. Multicopy arrays with the highest number of gene copies in each species

OrthoMCL ID	Chr	Description	Number in reference	Haploid number
		***L. major* (Friedlin)**		
OG4_64465	30	Class I nuclease-like protein	4	28
OG4_19653	19	ATG8B	9	25
OG4_18490	12	Hypothetical protein, conserved	12	24
OG4_12080	34	Amastin	23	23
OG4_10107	17	Elongation factor 1A	7	21
OG4_14254	8	Amastin	16	20
OG4_54550	12	Promastigote surface antigen protein 2	5	19
OG4_18012	12	Surface antigen protein, putative	14	17
OG4_10079	33	beta-tubulin	16	16
OG4_17568	32	Hypothetical protein, conserved	17	15
		***L. mexicana* (U1103)**		
OG4_10176	10	GP63, leishmanolysin	5	13
OG4_10079	8	beta-tubulin	2	7
OG4_12080	33	Amastin	3	6
OG4_10079	32	beta-tubulin	2	6
OG4_10075	13	alpha-tubulin	2	6
OG4_26438	**9**	ATG8C	2	5
OG4_19653	19	ATG8B	2	5
OG4_21149	**14**	Hypothetical protein, conserved	2	5
OG4_11934	29	Cysteine peptidase	2	4
OG4_10355	25	Eukaryotic initiation factor 5A	2	4
		***L. infantum* (JPCM5)**		
OG4_12080	34	Amastin	8	27
OG4_64460	34	Amastin	1	15
OG4_10176	10	GP63, leishmanolysin,	5	15
OG4_10107	**17**	Elongation factor 1A	7	12
OG4_14440	36	Glucose transporter, LMGT1	3	10
OG4_64457	29	Amastin	4	9
OG4_10075	13	alpha-tubulin	2	9
OG4_12706	10	Folate/biopterin transporter, putative	8	9
OG4_12080	**8**	Amastin	4	8
OG4_19653	19	ATG8B	3	8
		***L. braziliensis* (M2904)**		
OG4_24845	8	Amastin	6	36
OG4_10176	10	GP63, leishmanolysin	29	31
OG4_24845	20.1	Amastin	5	26
OG4_12080	20.1	Amastin	10	16
OG4_10079	33	beta-tubulin	3	15
OG4_10075	13	alpha-tubulin	2	15
OG4_14254	8	Amastin	4	13
OG4_10088	33	HSP83	3	12
OG4_28794	4	Hypothetical protein, conserved in *Leishmania*	4	12
OG4_12218	34	NADH-dependent fumarate reductase	4	12

Chromosomes shown in bold are supernumerary.

L. braziliensis, which shows that 30 of 35 chromosomes are clearly trisomic (Fig. 4E), three are tetrasomic (chromosomes 4, 5, 29) (Fig. 4F), and one hexasomic (chromosome 31) (Fig. 4G), closely matching read depth analysis (Fig. 3D). One chromosome (chromosome 14) has an ambiguous profile closely resembling that of the tetrasomic chromosomes (data not shown). Taken together, these data indicate that *L. braziliensis* strain M2904 is primarily triploid but contains several tetrasomic chromosomes, and intriguingly six copies of chromosome 31.

Real-time PCR was used to further validate chromosome copy number determinations (Supplemental Fig. S4A). Single-copy genes from *L. donovani* BPK206/0 were selected from representative disomic and supernumerary chromosomes. The real-time PCR cycle threshold (C_T) values confirmed that chromosomes 14 and 36 are disomic, chromosomes 15 and 33 are trisomic, and chromosome 8 is tetrasomic. These data match closely the chromosomal ploidy values predicted by read depth analysis (Fig. 3). To test if total DNA content varies between species in the same proportion as chromosome copy number, FACS analysis was performed (Supplemental Fig. S4B). The 2N DNA content for *L. major* Friedlin was used as a baseline and compared with the other species. *L. major* LV39 is identical to *L. major* Friedlin, whereas an increase in DNA content is observed for the 2N peak of *L. infantum* JPCM5 (110%) and *L. mexicana* U1103 (115%), which matches closely the increased DNA content predicted from calculating the total megabases of DNA for each species, based on the chromosome copy number shown by read depth coverage (Table 4). *L. braziliensis* M2904 has 160% of the DNA content of *L. major* Friedlin, indicating that *L. braziliensis* M2904 is triploid, which is confirmation for the data obtained with analysis of distribution of base frequencies for predicted heterozygous sites (Fig. 4). In comparison to *L. major* Friedlin, *L. donovani* BPK206/0 has 110% DNA content, and *L. donovani* LV9 has 102% DNA content, again matching closely the increased DNA content predicted from read depth coverage (Table 4).

Further validation of chromosome copy number was obtained by gene-specific deletions in *L. infantum*. *LinJ36.0640* encodes an SEC14-like protein found on the disomic chromosome 36, and two rounds of gene deletion resulted in the generation of a SEC14-like null mutant, as expected (Supplemental Fig. S4C). SEC14-like null mutants had the same growth characteristics in culture and infectivity dynamics to macrophages as wild-type *L. infantum*. *LinJ.36.0640* was originally identified as an *L. infantum*–specific gene (Peacock et al. 2007), but has now been identified in *L. mexicana*. In contrast, *LinJ.31.3030* encodes a phosphatase on the tetrasomic chromosome 31, and, in this case, four rounds of gene deletion were required to generate a phosphatase null mutant. Overall, these experiments validate using read depth coverage as a robust parameter to assess chromosome copy number variation within a *Leishmania* population.

Changes in ploidy and chromosome copy number have been reported in various *Leishmania* species following genetic manipulation of essential genes (Cruz et al. 1993; Hassan et al. 2001), during in vitro growth (Martinez-Calvillo et al. 2005), after genetic exchange (Akopyants et al. 2009), and following drug selection in vitro (Leprohon et al. 2009). *L. major* has 36 chromosomes (Wincker et al. 1998), and our data suggest that within a population, disomy for most chromosomes is common in both *L. major* Friedlin and *L. major* LV39. *L. major* Friedlin chromosome 1 has previously been reported as trisomic (Sunkin et al. 2000; Sterkers et al. 2011). This, however, was apparently not the case for the *L. major* Friedlin isolate used in this study, perhaps due to variations in culture methods, time in culture, growth conditions used in different laboratories (although the parasites used in all these analyses were originally derived from the same stock), and method of analysis (individual cell vs. cell populations). Chromosome 31, which has been reported to be supernumerary in *L. major* (Akopyants et al. 2009), was the only chromosome identified in *L. major* Friedlin in this study that was also supernumerary in all species and isolates analyzed,

Table 3. Unique multicopy arrays in each species, defined by ortholog group

OrthoMCL ID	Chr	Description	Number in reference	Haploid number
		***L. major* (Friedlin)**		
OG4_64465	30	Class I nuclease-like protein	4	28
OG4_54550	12	Promastigote surface antigen 2	5	19
OG4_18012	12	Surface antigen protein, putative	14	17
OG4_17568	32	Hypothetical protein, conserved	17	15
OG4_17051	34	Amastin	21	12
OG4_35275	12	Hypothetical protein	5	9
OG4_14297	34	Quinonoid dihydropteridine reductase	7	9
OG4_10345	12	Surface antigen protein, putative	4	7
OG4_83438	19	Hypothetical protein	3	4
OG4_10384	36	Hypothetical protein, conserved	2	4
		***L. mexicana* (U1103)**		
OG4_10442	23	Coronin, putative	1	3
OG4_54460	33	Hypothetical protein, conserved	2	3
OG4_68082	33	D-Isomer specific 2-hydroxyacid dehydrogenase- like protein	1	3
OG4_35685	34	Hypothetical protein, conserved	1	3
OG4_112308	11	ABC transporter-like protein	2	3
OG4_32333	8	Hypothetical protein, conserved	1	2
OG4_90898	33	D-Isomer specific 2-hydroxyacid dehydrogenase- like protein	1	2
OG4_83429	34	Hypothetical protein	1	2
OG4_38816	34	Hypothetical protein, conserved	1	2
OG4_83402	8	Hypothetical protein	1	2
		***L. infantum* (JPCM5)**		
OG4_64457	29	Amastin	2	5
OG4_64458	29	Histone H2A, putative	3	4
OG4_36933	15	Hypothetical protein	1	3
OG4_112234	**9**	Microtubule-associated protein-like	2	3
OG4_11537	27	Hypothetical protein, conserved	1	3
OG4_51380	29	Hypothetical protein, conserved	1	3
OG4_35505	27	Hypothetical protein, conserved	1	2
OG4_11081	34	Lipophosphoglycan biosynthetic protein (LPG2)	1	2
OG4_32453	32	Hypothetical protein, conserved	2	2
OG4_112225	2	Hypothetical protein, unknown function	1	2
		***L. braziliensis* (M2904)**		
OG4_10748	16	Hypothetical protein	6	10
OG4_50672	4	Surface antigen-like protein	3	9
OG4_14950	30	TATE DNA transposon	8	9
OG4_63547	8	β-Tubulin	4	9
OG4_26451	20.1	Hypothetical protein, conserved	2	7
OG4_31940	18	Hypothetical protein, conserved	1	6
OG4_83308	20.1	Amastin	2	6
OG4_112181	10	GP63, leishmanolysin	2	6
OG4_112174	2	Repeat gene hypothetical protein	2	5
OG4_47872	33	Expression-site associated gene (ESAG3)	3	5

Chromosomes shown in bold are supernumerary.

including the equivalent homologous chromosome 30 in *L. mexicana* (numbered differently due to the fusion of chromosomes 8 and 29). A recent FISH analysis proposing chromosomal mosaicism in individual *L. major* cells (Sterkers et al. 2011) is in broad agreement with our results at the population level in a variety of *Leishmania* strains and species (Fig. 3). The details of chromosomal mosaicism detected in specific chromosomes of *L. major*, however, differ in that we did not detect a high level of monosomy in chromosome 2 or trisomy in chromosomes 1, 5, and 17 (Sterkers et al. 2011).

Multicopy genes are found preferentially on disomic chromosomes

The distributions of the multicopy genes on each chromosome in the *Leishmania* reference genomes are shown in Figures 5 and 6. The bars (Fig. 6) are scaled by chromosome size and illustrate the relative presence of multicopy arrays across the genomes. The results, which show that variation exists between strains and species, are consistent with the finding that copy number variation is a major influence in speciation (Lynch and Conery 2003). They also highlight the differences in the striking distribution of multicopy arrays between the disomic (black) and supernumerary chromosomes (highlighted in gray). By way of illustration, in *L. mexicana* U1103, there are no duplications for the 309 genes on supernumerary chromosome 30, whereas there are 12 duplications within the 498 genes on disomic chromosome 34 (Fig. 5; Supplemental Fig. S5). The bias in this distribution was analyzed using a Monte Carlo simulation, which showed that there is a significant bias for the presence of gene arrays on disomic chromosomes in all of the species tested, including *L. major* Friedlin ($p = 7.45 \times 10^{-4}$), *L. infantum* JPCM5 ($p < 1 \times 10^{-6}$), *L. mexicana* U1103 ($p < 1 \times 10^{-6}$), *L. braziliensis* ($p = 7 \times 10^{-6}$), *L. mexicana* M379 ($p = 2.61 \times 10^{-3}$), *L. donovani* BPK206/0 ($p < 1 \times 10^{-6}$), and *L. donovani* LV9 ($p = 9.40 \times 10^{-5}$). These data suggest that some of these disomic chromosomes may have persisted as non-supernumerary because of fitness constraints, and we propose that this is gene-dosage-related. Gene-dosage sensitivity is known to be important in selection of copy number changes (Schuster-Bockler et al. 2010), and the scale of duplications has also been shown to influence fitness and the likelihood of being retained (DeLuna et al. 2008). There may, therefore, be selection against whole-chromosome duplications for those that have a higher proportion of dose-sensitive genes in *Leishmania* species.

We went on to analyze potential gene family biases in the multicopy gene families and found only a few domains significantly over-represented in the genes present on the supernumerary chromosomes. These genes fall into four domain groups within three biological classes (Supplemental Table S8): signaling/kinases (PF00069; protein kinase, PF00400; WD domain), ubiquitin (PF00240), and a Kinesin motor domain (PF00225). Since the increase of dose of genes containing these domains is the result of whole-chromosome duplications, the over-representation of these groups could suggest that these are non-toxic passengers to the chromosome duplications, rather than having been selected for by the duplication event; or equally the selection for chromosome duplication could be facilitated by the over-representation of these gene classes. These findings are consistent with previous studies of the functional biases in the types of genes that are able to survive duplications; in bacteria, yeast, insects, and man, higher numbers of genes coding for transcription factors, protein kinases, and certain classes of enzymes and transporters have been previously observed (Taylor and Raes 2004). The genes that have increased copy numbers due to gene duplication have a much longer and broader list of Pfam domains (Supplemental Tables S6, S8), indicating that there might be individual evolutionary pressures for the selection of each of these genes, rather than because they encode a particular type/class of product.

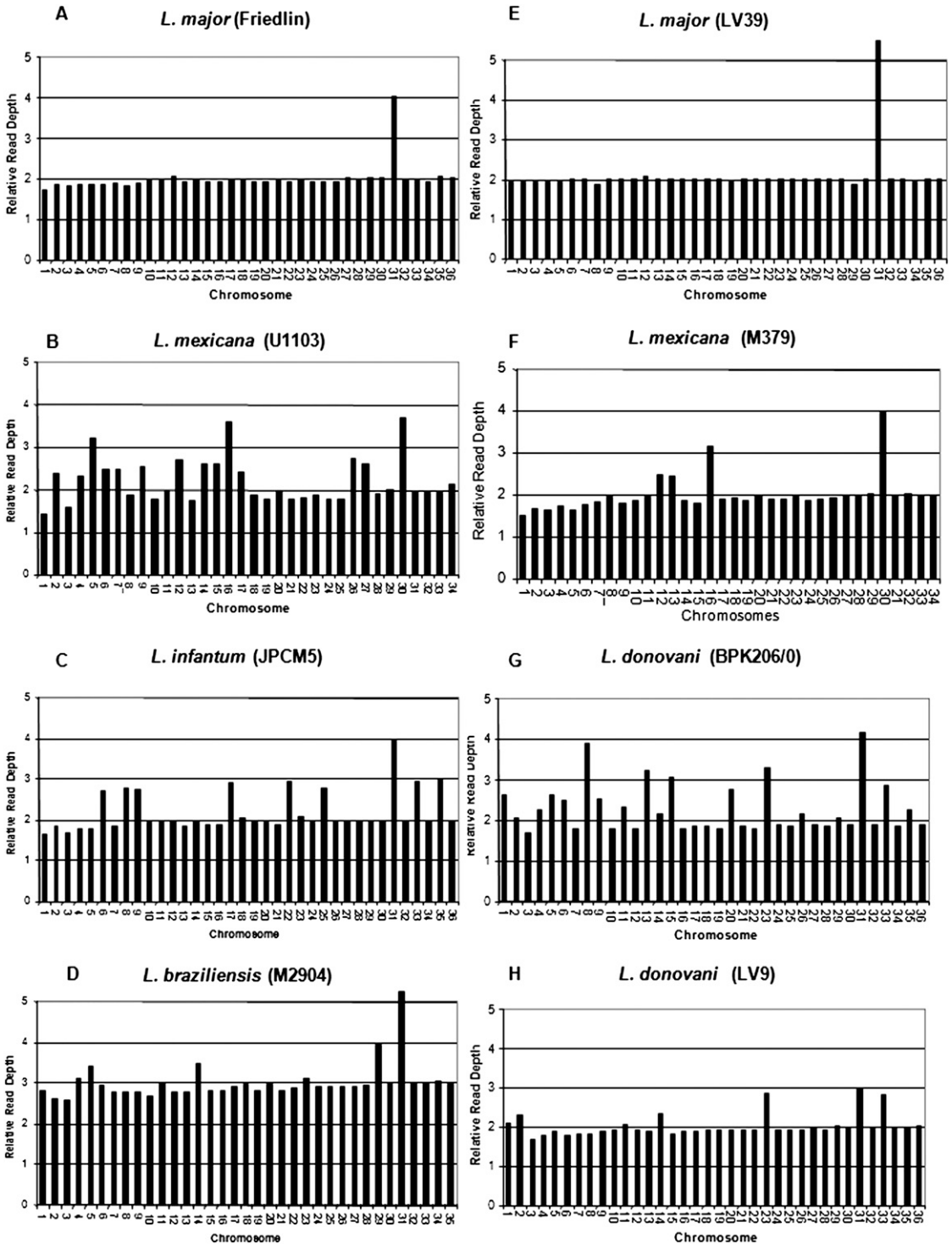

Figure 3. Chromosome copy number variation in *Leishmania* reference genomes. Read depth was scaled to give a value of 2 for disomic chromosomes. Median read depth over all chromosomes in the genome is indicated in brackets. (*A*) *L. major* Friedlin (52); (*B*) *L. mexicana* U1103 (106); (*C*) *L. infantum* JPCM5 (70); (*D*) *L. braziliensis* M2904 (80); (*E*) *L. major* LV39 (31); (*F*) *L. mexicana* M379 (87); (*G*) *L. donovani* BPK206/0 (86); (*H*) *L. donovani* LV9 (59).

In summary, these data lead us to hypothesize that *Leishmania* might increase mRNA levels in the absence of regulated promoter activity using two mechanisms: first, via gene duplications on disomic chromosomes, which can result in the generation of multicopy arrays of identical or near identical genes; second, gene copy number may increase through the formation of supernumerary chromosomes. While changes in gene expression in drug-resistant *Leishmania* have been associated with supernumerary chromosomes (Ubeda et al. 2008; Leprohon et al. 2009), our data on genome content across *Leishmania* populations, combined with data of

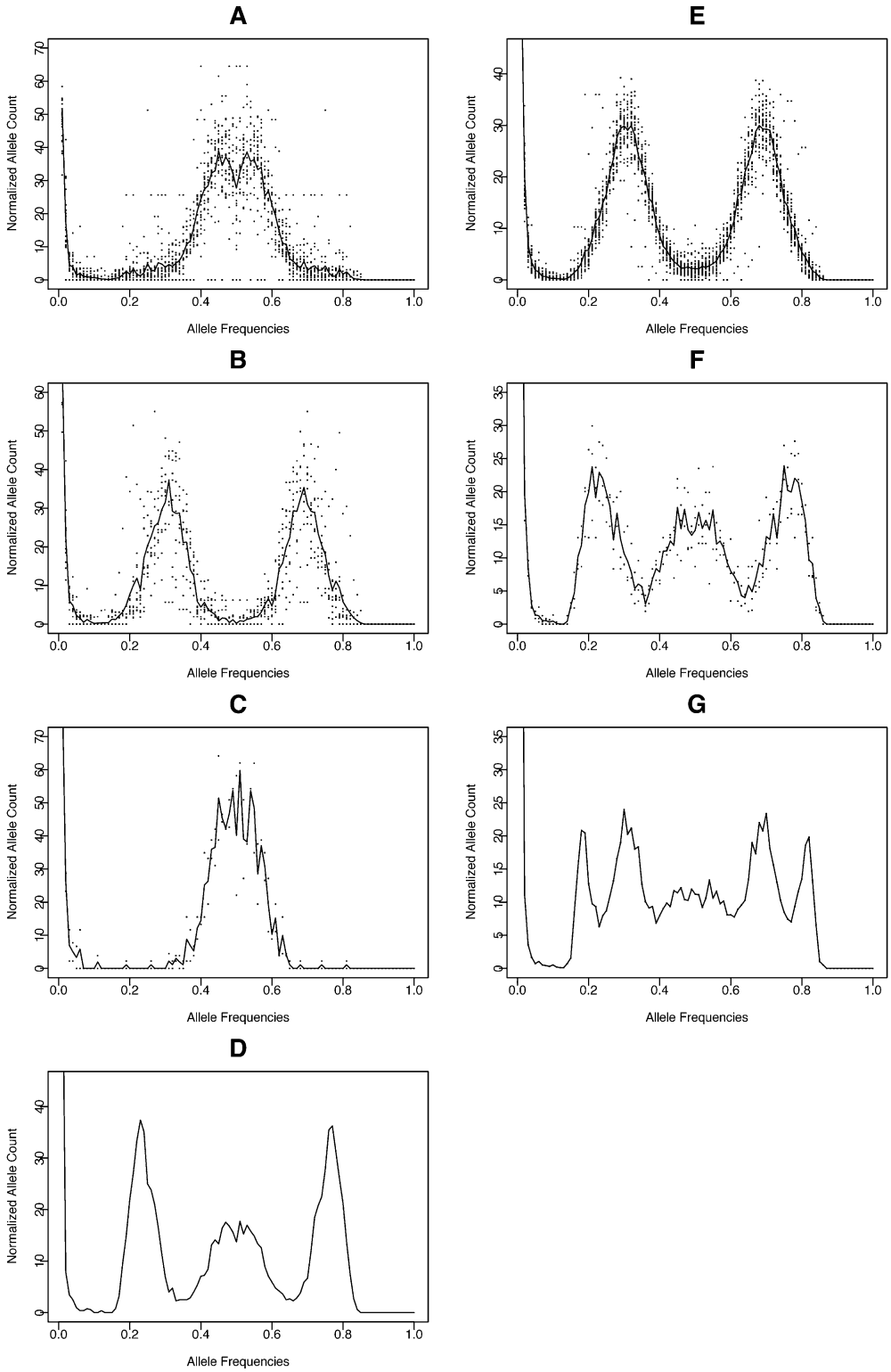

Figure 4. Distribution of normalized allele frequencies according to inferred ploidy for *L. mexicana* and *L. braziliensis* chromosomes. (*A*) *L. mexicana* disomic chromosomes: Chr 1, 3, 8, 10, 11, 13, 18–25, 28, 29, 31–34. (*B*) *L. mexicana* trisomic chromosomes: Chr 2, 4, 6, 7, 9, 12, 14, 15, 17, 26, 27. (C) *L. mexicana* tetrasomic chromosomes with disomic base frequency profiles: Chr 5, 16. (*D*) *L. mexicana* tetrasomic chromosome: Chr 30. (*E*) *L. braziliensis* trisomic chromosomes: Chr 1–3, 6–13, 15–28, 30, 32–35. (*F*) *L. braziliensis* tetrasomic chromosomes: Chr 4, 5, 29. (*G*) *L. braziliensis* hexasomic chromosome: Chr 31.

Figure 5. The genomic distribution of multicopy arrays across the chromosomes of *Leishmania* species. The multicopy genes are colored by number of copies: (white) 2, (yellow) 3, (orange) 4, and (red) >4. Chromosomes with the normal somy for that species are shown in black; for *L. major*, *L. infantum*, and *L. mexicana*, this is disomy, and for *L. braziliensis*, this is trisomy. (Red) The supernumerary chromosomes. The chromosomes are shown to scale; however, arrays are shown relative to the center point of the array, but are not to scale. (*A*) *L. braziliensis* M2904 (231); (*B*) *L. major* Friedlin (207); (*C*) *L. mexicana* U1103 (134); (*D*) *L. infantum* JPCM5 (216). The total number of multicopy genes for each species is shown in brackets.

Table 4. Summary of *Leishmania* chromosomes

Species	Strain	Number of chromosomes	Size (Mb) (sequenced, haploid)	Size (Mb) (adjusted for chromosome number)	Intermediate chromosomes	Trisomic chromosomes	Tetrasomic chromosomes or higher	% DNA increase relative to *L. major* Friedlin
L. major	Friedlin	36	32.8	67.8	—	—	31	0%
L. major	LV39	36	32.8[a]	70.8	—	—	31	+4.5%
L. infantum	JPCM5	36	32.1	72.8	—	6, 8, 9, 17, 22, 25, 33, 35	31	+7.5%
L. donovani	BPK206/0	36	32.1[a]	72.5	1, 5, 6, 9, 11, 35	13, 15, 20, 23, 33	8, 31	+7%
L. donovani	LV9	36	32.1[a]	66.5	2, 14	23, 31, 33	—	−2%
L. braziliensis	M2904	35	32.0	66.4	5, 14	29	31	−2% (triploid)
L. mexicana	U1103	34	32.1	67.8	2, 4, 6, 7, 8, 12, 14, 15, 17, 26, 27	5, 26	16, 30	0%
L. mexicana	M379	34	32.1[a]	ND	12,13	16	30	ND

[a]Estimate based on resequencing projects. All chromosomes disomic unless indicated.
(ND) Not determined.

others on the genome content of individual cells (Sterkers et al. 2011), suggest that aneuploidy allows rapid generation of diversity in *Leishmania* parasites growing normally, as well as in response to stress. For genes on supernumerary chromosomes, where dosage effects might have a detrimental impact on parasite growth and survival, mRNA levels could be regulated through an increase in mRNA degradation. Alternatively, the parasites might engage protein degradation pathways (Besteiro et al. 2007) in order to correct protein stoichiometry imbalances caused by aneuploidy. In yeast, mutations that promote fitness of aneuploid cells identified ubiquitin-proteasomal degradation as a mechanism to suppress the adverse affects of aneuploidy (Torres et al. 2010). Of note, one of the functional gene classes that are over-represented on supernumerary chromosomes in *Leishmania* is ubiquitin.

Our study shows for the first time the full scale of aneuploidy in the *Leishmania* genus. Aneuploidy appears to arise frequently and is well tolerated by *Leishmania*, and chromosome copy numbers can vary considerably between strains and species from diverse geographical regions, including recent isolates (see also Downing et al. 2011). It remains to be determined if the genotype is stable in both laboratory-adapted strains and in natural populations, but our allele frequency analysis of *L. mexicana* U1103 suggests that 13 of 15 supernumerary chromosomes are established in the population, indicating that aneuploidy does not have a high fitness cost. Also, the finding that the aneuploid state of the *L. major* Friedlin population analyzed in this study is different from another published analysis (Sunkin et al. 2000) indicates that aneuploidy is not completely stable. In *Saccharomyces cerevisiae* aneuploidy can provide a strong selective advantage under environmental stress (Rancati et al. 2008), as well as phenotypic variation conferred by changes in the proteome (Pavelka et al. 2010). *Leishmania* has a complex life cycle that involves insect and mammalian hosts, with many species being zoonotic. Complex chromosomal copy number changes may be well tolerated in *Leishmania* parasites, with their predominantly asexual replication. Furthermore, genome plasticity allows the parasite to adapt to different environments, including survival in a variety of mammalian hosts and under drug selection (Leprohon et al. 2009). Aneuploidy in *Leishmania* is likely to arise through unlicensed replication and/or mitotic non-disjunction, rather than through sexual recombination.

Next-generation sequencing has proved a powerful tool for characterizing gene and chromosome copy number variations in *Leishmania*. Recent advances for multiplexing of samples make rapid, extensive, and cheap genome analysis possible and will allow extensive analyses on genome plasticity in the large number of strains and species currently available globally (Lukes et al. 2007; Rougeron et al. 2009), as well as strains isolated directly from patients and vectors.

Methods

Parasites

L. mexicana U1103, (MHOM/GT/2001/U1103, clone 25), *L. donovani* BPK206/0 (MHOM/NE/2003/BPK206/0 clone 10), *L. donovani* LV9 (MHOM/ET/67/HU3), *L. infantum* JPCM5 (MCAN/ES/98/LLM-877), *L. major* Friedlin (MHOM/IL/81/Friedlin), *L. major* LV39 (MRHO/SU/59/P), *L. mexicana* (MNYC/BZ/62/M379), and *L. braziliensis* M2904 (MHOM/BR/75M2904) were grown in modified Eagle's medium (HOMEM medium) supplemented with 10% heat-inactivated Fetal Calf Serum (PAA Gold, PAA) at 25°C as described previously (Hilley et al. 2000).

Sequencing

L. mexicana reference genome U1103 was sequenced using a whole-genome shotgun strategy with Sanger methodology as described previously for *L. infantum* and *L. braziliensis* (Peacock et al. 2007). All *Leishmania* strains and species were sequenced by Illumina Genome Analyzer II.

Bioinformatics analysis and annotation

L. mexicana sequence reads were assembled using *phrap* (http://phrap.org). Automated in-house software (Auto-Prefinish) was used to identify primers and clones for additional sequencing to close physical and sequence gaps by oligo-walking. Manual base checking and finishing was carried out using Gap4 (http://www.mrc-lmb.cam.ac.uk/pubseq/manual/gap4_unix_1.html). The assembled contigs were iteratively ordered and orientated against the *L. major* genome sequence (Ivens et al. 2005) using the ABACAS software (Assefa et al. 2009). The manually curated annotation of the *L. major* reference genome was transferred to the assembled *L. mexicana* genome on the basis of BLASTP matches and positional information using custom Perl scripts. Subsequently, gene structure and functional annotation were manually inspected and further edited, where appropriate, using the Artemis and ACT software (Carver et al. 2008), as previously detailed (Peacock et al. 2007).

The *L. major* Friedlin, *L. infantum* JPCM5, and *L. braziliensis* M2904 reference genomes were corrected using high depth Illumina coverage. iCORN (Otto et al. 2010) was used to iteratively map and correct sequence disagreements between the Illumina reads and the

reference sequence. Orthologous genes were predicted by running OrthoMCL v 1.4 (Li et al. 2003) on the earlier versions of the reference genomes (performed against the GeneDB database May 2009). Clusters that did not contain all four *Leishmania* species, and singleton clusters were manually inspected using the Artemis Comparison Tool (ACT), to see if orthologous genes could be manually assigned to the cluster.

SNP predictions were generated using a combination of SSAHA2 (Ning et al. 2001) to map the reads, Samtools (Li et al. 2009) to generate the variant predictions, and in-house Perl scripts to further filter the results. SNP-dense regions (more than three SNPs in a 7-bp window) were excluded, as were SNPs within 100 bp of contig edges. SNPs with consensus quality and base quality scores ≥40, mapping quality scores ≥25, coverage ≥10 reads, and less than twice the median coverage of the chromosome were retained. Finally, SNPs in the bin contigs of *L. infantum*, *L. mexicana*, and *L. braziliensis* were excluded due to uncertainty of median read coverage for these regions.

Allele frequency distribution plots

Allele frequencies for *L. braziliensis* M2904 and *L. mexicana* U1103 were inferred from filtered Samtools pile-up results. For each predicted heterozygous site, the proportion of each of the four sites over the total read depth for the site was determined and rounded to the second decimal place. Allele frequencies were binned into categories from 0.1 to 1.0, and counts for each allele frequency were plotted in R for each chromosome. In order to facilitate cross-comparison of chromosomes, counts for each category were normalized by dividing the count for each allele frequency category by the sum of all frequency category counts for that chromosome. Arithmetic means were calculated for each "somy" group and plotted in R, along with scatterplots of normalized allele frequency counts for every chromosome in this group. Plots were grouped according to their number of peaks. Chromosome 14 from *L. braziliensis* did not fit any of these profiles and was excluded from the analysis.

Chromosome read depth analysis

Reads were mapped to the appropriate reference genomes using MAQ (Li et al. 2008) version 0.7.1 (http://maq.sourceforge.net/), under the guidance of a custom Perl script. Read sets were parsed into smaller paired sets of 2.5×10^6 reads or less and converted

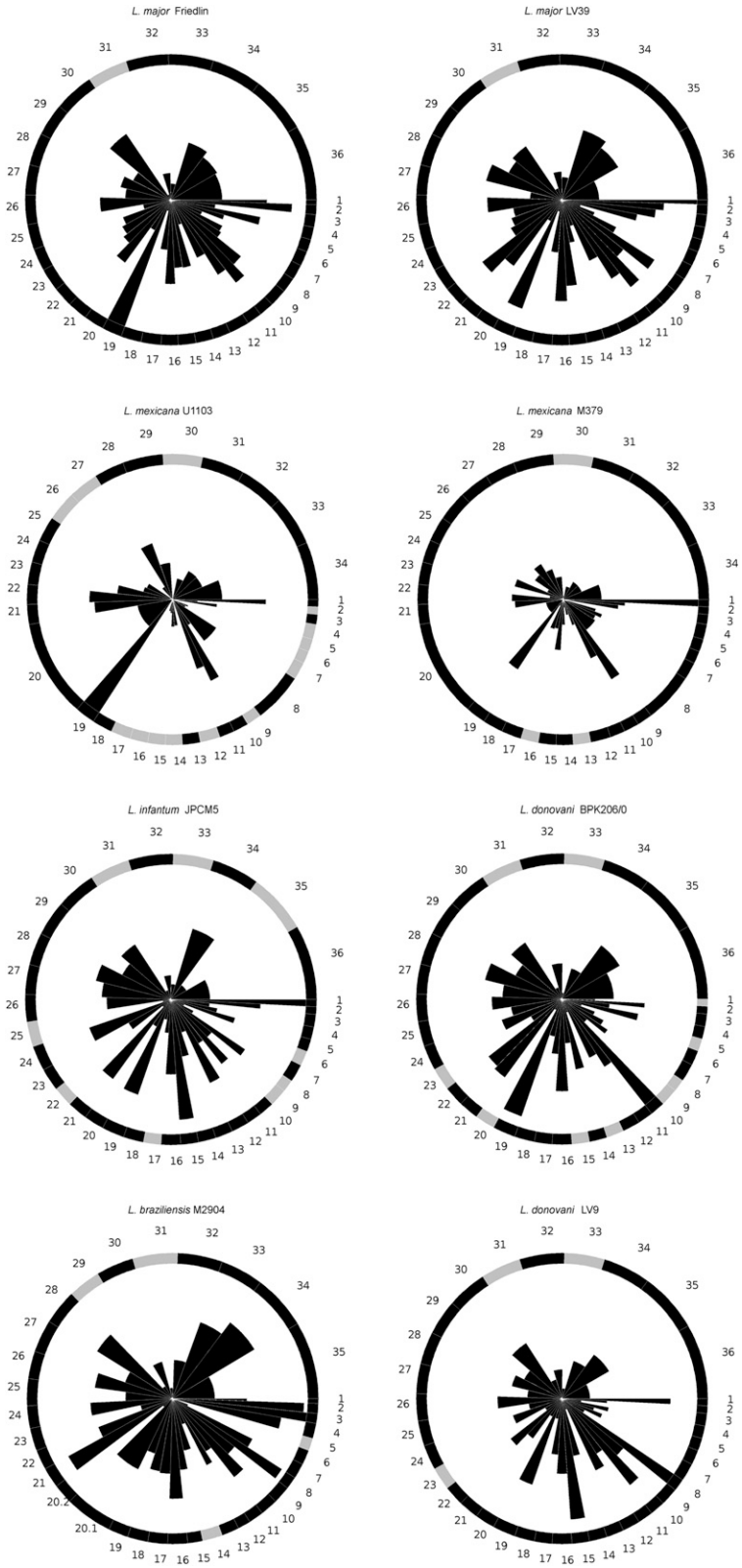

Figure 6. Mapping multicopy arrays onto disomic and supernumerary chromosomes across the *Leishmania* species. The outer circle shows the chromosomes colored as either disomic (black) or supernumerary (gray), as calculated using the average read depth across the genes on the chromosome. The inner circle is scaled by the number of multicopy arrays present per megabase on that chromosome. (*L. braziliensis* is triploid so the normal state for its chromosomes is trisomic.)

into binary format. The number of bases mapping to each position in each chromosome was recorded and used to determine the total number of read bases mapping to each chromosome and the median read depth for each chromosome. Observing that a majority of chromosomes displayed similar median read depths, and interpreting this as a nominal "ploidy" for the cells, a within-genome normalization was performed by setting the average of the read depth of the four longest disomic chromosomes to 2. The read depth for each chromosome was subsequently normalized to this value.

Gene read depth analysis

Alignment of the reads to the genome was performed using Bowtie (Langmead et al. 2009). To increase the robustness of the analysis, only reads that aligned to single regions of the genome and the best alignments (by quality score) were used. Cufflinks (Roberts et al. 2011) was used to generate fragments per kilobase of exon per million reads aligned (FPKM) values for each gene. This created a value of FPKM for the gene/median FPKM for that chromosome, multiplied by the ploidy estimate to give the raw haploid number. The genes were grouped by OrthoMCL ortholog ID version 4 (http://orthomcl.org) and also by chromosome, and annotated using Pfam entries from the Interpro annotations. Domains that were significantly over-represented in the genes in multicopy arrays versus the genomic background by species were identified using the hypergeometric distribution and a *P*-value threshold corrected for multiple testing of the terms ($P < 10^{-5}$).

Bias of multicopy arrays to disomic chromosomes

The representation of multicopy arrays on the non-disomic chromosomes was examined using a Monte-Carlo simulation. Using a custom script, each of the genes was placed at random into the genome, weighted for the sizes of the different chromosomes. If the array was placed on a non-disomic chromosome, this was scored, and if the score was the same or higher than the real non-disomic score for that species, the run was counted. Repeating this random analysis a million times for each species gave an empirical *P*-value for the disomic bias.

Base frequencies for heterozygous sites were inferred from filtered Samtools pile-up results. For each predicted heterozygous site, the number of each of the four possible alternative sites over the read coverage of this base was determined and rounded to the second decimal place by a custom Perl script. Base frequencies were binned into 101 categories of 0.01 each (0–1.00), and an approximate distribution of base frequencies for each chromosome of *L. braziliensis* and *L. mexicana* was plotted in Excel (Supplemental Fig. S3). Frequencies of zero were not included in the plot.

Further details on parasites, Illunina sequencing, bioinformatics analyses and annotation, SNP analysis, chromosome read depth analysis, and gene copy number analysis can be found in the Supplemental Methods.

Data access

The sequence data in this study have been submitted to the NCBI Sequence Read Archive (http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi) under accession numbers ERP000169, ERX005631, ERX005632, ERX005633, and ERX005636. The genome sequence of *L. mexicana* U1103 has been submitted to the EMBL Nucleotide Sequence Database (http://www.ebi.ac.uk/embl) under accession numbers FR799554–FR799587. New chromosome/contig data for resequenced *Leishmania* genomes have been submitted to the EMBL Nucleotide Sequence Database under accession numbers FR796397–FR796432, FR796433–FR796468, FR798975–FR799010, CADB0100001–CADB01000554, CADA01000002–CADA01000102, and CACT01000001–CACT01000040.

Acknowledgments

This work was supported by the Wellcome Trust (Grant numbers 076355, 085822, and 085775). We thank our colleagues in the sequencing and informatics groups at the Wellcome Trust Sanger Institute and the Sir Henry Wellcome Functional Genomics Centre at the University of Glasgow. We are grateful to Walide Saad for contributing to the generation of *L. infantum SEC14*-deficient mutants.

References

Akopyants NS, Kimblin N, Secundino N, Patrick R, Peters N, Lawyer P, Dobson DE, Beverley SM, Sacks DL. 2009. Demonstration of genetic exchange during cyclical development of *Leishmania* in the sand fly vector. *Science* **324:** 265–268.

Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. 2009. ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* **25:** 1968–1969.

Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, Lennard NJ, Caler E, Hamlin NE, Haas B, et al. 2005. The genome of the African trypanosome *Trypanosoma brucei*. *Science* **309:** 416–422.

Besteiro S, Williams RAM, Coombs GH, Mottram JC. 2007. Protein turnover and differentiation in *Leishmania*. *Int J Parasitol* **37:** 1063–1075.

Bradley DJ, Kirkley J. 1977. Regulation of *Leishmania* populations within the host. I. The variable course of *Leishmania donovani* infections in mice. *Clin Exp Immunol* **30:** 119–129.

Britto C, Ravel C, Bastien P, Blaineau C, Pagès M, Dedet JP, Wincker P. 1998. Conserved linkage groups associated with large-scale chromosomal rearrangements between Old World and New World *Leishmania* genomes. *Gene* **222:** 107–117.

Carver T, Berriman M, Tivey A, Patel C, Bohme U, Barrell BG, Parkhill J, Rajandream MA. 2008. Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* **24:** 2672–2676.

Chain PS, Grafham DV, Fulton RS, Fitzgerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C, et al. 2009. Genomics. Genome project standards in a new era of sequencing. *Science* **326:** 236–237.

Clayton C, Shapira M. 2007. Post-transcriptional regulation of gene expression in trypanosomes and leishmanias. *Mol Biochem Parasitol* **156:** 93–101.

Cruz AK, Titus R, Beverley SM. 1993. Plasticity in chromosome number and testing of essential genes in *Leishmania* by targeting. *Proc Natl Acad Sci* **90:** 1599–1603.

DeLuna A, Vetsigian K, Shoresh N, Hegreness M, Colon-Gonzalez M, Chao S, Kishony R. 2008. Exposing the fitness contribution of duplicated genes. *Nat Genet* **40:** 676–681.

Depledge DP, Evans KJ, Ivens AC, Aziz N, Maroof A, Kaye PM, Smith DF. 2009. Comparative expression profiling of *Leishmania*: Modulation in gene expression between species and in different host genetic backgrounds. *PLoS Negl Trop Dis* **3:** e476. doi: 10.1371/journal.pntd.0000476.

Depledge DP, MacLean LM, Hodgkinson MR, Smith BA, Jackson AP, Ma S, Uliana SR, Smith DF. 2010. *Leishmania*-specific surface antigens show sub-genus sequence variation and immune recognition. *PLoS Negl Trop Dis* **4:** e829. doi: 10.1371/journal.pntd.0000829.

Devault A, Banuls AL. 2008. The promastigote surface antigen gene family of the *Leishmania* parasite: differential evolution by positive selection and recombination. *BMC Evol Biol* **8:** 292. doi: 10.1186/1471-2148-8-292.

Downing T, Imamura H, Decuypere S, Clark TG, Coombs GH, Cotton JA, Hilley JD, de Doncker S, Maes I, Mottram JC, et al. 2011. Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance. *Genome Res* (this issue). doi: 10.1101/gr.123430.111.

Eschenlauer SC, Coombs GH, Mottram JC. 2006. *PFPI*-like genes are expressed in *Leishmania major* but are pseudogenes in other *Leishmania* species. *FEMS Microbiol Lett* **260:** 47–54.

Gelanew T, Kuhls K, Hurissa Z, Weldegebreal T, Hailu W, Kassahun A, Abebe T, Hailu A, Schönian G. 2010. Inference of population structure of *Leishmania donovani* strains isolated from different Ethiopian visceral leishmaniasis endemic areas. *PLoS Negl Trop Dis* **4:** e889. doi: 10.1371/journal.pntd.0000889.

Gomez MA, Contreras I, Halle M, Tremblay ML, McMaster RW, Olivier M. 2009. *Leishmania* GP63 alters host signaling through cleavage-activated protein tyrosine phosphatases. *Sci Signal* **2:** ra58. doi: 10.1126/scisignal.2000213.

Halle M, Gomez MA, Stuible M, Shimizu H, McMaster WR, Olivier M, Tremblay ML. 2009. The *Leishmania* surface protease GP63 cleaves multiple intracellular proteins and actively participates in p38 mitogen-activated protein kinase inactivation. *J Biol Chem* **284:** 6893–6908.

Hassan P, Fergusson D, Grant KM, Mottram JC. 2001. The CRK3 protein kinase is essential for cell cycle progression of *Leishmania mexicana*. *Mol Biochem Parasitol* **113:** 189–198.

Hilley JD, Zawadzki J, McConville MJ, Coombs GH, Mottram JC. 2000. *Leishmania mexicana* mutants lacking glycosylphosphatidyl (GPI):protein transamidase provide insights into the biosynthesis and functions of GPI-anchored proteins. *Mol Biol Cell* **11:** 1183–1195.

Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, Sisk E, Rajandream MA, Adlem E, Aert R, et al. 2005. The genome of the kinetoplastid parasite, *Leishmania major*. *Science* **309:** 436–442.

Jackson AP. 2010. The evolution of amastin surface glycoproteins in trypanosomatid parasites. *Mol Biol Evol* **27:** 33–45.

Jackson AP, Vaughan S, Gull K. 2006. Evolution of tubulin gene arrays in Trypanosomatid parasites: genomic restructuring in *Leishmania*. *BMC Genomics* **7:** 261. doi: 10.1186/1471-2164-7-261.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** R25. doi: 10.1186/gb-2009-10-3-r25.

Leprohon P, Legare D, Raymond F, Madore E, Hardiman G, Corbeil J, Ouellette M. 2009. Gene expression modulation is associated with gene amplification, supernumerary chromosomes and chromosome loss in antimony-resistant *Leishmania infantum*. *Nucleic Acids Res* **37:** 1387–1399.

Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res* **13:** 2178–2189.

Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18:** 1851–1858.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079.

Lukes J, Mauricio IL, Schonian G, Dujardin JC, Soteriadou K, Dedet JP, Kuhls K, Tintaya KW, Jirku M, Chocholova E, et al. 2007. Evolutionary and geographical history of the *Leishmania donovani* complex with a revision of current taxonomy. *Proc Natl Acad Sci* **104:** 9375–9380.

Lynch M, Conery JS. 2003. The origins of genome complexity. *Science* **302:** 1401–1404.

Martinez-Calvillo S, Yan S, Nguyen D, Fox M, Stuart K, Myler PJ. 2003. Transcription of *Leishmania major* Friedlin chromosome 1 initiates in both directions within a single region. *Mol Cell* **11:** 1291–1299.

Martinez-Calvillo S, Stuart K, Myler PJ. 2005. Ploidy changes associated with disruption of two adjacent genes on *Leishmania major* chromosome 1. *Int J Parasitol* **35:** 419–429.

Murray HW, Berman JD, Davies CR, Saravia NG. 2005. Advances in leishmaniasis. *Lancet* **366:** 1561–1577.

Ning Z, Cox AJ, Mullikin JC. 2001. SSAHA: A fast search method for large DNA databases. *Genome Res* **11:** 1725–1729.

Noben-Trauth N, Lira R, Nagase H, Paul WE, Sacks DL. 2003. The relative contribution of IL-4 receptor signaling and IL-10 to susceptibility to *Leishmania major*. *J Immunol* **170:** 5152–5158.

Otto TD, Sanders M, Berriman M, Newbold C. 2010. Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics* **26:** 1704–1707.

Otto TD, Dillon GP, Degrave WS, Berriman M. 2011. RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Res* **39:** e57. doi: 10.1093/nar/gkq1268.

Pavelka N, Rancati G, Zhu J, Bradford WD, Saraf A, Florens L, Sanderson BW, Hattem GL, Li R. 2010. Aneuploidy confers quantitative proteome changes and phenotypic variation in budding yeast. *Nature* **468:** 321–325.

Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail MA, Peters N, Adlem E, Tivey A, Aslett M, et al. 2007. Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. *Nat Genet* **39:** 839–847.

Rancati G, Pavelka N, Fleharty B, Noll A, Trimble R, Walton K, Perera A, Staehling-Hampton K, Seidel CW, Li R. 2008. Aneuploidy underlies rapid adaptive evolution of yeast cells deprived of a conserved cytokinesis motor. *Cell* **135:** 879–893.

Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. 2011. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol* **12:** R22. doi: 10.1186/gb-2011-12-3-r22.

Rougeron V, De MT, Hide M, Waleckx E, Bermudez H, Arevalo J, Llanos-Cuentas A, Dujardin JC, De DS, Le RD, et al. 2009. Extreme inbreeding in *Leishmania braziliensis*. *Proc Natl Acad Sci* **106:** 10224–10229.

Schuster-Bockler B, Conrad D, Bateman A. 2010. Dosage sensitivity shapes the evolution of copy-number varied regions. *PLoS ONE* **5:** e9474. doi: 10.1371/journal.pone.0009474.

Smith DF, Peacock CS, Cruz AK. 2007. Comparative genomics: From genotype to disease phenotype in the leishmaniases. *Int J Parasitol* **37:** 1173–1186.

Stager S, Smith DF, Kaye PM. 2000. Immunization with a recombinant stage-regulated surface protein from *Leishmania donovani* induces protection against visceral leishmaniasis. *J Immunol* **165:** 7064–7071.

Sterkers Y, Lachaud L, Crobu L, Bastien P, Pages M. 2011. FISH analysis reveals aneuploidy and continual generation of chromosomal mosaicism in *Leishmania major*. *Cell Microbiol* **13:** 274–283.

Sunkin SM, Kiser P, Myler PJ, Stuart K. 2000. The size difference between *Leishmania major* Friedlin chromosome one homologues is localized to sub-telomeric repeats at one chromosomal end. *Mol Biochem Parasitol* **109:** 1–15.

Taylor JS, Raes J. 2004. Duplication and divergence: The evolution of new genes and old ideas. *Annu Rev Genet* **38:** 615–643.

Torres EM, Dephoure N, Panneerselvam A, Tucker CM, Whittaker CA, Gygi SP, Dunham MJ, Amon A. 2010. Identification of aneuploidy-tolerating mutations. *Cell* **143:** 71–83.

Ubeda JM, Legare D, Raymond F, Ouameur AA, Boisvert S, Rigault P, Corbeil J, Tremblay MJ, Olivier M, Papadopoulou B, et al. 2008. Modulation of gene expression in drug resistant *Leishmania* is associated with gene amplification, gene deletion and chromosome aneuploidy. *Genome Biol* **9:** R115. doi: 10.1186/gb-2008-9-7-r115.

Wincker P, Ravel C, Blaineau C, Pages M, Jaufrett Y, Dedet JP, Bastien P. 1998. The *Leishmania* genome comprises 36 chromosomes conserved across widely divergent human pathogenic species. *Nucleic Acids Res* **24:** 1688–1694.

Yamage M, Debrabant A, Dwyer DM. 2000. Molecular characterization of a hyperinducible, surface membrane-anchored, class I nuclease of a trypanosomatid parasite. *J Biol Chem* **275:** 36369–36379.

Zhang WW, Matlashewski G. 2010. Screening *Leishmania donovani*-specific genes required for visceral infection. *Mol Microbiol* **77:** 505–517.

Zhang WW, Peacock CS, Matlashewski G. 2008. A genomic-based approach combining in vivo selection in mice to identify a novel virulence gene in *Leishmania*. *PLoS Negl Trop Dis* **2:** e248. doi: 10.1371/journal.pntd.0000248.

Received March 22, 2011; accepted in revised form August 23, 2011.

Chromosome and gene copy number variation allow major structural change between species and strains of *Leishmania*

Matthew B. Rogers, James D. Hilley, Nicholas J. Dickens, et al.

Genome Res. 2011 21: 2129-2142 originally published online October 28, 2011
Access the most recent version at doi:10.1101/gr.122945.111