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, Jon Wilkes, Paul A. Bates, Daniel P. Depledge, David Harris, Yerim Her, Pawel Herzyk, Hideo Imamura, Thomas D. Otto, Mandy Sanders, Kathy Seeger, J-C. Dujardins, Matthew Berriman, Deborah F. Smith, Christiane Hertz-Fowler, Jeremy C. Mottram

Supplementary Materials and Methods

Parasites.

All *Leishmania* species were tested by PCR-RFLP (Schonian et al. 2003) to confirm species. DNA was extracted from promastigotes in late log phase of growth (~$2x10^7$ cells ml^{-1}) using the DNEasy DNA Purification Kits (Qiagen). DNA quantity and quality was assessed by agarose gel electrophoresis and Nanodrop.

Illumina Sequencing.

DNA from *L. major* Friedlin, *L. major* LV39, *L. donovani* LV9, *L. infantum* JPCM5 and *L. braziliensis* M2904 was sheared into 200-300bp fragments using Covaris Adaptive Focused Acoustics technology (AFA) followed by end repair with T4 and Klenow DNA polymerases and T4 polynucleotide kinase to blunt-end the DNA fragments. A 3' A nucleotide was added to the repaired ends using Klenow exonuclease and dATP. NoPCR adapters containing primer sites for sequencing and flow cell surface annealing were then ligated (Kozarewa et al. 2009). Ligated fragments were run on an agarose gel, size selected and extracted. These libraries were quantified using Agilent Bioanalyser chip and kappa SYBR qPCR. Sequencing libraries were denatured with sodium hydroxide and diluted with a hybridisation buffer for loading onto individual lanes of an Illumina GAII flowcell. Cluster formation, primer hybridization and PE sequencing of 76bp cycles were performed using proprietary reagents according to manufacturer's recommended protocol (https://icom.illumina.com/). Data was analysed from the Illumina GAII sequencing machines using RTA1.6 analysis pipeline

Bioinformatics Analysis and Annotation.

The annotation of the *L. mexicana* genome can be viewed and searched via GeneDB (http://www.genedb.org/) (Hertz-Fowler et al. 2004) and TritrypDB (http://tritrypdb.org) (Aslett et al. 2010). Initial *L. mexicana* gene models were transferred from the *Leishmania major* genome using Rapid Annotation Transfer Tool (Otto et al., 2010). This resulted in 7,825 predicted gene models which were were then manually verified and edited using the Artemis annotation tool (Carver et al. 2008). An additional 425 gene models were added manually using the Artemis annotation tool, informed by codon usage, synteny with other *Leishmania* genomes, and Blast hits against the NCBI NR database. Careful attention was paid to regions predicted to be non-syntenic with other *Leishmania* genomes by Nucmer, and regions and large intergenic regions.

Iterative Correction Of Reference Nucleotides (iCORN) (Otto et al. 2010) was run for 8 iterations with *Leishmania major* at which point no new SNPs were called and new insertions existed in a balance with deletions from the previous iteration. *L. infantum* was run for 12 iterations at which point no new SNPs were predicted and insertions and deletions had equilibrated or reached a plateau at a single insertion/deletion per generation. *L. braziliensis* was run for 13 iterations, at which point no more SNPs were called on all but two chromosomes. *L. mexicana* was run for 14 iterations at which point no new SNPs were called on all but 1 chromosome. For both *L. braziliensis* and *L. mexicana*, ICORN continued to make insertions and deletions on most chromosomes up to the final iteration, but with a few exceptions these appeared to exist in balance with insertions/deletions from the previous iteration. In addition to re-sequencing the *Leishmania* reference genomes, we have made minor structural changes to the genomes of *L. infantum* and *L. braziliensis* by removing contigs from the 3' ends of chromosomes, corresponding to contigs that could not be ordered correctly in the previous versions. This has resulted in a *L. braziliensis* bin chromosome consisting of 103 contigs and 850,747 base pairs, and 201,216 base pairs of sequence in *L. infantum* chromosome consisting of 40 contigs.

The most obvious improvements have been the correction of genes containing frame shifts, particularly in the draft genomes. In both *L. infantum* and *L. braziliensis* respectively, 98 and 68 genes that previously contained frame shifts were converted to intact open reading frames. Far fewer gene model changes occurred in the finished genome of *L. major*, where only 3 genes containing frame shifts were corrected. This nevertheless allowed for the status of these genes as pseudogenes to be removed. Table S7 contains a summary of the substitutions, insertions and deletions performed by ICORN for each genome**.**

An unexpectedly large proportion of *L. braziliensis* reads (15%) did not map to the existing *L. braziliensis* reference genome using ICORN. These unmapped reads were assembled using Velvet v. 1.0.12 using a Kmer of 55, and a coverage cut-off of 4. This resulted in 1.24 Mb of sequence, consisting of 3,636 contigs with an average size of 342 base pairs. Although most of these contigs were too small to annotate uninterrupted reading frames, this file was used to run blast searches for *L. braziliensis* genes identified as uniquely absent using predicted orthologues in *L. major*. This search returned 34 hits with bit scores greater than 150, which were then excluded from the final table of uniquely absent *L. braziliensis* genes.

Orthologous genes were predicted by running OrthoMCL v 1.4 (Li et al. 2003) on the *L. mexicana* genes. The initial OrthoMCL results were loaded into an Excel spreadsheet and instances of differential gene distributions between *Leishmania* genomes were examined using a four-way sequence alignment visualised in the Artemis Comparison Tool (ACT). On the basis of ACT comparisons, further orthologues not predicted by OrthoMCL were manually assigned to orthologue clusters on the basis of conserved synteny and sequence similarity. Excluded from downstream analyses were spurious gene models such as 5' or 3' truncated sequences and some genes belonging to large widely distributed families such as histones and amastins as these clustered into large groups where it was not possible to establish any further hierarchy of orthology.

SNP Analysis.

SNPs were predicted using a combination of SSAHA2 to map Illumina reads to the appropriate reference genome, and SAMtools pileup to call SNPs from the mapped reads. Samtools varfilter.pl was used to exclude regions where the SNP density was higher than 3

SNPs in a 7 bp window, or the read depth exceeded 5000 reads. Samtools pileup files were further filtered on the following quality scores: 40 consensus, 40 SNP quality, 25 mapping score, 10 coverage. In addition to these parameters, SNPs called in regions where the read coverage was more than 2x the predicted median coverage for the entire chromosome were discarded to limit the effect of paralogy on SNP determination. Although reads could be mapped to the 'bin chromosomes' of *L. infantum*, *L. braziliensis* and *L. mexicana*, SNPs could not be called reliably on these chromosomes as these were composed of unordered contigs from various chromosomes with differing ploidy values. Variants affecting coding sequences were examined by converting the filtered Samtools pileup format files to variant call format (VCF) files, and examining these in Artemis.

We detected 10-fold more heterozygous sites (3,705) in *L. major* LV39 than in the Friedlin strain, and 72,879 homozygous SNPs. These SNPs will be useful for positional cloning of genes regulating traits such as virulence in hybrid progeny from *L. major* Friedlin and LV39 genetic crosses (Akopyants et al. 2009). We also predicted 101,088 homozygous SNPs in *L. donovani* LV9 compared with *L. infantum* JPCM5 and 10,007 heterozygous sites - numbers similar to those reported for a comparison between JPCM5 and 17 clinically isolated *L. donovani* lines (Downing et al. 2011). As seen in LV39, a far larger number of heterozygous sites are predicted in these two strains than in the reference genomes.

Thirteen out of the14 *L. major*-specific genes identified in the orthoMCL analysis appear intact in LV39, the one exception being hypothetical protein *LmjF.32.2470*, which contains a premature stop codon. For 5 genes the read coverage or mapping quality across the gene was insufficient to call SNPs with high confidence, making it possible that these genes contain undetected SNPs in unmapped regions.

Similarly, no SNPs resulting in internal stop codons were predicted for any of the 19 *L. infantum* unique genes in *L. donovani* LV9, although for 7 of the 19 unique genes there was insufficient read coverage or mapping quality issues prevented SNP calling across the entire length of the gene. The remaining genes were all intact and while most genes have synonymous SNPs, we predict that all are expressed to give functional proteins. Similarly in *L. donovani* BPK206/0, only one of the genes examine contained an internal stop codon (LinJ.19.1120), and in 7 of the 19, mapping problems prevented SNP calling across the entire length of the gene.

Chromosome read depth analysis.

If necessary, the quality score encoding of the reads was converted from Illumina 1.3+ to Sanger format. All short read sets were pair-ended (insert length ~200bp), with read lengths of 51 or 76 bases. Reads were mapped to the appropriate reference genomes using MAQ version 0.7.1, under the guidance of a custom perl script. Read sets were parsed into smaller paired sets of 2.5x10^6 reads or less and converted into binary format. Following the initial mapping, map files relating to a given sample were merged into a larger file, and this alignment output in pileup 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 normalisation was performed by setting the average of the read depth of the four longest 'euploidic' chromosomes to 2. The read depth for each chromosome was subsequently normalized to this value.

Generating a threshold to call a gene copy number gain

The *L. major* Friedlin genome has the most complete sequence coverage and is the best annotated of the *Leishmania* reference genomes. In order to calibrate the detection of multi-copy arrays we used the genes being annotated as either a single copy (haploid number 1) in the genome or multi-copy (haploid number 2).

Threshold	% Sensitivity	Positive Predictive Value
1.5	95	48.1
1.55	94.2	52.9
1.6	93.8	57.8
1.65	93	64.3
1.7	92.2	70.2
1.75	90.3	74
1.8	87.6	78.7
1.85	83.3	81.7
1.9	79.1	85
1.95	76	88.7
2	72.5	90.3

We chose the threshold that had the maximum positive-predictive value, so that we could be sure that those that were being called an array (copy number (cn) 2) in the genome were actually arrays. But there is an associated loss of sensitivity to the detection of a single-copy number increase. Exploring this further in the *L. major* genome, we looked at the actual numbers of arrays called 2 that were annotated as 1 in the genome and those that were called 1 that were annotated as a 2. The percentages are a proportion of all the genes/arrays that are 1 or 2 in the genome respectively.

Threshold	% 1 in ref called 2 (number)	% 2 in ref called 1 (number)	% Identity to the numbers in the *L. major* genome
1.5	3.4 (259)	6.6 (12)	95.3
1.55	2.8 (211)	7.2 (13)	95.9
1.6	2.3 (172)	7.2 (13)	96.4
1.65	1.7 (128)	8.3 (15)	96.9
1.7	1.3 (96)	9.4 (17)	97.3
1.75	1 (77)	12.2 (22)	97.5
1.8	0.7 (56)	16 (29)	97.6
1.85	0.6 (43)	21.5 (39)	97.7
1.9	0.4 (31)	27.6 (50)	97.7
1.95	0.3 (20)	32 (58)	97.7
2	0.2 (15)	37 (67)	97.7

Although there is a loss of sensitivity at a threshold of 2 and 37% of arrays that are called cn 2 in the genome are being called too low there are also 15 arrays that are not called 2 in the genome but are clearly a 2. However, only 0.2% genes, *i.e.* 15, that are annotated as single copy are called as higher copies. There is still a 97.7% overall agreement with the exact number of genes annotated in the *L. major*. We used copy number 2 and 1 for the sensitivity analyses as these will be the best annotated in the genome; the overall identity was calculated against all copy numbers.

In Table S3 we have given the raw and rounded copy number for all genes/arrays that are ≥1.5, the ones shaded in gray are in the 1.5-2.0 boundary. Some of these will be real arrays, especially those with 2 genes annotated in the reference genome, but the false-positive rate was too high at this level to keep them in the overall analyses.

References

1. Akopyants, N.S., Kimblin, N., Secundino, N., Patrick, R., Peters, N., Lawyer, P., Dobson, D.E., Beverley, S.M., and Sacks, D.L. 2009. Demonstration of genetic exchange during cyclical development of *Leishmania* in the sand fly vector. *Science* **324**:265-268
2. Aslett, M., Aurrecoechea, C., Berriman, M., Brestelli, J., Brunk, B.P., Carrington, M., Depledge, D.P., Fischer, S., Gajria, B., Gao, X., et al. 2010. TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Research* **38**:D457-D462
3. Downing, T., Imamura, H., Decuypere, S., Clark, T.G., Coombs, G.H., Cotton, J.A., Hilley, J.D., De Doncker, S., Maes, I., Mottram, J.C., et al. 2011. Whole genome sequencing of *Leishmania donovani* clinical lines provides insights into population structure and mechanisms of drug resistance. *Genome Res, in press*
4. Guindon, S., Lethiec, F., Duroux, P., and Gascuel, O. 2005. PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Research* **33**:W557-W559
5. Hertz-Fowler, C., Peacock, C.S., Wood, V., Aslett, M., Kerhornou, A., Mooney, P., Tivey, A., Berriman, M., Hall, N., Rutherford, K., et al. 2004. GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Research* **32**:D339-D343
6. Kozarewa, I., Ning, Z., Quail, M.A., Sanders, M.J., Berriman, M., and Turner, D.J. 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods* **6**:291-295
7. Li, L., Stoeckert, C.J., Jr., and Roos, D.S. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* **13**:2178-2189
8. Otto, T.D., Sanders, M., Berriman, M., and Newbold, C. 2010. Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics.* **26**:1704-1707
9. Schonian, G., Nasereddin, A., Dinse, N., Schweynoch, C., Schallig, H.D., Presber, W., and Jaffe, C.L. 2003. PCR diagnosis and characterization of *Leishmania* in local and imported clinical samples. *Diagn.Microbiol Infect Dis* **47**:349-358