Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance

Tim Downing, Hideo Imamura, Saskia Decuypere, et al.

Genome Res. 2011 21: 2143-2156 originally published online October 28, 2011
Access the most recent version at doi:10.1101/gr.123430.111

Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance

Tim Downing,[1,10] Hideo Imamura,[2,10] Saskia Decuypere,[2] Taane G. Clark,[3] Graham H. Coombs,[4] James A. Cotton,[1] James D. Hilley,[5] Simonne de Doncker,[2] Ilse Maes,[2] Jeremy C. Mottram,[5] Mike A. Quail,[1] Suman Rijal,[6] Mandy Sanders,[1] Gabriele Schönian,[7] Olivia Stark,[7] Shyam Sundar,[8] Manu Vanaerschot,[2] Christiane Hertz-Fowler,[1,9] Jean-Claude Dujardin,[2,11] and Matthew Berriman[1,11]

[1]*Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SA, United Kingdom;* [2]*Unit of Molecular Parasitology, Department of Parasitology, Institute of Tropical Medicine, 2000 Antwerp, Belgium;* [3]*London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, United Kingdom;* [4]*Strathclyde Institute of Pharmacy and Biomedical and Sciences, University of Strathclyde, Glasgow G4 0RE, United Kingdom;* [5]*Wellcome Trust Centre for Molecular Parasitology, Institute of Infection, Immunity and Inflammation, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8TA, Scotland, United Kingdom;* [6]*B.P. Koirala Institute of Health Sciences, Ghopa, Dharan, Nepal;* [7]*Institut für Mikrobiologie und Hygiene, Charité Universitätsmedizin Berlin, 10117 Berlin, Germany;* [8]*Institute of Medical Sciences, Banaras Hindu University, Varanasi, India*

Visceral leishmaniasis is a potentially fatal disease endemic to large parts of Asia and Africa, primarily caused by the protozoan parasite *Leishmania donovani*. Here, we report a high-quality reference genome sequence for a strain of *L. donovani* from Nepal, and use this sequence to study variation in a set of 16 related clinical lines, isolated from visceral leishmaniasis patients from the same region, which also differ in their response to in vitro drug susceptibility. We show that whole-genome sequence data reveals genetic structure within these lines not shown by multilocus typing, and suggests that drug resistance has emerged multiple times in this closely related set of lines. Sequence comparisons with other *Leishmania* species and analysis of single-nucleotide diversity within our sample showed evidence of selection acting in a range of surface- and transport-related genes, including genes associated with drug resistance. Against a background of relative genetic homogeneity, we found extensive variation in chromosome copy number between our lines. Other forms of structural variation were significantly associated with drug resistance, notably including gene dosage and the copy number of an experimentally verified circular episome present in all lines and described here for the first time. This study provides a basis for more powerful molecular profiling of visceral leishmaniasis, providing additional power to track the drug resistance and epidemiology of an important human pathogen.

[Supplemental material is available for this article.]

Leishmaniases are a complex of diseases that range from self-curing lesions to gross disfigurations and potentially deadly visceral disease. The diseases are caused by protozoan parasites that are transmitted by sandflies in 88 countries and infect an estimated 12 million people (www.who.int/leishmaniasis/en/). Parasites of the *Leishmania* genus are remarkably biologically, clinically, and epidemiologically diverse and present enormous differences in disease tropism. The mildest form is cutaneous leishmaniasis, which is caused by *Leishmania major* and other species, and is largely limited to lesions around the area of a sandfly bite—though a diffuse form can also occur. Disfiguring mucocutaneous leishmaniasis is due to the destruction of nasopharyngeal tissue by parasites such as *L. braziliensis*. More significantly, visceral leishmaniasis is caused by parasites of the *L. donovani* species complex that can spread to internal organs and cause death.

In 2005, sequencing the genome of *L. major* identified 8311 protein-coding genes and provided a framework for future comparative genomic studies (Ivens et al. 2005). The genome elucidated the full structural architecture of *Leishmania* chromosomes, which includes an unusual pattern of genes distributed in large directional clusters. Subsequently, the genomes of *L. braziliensis* and *L. infantum* were described—the latter is a member of the *L. donovani* complex (Peacock et al. 2007). A detailed comparison of these first three *Leishmania* genomes revealed a striking background of conservation at the gene-content level with almost complete synteny. Moreover, in stark contrast to the major phenotypic differences caused by each species, less than 50 genes were

[9]Present address: Centre for Genomic Research, Institute of Integrative Biology, Biosciences Building, University of Liverpool, Crown Street, Liverpool L69 7ZB, U.K.
[10]These authors contributed equally to this work.
[11]Corresponding authors.
E-mail jcdujardin@itg.be.
E-mail matthew.berriman@sanger.ac.uk.
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.123430.111. Freely available online through the *Genome Research* Open Access option.

differentially distributed, indicating that few genuinely species-specific genes may exist. This study, however, only honed in on the presence or absence of orthologs. It did not include other classes of diversity, such as structural polymorphisms within tandem arrays or the generation of episomes; both of which have subsequently been described as differing considerably between *Leishmania* lines and as contributing substantial differences to gene expression (Dujardin 2009; Leprohon et al. 2009a,b). Furthermore, the genome of *L. donovani* has thus far not been examined, despite it being arguably the most important *Leishmania* species in terms of public health, and being genetically distinct from *L. infantum* (Lukes et al. 2007). Both in terms of mortality and morbidity, *L. donovani* causes a major part of the leishmaniasis burden, in East Africa and the Indian subcontinent (Rijal et al. 2010).

In the present study, we aimed to explore the intraspecies genomic diversity of *Leishmania* in a clinical context. The first regional program of visceral leishmaniasis elimination is targeting *L. donovani* in Bangladesh, India, and Nepal (Mondal et al. 2009). However, emerging drug resistance is compromising current efforts. Understanding the *L. donovani* genome and its natural variation and genetic population structure in these regions is therefore essential to underpin and enhance public health surveillance and intervention strategies.

We focused specifically on *L. donovani* populations from the Terai region of Nepal and the state of Bihar in India. These populations constitute an ideal genomic model in which genetic variation between samples should be reduced: Indeed, these parasites are relatively homogeneous, likely as a result of a demographic bottleneck in the 1960s, after DDT spraying campaigns (WHO Expert Committee on the Control of the Leishmaniases 2010). In addition, these isolates were the subjects of several recent studies, providing a unique clinical, biological, and epidemiological background (Decuypere et al. 2005, 2008; Laurent et al. 2007; Alam et al. 2009; Rijal et al. 2010; Vanaerschot et al. 2010). Importantly, these parasites are highly variable in terms of susceptibility to antimonial drugs such as sodium stibogluconate (SSG; Rijal et al. 2007; Samant et al. 2007; Kumar et al. 2009).

Using a combination of 454 Life Sciences (Roche) and Illumina sequencing technologies, we have generated a high-quality and annotated draft genome for *L. donovani*. We have subsequently used this reference genome as a framework to map natural variation data collected from deep resequencing of 16 further Nepalese and Indian clinical lines. By discovering genome-wide single-nucleotide polymorphisms (SNPs) and structural variation (SV), we identified genes with differential patterns of diversity in antimonial resistant and susceptible samples. Both tests for adaptive evolution across SNPs and assessments of copy-number variation (CNV) highlighted subsets of protein-coding genes undergoing adaptive evolution in the *L. donovani* population. As expected, we saw a low level of SNP variation; but despite this, we report striking structural polymorphisms that may result in locus-specific changes to gene dosage. These include extensive chromosomal CNVs as well as the generation of extrachromosomal circular fragments.

Results

Generation of an annotated *L. donovani* reference genome

To investigate natural genetic variation in *L. donovani*, we selected a single cloned line (BPK282/0cl4) from Nepal to use as a reference and sequenced its genome by combining single and paired-end reads from the 454 GS FLX Titanium platform (median 22-fold coverage) with reads from Illumina Genome Analyzer (median 52-fold coverage). The 454 data were assembled de novo, and the Illumina data were used to iteratively correct errors in the consensus sequence and close gaps (Supplemental Fig. S1). The initial set of contigs—50% of which were 44.4 kb or greater (N50)—were aligned, ordered, and oriented against the *L. infantum* JPCM5 reference genome (Peacock et al. 2007) to produce an assembled high-quality draft *L. donovani* genome of length 32.4 Mb containing 2154 contigs and with a final N50 of 45.5 kb.

The reference genome was annotated by identifying orthologous genes within the *L. infantum* genome and transferring their annotation onto the *L. donovani* draft assembly. Of the 8395 genes in the *L. infantum* genome (Peacock et al. 2007), 8252 were transferred to *L. donovani* using RATT (Otto et al. 2011). These included 33 genes assigned to 15 *L. donovani* contigs that could not be uniquely associated with a chromosome; most of these genes occurred in arrays with homology with multiple regions in the *L. donovani* genome (Supplemental Tables S1, S2). A total of 143 genes were not transferred from *L. infantum* because they occurred within repetitive gene arrays, had multiple paralogs in *L. infantum*, matched multiple regions, or were located in regions for which local synteny was unresolved in *L. donovani*.

The *L. donovani* reference sequence was aligned to each of the other previously described *Leishmania* genomes (*L. infantum*, *L. major*, and *L. braziliensis*) (Peacock et al. 2007) and the recently sequenced *L. mexicana* genome (Rogers et al. 2011). This highlighted substantial genetic differentiation at the species level (Fig. 1A), with 156,274 nucleotide changes between the *L. infantum* and *L. donovani* reference genomes: two of these changes corrected in-frame stop codons in *L. infantum*-specific pseudogenes with high identity to their *L. donovani* orthologs (LdBPK_321870 and LdBPK_332420).

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

We generated an average of 66-fold sequence coverage for each of 17 *L. donovani* lines derived from unique visceral leishmaniasis patients, among whom there were different responses to antimonial therapy (Supplemental Table S3). By mapping the sequencing reads from these lines to the BPK282/0cl4 reference *L. donovani* genome sequence, we called SNPs using a conservative approach exploiting the read-depth coverage values across all lines to minimize systematic errors (Jiang et al. 2009; Lynch 2009). This approach identified SNPs at 3549 sites, of which 2933 were in noncoding regions, 220 were synonymous mutations in coding regions, and 396 caused changes at the protein sequence level. As expected, there was evidence of extensive purifying selection at genes: 17.4% of the SNP variation surveyed was in coding regions, which account for 47.0% of the genome. Coding-sequence mutations occurred in just 368 genes—263 had at least one protein-level polymorphism and 158 one or more silent changes. This was confirmed by neutral allele frequency distributions for coding and noncoding sites (Supplemental Fig. S2) and a higher rate of synonymous variants segregating at intermediate allele frequencies relative to nonsynonymous ones (Supplemental Fig. S3).

Despite a low level of genetic differentiation between the lines, principal component analysis (PCA) of genome-wide nucleotide variation (Fig. 1B), and a protein-level phylogenetic network (Fig. 1C; Supplemental Fig. S4) both revealed significant genetic differences in our population not detected by microsatellite profiling (Alam et al. 2009; Bhattarai et al. 2010). These

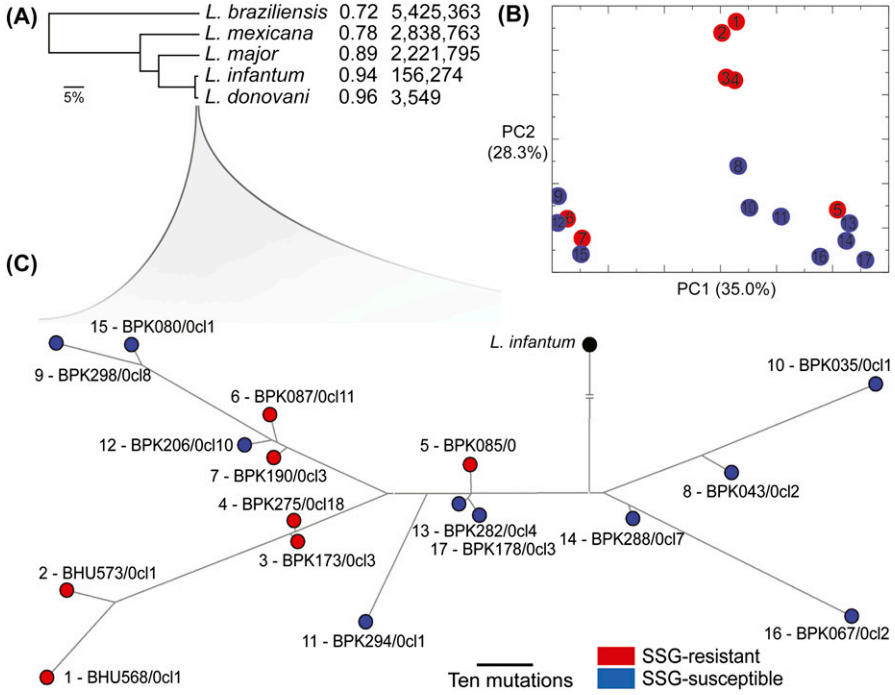

Figure 1. The phylogenomic context of functional variation in clinical *L. donovani* lines. (*A*) A neighbor-joining phylogenomic tree of *Leishmania* species. The relative similarity is shown (1% was equivalent to 108,000 mutations). The first column indicates the fraction of the known genomes orthologously aligned to *L. donovani* and the second is the number of SNPs identified (Table 2). (*B*) Principle component analysis (PCA) of genomic SNP variation in the 17 *L. donovani* lines resistant (red) and susceptible (blue) to SGG. The two most significant PCs distinguished three main groups, accounting for 63.3% of the total variation. The PCA *L. donovani* line numbers correspond to those shown in *C*, a median-joining phylogenetic network of genome-wide nonsynonymous sites variation (396 SNPs) for samples resistant (red) and susceptible (blue) to SSG from India (beginning with BHU) and Nepal (BPK). The branch lengths displayed are proportional to the number of differences between lines. The position of the ancestral node for *L. infantum* (black) was shortened: For comparison, the *L. donovani*:*L. infantum* branch length was 138 times that between the most divergent *L. donovani* (1, BHU568/0cl1; and 10, BPK035/0cl1).

distinct relationships resolved by genomic SNP data were closely mirrored by phylogenetic signals inferred from large CNVs (Supplemental Fig. S5), and supported the possibility of multiple events of emergence of SSG resistance (Laurent et al. 2007). Although geographic separation has played a significant role in genetic differentiation between species (Lukes et al. 2007), no evidence for this was found in the 17 lines, presumably due to the small geographic area sampled and extensive local human migration (Rijal et al. 2010). Similarly, little of the variability in SSG phenotype could be explained directly by either genome-wide variation or geographic distance.

The sites most differentiated between the SSG-resistant and SSG-susceptible lines across the genome, and so potentially involved in the resistance phenotype, were identified using F-statistics (Wright 1951). We found a total of 41 SNPs with $F_{ST} > 0.4$, and these most discriminating SNPs showed an excess of nonsynonymous mutations (Fisher's Exact test $P < 0.02$) (Table 1). No SNP completely discriminated between resistant from sensitive lines, but 34 out of these 41 were present solely and homozygously in four related SSG-resistant lines (BHU568/0cl1, BHU573/0cl1, BPK173/0cl3, and BPK275/0cl18; Supplemental Table S4). These four lines likely share a recent common ancestor, so while our data cannot resolve whether these variants were due to selective events or population history, it is possible that some of these SNPs may be related to SSG resistance in this group of isolates.

Identifying recent selective events this parasite population could be hindered by complex population history and modes of reproduction. *L. donovani* propagation is largely clonal in regions of endemic leishmaniasis, leading to high levels of inbreeding (Gelanew et al. 2010; Rougeron et al. 2010). Together with its likely small effective population size (Laurent et al. 2007), this means purifying selection should be less efficient at removing disadvantageous alleles, obscuring signals of selective pressure (Charlesworth 2009). Thus, the pattern in our sample of low diversity (nucleotide diversity, $\pi = 68.6$ per Mb) (Tajima 1983) and an excess of rare alleles ($F_S = -1.72$) (Fu 1997) could be due to a pattern of clonal reproduction as well as purifying selection, which had a limited power to remove mildly deleterious variants. Comparing the proportions of nonsynonymous and synonymous mutations in our 17 lines and between the four other *Leishmania* species (Fig. 2) showed that a substantial proportion (0.55) of the population-level variation was neutral or nearly neutral, confirming that selection was relatively weak within this cohort, though stronger on high-frequency variants (Supplemental Table S5; Supplemental Fig. S6).

Detecting ancient and recent adaptive evolution in *L. donovani*

L. donovani has been a parasite of humans for a long period of time, and so host-specific selective processes are likely to have shaped its genome (Stevens et al. 2001; Simpson et al. 2006). Furthermore, genes are likely to show similar adaptive changes in related species (Obbard et al. 2010). Consequently, by examing the effects of long-term adaptation between species as well as recent selective events in our population, novel insights into the functional importance of genes in *Leishmania* can be revealed.

We used a series of complementary tests to examine variation between five *Leishmania* genomes in the *L. donovani* lineage and within our set of 17 *L. donovani* lines (Supplemental Fig. S7). Twelve genes showed evidence of sustained positive selection across the *Leishmania* genus, as inferred from a high ratio of nonsynonymous to synonymous subsitutions ($\omega = d_N/d_S > 1$) (Table 2; Supplemental Fig. S8; Yang 2007). These include 10 not previously identified in a similar test comparing *L. infantum*, *L. major*, and *L. braziliensis* (Supplemental Tables S6, S7; Peacock et al. 2007). A *L. donovani* lineage- and site-specific test of ω values highlighted 11 individual sites in nine genes that have historically been under directional selection in *L. donovani* (Supplemental Table S8; Yang 2007). These two analyses did not make use of the information from our multiple *L. donovani* lines, so by comparing the relative rates of coding sequence changes within the 17 lines to those between *L. donovani* and other *Leishmania* species, an additional 15 genes showing positive selection were identified with the direction of selection test (Supplemental Table S9; Stoletzki and Eyre-Walker

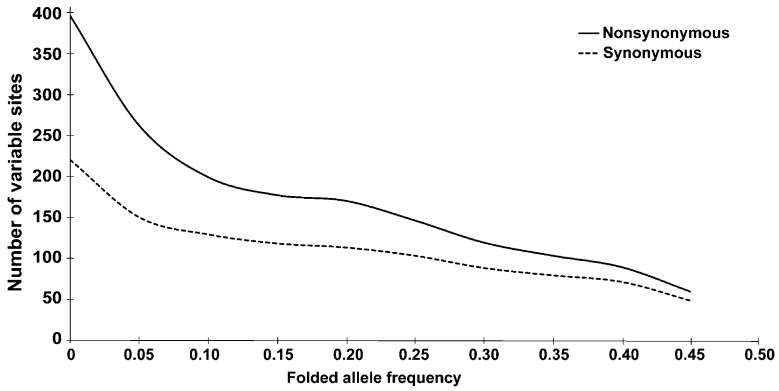

Figure 2. The neutrality of coding-site allele-frequency spectra. The allele-frequency spectrum for SNPs at nonsynonymous (black line) and synonymous (dashed line) sites among the 17 samples for the genome. Allele frequencies were computed using read-depth coverage values and were folded because the selective process may act on both the ancestral alleles inferred from *L. infantum* as well as the derived version with the *L. donovani* population. The ratio of nonsynonymous to synonymous SNPs approaches one as the allele frequencies increase, suggesting that higher frequency variants are subject to stronger purifying selection to remove deleterious alleles.

2011). To identify genes with evidence of positive selection that could not be aligned across all five genomes, pairwise comparisons between *L. donovani* and each of the four other *Leishmania* species identified 36 genes with consistently high D_N/D_S ratios (Supplemental Table S10). Finally, signatures of selection at genes were explored using a genome-wide sliding-window approach for standard measures of diversity within the 17 lines: π, θ_W, (Supplemental Fig. S9; Watterson 1975), and Tajima's D (Supplemental Table S11; Tajima 1989).

Many of the genes identified by these approaches encode hypothetical proteins, but a number of results stand out as being explainable in terms of known gene functions important to *Leishmania* biology. Molecular studies investigating antimonial resistance in *L. donovani* clinical isolates have implicated modifications to proteins topologically predisposed to faster adaptation (Kim et al. 2007; Sackton et al. 2007; Cui et al. 2009) such as surface proteins (Samant et al. 2007) and transporters (Mandal et al. 2010) in the SSG-resistance phenotype. Here, we provide evidence of ancient and recent adaptive evolution both between *Leishmania* species and in the *L. donovani* lineage at the highly immunogenic amastin surface glycoproteins (Wu et al. 2000; Stober et al. 2006; Jackson 2010). Diversity at these loci should be advantageous within our set of *L. donovani* lines, and, indeed, we detected an excess of derived high-frequency SNPs at one amastin locus, with half of the 12 protein-level SNPs in the 34D subfamily amastin gene (LdBPK_341150) segregating at elevated frequencies (Supplemental Table S12).

Evidence of adaptive evolution was also consistently observed in genes encoding surface proteins with roles in transport, including some that have roles in SSG resistance. A high level of diversity was found at an amino acid transporter locus (LdBPK_310350, $\pi = 0.9$ per kb) that has been found to be down-regulated in SSG-resistant *L. donovani* strains (Singh et al. 2010). Several genes previously associated with drug resistance showed signs of persistent positive selection in *Leishmania*, including

Table 1. Coding SNPs showing greatest differentiation between lines resistant and susceptible to SSG treatment

F_{ST}	C[a]	Position	S[b]	R[c]	Type[d]	Gene product	Site	Gene ID
0.54	34	1046281[e]	0.81	0.00	N	Conserved hypothetical protein	H523R[f]	LdBPK_342390
0.44	2	36464	1.00	0.43	N	Phosphatidylinositol 3-kinase	R3452H	LdBPK_020100
0.44	4	337729[g]	1.00	0.43	N	Rhomboid-like serine protease	A113V	LdBPK_040850
0.44	7	383016	1.00	0.43	Syn	Phosphoacetylglucosamine mutase	161V	LdBPK_070930
0.44	12	128965	1.00	0.43	N	Conserved hypothetical protein	Y1230C	LdBPK_120270
0.44	16	690776	1.00	0.43	N	Conserved hypothetical protein	Y1214C	LdBPK_161760
0.44	19	412141	1.00	0.43	Syn	4-coumarate CoA ligase	395V	LdBPK_190970
0.44	21	412303	1.00	0.43	Syn	Hypothetical protein	251P	LdBPK_211240
0.44	24	26882	1.00	0.43	N	Ankyrin/TPR repeat protein	A150V	LdBPK_240130
0.44	25	702972[g]	1.00	0.43	N	Conserved hypothetical protein	A61E	LdBPK_251890
0.44	31	553554[g]	1.00	0.43	Syn	Conserved hypothetical protein[h]	426T	LdBPK_311340
0.44	31	555080	1.00	0.43	N	Conserved hypothetical protein	S268P	LdBPK_311340
0.44	32	357021[g]	1.00	0.43	N	Conserved hypothetical protein	V125L	LdBPK_320990
0.44	35	1773762[g]	1.00	0.43	N	Conserved hypothetical protein	A54T	LdBPK_354470
0.44	36	886126[g]	1.00	0.43	Syn	Conserved hypothetical protein	347T	LdBPK_362330
0.42	33	1316522[e]	0.20	0.86	N	Conserved hypothetical protein	A390V	LdBPK_333140
0.40	34	493541[e]	0.78	0.14	Syn	Amastin-like surface protein	209A	LdBPK_341150

SNPs were ranked by F_{ST}: This set of 17 all had $F_{ST} > 0.4$ (bootstrapping $P < 0.02$) and were in coding sequence; noncoding SNPs with $F_{ST} > 0.4$ are in Supplemental Table S4.
[a]Chromosome.
[b]Frequency of ancestral allele in SSG-sensitive set.
[c]Frequency in SSG-resistant set.
[d]N stands for nonsynonmous and Syn for synonymous.
[e]Sites not among the 34 SNPs, where the derived allele was homozygous in SSG-resistant lines BHU568/0cl1, BHU573/0cl1, BPK173/0cl3, and BPK275/0cl18, and was not in any other lines.
[f]Predicted to be tolerated by protein impact software.
[g]SNPs confirmed by allele-specific genotyping.
[h]Contained two high-F_{ST} SNPs; locus had a positive direction of selection test value (Supplemental Table S9); predicted to have phosphatase 2C activity (InterProScan IPR001932).

Table 2. Substitutions in *Leishmania* genomes compared twith the *L. donovani* genome

Species		*L. infantum*	*L. major*	*L. mexicana*	*L. braziliensis*	*L. donovani*
Total substitutions		156,274	2,221,795	2,838,763	5,425,363	3549
Noncoding		108,956	1,492,716	1,887,762	3,056,798	2933
Coding		47,318	729,079	951,001	2,368,565	616
Nonsynonymous (D_N)		25,289	365,080	476,459	1,227,707	$p_N = 396$
Synonymous (D_S)		22,645	364,551	475,158	1,140,857	$p_S = 220$
Fraction of CDS SNPs		0.303	0.328	0.335	0.437	0.174
Divergence per kb		4.88	67.6	85.4	169.1	$\pi = 68.6$[a]
Genome size (bp)		32,055,853	32,855,403	33,254,784	32,091,314	32,444,998
Known orthologous sites[b]	bp	30,542,998	28,786,268	25,433,275	23,401,169	31,254,077
	%	94.1	88.7	78.4	72.1	96.3
D_N/D_S		1.11	1.00	1.00	1.07	$p_N/p_S = 1.80$
FI = $[D_N/D_S]/[p_N/p_S]$		0.62	0.56	0.56	0.60	-

[a]Nucleotide diversity per Mb.
[b]Compared with the *L. donovani* reference genome. D_N is the number of nonsynonymous changes between *L. donovani* and the given species, and D_S the synonymous mutations. p_N is the number of nonsynonymous variants segregating in the 17 lines, and p_S the synonymous ones. FI is the fixation index. Intraspecific measures of variability for the 17 *L. donovani* lines are shown for comparison.

ATP-binding cassette elements (Castanys-Muñoz et al. 2008; Leprohon et al. 2009b), histone genes (Singh et al. 2007), and the nucleoside transporter 1 gene (LdBPK_151230) (Vasudevan et al. 2001).

The A2 protein plays a role in promastigote–amastigote differentiation (Alcolea et al. 2010) and is expressed solely during amastigote-stage host infection in *L. donovani* (Ghedin et al. 1998; Zhang and Matlashewski 2001) but not *L. major*, and has been associated with the differing disease tropisms between the two species (Zhang et al. 2003). In contrast to the conserved A2 locus in *L. major* (Garin et al. 2005), functional CNVs have been previously observed in the *L. donovani* species complex, and we see significant sequence diversity and structural variation at this region even between our related lines. Whereas 3′ and 5′ A2-related genes possessed nine high-frequency protein-level SNPs, genes on the rest of that chromosome had only a total of four, suggesting a possible role for variability at this locus in A2 gene expression (Zhang et al. 2003).

Kinetoplastids use the thiol trypanothione and a redox intermediate, tryparedoxin, to defend against oxidative stress: genes involved in this pathway are implicated in resistance to antimonial drugs and in mediating host immune response. Tryparedoxin peroxidase diminishes the inhibition of thiol redox metabolism by antimonial drugs by reactive oxygen species reduction (Wyllie et al. 2004, 2008; König and Fairlamb 2007; Mandal et al. 2007), and is overexpressed in SSG-resistant clinical *L. donovani* (Wyllie et al. 2010). The tryparedoxin locus itself may be associated with both host immunogenicity and drug resistance in *Leishmania* (Stober et al. 2005), and in particular, with the murine B-cell response against *L. infantum* (Castro et al. 2004; Cabral et al. 2008). We discovered variation at three genes associated with redox metabolism: the tryparedoxin peroxidase gene (LdBPK_151140) showed signatures of selection on the *L. donovani* lineage (Supplemental Table S9) and also had a pattern of balanced variation in the 17 lines. One tryparedoxin gene (LdBPK_ 291220) had a protein-coding variant in a region of low diversity. In addition, a cAMP-specific phosphodiesterase locus (LdBPK_151540) displayed evidence of adaptive evolution (Supplemental Table S9): Reduced cAMP phosphodiesterase expression is associated with increased resistance to antioxidant stress, possibly mediated by components of thiol metabolism pathways in *L. donovani* (Bhattacharya et al. 2009).

Whole-chromosome copy-number variation

We used alignments of Illumina reads against the reference *L. donovani* sequence to assess chromosomal copy number. Using normalized read-depth (*s*, see Methods), 70.9% of chromosomes had a relative coverage equivalent to one, but a significant proportion of chromosomes deviated from this value by discrete multiples, namely, 1.5- (24.3%), 2- (4.6%), and 2.5-fold (0.2%) (Fig. 3). As validation, we quantified four single-copy genes from chromosomes 17, 34, 36 (mean $s = 1$ in all lines), and 31 (mean $s = 2$) by qPCR across the 17 lines. The amplification signal from genes on chromosome 31 was 2.2-fold greater than those on chromosomes 17, 34, and 36, confirming the double copy number of chromosome 31 (Mann-Whitney U, $P < 0.001$) (Mann and Whitney 1947). We then examined the read ratios that support allele calls, to establish whether the baseline coverage ($s = 1$) could be explained by monosomy (in which case, $s = 1.5$ and $s = 2.5$ would correspond to mixed populations) or disomy (in which case, $s = 1.5$ and $s = 2.5$ would correspond to trisomy and pentasomy, respectively). Considering that the lines here analyzed were cloned, monosomic chromosomes could be expected to be essentially homozygous at sequence level. This was not the case, as an average of 293 heterozygous SNPs were detected per strain on the nine chromosomes showing s = 1. Accordingly, we concluded that the latter chromosomes were disomic.

Fluorescent in situ hybridization (FISH) has been used to characterize variable aneuploidy for seven chromosomes from individual *L. major* parasites (Sterkers et al. 2010). Our sequencing approach complements and extends the observation and provided a quantification of chromosomal copy number, for all chromosomes, across a population of cells. Further analysis revealed three main categories of chromosomes: nine chromosomes (17, 18, 19, 21, 25, 28, 30, 34, 36) with disomy in all lines; chromosome 31 alone was tetrasomic in all 17 (Supplemental Fig. S10), and the 26 remaining chromosomes showed variable polysomy values between lines (Fig. 3). As a consequence of this polysomic diversity, the pattern of aneuploidy was unique for each line. Although experimentally induced SSG-resistant *L. infantum* strains showing aneuploidy have been associated with the drug-resistance phenotype (Leprohon et al. 2009a), no clear link was observed here: average *s* values of each individual chromosome in the SSG-resistant and SSG-susceptible strains were not significantly different (Supplemental Table S14). An ongoing investigation into genome stability in BPK282/0cl4 revealed that aneuploidy was stable for at least 32 passages after genome sequencing, though we cannot exclude changes between the time of isolation and the 30 passages prior to genome sequencing.

Structural variation in *L. donovani*

Within each chromosome, we used read-depth variation to identify loci with repeated genes (see Supplemental Table S15 for BPK282/0cl4) and to assess copy-number variation (CNV) in the 17 strains. Evidence of large duplications and deletions were discovered

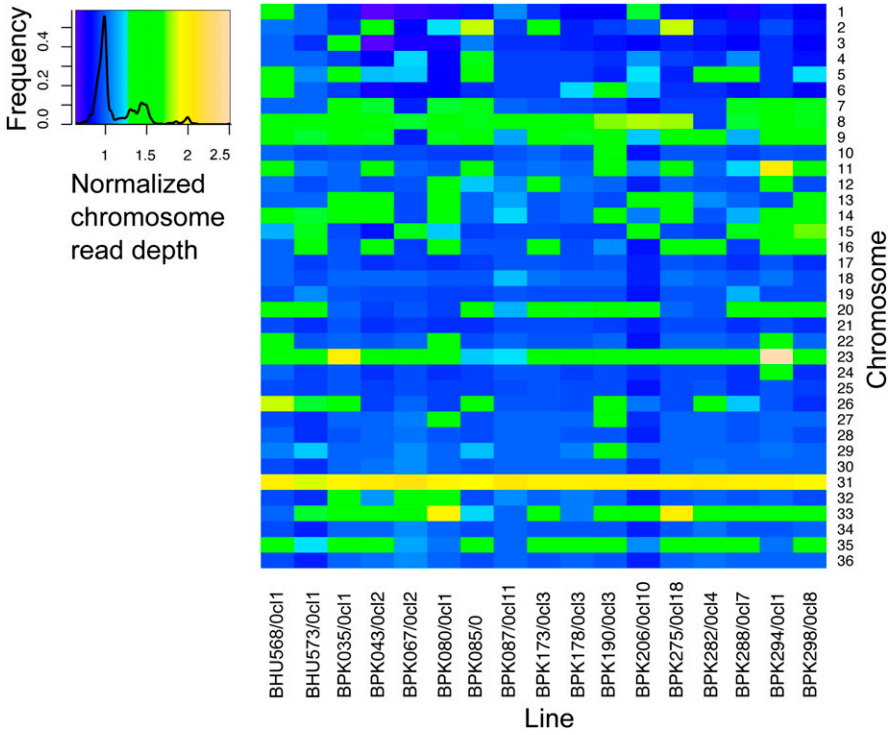

Figure 3. Aneuploidy in natural populations of *L. donovani*. The heatmap shows the copy-number status of the 36 chromosomes for the 17 lines as disomic (1, blue), trisomic (1.5, green), tetrasomic (2, yellow), and pentasomic (2.5, ivory). The color key shows the normalized chromosome read-depth and the distribution frequency.

(Supplemental Figs. S11–S14) as well as clustered single-copy genes flanked by either short direct repeats, ribosomal mobile elements (RIMES), or both. A subset of these protein-encoding genes, in addition to both the mini-exon and rDNA loci, significantly discriminated SSG-resistant from SSG-susceptible lines (Table 3).

Two tandem repeats were analyzed in detail: rDNA transcription units (chromosome 27) and mini-exon genes (chromosome 2). Copy number per chromosome ranged from six to 15 (mean 10.2 ± 3.4) for the rDNA genes, and from 26 to 146 (88 ± 28) for the mini-exon genes, consistent with reported values (Kebede et al. 1999; Inga et al. 1998). Gene dosage per cell was estimated taking into account the somy (copy number) values of the respective chromosomes. SSG-resistant parasites showed on average 44% fewer rDNA transcription units/cell than the sensitive ones (MWU P = 0.046), and 77% more mini-exon genes (MWU P = 0.004). Consequently, the ratio of mini-exon to rDNA genes tended to be higher in SSG-resistant lines (MWU P = 0.006). In addition, the SSG resistance locus (LdBPK_310950) was present in more copies in SSG-resistant lines than in susceptible ones (24 vs. 20, MWU P = 0.046). This reflects a pattern previously observed in *L. tarentolae*, where the ortholog of LdBPK_310950 was more highly expressed in samples selected for SSG resistance than in susceptible ones (Haimeur and Ouellette 1998).

Genome segments flanked by repeats can form extrachromosomal circular episomes through homologous recombination between regions adjacent to the repeats (Leprohon et al. 2009a). Here we observed two such regions: a 13.5-kb region on chromosome 23 encompassing four genes (the H-locus, including the ABC-thiol transporter MRPA gene) (Grondin et al. 1996), and a 15.8-kb segment on chromosome 36 containing four

Table 3. Major structural variants in *L. donovani* clinical lines differentiated SSG-resistant and SSG-susceptible groups

Chromosomal location	Loci and gene products	Gene ID	Median copy number	
			SSG-R	SSG-S
2: 267,501–267,539	Mini-exons	None[a]	288	156
8: 326,771–327,502	Amastin	LdBPK_080760	15	9
10: 213,241–213,495	GP63, leishmanolysin metallo-peptidase Clan MA(M) Family M8	LdBPK_100510[b]	31	17
27: 1,016,848–1,021,502	28S RNA gamma, 28S RNA alpha, 5.8S RNA, 18S RNA	None[c]	15	27
31: 354,449–359,110	SSG resistance protein	LdBPK_310950[d]	24	20
36: 2,527,016–2,542,828	Tartrate-sensitive acid phosphatase	LdBPK_366740[e]	31	23
	Hypothetical protein	LdBPK_366750[f]		
	MAPK homolog	LdBPK_366760[e]		
	Histidine secretory acid phosphatase	LdBPK_366770[g]		

Regions with a differential copy-number between antimony-resistant (SSG-R) and antimony-sensitive (SSG-S) lines (MWU, $P < 0.05$; Supplemental Table S14). Copy number of the respective loci or genes is shown as median per set. All are repetitive genes (GeneDB, www.genedb.org) except a candidate episome region referred to here as the *MAPK*-locus (mitogen-activated protein kinase) was flanked by short repeats and/or ribosomal mobile elements (RIMEs) containing a CNV encompassing five genes.
[a]Involved in mRNA splicing.
[b]Implicated in parasite virulence (see Matlashewski 2001).
[c]Functions in RNA translation.
[d]Overexpressed in antimonial-resistant *L. tarentolae* (Haimeur and Ouellette 1998).
[e]Mitogen-activated protein kinase; associated with survival in macrophage (Saha et al. 1985).
[f]Linked to SSG-resistance (Singh et al. 2007).
[g]A gene encoding a ubiquitin fusion degradation protein (LdBPK_366780) is at the 3′ end of the region.

genes referred to here as the *MAPK*-locus (Fig. 4). We examined the *MAPK*-locus to determine if the higher read-depth observed was due to the presence of episomal DNA by using an assay combining alkaline lysis (to extract circular DNA) and PCR amplification with primers situated at the edges of the chromosomal locus and oriented externally (Fig. 4). With this design, PCR amplification should be only possible when episomes are present, and assuming that circularization would occur through homologous recombination between the 822-bp region shared by two acid phosphatase genes located at both edges of the locus (LdBPK_366740 and LdBPK_366770), an amplicon of 1700 bp would be expected. We observed amplicons of that size (Supplemental Fig. S15) and the nucleotide-level sequence of the amplicon further verified this assay. Gene dosage per cell showed that there were more copies of the *MAPK*-locus in SSG-resistant than *MAPK*-sensitive lines (median 31 vs. 23, MWU $P < 0.05$) (Table 3). In contrast, no significant differences were observed between SSG-resistant and SSG-sensitive lines for the H-locus (median 8 vs. 7; Supplemental Table S14). The ongoing genome stability study above revealed that the read-depth of the *MAPK*-locus was stable from the initial sequencing time point to passage number 32, though again, we cannot exclude that changes might have occurred since the parasite was originally isolated.

We analyzed the influence of somy variation on gene transcription by comparing the coverage of cDNA and genomic DNA reads mapped to the reference strain BPK282/0cl4: A poor correlation was observed between the genomic and cDNA CDS coverage read-depth ($r^2 = 0.29$, *t*-test $P < 7.75 \times 10^{-4}$), but this increased when chromosome 31 was excluded ($r^2 = 0.54$, *t*-test $P < 4.75 \times 10^{-7}$)—the latter tetrasomic chromosome showed a similar cDNA read-depth as the disomic ones. Coverage for individual genes was proportional to the genomic copy number for repetitive genes (Supplemental Table S15), though with some notable exceptions (e.g., alpha tubulin or elongation factor 1-alpha loci), likely reflecting both the increased stability of the corresponding mRNAs at the life stage at which RNA was extracted, and also the complexity of transcription on *Leishmania*.

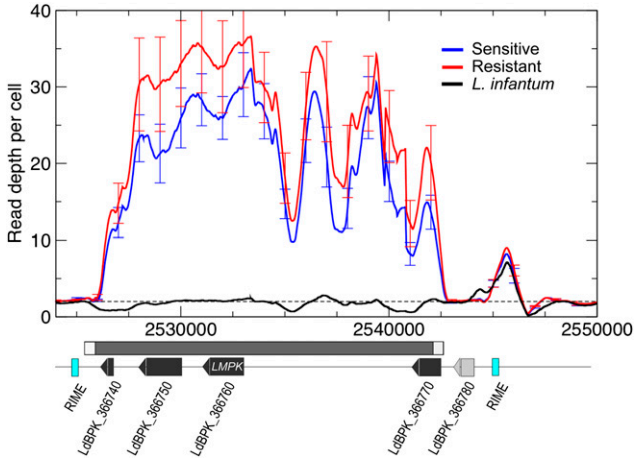

Figure 4. Copy-number variation spanning *MAPK*-locus on chromosome 36. Copy number per cell in *L. donovani* (red, SSG-R; blue, SSG-S) was much higher for the *MAPK* locus than the chromosome 36 median for *L. donovani* (black dashed line) and *L. infantum* (black). The amplified region is marked in gray and their white edges indicate the location of two direct repeats. Genes within and adjacent to the amplified region are given in black and light gray, respectively. The *MAPK* (mitogen-activated protein kinase) gene is labeled as *LMPK* (LdBPK_366760). Recombination events may occur through a highly homologous 220-bp region starting at the start codon of the tartrate-sensitive acid phosphatase (LdBPK_366740) and histidine secretory acid phosphatase (LdBPK_366770). LdBPK_366750 encodes a protein of unknown function. Two RIME (*L. donovani* ribosomal mobile elements; light-blue) flank the duplicated region. Read-depth values were computed using 1-kb sliding windows, and error bars are shown for each kilobase.

Discussion

An ability to underpin large epidemiological studies with genome-wide discovery of new variants is critically important. In this study, we demonstrated the feasibility of using deep genome sequencing to first produce a relevant reference genome, and second, to use it to characterize natural variation in clinically isolated *Leishmania donovani* parasites. Our study focused on the Indian subcontinent, where *L. donovani* sequence diversity is low and a high sensitivity to detect relevant variants is essential. Within this closely related set of parasites from a single disease focus, whole-genome sequencing resolved the details of population structure not visible using traditional multilocus typing methods (Alam et al. 2009).

We observed a low level of SNP variation between the 17 *L. donovani* lines sequenced here in comparison to other eukaryotes and *Leishmania* species (Laurent et al. 2007). An excess of rare nonsynonymous alleles were distributed across the genomes beyond levels predicted by interspecific comparisons, suggesting that a recent population recovery after a sustained bottleneck may explain this pattern of diversity (Nielsen 2001; Hermisson 2009), fitting the epidemiology of VL re-emergence and spreading after DDT spraying campaigns. This pattern may mask signals of adaptive evolution (Elyashiv et al. 2010). Consequently, genotype and phenotype changes in *L. donovani* should be examined in the context of a dynamic population structure as well as selective processes induced by drug treatment.

In contrast to the limited SNP variation among the 17 lines, extensive variation in chromosomal copy number was discovered. Aneuploidy is known to arise in *Leishmania*, but for only a small number of chromosomes and under experimental conditions, such as gene knockouts (Cruz et al. 1993) or induced in vitro drug resistance (Ubeda et al. 2008; Leprohon et al. 2009a). Aneuploidy has previously been observed in seven *L. major* chromosomes from individual parasite cells of a laboratory strain using FISH (Sterkers et al. 2010). In this study we documented the phenomenon for the first time on a genome-wide basis in clinical lines. Our results demonstrated extensive variation in aneuploidy within the 17 lines, implying global differences in gene dosage; however, the absence of a clear relationship with drug resistance contrasts with earlier experimental work (Leprohon et al. 2009a). Although aneuploidy was stable in culture, analysis of ploidy immediately after parasite isolation, or if possible, directly in the patient might yet reveal a correlation with a drug-resistance phenotype. Nonetheless, there was a remarkable disparity between the stable tetrasomy and disomy observed for one and nine chromosomes, respectively, and the highly variable somy of the 26 other chromosomes. Further work is needed to understand the origin of aneuploidy; it could be related to a chromosomal replication defect (Sterkers et al. 2010) or sexual recombination, followed by genome erosion as proposed in *Trypanosoma cruzi* (Lewis et al. 2009).

Besides aneuploidy, two other features of *Leishmania* genome plasticity contributed to gene-dosage differences on a finer scale. The first was the expansion and contraction of genes in tandem arrays. These are expressed through polycistronic transcription,

suggesting that the primary purpose of the arrayed structure is to regulate gene dosage. As a result, changes in copy number could contribute directly to variation in expression of antigen-related genes (Victoir and Dujardin 2002). The second feature was an episome that varied in copy number between our lines. Episomes have previously been detected after drug selection in vitro (Leprohon et al. 2009a), but for the first time, we document the amplification of the episomal *MAPK* locus, containing a *MAPK* gene (LdBPK_366760) likely to have a role in signal transduction (Wiese 2007) and an acid phosphatase gene (LdBPK_366770) that could be expressed during infection (Ellis et al. 1998). In one instance of induced drug resistance, the phenotype and the corresponding episome were unstable in the absence of the drug pressure (Leprohon et al. 2009a). However, resistance to antimonials is stable during in vitro maintenance (Laurent et al. 2007), suggesting that antimonial resistance has a negligible fitness cost (Vanaerschot et al. 2010), which may explain the stability of the *MAPK* episome in vitro. The biological role of gene dosage caused by this assortment of SVs remains unexplored. In trypanosomatids, gene expression is primarily regulated at the post-transcriptional level rather than at initiation (McDonagh et al. 2000), allowing up-regulation of expression by copy-number amplification (Victoir and Dujardin 2002). We have shown in the reference line that gene dosage strongly correlates with transcription levels, suggesting that there may indeed be functional consequences to this variability requiring exploration at proteome and metabolome levels.

We have also shown the contributing effects of both single-nucleotide and structural polymorphisms to the population genomics of these parasites. For instance, the SSG resistance locus (LdBPK_310950) had a substantial difference in copy number between the SSG-resistant and SSG-susceptible lines, and had only three distinct haplotypes despite the presence of six high-frequency amino acid substitutions. Moreover, separate genome-wide phylogenetic analyses based on SNPs or CNVs both converged on strikingly similar population structures. Thus, integrative approaches that combine SNP and SV data to maximize the variation that is included are likely to yield the greatest insights into the diversity of parasite biology.

Distinct patterns of diversity were also detected by both SNP and copy-number analyses in genes related to RNA stability and translation. A trend of ancient adaptive evolution was perceptible at six such genes, and extended regions of positive selection within the 17 lines were observed at loci encoding components of ribosomes and RNA-binding proteins (LdBPK_131120-1280, LdBPK_352230-2370). These nucleotide-level signatures of selection were paralleled by significant SV at two loci essential for translation: the rDNA and mini-exon genes. SSG-resistant lines had over twice as many mini-exon genes per rDNA transcription unit compared with SSG-sensitive ones. This intriguing disparity suggests that systemic adaptations, which substantially and globally alter the proteome, have a role in drug resistance—a concept that is supported by a previous metabolomic study, where 100 out of 340 detected metabolites differed in quantity between the drug resistant and sensitive phenotypes (t'Kindt et al. 2010).

In this study we have attempted to link genome-wide variation of *L. donovani* strains to their differential responses to drug exposure. We found evidence that drug resistance has multiple origins even in closely related lines, mirroring results from other parasites (Hawkins et al. 2008), and the associated differing genomic signatures in our small sample of isolates is compatible with a pleomorphic response. A full understanding of the emergence of drug resistance even in this restricted region will require validation in more strains. Even then, the genetic switches associated with the phenotype diversity may be specific to this population, so a broader sampling of genetic variation in *L. donovani* will be needed to comprehensively describe population-level *Leishmania* variability in a clinical context, particularly given that the substantial genetic divergence between our reference Nepalese *L. donovani* BPK282/0cl4 and the reference Spanish *L. infantum* JPCM5 strain suggests a significant regional component to diversity in the *L. donovani* species complex. Beyond deeper sampling, improved genome assemblies will improve our ability to identify SVs, enhancing combined analyses of SNPs and SVs, including identifying episomes and variable chromosome copy number.

In terms of molecular adaptations of antimonial-resistant parasites, our results support the concept of a systemic and pleomorphic response by *L. donovani*. On one hand, resistant and sensitive parasites differ by a series of genetic features, some of them being possibly related directly with an adaptation to antimonial pressure (like detoxifying enzymes or pumps), others being connected with housekeeping adaptations (like the hits on translation machinery here evidenced). The concept of a systemic adaptation is supported by metabolomic studies undertaken on strains studied here (t'Kindt et al. 2010). It is likely that there is an evolutionary cascade of adaptations, possibly to compensate the cost of a previous one or to provide a more adequate frame for its expression (for instance through adaptation of the translation machinery). Further work at population and experimental levels is required to elucidate this chain of events, but the power of full genome sequencing and other "omics" is obviously unique to address these questions in a systems biology approach. On the other hand, the emergence of multiple origins of drug resistance in this population together with different genomic signatures encountered among antimonial-resistant parasites is compatible with a pleomorphic response of *L. donovani*. This is supported by previous studies targeting specific genes and pointing out different patterns of overexpression in antimonial-resistant parasites (Adaui et al. 2011).

This study lays the groundwork for more sensitive and powerful approaches to collating genomic diversity, as well as comprehensively describing population-level *Leishmania* variability in a clinical context. Within the region *L. donovani*, infection levels are both endemic and epidemic. Changes in drug policy in the Indian subcontinent (from SSG to miltefosine and probably to combination treatment in the near future) are likely to result in dynamic selective pressures that mould the genome of this parasite (MacLean et al. 2010). Continuous surveillance must be maintained to monitor the threat from the ongoing emergence of drug resistance. The *L. donovani* reference genome provides a tool to identify and analyze new variants as they emerge. With available second-generation sequencing technologies, the approach of developing a reference and then studying the local population genetics from within any study region, and for any parasite, is now readily possible.

Methods

Sample collection

All 17 *L. donovani* isolates were obtained from Nepalese and Indian patients with visceral leishmaniasis (Supplemental Table S3; Rijal et al. 2010). Species identity of the 17 lines was confirmed by PCR–RFLP analysis of cysteine proteinase b genes (Quispe-Tintaya et al. 2004), and 16 isolates were cloned by the microdrop method (Van

Meirvenne et al. 1975). The 16 cloned lines and uncloned BPK085/0 used in present study are referred to as lines. A line (BPK282/0cl4, also known as BPK282/0 clone 4) was chosen as a biological reference for determination of the *L. donovani* genome by the Kaladrug-R and Gemini consortia (www.leishrisk.net/kaladrug, www.leishrisk.net/gemini) based on its characteristics of being drug-sensitive, easy to manipulate in vitro, and a representative of the most frequent microsatellite group in the Indian subcontinent (Alam et al. 2009). In addition, it is sensitive to SSG and miltefosine, has good infectivity both in animals in vivo and in macrophages in vitro, and was isolated from a cured patient in the Nepalese endemic area with a documented treatment outcome.

Sequencing and assembly of genome sequence data

A reference *L. donovani* genome sequence for isolate BPK282/0cl4 was produced using 1.29 million single-end and 3.58 million paired-end 454 reads with an average length of 167 bp and a mean insert size of 3 kb, corresponding to 815 Mb of DNA (Supplemental Methods). A total of 96% of the reads were assembled using Newbler (Quinn et al. 2008) into an initial set of reference contigs and scaffolds, with N50 values of 22.3 and 680 kb, respectively.

Three complementary genome assembly tools were used to improve the accuracy of the BPK282/0cl4 contig sequences (Supplemental Fig. S1). First, homopolymer errors induced by the 454 pyrosequencing were corrected by the Illumina reads, which had a median depth of 52-fold for BPK282/0cl4: These shorter 76-base reads had a more uniform genome-wide distribution, and thus fewer homopolymer errors. Miscalled single bases and indels in the 454 contigs were corrected with iCORN (Iterative Correction of Reference Nucleotides) (Otto et al. 2010), run for 14 iterations. Second, the contigs were iteratively extended by locally assembling Illumina data into gaps using IMAGE (Iteratively Mapping and Assembly for Gap Elimination) (Tsai et al. 2010) for 18 iterations. In order to eliminate additional errors introduced by IMAGE, iCORN was repeated for five further iterations. Third, the refined contigs ordered and oriented against the closely related and largely orthologous *L. infauntum* genome (Peacock et al. 2007) using ABACAS (Algorithm Based Automatic Contiguation of Assembled Sequences) (Assefa et al. 2009).

Quality control and read mapping of multiple lines

Excluding PCR duplicate reads, 41.0 Gb of sequence was generated for the 17 lines using the Illumina platform: a mean of 31.7 million reads per Illumina run (Supplemental Methods). On average, 27.3 million reads per lane (86.1%) were correctly paired and mapped to chromosomes, and 1.7 million per lane (5.4%) to contigs not assigned to chromosomes or to kDNA (Supplemental Tables S16, S17).

Variability between Illumina sequencing runs made it necessary to develop approaches to ensure that the data analyzed was of high quality (Malhis and Jones 2010). Using the expected GC content distribution from *L. infantum* (Peacock et al. 2007), the GC distributions for the reads for each line were examined: The median was 0.62 for most data sets, but poor-quality ones had lower values (Supplemental Fig. S16). Empirical examination of the sequence quality parameters and read-depth coverage with Samtools v0.1.6 (Li et al. 2009) showed that the mode and median of base-quality score (BQ) and coverage could distinguish informative from substandard sequencing data (Supplemental Fig. S17). The latter had lower median coverage (<17 vs. 41+), coverage mode (<8 vs. 25+), median BQ (<58 vs. 72+), BQ mode (<34 vs. 65+). BQ was the phred-style log-scaled adjusted probability that the given base was incorrect.

Reads for each line were mapped to the reference chromosomes with SSAHA2 (Ning et al. 2001) using the Smith-Waterman algorithm (Smith and Waterman 1981). All nonmapping and artificial duplicate reads were removed, as were sites with less than threefold coverage, or with low BQ (<26) or mapping quality scores (<31). Mapping quality reflected the probability of a given read mapping uniquely to the genome sequence, such that reads derived from DNA repeats with multiple high-orthology hits were excluded.

Classifying kinetoplast DNA

Homology between all known minicircle (381) and maxicircle (112) DNA sequences from GenBank (www.ncbi.nlm.nih.gov) and all contigs that could not be assigned to chromosomes was assessed using MegaBLAST (Zhang et al. 2000) for pairs whose E value <0.01 and score >100 to remove short segmental partial hits. The *L. infantum* and *L. major* genomes were included to identify contigs that were chromosomal. This assigned 26 contigs (22.0 kbp) to minicircles, two (19.0 kbp) to maxicircles; those that hit one or more of the four categories (*L. infantum*, *L. major*, minicircle kDNA, and maxicircle kDNA) were assigned as miscellaneous. The amount of reads mapped to kDNA varied among the samples from ~10%–1% of the total, reflecting the difficulty in quantitatively sequencing kDNA, particularly for minicircles due to their small length.

Gene annotation verification

Annotating the *L. donovani* genome was aided by the shared genomic characteristics of *Leishmania* species with respect to codon usage bias (Myler et al. 1999), its lack of *cis*-splicing and conserved synteny (Peacock et al. 2007), as well as the relatively close genetic relationship between *L. infantum* and *L. donovani*. *L. infantum* gene models (June 2010) were transferred to *L. donovani* based on sequence conservation and synteny with RATT (Rapid Annotation Transfer Tool) (Otto et al. 2011). The resultant *L. donovani* gene models were manually inspected and corrected where appropiate using Artemis (Rutherford et al. 2000; Carver et al. 2008) and the Artemis Comparison Tool (Carver et al. 2005). A scan for novel open reading frames (ORFs) longer than 450 bp with distinct patterns of codon usage and sequence similarity with other genes in GenBank did not identify any new gene models.

SNP identification and validation

A series of SNP quality validation steps were implemented to exclude sequencing errors and mapping artefacts. Candidate variable sites and indels were initially identified with Samtools from loci having a read-depth of <1000-fold and a maximum of three SNPs in any given seven-base region. Perl scripts were used to select only those candidate SNPs that met the following criteria: (1) the read coverage for the variant base on the forward or reverse strand >3; (2) a high SNP quality (>30); (3) genotypes were known in at least half of the 34 genotypes; (4) at least one read-length (76 bp) distant from contig gaps; and (5) normalized read-depth coverage was within two standard deviations of the chromosome's normalized median value. SNP quality was the average BQ-weighted mapping quality for all reads mapping to the site. If a candidate SNP was heterozygous, the variant BQ was checked to ensure it was not significantly worse than the reference genotype BQ.

In order to identify repetitive, arrayed, and paralogous DNA segments, read-depth coverage was normalized for each chromosome by first measuring the coverage distribution at every position

to which at least one read mapped (thus avoiding bias due to gaps). After excluding sites whose coverage was more than one standard deviation from the initial chromosome median, an optimized chromosome median and standard deviation were determined for interchromosomal comparisons of coverage, so that chromosomes whose coverage was consistently higher or lower than other chromosomes could be highlighted.

To counter elevated false-positive detection rates in repetitive and paralogous regions, the average local uniqueness was determined using a sliding window approach. Up- and downstream from each site (on both strands), the minimum average point at which a local region became unique (*k* bases) was evaluated. A total of 97.9% of regions not containing gaps were unique for $k < 30$ (Supplemental Fig. S18), and for this threshold ($k < 30$) the relative rates of singleton discovery in screened (30%) and unscreened (45%) SNP sets were stable (Supplemental Fig. S19). All sites were unique for $k > 675$, indicating that this was the maximum repeat size in this version of the *L. donovani* genome.

In order to examine the validity of SNPs in both polymorphic and paralogous regions, candidate SNPs were examined using PCR of four amplicons (Supplemental Table S18) and with SNP genotyping. The SNP genotyping assayed 289 possible SNPs as representatives of an initial set of 7642 variants in the 17 strains and was conducted in duplicate using mass spectrophotometry of allele-specific amplified DNA on the Sequenom platform (Buetow et al. 2001).

Population-wide allele frequencies were estimated using read-depth coverage values for each of the 17 lines pooled together, in order to minimize epistatic biases associated with next generation sequencing strategies (Jiang et al. 2009; Lynch 2009). A scale of SNP likelihood was used for interpreting the SNP genotyping results using the aligned *L. infantum* genome to determine the derived allele for each of the 289 genotyped sites (Supplemental Fig. S20). This approach was used to calculate the derived and folded allele frequencies for the 3549 validated SNPs for the 17 lines.

Structural variation discovery

In devising the CNV detection criteria, we aimed to develop sensitive methodology to identify real biological signals of CNVs in the context of stochastic variability of read coverage and misassembly (Lander and Waterman 1988). Median normalized read-depth was measured at bases located within one standard deviation of the initial chromosomal median to remove outliers caused by assembly gaps, spurious high-coverage regions, and CNVs. Normalized read-depth per haploid genome (d) was defined as a raw depth (d_r), divided by the median depth of its chromosome (d_{ch}): so that $d = d_r/d_{ch}$.

To get chromosomal read-depths, the sequence data for each sample was normalized using the median depth of 36 chromosomes of a line (d_{mch}) to make comparisons between samples, and the normalized median depths of disomic, trisomic, tetrasomic, and pentasomic chromosomes become ~1.0, 1.5, 2.0, and 2.5, respectively. So for a normalized chromosome read-depth or somy value, $s = (d_{ch}/d_{mch})$ a read-depth per cell (d_{pc}) was defined as $d_{pc} = 2ds = 2d[d_{ch}//d_{mch}] = 2[d_r/d_{mch}]$, thus reflecting differences in somy number.

The detection of structural diversity was based on the assumption that all genomic regions have equal probabilities of being sequenced by any read, such that coverage was a Poisson-distributed variable approaching a normal distribution as coverage increases based on the central limit theory (Sebat 2007). Consequently, SVs were detected with the read-depth coverage distribution, and read-pair insert size distributions, microhomologies (Hastings et al. 2009), and the densities of singleton reads whose mate pair was unmapped were used to confirm or reject candidate SVs.

In order to increase the specificity of SV analysis, candidates were excluded if they met any of the following criteria: (1) their length <1 kb or their normalized read-depth coverage was not within two standard deviations of the normalized median for the chromosome; (2) they were at chromosome ends; (3) they had a pattern of uniform heterozygosity in the 17 lines; (4) they were in missing paralogous regions in comparison with *L. infantum*; or (5) they were located in a region with low uniqueness (Supplemental Fig. S18). In addition, SVs found in samples with high read-depth variation (BPK043/0cl2, BPK067/0cl2, and BPK298/0cl8) were not considered unless they were longer than 3 kb. *L. donovani* candidate regions were compared against *L. infantum* chromosomes using BLAST (Altschul et al. 1990) and visualized in ACT (Artemis Comparison Tool) (Carver et al. 2008). The depth coverage of *L. infantum* reads mapped against BPK282/0cl4 also identified false candidate SVs that were present in *L. infantum*. Small-scale SVs were assessed using *d* to avoid calling SVs due to somy differences.

To determine genetic distances in the 17 lines to construct a neighbor-joining tree, effective normalized copy numbers were estimated for each CNV from the normalized coverage levels. These copy numbers were adjusted for effective normalized read-depths (d_{n2}) relative to the minimum (min_d) and maximum (max_d) values in the set of samples, such that $d_{n2} = (d–min_d)/(max_d–min_d)$ in order to assess CNV variation in a quantitative manner. This approach precluded CNVs with extreme copy-number values from skewing the analysis.

Experimental validation of the candidate episome

In order to prepare extrachromosomal circular DNA, an estimated 107 promastigotes of *L. donovani* line BPK275/0cl18 were selectively denaturated by alkaline lysis (Sen et al. 1992). In order to verify that extrachromosomal circular DNA contained amplicons generated by circularization of the *MAPK* locus, externally oriented primers homologous to the locus edges (TCTTGGCACGGCATC AGCAG and CATGGCGCAGTGACCTTCAG) were used for 40 PCR amplification cycles (Supplemental Methods). The nucletoide sequence of the amplified sequences was determined with an ABI 3730 automated sequencer (PerkinElmer).

Comparing the *L. donovani* genome to orthologous *Leishmania* sequences

In order to compare orthologous sites between *Leishmania* species, a series of tools were used to first identify conserved sequences; second, expand the orthologous regions between genomes; and third, align the orthologous pair (Nygaard et al. 2010). The initial identification of conserved regions was conducted with Genscan (Burge and Karlin 1997) to find genes to serve as syntenic anchors for BLAT comparisons conducted in protein space (Kent 2002), which were subsequently computed by Mercator (Dewey 2007). Mercator iteratively expanded the local pairwise homologous regions across the chromosomes to produce maps of syntenic regions relative to reference *L. donovani* BPK282/0cl4 (*L. infantum*: 87 syntenic regions; *L. major*: 102; *L. mexicana*: 95; *L. braziliensis*: 237). Alignment data produced by Mercator were combined with a phylogenetic tree of the species-informed base-level alignments of the orthologous regions (Mavid) (Bray and Pachter 2004). FSA (Fast Statistical Alignment) (Bradley et al. 2009) estimated the local level of homology of these synteny block alignments using MUMmer v3.0 (Kurtz et al. 2004), and inserted gaps where ambiguities arose to ensure that any predictions of orthology made were conservative.

Substitutions at nonsynonymous, synonymous, and intergenic sites were determined, as were the derived alleles in the 17 *L. donovani* lines and levels of divergence between species. To obtain a comparative measure of variation between *Leishmania* species, a neighbor-joining phylogenetic tree was constructed for all genomic orthologous sites with Treepuzzle v5.2 (Schmidt et al. 2002). Genomic scans for highly divergent regions between species indicated faster rates of ancestral change—these can be symptomatic of recent selective sweeps (Nielsen 2001).

The pairwise ratio d_N/d_S (ω) was calculated for each CDS alignment using the codeml implementation of the PAML 4.4 package (Yang 1997, 2007), where d_N was the number of nonsynonymous mutations per nonsynonymous site, and d_S the equivalent for synonymous sites. If synonymous and nonsynonymous mutations were neutral, it was expected that $\omega = d_N/d_S = 1$ (Yang 2002). Departures from this, where $\omega > 1$ ($d_N > d_S$), suggested that nonsynonymous mutations could be advantageous and were maintained under directional selection. If $\omega < 1$ ($d_N < d_S$), then purifying selection may have eliminated deleterious nonsynonymous variants (Yang 2002). One-ratio models estimated ω across the *Leishmania* lineage for each gene and this was compared with models with ω fixed as 1 using likelihood ratio tests (LRTs; Supplemental Methods). In addition, LRTs between variable-ω and fixed-ω branch-site models were conducted to find specific sites under selection in the *L. donovani* lineage (Nielsen and Yang 1998). These tests were complementary to branch-specific approaches published for *L. infantum*, *major*, and *braziliensis* (Peacock et al. 2007).

Population diversity and structure analysis

Several complementary approaches were used to investigate the genetic structure releveant to the clinical characteristics of the lines. Median-joining networks of SNPs segregating at nonsynonymous and synonymous sites were constructed using Network v4.2.0.1, excluding noninformative uniformly heterozygous sites (Bandelt et al. 1999, www.fluxus-technology.com). Genetic differentiation between SSG-resistant and SSG-susceptible groups was assessed using F_{ST} (Wright 1951) based on the relative nucleotide diversity (π) (Tajima 1983): sites with $F_{ST} > 0.4$ (1.2%) had gene-flow estimates 16 times lower than those with $F_{ST} = 0.05$ (Holsinger and Weir 2009). Possible associations between genomic F_{ST} values, geographic great circle distance, and SSG phenotypes for each pair of lines were also explored using Mantel-test approach (Mantel 1967).

As a result of the low SNP-level diversity and the unique nature of genome-wide haplotypes, allele frequency-based tests were more instructive than haplotype-focused ones. To reflect the lack of linkage information between genotypes, population metrics were determined for each SNP as paired rather than individual genotypes to avoid artifically constructing a balanced pattern of diversity. These summary statistics included Watterson's θ_W, π, Fu's F_S, and Tajima's D (Tajima 1989); the latter reflected the standardized sample size-corrected difference between θ_W and π, and should be minimal under neutrality. Sample sizes were corrected for unknown genotypes at each SNP site. In order to distinguish regions with differential diversity signals due to recent demographic history, only extended regions were considered in a sliding-window scan of intraspecific values of θ_W, π, and D.

The population-level spectrum of evolutionary change was calculated for each gene as p_N/p_S, the ratio of nonsynonymous (p_N) to synonymous mutations (p_S). Adjusting this ratio for L_N (the number of nonsynonymous sites) and L_S (synonymous), estimated using KaKs_Calculator (v1.2) (Zhang et al. 2006), gives the fraction of neutral alleles segregating: $f = [p_N/p_S]/[L_N/L_S]$, assuming little temporal variation of selection (Smith and Eyre-Walker 2002). Using p_N/p_S values for each gene and the ratio of fixed nonsynonymous (D_N) and synonymous (D_S) substitutions measured between the orthologous alignments of the *L. donovani* and each of the other *Leishmania* genomes, a series of McDonald–Kreitman-based tests were implemented (McDonald and Kreitman 1991) for a fixation index, $FI = [D_N/D_S]/[p_N/p_S]$. Because nonsynonymous and synonymous sites were intercalated in coding sequences and so closely share genealogical histories, the absolute numbers of polymorphisms were examined rather than the rates alone, a viable approach to investigating diversity in structurally variable genes of indeterminant lengths (Nei et al. 2010). If $FI > 1$, it indicated an elevated ancestral fixation rate of protein-level changes in *L. donovani* (McDonald and Kreitman 1991). If $FI < 1$ with a high p_N/p_S, this was likely to reflect relaxed selective constraint within the population; or if D_N/D_S was low, ancestral purifying selection (Eyre-Walker 2002). The expected contingency table values of D_N, D_S, p_N, and p_S for each gene were determined and summed to calculate an expected fixation index (eFI) (Gojobori et al. 2007). The difference between FI and eFI, $\alpha = FI/eFI–1$, was calculated to evaluate the genomic scale of the abundance of adaptive and deleterious alleles segregating in the population (Axelsson and Ellegren 2009).

To combat the challenge of searching for evidence of selective processes in a data set with limited rates of intraspecific polymorphism, the direction of selection (DoS) test was conducted as $DoS = D_N/[D_N + D_S]–p_N/[p_N + p_S]$ (Stoletzki and Eyre-Walker 2011) for genes with $p_N > 0$: This also had the advantage of being less prone to biases from purifying selection. In addition, genes likely to possess a high fraction of adaptive variants were further identified by determining $\alpha = 1–1/FI$ (Smith and Eyre-Walker 2002).

Sites variable between species were determined as fixed if they were conserved in the *L. donovani* population: As a consequence, D_N/D_S values calculated for less-divergent species (such as *L. infantum*) may be biased by ancestral polymorphisms (Charlesworth 2009; Peterson and Masel 2009), and equally, saturation limitations increase for highly divergent species. Moreover, using counts rather than relative rates (D_N vs. d_N) of the numbers and types of mutations at structurally variable genes may be a viable compromise to omitting such loci.

Data access

The annotated reference *L. donovani* genome for BPK282/0cl4 is available at the EMBL Nucleotide Sequence Database (http://www.ebi.ac.uk/embl) (accession nos. FR799588–FR799623) and from GeneDB (http://www.genedb.org/Homepage/Ldonovani_BPK282A1) and the strain-specific genotypes are browsable via MapSeq (http://www.genedb.org/mapseq/Ldonovani).

Acknowledgments

We thank the core sequencing and genotyping groups, as well as the Parasite Genomics team from the Sanger Institute, especially Martin Aslett, Raeece Naeem, and Matthew Rogers. This work was funded by the Wellcome Trust (grant nos. WT 085775/Z/08/Z and 076355), the Kaladrug (grant no. EC-FP7-222895, www.leishrisk.net/kaladrug), the Gemini consortia (grant no. ITMA SOFI-B, www.leishrisk.net/gemini), and the Baillet-Latour Foundation.

Authors' contributions: C.H.F., G.H.C., G.S., H.I., J.C.D., M.B., S.D., and T.D. designed the study. I.M., S.D.D., S.D., S.R., and S.S. collected, documented, and maintained samples. M.A.Q. and M.S. conducted sequencing. M.V., O.S., and S.D.D. completed PCR assays. J.D.H. and J.C.M. carried out SNP validation. H.I. carried out assembly and mapping. H.I. and T.D. screened se-

quence data. T.D. aligned species' genomes. H.I., J.C.D., T.G.C., and T.D. performed analyses. H.I., J.A.C., J.C.D., M.B., and T.D. wrote the manuscript.

References

Adaui V, Castillo D, Zimic M, Gutiérrez A, Decuypere S, Vanaerschot M, De Doncker S, Llanos-Cuentas A, Arévalo J, Dujardin JC. 2011. Comparison of gene expression patterns among *Leishmania braziliensis* clinical isolates showing a different *in vitro* susceptibility to pentavalent antimony. *Parasitology* **138:** 183–193.

Alam MZ, Kuhls K, Schweynoch C, Sundar S, Rijal S, Shamsuzzaman AK, Raju BV, Salotra P, Dujardin JC, Schönian G. 2009. Multilocus microsatellite typing (MLMT) reveals genetic homogeneity of *Leishmania donovani* strains in the Indian subcontinent. *Infect Genet Evol* **9:** 24–31.

Alcolea PJ, Alonso A, Gómez MJ, Sánchez-Gorostiaga A, Moreno-Paz M, González-Pastor E, Toraño A, Parro V, Larraga V. 2010. Temperature increase prevails over acidification in gene expression modulation of amastigote differentiation in *Leishmania infantum*. *BMC Genomics* **11:** 31. doi: 10.1186/1471-2164-11-31.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403–410.

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

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:** 1073–1079.

Bandelt HJ, Forster P, Röhl A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* **16:** 37–48.

Bhattacharya A, Biswas A, Das PK. 2009. Role of a differentially expressed cAMP phosphodiesterase in regulating the induction of resistance against oxidative damage in *Leishmania donovani*. *Free Radic Biol Med* **47:** 1494–1506.

Bhattarai NR, Dujardin JC, Rijal S, De Doncker S, Boelaert M, Van der Auwera G. 2010. Development and evaluation of different PCR-based typing methods for discrimination of *Leishmania donovani* isolates from Nepal. *Parasitology* **137:** 947–957.

Bradley RK, Roberts A, Smoot M, Juvekar S, Do J, Dewey C, Holmes I, Pachter L. 2009. Fast statistical alignment. *PLoS Comput Biol* **5:** e1000392. doi: 10.1371/journal.pcbi.1000392.

Bray N, Pachter L. 2004. MAVID: Constrained ancestral alignment of multiple sequences. *Genome Res* **14:** 693–699.

Buetow KH, Edmonson M, MacDonald R, Clifford R, Yip P, Kelley J, Little DP, Strausberg R, Koester H, Cantor CR, et al. 2001. High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism markers by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Proc Natl Acad Sci* **98:** 581–584.

Burge C, Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268:** 78–94.

Cabral SM, Silvestre RL, Santarém NM, Tavares JC, Silva AF, Cordeiro-da-Silva A. 2008. A *Leishmania infantum* cytosolic tryparedoxin activates B cells to secrete interleukin-10 and specific immunoglobulin. *Immunology* **1234:** 555–565.

Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J. 2005. ACT: the Artemis Comparison Tool. *Bioinformatics* **21:** 3422–3423.

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

Castanys-Muñoz E, Pérez-Victoria JM, Gamarro F, Castanys S. 2008. Characterization of an ABCG-like transporter from the protozoan parasite *Leishmania* with a role in drug resistance and transbilayer lipid movement. *Antimicrob Agents Chemother* **5210:** 3573–3579.

Castro H, Sousa C, Novais M, Santos M, Budde H, Cordeiro-da-Silva A, Flohé L, Tomás AM. 2004. Two linked genes of *Leishmania infantum* encode tryparedoxins localised to cytosol and mitochondrion. *Mol Biochem Parasitol* **136:** 137–147.

Charlesworth B. 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* **103:** 195–205.

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

Cui Q, Purisima EO, Wang E. 2009. Protein evolution on a human signaling network. *BMC Syst Biol* **3:** 21. doi: 10.1186/1752-0509-3-21.

Decuypere S, Yardley V, De Doncker S, Laurent T, Rijal S, Chappuis F, Dujardin JC. 2005. Gene expression analysis of the mechanism of natural Sb(V) resistance in *Leishmania donovani* isolates from Nepal. *Antimicrob Agents Chemother* **49:** 4616–4621.

Decuypere S, Vanaerschot M, Rijal S, Yardley V, Maes L, De Doncker S, Chappuis F, Dujardin JC. 2008. Gene expression profiling in *Leishmania*: overcoming technical variation and exploiting biological variation. *Parasitology* **135:** 183–194.

Dewey CN. 2007. Aligning multiple whole genomes with Mercator and MAVID. *Methods Mol Biol* **395:** 221–236.

Dujardin JC. 2009. Structure, dynamics and function of *Leishmania* genome: resolving the puzzle of infection, genetics and evolution? *Infect Genet Evol* **9:** 290–297.

Ellis SL, Shakarian AM, Dwyer DM. 1998. *Leishmania*: amastigotes synthesize conserved secretory acid phosphatases during human infection. *Exp Parasitol* **89:** 161–168.

Elyashiv E, Bullaughey K, Sattath S, Rinott Y, Przeworski M, Sella G. 2010. Shifts in the intensity of purifying selection: An analysis of genome-wide polymorphism data from two closely related yeast species. *Genome Res* **20:** 1558–1573.

Eyre-Walker A. 2002. Changing effective population size and the McDonald-Kreitman test. *Genetics* **162:** 2017–2024.

Fu YX. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147:** 915–925.

Garin YJ, Meneceur P, Pratlong F, Dedet JP, Derouin F, Lorenzo F. 2005. A2 gene of Old World cutaneous *Leishmania* is a single highly conserved functional gene. *BMC Infect Dis* **51:** 18. doi: 10.1186/1471-2334-5-18.

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

Ghedin E, Charest H, Matlashewski G. 1998. A2rel: a constitutively expressed *Leishmania* gene linked to an amastigote-stage-specific gene. *Mol Biochem Parasitol* **93:** 23–29.

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* **104:** 3907–3912.

Grondin K, Roy G, Ouellette M. 1996. Formation of extrachromosomal circular amplicons with direct or inverted duplications in drug-resistant *Leishmania tarentolae*. *Mol Cell Biol* **16:** 3587–3595.

Haimeur A, Ouellette M. 1998. Gene amplification in *Leishmania tarentolae* selected for resistance to sodium stibogluconate. *Antimicrob Agents Chemother* **42:** 1689–1694.

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009. Mechanisms of change in gene copy number. *Nat Rev Genet* **10:** 551–564.

Hawkins VN, Auliff A, Prajapati SK, Rungsihirunrat K, Hapuarachchi HC, Maestre A, O'Neil MT, Cheng Q, Joshi H, Na-Bangchang K, et al. 2008. Multiple origins of resistance-conferring mutations in *Plasmodium vivax dihydrofolate reductase*. *Malar J* **7:** 72. doi: 10.1186/1475-2875-7-72.

Hermisson J. 2009. Who believes in whole-genome scans for selection? *Heredity* **103:** 283–284.

Holsinger KE, Weir BS. 2009. Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nat Rev Genet* **10:** 639–650.

Inga R, De Doncker S, Gomez J, Lopez M, Garcia R, Le Ray D, Arevalo J, Dujardin JC. 1998. Relation between variation in copy number of ribosomal RNA encoding genes and size of harbouring chromosomes in *Leishmania* of subgenus Viannia. *Mol Biochem Parasitol* **92:** 219–228.

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

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

Jiang R, Tavaré S, Marjoram P. 2009. Population genetic inference from resequencing data. *Genetics* **181:** 187–197.

Kebede A, De Doncker S, Arevalo J, Le Ray D, Dujardin JC. 1999. *Leishmania* of subgenus Viannia: size polymorphism of chromosomes bearing mini-exon genes in natural populations is due to rearrangement of the mini-exon gene array. *Int J Parasitol* **29:** 549–557.

Kent WJ. 2002. BLAT—the BLAST-like alignment tool. *Genome Res* **12:** 656–664.

Kim PM, Korbel JO, Gerstein MB. 2007. Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. *Proc Natl Acad Sci* **104:** 20274–20279.

König J, Fairlamb AH. 2007. A comparative study of type I and type II tryparedoxin peroxidases in *Leishmania major*. *FEBS J* **274:** 5643–5658.

Kumar D, Kulshrestha A, Singh R, Salotra P. 2009. *In vitro* susceptibility of field isolates of *Leishmania donovani* to Miltefosine and amphotericin B: correlation with sodium antimony gluconate susceptibility and

implications for treatment in areas of endemicity. *Antimicrob Agents Chemother* **532:** 835–838.
Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biol* **52:** R12. doi: 10.1186/gb-2004-5-2-r12.
Lander ES, Waterman MS. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2:** 231–239.
Laurent T, Rijal S, Yardley V, Croft S, De Doncker S, Decuypere S, Khanal B, Singh R, Schönian G, Kuhls K, et al. 2007. Epidemiological dynamics of antimonial resistance in *Leishmania donovani*: genotyping reveals a polyclonal population structure among naturally-resistant clinical isolates from Nepal. *Infect Genet Evol* **7:** 206–212.
Leprohon P, Légaré D, Raymond F, Madore E, Hardiman G, Corbeil J, Ouellette M. 2009a. Gene expression modulation is associated with gene amplification, supernumerary chromosomes and chromosome loss in antimony-resistant *Leishmania infantum*. *Nucleic Acids Res* **375:** 1387–1399.
Leprohon P, Légaré D, Ouellette M. 2009b. Intracellular localization of the ABCC proteins of *Leishmania* and their role in resistance to antimonials. *Antimicrob Agents Chemother* **536:** 2646–2649.
Lewis MD, Llewellyn MS, Gaunt MW, Yeo M, Carrasco HJ, Miles MA. 2009. Flow cytometric analysis and microsatellite genotyping reveal extensive DNA content variation in *Trypanosoma cruzi* populations and expose contrasts between natural and experimental hybrids. *Int J Parasitol* **39:** 1305–1317.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079.
Lukes J, Mauricio IL, Schönian G, Dujardin JC, Soteriadou K, Dedet JP, Kuhls K, Tintaya KW, Jirku M, Chocholová E, et al. 2007. Evolutionary and geographical history of the *Leishmania donovani* complex with a revision of current taxonomy. *Proc Natl Acad Sci* **104:** 9375–9380.
Lynch M. 2009. Estimation of allele frequencies from high-coverage genome-sequencing projects. *Genetics* **182:** 295–301.
MacLean RC, Hall AR, Perron GG, Buckling A. 2010. The population genetics of antibiotic resistance: integrating molecular mechanisms and treatment contexts. *Nat Rev Genet* **11:** 405–414.
Malhis N, Jones SJ. 2010. High quality SNP calling using Illumina data at shallow coverage. *Bioinformatics* **268:** 1029–1035.
Mandal G, Wyllie S, Singh N, Sundar S, Fairlamb AH, Chatterjee M. 2007. Increased levels of thiols protect antimony unresponsive *Leishmania donovani* field isolates against reactive oxygen species generated by trivalent antimony. *Parasitology* **134:** 1679–1687.
Mandal S, Maharjan M, Singh S, Chatterjee M, Madhubala R. 2010. Assessing aquaglyceroporin gene status and expression profile in antimony-susceptible and -resistant clinical isolates of *Leishmania donovani* from India. *J Antimicrob Chemother* **65:** 496–507.
Mann HB, Whitney DR. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat* **18:** 50–60.
Mantel N. 1967. Ranking procedures for arbitrarily restricted observation. *Biometrics* **23:** 65–78.
Matlashewski G. 2001. *Leishmania* infection and virulence. *Med Microbiol Immunol*. **190:** 37–42.
McDonagh PD, Myler PJ, Stuart K. 2000. The unusual gene organization of *Leishmania major* chromosome 1 may reflect novel transcription processes. *Nucleic Acids Res* **28:** 2800–2803.
McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* **351:** 652–654.
Mondal D, Singh SP, Kumar N, Joshi A, Sundar S, Das P, Siddhivinayak H, Kroeger A, Boelaert M. 2009. Visceral leishmaniasis elimination programme in India, Bangladesh, and Nepal: reshaping the case finding/case management strategy. *PLoS Negl Trop Dis* **3:** e355. doi: 10.1371/journal.pntd.0000355.
Myler PJ, Audleman L, deVos T, Hixson G, Kiser P, Lemley C, Magness C, Rickel E, Sisk E, Sunkin S, et al. 1999. *Leishmania major* Friedlin chromosome 1 has an unusual distribution of protein-coding genes. *Proc Natl Acad Sci* **96:** 2902–2906.
Nei M, Suzuki Y, Nozawa M. 2010. The neutral theory of molecular evolution in the genomic era. *Annu Rev Genomics Hum Genet* **11:** 265–289.
Nielsen R. 2001. Statistical tests of selective neutrality in the age of genomics. *Heredity* **86:** 641–647.
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:** 1725–1729.
Nygaard S, Braunstein A, Malsen G, Van Dongen S, Gardner PP, Krogh A, Otto TD, Pain A, Berriman M, McAuliffe J, et al. 2010. Long- and short-term selective forces on malaria parasite genomes. *PLoS Genet* **6**. doi: 10.1371/journal.pgen.1001099.
Obbard DJ, Jiggins FM, Bradshaw NJ, Little TJ. 2010. Recent and recurrent selective sweeps of the antiviral RNAi gene Argonaute-2 in three species of *Drosophila*. *Mol Biol Evol* **28:** 1043–1056.
Otto TD, Sanders M, Berriman M, Newbold C. 2010. Iterative correction of reference nucleotides iCORN using second generation sequencing technology. *Bioinformatics* **26:** 1704–1707.
Otto TD, Dillon GP, Degrave WS, Berriman M. 2011. RATT: rapid annotation transfer tool. *Nucleic Acid Res* **39:** e57. doi: 10.1093/nar/gkg1268.
Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail MA, Peters N, Adlem E, Tivey A, Aslett M, et al. 2007. Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. *Nat Genet* **39:** 839–847.
Peterson GI, Masel J. 2009. Quantitative prediction of molecular clock and ka/ks at short timescales. *Mol Biol Evol* **26:** 2595–2603.
Quinn NL, Levenkova N, Chow W, Bouffard P, Boroevich KA, Knight JR, Jarvie TP, Lubieniecki KP, Desany BA, Koop BF, et al. 2008. Assessing the feasibility of GS FLX Pyrosequencing for sequencing the Atlantic salmon genome. *BMC Genomics* **9:** 404. doi: 10.1186/1471-2164-9-404.
Quispe Tintaya KW, Ying X, Dedet JP, Rijal S, De Bolle X, Dujardin JC. 2004. Antigen genes for molecular epidemiology of leishmaniasis: polymorphism of cysteine proteinase B and surface metalloprotease glycoprotein 63 in the *Leishmania donovani* complex. *J Infect Dis* **189:** 1035–1043.
Rijal S, Yardley V, Chappuis F, Khanal B, Singh R, Boelaert M, De Doncker S, Croft S, Decuypere S, Dujardin JC. 2007. Antimonial treatment of visceral leishmaniasis: are current in vitro susceptibility assays adequate for prognosis of in vivo therapy outcome? *Microbes Infect* **9:** 529–535.
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:** 21–28.
Rogers MB, Hilley JD, Dickens NJ, Wilkes J, Bates PA, Depledge DP, Harris D, Her Y, Herzyk P, Imamura H, et al. 2011. Chromosome and gene copy number variation allow major structural change between species and strains of *Leishmania*. *Genome Res* (this issue). doi: 10.1101/gr.122945.111.
Rougeron V, De Meeûs T, Kako Ouraga S, Hide M, Bañuls AL. 2010. "Everything you always wanted to know about sex but were afraid to ask" in Leishmania after two decades of laboratory and field analyses. *PLoS Pathog* **68:** e1001004. doi: 10.1371/journal.ppat.1001004.
Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B. 2000. Artemis: sequence visualization and annotation. *Bioinformatics* **16:** 944–945.
Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG. 2007. Dynamic evolution of the innate immune system in *Drosophila*. *Nat Genet* **39:** 1461–1468.
Saha AK, Das S, Glew RH, Gottlieb M. 1985. Resistance of leishmanial phosphatases to inactivation by oxygen metabolites. *J Clin Microbiol* **22:** 329–332.
Samant M, Sahasrabuddhe AA, Singh N, Gupta SK, Sundar S, Dube A. 2007. Proteophosphoglycan is differentially expressed in sodium stibogluconate-sensitive and resistant Indian clinical isolates of *Leishmania donovani*. *Parasitology* **134:** 1175–1184.
Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18:** 502–504.
Sebat J. 2007. Major changes in our DNA lead to major changes in our thinking. *Nat Genet* **39:** S3–S5.
Sen S, Rani S, Freireich EJ, Hewitt R, Stass SA. 1992. Detection of extrachromosomal circular DNA sequences from tumor cells by an alkaline lysis, Alu-polymerase chain reaction technique. *Mol Carcinog* **5:** 107–110.
Simpson AG, Stevens JR, Lukes J. 2006. The evolution and diversity of kinetoplastid flagellates. *Trends Parasitol* **224:** 168–174.
Singh N, Almeida R, Kothari H, Kumar P, Mandal G, Chatterjee M, Venkatachalam S, Govind MK, Mandal SK, Sundar S. 2007. Differential gene expression analysis in antimony-unresponsive Indian kala azar visceral leishmaniasis clinical isolates by DNA microarray. *Parasitology* **134:** 777–787.
Singh R, Kumar D, Duncan RC, Nakhasi HL, Salotra P. 2010. Overexpression of histone H2A modulates drug susceptibility in *Leishmania* parasites. *Int J Antimicrob Agents* **36:** 50–57.
Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. *J Mol Biol* **147:** 195–197.
Smith NG, Eyre-Walker A. 2002. Adaptive protein evolution in *Drosophil*a. *Nature* **415:** 1022–1024.
Sterkers Y, Lachaud L, Crobu L, Bastien P, Pagès M. 2010. FISH analysis reveals aneuploidy and continual generation of chromosomal mosaicism in *Leishmania major*. *Cell Microbiol* **13:** 274–283.
Stevens JR, Noyes HA, Schofield CJ, Gibson W. 2001. The molecular evolution of Trypanosomatidae. *Adv Parasitol* **48:** 1–56.

Stober CB, Lange UG, Roberts MT, Alcami A, Blackwell JM. 2005. IL-10 from regulatory T cells determines vaccine efficacy in murine *Leishmania major* infection. *J Immunol* **175:** 2517–2524.
Stober CB, Lange UG, Roberts MT, Gilmartin B, Francis R, Almeida R, Peacock CS, McCann S, Blackwell JM. 2006. From genome to vaccines for leishmaniasis: screening 100 novel vaccine candidates against murine *Leishmania major* infection. *Vaccine* **24:** 2602–2616.
Stoletzki N, Eyre-Walker A. 2011. Estimation of the neutrality index. *Mol Biol Evol* **28:** 63–70.
Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105:** 437–460.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123:** 585–595.
t'Kindt R, Scheltema RA, Jankevics A, Brunker K, Rijal S, Dujardin JC, Breitling R, Watson DG, Coombs GH, Decuypere S. 2010. Metabolomics to unveil and understand phenotypic diversity between pathogen populations. *PLoS Negl Trop Dis* **4:** e904. doi: 10.1371/journal.pntd.0000904.
Tsai IJ, Otto TD, Berriman M. 2010. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol* **11:** R41. doi: 10.1186/gb-2010-11-4-r41.
Ubeda JM, Légaré D, Raymond F, Ouameur AA, Boisvert S, Rigault P, Corbeil J, Tremblay MJ, Olivier M, Papadopoulou B, et al. 2008. Modulation of gene expression in drug resistant *Leishmania* is associated with gene amplification, gene deletion and chromosome aneuploidy. *Genome Biol* **9:** R115. doi: 10.1186/gb-2008-9-7-5115.
Vanaerschot M, Maes I, Ouakad M, Adaui V, Maes L, De Doncker S, Rijal S, Chappuis F, Dujardin JC, Decuypere S. 2010. Linking in vitro and in vivo survival of clinical *Leishmania donovani* strains. *PLoS ONE* **5:** e12211. doi: 10.1371/journal.pone.0012211.
Van Meirvenne N, Janssens PG, Magnus E. 1975. Antigenic variation in syringe passaged populations of *Trypanosoma* (Trypanozoon) *brucei*. 1. Rationalization of the experimental approach. *Ann Soc Belg Med Trop* **55:** 1–23.
Vasudevan G, Ullman B, Landfear SM. 2001. Point mutations in a nucleoside transporter gene from *Leishmania donovani* confer drug resistance and alter substrate selectivity. *Proc Natl Acad Sci* **98:** 6092–6097.
Victoir K, Dujardin JC. 2002. How to succeed in parasitic life without sex? Asking *Leishmania*. *Trends Parasitol* **18:** 81–85.
Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* **7:** 256–276.
Wiese M. 2007. *Leishmania* MAP kinases–familiar proteins in an unusual context. *Int J Parasitol* **37:** 1053–1062.
World Health Organization. 2010. *Control of the leishmaniases: Report of a meeting of the WHO Expert Committee on the Control of Leishmaniasis, Geneva, 22–26 March 2010*. WHO Technical Report Series, no. 949. WHO, Geneva, Switzerland.
Wright S. 1951. The genetical structure of populations. *Ann Eugen* **15:** 323–354.
Wu Y, El Fakhry Y, Sereno D, Tamar S, Papadopoulou B. 2000. A new developmentally regulated gene family in *Leishmania* amastigotes encoding a homolog of amastin surface proteins. *Mol Biochem Parasitol* **110:** 345–357.
Wyllie S, Cunningham ML, Fairlamb AH. 2004. Dual action of antimonial drugs on thiol-redox metabolism in the human pathogen *Leishmania donovani*. *J Biol Chem* **279:** 39925–39932.
Wyllie S, Vickers TJ, Fairlamb AH. 2008. Roles of trypanothione S-transferase and tryparedoxin peroxidase in resistance to antimonials. *Antimicrob Agents Chemother* **524:** 1359–1365.
Wyllie S, Mandal G, Singh N, Sundar S, Fairlamb AH, Chatterjee M. 2010. Elevated levels of tryparedoxin peroxidase in antimony unresponsive *Leishmania donovani* field isolates. *Mol Biochem Parasitol* **173:** 162–164.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13:** 555–556.
Yang Z. 2002. Inference of selection from multiple species alignments. *Curr Opin Genet Dev* **12:** 688–694.
Yang Z. 2007. PAML 4: a program package for phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24:** 1586–1591.
Zhang WW, Matlashewski G. 2001. Characterization of the A2-A2rel gene cluster in *Leishmania donovani*: involvement of A2 in visceralization during infection. *Mol Microbiol* **39:** 935–948.
Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. *J Comput Biol* **7:** 203–214.
Zhang WW, Mendez S, Ghosh A, Myler P, Ivens A, Clos J, Sacks DL, Matlashewski G. 2003. Comparison of the A2 gene locus in *Leishmania donovani* and *Leishmania major* and its control over cutaneous infection. *J Biol Chem* **278:** 35508–35515.
Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, Yu J. 2006. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics* **44:** 259–263.

Received March 18, 2011; accepted in revised form August 23, 2011.