DNA Research **20**, 567–581, (2013) doi:10.1093/dnares/dst031
Advance Access publication on 15 July 2013

The Genome Sequence of *Leishmania (Leishmania) amazonensis*: Functional Annotation and Extended Analysis of Gene Models

Fernando Real[1,†], Ramon Oliveira Vidal[2,†], Marcelo Falsarella Carazzolle[2,3,†], Jorge Maurício Costa Mondego[4], Gustavo Gilson Lacerda Costa[3], Roberto Hirochi Herai[5], Martin Würtele[6], Lucas Miguel de Carvalho[2], Renata Carmona e Ferreira[1], Renato Arruda Mortara[1], Clara Lucia Barbiéri[1], Piotr Mieczkowski[7], José Franco da Silveira[1], Marcelo Ribeiro da Silva Briones[1], Gonçalo Amarante Guimarães Pereira[2], and Diana Bahia[1,8,*]

Departamento de Microbiologia, Imunologia e Parasitologia, Escola Paulista de Medicina, Universidade Federal de São Paulo − EPM/UNIFESP, Rua Botucatu 862, 6º andar, 04023-062 São Paulo, Brazil[1]; Laboratório Nacional de Biociências, LNBio/CNPEM, Campinas, Brazil[2]; Laboratório de Genômica e Expressão, LGE/UNICAMP, Campinas, Brazil[3]; Centro de Pesquisa e Desenvolvimento de Recursos Genéticos Vegetais, Instituto Agronômico de Campinas − IAC, Campinas, Brazil[4]; Department of Pediatrics, School of Medicine, University of California, San Diego, CA, USA[5]; Departamento de Ciência e Tecnologia, Universidade Federal de São Paulo − UNIFESP, São José dos Campos, Brazil[6]; Department of Genetics, School of Medicine, University of North Carolina, Chapel Hill, NC, USA[7] and Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais-ICB/UFMG, Minas Gerais, Brazil[8]

*To whom correspondence should be addressed: Tel. +55 11 5576-4532. Fax. +55 11 5571-1095.
E-mail: dianabahia@hotmail.com

Edited by Naotake Ogasawara
(Received 24 January 2013; accepted 17 June 2013)

Abstract

We present the sequencing and annotation of the *Leishmania (Leishmania) amazonensis* genome, an etiological agent of human cutaneous leishmaniasis in the Amazon region of Brazil. *L. (L.) amazonensis* shares features with *Leishmania (L.) mexicana* but also exhibits unique characteristics regarding geographical distribution and clinical manifestations of cutaneous lesions (e.g. borderline disseminated cutaneous leishmaniasis). Predicted genes were scored for orthologous gene families and conserved domains in comparison with other human pathogenic *Leishmania* spp. Carboxypeptidase, aminotransferase, and 3′-nucleotidase genes and ATPase, thioredoxin, and chaperone-related domains were represented more abundantly in *L. (L.) amazonensis* and *L. (L.) mexicana* species. Phylogenetic analysis revealed that these two species share groups of amastin surface proteins unique to the genus that could be related to specific features of disease outcomes and host cell interactions. Additionally, we describe a hypothetical hybrid interactome of potentially secreted *L. (L.) amazonensis* proteins and host proteins under the assumption that parasite factors mimic their mammalian counterparts. The model predicts an interaction between an *L. (L.) amazonensis* heat-shock protein and mammalian Toll-like receptor 9, which is implicated in important immune responses such as cytokine and nitric oxide production. The analysis presented here represents valuable information for future studies of leishmaniasis pathogenicity and treatment.

Key words: genome; *Leishmania amazonensis*; interactome; amastin; heat-shock protein

1. Introduction

Leishmaniases are neglected infectious diseases caused by parasites belonging to the Trypanosomatidae family and the *Leishmania* genus. Leishmaniases are prevalent in tropical countries; ∼12 million people are affected by these diseases worldwide with 350 million people at risk of infection and an estimated yearly incidence of 2 million cases.[1] *Leishmania* spp. are digenetic parasites that develop as promastigotes in the gut of phlebotominae sandflies and as intracellular amastigotes in

† The authors agree that the first three authors should be regarded as joint first authors.

the macrophages of vertebrate hosts. The Lainson and Shaw classification[2] subdivides the *Leishmania* genus into two subgenera based on the localization of promastigotes in the insect alimentary tract. The subgenus *Leishmania* comprises species limited to the midgut and foregut of the sand fly, whereas the subgenus *Viannia* includes species that develop a prolonged phase in the hindgut with later migration of flagellates to the midgut and foregut of the vector's alimentary tract. More recently, a third subgenus has been included in *Leishmania* classification, the subgenus *Sauroleishmania*, which comprises species that exclusively parasitize lizards.[3] A brief classification of *Leishmania* subgenera and species associated with the diverse array of leishmaniasis clinical manifestations (cutaneous, mucocutaneous, and visceral forms) is provided in Fig. 1A.

Leishmaniases are primarily zoonotic diseases, and a variety of mammals acts as reservoirs of *Leishmania* species. Specifically, rodents, edentates, and marsupials typically harbor cutaneous leishmaniasis, whereas wild canines and domestic dogs are the main reservoirs of zoonotic visceral leishmaniases. In human hosts, disease outcomes are determined by a combination of parasitic properties (dermotropic versus viscerotropic species) and host factors, such as genetic variability and immune responses to infection.[4,5] Among the causative species of cutaneous leishmaniases in Brazil, recent data indicate that 8% are attributed to *Leishmania (L.) amazonensis*.[6] This species can cause simple and diffuse forms of cutaneous leishmaniasis (DCL) and was implicated recently in borderline disseminated cutaneous leishmaniasis, an intermediate form of disease.[7]

Infections with *Leishmania* species belonging to the *Leishmania (L.) mexicana* complex involve the dermal infiltration of macrophages that harbor parasites in large parasitophorous vacuoles (PVs). Most *Leishmania* species including *Leishmania (L.) major*, *Leishmania (L.) donovani*, and *Leishmania (V.) braziliensis* lodge intracellularly within small membrane-bound PVs that typically contain a single parasite and undergo fission as the amastigotes divide.[8] In contrast, amastigotes of *L. (L.) amazonensis* and *L. (L.) mexicana* are housed in large numbers within spacious PVs (Fig. 1B) that fuse together.[9] These enlarged PVs may subvert host cell defenses by facilitating conditions of relatively diluted hydrolytic enzymes.[10,11] The mechanistic basis of spacious PV development remains unknown and likely is triggered by unidentified parasitic factors produced by *L. (L.) amazonensis* and other species from the *L. (L.) mexicana* complex, such as *L. (L.) mexicana* and *Leishmania (L.) pifanoi*.[12]

During the past decade, several reports have attempted to elucidate the factors used by *Leishmania* to interact with its vertebrate host and establish an infection. Like other kinetoplastids, gene expression in *Leishmania* is regulated mainly at the post-transcriptional level by RNA stability, rather than by promoters.[13] Genes are organized into polycistronic transcriptional units, and protein-encoding genes are co-transcribed by RNA polymerase II. Precursor mRNAs subsequently are trans-spliced and polyadenylated.[14–16]

The number of chromosomes has been established for several *Leishmania* species.[17–21] The molecular karyotypes of Old World *Leishmania* species [*L. (L.) infantum*, *L. (L.) donovani*, *L. (L.) major*, and *L. (S.) tarentolae*] each comprise 36 chromosomes,[17] whereas the New World species, *L. (V.) braziliensis*, and *L. (L.) mexicana* have 35 and 34 chromosomes, respectively, due to fusion events involving 2–4 chromosomes.[18,19]

The genomes of two Old World *Leishmania* species, *L. (L.) major* Friedlin, and *L. (L.) infantum* JPCM5, and one New World species, *L. (V.) braziliensis* M2904, have been sequenced and annotated.[19,22] Recently, the genomes of *L. (L.) mexicana*, 16 clinical isolates of *L. (L.) donovani*, and the lizard parasite, *L. (S.) tarentolae*, were sequenced and assembled using high-throughput DNA sequencing technologies.[20,23,24] Despite evolutionary divergence within the *Leishmania* genus, *Leishmania* comparative genomics suggests a high degree of synteny.[19,20,23,24] *Leishmania* spp. from the *Leishmania* and *Viannia* subgenera exhibit highly conserved gene sequences with remarkably few genes or paralog groups that are unique to any given species. However, *L. (S.) tarentolae* lacks genes associated with the intracellular life stages of human pathogenic *Leishmania* spp.[20] On the other hand, the *L. (V.) braziliensis* genome includes features that are lacking in the genomes of Old World *Leishmania* spp., such as transposable elements and RNA interference (RNAi) machinery.[19,25,26]

To obtain a broader understanding of the pathogenesis of leishmaniasis, we sequenced the genome of the New World species, *L. (L.) amazonensis*. Using a comparative bioinformatics approach with other available *Leishmania* genomes, we searched for conserved domains and orthologous gene families among predicted proteins of *L. (L.) amazonensis*. In addition, we inferred the phylogeny of the surface glycoprotein, amastin, and generated a hybrid protein interactome to identify potential interactions between *L. (L.) amazonensis* secreted proteins and mammalian host factors.

2. Materials and methods

All the procedures employed in this study, except for phylogenetic analyses, are summarized in the workflow presented in Fig. 2. The workflow was divided into genome assembly and annotation steps, and functional and extended analyses of gene models.

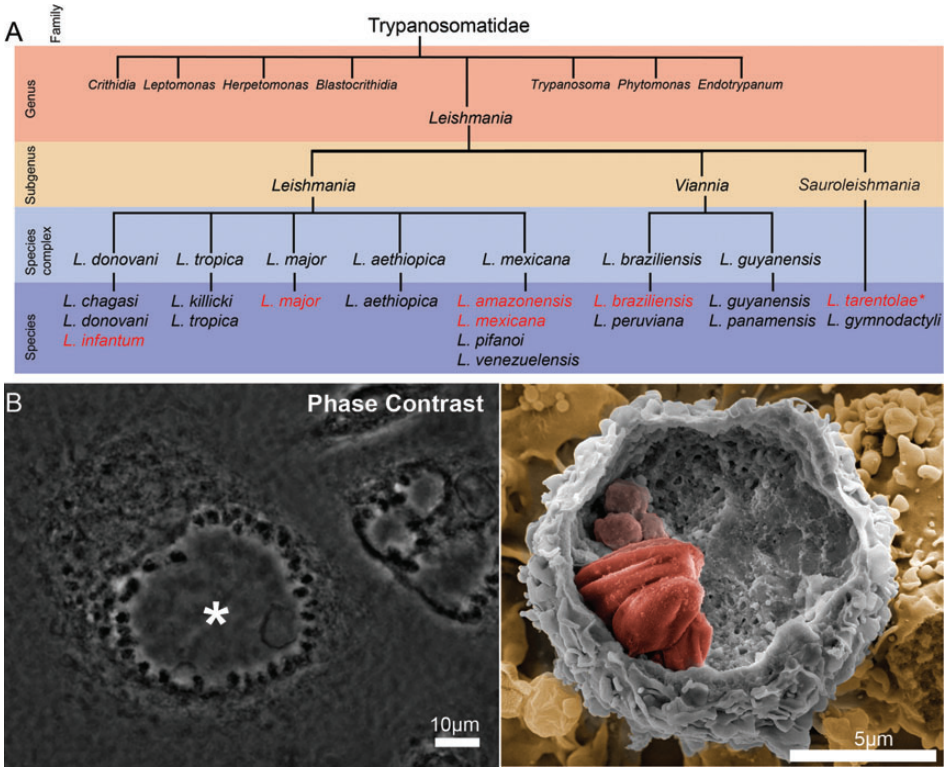

Figure 1. Overview of the *L. (L.) mexicana* complex. (A) Classification of the *Leishmania* genus, subgenus and species complex (adapted from the WHO reports and Bates, 2007). *Leishmania (L.) amazonensis* and *L. (L.) mexicana* belong to the *L. (L.) mexicana* complex, subgenus *Leishmania*, and are causative agents of New World cutaneous leishmaniasis in which diffuse or disseminated lesions are hallmarks. The genomes of the species marked in red were employed in the present comparative analyses [*L. (S.) *tarentolae* was employed only in the amastin phylogenetic study]. (B) Large parasitophorous vacuoles (PVs) of *L. (L.) amazonensis*. Phase contrast microscopy image (left) of a bone marrow-derived macrophage containing a spacious PV (asterisk) lined with rounded amastigotes. Bar = 10 μm. Field-emission scanning electron micrograph (right) of an amastigote-hosting macrophage. The fractured sample indicated that amastigote forms (in red) were contained in a spacious PV. Bar = 5 μm.

2.1. Genomic DNA extraction

Leishmania (L.) amazonensis (MHOM/BR/71973/M2269 strain) was extracted from the cutaneous lesions of a patient from Cafezal city, in the state of Pará, Brazil, in 1973. Since then, this strain has been maintained in the laboratory by inoculating hamsters and mice and by axenic culture. Parasites were cultivated in M199 culture medium supplemented with 10% fetal bovine serum. Genomic DNA was extracted from 10^9 promastigotes after the parasites were incubated in lysis buffer [50 mM Tris–HCl (pH 8.0), 62.5 mM EDTA (pH 9.0), 2.5 M LiCl, 4% Triton X-100] at 37°C for 5 min. DNA was purified using phenol–chloroform extraction (1:1 v/v) and ethanol precipitation. The resulting pellets were resuspended in 50 μl of 10 mM TE [Tris–HCl (pH 8.0), 1 mM EDTA] containing 0.6 μg/μl of RNase A (Life Technologies Corporation, USA), and were incubated at 37°C for 30 min. Genomic DNA was precipitated with 2.5 v of 100% ethanol and 0.3 M sodium acetate, centrifuged at 15 700*g* at 4°C for 15 min, and resuspended in DNase-free water.

2.2. Genome sequencing and assembly

Leishmania (L.) amazonensis DNA sequences were obtained using the whole-genome shotgun strategy[27] on a combination of 454 GS-FLX Titanium (Roche) and Solexa (Illumina, Inc.) instruments from the University of North Carolina (UNC, USA) sequencing facility. The GS-FLX sequencer generated single-end fragment reads (454 reads) with a mean length of 315 bp. The Illumina Genome Analyzer generated 76-bp paired-end fragment reads (Solexa reads) with an average insert size of 400 bp. Using an Perl script developed in-house, Illumina reads were filtered out if the average Phred quality score was lower than 20. For 454 sequences, reads comprising fewer than 100 bp or more than 500 bp (5% of each side of the normal distribution of read sizes) and reads with more than 1 unknown nucleotide (N) also were filtered out. The Solexa reads were assembled into longer scaffolds using the Velvet 0.7.56 *de novo* assembler[28] with a k-mer parameter of 43. This value was calculated using the Velvet Optimizer script (Victorian Bioinformatics Consortium, Monash University, Australia),

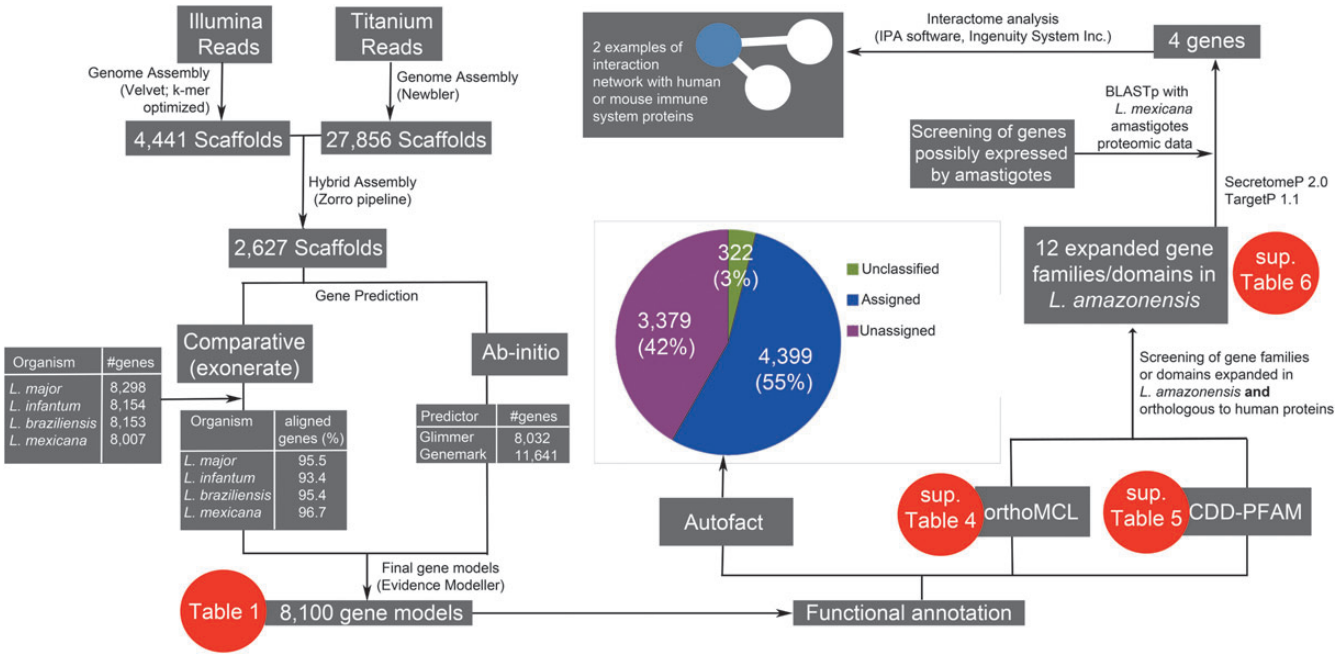

Figure 2. Bioinformatics analysis workflow used in the present study. Sequenced reads from the *L. (L.) amazonensis* genome were assembled into 2627 scaffolds and 8100 genes were predicted using comparative and *ab initio* prediction tools. The functional analysis of these predicted genes included: (i) AutoFACT functional annotation, which revealed that 45% of the predicted genes were unclassified or with unassigned function; (ii) screening for orthologous families of genes among *Leishmania* spp. (OrthoMCL); and (iii) screening for information about conserved protein domains deposited in CDD and PFAM databases. Expanded or exclusive orthologous proteins, or those conserved domains detected in the *L. (L.) amazonensis* genome were selected for interactome analysis with mammalian host proteins. This selection involved screening for possibly secreted proteins (using SecretomeP and TargetP) that also were orthologous to immune function-related proteins in humans and mice.

which tested a range of k-mers from 31 to 69. Newbler software[29] was then used to assemble the 454 reads. The N_{50} scaffold and contig lengths of the *L. (L.) amazonensis* assembly were 22 275 and 17 272 bp, respectively. Solexa scaffolds and 454 contigs were combined by the Zorro assembler (Laboratório de Genômica e Expressão, UNICAMP, Brasil, http://lge.ibi.unicamp.br/zorro/downloads/Zorro2.2/DOC) to create the final hybrid genome assembly. The hybrid assembly was performed by combining the 4441 Illumina scaffolds and the 27 856 contigs from the 454 reads. Briefly, the Zorro pipeline consisted of (i) masking repeat regions in the contigs, (ii) detecting overlaps, (iii) unmasking repeat regions, and (iv) assembling hybrid contigs. In Phase 1, the repeat regions were determined by counting the occurrences of k-mers in the 454 reads. The assembler then masked k-mers in the contigs that occurred at high frequency. The absence of repeats enabled accurate overlap detection in Phase 2. The contigs were then unmasked in Phase 3, and the correct hybrid assembly was obtained in Phase 4 by merging all overlapping contigs into hybrid contigs. Overlap detection and consensus generation were performed using the Minimus package.[30] The hybrid contigs were ordered and oriented with the Bambus program[31] using paired-end information and manual verification, which yielded 2627 scaffolds.

2.3. *Gene identification and annotation*

2.3.1. Alignment of Leishmania *spp. proteins with the* L. (L.) amazonensis *genome* Predicted proteins for *L. (V.) braziliensis* (8153 proteins), *L. (L.) infantum* (8154 proteins), *L. (L.) major* (8298 proteins), and *L. (L.) mexicana* (8007 proteins) were mapped onto the *L. (L.) amazonensis* genome sequence using the Exonerate program (v. 2.2.0),[32] which performs genomic searches and spliced alignments in a single run.

2.3.2. Ab initio prediction of gene models Glimmer, v. 3.02 [33] and Genemark.hmm, v. 3.3 [34] programs were used to perform *ab initio* gene predictions. The Glimmer long-orfs program was trained on non-overlapping open reading frames (ORFs) exceeding 200 bp in length. Genemark.hmm was executed in self-training mode by considering ORFs of at least 200 bp in length.

2.3.3. Combined gene models The above resources were used to automatically create *L. (L.) amazonensis* gene models using EVidenceModeler (EVM, v. r03062010) software.[35] For individual outputs generated by Exonerate, Glimmer and Genemark.hmm were given values reflecting our data confidence to define gene structure. We considered the following confidence values: Glimmer = 3, Genemark = 3,

nucleotide-to-protein match = 6. The gene model was considered reliable only if it was predicted by all *ab initio* software or if it had similarity with one of the compared species. A final set of predicted gene models was then selected and manually inspected. Incomplete genes or genes overlapping gap regions were inspected with input from blastx alignments against protein databases and scaffold edges.

2.3.4. Automatic annotation and analysis of protein domains The automatic annotation software, AutoFACT (v. 3.4),[36] was used for functional annotation of gene models. AutoFACT enables transitive annotation based on sequence similarity searches in several databases. We used the blastp algorithm (*e*-value 1×10^{-5}) to model gene alignments against the following protein databases: non-redundant proteins (NCBI RefSeq database, downloaded 8/6/2010), Swiss-Prot (only manually curated proteins, downloaded 4/15/2010),[37] UniRef90, and UniRef100 (UniProt databases of clustered protein families, downloaded 4/15/2010), CDD (conserved protein domains, downloaded 4/15/2010),[38] PFAM (protein domains, downloaded 4/15/2010),[39] and KEGG (metabolic pathways, downloaded 4/15/2010).[40] We set AutoFACT to consider the following order of importance for annotation: UniRef100, UniRef90, KEGG, non-redundant proteins, and CDD. Data from CDD–PFAM analyses were extracted from AutoFACT and were used for comparative analyses of CDD–PFAM protein domains among *L. (L.) amazonensis*, *L. (V.) braziliensis*, *L. (L.) infantum*, *L. (L.) major*, and *L. (L.) mexicana*. These data also were evaluated using reverse PSI-BLAST (RPS-BLAST, *e*-value cutoff of 1×10^{-5}).

2.3.5. Orthologous gene analysis A data set composed of all the *Leishmania* spp. gene models was created and compared all-against-all using blastp (*e*-value cutoff of 1×10^{-5}). The results were submitted to OrthoMCL (v. 1.4) software,[41] which clustered the proteins into orthologous and paralogous families. We applied the default software parameters, including an inflation index of 1.5. The inflation index regulates the cluster tightness (granularity) associated with sensitivity and selectivity.[42] Clusters of proteins that presented bidirectional similarities between at least two *Leishmania* species were considered orthologs.

2.3.6. Pseudogene identification *Leishmania* and *Trypanosoma* spp. proteins were aligned against the *L. (L.) amazonensis* genome using blastx with an *e*-value threshold of 1×10^{-5}. The coordinates of the first hit alignment for all proteins were converted to GFF file format using an in-house Perl script. The BEDTools package[43] was used to identify *L. (L.) amazonensis* regions exceeding 200 bp that showed similarity with *Leishmania* or *Trypanosoma* spp. proteins without overlapping with the gene model predictions. These regions were compared against the NCBI non-redundant database using the blastx program and manual annotation.

2.3.7. Calculation of the codon adaptation index The CodonW v. 1.4.4 software (http://bioweb.pasteur.fr/seqanal/interfaces/codonw.html) was employed to calculate the codon usage indices of each *L. (L.) amazonensis* predicted gene. The codon adaptation index (CAI) estimates the extent of bias toward codons known to be preferred in highly expressed genes.[44] This index ranges from 0 to 1.0 with higher values indicating stronger codon usage bias and a higher expression level. The frequency of codon usage in highly expressed genes defines the relative fitness values for each synonymous codon. These values were calculated from the relative synonymous codon usage rather than from the raw codon usage and therefore were essentially independent of amino acid composition. Because fitness values are highly species specific, we first identified a set of highly expressed genes in *L. (L.) amazonensis*. This set was input into the calculation of the CAI. The effective number of codons (N_c) also was used to quantify the codon usage bias of each gene. The N_c ranges from 20 for a gene with extreme bias (using only 1 codon per amino acid) to 61 for a gene with no bias (using synonymous codons equally).[45,46] Sequences for which N_c values are less than 30 were considered highly expressed, whereas sequences with N_c values exceeding 55 were considered poorly expressed genes.

2.4. Phylogeny of amastin surface proteins

A phylogeny of *Leishmania* amastin proteins was built from a set of *L. (L.) amazonensis* amastins predicted by our assembled genome and from all predicted amastin proteins in the *T. cruzi*, *T. brucei*, *L. (L.) major*, *L. (L.) infantum*, *L. (V.) braziliensis*, *L. (L.) mexicana*, and *L. (S.) tarentolae* genomes. These amastin sequences [except *L. (L.) amazonensis* amastins] were extracted from TriTrypDB[47] (accessed 6/28/2012) by searching for 'amastin' or 'amastin-like' entries.

The 181 amastin/amastin-like protein sequences encoded by *L. (L.) braziliensis*, *L. (L.) infantum*, *L. (L.) major*, *L. (L.) mexicana*, and *L. (S.) tarentolae* and 24 amastin/amastin-like protein sequences from *L. (L.) amazonensis* were aligned using Geneious software (v. 5.6.3),[48] in which an embedded MUSCLE software was applied.[49] The alignment was performed using default parameters. The phylogenetic tree was inferred by Bayesian methods using MrBayes v. 3.1.2[50] with tree parameter optimization during the generations. A Bayesian tree was inferred based on 1×10^7 generations with a burn-in value of 75 000. Data were saved every 100 generations and were run in 4 chains

during 2 runs. The Whelan and Goldman (WAG) substitution matrix was used for the protein alignment.[51]

2.5. *Interactome of mammalian proteins and* Leishmania *secreted proteins*

From the CDD and OrthoMCL analyses, protein families that were found to be expanded in *L. (L.) amazonensis* [i.e. more gene/domain copies when compared with *L. (V.) braziliensis*, *L. (L.) infantum*, and *L. (L.) major*] were also screened for orthologous in human protein databases. To identify human proteins that are orthologous to *L. (L.) amazonensis* proteins, we performed a blastp search (*e*-value cutoff of 1×10^{-10}) against the human protein databases using the parasite's expanded proteins as query. These selected gene models were then analyzed using TargetP 1.1 and SecretomeP 2.0 prediction tools.[52,53] Our objective was to identify proteins that could be secreted to the extracellular compartment or exposed for interaction with host cell components. We considered as putative secreted proteins those gene products for which TargetP returned a signal peptide value exceeding 0.8 and for which other values, such as the mitochondrial targeting peptide and the chloroplast transit peptide, were below 0.2. *L. (L.) amazonensis* proteins implicated in non-classical secretion or ectodomain shedding were predicted using SecretomeP v. 2.0 with the recommended threshold of 0.5.

Considering that amastigote is the developmental form which maintains a durable relationship with the mammalian host cell, an additional step was included in the screening. Using the screened gene models above mentioned, we performed a blastp analysis (*e*-value cutoff of 1×10^{-5}) against *L. (L.) mexicana* proteins expressed by amastigotes and identified in proteomic databases.[54] *L. (L.) amazonensis* gene models that attend to these four criteria were considered for hybrid interactomes: (i) expanded or exclusive in *L. (L.) amazonensis*; (ii) orthologous to human proteins; (iii) possibly secreted (start codon was manually verified); and (iv) possibly expressed by amastigotes.

Some of the screened gene products could allow for interactions and interferences with native mammalian interactomes. Interactome networks for parasite-secreted proteins and for human and mouse proteins were created using IPA software (Ingenuity Systems, Redwood City, CA, USA) which was configured to build interactomes considering only those proteins expressed by cells of the immune system.

3. Results and discussion

3.1. *Annotation of* Leishmania (L.) amazonensis *gene models*

Approximately 37 million 76-bp paired-end reads (average insert size, 400 bp) were assembled into 4411 scaffolds (coverage, $\sim 94\times$) using Velvet *de novo* assembler software. The 454 reads (179 112 reads; average read length, 312 bp) were assembled using Newbler software into 27 856 contigs (coverage, $\sim 2\times$. The final assembly was performed using Zorro, which combined scaffolds from Solexa and contigs from 454 sequencing to generate 2627 scaffolds ($\geq$1000 bp in length) that specified an *(L.) amazonensis* genome size of 29.6 Mb.

Ab initio gene prediction was conducted using Glimmer and Genemark.hmm programs, which gave 8032 and 11 641 gene models (gene lengths $\geq$150 nt), respectively. Comparative gene predictions consisted of proteins from *L. (V.) braziliensis*, *L. (L.) infantum*, *L. (L.) mexicana*, and *L. (L.) major* aligned against the final *L. (L.) amazonensis* genome assembly. The final set of 8168 gene models was created using a combination of *ab initio* and comparative gene model analyses, EVM software to identify consensus gene structures supported by these two approaches, and manual annotation. Incomplete genes and genes overlapping gap regions were discarded from further analyses. Following manual annotation, 8100 gene models remained with an average length of 1793 bp; this is consistent with other sequenced *Leishmania* species. These data are summarized and compared with other *Leishmania* genomes in Table 1.

The final set of gene models were annotated automatically by comparing them against protein databases (blastp) and summarizing with AutoFACT software. The results indicated that 55% of the gene models showed high sequence similarity (*e*-value 1×10^{-10}) to functionally annotated proteins, whereas 42% of the models were similar to unassigned proteins (i.e. proteins with no functional annotation). The remaining 3% yielded no hits with any databases and were considered unclassified (Fig. 2).

In agreement with previous reports,[19,20,22–24] our genome sequence analysis indicated that more than 90% of the 8100 *L. (L.) amazonensis* genes are shared with other human pathogenic *Leishmania* spp. with little variation in orthologous gene content. Despite varying clinical manifestations and features of lesions, *Leishmania* spp. harbor a conserved genomic core encoding functions ranging from fundamental biological processes to complex host–parasite interaction networks.

We performed a comparative annotation of tRNAs from the *L. (L.) amazonensis*, *L. (L.) infantum*, *L. (V.) braziliensis*, and *L. (L.) major* genomes. Using tRNAscan-SE software,[55] the annotation yielded very similar numbers among the studied *Leishmania* species (Supplementary data, Table S1). Given the number of tRNAs for each codon in *L. (L.) amazonensis*, we calculated the CAI for all annotated gene models (Supplementary data, Table S2). The index measures

Table 1. Summary of the information obtained from the genome sequences of *Leishmania* spp.

	L. (L.) amazonensis	*L. (L.) mexicana*	*L. (L.) major*	*L. (L.) infantum*	*L. (V.) braziliensis*	*L. (L.) donovani*	*L. (S.) tarentolae*
Contigs	3199	35	36	37	1041	2154	N/A
Genome length (Mb)	29.6	32.1	32.8	32.1	33.7	32.4	30.4
chromosomes	34*	34	36	36	35	36	36
Number of predicted genes	8100	8007	8298	8216	8153	8252	8201
Gene density (genes/Mb)	273	256	260	252	228	254	270
G + C content (%)	58.5	60.5	59.7	59.3	57.8	>60	57.2
CDS G + C content (%)	61	61.23	62.5	62.45	60.38	61	58.4
References	Current study	Rogers *et al.*[24]	Ivens *et al.*[22]	Peacock *et al.*[19]	Peacock *et al.*[19]	Downing *et al.*[23]	Raymond *et al.*[20]

The number of *L. (L.) amazonensis* chromosomes (*) was inferred by mapping against *L. (L.) mexicana* chromosomes using the software SSAHA2 with all the *L. (L.) amazonensis* reads generated by Illumina.
N/A, not available.

the codon usage bias presented by highly expressed genes and can be comparatively employed to score native genes (higher CAI, more adapted to the pattern of codon usage) and possibly foreign or transferred genes (lower CAI, less adapted to the pattern of codon usage).[56] Ranging from 0 (non-expressing gene, likely pseudogene) to 1 (highly expressed gene), we found that *L. (L.) amazonensis* predicted genes have a mean CAI value of 0.49 and median of 0.48 (Supplementary data, Fig. S1A). These data will be useful for future analyses of recent events of horizontal gene transfer in *Leishmania*.

Additionally, we scored regions in the *L. (L.) amazonensis* genome that showed similarities with predicted genes in other *Leishmania* and trypanosomatid species but were not identified as ORFs due to premature stop codons or frame shifts. This approach yielded 36 genomic regions corresponding to potential pseudogenes (Supplementary data, Table S3). One *L. (L.) amazonensis* pseudogene identified in our analysis was a fragment of argonaute 1 (AGO1), which is involved in the RNAi machinery. In the genome of *L. (L.) amazonensis*, we did not detect the known trypanosomatid argonaute and dicer variants, AGO1, DCL1 or DCL2. In addition, proteins containing two RNAse III domains (characteristic of dicer) or PAZ and Piwi domains (characteristic of argonaute) were not detected in this analysis. Our results suggest that RNAi pathways are absent in *L. (L.) amazonensis*, corroborating the hypothesis that RNAi via dicer and argonaute has been lost from the *Leishmania* subgenus following its divergence from the *Viannia* subgenus.[26]

Leishmania generally is considered to be a diploid organism because it carries two copies of most of its homologous chromosomes.[19,22,24,57,58] However, there is increasing evidence suggesting that aneuploidy can occur in *Leishmania* species.[23,24,58,59] Chromosome copy numbers can vary considerably among strains and species from different geographic regions, even among recent isolates.[23,24] We mapped the *L. (L.) amazonensis* reads generated by Illumina against the *L. (L.) mexicana* chromosomes using the software SSAHA2.[60] The median of coverage along each chromosome indicated a probable extra copy of chromosomes 7 and 26 and 3 extra copies of chromosome 30 in *L. (L.) amazonensis* (Supplementary data, Fig. S1B). In *L. (L.) amazonensis*, the exact number of chromosomes has not been defined. Preliminary studies using pulsed field gel electrophoresis have reported that the *L. (L.) amazonensis* karyotype consists of 25 chromosomal bands ranging in size from 0.2 to 2.2 Mb.[61] Some bands exhibited variable ethidium bromide staining intensities possibly due to co-migration of chromosomes of similar sizes. Further studies will be needed to define the number of chromosomes and ploidy in *L. (L.) amazonensis*.

3.2. *Functional analysis of gene models*

In our functional analysis of gene models, we focused on the common factors, rather than the species-specific factors, predicted in *L. (L.) amazonensis* and *L. (L.) mexicana* genomes and not predicted in the genomes of the other species. We chose one genome data set for each disease outcome to compare with *L. (L.) amazonensis* and *L. (L.) mexicana*. Specifically, we chose leishmaniasis causative agents representative of cutaneous [*L. (L.) major*], mucocutaneous [*L. (V.) braziliensis*], and visceral [*L. (L.) infantum*] infections.

We searched for *L. (L.) amazonensis* genes that could be expanded or contracted in terms of gene or domain copies compared with the other species, particularly *L. (L.) mexicana*. We present a discussion of

some of these expanded genes/domains that could participate in: (i) parasite tropism in host organisms via adhesion molecules or amastin surface proteins; (ii) the development of large PVs by lipid synthesis; and (iii) intracellular establishment by enzymes related to nutritional acquisition and resistance to host intracellular defenses, such as oxidative burst.

3.2.1. Expanded and contracted orthologous gene families The OrthoMCL software was applied to genome data sets from *L. (L.) amazonensis*, *L. (L.) mexicana*, *L. (L.) infantum, L. (V.) braziliensis*, and *L. (L.) major* to identify *Leishmania* orthologous gene families. We identified 7826 orthologous gene families [7488 orthologous gene families containing $\geq$1 *L. (L.) amazonensis* protein] with 468 families comprising at least 7 members. Most (6784) of the orthologous gene families were shared by other *Leishmania* spp. Eight families were found only in *L. (L.) amazonensis* and *L. (L.) mexicana* and 23 families were unique to *L. (L.) amazonensis* (Fig. 3A, Supplementary data, Table S4). The five largest families identified across human pathogenic *Leishmania* spp. using OrthoMCL also were identified in *L. (L.) amazonensis*: dyneins (Family 1), glycoprotein GP63 (Family 3, leishmanolysins), histone H4 (Family 4), ABC transporters (Family 8), and amastin proteins (Family 14; Supplementary data, Table S4). We identified several families associated with 2-fold or more gene copies in *L. (L.) amazonensis* than in any of the other human pathogenic species. These families include aminotransferases (family 256), 60S ribosomal protein L37 (family 216), and hypothetical proteins (families 323, 5508, and 7732). Among the families with at least 2-fold fewer gene copies in *L. (L.) amazonensis* we highlighted the GP63 gene family. The list of contracted gene families requires further investigation because some of them could be present in unassembled regions.

Comparative genome analyses of *L. (L.) amazonensis* and *L. (L.) mexicana* indicated an expansion in the gene family encoding a class-IV branched-chain amino acid aminotransferase (OrthoMCL family 256) that consisted of 270–415 amino acid residues and shared few regions of sequence similarity.[62] Branched-chain aminotransferases catalyze the synthesis of leucine, isoleucine, and valine, and may be used to fulfill the parasite's nutritional requirements. They also may be involved in parasite sterol and prenol lipid synthesis because leucine is the main isoprenoid precursor for *L. (L.) mexicana* promastigotes and amastigotes.[63,64]

A 3′-nucleotidase/nuclease (OrthoMCL family 7761) gene was predicted as being exclusive to the *L. (L.) amazonensis* and *L. (L.) mexicana* genomes. This gene encodes an enzyme responsible for nucleic acid hydrolysis that was found to be dramatically upregulated on the cell surface of the trypanosomatid, *Chritidia luciliae*, under purine starvation conditions.[65] An ecto-3′-nucleotidase/nuclease was detected experimentally in *L. (L.) amazonensis*; this component has important implications for parasite nutrition, adhesion to host cells, and infectivity.[66]

Substantial differences in the gene copy numbers between certain *Leishmania* spp. may account for the observed phenotypic variability in terms of pathogenesis and virulence. Several genome features could contribute to quantitative variation in gene copies among *Leishmania* spp. The expansion and contraction of genes in tandem arrays could result in up- or down-regulation of gene expression associated with copy-number variation. In addition, extensive variation in aneuploidy frequencies within parasite populations has been reported for several *Leishmania* spp. and for different *Leishmania* isolates within the same species.[23,24,59]

3.2.2. Expanded and contracted conserved domains Another comparison between *Leishmania* genomes was based on the identification of predicted protein domains in CDD–PFAM databanks. The result of CDD–PFAM analysis, included in the AutoFACT annotation, was retrieved and manually evaluated. In total, 2509 protein domains were identified; 2186 of these domains were detected in at least one *L. (L.) amazonensis* protein (Supplementary data, Table S5). Most (1881) of the identified domains were shared by other *Leishmania* spp.; 20 domains were found exclusive to *L. (L.) mexicana* complex [*L. (L.) amazonensis*, and *L. (L.) mexicana*], and 26 domains were unique to *L. (L.) amazonensis* (Fig. 3B, Supplementary data, Table S5). The expanded and contracted protein domains in *L. (L.) amazonensis* were evaluated by the same criteria applied in OrthoMCL (i.e. variations in gene copies when compared with other *Leishmania* genomes). Among the most prevalent domains were the heat-shock protein (HSP)70 chaperone (CDD: 143803) and the vacuolar protein sorting-associated protein MRS6 (CDD: 34648). The following protein domains were detected only in *L. (L.) amazonensis* and *L. (L.) mexicana*: thioredoxin domain (CDD: 32932), tat-binding protein 1 (CDD: 148614), sterile alpha motif (SAM) domain of bicaudal C homolog protein 1 (BCC1, CDD: 188919), hydrolase (CDD: 188206), and ATPase (CDD: 190944).

As examples of contracted protein domains identified in *L. (L.) amazonensis*, we highlight: amastins (CDD: 140228), UDP-GlcNAc-dependent glycosyltransferase (CDD: 140237), leishmanolysin peptidase M8 (CDD: 189994), cathepsin (CDD: 185513), tryparedoxin peroxidase (CDD: 140280), non-long terminal repeat reverse transcriptases (RTs-nLTR, CDD: 73156), rim ABC transporters (CDD: 185513), adenylyl/guanylyl cyclase (CDD: 128359), and paraflagellar rod protein (CDD: 140353).

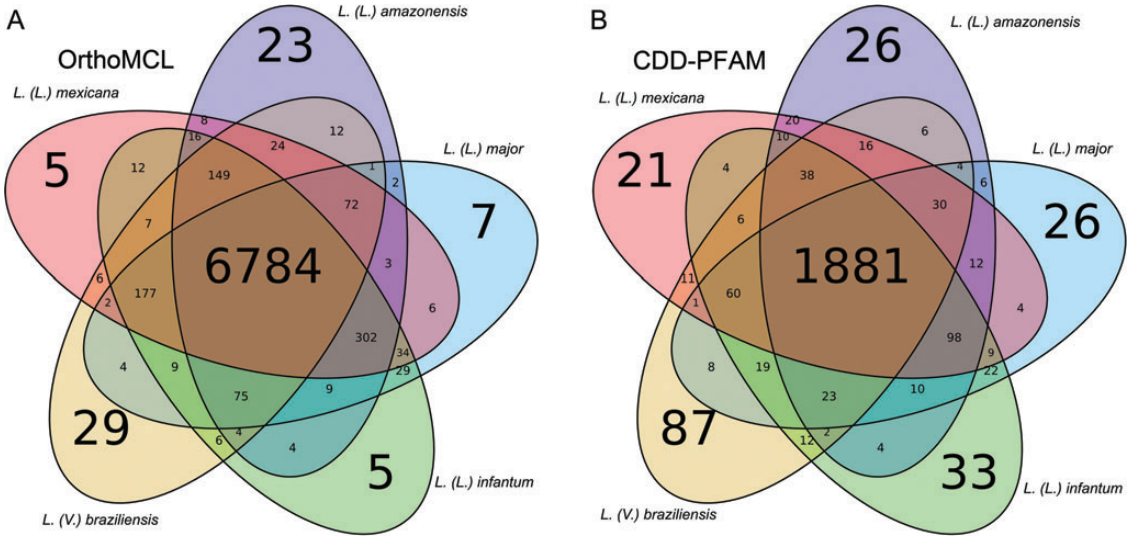

Figure 3. Diagrammatic representation of (A) species-specific orthologous gene families (OrthoMCL analysis) and (B) conserved domains (CDD−PFAM analysis). A core of 6784 orthologous families and 1881 domains was conserved in all studied *Leishmania* species [*L. (L.) amazonensis*, *L. (L.) mexicana*, *L. (L.) major*, *L. (L.) infantum*, and *L. (V.) braziliensis*]. We detected 8 orthologous families and 20 conserved domains that were exclusive to *L. (L.) mexicana* complex. A complete list of orthologous families and conserved domains is presented in Supplementary data, Table S4 and S5, respectively.

The present study detected a thioredoxin domain unique to *L. (L.) amazonensis* and *L. (L.) mexicana* (NCBI accession COG3118). Thioredoxin functions as a hydrogen donor or disulfide reductase and is involved in the response to oxidative stress and in protein folding.[67] Reactive oxygen species can be scavenged directly by thioredoxin or by thioredoxin-related dehydrogenases. Parasitic thiol and dithiol proteins may buffer the redox environment of PVs; this could account for the resistance of *L. (L.) amazonensis* to nitric oxide (NO) production in interferon (IFN)-γ-activated macrophages.[68–70]

3.3. *Extended gene model analysis*

3.3.1. *Amastin phylogeny suggests specialized amastins in the* Leishmania (L.) mexicana *complex*

Previous *Leishmania* phylogenetic analyses, based on comparisons of isoenzymes, DNA sequences, and HSP profiles among species agreed with the adopted Linnean classification and with complexes proposed by Lainson and Shaw in 1987.[4,71,72] The phylogenies from these studies indicated that *L. (L.) amazonensis* has an evolutionary proximity to *L. (L.) mexicana*, a finding that was interpreted as these species comprising a monophyletic clade. These studies also indicate that parasites responsible for cutaneous/mucocutaneous lesions could be as different and divergent from one another as they are from the parasites that cause visceral leishmaniasis. For instance, the phylogenetic distance between the cutaneous-associated species, *L. (L.) amazonensis* and *L. (L.) major*, is similar to the distance between *L. (L.) amazonensis*/*L. (L.) major* and *L. (L.) donovani*, which causes visceral leishmaniasis. Therefore, the same disease outcomes in *Leishmania* mammalian hosts can result from a variety of evasive strategies and factors distinctively featured by *Leishmania* spp.

Amastin belongs to a multi-gene family in *Leishmania* that encodes small surface proteins of ∼200 amino acids. Several members of the amastin gene family are dispersed throughout the genomes of all *Leishmania* species and exhibit various expression patterns.[73] Phylogenetic analysis of trypanosomatid amastins defined four subfamilies of amastin (α, β, γ, and δ) with distinct genomic organizations and expression patterns during the cell cycles of *T. cruzi* and *Leishmania* spp.[74] δ-amastins comprise the largest and most diverse amastin subfamily. In *T. cruzi*, δ-amastin expression was associated with parasite infectivity to host cells.[75] In *Leishmania*, the amastin N-terminal signature peptides are among the most immunogenic of all leishmanial surface antigens in mice[76] and generate strong immune responses in humans with visceral leishmaniasis.[77] DNA microarray analyses have implicated amastin in the intracellular survival of the parasite.[78] Amastin gene expression was detected predominantly in amastigotes of several *L. (L.) donovani* strains isolated from patients with visceral and post-kala-azar dermal leishmaniasis.[79] The roles of amastin genes in parasite homeostasis and growth inside acidic PVs also have been addressed.[73,80] As transmembrane proteins, amastins could contribute to proton or ion trafficking across the membrane to adjust cytoplasmic pH under the harsh conditions of

the phagolysosome. As a surface epitope, amastin may be recognized by opsonizing host IgG antibodies and could promote parasite uptake by host macrophages (via Fc receptors) and subsequent release of interleukin (IL)-10.[81] We speculate that amastins could be involved in certain peculiar characteristics of *L. (L.) amazonensis*, such as its propensity to induce DCL, its development inside spacious PVs, and its resistance to the highly oxidative phagolysosomal environment in host cells.

The TriTrypDB provides a set of annotated amastin and amastin-like surface proteins in the *Leishmania* and *Trypanosoma* genomes. Our search of the TriTrypDB yielded 181 annotated genes encoding amastin or amastin-like proteins in *L. (L.) braziliensis*, *L. (L.) infantum*, *L. (L.) major*, *L. (L.) mexicana*, and *L. (S.) tarentolae*. The *L. (L.) amazonensis* genome presented in this study identified 12 orthologous groups annotated as amastin or amastin-like proteins (families 14, 19, 20, 3539, 3935, 5852, 6119, 6120, 6543, 7556, 7771, and 7778). All families corresponded to 1 representative gene model, except family 19 (3 gene models were associated with this orthologous group). Thus, our OrthoMCL analysis predicted 14 amastin/amastin-like proteins in the *L. (L.) amazonensis* genome. All of these proteins also were identified in the CDD−PFAM analysis. The CDD−PFAM list also identified gene models containing amastin domains that were not identified in the OrthoMCL analysis. We gathered all gene models identified as amastin/amastin-like proteins (OrthoMCL) or containing amastin domains (CDD−PFAM) and built a list of 24 *L. (L.) amazonensis* predicted amastin/amastin-like surface proteins to perform our phylogenetic analysis.

Alignment of the *Leishmania*-annotated amastins with 24 amastin/amastin-like proteins identified in the *L. (L.) amazonensis* genome (OrthoMCL and CDD−PFAM combined scoring) allowed us to build an amastin phylogenetic tree (Fig. 4). By placing the phylogenetic root halfway between the two most divergent/distant amastins (midpoint rooting) we identified clades composed of species-related amastins. Early branching clades could represent a class of amastin surface proteins conserved in *Leishmania* prior to its radiation (Fig. 4, blue branches). These *Leishmania* pre-speciation amastins are gathered in α, β, and γ subfamily clades.[74] The presence of species-specific clades of δ-amastins at terminal tree branches (Fig. 4, red, green, yellow, and purple branches) suggests that several amastin surface proteins appeared because of environmental selective pressures or pathogen speciation. At least in part, this could be associated with the diverse leishmaniasis outcomes of different *Leishmania* species. It is possible to identify amastin subfamilies unique to *L. (L.) major*, *L. (V.) braziliensis*, and *L. (L.) infantum*, and three groups in which *L. (L.) amazonensis* and *L. (L.) mexicana* amastins represent a distinct clade of amastins (Fig. 4, red branches). These amastins could play a role in the unusual housing of these parasites within spacious PVs of infected macrophages.

3.3.2. Secreted Leishmania *HSPs could interfere with native host interactomes* One of the most striking features of the *L. (L.) mexicana* complex is the development of giant PVs in infected macrophages that harbor amastigotes. We speculate that the formation of large PVs may be related to factors secreted by the parasite, the subversion of host native vesicular trafficking, and potentially the production/incorporation of parasitic components into PV membranes. *L. (L.) amazonensis* amastigotes interact with the internal membranes of PVs via their posterior poles.[8] The posterior pole behaves like an adhesion site between the parasite and PV membranes, although no adhesion factors or junction components have been identified to date. De Souza Leão *et al.*[82] suggested that in *L. (L.) amazonensis*-infected macrophages, the internalization and degradation of major histocompatibility complex (MHC) class II molecules by amastigotes occur through their posterior poles. This degradation could be performed by secreted components inserted into the PV membrane. The posterior pole also may be interpreted as part of a parasitic secretory pathway in which secreted proteins directly encounter the host cell cytosol, bypassing the acidic milieu of the PV. Once in the host cell cytosol, secreted factors may be transferred to the host cell nucleus and/or plasma membrane, affecting gene expression, cellular functions and metabolic processes. However, the classically described site for parasite exocytosis and endocytosis is the anterior pole where the flagellar pocket is located. The flagellar pocket faces the lumen of the PV, and most secreted *Leishmania* proteins, regardless of their association with exosomes, are expected to reach the acidic (pH 4.5−5.0) PV milieu from there.[83]

We hypothesized that *Leishmania* secreted factors could mimic mammalian factors, thus perturbing native host protein interactions. To identify possible interactions between parasitic and mammalian host factors, we constructed hybrid protein interaction networks in which human and mouse databases were compared against the list of proteins that are potentially secreted by *L. (L.) amazonensis* amastigotes. Our CDD−PFAM and OrthoMCL analyses identified nine conserved domains and three orthologous gene families that were exclusive to or expanded in *L. (L.) amazonensis* and/or *L. (L.) mexicana* and are also orthologous to human proteins (Supplementary data, Table S6). The CDD 143803, an HSP 70 domain, exists as six copies in *L. (L.) amazonensis*, four copies in *L. (L.) mexicana*, and three copies in the other species' genomes. We considered this as an expanded number of HSP70 domains

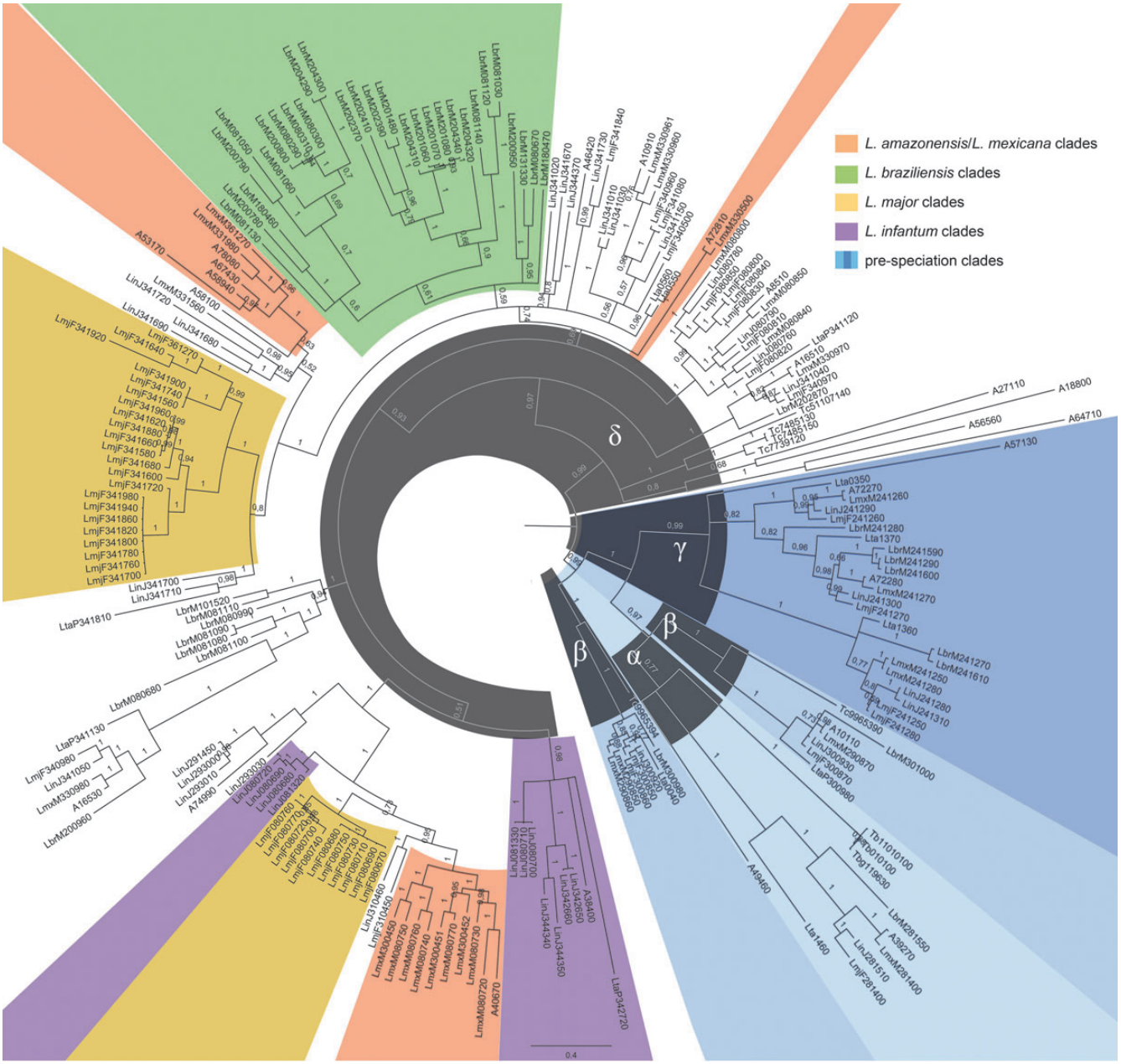

Figure 4. Bayesian consensus phylogeny of amastin surface proteins. The phylogram is represented by a consensus of 214 amastin sequences. The root was inferred using midpoint rooting. WAG was used as the substitution matrix for the protein alignment. Posterior probabilities exceeding 0.5 are shown in the branches. The tree topology suggests early branching of similar amastins shared by different species (blue). These branches were classified as *Leishmania* pre-speciation amastins, composed by α, β, and γ subfamily clades. We highlighted the terminal taxa (late branching or apomorphic) of species-specific δ-amastin clades of *L. (L.) major* (yellow), *L. (V.) braziliensis* (green), and *L. (L.) infantum* (purple). Complex-specific clades of *L. (L.) amazonensis* and *L. (L.) mexicana* amastin surface proteins are in red. The scale of the generated tree (see 0.4 bar) represents the number of substitutions per sequence position. The classification of amastin clades in subfamilies α, β, γ, and δ was based on the amastin phylogeny performed by Jackson *et al*. (2010).

in the *L. (L.) mexicana* complex. The six gene models in which these domains were identified (A42670, A6630, A68920, A73510, A30200, and A45910) were submitted to the TargetP and SecretomeP servers and gene models A30200 and A45910 were predicted for secretion. Additionally, A30200 and A45910 were similar to two *L. (L.) mexicana* proteins identified in a proteomic data analysis of *L. (L.) mexicana* amastigotes[54] (LmxM.28.2770 and LmxM.34.4710, respectively; Supplementary data, Table S6). Although these genes are not exclusive to the *L. (L.) mexicana* complex, their similarity to two products from amastigote proteomic data is suggestive that, at least in *L. (L.) amazonensis* and *L. (L.) mexicana*, these products are expressed by the intracellular form of the parasite and could be secreted within host cells.

The A30200 and A45910 gene models are candidates for the construction of hybrid interactomes, given that they present a protein domain expanded in both *L. (L.) amazonensis* and *L. (L.) mexicana* (CDD 143803), are predicted to be secreted and are similar to proteins identified in the proteome of *L. (L.) mexicana* amastigotes. The A30200 and A45910 gene models present a considerable similarity with the mammalian hypoxia up-regulated protein 1 (HYOU1) and the HSP 70 kDa protein 5 (HSPA5), respectively. We created HYOU1 and HSPA5 interactome subsets to identify host components that could be affected by the secretion of both A30200 and A45910 gene products. This analysis led us to identify toll-like receptor (TLR) 9 and IL-6 as putative targets of the *L. (L.) amazonensis* A30200 and A45910 gene products (Fig. 5A and B). The HYOU1 and HSPA5 both directly interact with TLR9, a receptor implicated in the recognition of CpG DNA motifs and present in endolysosomal compartments where it is activated by proteolytic cleavage.[84] TLR9 is preferentially expressed in the granulomas of human cutaneous leishmaniasis caused by *L. (V.) braziliensis*,[85] and TLR9-deficient mice are more susceptible to *L. (L.) major* infection.[86,87] Thus, TLR9 is implicated in the immune response against *Leishmania*. The interaction between host TLR9 and the putative secreted *L. (L.) amazonensis* HYOU1/HSPA5-mimic could block TLR9 function and favor intracellular establishment of the parasite. TLR9 also is implicated in the production of NO via NO synthase 2, tumor necrosis factor, IL-6, and IL-12B. The production of IL-6 is inhibited in dendritic cells differentiated from monocytes in the presence of *L. (L.) amazonensis*[88] and is present at low levels in the sera of Chiclero's ulcer patients infected for 3–8 months.[89] Linares *et al.*[90] reported that *in vitro* infection with *L. (L.) amazonensis* amastigotes decreases NO production by macrophages stimulated with IFN-γ plus lipopolysaccharide. Thus, although hypothetical and genome based, our proposed interactome can be used to identify components implicated in the establishment of *Leishmania* infection of mammalian host cells. Moreover, the interactome provides a model for studying *Leishmania*-secreted proteins and their influence on important effectors of the host cell immune response.

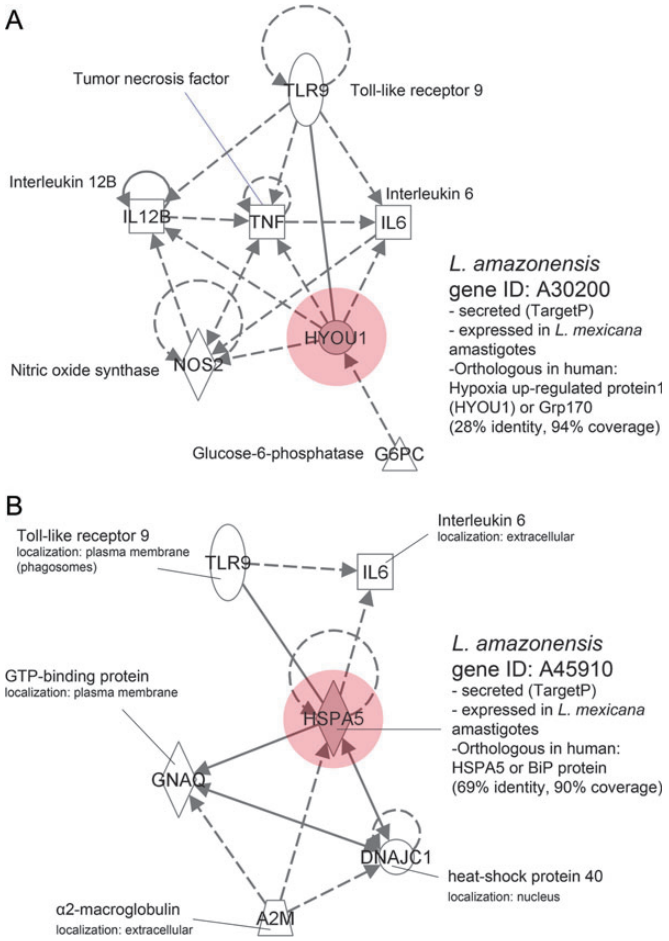

Figure 5. Interactomes of potentially secreted *L. (L.) amazonensis* [A30200 (A) and A45910 (B)] and mammalian immune cell proteins. The secreted parasite gene products are represented by red nodes in the interactome. The expression statuses of these parasite proteins during the amastigote stage were inferred using blastp with the proteomic database of *L. (L.) mexicana* amastigotes. The secreted components of *L. (L.) amazonensis* amastigotes share 28% identity and 94% coverage (A30200, A) and 69% identity and 90% coverage (A45910, B) with the mammalian HYOU1 and HSPA5 proteins, respectively. Both secreted components could directly interact with TLR9. We propose that orthologs of mammalian HYOU1 and HSPA5 are secreted by *L. (L.) amazonensis* amastigotes, interfering with host cell functions such as signaling and the production of NO and ILs. Arrows represent direct interactions and dashed arrows represent indirect interactions. The interactome was built using Ingenuity software, considering only proteins expressed in human and mouse immune cells and considering experimentally identified protein–protein interactions.

3.4. Conclusions

We present the genome of the protozoan *L. (L.) amazonensis* together with functional annotations and extended analyses focused on host–parasite interactions. We examined the genome sequences of *L. (L.) amazonensis* and *L. (L.) mexicana* for potentially expressed genes at expanded copy numbers. Confirming that a few *Leishmania* species-specific genes may exist despite striking conservation at the gene level, we report conserved domains, orthologous gene families, and amastin surface proteins unique to *L. (L.) amazonensis* and *L. (L.) mexicana*. Additionally, we propose an innovative approach to interactome analysis that emphasizes the role of parasite secreted proteins in host interaction networks.

4. Availability

The *Leishmania (Leishmania) amazonensis* Genome Database is available at the URL http://www.lge.ibi.unicamp.br/leishmania. This Whole Genome Shotgun

project was deposited at DDBJ/EMBL/GenBank under the accession APNT00000000 (SUBID SUB120161, BioProject PRJNA173202). The version described in this paper is the first version, APNT01000000. While revising this manuscript, we realized that another Brazilian group from Instituto Oswaldo Cruz (IOC)—Fiocruz, Rio de Janeiro, is sequencing the genome of *L. (L.) amazonensis*.

Acknowledgements: F.R. and D.B. would like to thank Dr Michel Rabinovitch for stimulating advice and Dr Andrew Jackson for providing amastin alignments. The authors also thank Dr Angela Kaysel Cruz and Dr Colin Bowles for kindly revising the manuscript and response to reviewers and BioMed Proofreading (http://www.biomedproofreading.com) for English editing services.

Supplementary Data: Supplementary Data are available at www.dnaresearch.oxfordjournals.org.

Funding

The authors acknowledge the financial support of Fundação de Amparo à Pesquisa do Estado de São Paulo, FAPESP (Jovem Pesquisador 07/50551-2). F.R. is recipient of a FAPESP post-doctorate fellowship (10/19335-4). D.B., J.F.S., R.A.M., M.S.B., G.A.P. are recipients of a Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq, fellowship.

References

1. WHO. 2010, Control of the Leishmaniasis. *WHO Technical Report Series*. WHO Press: Geneva.
2. Lainson, R. and Shaw, J.J. 1987, *The leishmaniases in biology and medicine. Evolution, classification and geographical distribution*.
3. Bates, P.A. 2007, Transmission of *Leishmania metacyclic* promastigotes by phlebotomine sand flies, *Int. J. Parasitol.*, **37**, 1097–106.
4. Dedet, J.P., Pratlong, F., Lanotte, G. and Ravel, C. 1999, Cutaneous leishmaniasis. The parasite, *Clin. Dermatol.*, **17**, 261–8.
5. Murray, H.W., Berman, J.D., Davies, C.R. and Saravia, N.G. 2005, Advances in leishmaniasis, *Lancet*, **366**, 1561–77.
6. Camara Coelho, L.I., Paes, M., Guerra, J.A., et al. 2011, Characterization of *Leishmania* spp. causing cutaneous leishmaniasis in Manaus, Amazonas, Brazil, *Parasitol. Res.*, **108**, 671–7.
7. Silveira, F.T., Lainson, R. and Corbett, C.E. 2005, Further observations on clinical, histopathological, and immunological features of borderline disseminated cutaneous leishmaniasis caused by *Leishmania (Leishmania) amazonensis*, *Mem Inst Oswaldo Cruz*, **100**, 525–34.
8. Real, F. and Mortara, R.A. 2012, The diverse and dynamic nature of *Leishmania parasitophorous* vacuoles studied by multidimensional imaging, *PLoS Negl. Trop. Dis.*, **6**, e1518.
9. Real, F., Pouchelet, M. and Rabinovitch, M. 2008, *Leishmania (L.) amazonensis*: fusion between parasitophorous vacuoles in infected bone-marrow derived mouse macrophages, *Exp. Parasitol.*, **119**, 15–23.
10. Alpuche-Aranda, C.M., Racoosin, E.L., Swanson, J.A. and Miller, S.I. 1994, Salmonella stimulate macrophage macropinocytosis and persist within spacious phagosomes, *J. Exp. Med.*, **179**, 601–8.
11. Real, F., Mortara, R.A. and Rabinovitch, M. 2010, Fusion between *Leishmania amazonensis* and *Leishmania major* parasitophorous vacuoles: live imaging of coinfected macrophages, *PLoS Negl. Trop. Dis.*, **4**, e905.
12. Ndjamen, B., Kang, B.H., Hatsuzawa, K. and Kima, P.E. 2010, *Leishmania parasitophorous* vacuoles interact continuously with the host cell's endoplasmic reticulum; parasitophorous vacuoles are hybrid compartments, *Cell Microbiol.*, **12**, 1480–94.
13. Clayton, C. and Shapira, M. 2007, Post-transcriptional regulation of gene expression in trypanosomes and leishmanias, *Mol. Biochem. Parasitol.*, **156**, 93–101.
14. Martinez-Calvillo, S., Yan, S., Nguyen, D., Fox, M., Stuart, K. and Myler, P.J. 2003, Transcription of *Leishmania major* Friedlin chromosome 1 initiates in both directions within a single region, *Mol. Cell*, **11**, 1291–9.
15. Haile, S. and Papadopoulou, B. 2007, Developmental regulation of gene expression in trypanosomatid parasitic protozoa, *Curr. Opin. Microbiol.*, **10**, 569–77.
16. Martinez-Calvillo, S., Vizuet-de-Rueda, J.C., Florencio-Martinez, L.E., Manning-Cela, R.G. and Figueroa-Angulo, E.E. 2010, Gene expression in trypanosomatid parasites, *J. Biomed. Biotechnol.*, **2010**, 525241.
17. Wincker, P., Ravel, C., Blaineau, C., et al. 1996, The *Leishmania* genome comprises 36 chromosomes conserved across widely divergent human pathogenic species, *Nucleic Acids Res.*, **24**, 1688–94.
18. Britto, C., Ravel, C., Bastien, P., et al. 1998, Conserved linkage groups associated with large-scale chromosomal rearrangements between Old World and New World Leishmania genomes, *Gene*, **222**, 107–17.
19. Peacock, C.S., Seeger, K., Harris, D., et al. 2007, Comparative genomic analysis of three *Leishmania* species that cause diverse human disease, *Nat. Genet.*, **39**, 839–47.
20. Raymond, F., Boisvert, S., Roy, G., et al. 2012, Genome sequencing of the lizard parasite *Leishmania tarentolae* reveals loss of genes associated to the intracellular stage of human pathogenic species, *Nucleic Acids Res.*, **40**, 1131–47.
21. Rovai, L., Tripp, C., Stuart, K. and Simpson, L. 1992, Recurrent polymorphisms in small chromosomes of *Leishmania tarentolae* after nutrient stress or subcloning, *Mol. Biochem. Parasitol.*, **50**, 115–25.
22. Ivens, A.C., Peacock, C.S., Worthey, E.A., et al. 2005, The genome of the kinetoplastid parasite, *Leishmania major*, *Science*, **309**, 436–42.
23. Downing, T., Imamura, H., Decuypere, S., et al. 2011, Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population

structure and mechanisms of drug resistance, *Genome Res.*, **21**, 2143–56.
24. Rogers, M.B., Hilley, J.D., Dickens, N.J., et al. 2011, Chromosome and gene copy number variation allow major structural change between species and strains of *Leishmania*, *Genome Res.*, **21**, 2129–42.
25. Smith, D.F., Peacock, C.S. and Cruz, A.K. 2007, Comparative genomics: from genotype to disease phenotype in the leishmaniases, *Int. J. Parasitol.*, **37**, 1173–86.
26. Lye, L.F., Owens, K., Shi, H., et al. 2010, Retention and loss of RNA interference pathways in trypanosomatid protozoans, *PLoS Pathog.*, **6**, e1001161.
27. Messing, J., Crea, R. and Seeburg, P.H. 1981, A system for shotgun DNA sequencing, *Nucleic Acids Res.*, **9**, 309–21.
28. Zerbino, D.R. and Birney, E. 2008, Velvet: algorithms for de novo short read assembly using de Bruijn graphs, *Genome Res.*, **18**, 821–9.
29. Quinn, N.L., Levenkova, N., Chow, W., et al. 2008, Assessing the feasibility of GS FLX Pyrosequencing for sequencing the Atlantic salmon genome, *BMC Genomics*, **9**, 404.
30. Sommer, D.D., Delcher, A.L., Salzberg, S.L. and Pop, M. 2007, Minimus: a fast, lightweight genome assembler, *BMC Bioinformatics*, **8**, 64.
31. Pop, M., Kosack, D.S. and Salzberg, S.L. 2004, Hierarchical scaffolding with Bambus, *Genome Res.*, **14**, 149–59.
32. Slater, G.S. and Birney, E. 2005, Automated generation of heuristics for biological sequence comparison, *BMC Bioinformatics*, **6**, 31.
33. Salzberg, S.L., Delcher, A.L., Kasif, S. and White, O. 1998, Microbial gene identification using interpolated Markov models, *Nucleic Acids Res.*, **26**, 544–8.
34. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O. and Borodovsky, M. 2005, Gene identification in novel eukaryotic genomes by self-training algorithm, *Nucleic Acids Res.*, **33**, 6494–506.
35. Haas, B.J., Salzberg, S.L., Zhu, W., et al. 2008, Automated eukaryotic gene structure annotation using EVidence Modeler and the program to assemble spliced alignments, *Genome Biol.*, **9**, R7.
36. Koski, L.B., Gray, M.W., Lang, B.F. and Burger, G. 2005, AutoFACT: an automatic functional annotation and classification tool, *BMC Bioinformatics*, **6**, 151.
37. Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R. and Wu, C.H. 2007, UniRef: comprehensive and non-redundant UniProt reference clusters, *Bioinformatics*, **23**, 1282–8.
38. Marchler-Bauer, A. and Bryant, S.H. 2004, CD-Search: protein domain annotations on the fly, *Nucleic Acids Res.*, **32**, W327–31.
39. Bateman, A., Birney, E., Cerruti, L., et al. 2002, The Pfam protein families database, *Nucleic Acids Res.*, **30**, 276–80.
40. Kanehisa, M. and Goto, S. 2000, KEGG: kyoto encyclopedia of genes and genomes, *Nucleic Acids Res.*, **28**, 27–30.
41. Chen, F., Mackey, A.J., Stoeckert, C.J. Jr and Roos, D.S. 2006, OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups, *Nucleic Acids Res.*, **34**, D363–8.
42. Chen, F., Mackey, A.J., Vermunt, J.K. and Roos, D.S. 2007, Assessing performance of orthology detection strategies applied to eukaryotic genomes, *PloS One*, **2**, e383.
43. Quinlan, A.R. and Hall, I.M. 2010, BEDTools: a flexible suite of utilities for comparing genomic features, *Bioinformatics*, **26**, 841–2.
44. Sharp, P.M. and Li, W.H. 1987, The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications, *Nucleic Acids Res.*, **15**, 1281–95.
45. Sharp, P.M., Tuohy, T.M. and Mosurski, K.R. 1986, Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes, *Nucleic Acids Res.*, **14**, 5125–43.
46. Comeron, J.M. and Aguade, M. 1998, An evaluation of measures of synonymous codon usage bias, *J. Mol. Evol.*, **47**, 268–74.
47. Aslett, M., Aurrecoechea, C., Berriman, M., et al. 2010, TriTrypDB: a functional genomic resource for the Trypanosomatidae, *Nucleic Acids Res.*, **38**, D457–62.
48. Drummond, A.J., Ashton, B., Buxton, S., et al. 2011, Geneious v5.6.3. http://www.geneious.com/ (June 2012, date last accessed).
49. Edgar, R.C. 2004, MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res.*, **32**, 1792–7.
50. Ronquist, F. and Huelsenbeck, J.P. 2003, MrBayes 3: Bayesian phylogenetic inference under mixed models, *Bioinformatics*, **19**, 1572–4.
51. Whelan, S. and Goldman, N. 2001, A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach, *Mol. Biol. Evol.*, **18**, 691–9.
52. Emanuelsson, O., Brunak, S., von Heijne, G. and Nielsen, H. 2007, Locating proteins in the cell using TargetP, SignalP and related tools, *Nat. Protoc.*, **2**, 953–71.
53. Bendtsen, J.D., Jensen, L.J., Blom, N., Von Heijne, G. and Brunak, S. 2004, Feature-based prediction of non-classical and leaderless protein secretion, *Protein Eng Des Sel: PEDS*, **17**, 349–56.
54. Paape, D., Barrios-Llerena, M.E., Le Bihan, T., Mackay, L. and Aebischer, T. 2010, Gel free analysis of the proteome of intracellular *Leishmania mexicana*, *Mol. Biochem. Parasitol.*, **169**, 108–14.
55. Lowe, T.M. and Eddy, S.R. 1997, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, *Nucleic Acids Res.*, **25**, 955–64.
56. Castillo-Ramirez, S., Vazquez-Castellanos, J.F., Gonzalez, V. and Cevallos, M.A. 2009, Horizontal gene transfer and diverse functional constrains within a common replication-partitioning system in Alphaproteobacteria: the repABC operon, *BMC Genomics*, **10**, 536.
57. Bastien, P., Blaineau, C. and Pages, M. 1992, *Leishmania*: sex, lies and karyotype, *Parasitol. Today*, **8**, 174–7.
58. Mannaert, A., Downing, T., Imamura, H. and Dujardin, J.C. 2012, Adaptive mechanisms in pathogens: universal aneuploidy in Leishmania, *Trends Parasitol.*, **28**, 370–6.
59. Sterkers, Y., Lachaud, L., Bourgeois, N., Crobu, L., Bastien, P. and Pages, M. 2012, Novel insights into genome plasticity in Eukaryotes: mosaic aneuploidy in *Leishmania*, *Mol. Microbiol.*, **86**, 15–23.
60. Ning, Z., Cox, A.J. and Mullikin, J.C. 2001, SSAHA: a fast search method for large DNA databases, *Genome Res.*, **11**, 1725–9.

61. Gentil, L.G., Lasakosvitsch, F., Silveira, J.F., Santos, M.R. and Barbieri, C.L. 2007, Analysis and chromosomal mapping of *Leishmania* (Leishmania) amazonensis amastigote expressed sequence tags, *Mem Inst Oswaldo Cruz*, **102**, 707–11.
62. Hutson, S. 2001, Structure and function of branched chain aminotransferases, *Prog Nucleic Acid Res. Mol. Biol.*, **70**, 175–206.
63. Ginger, M.L., Chance, M.L. and Goad, L.J. 1999, Elucidation of carbon sources used for the biosynthesis of fatty acids and sterols in the trypanosomatid *Leishmania mexicana*, *Biochem. J.*, **342**, 397–405.
64. Arruda, D.C., D'Alexandri, F.L., Katzin, A.M. and Uliana, S.R. 2008, *Leishmania amazonensis*: biosynthesis of polyprenols of 9 isoprene units by amastigotes, *Exp. Parasitol.*, **118**, 624–8.
65. Neubert, T.A. and Gottlieb, M. 1990, An inducible 3′-nucleotidase/nuclease from the trypanosomatid *Crithidia luciliae*. Purification and characterization, *J. Biol. Chem.*, **265**, 7236–42.
66. Paletta-Silva, R., Vieira, D.P., Vieira-Bernardo, R., et al. 2011, Leishmania amazonensis: characterization of an ecto-3′-nucleotidase activity and its possible role in virulence, *Exp. Parasitol.*, **129**, 277–83.
67. Holmgren, A. and Lu, J. 2010, Thioredoxin and thioredoxin reductase: current research with special reference to human disease, *Biochem. Biophys. Res. Commun.*, **396**, 120–4.
68. Scott, P. and Sher, A. 1986, A spectrum in the susceptibility of leishmanial strains to intracellular killing by murine macrophages, *J. Immunol.*, **136**, 1461–6.
69. Krauth-Siegel, R.L. and Comini, M.A. 2008, Redox control in trypanosomatids, parasitic protozoa with trypanothione-based thiol metabolism, *Biochim. Biophys. Acta*, **1780**, 1236–48.
70. de Souza Carmo, E.V., Katz, S. and Barbieri, C.L. 2010, Neutrophils reduce the parasite burden in *Leishmania* (Leishmania) amazonensis-infected macrophages, *PloS One*, **5**, e13815.
71. Asato, Y., Oshiro, M., Myint, C.K., et al. 2009, Phylogenic analysis of the genus *Leishmania* by cytochrome b gene sequencing, *Exp. Parasitol.*, **121**, 352–61.
72. Fraga, J., Montalvo, A.M., De Doncker, S., Dujardin, J.C. and Van der Auwera, G. 2010, Phylogeny of *Leishmania* species based on the heat-shock protein 70 gene, *Infect. Genet. Evol.*, **10**, 238–45.
73. Rochette, A., McNicoll, F., Girard, J., et al. 2005, Characterization and developmental gene regulation of a large gene family encoding amastin surface proteins in *Leishmania* spp, *Mol. Biochem. Parasitol.*, **140**, 205–20.
74. Jackson, A.P. 2010, The evolution of amastin surface glycoproteins in trypanosomatid parasites, *Mol. Biol. Evol.*, **27**, 33–45.
75. Cruz, M.C., Souza-Melo, N., da Silva, C.V., et al. 2012, Trypanosoma cruzi: role of delta-Amastin on extracellular amastigote cell invasion and differentiation, *PloS One*, **7**, e51804.
76. Stober, C.B., Lange, U.G., Roberts, M.T., et al. 2006, From genome to vaccines for leishmaniasis: screening 100 novel vaccine candidates against murine *Leishmania major* infection, *Vaccine*, **24**, 2602–16.
77. Rafati, S., Hassani, N., Taslimi, Y., Movassagh, H., Rochette, A. and Papadopoulou, B. 2006, Amastin peptide-binding antibodies as biomarkers of active human visceral leishmaniasis, *Clin. Vaccine Immunol.*, **13**, 1104–10.
78. Salotra, P., Duncan, R.C., Singh, R., Subba Raju, B.V., Sreenivas, G. and Nakhasi, H.L. 2006, Upregulation of surface proteins in *Leishmania donovani* isolated from patients of post kala-azar dermal leishmaniasis, *Microbes Infect.*, **8**, 637–44.
79. Rochette, A., Raymond, F., Ubeda, J.M., et al. 2008, Genome-wide gene expression profiling analysis of *Leishmania major* and *Leishmania infantum* developmental stages reveals substantial differences between the two species, *BMC Genomics*, **9**, 255.
80. Azizi, H., Hassani, K., Taslimi, Y., Najafabadi, H.S., Papadopoulou, B. and Rafati, S. 2009, Searching for virulence factors in the non-pathogenic parasite to humans *Leishmania tarentolae*, *Parasitology*, **136**, 723–35.
81. Naderer, T. and McConville, M.J. 2008, The *Leishmania*-macrophage interaction: a metabolic perspective, *Cell Microbiol.*, **10**, 301–8.
82. De Souza Leao, S., Lang, T., Prina, E., Hellio, R. and Antoine, J.C. 1995, Intracellular *Leishmania amazonensis* amastigotes internalize and degrade MHC class II molecules of their host cells, *J. Cell Sci.*, **108**, 3219–31.
83. Silverman, J.M., Chan, S.K., Robinson, D.P., et al. 2008, Proteomic analysis of the secretome of *Leishmania donovani*, *Genome Biol.*, **9**, R35.
84. Mouchess, M.L., Arpaia, N., Souza, G., et al. 2011, Transmembrane mutations in Toll-like receptor 9 bypass the requirement for ectodomain proteolysis and induce fatal inflammation, *Immunity*, **35**, 721–32.
85. Tuon, F.F., Fernandes, E.R., Pagliari, C., Duarte, M.I. and Amato, V.S. 2010, The expression of TLR9 in human cutaneous leishmaniasis is associated with granuloma, *Parasite Immunol.*, **32**, 769–72.
86. Abou Fakher, F.H., Rachinel, N., Klimczak, M., Louis, J. and Doyen, N. 2009, TLR9-dependent activation of dendritic cells by DNA from *Leishmania major* favors Th1 cell development and the resolution of lesions, *J. Immunol.*, **182**, 1386–96.
87. Carvalho, L.P., Petritus, P.M., Trochtenberg, A.L., et al. 2012, Lymph node hypertrophy following *Leishmania major* infection is dependent on TLR9, *J. Immunol.*, **188**, 1394–401.
88. Favali, C., Tavares, N., Clarencio, J., Barral, A., Barral-Netto, M. and Brodskyn, C. 2007, *Leishmania amazonensis* infection impairs differentiation and function of human dendritic cells, *J. Leukoc. Biol.*, **82**, 1401–6.
89. Lezama-Davila, C.M. and Isaac-Marquez, A.P. 2006, Systemic cytokine response in humans with chiclero's ulcers, *Parasitol Res.*, **99**, 546–53.
90. Linares, E., Augusto, O., Barao, S.C. and Giorgio, S. 2000, *Leishmania amazonensis* infection does not inhibit systemic nitric oxide levels elicited by lipopolysaccharide in vivo, *J. Parasitol.*, **86**, 78–82.