*Supplementary data for paper entitled "*Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance*" by Downing & Imamura et al. (submitted as GENOME/2011/123430).*

Supplementary Data

Table of Contents

Supplementary Results (text)

Genome sequencing and assembly

An initial set of 4,455 contigs with an N50 (minimum contig length required for covering at least 50% of the genome) of 22.3 kb were assembled *de novo* from the 454 sequence data using Newbler. iCORN (Otto et al. 2010), a nucleotide-level improvement process, iteratively corrected a total of 1,363 miscalled bases, 8,828 insertions and 2,230 deletions. IMAGE (Tsai et al. 2010) eliminated over a thousand gaps, thus reducing the number of contigs to 3,326 and increasing the number of known bases in the reference *L. donovani* genome from 30.5 to 31.8 Mbp. This additional 1.3 Mbp allowed further enhancements by iCORN, correcting 564 base errors, 197 insertions and 97 deletions. ABACAS (Assefa et al. 2009) was used to map 2,154 contigs to the 36 chromosomes of *L. infantum* and inserted 100bp gaps between these contigs, and excluding these gaps there are 848 remaining gaps with total length 1.3 Mbp and a mean length of 1,544 bp. As a result of these genome improvement steps, the final N50 for chromosomes was elevated to 45.5 kb.

In the reference *L. donovani* genome of length 32,444,998 bp, 1,309,308 bp bases were undetermined in the 36 chromosomes, kDNA and unassigned contigs. The longest kDNA contigs were 1,459 bp (minicircle) and 18,318 bp (maxicircle). A set of 1,172 contigs (total length of 608,394 bp) remained unassigned to chromosomes, giving a total of 3,326 chromosomal contigs. An estimated 622 of the unassigned contigs (length 317,556 bp) could be linked to particular chromosome if assembly quality thresholds were reduced. 98.4% of BPK282/0cl4 reads mapped to these 3,326 contigs: 97.2% as a pair. 93.9% of BPK282/0cl4 reads could be mapped to *L. infantum* chromosomes: 93.7% of these were in proper pair whose insert size was within the expected range designed for sequencing. The fraction of singleton reads whose mate read unmapped was 3.96%. In the reference *L. donovani* genome, the overall average read depth coverage was 69.7: the median was 60 and mode 55.

Gene model transfer and annotation

In order to identify genes unique to *L. donovani* and classify the unassigned contigs, a search of the most related (*L. infantum*) and the most completed (*L. major*) genomes was conducted (BLAST, Altschul et al. 1990). Of the 1,172 non-genomic contigs, 908 had lengths greater than 100 bp: alignments of these to the *L. infantum*, *L. major* genomes and the kDNA contigs showed that the vast majority were homologous to regions in *L. infantum* alone (71%) and or in *L. major* exclusively (5%). Additionally, 6% showed significant similarity to multiple groups and 11% may have been sequencing contaminants. The remaining 7% of contigs were classed as kDNA minicircle (4%) and maxicircle (3%) sequence. *L. donovani* contigs 00711 (971 bp) and 04329 (116 bp) had no homologous region in *L. infantum* but did compared to *L. major* for a putative kinesin K39 gene (Supplementary Figure 21); however this locus may be present in *L. infantum* since it showed extensive structural variation.

Genomic patterns of SNP diversity within *L. donovani* lines

Tajima's D at nonsynonymous compared to synonymous sites was considerably more negative (-0.07 vs 0.23; Tajima 1989), and the former had a lower excess of rare variants, as measured by Fu's F_S (-1.17 vs -1.86; Fu 1997). Taking set of lines as a pair, comparisons of greater circle geographic distance, absolute SSG phenotype difference and the mean number of nucleotide differences per kb revealed no linear

association, supported by non-significant Mantel test p values. Signal peptide (SignalP v3.0; www.cbs.dtu.dk/services/SignalP/; Bendtsen et al. 2004) and transmembrane (TMHMM v2.0c; www.cbs.dtu.dk/services/TMHMM/; Krogh et al. 2001) regions were determined for each translated *L. donovani* gene: 622 (7.8%) had evidence of signal peptide domains and 1,408 (17.6%) transmembrane domains.

We investigated the effects of genetic diversity on the fitness of the 17 lines by calculating the ratios of coding SNPs that were advantageous, neutral or deleterious variants. There was a significant excess of protein-level variants in the 17 *L. donovani* lines (P_N/P_S = 396/220 = 1.8; Supplementary Table 4) compared to the substitution rates in alignments of the *L. donovani* reference with each of the four other *Leishmania* species ($1.0 \leq D_N/D_S \leq 1.1$, Fisher's Exact test $p < 0.002$), given the neutral expectation that the FI = $[D_N/D_S]/[P_N/P_S] = 1$ (McDonald and Kreitman 1991).

The number of genomic nonsynonymous (L_N = 11,102,501.8) and synonymous (L_S = 3,373,689.4) sites (total known coding sites 14,476,191 for 7,935 chromsomal genes) was estimated as $f = [P_N/P_S]/[L_N/L_S]$ to obtain the fraction neutral segregating substitutions (Supplementary Table 4). *f* includes slightly advantageous and slightly deleterious variants, where the selective coefficient (s) of the mutation scaled for the effective population size (N_e) was $-1 \leq N_e s \geq 1$, where $s < 0$ for deleterious alleles and $s > 0$ for beneficial ones (Smith and Eyre-Walker 2002). For folded allele frequency ≤ 0.4, *f* was 0.38. An excess of segregating relative to fixed nonsynonymous variants was supported by calculating both the expected values of P_N/P_S relative to D_N/D_S (Gojobori et al. 2007) and the effective FI (Axelsson and Ellegren 2009; Supplementary Figure 5), further evidencing that a portion of population-level mutations were deleterious rather than advantageous. Highly deleterious mutations were lethal and so were unlikely to be observed. Sharp selective sweeps proportionally reduce the incidence of high- and mid-frequency alleles more strongly than rare ones (Fu 1997); the same signature is observed during rapid demographic growth or recovery from an acute population decrease (Zeng et al. 2006).

Known genes with extreme patterns of diversity in *L. donovani*

A single gene (LdBPK_311340) that had more than one SNP discriminating between the SSG groups has high homology to protein phosphatase 2C (PP2C) at its 3' end (IPR001932; GO:0003824). In *Leishmania* PP2C is expressed in both promastigote and amastigote life stages (Burns et al. 1993), and regulates stress-regulated signalling in mammals, yeast and plants (Maeda et al. 1995, Takekawa et al. 1998, Hanada et al. 1998, Rodriguez et al. 1998).

Interestingly, the most highly differentiated SNP (H523R, F_{ST} = 0.54) in a hypothetical gene (LdBPK_342390) was predicted to be functionally and structurally tolerated: amino acids H and R are both positively charged and polar but only the latter is aromatic and hydrophobic. The 7 SSG-resistant lines all possessed the H allele (CGC), while all the susceptible lines had R (CAC), allowing for three heterozygotes (BPK178/0cl3, BPK282/0cl4 and BPK294/0cl1); this gene also had a high estimated abundance of segregating advantageous variants based on its ancestral substitution rate. However, alignments of the local *L. donovani* and *L. infantum* genomic regions indicate it may be structurally variable, complicating SNP detection. This gene product is constitutively expressed in *L. major* promastigote and amastigote life-cycle stages (Leifso et al. 2007).

No variation was identified within the *L. donovani* lines at certain genomic regions previously used to distinguish between *Leishmania* species and complexes: cytochrome oxidase II (Ibrahim and Barker et al. 2001), glycoprotein 63 (Mauricio et al. 2007), cysteine protease genes (Hide et al. 2007), cytochrome b (Luyo-Acero et al. 2004) and elongation factor 1 (Momen and Cupolillo 2000). However, two nonsynonymous intermediate-frequency mutations (K116R, folded allele frequency 0.26; R319H, 0.22) were observed at heat-shock protein (HSP) 70 (LdBPK_302480); HSP-70 genes have been associated with antimony resistance in *L. donovani* (Vergnes et al. 2007)*, tarentolae* and *infantum* (Brochu et al. 2004). Further downstream to HSP-70, a high-frequency (0.45) synonymous SNP was discovered in glyceraldehyde-3-phosphate dehydrogenase (GAPDH, LdBPK_302990; Momen & Cupolillo 2000). Although this suggests HSP-70 and GAPDH have useful phylogenetic power for discriminating *Leishmania* in the Indian subcontinent, evidence of local structural diversity may complicate SNP ascertainment.

Additionally, a transport-related gene involved in pterin and folate metabolism (pteridine transporter, LdBPK_101450; Kumar et al. 2007) contained a high-frequency synonymous SNP – a related protein pteridine reductase could be a drug target for leishmaniasis treatment (Kaur et al. 2010).

A mutation (D461N) was observed in lines BPK043/0cl2 and BPK035/0cl1 only in the P299 gene (LdBPK_080630), whose product confers SSG and miltefosine resistance at high expression in *L. infantum* (Choudhury et al. 2008) and is predominantly expressed at the promastigote stage (Rosenzweig et al. 2008).

A surface-related kinase gene (PI3K, LdBPK_020100) has high rates of nonsynonymous substitution in *Leishmania* species (Peacock et al. 2007) and here its sole intraspecific SNP was one of the 34 discriminating the four SSG-R lines from the remainder (R3452H).

Genes linked to signalling were associated with selective processes on the *L. donovani* lineage, as were three related to protein degradation – one (LdBPK_080960) was predicted to have a high number of mutations beneficial to the parasite. The same gene also had one of its four nonsynonymous changes in the 17 lines in its predicted signal peptide domain: it is possible this may be involved in differential expression of the active protein.

5 kb sliding windows for genes were at least in the top 5% most divergent for each species compared to *L. donovani*. The sliding window approach identified seven genes in the SNP-dense regions that also represented candidate structurally diverse genomic segments, two of which had evidence of relaxed selective constraint or subject to directional selection (hypothetical multi-pass gene, LdBPK_242320; conserved hypothetical gene, LdBPK_242330; Supplementary Figure 22). The remaining five gene encoded: a ß tubulin (LdBPK_081290), a conserved hypothetical proteins (LdBPK_120930, LdBPK_333220), a cytochrome c oxidase subunit VI (LdBPK_212080) and a sm-f snRNP core complex protein (LdBPK_354530).

Structural variation discovery

To validate the sequencing results by QPCR, genes located on chromosome 17, 32 and 34 (diploid in all lines) and chromosome 31 (tetraploid in all lines) were quantified by QPCR in the same DNA samples as those used for sequencing. These specific chromosomes were chosen because of their stable character in all samples according to the sequencing results. The crossing point (Cp) values resulting from the QPCR analysis are a measure for the relative quantity of target DNA in the samples, the lower the Cp, the more copies present in the original sample. The genes located on chromosome 17, 32 and 34 showed an average Cp of 17.84 (95% CI: 17.60 – 18.08) while the gene located on chromosome 31 had an average Cp of 16.72 (95% CI: 16.32 - 17.12). These two groups were found to differ significantly (Mann Whitney U-test $p < 0.001$) with 1,12 Cp, equals a 2.17-fold difference in quantity between the diploid chromosomes (17,34 and 36) and the tetraploid chromosome 31. These results therefore validate the quantitative aspect of our sequencing results.

Extensive homology was observed at the MAP kinase locus on chromosome 36 for the pair of phosphatase genes at the 5' (TSAP: tartrate-sensitive acid phosphatase, LdBPK_366740) and 3' (HSAP; histidine secretory acid phosphatase, LdBPK_366770) ends of the episome region. Alignments of each gene for *L. donovani*, *infantum*, *mexicana* and *braziliensis* support this (T-Coffee, Notredame et al. 2000; *L. mexicana* had no orthologous region). The 1,089 bp HSAP gene had homology of 69% – but this rose to 85% for the first 220 bp. Similarly, the 282 bp TSAP gene had homology of 66%, which increased to 85% for the starting 220 bp. Aligning the entire TSAP and HSAP loci for all species resulted in a low homology of 17%; however, this was 84% for the first 220 bp. Consequently, 5' 220 bp could involved in the process of episome generation through recombination: this is evidenced by the observation of only one such instance of the 220 bp region in the episome fragments, the other presumably remains in the genomic DNA.

The most prominent large-scale duplications classified from a genome-wide scan in 5 kb blocks as potential false positive structurally variable regions may be caused by collapsed paralogous regions, however, evaluating the significance of these duplications requires further refinement of the reference sequence. These were on chr11 at 490-505 kb (including LdBPK_111210 encoding an ABC protein subfamily A); chr12 at 380-385 kb (including LdBPK_120661, a hypothetical gene); chr15 at 175-180 kb; chr29 at 820-825 kb (including LdBPK_291890, a paraflagellar rod protein 1 D); chr33 at 110-115 kb (including LdBPK_330350, a heat shock protein 83); and chr34 at 725-730 kb (including LdBPK_341700, an amastin gene).

Amastin gene family variation

A total of 56 likely amastin-related loci were identified for the major sub-families (α-amastins 28; β-amastins 30; γ-amastins 24A; δ-amastins 8A, 8B, 24B, 31, 34A, 34B, 34D, 34E, 34F and 36) with the exceptions of δ-amastins 18 and 34C as well as γ-amastins 17 (Jackson 2010; Supplementary Table 12). Phylomic evidence of frequent historical gene conversion for certain sub-families demonstrated the disparate modes of diversity generation in the different groups (Supplementary Figure 23). Amastin loci had a total of 152 SNPs in the 17 lines, with a perceptible conservation of the 3' UTRs. Among the 32 coding SNPs, 12 of the 17 protein-level SNP were in just one gene (LdBPK_341150; 34D). A T-Coffee alignment of 224 sequences representing the 53 amastin-related genes (out of 56) with clear orthologues in other *Leishmania* species demonstrated the scope of likely gene conversion in the different amastin

gene sub-families (data not shown). Aligning 83 amastin gene-coding genes provided phylogenetic evidence of possible gene conversion events in *L. donovani* both between (34B and 34D) and within classes (8B, 28, 34E and 34F) based on the differential clustering of genes according to species rather than orthologues.

Patterns of both high conservation and diversity emerged from the different sub-families: the UTR and coding sequences of groups 8A, 8B, 24A, 24B, 28, 30, 31, 34A, 34B, 34F and 36. In stark contrast, classes 34D, 4-34 and to a lesser extent, 34E showed an abundance of intermediate-frequency SNPs, with the variants for each clustered in a single gene. In the context of the local substitution rates and these high dervied allele frequencies, variation in 34D and 34E appeared not to be due to relaxed constraint: this explanation was still viable for class 4-34. Combining these rates with the gene lengths, it was possible to estimate that variation in family 34E was more ancient and thus has been subject to stronger purifying selection than that at 34D or 4-34; however this speculation must be augmented by the differing gene conversion rates between and within amastin families (Jackson 2010).

Supplementary Methods (text)

Sample collection

In vitro antimonial testing of the lines was completed as described previously (Rijal et al. 2007). *In vitro* SSG susceptibility for each line was expressed as an activity index, which was defined as the ratio of the ED50 of a tested line versus the ED50 of the SSG-sensitive reference line (BPK282/0cl4), which was included in each assay. Isolates with an activity index ≤ 2 were considered sensitive to the drug, while isolates with an activity index ≥ 3 or higher were considered to be resistant.

DNA isolation

Cells harvested on the second day of their stationary phase were prepared in HOMEM + 20 % fetal calf serum cultures, and were checked for contamination and other irregularities by inverted microscopy before pellet preparation (including three wash steps with PBS). In order to lyse the cells, the blood and cell culture DNA maxi kit (Qiagen) was used to dissolve pooled pellets of cold PBS (12-13 pellets per 7.5 mL). 7.5 mL ice-cold Buffer C1 and 22.5 mL ice-cold milliQ water were added to each set, and after invertion the samples were kept on ice for 10 minutes. All subsequent steps were followed as described in the kit protocol using the maximum of the suggested incubation times. To isolate the genomic DNA, the protocol in the Qiagen Genomic DNA Handbook was followed: steps 1-4 were completed as described. DNA was eluted with 10.5 mL of isopropanol. Next, the samples were centrifuged at 16 krcf for 15 minutes at 4 °C, and 70% cold ethanol was added. The DNA pellets were resuspended in TE-buffer at 55 °C while shaking at 750 rpm in a Thermomixer. The genomic DNA isolated for line BPK282/0cl4 had a volume of 800 µL with a concentration of 252.4 mg/L, a spectrophotometric absorbance A_{260}/A_{280} value of 1.95 and a median estimated fragment size of 30 kb.

Sample preparation for cDNA and genomic DNA comparison

Pellets were prepared for RNA and DNA analysis as described above in DNA isolation and the RNAqueous-Micro kit (Ambion) was used according to the manual instructions for the RNAqueous-Micro procedure. 100 µL of lysis buffer was used elution for each 10 µL of 20 sample tubes twice for line BPK282/0cl4 created from the pellets. DNase treatment was implemented by adding 1.6 µL of DNase I buffer and 2 µL of DNase inactivation reagent to each tube. This produced 260 µL of DNase-treated RNA at a concentration of 620.37 µg/µL (Nanodrop).

Genomic DNA sequencing and comparative *de novo* assembly

For pyrosequencing on the 454 Life Sciences Genome Sequencer FLX Titanium platform, DNA for sample BPK282/0cl4 was quantified with a Picogreen, and then sheared into fragments using a Hydroshear (Wheeler et al. 2008). The paired-end template library was prepared as suggested by the manufacturer; for more information, see Margulies et al. (2005) and Wheeler et al. (2008). In brief, after end-blunting, circularization adapters were ligated to the fragments and were size-selected with AMPure SPRI beads. After end repair, circularization adapters were ligated to the fragment ends, and these were then circularized using Cre/loxP site-specific recombination and linear molecules degraded using exonuclease. The circularized DNA was nebulised, end repaired, and then linker positive molecules were selected via biotin streptavidin association and library adapters were ligated (as primer sites for amplification and sequencing). This was followed by amplification,

size-selection for fragments of 400-750 bp as above and denaturation to ssDNA. The sample ssDNA was then quantified and diluted. This was followed by emulsion PCR where the DNA was annealed to beads: those with one DNA molecule were separated out in micro reactors. Following clonal amplification steps, DNA carrying beads were enriched via biotin streptavidin association and sequencing primers were annealed. Sequencing reagents were sequentially flowed over the sample (for 800 repitions, 200 per nucleotide) such that polymerases extended the existing DNA strand by adding nucleotides, generating a light signal whose strength and frequency was proportional to the number and type of nucleotides incorporated (Kircher and Kelso 2010). The data was processed and analysed by Genome Sequencer System software.

For Illumina reversible terminator sequencing, the protocol was followed (Quail et al. 2008, 2009). Following concentration quantification, the 17 samples were each sheared into 200-300 bp fragments using Adaptive Focused Acoustics technology (Covaris). This was followed by end repair with T4 and Klenow DNA polymerase and blunt-ending of the fragments with T4 polynucleotide kinase. A single 3' A nucleotide was added to the repaired ends using Exo-Minus Klenow DNA polymerase and dATP to deter concatemerization of templates, limit adapter sequence dimers and increase the efficiency of adapter ligation. Adapters specific to this PCR-free DNA sequencing were then ligated (Kozarewa et al. 2009): these contained primer sites for sequencing and flowcell surface annealing. Ligated fragments were run on an agarose gel, size selected and DNA extracted as per the Qiagen gel extraction kit protocol with dissolution of gel slices at room temperature to avoid heat-induced bias.

The DNA libraries' concentrations were quantified using Agilent Bioanalyser chip and Kapa Illumina SYBR Fast QPCR kit. The libraries were denatured with 0.1 M sodium hydroxide and diluted to 6 pM in a hybridisation buffer to allow the template strands to hybridise to adapters attached to the flowcell surface. Cluster amplification was performed on the Illumina cluster and cBOT stations using the cluster generation V4 kit following the manufacturer's protocol and then SYBRGreen quality controls were performed to measure cluster densities and determine whether if flowcells were suitable for sequencing (Bentley et al. 2008). The DNA was then linearised, blocked and hybrdised using the R1 sequencing primer. The hybridized flowcells were loaded onto the Illumina Genome Analyser (GA) IIX for 76 cycles of sequencing-by-synthesis using the SBS sequencing V4 and V5 kits. The linearization, blocking and hybrdization steps were repeated *in situ* to regenerate clusters, and release the complementary strand for sequencing to which R2 sequencing primer could hybridise; this was followed by another 76 cycles of sequencing to produce paired-end reads (Quail et al. 2008). These steps were performed using proprietary reagents according to manufacturer's recommended protocol (https://icom.illumina.com). Illumina data was analysed using the RTA1.6 and 1.8 analysis pipelines. Illumina sequencing for reference line BPK282/0cl4 was repeated in order to increase accuracy.

Extraneous duplicate reads were removed by including the kDNA contigs in mapping the sequence data for each of the 17 lines to the reference chromosomes, thus increasing the quality of the resultant mapped reads (Supplementary Tables 14 & 15).

Quality control, screening and experimental validation of variation

GC content for *Leishmania* species approaches 60% and thus was a useful tool for eliminating obvious contamination (Supplementary Figure 16). In order to calculate the chromosome-wide coverage distributions, regions with at least one high-quality read mapped were examined. Because empirical coverage distributions frequently conformed to a Poisson rather than a Normal distribution, the median was more representative of the genome-wide coverage distribution than the mean. Consequently, sites within one standard deviation of the median were investigated, yielding a normalised median depth and standard deviation for variant validation.

Measuring local uniqueness was akin to finding motifs of size *k* and was achieved by sorting and then counting the frequency of all sites for *k*. Selecting the non-unique sites, *k* was assessed using a local sliding window along for regions up- and down-stream of each site (on both strands) in order to evaluate the average point at which the local region became unique (Supplementary Figures 17 and 18).

Experimental verification of SNPs by PCR and genotyping

Three forms of experimental SNP validation were implemented: PCR amplification of *L. donovani* BPK206/0cl10 and *L. infantum* JPCM5; PCR with 10 Indian and Nepalese *L. donovani* lines; and Sequenom genotyping of the 17 lines.

A cAMP specific phosphodiesterase (LdBPK_151540) and a cysteine peptidase Clan CA family C2 (LdBPK_270510) genes were selected for PCR amplification in 10 *L. donovani* strains to evaluate the sensitivity to calling SNPs in candidate paralogous regions. Similarly, the high homology of an amastin locus (LdBPK_341150) to other amastin genes and proximity to a contig edge provided a basis on which to optimise SNP screening at ambiguous repetitive regions. PCR primers were constructed for four amplicons spread across these genes using Primer3 software (http://frodo.wi.mit.edu) to have lengths of 20-23 bases, an annealing temperature of approximately 60-63^{o}C, GC contents of about 50% and a GC clamp (Supplementary Table 16). These regions were amplified using KAPA2G Robust PCR kits in both forward and reverse directions for ten samples, two from India (BHU568/0cl1, BHU573/0cl1) and eight from Nepal (BPK282/0cl4, BPK080/0cl1, BPK288/0cl7, BPK298/0cl8, BPK085/0, BPK173/0cl3, BPK294/0cl1, BPK085/0). PCR parameters were optimised for each primer pair. The PCR cycle used was: 30 seconds at 95^{o}C; 25 cycles of 30 seconds strand denaturation at 95^{o}C, 30 seconds polymerase annealing at the T_M, and 60 seconds sequence elongation at 72^{o}C; and a final step of 60 seconds at 72^{o}C. A total of 75 amplifications attempts worked out of 80 (94%).

Amplification of four loci was also successfully completed for BPK206/0cl10 and JPCM5, validating 35 SNPs that were initially detected by mapping BPK206/0cl10 paired-end reads to *L. infantum* reference genome JPCM5: 7 SNPs at LdBPK_060740 (hypothetical gene); 10 SNPs at LdBPK_130190 (hypothetical gene); 7 SNPs at LdBPK_160140 (hypothetical gene); and 11 SNPs at LdBPK_353950 (putative PFPI/DJ-1-like gene). See Supplementary Table 16 for experimental details. One SNP at LdBPK_130190 was not validated due to sequence ambiguity. The genotypes of each of the 35 SNPs were successfully predicated by the Mercator-directed alignment of the *L. infantum* JPCM5 and *L. donovani* reference genomes, thus producing experimental validation of SNP predictions between species derived from both mapping of reads and alignment of assembled genome sequences.

Signal detection of the 289 genotyped SNPs on the Sequenom iplex platform was successful in 99% of cases. In order to assess the relative predictive power of the various parameters used to refine the likely segregating sites, the correlations between the genotyping data with the allele frequencies (AF) determined for the pooled read-depth coverage values of the 17 lines were explored. The numbers of SNPs detected by the genotyping to be true variants increased as the sample size increased based on the inference of the ancestral allele from *L. infantum*.

Principle component analysis algorithm

The genome-wide genetic clustering of lines was also assessed using principle component analysis (PCA) implemented with an in-house Python script. The aggregate AF distance (AFd) was defined as:

$$AFd_{nl} = \sum_{n=1}^{n=N} \sum_{l=1}^{l=N} \sum_{i=1}^{i=S} (A_{in} - A_{il})^2 + (C_{in} - C_{il})^2 + (G_{in} - G_{il})^2 + (T_{in} - T_{il})^2$$

where n and l were line indexes ranging from 1 to N (the number of lines) and i was a SNP index ranging from 1 to S (the number of SNP sites). A, C, G and T were the weighted AF corresponding to the n and l indexes. The weighted AF was defined as $A_{xy} = \sum_{k=1}^{k=D_A} W_A$ where x and y were dummy line indexes and D_A was read depth of a given base A and W_A was a weighted count of A which depends on its BQ. The weight was defined to be $W_A = 1$ for $BQ \geq 25$ and $W_A = 10^{BQ}/10^{2.5}$ for BQ < 25 so that the effect of low-quality bases was reduced. The weighted AF of C, G and T nucleotides were similarly defined.

Selection and population diversity analyses

For the PAML analysis in codeml, the one-ratio (M0) model was used to calculate locus-specific $\omega = d_N/d_S$ values for the set of five species (Yang 1997, Yang 2007) – the models implemented were sensitive to low numbers of species (Anisimova et al. 2001). All five *Leishmania* species were examined for each gene only where the ortholog could clearly be aligned. Comparing models with $\omega = 1$ (neutral) and $\omega > 1$ or $\omega < 1$ (non-neutral) complements previously applied lineage-specific tests for positive selection using *L. infantum*, *major* and *braziliensis* (Peacock et al. 2007). However, genes known to have been subject to positive selection may have $0.5 < \omega < 1$ due to purifying selection when ω values are measured across the entire locus (Swanson et al. 2004) – thus our test for $\omega > 1$ selected only those with the strongest excess of nonsynonymous fixation. Nonetheless, using a genus-wide approach should identify candidates under the strongest ancestral selective pressures.

Branch-site specific models (variable and fixed) estimated ω for each site across the whole gene using a random sites Bayes empirical Bayes (BEB) model seeking specifically sites with elevated ω values in the *L. donovani* lineage (Yang 2007, Yang et al. 2005). The variable model had the same ß-distributed ω classes as the fixed model as well as an additional class where ω could be > 1. The posterior likelihoods for each model were then compared using a likelihood ratio test (LRT) to see if the variable one was significantly more favoured in a χ^2-test. The number of degrees of freedom was the difference between the numbers of parameters of the models. A BEB examination of each variable amino acid position determined the posterior Bayesian probability of the given ω value. A high posterior probability (at least 0.95) for a site's variable ω value being > 1 suggested a particular site could have been under

positive selection, if the variable was significantly favoured (Nielsen and Yang 1998).

A total of 6,416 non-overlapping 5 kb windows were used in the sliding window scan between species. Certain regions were not orthologous: for *L. mexicana* the start of chr20 to 395 kb did not align, neither did the region at 410-750 kb; nor did any of chr36. For *L. major*, sites at 25-95 kb on chr32 and 220-995 kb on chr34 did not align.

Small indel detection

Small indels (< 10 b) were identified and screened with Ssaha2 (Ning et al. 2001) mapping and Samtools (Li et al. 2009), using the same quality parameters for the SNP screening analysis. In addition, each indel was supported by both forward and reverse strands and by at least 25% of the coverage with a minimum read-depth of 8 to avoid false calls. The allele frequency spectrum for indels was estimated based on the most frequent allele since ancestry was unknown. Candidate SVs were checked for all lines using the Artemis Bamview tool for manual visualisation (Carver et al. 2010), and those adjacent to contig edges were excluded.

Determining high copy number variation loci

Larger scale SVs in size and copy number were analysed based on per cell read depth to reflect gene dosage effect. We identified significantly amplified copy number per cell regions using 5 kb non-overlapping windows. We first identified the maximum read depth block at a given position among 17 lines (MDP17). We selected 24 SV candidate blocks out of 6,467 MDP17s based on the criteria that the blocks had their significance z-score greater than 4, where the z-score was defined as $z = (X - \mu)/\sigma$, where z was a z-score, X was the maximum depth value among 17 lines, μ was the average value of all Xs and σ was the standard deviation among Xs. The validity of 24 blocks were checked individually excluding ones contains assembly possible collapsed paralogs, and we identified simplified blocks after combining adjacent blocks as biologically relevant duplications.

Miniexon and ribosomal RNA gene copy number was examined in particular. There were two copies of ribosomal RNA (rRNA) both of which come with a SSU (small subunit) and LSU (large subunit) in chromosome 27. The read depth for these units were individually measured and average over the units to define rRNA copy number since the read depth variation among these units of a given line was small. Copy number values of mini-exons were estimated from read depth of CNV blocks encompassing two mini-exons present on chromosome 2: these values were

Identifying candidate episome regions

Episomal duplications in *Leishmania* have two characteristic features: a significantly large read depth over 10-50kb and direct repeats flanking a duplicated region (Leprohon et al. 2009). We identified episomal duplication candidates based on these two criteria, but we also were mindful of the fact that highly duplicated regions tend to have misassemblies and gaps, hence the direct repeat may not be complete. The presence of direct repeats was checked by aligning the *L. donovani* reference to both itself and to *L. infantum* with BLAST (Altschul et al. 1990). Duplicated regions adjacent to large gaps with multiple paralogous regions were discarded. A somy-level heatmap was created with heatmap.2 in R with the gplots package. To construct a 3D depth coverage plot, the median coverage depth measured in 5 kb blocks for all lines were plotted using Mayavi (Ramachandran and Varoquaux 2011) in Python.

PCR details of experimental validation of candidate episome

DNA presence was confirmed after alkaline lysis by gel electrophoresis. PCR was carried out in a volume of 50 µl, with 5mM each of dATP, dCTP, dGTP and dTTP (Eurogentec SA), 100 µM of each primer, 25 mM magnesium chloride and 5 ul Hotstart+ polymerase (Qiagen). The PCR reaction started with 5 min at 95°C, then 40 repeats of 1 min at 94°C, 1 min at annealing temperature, and 1 min at 72°C. A final extension step of 10 min at 72°C completed the reaction. Sequences of the PCR amplicons were generated using the dideoxy nucleotide chemistry with the ABI PRISM1 BigDyeTM Terminator cycle sequencing kit (PerkinElmer).

Quantification of gene dosage and transcription variability

In order to measure the correlation between the cDNA and genomic DNA of reference strain BPK282/0cl4, the samples were prepared simultaneously as outlined above. *Leishmania* mRNA is co-transcribed in polycistronic transcription units (Martínez-Calvillo et al. 2010), and therefore cDNA abundance was determined using the median read-depth using the same method described for read-depth analysis above. For chromosome level comparisons, the normalised chromosome read-depths were examined. To compare the local copy number variation, the read-depths per cell used; mini-exon loci were excluded since multiple biological factors affect their copy number (Martínez-Calvillo et al 2010). Once the reference genome sequence is further refined, more structural diversity can be identified since there were highly polymorphic regions spanning gaps in the current assembly. The insert size of mapped read-pairs can be used to identify indels, but assembly gaps and misassemblies in the reference genome made this approach insensitive to distinguishing true indels from false ones.

Quantitative PCR

Four genes located on chromosomes 17 (CBS), 31 (AQP1), 34 (SAT) and 36 (CS) were quantified by quantitative PCR (QPCR) in the 14 of the 17 samples sequenced (BHU568/0cl1, BHU573/0cl1, BPK043/0cl2, BPK067/0cl2, BPK080/0cl1, BPK085/0, BPK087/0cl11, BPK173/0cl3, BPK178/0cl3, BPK190/0cl3, BPK275/0cl18, BPK282/0cl4, BPK288/0cl7 and BPK294/0cl1; BPK035/0cl1, BPK298/0cl8 and BPK206/0cl10 were not assessed). These QPCR reactions have been optimized as described in a previous gene expression study on *L. donovani* (de Paiva Cavalcanti et al. 2009) and showed an amplification efficiency of 2. QPCRs were run with the iQ Sybr Green Supermix (Bio-Rad), 2 µL of DNA (12 ng/µl), and the primers in a 25 µl format with the following thermal profile: (i) initial denaturation for 5 min at 95°C, (ii) 36 cycles of denaturation for 30 s at 95°C, annealing for 15 s at 60°C and extension for 15 s at 72°C. Melt curve analysis was done at the end of all QPCRs to verify PCR product quality and primer dimer formation. All QPCRs were run on the LightCycler 480 (Roche) in 96 well plates and raw data (Cp-values) were used to compare the Cp-values of chromosome 17, 32 and 34 (diploid according to sequencing data) to those of chromosome 31 (tetraploid according to sequencing data). GraphPpad Prism 5 was used for statistical analysis of the data.

Supplementary Tables

Supplementary Table 1. *L. donovani* contigs not assigned to chromosomes to which gene models were transferred.

Contig ID	N [1]	Orthol chrom [2]	Gene products
00015_1587	1	27, 31	Amino acid permease
03211_1654	1	10	GP63, leishmanolysin
01230_494	1	7	60S ribosomal gene L7a
01189_1398	1	19	ATG8/AUT7/APG8/PAZ2
01698_1408	1	7, 17	Vacuolar-type Ca2+-ATPase
01232_1813	1	30	40S ribosomal S30
01091_59076	15	14, 27, 31, 32	Radial spoke 3; Replication factor C subunit 4; Protein kinase; 12 hypothetical proteins
00714_848	1	36 at 426,239 bp	Ribosomal protein L24m
01226_2141	1	22	Conserved hypothetical protein
00890_1428	1	8, 31	Amastin-like surface protein
02855_12186	4	7	RNA binding protein; nucleolar RNA-binding protein; 2 hypothetical proteins
00336_505	1	8, 34	Amastin-like surface protein
03073_1073	1	8, 9, 34	Amastin-like surface protein
01542_1060	2	-	snoRNAs [3]
03425_1467	1	-	snoRNA

[1] Number of genes transferred to the contig. [2] *L. donovani* chromosomes orthologous to the contigs. [3] Small nucleolar RNAs.

Supplementary Table 2. Chromosomal distributions of gene transfer completion in *L donovani* from the 8,395 genes in *L. infantum*.

Chr [1]	Genes transferred	
	Successfully	Not
Total	8,252	143
0 [2]	-	16
1	83	1
2	70	1
3	99	-
4	124	-
5	132	-
6	135	1
7	126	1
8	115	3
9	175	2
10	144	3
11	139	2
12	102	9
13	162	-
14	158	2
15	164	6
16	174	-
17	157	6
18	167	1
19	162	6
20	182	2
21	228	2
22	164	4
23	207	5
24	248	-
25	258	-
26	291	1
27	250	3
28	318	6
29	300	3
30	384	1
31	339	4
32	408	4
33	346	3
34	435	14
35	531	20
36	742	11

[1] Chromosome. [2] *L. infantum* contigs containing genes unassigned to chromosomes. The number successfully transferred includes 33 genes spread across 15 contigs not yet assigned to chromosomes. The numbers of genes transferred successfully includes 137 assigned to chromosomes but with a degree of uncertainty regarding the precise nucleotide-level positions due to multiple gene models – consequently, these are not included in the current draft of the published annotation files.

Supplementary Table 3. List of the 17 lines sequenced (BPK282/0cl4 is the reference genome line).

Sample (WHO code)	Alternate names	Geographic origin [1]	Phenotypes [2]		Passage number [3]	Treatment outcome [4]
			SSG	MIL		
MHOM/IN/09/BHU568A1	BHU568/0cl1	Bihar, Muzzafarpur *	6 R	S	8+	Relapsed
MHOM/IN/09/BHU573A1	BHU573/0cl3	Bihar, Sitamadhi *	6 R	S	8+	Cured
MHOM/NP/02/BPK035A1	BPK035/0cl1	Saptari, Bhagani	0 S	-	42	Relapsed
MHOM/NP/02/BPK043A1	BPK043/0cl2	Sunsari, Chatra	0 S	-	43	Cured
MHOM/NP/02/BPK067A1	BPK067/0cl2	Morang, Sundarpur	1 S	-	34	Dead
MHOM/NP/02/BPK080A1	BPK080/0cl1	Sunsari, Ithahari	1 S	-	31	Relapsed
MHOM/NP/02/BPK173A1	BPK173/0cl3	Sarlahi, Pirari	6+ R	S	28	Non responder
MHOM/NP/02/BPK178A1	BPK178/0cl3	Sunsari, Inaruwa	2 S	-	34	Cured
MHOM/NP/03/BPK190A1	BPK190/0cl3	Morang, Govindpur	6+ R	S	51	Non responder
MHOM/NP/03/BPK206A1	BPK206/0cl10	Sunsari, Ithahari	1 S	S	49	Cured
MHOM/NP/03/BPK275A1	BPK275/0cl18	Morang, Schanischare	6+ R	S	44	Non responder
MHOM/NP/03/BPK282A1	BPK282/0cl4	Sunsari, Inaruwa	1 S	S	47	Cured
MHOM/NP/03/BPK288B2	BPK288/0cl7	Saptari, Sitapur	1 S	-	37	Cured
MHOM/NP/03/BPK294A1	BPK294/0cl1	Siraha, Bishnupur	1 S	S	29	Cured
MHOM/NP/03/BPK298A1	BPK298/0cl8	Sunsari, Ithahari	1 S	-	41	Cured
MHOM/NP/02/BPK085B2	BPK085/0 (isolate)	Saptari, Kamalpur	6+ R	-	44	Cured
MHOM/NP/02/BPK087A1	BPK087/0cl11	Morang, Bhodaha	3 R	S	34	Cured

All Nepalese lines were previously typed by MLMT (Bhattarai et al. 2010) and belong to the type M1 except BPK035/0cl1 (M4), BPK043/0cl2 (M5), BPK067/0cl2 (M6), BPK288/0cl7 (M7) and BPK294/0cl1 (M8). All lines were cloned from the parental isolate, except BPK085/0. [1] All samples are Nepalese except where noted (*) – see Supplementary Figure 25. [2] Acitivity index: SSG stands for stibogluconate and MIL for miltefosine; R denotes a drug-resistant phenotype, S susceptible; the remainder were not determined. [3] The number of passages between isolation from patients to DNA elution. [4] Clinical classification (Rijal et al. 2010).

Supplementary Table 4. Noncoding SNPs showing high differentiation between lines resistant and susceptible to SSG treatment.

F_{ST}	Chromosome	Position	S [1]	R [2]
0.54	34	493493 [3,4]	0.88	0.14
0.48	10	296325 [3]	0.17	0.86
0.44	18	119734 [3]	1.00	0.50
0.44	6	279131	1.00	0.43
0.44	9	417791	1.00	0.43
0.44	12	520300	1.00	0.43
0.44	14	579004	1.00	0.43
0.44	20	606613	1.00	0.43
0.44	25	833936	1.00	0.43
0.44	26	743394	1.00	0.43
0.44	26	943103	1.00	0.43
0.44	27	681727	1.00	0.43
0.44	28	397634	1.00	0.43
0.44	29	579621	1.00	0.43
0.44	29	717243	1.00	0.43
0.44	29	1027694	1.00	0.43
0.44	32	849195	1.00	0.43
0.44	33	1240250	1.00	0.43
0.44	34	684395	1.00	0.43
0.44	35	1257940	1.00	0.43
0.44	35	1619958	1.00	0.43
0.44	36	273712	1.00	0.43
0.44	36	2266023	1.00	0.43
0.42	27	162368 [3]	0.33	1.00

SNPs ranked by F_{ST} all have $F_{ST} > 0.4$ (bootstrapping $p < 0.02$) and are noncoding. [1] Frequency of ancestral allele in SSG-S set. [2] Frequency of ancestral allele in SSG-resistant set. [3] Sites not in the set of 34 SNPs where the derived allele was homozygous in SSG-R lines BHU568/0cl1, BHU573/0cl1, BPK173/0cl3 and BPK275/0cl18 but were not in other samples. [4] Three bases after stop codon of amastin gene from sub-family 34D (LdBPK_341150), which had a synonymous high-F_{ST} SNP.

Supplementary Table 5. Fitness effects of substitutions in *L. donovani* compared to rates in orthologous *Leishmania* species regions.

Species	*L. infantum*	*L. major*	*L. mexicana*	*L. braziliensis*	*L. donovani*
$E[D_N/D_S]$	1.12	1.00	1.00	1.08	$E[P_N/P_S] = 1.40$
eFI	0.80	0.72	0.72	0.77	-
α = FI/eFI - 1	-0.23	-0.22	-0.22	-0.22	-
α(Prior)	-0.61	-0.80	-0.80	-0.67	$P_N/P_S = 1.80$
FI [1]	0.74	0.67	0.67	0.72	-
α(Prior) [1]	-0.35	-0.50	-0.50	-0.40	$P_N/P_S = 1.50$
FI [2]	0.89	0.80	0.80	0.86	-
α(Prior) [2]	-0.14	-0.27	-0.27	-0.18	$P_N/P_S = 1.25$

$E[D_N/D_S]$ is the expected value of D_N/D_S; $E[P_N/P_S]$ that for segregating sites (Gojobori et al. 2007). eFI is the effective FI (fixation index; Axelsson and Ellegren 2009). α(Prior) = 1 – 1/FI; the fraction of advantageous mutations (if negative, it suggests an abundance of deleterious variants). [1] For SNPs segregating in the *L. donovani* population with a folded allele frequency > 0.2. [2] For folded allele frequency > 0.4.

Supplementary Table 6. Genes with evidence of positive selection in the *Leishmania* genus.

ω	P value	d_N	d_S	Gene product	Gene ID
1.70	0.019	0.08	0.05	Hypothetical protein	LdBPK_020720
1.57	0.045	0.19	0.12	Conserved hypothetical protein	LdBPK_041180
1.83	0.045	0.19	0.11	Conserved hypothetical protein [1]	LdBPK_100340
1.36	0.040	0.17	0.12	Conserved hypothetical protein	LdBPK_120140
8.63	0.017	0.07	0.01	Calpain-like cysteine peptidase Clan CA family C2	LdBPK_201210
1.94	0.023	0.27	0.14	Conserved hypothetical protein	LdBPK_220620
1.81	0.022	0.25	0.14	Conserved hypothetical protein	LdBPK_240080
2.32	<0.001	0.18	0.08	Acyl phosphatase	LdBPK_252040
1.73	0.029	0.02	0.02	Histone H2A	LdBPK_291850
2.32	<0.001	0.12	0.05	Histone H2A	LdBPK_291870
1.90	0.013	0.22	0.12	Conserved hypothetical protein	LdBPK_310660
2.30	0.020	0.24	0.11	60S ribosomal protein L22 [1]	LdBPK_364640

Genes were determined to show evidence of positive selection where the variable PAML one-ratio model had ω (d_N/d_S) > 1 across the entire gene sequences and was significantly ($p < 0.05$) more likely than a fixed one-ratio model with $\omega = 1$ for alignments of each orthologous *Leishmania* sequence (Yang 2007). Likelihood ratio test (LRT) p values were determined from a chi-square distribution with 1 degree of freedom. Genes with $d_S = 0$ were not included. Differential histone gene expression is associated with antimonial resistance in *L. donovani* (Singh et al. 2010). Genes containing repetitive sequence may have artefactually high ω values due to biases in mutation ascertainment and the repetitive region length. [1] Previously shown to have variable ω among *Leishmania* species (Peacock et al. 2007).

Supplementary Table 7. Genes subject to selective processes from this study that also had variable $\omega = d_N/d_S$ values between *Leishmania* species in a previous study (Peacock et al. 2007).

Gene product	Gene ID	Tests
Phosphatidylinositol 3 kinase-like	LdBPK_020100	High F_{ST} SNP
Cathepsin L-like protease [1]	LdBPK_080960	Direction of selection
Conserved hypothetical protein	LdBPK_100340	PAML one-ratio
Conserved hypothetical protein	LdBPK_300400	Direction of selection
Hypothetical protein	LdBPK_301140	$D_N/D_S > 2$
Conserved hypothetical protein	LdBPK_301640	High-frequency CNV
Kinetoplastid membrane protein 11	LdBPK_352250	Negative Tajima's D
Kinetoplastid membrane protein 11	LdBPK_352260	Negative Tajima's D
Hypothetical protein	LdBPK_342360	Direction of selection
Vacuolar sorting associated protein	LdBPK_360490	PAML branch-site
60S ribosomal protein L	LdBPK_364640	PAML one-ratio

Note that genes with strong signatures of adaptation or differentiation within *L. donovani*, or on the *L. donovani* lineage, or within *Leishmania* were included. [1] Predicted intermediate-frequency nonsynonymous SNP A20D in signal peptide domain (amino acid sites 1-27).

Supplementary Table 8. Genes with sites under selection on the *L. donovani* lineage.

Gene product	Gene ID	P value	Sites
Conserved hypothetical protein	LdBPK_351510	< 0.001	107L, 108E
Conserved hypothetical protein	LdBPK_312610	< 0.001	470V, 546V
Vacuolar sorting associated protein [1]	LdBPK_360490	< 0.001	381N
Conserved hypothetical protein	LdBPK_354250	< 0.001	257A
Conserved hypothetical protein	LdBPK_030730	< 0.001	200F
Amastin-like surface protein	LdBPK_341730	< 0.001	148N
Conserved hypothetical protein	LdBPK_241970	0.001	652D
Dynein intermediate chain	LdBPK_332770	0.004	644R
Conserved hypothetical protein	LdBPK_180970	0.028	623V

Genes were ranked by likelihood ratio test (LRT) scores. Branch-site fixed and variable models were computed in PAML (Yang 2007) to estimate $\omega = d_N/d_S$ for sites subject to positive selection on the *L. donovani* branch where the variable model was significantly more favoured by a LRT (χ^2 p value) and the Bayes Empirical Bayes probability of the sites shown having $\omega > 1$ was greater than 0.95 (Yang et al. 2005). [1] Previously shown to have variable ω among *Leishmania* species (Peacock et al. 2007).

Supplementary Table 9. Genes with mean positive direction of selection (DoS) test values between the 17 *L. donovani* lines the *Leishmania* species and with one or more intraspecific nonsynonymous mutation.

Gene product	Gene ID	*L. infantum*	*L. major*	*L. mexicana*	*L. braziliensis*	α(Prior)
Conserved hypothetical protein	LdBPK_040410	0.20	0.03	0.05	0.08	0.52 [1]
Cathepsin L protease [2]	LdBPK_080960	0.06	0.21	0.22	0.19	1.27 [1]
Hypothetical protein [3]	LdBPK_120667	0.25	-0.09	-0.03	0.13	0.22
Tryparedoxin peroxidase	LdBPK_151140	0.50	0.22	0.39	0.08	2.24 [1]
Nucleoside transporter 1	LdBPK_151230	0.50	-0.03	0.08	0.17	0.41
CAMP specific phosphodiesterase	LdBPK_151540	0.36	0.30	0.28	0.30	3.98 [1]
Receptor-type adenylate cyclase a	LdBPK_170120	0.36	0.22	0.22	-	2.21 [1]
Receptor-type adenylate cyclase [4]	LdBPK_170130	0.15	-0.04	-0.04	0.00	0.05
Hypothetical transmembrane protein	LdBPK_240440	0.15	-0.10	-0.03	0.08	0.03
Conserved hypothetical protein	LdBPK_300400	0.00	0.00	-0.04	0.06	0.02
Conserved hypothetical protein [5]	LdBPK_311340	-0.15	0.04	0.06	0.11	-0.03
Conserved hypothetical protein	LdBPK_331070	-0.02	0.06	0.03	0.10	0.15
Hypothetical protein [6]	LdBPK_342360	-0.11	0.27	0.21	0.27	1.10 [1]
Conserved hypothetical protein [7]	LdBPK_342390	0.21	0.18	0.14	0.12	0.99 [1]
Mt aminomethyltransferase	LdBPK_364000	0.00	0.12	0.67	0.10	0.44

Genes ordered by gene ID whose mean direction of selection test (DoS) value was positive, such that $DoS = D_N/[D_N + D_S] – P_N/[P_N + P_S]$ so that the higher the DoS value, the greater the ancestral relative rate of nonsynonymous mutation fixation (Stoletzki and Eyre-Walker 2010). Only genes with $P_N > 0$ were assessed. The fraction of segregating beneficial protein-level variants in the 17 lines estimated as $\alpha(Prior) = 1 – 1/FI$ for each of the comparison species (Smith and Eyre-Walker 2002). [1] High α(Prior) where the mean > 0.5. [2] Predicted intermediate-frequency nonsynonymous SNP A20D in signal peptide domain (amino acid sites 1-27). [3] Located in an extended region of elevated π. [4] Predicted singleton nonsynonymous SNP W841C in transmembrane domain (amino acid sites 819-841). [5] Had two SNPs (S268P, 426T) with $F_{ST} = 0.44$ (Table 2). [6] Previously shown to have variable ω among *Leishmania* species (Peacock et al. 2007). [7] Had the SNP (H532R) with the highest F_{ST} value (0.54).

Supplementary Table 10. Genes with high relative rates of nonsynonymous polymorphism fixation ($D_N/D_S > 2$) in *Leishmania*.

Gene product	Gene ID	ω	D_N/D_S			
			L. infantum	*L. major*	*L. mexicana*	*L. braziliensis*
Aminoacylase N-acyl-L-amino acid amidohydrolase	LdBPK_201730	-	High	4	High	High
Conserved hypothetical protein	LdBPK_291660	2.38	High	11	9.67	2.94
Histone H2A	LdBPK_291870	2.32	7	3.33	4.27	3.39
Conserved hypothetical protein	LdBPK_220620	1.94	High	3.83	2.71	3.35
Conserved hypothetical protein	LdBPK_292340	1.93	High	4.33	3.8	4.29
Conserved hypothetical protein	LdBPK_240080	1.81	6	5	4.5	4.4
Conserved hypothetical protein	LdBPK_240430	1.65	3	3.5	6.75	2.44
Conserved hypothetical protein	LdBPK_352880	1.59	High	2.63	3.67	2.29
Conserved hypothetical protein	LdBPK_120140	1.40	7	3.9	3	2.52
Phosphoglycan ß 13	LdBPK_210010	1.29	5	3	3	3
Calcium-binding protein	LdBPK_301300	1.26	5	2.5	2.69	2.48
Conserved hypothetical protein	LdBPK_353700	1.13	2.57	2.11	2.08	2.42
Conserved hypothetical protein	LdBPK_354610	1.07	4	2.81	2.19	2.18
40S ribosomal protein S3 [2]	LdBPK_151010	1.04	High	2.27	2.06	2.1
U1 small nuclear ribonucleoprotein	LdBPK_161690	1.03	High	3.25	2.81	2.49
Conserved hypothetical protein	LdBPK_310270	1.03	High	2.2	2.32	2.35
Conserved hypothetical protein	LdBPK_354270	0.99	High	2.38	2.48	6
Conserved hypothetical protein	LdBPK_312150	0.98	5	2.1	2.43	2.75
Conserved hypothetical protein	LdBPK_313130	0.97	High	2.24	2.07	2.19
Conserved hypothetical protein	LdBPK_354280	0.91	7	2.4	2.12	2.27
Conserved hypothetical protein	LdBPK_262180	0.86	3.63	2.44	2.14	2.32
Conserved hypothetical protein	LdBPK_302130	0.86	High	2.71	5.57	2.36
Conserved hypothetical protein	LdBPK_302010	0.85	5.33	2.15	2.18	2.17
Conserved hypothetical protein	LdBPK_301140	0.81	2.56	2.1	2.09	2.03
Conserved hypothetical protein	LdBPK_330190	0.80	High	3.67	2.29	2.7
Phosphatidylinositol-4-phosphate 5-kinase	LdBPK_262520	0.77	High	2.78	3.05	2.83
Conserved hypothetical protein	LdBPK_090750	0.76	High	2.76	2.79	3.38
Conserved hypothetical protein	LdBPK_070040	0.71	3	3.05	2.32	2.1
Conserved hypothetical protein	LdBPK_010650	0.65	2.5	2.11	2.16	2.13
Conserved hypothetical protein	LdBPK_160510	0.63	High	3.88	3.17	2.96
Phosphoglycan ß 12	LdBPK_020190	0.50	2.25	2.34	2.35	3
Conserved hypothetical protein	LdBPK_191150	0.41	High	2.28	2.15	2.48
Conserved hypothetical protein	LdBPK_150210	0.32	2.67	3.94	2.82	3.73
Hypothetical protein	LdBPK_191690	-	4	4	4	4
Conserved hypothetical protein	LdBPK_013360	-	High	3.85	4.03	4.47

Proteins were ranked by $\omega = d_N/d_S$ value where L_N and L_S could be estimated. D_N/D_S values were calculated in alignments for each species with *L. donovani*. Proteins for which $D_S = 0$ have D_N/D_S ratios denoted as high. Highly divergent regions not present in all five species, candidate pseudogenes and genes whose lengths were less than 300 bp were not included.

Supplementary Table 11. Known genes in regions with extended outlying characteristics from intraspecific genome scans.

Extended local signal	Gene ID	Gene product
Positive Tajima's D	LdBPK_220530	Glucoamylase protein
Positive Tajima's D	LdBPK_220540	Prefoldin 5 protein
Positive Tajima's D	LdBPK_220560	3'a2rel-related protein
Positive Tajima's D	LdBPK_220590	NADH-cytochrome b5 reductase
Positive Tajima's D	LdBPK_220600	NADH-cytochrome b5 reductase
Positive Tajima's D	LdBPK_220630	Protein kinase serine/threonine sos2
Positive Tajima's D	LdBPK_220660	5'a2rel-related protein
Negative Tajima's D	LdBPK_190540	Metallo-peptidase Clan MG Family [1]
Negative Tajima's D	LdBPK_190590	Protein kinase
Negative Tajima's D	LdBPK_131120	40S ribosomal protein S4
Negative Tajima's D	LdBPK_131130	40S ribosomal protein S4
Negative Tajima's D	LdBPK_131170	DNA/RNA non-specific endonuclease
Negative Tajima's D	LdBPK_131180	XPA-interacting protein
Negative Tajima's D	LdBPK_131280	Endomembrane protein
Negative Tajima's D	LdBPK_352230	60S ribosomal protein L12
Negative Tajima's D	LdBPK_352240	RNA-binding protein
Negative Tajima's D	LdBPK_352250	Kinetoplastid membrane protein-11
Negative Tajima's D	LdBPK_352260	Kinetoplastid membrane protein-11
Negative Tajima's D	LdBPK_352320	ATP-dependent RNA helicase protein
Negative Tajima's D	LdBPK_352370	Protein kinase
High π and θ_W	LdBPK_310350	Amino acid transporter aATP11 [2]
High π and θ_W	LdBPK_310360	Amino acid transporter aATP11
High π and θ_W	LdBPK_310370	Amino acid transporter aATP11

Hypothetical or unknown genes are not shown. Genes with high π and θ_W were detected in order to identify variable regions for which Tajima's D was neutral. [1] Methionine aminopeptidase. [2] Predicted rare nonsynonymous SNP A343S in transmembrane domain (amino acid sites 327-349). Scans were conducted to find consecutive sliding windows of high diversity (π and θ_W) and Tajima's D values.

Supplementary Table 12. Genomic regions and variability of known amastin gene families in *L. donovani*.

Family	N [1]	C [2]	Start (bp)	End (bp)	L [3]	Gene ID	SNPs [4]				
							Inter	P_N	P_S	5'	3'
8A	1	8	93231	94547	1.3	LdBPK_080270				1	
8B	11	8	309435	354558	45.1	LdBPK_080710-0820	8	1 [7]			
24A	8	24	450887	481039	30.2	LdBPK_241250-1320	2				
24B	7	24	818719	841737	23.0	LdBPK_242230-2290					
28	3	28	525151	539992	14.8	LdBPK_281500-1520	1				
30	3	30	270334	279025	8.7	LdBPK_300910-0930					
31	2	31 [5]	157148	157531	1.0	LdBPK_310460-0470					
34A	1	34	202867	205572	2.7	LdBPK_340530					
34B	7	34 [5]	412204	452033	39.9	LdBPK_341000-1060	14				1
34D	2	34	493496	496310	2.8	LdBPK_341150-1160	11	12 [8]	11 [9]	1	
34E	7	34	717420	754169	36.8	LdBPK_341670-1750	49	1 [10]	3 [11]		
34F	4	34	1146524	1160433	13.9	LdBPK_342630-2660	28				3
36	2	36	475562	481235	5.7	LdBPK_361330-1340	1		1 [12]		
4-34 [6]	2	34 [6]	1727574	1728152	0.8	LdBPK_044340-4350		3 [13]		1	

[1] Number of loci in family. [2] Chromosome. [3] Length of region (kb) including unassigned contigs. [4] Families showing variation within the 17 *L. donovani* lines at intergenic (Inter), nonsynonymous (P_N), synonymous (P_S), 5' UTR (5') and 3' UTR (3') sites. 5' and 3' UTRs were defined as 1 kb adjacent to the gene family region. [5] As a result of shared homology with multiple genomic regions, providing evidence of recombination between amastin families, LdBPK_310460, LdBPK_341030 and LdBPK_044350 were located on non-assigned extra-chromosomal contigs. [6] LdBPK_044340 and LdBPK_044350 have been assigned to a proposed candidate subfamily named 4-34. [7] In amastin gene LdBPK_080710 with a folded allele frequency of 0.01. [8] In amastin gene LdBPK_341150 with mean folded allele frequency = 0.24 ± 0.18. [9] In LdBPK_341150 with mean folded allele frequency = 0.38 ± 0.09. [10] In amastin gene LdBPK_341690 with a folded allele frequency = 0.47. [11] In LdBPK_341690 with mean folded allele frequency = 0.33 ± 0.09. [12] In conserved hypothetical gene LdBPK_361340 with folded allele frequency = 0.04. [13] In amastin gene LdBPK_044340 with mean folded allele frequency = 0.31 ± 0.04. An additional candidate amastin gene (LdBPK_291450) was noted but not included here: it showed no variation within *L. donovani*.

Supplementary Table 13. Association of changes in chromosome copy number between SSG-resistant and -sensitive strains.

Chrom	SSG-sensitive average ploidy	SSG-resistant average ploidy	Difference (resistant minus sensitive)	MWU-test p value
1	0.832	0.982	-0.149	0.016
2	0.999	1.295	-0.296	0.018
3	0.852	0.933	-0.081	0.007
4	0.963	1.001	-0.038	0.217
5	1.173	1.071	0.103	0.217
6	0.928	1.076	-0.148	0.032
7	1.169	1.027	0.142	0.279
8	1.406	1.507	-0.101	0.142
9	1.282	1.372	-0.090	0.142
10	0.952	1.028	-0.076	0.049
11	1.239	1.297	-0.059	0.102
12	1.065	1.109	-0.044	0.102
13	1.210	1.071	0.138	0.163
14	1.225	1.258	-0.033	0.423
15	1.215	1.033	0.182	0.217
16	1.158	1.192	-0.034	0.142
17	0.929	0.968	-0.039	0.010
18	0.988	1.024	-0.036	0.164
19	0.968	0.979	-0.011	0.247
20	1.177	1.345	-0.169	0.039
21	0.925	0.942	-0.017	0.102
22	1.072	1.049	0.023	0.383
23	1.627	1.444	0.183	0.423
24	0.979	0.963	0.016	0.049
25	0.941	0.956	-0.015	0.263
26	1.085	1.265	-0.180	0.131
27	1.043	1.044	-0.002	0.170
28	0.983	0.978	0.005	0.441
29	1.017	1.138	-0.120	0.124
30	1.007	0.981	0.026	0.046
31	2.003	1.983	0.020	0.312
32	1.152	0.984	0.169	0.053
33	1.494	1.365	0.129	0.119
34	0.996	0.964	0.032	0.028
35	1.338	1.385	-0.047	0.313
36	1.005	0.966	0.039	0.009

Mann-Whitney U-tests were performed to compare non-parametrically the average ploidy levels (s) in SSG-resistant and -sensitive strains using a threshold of biological significance: differences of average s > 0.5 (thus at least a change from disomy to trisomy in the population of strains). The test results indicated a statistically significant difference of s in several chromosomes ($p < 0.05$), but the differences were all below 0.5.

Supplementary Table 14. Gene dosage of copy number variable regions on clinical *L. donovani* lines.

Region		Gene Identifiers	Gene names and/or functions	Copy number SSG-R/SSG-S (median)
Chr	Position			
11	263,794-269,716	LdBPK_110690	Hypothetical protein	3.1/2.4; deleted BPK067/0cl2
23	86,402-100,043 (H-locus) [1]	LdBPK_230270	Hypothetical protein	7.9/7.3; duplicated in all 17 lines
		LdBPK_230280	Terbinafine resistance gene (yip1); resistance an inhibitor of ergosterol biosynthesis	
		LdBPK_230290	MRPA (PGPA); ABC-thiol transporter, SSG resistance [1]	
		LdBPK_230300	Arginosuccinate synthase; arginine biosynthesis	
23	439,984-444,339	Noncoding; low complexity repeats	None	3.6/4.1; duplicated in BPK080/0cl1, BPK087/0cl11, BPK178/0cl3 and BPK190/0cl3
26	345,382-400,300	LdBPK_261090-1200	All 12 have hypothetical proteins as products, except an aldo/keto reductase (for LdBPK_261190)	2.5/1.9; duplicated in BPK067/0cl2
29	16,594-22,201 [2]	LdBPK_290050	Hypothetical protein	2.3/2.0; partial deletion in BPK178/0cl3, BPK282/0cl4 and BPK288/0cl7
		LdBPK_290060	Tryptophanyl-tRNA synthetase	
30	566,243-586,196	LdBPK_301640-1670	All four encode hypothetical proteins [4]	3.9/3.9; duplicated in all 17 except BPK294/0cl1
31	149,819-158,702	LdBPK_310470	Protein with a c2 domain	7.3/6.2
31	661,662-673,839 [3]	LdBPK_311470	Hypothetical protein; adjacent to pentamidin resistance gene (Coelho et al 2003; LdBPK_311460)	18.0/13.0
31	1,046,005-1,054,726 [2]	LdBPK_312180	Hypothetical gene	4.1/4.2; duplicated in BPK080/0cl1, BPK087/0cl11, BPK190/0cl3, BPK206/0cl10 and BPK298/0cl8
		LdBPK_312190	Hypothetical gene	
31	1,451,676-1,456,563	Noncoding, no repeats	None	4.2/4.7; duplicated in BPK085/0, BPK178/0cl3, BPK282/0cl4 and BPK288/0cl7

No significant biologically relevant difference in copy number was observed between antimony-resistant and -sensitive lines except on locus chromosome 31 (at 661,662-673,839, encompassing LdBPK_311470). Copy number is expressed by cell (average number per phenotype) to measure possible impact on SSG resistance; small-scale structural mutations were identified using normalized read-depth per haploid. [1] High expression in antimony-resistant *L. infantum* lines (Leprohon et al. 2006); MRPA gene LdBPK_230290 is associated with SSG-resistance (Leprohon et al. 2009a). [2] Locus is flanked by RIMEs. [3] The single copy gene in the duplication was not flanked by short repeats and RIMEs: it showed read-depth coverage variation between SSG-R and -S lines (18 vs 13); the upstream pentamidin resistance gene displayed no coverage variability. [4] LdBPK_301640 had variable ω among *Leishmania* species (Peacock et al. 2007).

Supplementary Table 15. Coverage for gDNA and cDNA for strain BPK282/0cl4 for highly copy number variable genes.

Gene ID	Read-depth coverage			Gene product
	gDNA		cDNA	
	Per cell	Per haploid genome		
LdBPK_080720	17	9.8	10.8	Amastin
LdBPK_100510	20	10.6	47.1	GP63 leishmanolysin metallo-peptidase Clan MA(M) Family M8
LdBPK_120661	19.2	10.0	28.3	Conserved hypothetical protein
LdBPK_120667	19	9.9	28.0	Hypothetical protein
LdBPK_130330	19.8	7.5	125.1	Alpha tubulin
LdBPK_170170	27.6	14.8	115.4	Elongation factor 1-alpha
LdBPK_191350	13.4	7.1	6.4	Glycerol uptake protein
LdBPK_230270	9.4	3.2	6.7	Conserved hypothetical protein
LdBPK_230280	8.4	2.8	8.8	Terbinafine resistance locus protein
LdBPK_230290	7.8	2.7	7.5	ABC-thiol transporter
LdBPK_230300	8	2.7	2.5	Argininosuccinate synthase
LdBPK_231240	8	2.7	9.8	Hydrophilic surface protein 2
LdBPK_283000	12.8	6.5	86.8	Heat-shock protein hsp70
LdBPK_291890	12	6.1	51.5	Paraflagellar rod protein 1D
LdBPK_300710	7.2	3.6	9.9	40S ribosomal protein S30
LdBPK_310350	8.6	2.1	0.6	Amino acid transporter aATP11
LdBPK_310470	14	3.5	5.9	Tuzin
LdBPK_310910	7.6	1.9	1.4	Amino acid transporter
LdBPK_311930	8	2.0	49.8	Ubiquitin-fusion protein
LdBPK_313190	8.6	2.2	8.6	Iron/zinc transporter protein
LdBPK_330370	23.2	8.5	53.1	Heat shock protein 83-1
LdBPK_341730	7.4	3.7	1.9	Amastin
LdBPK_342660	11	5.5	4.1	Amastin
LdBPK_351870	9	2.9	64.9	60S ribosomal protein L5
LdBPK_360210	7.6	3.8	34.3	Elongation factor 2
LdBPK_366560	12.4	6.1	34.7	Glucose transporter
LdBPK_366740	8	4.0	2.4	Tartrate-sensitive acid phosphatase
LdBPK_366750	26.4	13.0	17.3	Hypothetical protein
LdBPK_366760	32.4	15.9	20.6	Mitogen activated protein kinase
LdBPK_366770	18.2	8.9	9.6	Histidine secretory acid phosphatase

A comparison of normalised read-depth coverage values of genomic DNA (gDNA) and cDNA for strain BPK282/0cl4, for the 30 protein encoding genes showing most variable copy number among the 17 lines of current study.

Supplementary Table 16. Sequencing and mapping output for each line excluding kDNA contigs.

Line	Genomic bases	Coverage		Millions of reads [1]							
		Median	Mean	Total	Mapped		Mapped & paired		Singleton		Both mapped
					Number	%	Number	%	Number	%	
BHU568/0cl1	31,236,355	33	39.0	18.34	16.96	92.5	15.41	90.86	0.91	5.0	16.06
BHU573/0cl1	31,241,223	43	49.5	22.81	21.42	93.9	19.54	91.22	0.95	4.2	20.47
BPK035/0cl1	31,242,084	75	89.0	45.66	40.29	88.3	32.38	80.37	3.83	8.4	36.46
BPK043/0cl2	31,229,768	49	56.2	28.79	25.25	87.7	21.39	84.71	2.32	8.1	22.93
BPK067/0cl2	31,240,611	60	73.0	39.76	34.13	85.8	28.06	82.22	3.74	9.4	30.39
BPK080/0cl1	31,242,516	73	84.8	40.14	37.19	92.7	34.16	91.85	2.05	5.1	35.14
BPK173/0cl3	31,241,114	52	59.3	24.60	23.45	95.3	21.87	93.26	0.94	3.8	22.51
BPK178/0cl3	31,242,142	67	78.0	26.57	25.58	96.3	24.11	94.25	0.74	2.8	24.84
BPK190/0cl3	31,240,232	76	85.6	26.41	25.29	95.8	23.94	94.66	0.76	2.9	24.53
BPK206/0cl10	31,244,049	43	49.3	41.96	36.21	86.3	30.15	83.26	3.82	9.1	32.38
BPK275/0cl18	31,244,867	72	83.4	38.74	37.12	95.8	34.74	93.59	1.30	3.4	35.82
BPK282/0cl4	31,238,467	61	69.7	22.32	21.10	94.6	19.65	93.13	1.01	4.5	20.09
BPK288/0cl7	31,243,837	73	84.3	36.66	35.65	97.2	34.22	95.99	0.81	2.2	34.85
BPK294/0cl1	31,240,440	54	63.0	32.09	30.12	93.9	28.22	93.69	1.27	4.0	28.85
BPK298/0cl8	31,245,712	34	43.8	40.27	36.93	91.7	33.69	91.23	2.20	5.5	34.72
BPK085/0	31,242,735	49	55.0	30.65	27.77	90.6	24.47	88.12	2.18	7.1	25.59
BPK087/0cl11	31,238,069	54	60.1	23.83	20.43	85.7	15.77	77.19	2.33	9.8	18.10

[1] The total number of Illumina reads excluding those identified as PCR duplicates.

Supplementary Table 17. Sequencing and mapping output for each line including kDNA contigs.

Line	Millions of reads [1]							
	Total	Mapped		Mapped & paired		Singleton		Both mapped
		Number	%	Number	%	Number	%	
BHU568/0cl1	18.03	17.49	97.0	16.56	94.68	2.12	2.1	17.11
BHU573/0cl1	22.59	22.03	97.5	20.79	94.37	1.83	1.8	21.61
BPK035/0cl1	43.26	41.34	95.6	36.18	87.52	3.53	3.5	39.82
BPK043/0cl2	27.30	26.11	95.6	24.28	92.99	3.11	3.1	25.26
BPK067/0cl2	37.97	35.81	94.3	32.29	90.17	4.54	4.5	34.09
BPK080/0cl1	39.61	38.30	96.7	36.59	95.54	2.67	2.7	37.24
BPK173/0cl3	24.62	23.72	96.4	22.33	94.14	3.18	3.2	22.94
BPK178/0cl3	26.58	25.89	97.4	24.65	95.21	2.07	2.1	25.34
BPK190/0cl3	26.34	25.83	98.1	24.94	96.55	1.55	1.6	25.42
BPK206/0cl10	39.81	37.66	94.6	34.34	91.18	4.31	4.3	35.94
BPK275/0cl18	38.67	37.48	96.9	35.39	94.42	2.70	2.7	36.43
BPK282/0cl4	22.27	21.35	95.8	20.13	94.29	3.66	3.7	20.53
BPK288/0cl7	36.67	35.90	97.9	34.62	96.43	1.80	1.8	35.24
BPK294/0cl1	31.31	30.81	98.4	29.96	97.24	1.37	1.4	30.38
BPK298/0cl8	39.21	38.04	97.0	36.34	95.53	2.40	2.4	37.10
BPK085/0	30.32	28.69	94.6	26.38	91.95	4.73	4.7	27.25
BPK087/0cl11	23.33	21.65	92.8	18.32	84.62	5.43	5.4	20.38

[1] The total number of Illumina reads excluding those identified as PCR duplicates; because the kDNA contigs were included, more duplicate reads could be identified.

Supplementary Table 18. Primer and amplicon details for genes amplified by PCR.

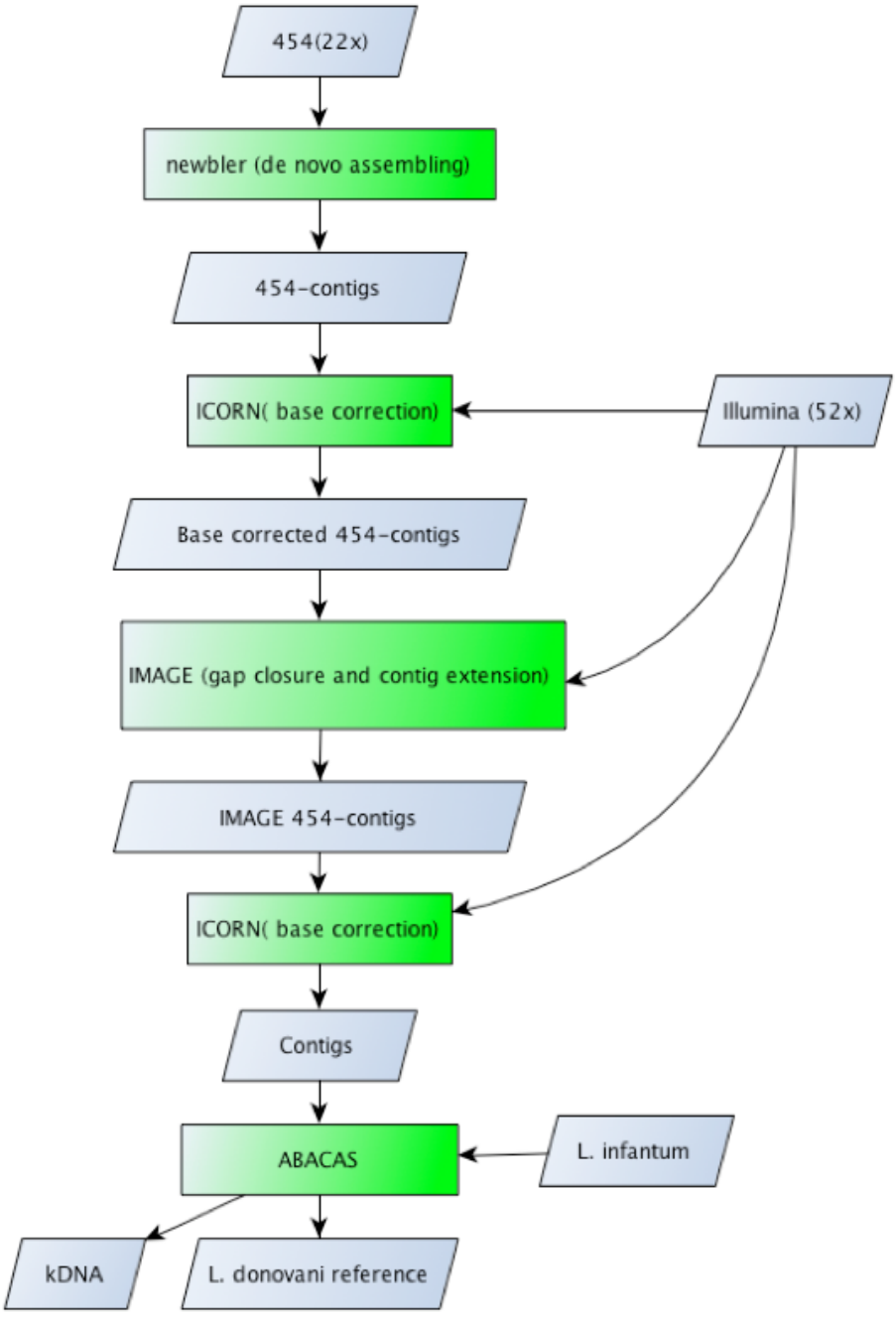
Gene product	C [1]	Gene location		Amplified region		Length (bp)	SNP number	
		Start	End	Start	End		Pre [2]	Post [3]
CAMP specific phosphodiesterase	15	588864	591686	589830	590295	465	2 (9)	1 (7)
5’ primer = AGACTATGTGCACTACGCAACG; 3’ = GATGGAATCTTCGAGGTTGATG; T_M [4] = 63°C								
Cysteine peptidase Clan CA family C2	27	172455	189107	174120	174485	365	0 (21)	0 (1)
5’ primer = CACCGACGTTCATATACTTCTCC; 3’ = GTGTGTCTACGCCTTCGCTATC; T_M = 60°C								
Cysteine peptidase Clan CA family C2	27	172455	189107	189325	189790	465	0 (21)	0 (1)
5’ primer = AATTACCGAGATGTGGCCTTC; 3’ = ACGTTGTGGTTCGAGTTGTTC; T_M = 60°C								
Amastin	34	493496	494167	493386	493873	487	23 (42)	9 (23)
5’ primer = AGACTATGTGCACTACGCAACG; 3’ = GATGGAATCTTCGAGGTTGATG; T_M = 50°C								
^ Hypothetical protein LdBPK_060740	6	283306	283827	283209	283835	627	n/a	7
5' primer = CACACTATCTCCAACAATGGTCACAGAGGCGGGGG; 3’ = CACGCTATCTCCAACAATGCTCACAGAGACGGGGG 5' primer = GGTGCTACGGCTATGGTC; 3’ = GGTGCTATGGCTATGGTC 5' primer = GGCCTTCGCCCAGTA; 3’ = GGCCTTCACCCAGTA 5' primer = CGAGCTGAGCGTT; 3’ = CGAGCTCAGCGTT 5' primer = GTCTGACGACGCTGG; 3’ = GTCTGACAACGCTGG 5' primer = GCCAGGCCTGTCCAT; 3’ = GCCAGGCGTGTCCAT								
Hypothetical protein LdBPK_130190	13	57839	58585	57912	58681	770	n/a	10
5' primer = GCGTCCGCAT; 3' = CGTCAGCAT 5' primer = TTGACGCTTGGAGACG; 3' = TTGACGCATGGAGACG 5' primer = CGCGCCGTGTAGTAGGCG; 3' = CGCGCCGTATAGTAGGCG 5' primer = CCCTCGAGCGTGTGAC; 3' = CCCTCGAACGTGTGAC 5' primer = TGCAGGAGCTCGTTGTT; 3' = TGCAGGAGCGCGTTGT 5' primer = CGCAACGCCGCTGCGG; 3' = CGCAGCGCCGTTGCGG 5' primer = TGCCGCTGCTGCACCGACTCT; 3' = TGCCGCTGCTGTACCGACTCT 5' primer = CCCGGGGCCATGCTGCCG; 3' = CCCGGGGCCCTGCTGCC 5' primer = GTGGAGGTCTCGTA; 3' = GTGGAGGCCTCGTA 5' primer = CAGCCACAGATG; 3' = CATCCACAGATG								
Hypothetical protein LdBPK_160140	16	34734	35360	34864	35433	570	n/a	7
5' primer = TGCAGTGCACAGGAGAGGTCC; 3' = TGCAGTGCACGGGAGAGGTCC 5' primer = ACCGACTGTGGGTGCTGC; 3' = ACCGACTGTAGGTGCTGC 5' primer = TTCATAGCAGCCTTTC; 3' = TTCATAGAAGCCTTTC 5' primer = GCTGTGGGTTCATCTAAC; 3' = GCTGTGGGCTCATCTAAC 5' primer = GCAGTGTCGAAGTTGGCAAT; 3' = GCAGTGCCGAAGTTAGCAAT 5' primer = CTCCTCCACTGCACTCT; 3' = CTCCTCCATTGCACTCT								
PFPI/DJ-1-like LdBPK_353950	35	1553851	1553883	1553261	1553944	684	n/a	11
5' primer = TTCTGTCTTTCGAAAAA; 3' = TTCTGTCCTTTGAAAAA 5' primer = TGACTTCGTGGAGGACGA; 3' = TGACTTCGTAGAGGACGA 5' primer = CCAGACATATGTGGAGG; 3' = CCAGACATATATGGAGG 5' primer = TTAGGAACTCATTT; 3' = TTAGGAACCCATTT 5' primer = AAGTGGATGTAGTA; 3' = AAGTGAATGTAGTA 5' primer = CGCCAAAGTTCATTCG; 3' = CGCCAAAGTACATTCG 5' primer = TCGCCGCTTCTTTAACC; 3' = TCGCCACTTCTTCAACC 5' primer = CCCGTTGCTGCCACTT; 3' = CCCGTTGCGGCCACTT 5' primer = TGCAGTCTCGCCGGGCATGATAA; 3' = TGCAGTATCGCCGGGAATGATAA 5' primer = TCGGGTCACTAGGCGC; 3' = TCGGGTGACTAGGCGC								

Legend to Supplementary Table 17:

The primer sequences and annealing temperatures used are listed below each gene. [1] Chromosome. [2] Prior to screening SNPs, includes all possible variants in the amplicon – the values in parentheses refer to the number of (unscreened) SNPs in the entire gene. [3] After screening. [4] T_M is the annealing temperature. Regions amplified to verify SNPs within the 17 lines are marked as *****. Regions amplified to verify SNPs between *L. donovani* and *infantum* are marked as ^. For the latter, SNPs were validated at sites 283309, 283318, 283484, 283570, 283605, 283696 and 283735 for LdBPK_060740; at 58012, 58046, 58061, 58158, 58161, 58164, 58293, 58364, 58478, 58506 and 58581 for LdBPK_130190; at 34964, 35069, 35077, 35097, 35240, 35314 and 35333 for LdBPK_160140; and at 1553361, 1553452, 1553461, 1553539, 1553560, 1553567, 1553606, 1553658, 1553673, 1553709 and 1553844 for LdBPK_353950.

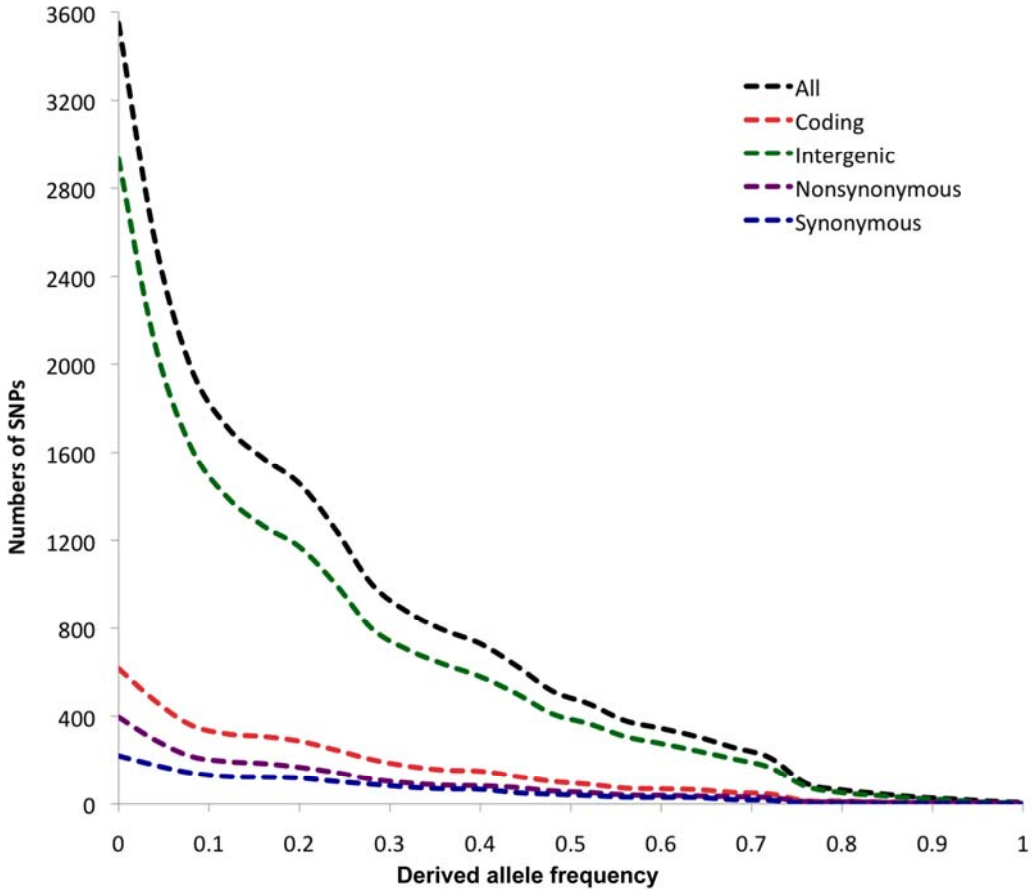
Supplementary Figures

Supplementary Figure 1. The assembly, mapping and sequence improvement processes to create a reference *L. donovani* genome.



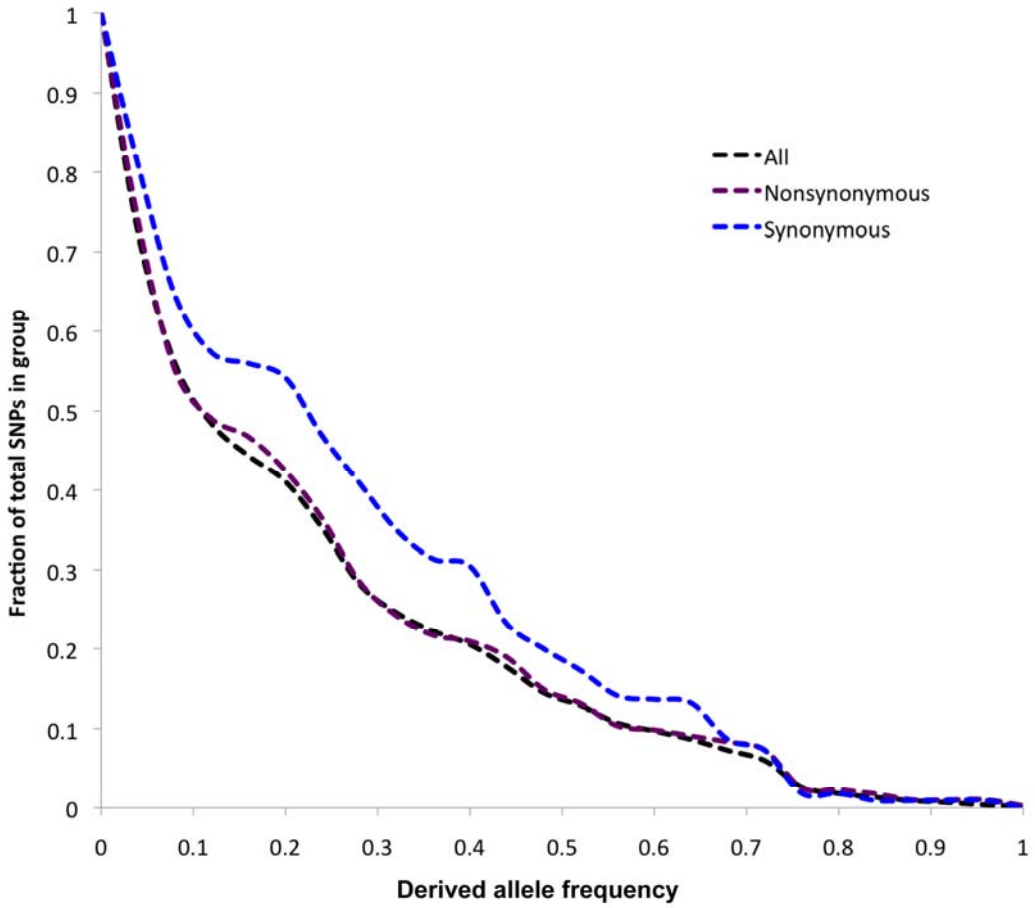
Datasets with coverage values are shown as grey; bioinformatic tools are shown as green. The vast majority of the Illumina data was incorporated with the 454 contigs at the first ICORN (Otto et al. 2010) step; however additional reads were used when the contigs were re-mapped with IMAGE (Tsai et al. 2010) and in the second ICORN step.

Supplementary Figure 2. SNP population genomic derived allele frequency spectra.



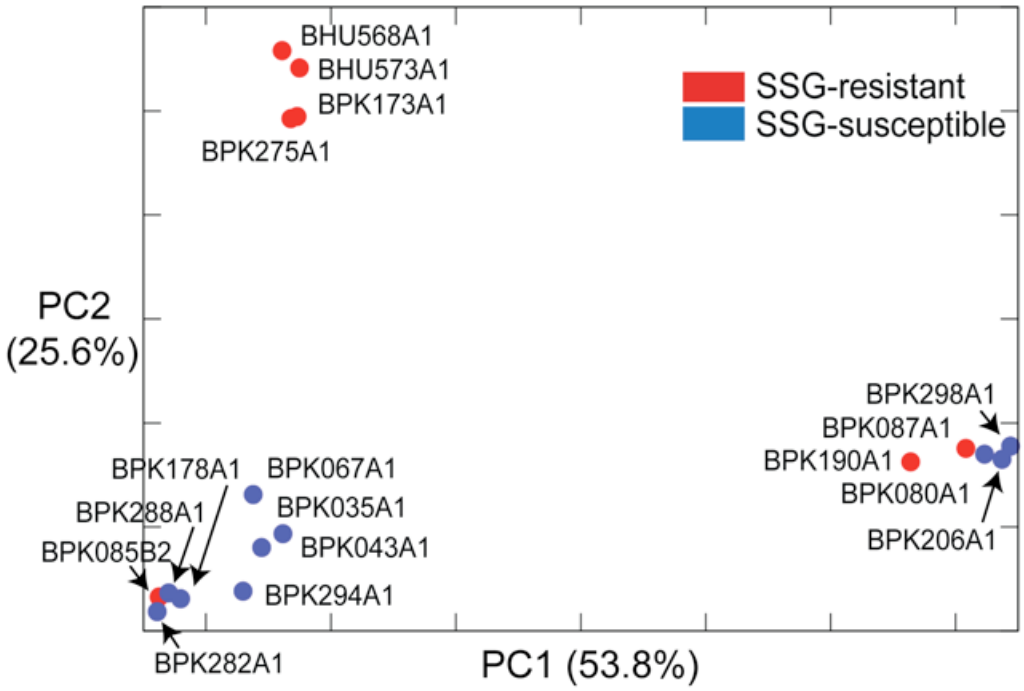
Derived SNPs were ascertained using from read-depth coverage values using the *L. infantum* allele as the ancestral allele. The numbers of SNPs in the genome (3,549, black) as well as coding (616, red), intergenic (2,933, green), nonsynonymous (396, grey) and synonymous (220, blue) sites are shown. The derived allele frequency decay for each site class followed a pattern of exponential decay, which was evidence of purifying selection acting more acutely on high-frequency variants, as expected.

Supplementary Figure 3. SNPs as a fraction of the total in each group relative to their derived allele frequency assigned using *L. infantum*.



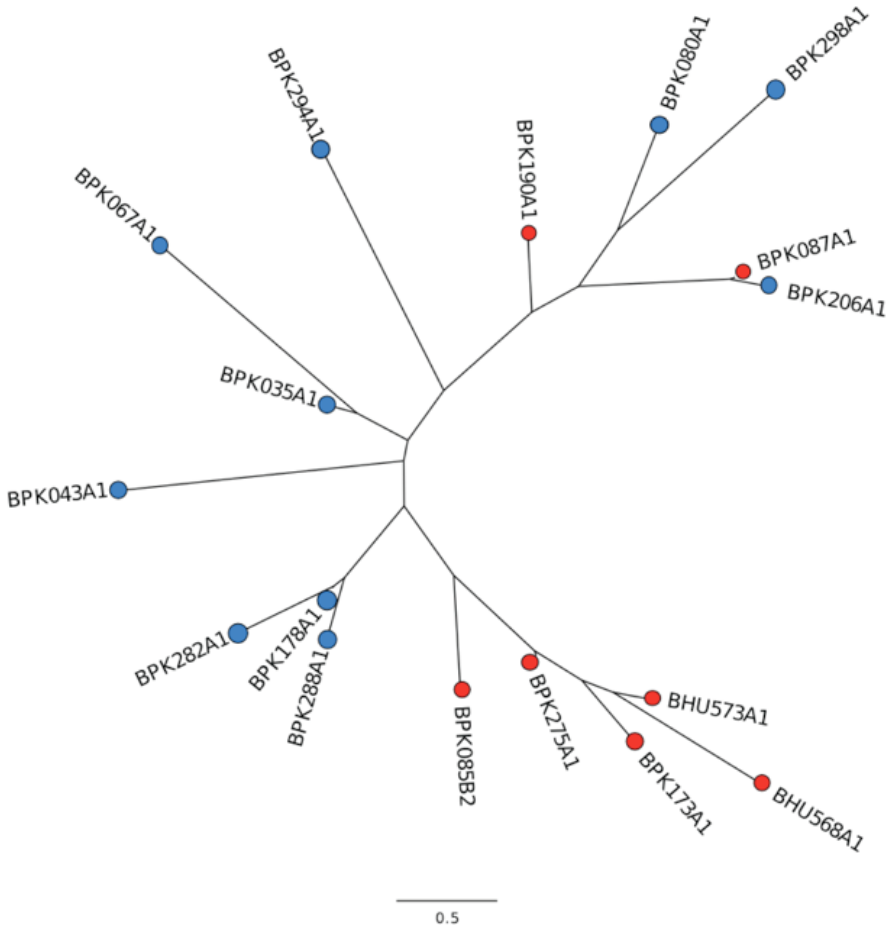
SNP numbers relative to their incidence in the population 17 lines are shown for the genome ("All", black) nonsynonymous regions (grey) and synonymous (blue) sites. Derived alleles for the *L. donovani* strains were inferred using the *L. infantum* genome sequence allele as the ancestral type. The pattern of higher derived allele frequencies for synonymous sites compared to nonsynonymous suggests purifying selection may be stronger on the latter.

Supplementary Figure 4. A Principle Component Analysis (PCA) plot of nonsynonymous variation within the *L. donovani* lines resistant (red) and susceptible (blue) to SGG.



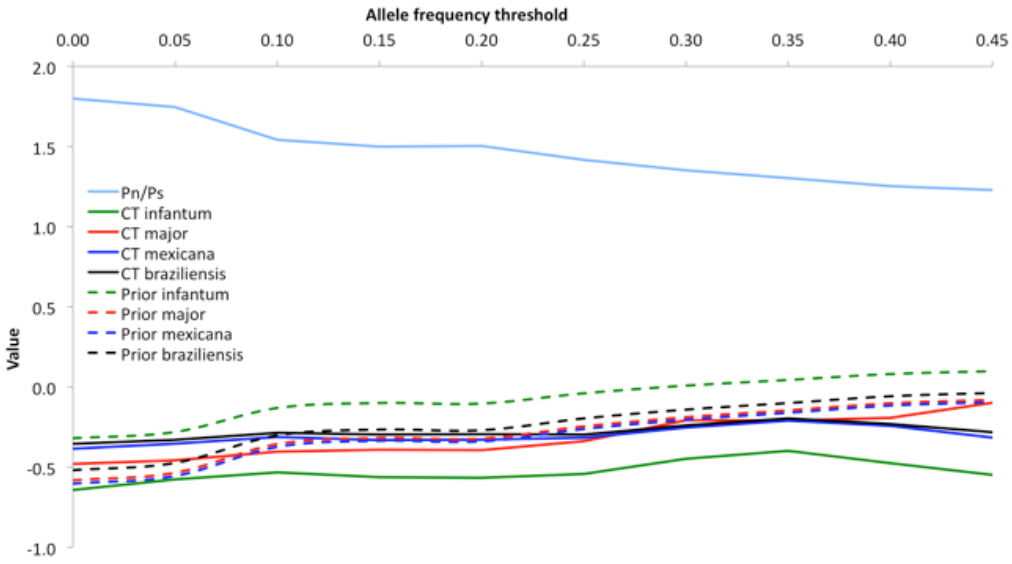
The two most significant components account for 79.4% (53.8% and 25.6%) of the total variation. PCA of the 17 lines' genomes also produced a PC3 accounting for 17.8% of the total variation, and a smaller PC4 (9.3%); these did not noticeably differentiate the lines.

Supplementary Figure 5. Phylogenetic relationship of *L. donovani* clinical lines for large CNVs.



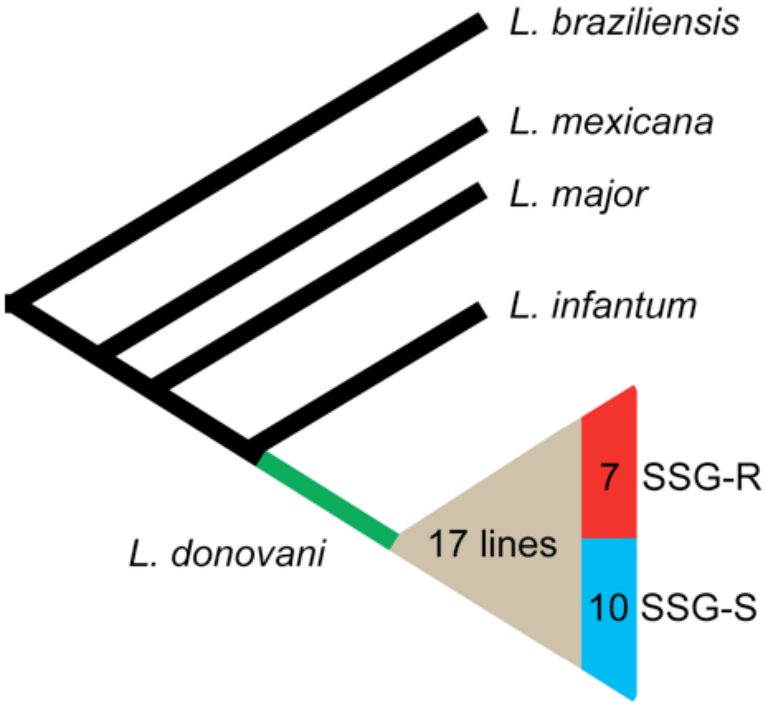
A neighbour-joining tree was constructed using the copy numbers for each large CNV (1 kb in size or longer) from the normalised coverage levels to determine genetic distances. The copy numbers were adjusted for normalised read-depths (d_{n2}) relative to the minimum (min_d) and maximum (max_d) values in the set of samples such that $d_{n2} = (d - min_d)/(max_d - min_d)$. The branch lengths displayed are proportional to the genetic distances between lines. Strains from India begin with BHU and those from Nepal with BPK. Strains BHU568/0cl1, BHU573/0cl1, BPK173/0cl3, BPK275/0cl18 and BPK085/0 clustered closely; all were SSG-resistant lines (all of the remainder except BPK190/0cl3 and BPK087/0cl11 were SSG-susceptible).

Supplementary Figure 6. The neutrality of coding sequence variants relative to the read-depth coverage folded allele frequency spectrum.



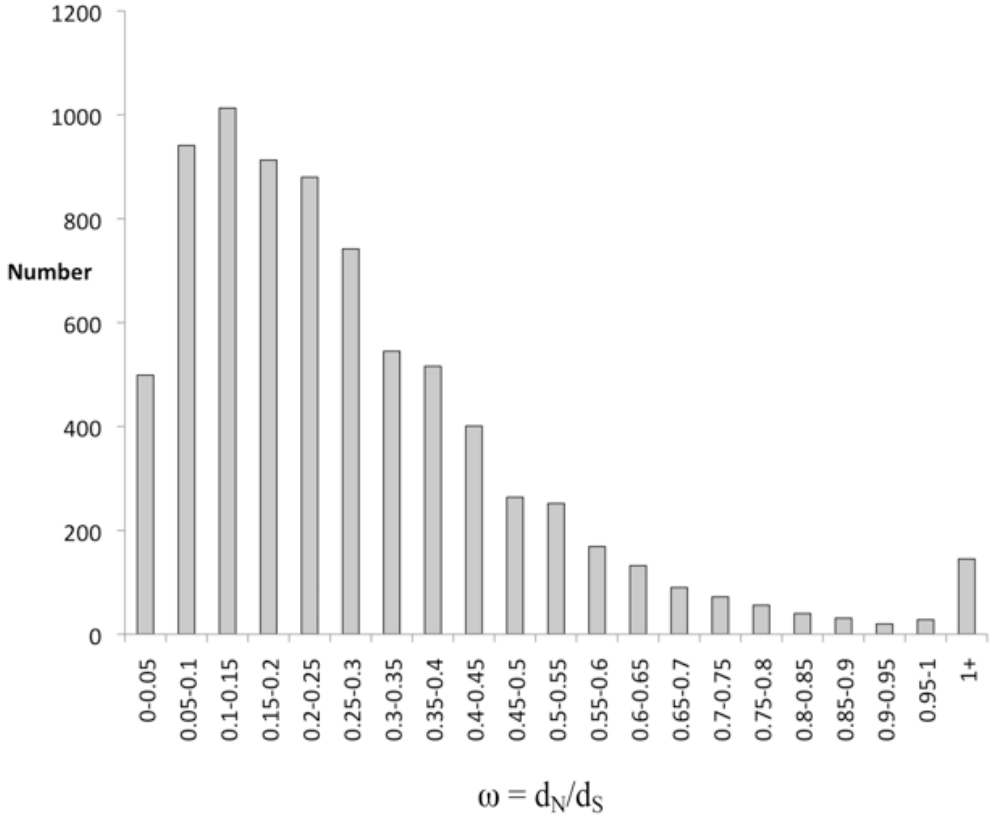
The dyanmics of folded allele frequency (DAF) spectra for mutations at nonsynonmous (P_N) and synonymous (P_S) sites as P_N/P_S (light blue) as well as α, the relative abundance of adaptive substitutions, from contingency table (CT, Gojobori et al. 2007) values and the averaged rate of $\alpha = 1 - 1/FI$ across all loci (Prior, Smith and Eyre-Walker 2002) for each species: *L. infantum* (green), *L. major* (red), *L. mexicana* (blue) and *L. braziliensis* (black). The allele frequency spectrum was folded to allow for segregating rare deleterious ancestral alleles.

Supplementary Figure 7. Differential tests of adaptive evolution in *Leishmania* (black), the *L. donovani* lineage (green), within the 17 *L. donovani* lines (grey) and between the SSG-resistant (red) and -susceptible (blue) lines.



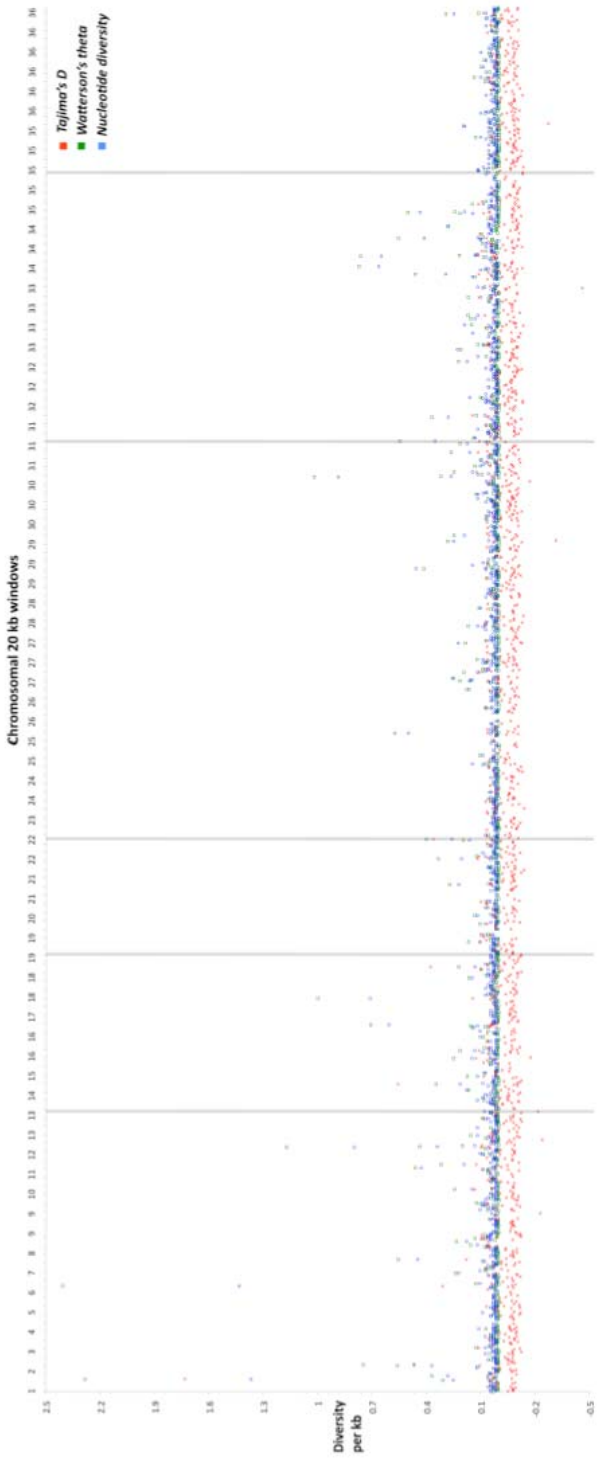
The variation between species is shown (black) in its phylogenetic context to that on the *L. donovani* branch (green) and within the 17 *L. donovani* lines (grey). Branch lengths are figurative and are not proportional to phylogenetic distance. SSG stands for sodium stibogluconate, R for resistant and S for susceptible.

Supplementary Figure 8. The genome-wide distribution of gene ω values.



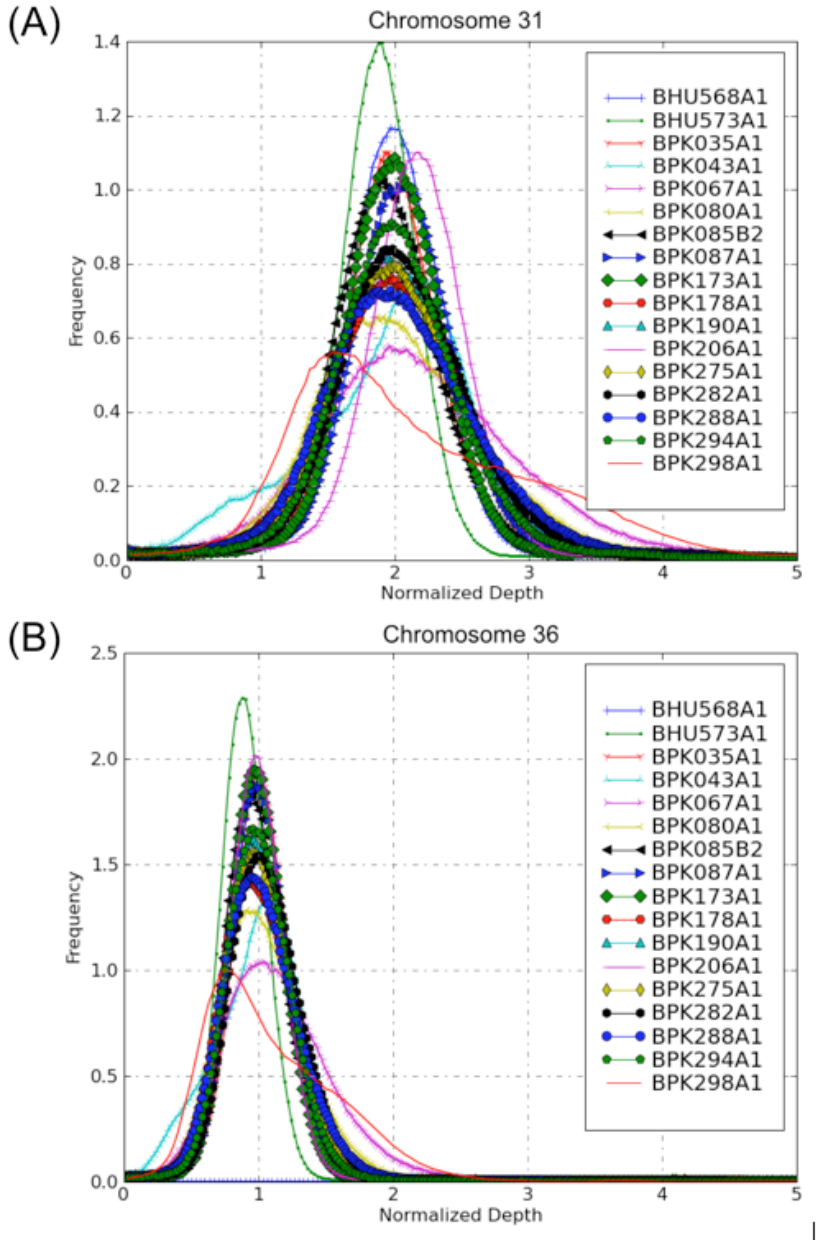
The one-ratio model in the codeml implementation of PAML (Yang 2007) was used to conservatively estimate $\omega = d_N/d_S$ for each chromosomal gene. Genes with $\omega > 1$ were binned into a "1+" category. Candidate pseudogenes and genes with $d_S = 0$ or $d_N = 0$, as well as those of less than 50 known amino acid sites were omitted.

Supplementary Figure 9. Genomic sliding window of variation in the 17 lines.



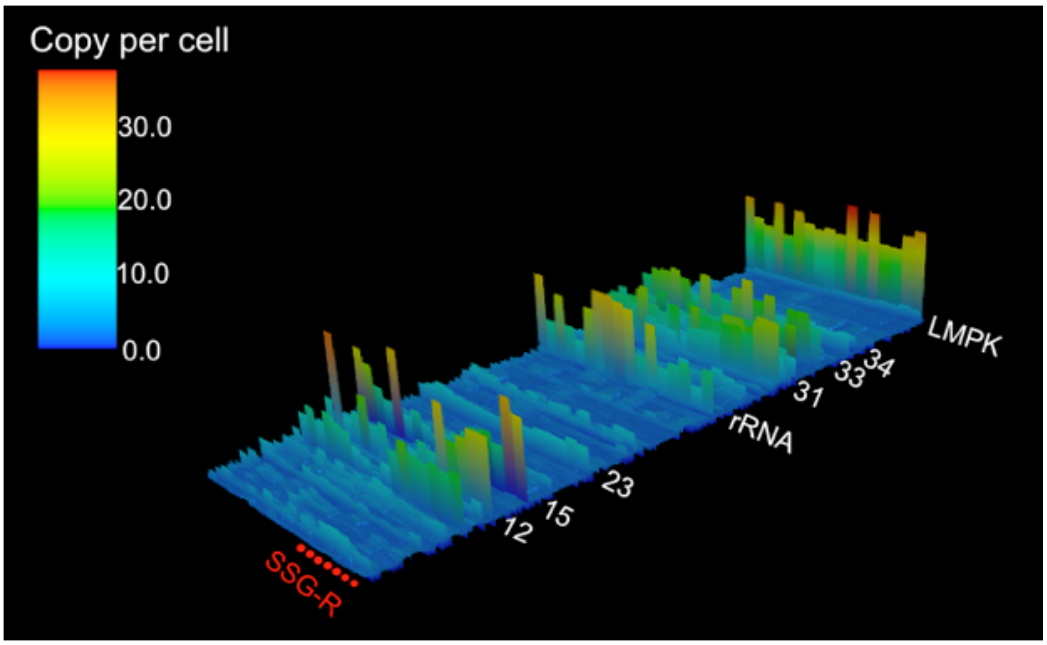
Each window was non-overlapping and 20 kb in length. Nucleotide diversity (π, blue) and Watterson's theta (θ_W, green) are shown per kb and Tajima's D is shown for each 20 kb window. A longer window size was used to avoid low numbers of SNPs. Regions near chromosomes ends were excluded. Shaded areas indicate extended regions of outlying diversity values.

Supplementary Figure 10. Ploidy variation in *L. donovani* lines: (A) tetrasomy of the chromosome 31 and (B) disomy of the chromosome 36.



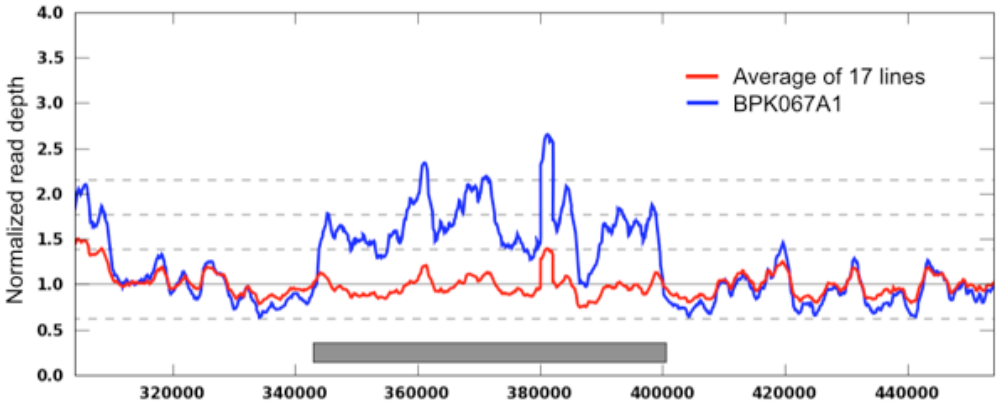
The horizontal and vertical axes show normalised read depth and the frequency of normalised read depth of a given chromosome. The median depth that separates the area under a curve into two equal regions indicates the ploidy status: 1 and 2 indicates disomy and tetrasomy state of the chromosomes, respectively. The histogram was normalised so that the area under each curve had the same size to accommodate the difference in the yields of individual sequencing in order to illustrate general trend among all lines. Sequencing runs with high read output have wider distributions since depth is Possion-distributed. The distribution of some lines (BPK298/0cl8) deviated slightly from a Possion distribution not because sequencing sampling effects.

Supplementary Figure 11. Genomic sliding windows of inter-chromosomal structural diversity in the 17 *L. donovani* lines.



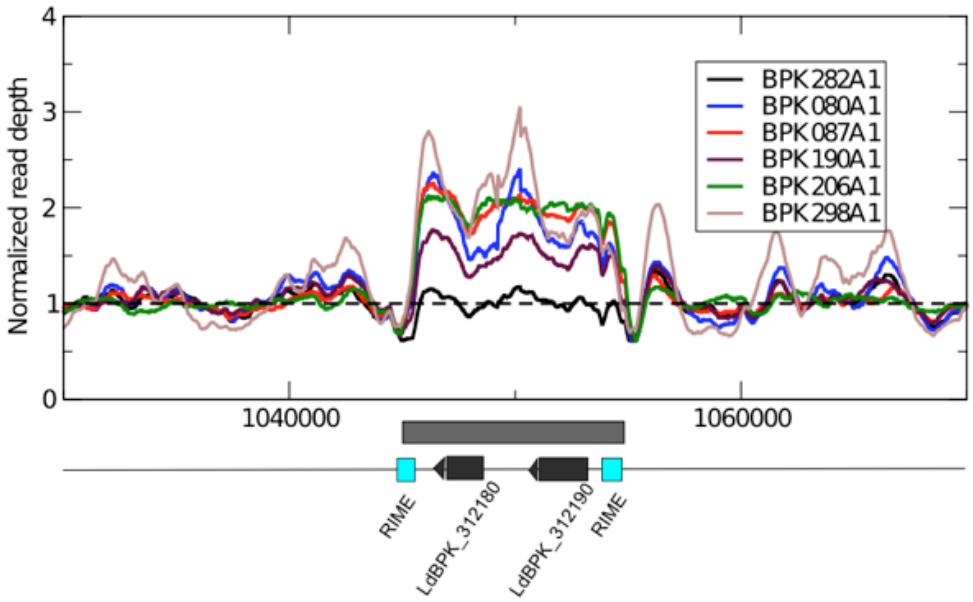
Antimony-resistant lines (SSG-R) are marked with a red circle: going from right to left, these are BHU568/0cl1, BHU573/0cl1, BPK085/0, BPK087/0cl11, BPK173/0cl3, BPK190/0cl3, BPK275/0cl18, and sensitive lines BPK035/0cl1, BPK043/0cl2, BPK067/0cl2, BPK080/0cl1, BPK178/0cl3, BPK206/0cl10, BPK282/0cl4, BPK288/0cl7, BPK294/0cl1 and BPK298/0cl8. Read-depth per cell (per diploid genome adjusted for ploidy variability) of successive 5-kb segments of large SVs for the 36 chromosomes is shown. Prominent large-scale duplications are marked with their locus or chromosome name. The large duplications whose origin could be explained by collapsed paralogs are not shown, including blocks containing amastin genes at chromosomes 12 and 34. The peaks of the mini-exon loci are not apparent in this figure because gaps surrounding the mini-exon region suppressed the median depth. Chromosomes 12, 15, 23, 31 and 33 are highlighted, as is the rRNA gene cluster on chromosome 27 and the extended copy number variable region on chromosome 36, the MAPK-locus (LMPK).

Supplementary Figure 12. A large duplication of 50kb located in chromosome 26.



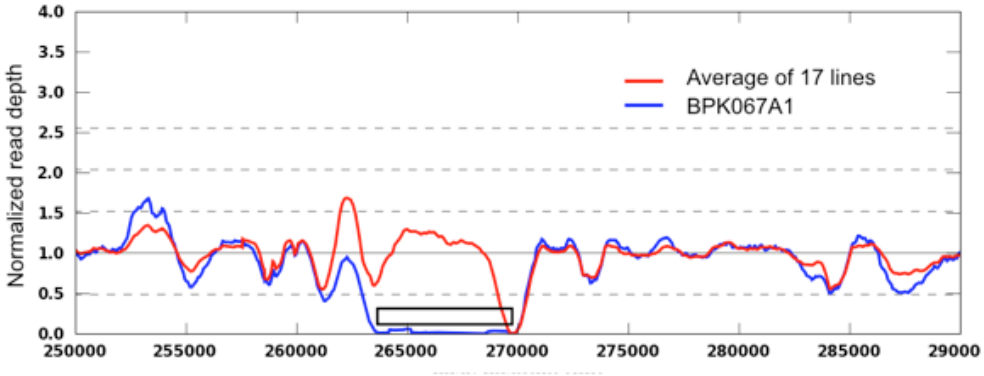
This large duplication of 50kb located was on chromosome 26 of BPK067/0cl2 in which 12 gene models were located (LdBPK_261090 to LdBPK_261200). The dark grey stripe represents the duplicated region and the light grey line passes through one represents the normalised mean depth. The grey dashed lines indicate standard deviation of BPK067/0cl2 to show the significance of the depth variation. Coverage of BPK067/0cl2 (blue) and the average for the 17 lines (red) is shown. The x-axis indicates genomic position in bases.

Supplementary Figure 13. A group-specific duplication on chromosome 31.



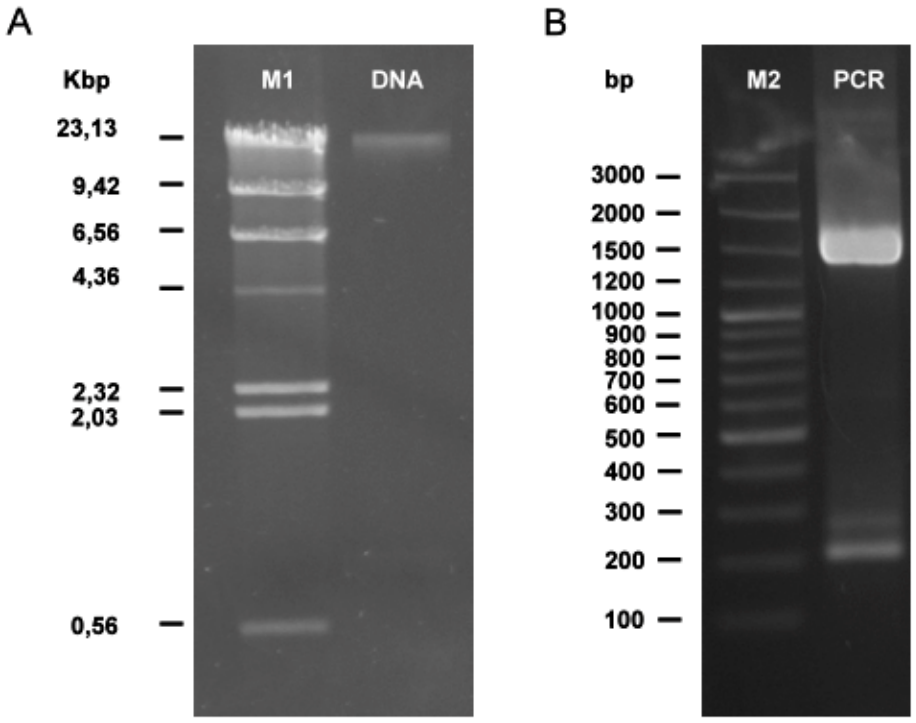
The duplication (shown as a grey stripe) contained two genes of unknown function was flanked by two RIMEs (ribosomal mobile elements). The normalised read depths are shown in colour specified in the legend for BPK282/0cl4 (black), BPK080/0cl1 (blue), BPK087/0cl11 (red), BPK190/0cl3 (purple), BPK206/0cl10 (green) and BPK298/0cl8 (brown). This duplication was only found in the kDNA A class group consisting of BPK080/0cl1, BPK087/0cl11, BPK190/0cl3, BPK206/0cl10 and BPK298/0cl8. The depth of non-duplicated BPK282/0cl4 and the depth base line (black dash) were shown for comparison. The black dash line passing through one is the baseline of normalised read depth. The x-axis indicates genomic position in bases.

Supplementary Figure 14. A 6kb deletion in chromosome 11 of BPK067/0cl2.



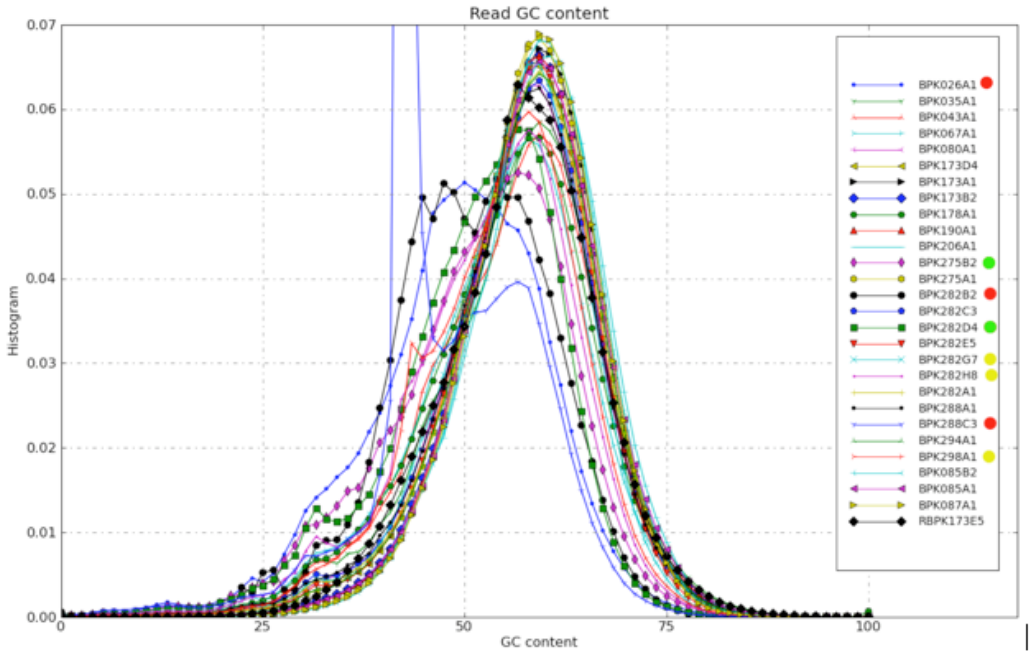
The 6kb deletion in chromosome 11 of BPK067/0cl2 is marked with a black frame box. No other lines had this deletion. LdBPK_110690 of unknown function was located at the 3' end of the deleted region. The grey dashed lines indicate the gird of the standard deviation of BPK067/0cl2. The grey line passes through 1.0 is the baseline genomic coverage. Coverage of BPK067/0cl2 (blue) and the average for the 17 lines (red) is shown. The horizontal axis indicates genomic position.

Supplementary Figure 15. Verification of candidate episome region by gel electrophoresis.



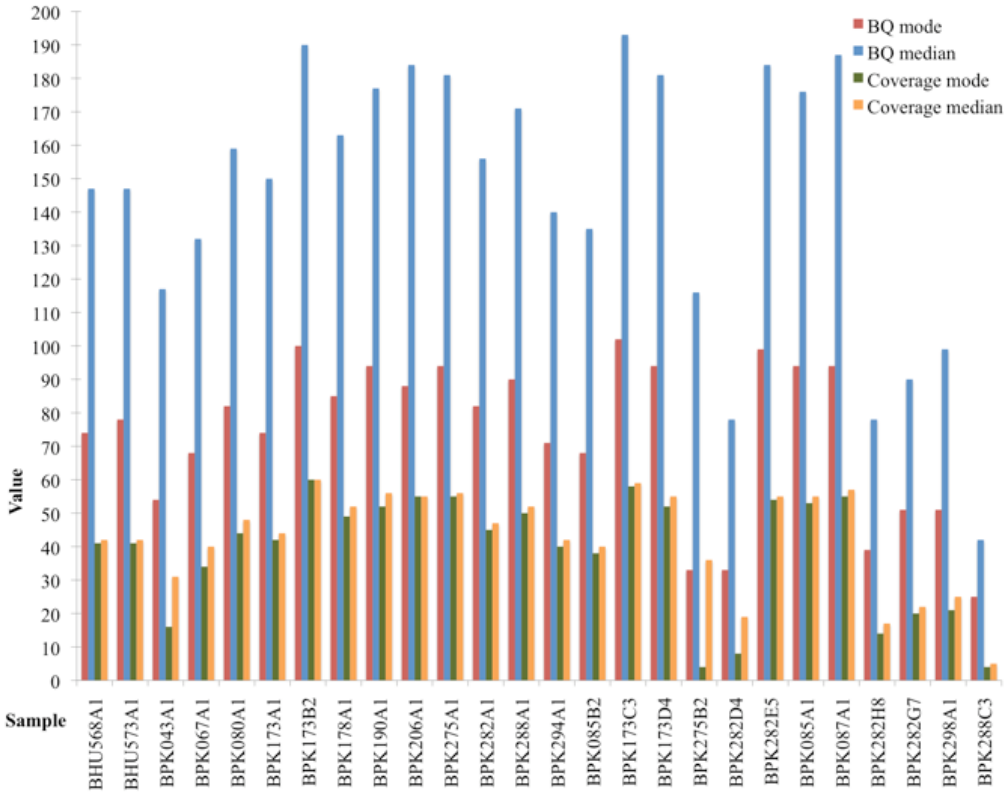
(A) After alkaline lysis, a circular DNA amplicon of 18 kb was isolated from line BPK275/0cl18. This DNA was amplified with primers epi-ASP1 and epi-SP2, situated at the edges of the MAPK locus and oriented to the regions outside of this locus: (B) this generated a 1.7 kb fragment.

Supplementary Figure 16. A histogram of GC content of read for all lines.



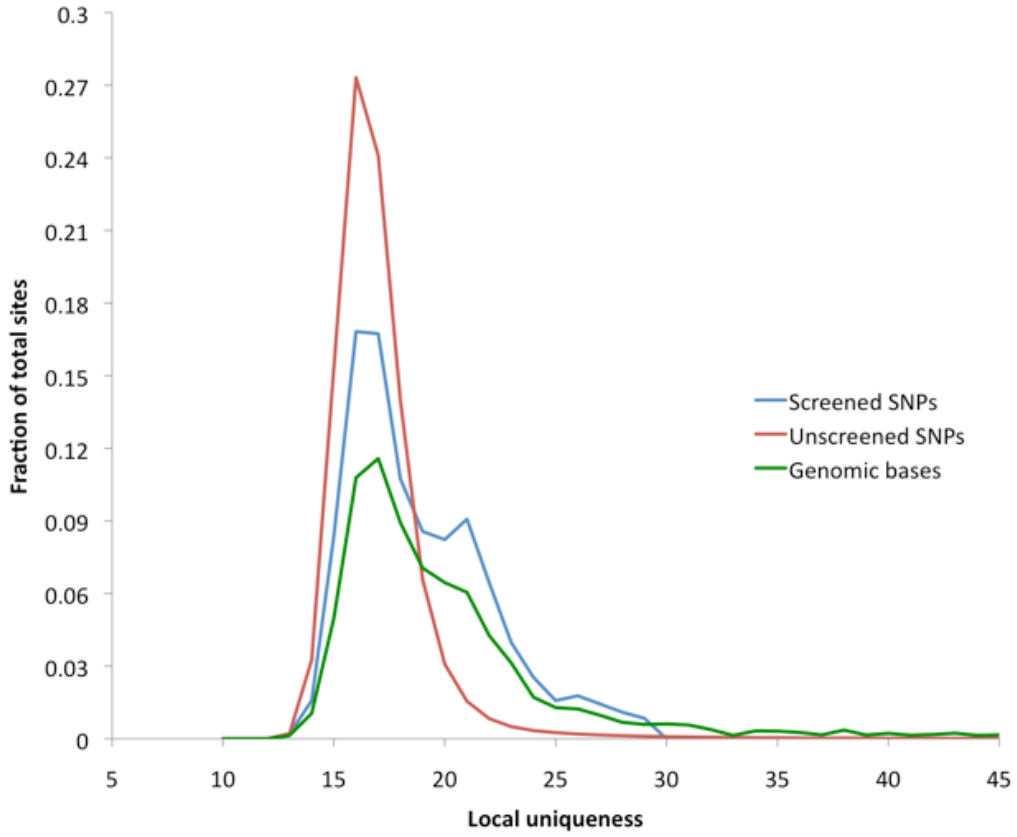
GC content for each line was calculated for all sequence output for each read individually, excluding those with unknown bases; this is shown as a frequency histogram on the y-axis. This indicated runs that were contaminated or of low quality where the median value was much lower than 62 (BPK288C3, BPK026A1 and BPK282B2: shown with a red marker).

Supplementary Figure 17. The base quality (BQ) and coverage mode and medians for chromosome 1: these metrics distinguished informative and non-informative Illumina libraries.



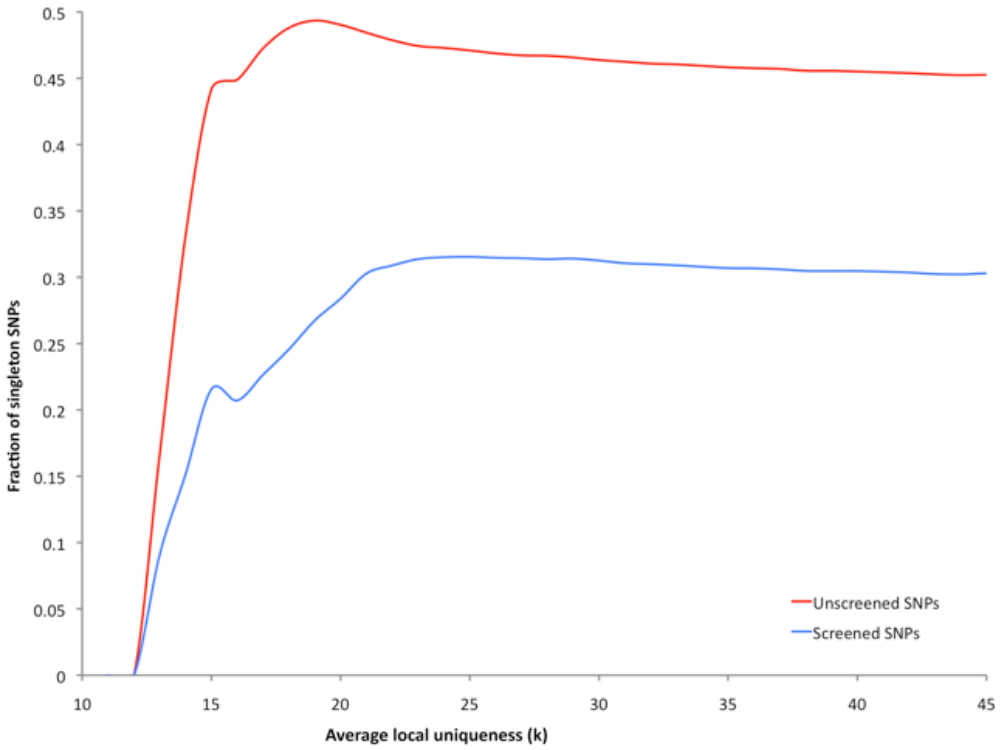
A number of cloned line not included in the main study were examined for data interpretation: BPK173B2, BPK173C3 and BPK173D4 were from the same patient as BPK173/0cl3. BPK275B2 was from the same patient as BPK275/0cl18. BPK282C3, BPK282D4, BPK282E5, BPK282G7 and BPK282H8 were from the same isolateclas BPK282/0cl4. BPK085A1 was from the same isolate as BPK085/0. BPK275B2, BPK282D4 and BPK282H8 represented libraries below sufficent quality thresholds.

Supplementary Figure 18. The fraction of screened SNP (blue), genomic (red) and unscreened SNP (green) sites (y-axis) distributed according to their minimum average local uniqueness value (*k*, x-axis).

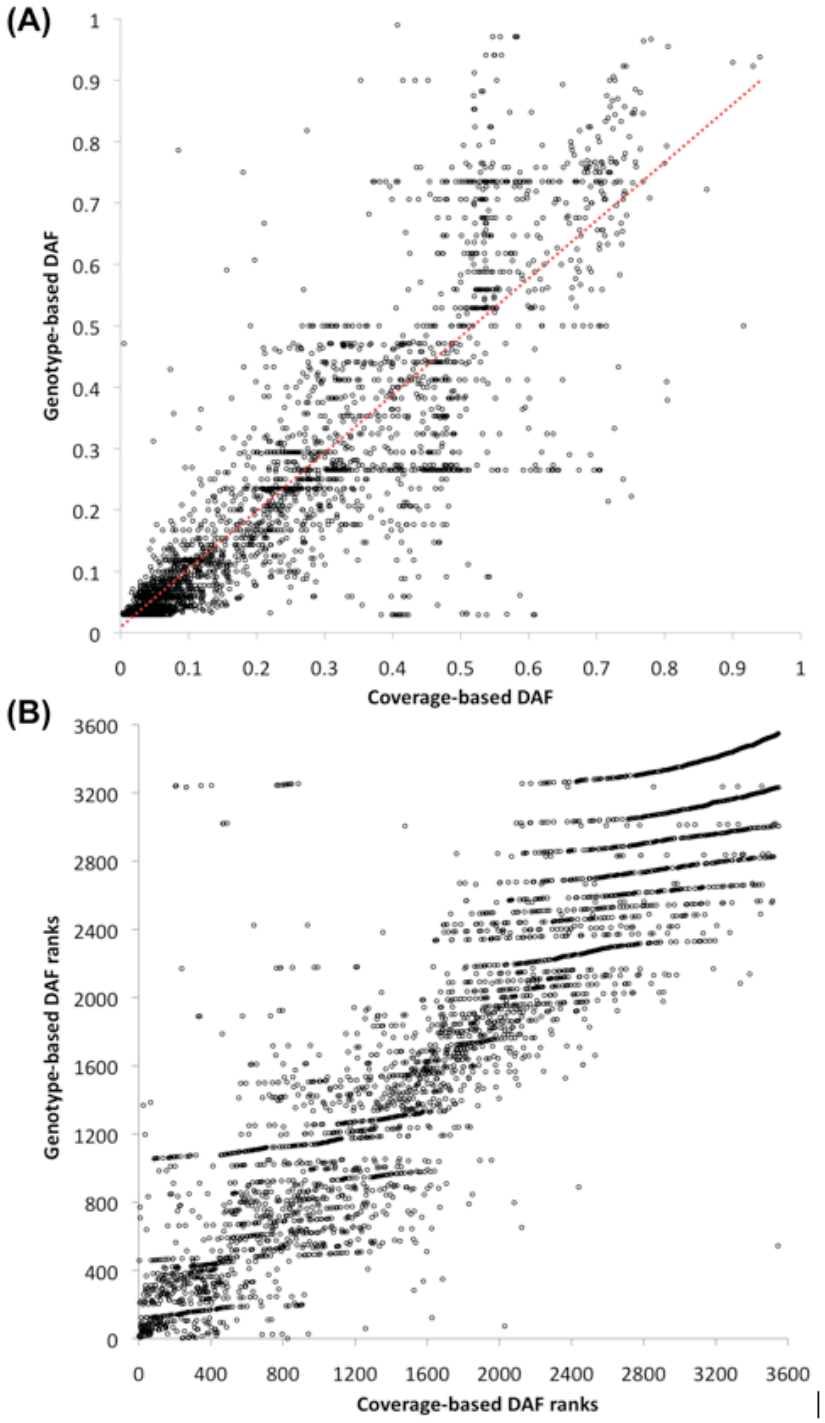


The mode of *k* for each was 16, except for genomic bases (17). By omitting all SNPs with $k > 30$, the mean local uniqueness reflects that of the 59 bases surrounding a given SNP. For $k > 290$, values were determined in 25 bp steps since the read library insert segments were approximately 300 bp long; *k* could not be accurately ascertained using the approach by DNA libraries whose median insert size < *k*.

Supplementary Figure 19. The fraction of singleton SNPs for screened (blue) and unscreened (red) datasets relative to their minimum average local uniqueness (*k*).

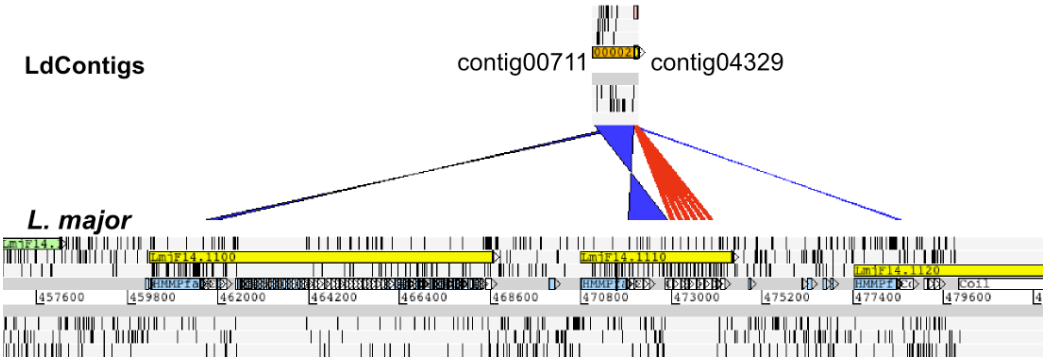


Supplementary Figure 20. The correlation of SNP derived allele frequencies (AF) determined by examining genotypes individually before calculating derived AF across the population (Genotype-based) versus pooling the population-wide read-depths as a set to obtain DAF values (Coverage-based).



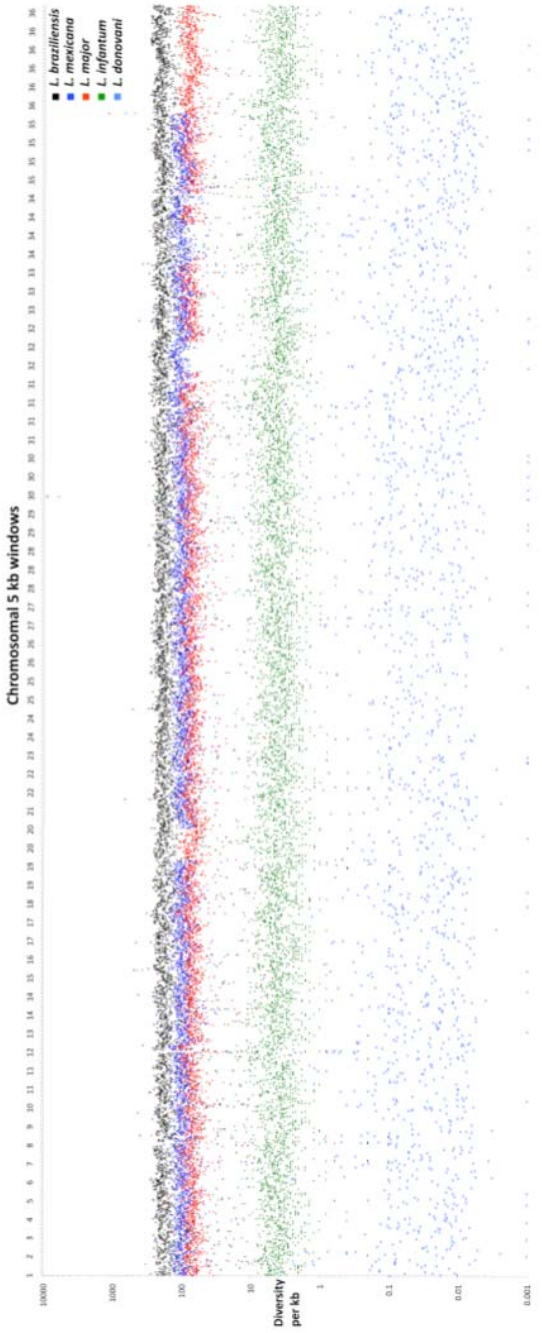
Each of the 3,549 SNPs is represented by a black dot. (A) Correlation was assessed for Pearson's linear coefficient (red dotted line, $r^2 = 0.77$, $p < 10^{-6}$) and (B) Spearman's rank-based approach ($r_S = 0.92$, $t = 136.5$, $p < 10^{-6}$).

Supplementary Figure 21. Alignment of two *L. donovani* contigs not homologous to anywhere in *L. infantum* genome but highly similar to the *L. major* genomic region (including a putative kinesin K39 gene, LmjF14.1110).



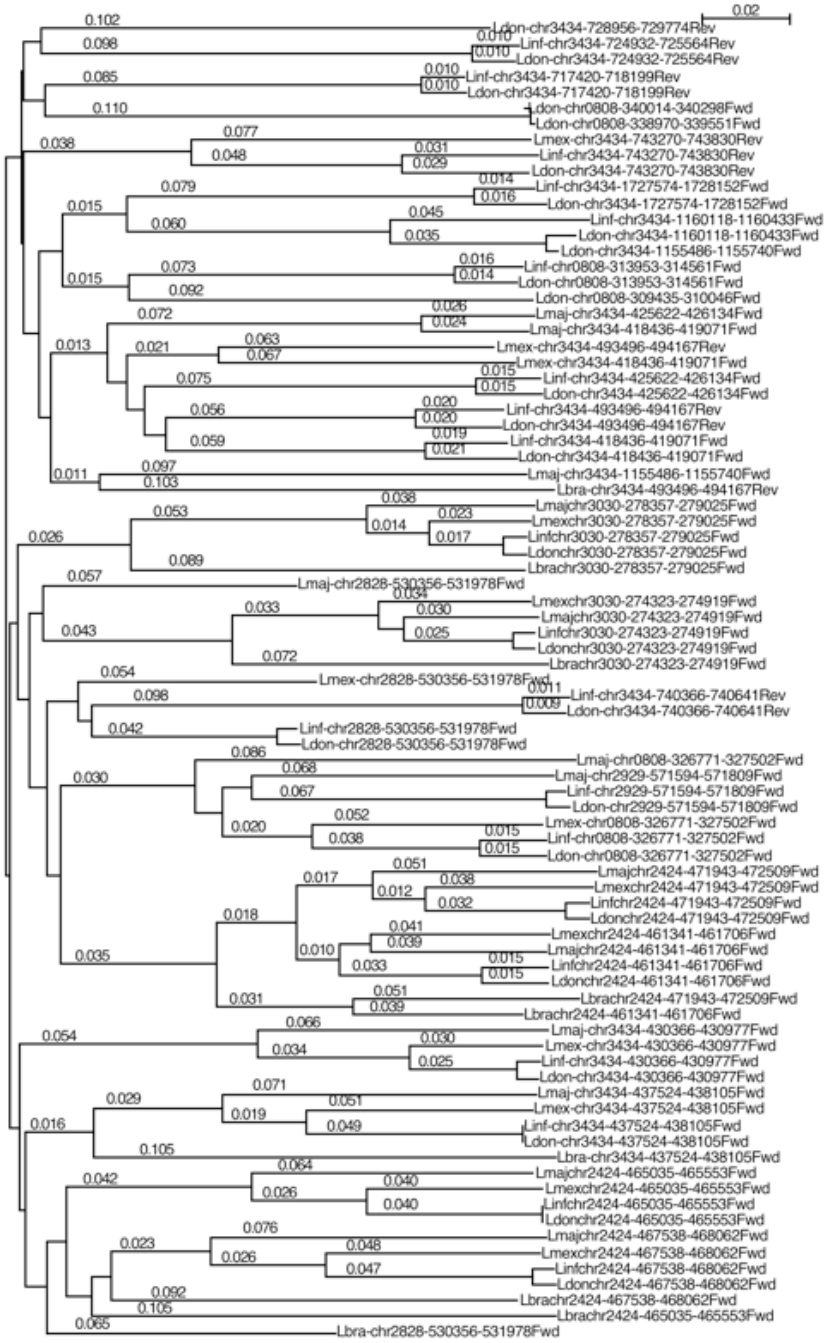
The alignment was performed using ACT (Carver et al. 2005) for *L. donovani* contigs 00711 (971 bp) and 04329 (116 bp) against *L. major*. Regions of high homology according to Blastn hits are shown in red and yellow. These contigs had no homologous region in *L. infantum*; however, this was no evidence of their absence from that genome.

Supplementary Figure 22. Genomic sliding window of variation between pairwise alignment of regions orthologous in the *L. donovani* reference and other sequenced *Leishmania* genomes.



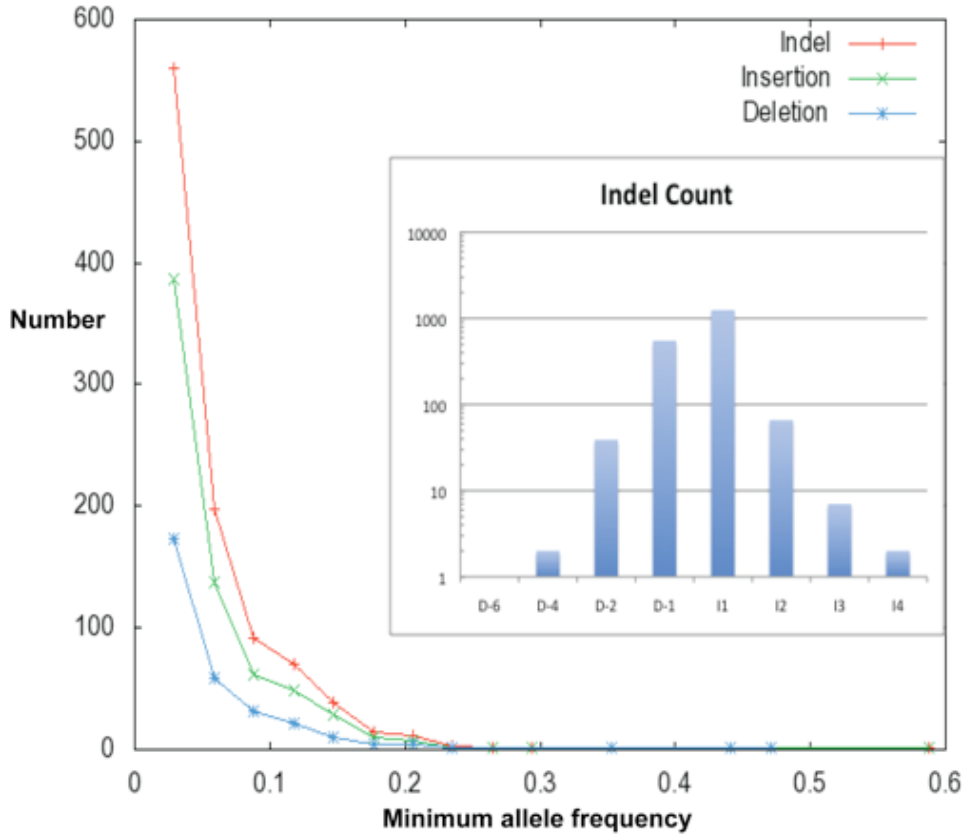
The windows were non-overlapping 5 kb shown on a log-scale for *L. braziliensis* (black), *L. mexicana* (dark blue), *L. major* (red), *L. infantum* (green) and *L. donovani* (light blue). *L. donovani* diversity is represented as nucleotide diversity per kb. SNP concentrations in the top 5% were those with 9.6+ per kb for *L. infantum*, 104.8+ for *L. major*, 139.0+ for *L. mexicana*, and 248.6+ for *L. braziliensis*.

Supplementary Figure 23. A genome-wide neighbour-joining phylogenetic tree of amastin genes in *L. donovani*.



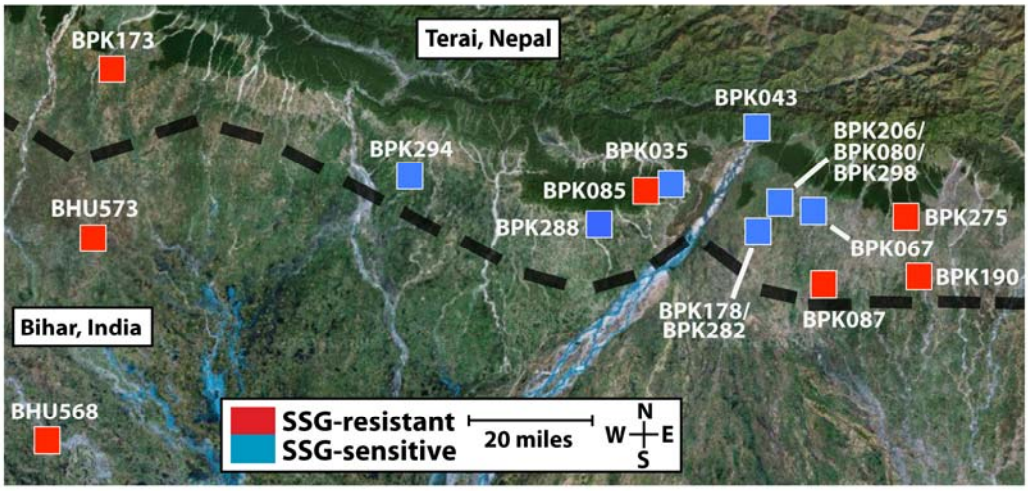
The tree was constructed from a T-Coffee alignment using NJplot of all known and resolved amastin-related genes. The branch lengths and values were proportional to the number of mutations. Genes are shown for each species (Ldon, *L. donovani*; Linf, *L. infantum*; Lmaj, *L. major*; Lmex, *L. mexicana*; Lbra, *L. braziliensis*) for the chromosome (chr), the gene start and end genomic bases, and the gene orientation (Fwd, forward; Rev, reverse). Distinct gene conversion event were apparent where the species tree was altered by the presence of high homology between paralogues, for example between LdBPK_342650 and LdBPK_342660.

Supplementary Figure 24. Insertion and deletion (indel) population genomic allele frequency spectra.



The frequency decay of the total numbers of indels (red), insertions (green) and deletions (blue) followed the expected exponential pattern. Indels were determined by comparison with the *L. donovani* reference genome. The inset shows the counts and types of indels observed on a log-scale: D representing deletions of a denoted length and I showing insertions. The genomic pattern of an excess of insertions over deletions was expected due to the significantly more deleterious consequences of a deletion compared to an insertion. The sharp decline of this distribution for was similar to others obtained for human CNVs (International HapMap 3 Consortium 2010).

Supplementary Figure 25. Geographic distribution of *L. donovani* samples from Nepal and India.



The strains sampled from Nepal (15) and the Indian state of Bihar (2) – the black line represents the border between the countries. SSG-resistant (red) and -sensitive (blue) samples are shown. Although the geographic range was small, there were considerable differences between phenotypes during *in vitro* SSG treatment.

Supplementary references

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*. **215(3)**:403-10.

Anisimova M, Bielawski JP, Yang Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol*. **18(8)**:1585-92.

Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. 2009 ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* **25(15)**:1968-9.

Axelsson E, Ellegren H. 2009. Quantification of adaptive evolution of genes expressed in avian brain and the population size effect on the efficacy of selection. *Mol Biol Evol*. **26**:1073-9.

Bendtsen JD, Nielsen H, von Heijne G, Brunak S. 2004. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol*. **340**:783-795.

Bentley SD, Corton C, Brown SE, Barron A, Clark L, Doggett J, Harris B, Ormond D, Quail MA, May G, et al. 2008. Genome of the actinomycete plant pathogen *Clavibacter michiganensis* subsp. sepedonicus suggests recent niche adaptation. *J Bacteriol*. **190(6)**:2150-60.

Brochu C, Haimeur A, Ouellette M. 2004. The heat shock protein HSP70 and heat shock cognate protein HSC70 contribute to antimony tolerance in the protozoan parasite *Leishmania*. *Cell Stress Chaperones*. **93**:294-303.

Burns JM Jr., Parsons M, Rosman DE, Reed SG. 1993. Molecular cloning and characterization of a 42-kDa protein phosphatase of *Leishmania chagasi*. *J Biol Chem.* **263**:17155-17161.

Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J. 2005. ACT: the Artemis Comparison Tool. *Bioinformatics*. **21(16)**:3422-3.

Carver T, Böhme U, Otto TD, Parkhill J, Berriman M. 2010 BamView: viewing mapped read alignment data in the context of the reference sequence. *Bioinformatics* **26(5)**:676-7.

Choudhury K, Zander D, Kube M, Reinhardt R, Clos J. 2008. Identification of a *Leishmania infantum* gene mediating resistance to miltefosine and SbIII. *Int J Parasitol*. **3812**:1411-23.

Coelho AC, Beverley SM, Cotrim PC. 2003. Functional genetic identification of PRP1, an ABC transporter superfamily member conferring pentamidine resistance in *Leishmania major*. *Mol Biochem Parasitol*. **130(2)**:83-90.

de Paiva Cavalcanti M, Felinto de Brito ME, de Souza WV, de Miranda Gomes Y, Abath FG. 2009. The development of a real-time PCR assay for the quantification of *Leishmania infantum* DNA in canine blood. *Vet J.* **182(2)**:356-8.

Fu YX. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147**:915-25.

Gojobori J, Tang H, Akey JM, Wu CI. 2007. Adaptive evolution in humans revealed by the negative correlation between the polymorphism and fixation phases of evolution. *Proc Natl Acad Sci U S A*. **104(10)**:3907-12.

Hanada M, Kobayashi T, Ohnishi M, Ikeda S, Wang H, Katsura K, Yanagawa Y, Hiraga A, Kanamaru R, Tamura S. 1998. Selective suppression of stress-activated protein kinase pathway by protein phosphatase 2C in mammalian cells. *FEBS Lett.* **437(3)**:172-6.

Hide M, Bañuls AL. 2008. Polymorphisms of cpb multicopy genes in the *Leishmania* (*Leishmania*) *donovani* complex. Trans R Soc Trop Med Hyg. **102(2)**:105-6.

Ibrahim ME, Barker DC. 2001. The origin and evolution of the *Leishmania donovani* complex as inferred from a mitochondrial cytochrome oxidase II gene sequence. *Infect*

Genet Evol. **1(1)**:61-8.
International HapMap 3 Consortium, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, et al. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature.* **467(7311)**:52-8.
Jackson AP. 2010. The evolution of amastin surface glycoproteins in trypanosomatid parasites. *Mol Biol Evol*. **27(1)**:33-45.
Kaur J, Kumar P, Tyagi S, Pathak R, Batra S, Singh P, Singh N. 2010. An *in silico* screening, structure-activity relationship and biologic evaluation of selective pteridine reductase inhibitors targeting visceral leishmaniasis. *Antimicrob Agents Chemother*. Nov 29th. Epub ahead of print.
Kircher M, Kelso J. 2010. High-throughput DNA sequencing--concepts and limitations. *Bioessays* **32(6)**:524-36.
Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods*. **6(4)**:291-5.
Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. 2001. Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes. *J. Mol. Biol*. **305**:567-580.
Leifso K, Cohen-Freue G, Dogra N, Murray A, McMaster WR. 2007. Genomic and proteomic expression analysis of *Leishmania* promastigote and amastigote life stages: the *Leishmania* genome is constitutively expressed. *Mol Biochem Parasitol.* **152(1)**:35-46.
Leprohon P, Légaré D, Girard I, Papadopoulou B, Ouellette M. 2006. Modulation of *Leishmania* ABC protein gene expression through life stages and among drug-resistant parasites. *Eukaryot Cell*. **5(10)**:1713-25.
Leprohon P, Légaré 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*. **375**:1387-99.
Luyo-Acero GE, Uezato H, Oshiro M, Takei K, Kariya K, Katakura K, Gomez-Landires E, Hashiguchi Y, Nonaka S. 2004. Sequence variation of the cytochrome b gene of various human infecting members of the genus Leishmania and their phylogeny. *Parasitology.* **128(Pt 5)**:483-91.
Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature.* **437(7057)**:376-80.
Maeda T, Wurgler-Murphy SM, Saito H. 1994. A two-component system that regulates an osmosensing MAP kinase cascade in yeast. *Nature* **369(6477)**:242-5.
Martínez-Calvillo S, Vizuet-de-Rueda JC, Florencio-Martínez LE, Manning-Cela RG, Figueroa-Angulo EE. 2010. Gene Expression in Trypanosomatid Parasites. *J Biomed Biotechnol*. **2010**:525241.
Mauricio IL, Gaunt MW, Stothard JR, Miles MA. 2007. Glycoprotein 63 (gp63) genes show gene conversion and reveal the evolution of Old World *Leishmania. Int J Parasitol*. **37(5)**:565-76.
Momen H, Cupolillo E. 2000. Speculations on the origin and evolution of the genus *Leishmania. Mem Inst Oswaldo Cruz*. **95(4)**:583-8.
Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**:929-936.
Ning Z, Cox AJ, Mullikin JC. 2001. SSAHA: a fast search method for large DNA

Databases. *Genome Res*. **11(10)**:1725-9.
Otto TD, Sanders M, Berriman M, Newbold C. 2010. Iterative Correction of Reference Nucleotides iCORN using second generation sequencing technology. *Bioinformatics.* **26(14)**:1704-7.
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(7)**:839-47.
Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ. 2008. A large genome center's improvements to the Illumina sequencing system. *Nat Methods*. **5(12)**:1005-10.
Quail MA, Swerdlow H, Turner DJ. 2009. Improved protocols for the illumina genome analyzer sequencing system. *Curr Protoc Hum Genet*. **18**:18.2.
Ramachandran P, Varoquaux G. 2011. Mayavi: a package for 3D visualization of scientific data. Computing in Science and Engineering (in press).
Rodriguez PL. 1998. Protein phosphatase 2C PP2C function in higher plants. *Plant Mol Biol* **386**:919-927.
Rijal S, Uranw S, Chappuis F, Picado A, Khanal B, Paudel IS, Andersen EW, Meheus F, Ostyn B, Das ML, et al. 2010. Epidemiology of *Leishmania donovani* infection in high-transmission foci in Nepal. *Trop Med Int Health.* **15 Suppl 2**:21-8.
Rosenzweig D, Smith D, Opperdoes F, Stern S, Olafson RW, Zilberstein D. 2008. Retooling *Leishmania* metabolism: from sand fly gut to human macrophage. *FASEB J.* **222**:590-602.
Smith NG, Eyre-Walker A. 2002. Adaptive protein evolution in Drosophila. *Nature* **415(6875)**:1022-4.
Swanson WJ, Wong A, Wolfner MF, Aquadro CF. 2004. Evolutionary expressed sequence tag analysis of *Drosophila* female reproductive tracts identifies genes subjected to positive selection. *Genetics* **168(3)**:1457-65.
Takekawa M, Maeda T, Saito H. 1998. Protein phosphatase 2Calpha inhibits the human stress-responsive p38 and JNK MAPK pathways. *Embo J* **1716**:4744-4752.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123(3)**:585-95.
Tsai IJ, Otto TD, Berriman M. 2010. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol.* **11(4)**:R41.
Vergnes B, Gourbal B, Girard I, Sundar S, Drummelsmith J, Ouellette M. 2007. A proteomics screen implicates HSP83 and a small kinetoplastid calpain-related protein in drug resistance in *Leishmania donovani* clinical field isolates by modulating drug-induced programmed cell death. *Mol Cell Proteomics*. **61**:88-101.
Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature.* **452(7189)**:872-6.
Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol*. **22(4)**:1107-18.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in BioSciences* **13**:555-556.
Yang Z. 2007. PAML 4: a program package for phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**:1586-1591
Zeng K, Fu YX, Shi S, Wu CI. 2006. Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* **174(3)**:1431-9.