The Comparative Genomics and Phylogenomics of *Leishmania amazonensis* Parasite

Diogo A. Tschoeke¹,²*, Gisele L. Nunes²,⁴*, Rodrigo Jardim¹,², Joana Lima², Aline S. R. Dumaresq², Monete R. Gomes², Leandro de Mattos Pereira², Daniel R. Loureiro¹, Patricia H. Stoco³, Herbert Leonel de Matos Guedes⁴,⁵, Antonio Basílio de Miranda¹,², Jeronimo Ruiz¹,⁶, André Pitaluga⁷, Floriano P. Silva Jr¹,⁸, Christian M. Probst¹,⁹, Nicholas J. Dickens⁵, Jeremy C. Mottram⁵, Edmundo C. Grisard⁵ and Alberto M. R. Dávila¹,²

¹Polo de Biologia Computacional e Sistemas, Instituto Oswaldo Cruz (Fiocruz/IOC), Rio de Janeiro, RJ, Brazil. ²Laboratório de Biologia Computacional e Sistemas, Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro, RJ, Brazil. ³Laboratório de Biologia Molecular de Parasitas e Vetores, Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro, RJ, Brazil. ⁴Laboratório de Protozoologia, Universidade Federal de Santa Catarina, Florianópolis, SC, Brazil. ⁵Laboratório de Inflamação Instituto de Biofísica Carlos Chagas Filho, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brazil. ⁶Wellcome Trust Centre for Molecular Parasitology, Institute of Immunity, Infection and Inflammation, College of MVLS, University of Glasgow, Glasgow, UK. ⁷Instituto René Rachou (Fiocruz/IRR), Belo Horizonte, MG, Brazil. ⁸Laboratório de Biologia Molecular de Parasitas e Vetores, Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro, RJ, Brazil. ⁹Laboratório de Bioquímica de Proteínas e Peptídeos, Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro, RJ, Brazil. ¹⁰Instituto Carlos Chagas (Fiocruz/ICC), Curitiba, PR, Brazil. *Authors contributed equally.

**ABSTRACT:** Leishmaniasis is an infectious disease caused by *Leishmania* species. *Leishmania amazonensis* is a New World *Leishmania* species belonging to the Mexicana complex, which is able to cause all types of leishmaniasis infections. The *L. amazonensis* reference strain MHOM/BR/1973/M2269 was sequenced identifying 8,802 codifying sequences (CDS), most of them of hypothetical function. Comparative analysis using six *Leishmania* species showed a core set of 7,016 orthologs. *L. amazonensis* and *Leishmania mexicana* share the largest number of distinct orthologs, while *Leishmania braziliensis* presented the largest number of inparalogs. Additionally, phylogenomic analysis confirmed the taxonomic position for *L. amazonensis* within the "Mexicana complex", reinforcing understanding of the split of New and Old World *Leishmania*. Potential non-homologous isofunctional enzymes (NISE) were identified between *L. amazonensis* and *Homo sapiens* that could provide new drug targets for development.

**KEYWORDS:** *Leishmania amazonensis*, comparative genomics, phylogenomics

**CITATION:** Tschoeke at al. The Comparative Genomics and Phylogenomics of *Leishmania amazonensis* Parasite. Evolutionary Bioinformatics 2014:10 131–153 doi: 10.4137/EBO.S13759.

**ACADEMIC EDITOR:** Jike Cui, Associate Editor

**FUNDING:** Funding for JCM and NJD was provided by the Wellcome Trust [085349]. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

**COMPETING INTERESTS:** Authors disclose no potential conflicts of interest.

**COPYRIGHT:** © the authors, publisher and licensee Libertas Academica Limited. This is an open-access article distributed under the terms of the Creative Commons CC-BY-NC 3.0 License.

**CORRESPONDENCE:** davela@ficocruz.br

This paper was subject to independent, expert peer review by a minimum of two blind peer reviewers. All editorial decisions were made by the independent academic editor. All authors have provided signed confirmation of their compliance with ethical and legal obligations including (but not limited to) use of any copyrighted material, compliance with ICMJE authorship and competing interests disclosure guidelines and, where applicable, compliance with legal and ethical guidelines on human and animal research participants.

**Introduction**

Leishmaniasis is an infectious disease caused by parasites of the genomes *Leishmania*. It has a worldwide impact with considerable morbidity and mortality rates, especially in the developing countries. The lack of a vaccine and effective treatments is of great concern, since most of the drugs available are toxic and usually lead to side effects.¹ The disease is distributed across 88 countries, and it is estimated that more than 12 million people are currently infected with *Leishmania*. Around 350 million people are living in endemic areas (poor rural and suburban zones),²,³ and only a few countries (Afghanistan, Algeria, Brazil, India, Iran, Nepal, Peru, Saudi Arabia, Sudan, and Syria)² account for about 90% of the global cases. In Brazil, cutaneous leishmaniasis (CL) is endemic and...
caused by at least six Leishmania species from the subgenus Viannia and Leishmania. The main agents of CL in the south of the Amazon basin are Leishmania (Viannia) braziliensis and Leishmania amazonensis, showing no differences in clinical manifestations.\textsuperscript{4,5}

At least 20 Leishmania species are currently known to infect humans and can cause a variety of clinical manifestations depending on the species and the host immune response, ranging from cutaneous lesions to fatal visceral leishmaniasis (VL).\textsuperscript{6–9} The most severe is VL, caused by the Leishmania donovani complex, in which the parasites affect mainly the liver and spleen, resulting in host immunosuppression, progressive fever, weight loss, and anemia.\textsuperscript{2,0} In CL, the parasites cause localized long-term ulceration, inducing chronicity, latency, and, depending on the species, tendency to metastasize in the human host.\textsuperscript{11} Mucocutaneous leishmaniasis (MCL), caused mainly by L. braziliensis, induces the destruction of nasopharyngeal tissue with hideous disfiguring lesions. Diffuse cutaneous leishmaniasis (DCL), caused by L. amazonensis, Leishmania guyanensis, and Leishmania aethiopica, is a long-lasting disease because of a deficient cellular-mediated immune response presenting a progressive primary lesion and multiple metastatic lesions.\textsuperscript{12–13}

L. amazonensis is associated with a variety of clinical manifestations, CL, DCL (rare manifestation), MCL, VL, and post-kala-azar dermal leishmaniasis (PKDL), a systemic cutaneous form that occurs in some patients following VL treatment and apparent cure.\textsuperscript{6} In Brazil, PH8 (IFLA/BR/67/PH8) is one of the most studied L. amazonensis strains as it is a component of Leishvaccine.\textsuperscript{14} However, this strain was isolated from sand fly and an isolate from human disease was chosen. In this study, we present the genome of the L. amazonensis (MHOM/BR/71973/M2269), which was isolated from a single human cutaneous lesion.

Many advances have occurred during the last decade in the genomic area, mostly after the development of the high-throughput sequencing methods. Several trypanosomatids reference genomes have already been sequenced, among them are Trypanosoma cruzi, Trypanosoma brucei, Leishmania major, Leishmania infantum, L. braziliensis, and Leishmania mexicana.\textsuperscript{15–21} L. major was the first Leishmania genome sequenced showing 32.8 Mbp size and 8,311 predicted protein-coding genes (coding sequences (CDS)).\textsuperscript{15} A comparative analysis using other three Leishmania species (L. major, L. infantum, and L. braziliensis) was carried out in 2007 and revealed a highly conserved genomic organization, in which the genomes display 8,300 genes in average of which more than 99% of the genes were highly syntenic. However, around 200 genes were differentially distributed among the three compared genomes, showing 47, 27, and 5 exclusive genes (species specific or unique) for L. braziliensis, L. infantum, and L. major, respectively.\textsuperscript{18} The most recently sequenced Leishmania genome was L. mexicana published in 2011\textsuperscript{21} and, during our sequencing effort, the L. amazonensis genome was described by Real and colleagues\textsuperscript{22} in 2013.

Considering the “TriTryp” genomes (T. cruzi, T. brucei, and L. major), approximately 6,200 genes are conserved among the three species and 94% of these genes were syntenic.\textsuperscript{23} Most of the species-specific genes are located in non-syntenic regions/chromosomes and consist of members of large surface antigen families.\textsuperscript{24} Such observation may indicate that differences detected among these parasites’ pathogenesis are likely. In this study, we performed a detailed comparative analysis of L. amazonensis M2269 genome with the L. mexicana U1103, L. major Friedlin, L. infantum JPC5, L. donovani BPK282A, and L. braziliensis M2604 genomes retrieved from GenBank, identifying new features of this parasite genome and also performing a phylogenomic analysis of the genus.

Material and Methods

DNA preparation and sequencing. L. amazonensis reference strain MHOM/BR/1973/M2269, provided by Dr Paul Bates, was used in this study. Genomic DNA was extracted using a Qiagen QIAamp DNA Kit, according to the manufacturer’s instructions. The extracted DNA was sequenced in a Solexa sequencer (Illumina) using paired-end reads of 50 + 50 bases.

Assembly, sequence analysis, and annotation. Genomic sequences of 66,869,406 reads were trimmed for platform-dependent systematic errors and then quality was evaluated using Phred (cutoff Q = 26).\textsuperscript{25,26} High-quality reads (~28,52× genome coverage) were assembled using Velvet version 0.7.55 software\textsuperscript{27} resulting in 10,721 contigs with mean contig length of 2,817 bp and N50 value of 6,946. De novo and reference genome assembly strategies were applied using the L. mexicana genome (GenBank Assembly ID: GCA_000234665.4 and RefSeq Assembly ID: GCF_000234665.1) and L. major genome (GenBank Assembly ID: GCA_000002725.2 and RefSeq Assembly ID: GCF_000002725.2). Assemblies were merged using in-house developed Perl scripts, and contigs were generated by comparing the assembled scaffolds and contigs with the Leishmania genomes available at the GenBank.

The multifasta files of the assembled L. amazonensis genome were submitted to STINGRAY pipeline (Wagner, et al., 2014)\textsuperscript{28} (http://stingray.biowebdb.org), an improved version of the original GARSA\textsuperscript{29} system, for semi-automatic annotation. The STINGRAY pipeline and a TblastX\textsuperscript{30} approach were used by transferring the L. mexicana\textsuperscript{21} annotation to L. amazonensis, which was further improved by the identification of conserved domains.

Protein families (Pfam) and domain identification. Pfam–A (v. 26.0)\textsuperscript{31,32} and Hmmer 3.0\textsuperscript{33} were used against the 8,802 predicted protein sets using hmmssearch program with an e-value 1e − 5 and other default parameters.

Gene ontology (GO) inference. The L. amazonensis proteins were also analyzed using GO.\textsuperscript{34} Briefly, similarity analysis was performed with the STINGRAY pipeline,\textsuperscript{29} using Blastp (v. 2.2.23)\textsuperscript{35} against the GO database (go_20130223-seqdb.fasta), and then proteins were classified within one of three
GO categories, as follows: (i) biological process, (ii) molecular function, and (iii) cellular component.

Conserved domain identification. Conserved domains were identified using RpsBlast (v. 2.2.23)\textsuperscript{30} analysis on the 8,802 proteins inferred in \textit{L. amazonensis} against seven databases simultaneously (CDD.v3.10–Conserved Domain Database, COG.v1.0–Cluster of Orthologous Groups, KOG.v1.0–Cluster of Eukaryotic Orthologous Groups, Pfam.v26.0–Protein Family, PRK.v6.0, SMART.v6.0, and TIGR.v13.0) with an e-value 1e − 05.

Identification of orthologous and paralogous groups. The identification of orthologous proteins was performed using results generated by the OrthoMCL v.1.4 software.\textsuperscript{35} Orthologous proteins shared by all six \textit{Leishmania} species (\textit{L. major}, \textit{L. infantum}, \textit{L. donovani}, \textit{L. braziliensis}, \textit{L. mexicana}, and \textit{L. amazonensis}) were inferred. The protein function of those inferred orthologs was semi-automatically transferred from previously annotated \textit{Leishmania} genomes. Inparalogous and recent paralogous proteins in \textit{L. amazonensis} were also identified inside the output file generated by the OrthoMCL software.

The orthologous proteins shared among the different \textit{Leishmania} species as well as the inparalogous proteins from \textit{L. amazonensis} and other species were used to generate a Venn diagram using R software.\textsuperscript{36}

Putative orphan proteins identification. To find putative orphan proteins, ie, not homologous to any protein in this study, a first list of protein identifiers was generated and used as input to OrthoMCL to build a second list with protein identifiers clustered by OrthoMCL. Then, these two lists (submitted versus clustered) were compared using a script written in Ruby language to obtain the identifiers of the putative orphan proteins. Since these potential orphans are based on a universe of only six \textit{Leishmania} genomes and to minimize possible misclassification, we performed a BlastP search (v. 2.2.28+\textsuperscript{39}) with these putative orphan proteins against RefSeq database (r.56 18, 132, 578 sequences). These steps allowed us to identify proteins that were classified as putative orphans having similarity to prokaryotic or other eukaryotic (non-\textit{Leishmania}) protein. Finally, proteins without any match to Refseq database were considered as orphan proteins in this study.

Phylogenomics. The phylogenomic tree was inferred based on the studies of Ocaña and Dávila,\textsuperscript{37} and Ciccarelli and colleagues.\textsuperscript{38} Thirty-one universal orthologous (UO) genes showing 1:1 orthologous relationships were used. These UO genes originally identified by Ciccarelli et al.\textsuperscript{38} showed the following characteristics: (i) were present in all complete genomes available at GenBank until 2006 and (ii) were not involved in horizontal transfer. Since these 31 UO genes are directly connected to the protozoan genome available at RefSeq and ProtozoaDB,\textsuperscript{39} they were mapped to the \textit{L. major} proteins using both (a) the best blast hits (e-value 1e − 50) and (b) the manual verification of the annotation (the RefSeq annotation of the best hits needed to match the UO annotation).

Once mapped, the \textit{L. major} protein sequences corresponding to these 31 UO genes were searched in the orthologous groups identified in the six \textit{Leishmania} species by OrthoMCL. We also mapped the 31 orthologs in 28 distinct protozoa species, as follows: \textit{Anonymonas deanei}, \textit{Babesia bovis}, \textit{Babesia equi}, \textit{Babesia microti}, \textit{Cryptosporidium muris}, \textit{Dictyostelium discoideum}, \textit{Entamoeba diapai}, \textit{Entamoeba histolytica}, \textit{Entamoeba invadens}, \textit{Giardia lamblia}, \textit{L. amazonensis}, \textit{L. braziliensis}, \textit{L. donovani}, \textit{L. infantum}, \textit{L. major}, \textit{L. mexicana}, \textit{Naegleria gruberi}, \textit{Neospora caninum}, \textit{Plasmodium berghei}, \textit{Plasmodium cynomolgi}, \textit{Plasmodium falciparum}, \textit{Plasmodium knowlesi}, \textit{Plasmodium vivax}, \textit{Polypodendylium pallidum}, \textit{Strigomonas culicis}, \textit{Tetrabymena thermophila}, \textit{Theileria annulata}, \textit{Theileria orientalis}, \textit{Theileria parva}, \textit{Toxoplasma gondii}, \textit{Trichomonas vaginalis}, \textit{T. brucei}, \textit{T. cruzi}, and \textit{Tryptansoma vivax}. Finally, each of these 31 mapped orthologs were exported as multifasta files and aligned using Mafft v5.861,\textsuperscript{40} using the default parameters.

A supermatrix tree was obtained using concatenated multiple alignments from entire protein sequences. The individual alignments were concatenated using an in-house perl script, resulting in a global supermatrix of 9,450 positions for the six species. The resulting supermatrix was used to generate the phylogenomic tree with MEGA 5,\textsuperscript{41} inferred by Maximum Likelihood using 1,000 bootstrap replicates. We opted to use the JTT model in the single (concatenated) alignment, which was also the model adopted in the phylogenomics studies of Ciccarelli et al.\textsuperscript{38} and Ocaña and Dávila.\textsuperscript{37} Jones, Taylor and Thornton (JTT) model assumes that there are two classes of sites, one class being invariable and the other class being free to change.\textsuperscript{42}

Intragenomic and intergenomic non-homologous iso-functional enzymes (NISE) identification. To identify in the genome of \textit{L. amazonensis} possible cases of intra- and intergenic NISE between this genome and the \textit{Homo sapiens} genome, we applied methodologies previously described.\textsuperscript{43–45} Briefly, protein sequences of enzymes with the same functional activity were downloaded and grouped according to its functional activity as determined by the classification from the International Union of Biochemistry and Molecular Biology, the Enzyme Commission (EC) number.\textsuperscript{46} Protein sequences and functional classification were obtained from KEGG (Kyoto Encyclopedia of Genes and Genomes, version 58.1).\textsuperscript{47} After grouping, we performed a step to confirm the functional activity assigned by KEGG. First, we removed sequences with less than 60 amino acids from the 8,802 \textit{L. amazonensis} predicted proteins, since they may represent protein fragments, resulting in a data set of 8,575 predicted proteins. Then, the protein primary structures inside each protein functional group were compared in a pairwise, all-against-all manner, using Blastp. Functional activities were confirmed via the AnEnPi’s module,\textsuperscript{43} which classifies the enzymes in accordance to the EC number. This classification is obtained after parsing the results of Blastp, using the data set of predicted proteins from \textit{L. amazonensis} as query and the groups
previously obtained as subjects. A restrictive e-value of 10^{-20} was used as a threshold^{34,45,48} to include a primary structure in a group or cluster. Proteins were considered to be NISE if, inside each group of functionally related enzymes, they were allocated in different clusters after parsing the results from Blastp. Possible analogy cases were verified by the examination of their folding categories as classified by the SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/)^{49} and SUPERFAMILY databases (supfam.cs.bris.ac.uk).^{50} Further refinement of confirmed NISEs was achieved by three-dimensional (3D) structure prediction of _L. amazonensis_ proteins by homology modeling and structural comparison with their human analogous counterparts (see below).

**Homology modeling of _L. amazonensis_ proteins and comparative structural analysis with human proteins.** Modeling of the 3D structure of the selected NISEs of _L. amazonensis_ was performed by the satisfaction of spatial restraints method implemented in the program Modeller 9v10.^{51} Global pairwise sequence alignment between the target _L. amazonensis_ sequences and the respective templates was performed with the needle (Needleman–Wunsch) program within the EMBOSS v.6.3.1 package.^{52} The models were constructed using as templates the atomic coordinates of PDB IDs listed in Supplementary Table S2 for each of the selected analogy cases. Ten models were generated for each protein target sequence, and the model with the most favorable DOPE-score and the lowest Modeller objective function value was subjected to external assessment of the stereochemical and overall structural quality within the Structural Analysis and Verification Server (SAVES v.4) (http://services.mbi.ucla.edu/SAVES/). All models selected for further analysis had at least 95% of residues in the most favorable and additionally allowed regions of Ramachandran plots along with other reasonable stereochemical quality parameters. Inspection of molecular structures and other structural analysis was performed by SYBYL X-1.3 software (Tripos L.P., St. Louis, MO).

**_L. amazonensis_ genome functional categorization.** To briefly know the genome content of _L. amazonensis_, we performed a functional categorization through similarity analysis using Blast and RpsBlast programs against the database of orthologous genes in prokaryotes (COG/NCBI) and eukaryotic orthologous genes (KOG/NCBI),^{53,54} which are classified in functional categories (ftp://ftp.ncbi.nih.gov/pub/COG/COG/fun.txt). To infer to which functional category each protein belongs, a cutoff e-value of 1e−5 was used in both programs and databases. Plots of the functional categories were created with R software.

_L. amazonensis_ proteome was also characterized by Pfam (v. 26.0)^{52} and by CDD (v 3.10) through RpsBlast. A further analysis was performed using in-house perl scripts to identify (i) which genes were identified only by Pfam with Hmmer 3, (ii) which ones were identified only by “Conserved Domains” (CDD) and (iii) which ones were characterized by both of them (Pfam and CDD).

**Leishmania core proteome identification.** The _Leishmania_ spp. core proteins (LCP) were identified and analyzed among the orthologous groups and defined as orthologous proteins shared by all the six _Leishmania_ species studied. To find the LCP, the OrthoMCL results were analyzed, and only orthologs shared by the “6 taxa” were chosen. LCP functions were accessed through annotation provided with the sequences.

**_L. amazonensis_ database.** The contigs generated from the assembly of sequencing reads, and the genes and proteins found from these contigs are all available for public consulting in the STINGRAY pipeline (http://stingray.biowebdb.org). Furthermore, _L. amazonensis_ contigs were submitted to GenBank under BioProject ID PRJNA221875, locus_tag prefix Q771.

**_L. amazonensis_ RNA Interference (RNAi) Machinery.** Identification of RNAi genes in _L. amazonensis_ genome was performed through Blast analysis using as query RNAi genes from _Leishmania_ spp. and _T. brucei_ genes annotated as participants of the RNAi pathway in GenDB database^{55} (www.genedb.org). The genes related to RNAi machinery in _L. amazonensis_ were then submitted to a phylogenetic analysis using MEGA5,^{41} and a tree was inferred by Neighbor-Joining using 1,000 bootstrap replicates.

**Synteny Analysis: _L. mexicana Versus L. amazonensis_.** The synteny analysis was performed using the ABACAS^{56} (v. 1.3.1) pipeline, and the results were visualized with ACT (Artemis Comparison Tool),^{57} v. 12.0.0. The following steps were carried out to accomplish the analysis: (i) the 34 _L. mexicana_ chromosomes were concatenated in a single fasta file; (ii) the multifasta of the 8,552 putative _L. amazonensis_ CDS was compared to that of the _L. mexicana_ genome (chromosome) using ABACAS; and (iii) the _L. amazonensis_ genome on the ABACAS output had the CDS reordered, and the resulting comparison was visualized on ACT.

**Results**

**Sequencing, assembly, and genome characteristics.** The _L. amazonensis_ genome assembly was obtained via a reference-guided approach where the obtained contigs were aligned against the reference _L. mexicana_ genome. The assembly resulted in 10,305 contigs, with approximately 59% GC content. The smallest and largest detected contigs had 96 and 141,211 bases, respectively, with a mean of 2,879 bp and median of 853 bp (Table 1). _L. amazonensis_ genome presented 8,802 protein-coding genes after analysis with TblastX against _L. mexicana_ and Refseq databases. The largest coding region had 19,872 bp and the smallest only 66 bp with median and mean of 1,637 bp and 1,209 bp, respectively (Fig. 1). The GC content for coding regions was 61.1%. Of these 8,802 proteins, 5,554 were putative proteins, and 887 were not clustered by OrthoMCL and then were
Leishmania amazonensis genome analysis

Table 1. Summary of the Leishmania amazonensis assembly and genome.

| Contigs  | 10,305 |
|---------|--------|
| Sum of consensus sequences length | 29,670,588 bases |
| Number of scaffolds >1 K nt | 4827 (46.8%) |
| Number of scaffolds >10 K nt | 732 (7.1%) |
| Number of scaffolds >100 K nt | 2 (0.02%) |
| Coding genes: CDS | 8,802 |
| Chromosome | 34 |
| %GC content: Contigs/CDS | 59%/61.125% |

Size: Contigs/CDS

| Max (bases) | 141,211/(19,872) |
| Min (bases) | 96/(66) |
| Mean (bases) | 2,879/(1,637) |
| Median (bases) | 853/(1,209) |
| N50 scaffold length | 8,346 |

CDS ontology

| Molecular function | 4,065 |
| Biological process | 4,007 |
| Cellular component | 4,054 |
| Protein families (PFAM) | 3,075 |
| Conserved domains (CDD) | 6,144 |
| Annotated as “Hypothetic protein” | 5,554 |
| Putative orphans (OrthoMCL) | 887 |

analyzed with Blast against Refseq database with e-value 1e−5, resulting in 14 proteins classified as orphans in this study (Table 2). Furthermore, while some genes were found occurring in single copy, such as ribosomal protein S2 (rpS2) and ribosomal protein L7, other genes exhibited multiple copies, such as ATP-binding cassette (ABC) transporter (50 copies) and calpains (44 copies). Nonetheless, 63% of the CDS were annotated as proteins with hypothetical function (Table 3).

Functional Analysis of L. amazonensis Proteins

Taking into consideration the GO classification, the most frequent molecular functions of L. amazonensis proteins were protein binding (9% or 1,153/12,328), nucleotide binding (8% or 947/12,328), metal ion binding (5% or 653/12,328), receptor activity (4% or 473/12,328), DNA binding (4% or 468/12,328), signal transducers activity (4% or 453/12,328), and binding (4% or 451/12,328) (Fig. 2A). The most representative functions related to biological process were signal transduction (3% or 308/11,799); transmembrane transport (3% or 306/11,799); regulation of transcription, DNAdependent (2% or 251/11,799); and transport (2% or 192/11,799) (Fig. 2B). The last GO category, Cellular Component, had most frequent components related to: cytoplasm (12% or 1,441/12,111), membrane (10% or 1,168/12,111), nucleus (7% or 856/12,111), intracellular (7% or 810/12,111), and plasma membrane (6% or 694/12,111) (Fig. 2C). The most abundant protein-coding genes detected in the L. amazonensis genome were ABC transporter, kinesin, ATP-dependent RNA helicase, heat shock proteins (HSPs), protein kinase, dynein heavy chain, calpains, and amastin surface glycoprotein.

Figure 1. Average size (bases) from putative CDS identified in L. amazonensis genome.
(Table 3). Pfam and conserved domain (CDD) analyses were performed to identify the families/domains present in the 8,802 putative proteins. Of these, 3,075 proteins were assigned to the family level using Pfam, representing a total of 1,004 different families in the *L. amazonensis* genome. The largest family assigned by Pfam was kinase, which contains 69 entries for Pkinase_Tyr and 64 for Pkinase, accounting for about 2% of the total families detected (Figs. 3 and 4). The families TPR, zf-C3HC4_2, DnaJ, RRM_1, AAA22, Helicase C, URR1, URR6, and AAA25 range in size from 58 to 36 proteins, and 617 families were represented by a single protein (Fig. 3). The RpsBlast analysis, used to find the CDDs, characterized 6,144 domains (Fig. 5), in which approximately 1,800 were found in single copies. The domains most frequently found in *L. amazonensis* proteins were SMC_prok_B (chromosome segregation protein SMC) with 131 hits, PHA03247 (large tegument protein UL36) with 126 copies, and PRK07003 (DNA polymerase III subunits gamma and tau) with 113 copies. Altogether, the fact that more than 60% (5,554/8,802) of the proteins were annotated as hypothetical, 6,144 domains were found using CDD, and 1,004 different families were identified by Pfam highlights the great and unknown diversity of *Leishmania* spp. functionality. The combination of Pfam and CDD results (Fig. 4) showed 2,483 proteins simultaneously assigned to some Pfam family and CDD, with Pkinase_Tyr being the most frequent family found that has some CDD associated. Nevertheless, nearly 5,500 *L. amazonensis* proteins were not functionally annotated or were not assigned to any protein family, which is consistent with other *Leishmania* genomes. The functional analysis of *L. amazonensis* according to KOG and COG categories confirmed the specificity of its proteins, since R category (general function prediction only) was the most abundant category found (Fig. 6).

### Table 2. Resume table of most abundant and single copy genes/domains found in *Leishmania amazonensis* genome analysis.

| MOST ABUNDANT GENES/DOMAINS | SINGLE COPY GENES/DOMAINS |
|-----------------------------|--------------------------|
| ABC transporter             | rpS2                     |
| Amastin surface glycoprotein| rpS5                     |
| ATP-dependent RNA helicase  | rpS8                     |
| Calpains                    | rpS10                    |
| Dynein heavy chain          | rpS12                    |
| Heat Shock Proteins (HSPs)  | rpL7                     |
| Kinesin                     | rpL12                    |
| Protein kinase              | rpL13                    |
| WD40                        | rpL19                    |
| Chaperone DNAJ              | rpL23                    |

**Notes:** Most abundant genes/domains found in the initial *Leishmania amazonensis* genome analysis. Genes/domains found in single copy during the analysis. 40S ribosomal proteins (rpS) and 60S ribosomal proteins (rpL).

### Table 3. List of orphans proteins found in *Leishmania amazonensis* with their respective identification, description and length (aa).

| IDENTIFICATION   | DESCRIPTION               | LENGTH |
|------------------|---------------------------|--------|
| LAJMNGS001H06.b.195 | Unspecified product       | 98     |
| LAJMNGS002H09.b.421  | Unspecified product       | 150    |
| LAJMNGS005H02.b.1027 | Hypothetical protein, conserved | 79     |
| LAJMNGS006F03.b.1178 | Hypothetical protein, unknown function | 771    |
| LAJMNGS018E09.b.3196 | Carboxypeptidase, putative | 325    |
| LAJMNGS018H07.b.3264 | Hypothetical protein       | 951    |
| LAJMNGS027A04.b.4532  | Hypothetical protein       | 167    |
| LAJMNGS030G04.b.5103  | Unspecified product        | 48     |
| LAJMNGS031F02.b.5255  | Unspecified product        | 212    |
| LAJMNGS038C10.b.6191  | Unspecified product        | 37     |
| LAJMNGS038E01.b.6205  | Unspecified product        | 94     |
| LAJMNGS051A11.b.7995  | Hypothetical protein       | 139    |
| NODE_5216_1           | Hypothetical protein, Unknown function | 68     |
| NODE_20256_1          | Unspecified product        | 81     |
Comparative Analysis

A comparative analysis to identify orthologous proteins among the six different *Leishmania* genomes was performed using OrthoMCL. Most of the *L. amazonensis* proteins revealed to be orthologs were observed on all genomes evaluated, consisting of the *Leishmania* spp. core genome. A total of 7,016 (79.7%) orthologous groups were found among *L. amazonensis*, *L. donovani*, *L. mexicana*, *L. infantum*, *L. braziliensis*, and *L. major* (Fig. 7) (Supplementary File, Table S1). Within LCP, approximately 4,800 (68.4%) orthologs were annotated as hypothetical proteins; however, among those who have a defined function, we found proteins such as amastin, calpain-like cysteine peptidase, 40S ribosomal protein S16, RNA helicase, protein kinase, dynein heavy chain, activated protein kinase c receptor (LACK), ABC transporter, tuzin, and DNA primase large subunit. Considering genes shared between two *Leishmania* species, we found 18 orthologous protein groups between *L. amazonensis* and *L. mexicana*, which are closely related and belong to the *L. mexicana* complex (Table 4 and Supplementary File). Within these 18 orthologous groups, 7 proteins had an identified function (kinetoplast-associated protein, 3-hydroxyisobutyryl-coenzyme a hydrolase-like protein, viscerotropic Leishmaniasis antigen, ribosomal protein L1a, amastin, viscerotropic Leishmaniasis antigen, and flagellar calcium-binding protein) and 12 were classified with hypothetical function. The comparison of the most distant species inside *Leishmania* genus, *L. (L.) amazonensis* versus *L. (V.) braziliensis*, showed that nine proteins were exclusive and shared by both, among which four had known function: heat shock 70-related protein 1, beta tubulin, tyrosine/dopa decarboxylase, and oxidoreductase (Table 4). When inparalogous proteins were evaluated in *L. amazonensis*, one paralog was found: triacylglycerol lipase-like protein (Table 4).

A phylogenomic analysis was performed based on 31 UO genes to confirm that *Leishmania* species are closely related, mainly regarding the *Leishmania* and the *Vianna* subgenus. Figure 6 shows the relationship between the species from *L. mexicana* (*L. amazonensis* and *L. mexicana*) and *L. donovani*. 

**Figure 3.** Conserved domains identification generated by RpsBlast with CDD database. 
**Notes:** Only 20 most domains were represented in the legend. Remaining families are grouped into green square and uncharacterized proteins are in purple.
(L. donovani and L. infantum) complexes and also other 28 protozoa species. Even though these orthologous genes are very close, the differences among groups could still be observed in the generated dendrogram (Fig. 8). This result was supported by an alignment using one of the UOs (DNA-directed RNA polymerase; Fig. 9), where L. amazonensis and L. mexicana have very similar sequences, while L. braziliensis has the most divergent sequence, indeed presenting gaps in the multiple alignment. L. braziliensis, which belongs to Viannia subgenus, was in fact positioned in a different clade, underlining their differences and reflecting the divergence observed in the alignment (Fig. 9). At the same Leishmania clade, another five Kinetoplastida species are found, reflecting the monophyletic nature of this genus.

NISE

After the initial clustering of 4,215 ECs available in KEGG with AnEnPi, 412 ECs present in more than one cluster were detected. This group of 412 ECs was parsed for L. amazonensis sequences allocated in different clusters with the same enzymatic activity. Using this procedure, it was possible to identify 25 potential cases of NISEs when L. amazonensis was compared to H. sapiens (termed “intergenomic NISE”). In addition, 31 potential cases of NISEs were identified when L. amazonensis protein sequences (termed “intragenomic NISE”) were searched. The presence of NISEs was detected in five of the six main EC classes, such as Oxidoreductases (EC 1), Transferases (EC 2), Hydrolases (EC 3), Lyases (EC 4), and Isomerases (EC 5), but no cases of functional analogy on Ligases (EC 6) were found (Supplementary File, Tables S2–S4).

Complementary analyses based on the SUPERFAMILY database at potential NISEs excluded few cases, where the status of “Predicted NISE” was given to cases with no significant hits on the SUPERFAMILY database, since although the sequences were allocated in different clusters, we could not confirm the structural differences (Supplementary File, Tables S2 and S3). Among the 25 potential intergenomic NISEs, 14 cases were confirmed and 1 case was considered as predicted NISE. Among the 31 potential intragenomic NISEs, 15 cases were confirmed and 1 was considered as predicted NISE (partially demonstrated on Tables S2 and S3). It is important to emphasize that the approach considered here is very restrictive since only NISEs under the same EC that had different folds (not sharing any type of fold) were taken into account.

Further structural characterization was performed using three confirmed intergenomic NISE cases that showed clear homology (above 30% sequence identity) with a protein with known 3D structure deposited in PDB and that had a solved structure for its human analogous counterpart: LAJMNGS050H11.b.7960 (EC1.1.1.2—putative
| ORTHOMCL | L. AMazonensis ACCESSION | PFAM ANNOTATION | CDD ANNOTATION | PROTEIN DESCRIPTION | L. AM | L. ME | L. DO | L. BR |
|----------|-------------------------|-----------------|----------------|---------------------|------|------|------|------|
| ORTHOMCL7819 | LAJMNGS015A07.b.2588  
LAJMNGS029B11.b.4872 |                  |                | Triacylglycerol lipase |     | X    |      | X    |
| ORTHOMCL7785 | NODE_9861_1  
g|401424225 |                | Kinetoplast-associated protein | X    | X    |      |      |
| ORTHOMCL7789 | NODE_20602_1  
g|401430272 | pfam07344, Amastin | Unspecified product | X    | X    |      |      |
| ORTHOMCL7794 | NODE_11369_4  
g|401427459 | pfam13766, ECH_C,  
2-enoyl-CoA Hydratase | 3-hydroxyisobutyryl-coenzyme a hydrolase | X    | X    |      |      |
| ORTHOMCL7802 | LAJMNGS046H11.b.7373  
g|401414833 |                | Viscero tropic Leishmania antigen | X    | X    |      |      |
| ORTHOMCL7803 | LAJMNGS046G07.b.7351  
g|401418572 | cd00051, EFh, EF-hand,  
calcium binding motif | Flagellar calcium-binding protein,  
putative | X    | X    |      |      |
| ORTHOMCL7808 | LAJMNGS033H06.b.5610  
g|401426307 | PTZ00201, amastin surface  
glycoprotein | Ama tin-like protein | X    | X    |      |      |
| ORTHOMCL7813 | LAJMNGS029D07.b.4914  
g|401427209 | PF13415.1Kelch_3 | Hypothetical protein | X    | X    |      |      |
| ORTHOMCL7814 | LAJMNGS029D03.b.4903  
g|401430342 | CD00806.14PUF | Unspecified product | X    | X    |      |      |
| ORTHOMCL7822 | LAJMNGS010A05.b.1767  
g|401415906 | PTZ00428, 60S ribosomal  
protein L4 | Ribosomal protein L1a, putative | X    | X    |      |      |
| ORTHOMCL7717 | NODE_33600_1  
g|39801921  
g|39801923 | Amino acid permease | X    | X    |      |      |
| ORTHOMCL7788 | NODE_21871_1  
g|398010889 | PTZ00201, amastin | Ama tin-like protein | X    | X    |      |      |
| ORTHOMCL7790 | NODE_20189_1  
g|398010239 | cd03213, ABCG_EPDR | ATP-binding cassette protein  
subfamily G, member 1 | X    | X    |      |      |
| ORTHOMCL7792 | NODE_12712_1  
g|398023645 | PTZ00263, protein kinase A | Protein kinase A catalytic subunit  
isomorph 2 | X    | X    |      |      |
| ORTHOMCL7795 | NODE_10493_4  
g|398023914 |                  | Phosphoglycan beta 1,3  
galactosyltransferase 4 | X    | X    |      |      |
| ORTHOMCL7800 | LAJMNGS047C07.b.7405  
g|39801940 | COG1788, Acyl CoA:  
acetate/3-ketoacid  
succinyl-coa:3-ketoacid-coenz yme transferase- like protein | Multidrug resistance  
protein, putative  
putative, ABC transporter | X    | X    |      |      |
| ORTHOMCL7801 | LAJMNGS047B12.b.7397  
g|398015472 | PTZ00243, ABC transporter | Calpain-like cysteine peptidase | X    | X    |      |      |
| ORTHOMCL7809 | LAJMNGS033B09.b.5509  
g|398014546 |                  | Vac uolar-type Ca2 -ATPase,  
putative | X    | X    |      |      |
| ORTHOMCL7811 | LAJMNGS030F01.b.5087  
g|398010628 |                  | Bet a-fructosidase, invertase,sucrose  
hydrolase | X    | X    |      |      |

Table 4. Identification of orthologous groups between L. amazonensis and Leishmania species and inparalogous from L. amazonensis only characterized orthologs are listed.
**Leishmania amazonensis** genome analysis

| Gene Id | Description                                                                 |
|---------|-----------------------------------------------------------------------------|
| DE_11708_1 | gi|154341831 | Heat shock 70 kDa - related protein 1, mitochondrial precursor, putative |
| DE_11708_2 | gi|154341835 | Hypothetical protein, conserved |
| DE_11708_3 | gi|154341839 | Putative 2,4-dienoyl-coa reductase FADH1 |
| DE_11708_4 | gi|154341843 | Putative isopentenyl-diphosphate delta-isomerase |

Finally, a search for the intergenic NISE detected in this study was performed against drug target databases such as TDR targets, TTD, and DrugBank, verifying that some of these NISE are already under study as potential drug targets against other pathogens. The complete list containing such targets and the pathogens is in Supplementary File (Supplementary Table S5).

**RNAi Pathway in *L. amazonensis***

Some RNAi pathway-related genes are present in *L. amazonensis* (Table 7). Dicer seems to be missing in trypanosomatids that lack a functional RNAi pathway. We were unable to detect Dicer in *L. amazonensis* genome or any sequence bearing the characteristic Rnc (dsRNA-specific ribonuclease) domain. However, the presence of a possible functional Dicer homolog with very divergent sequence is not definitely discarded, and more studies need to be carried out. Nine DEAD/H box RNA helicase and two ribonuclease III genes with putative relationship to RNAi pathway were identified in *L. amazonensis* (Table 7). Although Dicer was not identified, some Dicer-related genes were characterized. Four ERI sequences were identified in *L. amazonensis* genome data set (LAJMNGS009D01.b.1653, LAJMNGS023D01.b.3956, LAJMNGS034E11.b.5717, and LAJMNGS035F02.b.5853) (Table 7). Two genes of the RNA-induced silencing complex (RISC; a major effector complex of the RNAi pathway) were also identified: tudor and piwi (argonaute family) (Table 7).

The *L. amazonensis* argonaute-like gene identified (LlPWI1) is phylogenetically related to TbPWI1, which is not involved in RNAi. The full sequence of the LlPWI1 gene in *L. amazonensis* and its orthologs was submitted for phylogenetic analysis (Fig. 11). The neighbor-joining tree clearly distinguishes two functionally different forms of argonaute family proteins based on *T. brucei*, only *L. braziliensis* possesses the two forms of argonaute family genes (ACI22628 and XP_001564757), which are related to TbAGO1 and TbPWI1, respectively.

**Synteny Analysis Between *L. mexicana* and *L. amazonensis***

The results of the synteny analysis between the *L. mexicana* and *L. amazonensis* genomes (Fig. 12) showed no synteny
breaks or inversions. The red line connecting the *L. mexicana* (upper) and *L. amazonensis* sequences (bottom) represents a good match (more than 92% of identity), and 99.87% of the *L. amazonensis* sequences (8,541/8,552) have a good match with the *L. mexicana* sequence. Only three *L. amazonensis* sequences (succinyl-coa:3-ketoacid-coenzyme a transferase-like protein and two unspecified products) did not match with *L. mexicana* chromosome sequence.

**Discussion**

The present assembly of the *L. amazonensis* genome resulted in 29,670,588 bases, consisting of 8,802 putative CDS with a GC content of 58.5% for the contigs and 61.12% for the CDS, while the present assembly of the *L. mexicana* chromosome sequence. Besides the 50 ABC copies annotated by us in *L. amazonensis*, we observed 33 copies in *L. mexicana*. Possibly, some ABC transporter genes in *L. amazonensis* may be incomplete and the number overestimated because of the presence of the same gene on multiple contigs. Forty-four calpains were found in *L. amazonensis*. Mottram and colleagues found 27 in *L. major*, and Ersfeld and colleagues found 24 copies in *T. cruzi* and 18 copies in *T. brucei*. Calpains are involved in the remodeling of cytoskeletal or membrane attachments and have been found mostly in invertebrates and lower eukaryotes. The importance of cytoskeleton remodeling during *Leishmania* spp. differentiation may explain the high number of Calpain genes in these parasites.

*Figure 5.* Functional category by KOG and COG for *Leishmania amazonensis* proteins INFORMATION STORAGE AND PROCESSING: [J] Translation, ribosomal structure and biogenesis, [K] RNA processing and modification, [L] Replication, recombination and repair, [M] Chromatin structure and dynamics, CELLULAR PROCESSES AND SIGNALING: [N] Cell cycle control, cell division, chromosome partitioning, [O] Nuclear structure, [P] Defense mechanisms, [Q] Signal transduction mechanisms, [R] Cell wall/membrane/envelope biogenesis, [S] Cell motility, [T] Cytoskeleton, [U] Extracellular structures, [V] Intracellular trafficking, secretion, and vesicular transport, [W] Posttranslational modification, protein turnover, chaperones. METABOLISM: [X] Energy production and conversion, [Y] Carbohydrate transport and metabolism, [Z] Amino acid transport and metabolism, [F] Nucleotide transport and metabolism, [H] Coenzyme transport and metabolism, [L] Lipid transport and metabolism, [P] Inorganic ion transport and metabolism, [Q] Secondary metabolites biosynthesis, transport and catabolism. PODERLY CHARACTERIZED: [R] General function prediction only, [S] Function unknown.
Calpain is essential for the parasite and has a great potential for drug target. It was demonstrated that MDL 28170, a calpain inhibitor, showed a high antileishmanial activity against *L. amazonensis*.

Other interesting genes found were tuzins and amastins. Eight tuzin copies were found in the *L. amazonensis* genome with a moderate diversity. For comparison, *L. mexicana* and *L. tarentolae* have 4 copies, and *L. infantum* 6, and *L. major* the highest diversity with 28 copies. Among the tuzin copies in *L. amazonensis*, one copy forms an ortholog group only with *L. mexicana* (Table 4), given the fact that *L. amazonensis* and *L. mexicana* belong to the same taxonomic complex. Amastins belong to a large family of surface proteins unique to the parasite.

---

**Table 5.** Intergenomic NISEs, their official enzyme names, sequences IDs, Uniprot IDs for human sequences, PDB structures and the identity for each sequence.

| EC         | ENZYME NAME (OFFICIAL)                   | ORGANISM          | SEQUENCES IDs (*) | UNIPROT ACCESS | PDB [BEST HIT] [**] | IDENTITY [PDB] |
|------------|-----------------------------------------|-------------------|-------------------|----------------|----------------------|----------------|
| 1.1.1.2    | Alcohol dehydrogenase (NADP(+)          | *L. amazonensis*  | LAJMNGS050H11.b.7960 | N/A            | 1UUF                 | 160/332 (48%)  |
|            |                                         | *H. sapiens*      | hsa:10327         | P14550         | 2ALR                 | Structure solved |
| 1.3.1.34   | 2,4-dienoyl-CoA reductase (NADPH)       | *L. amazonensis*  | LAJMNGS010C07.b.1806 | N/A            | 1PS9                 | 294/730 (40%)  |
|            |                                         | *H. sapiens*      | LAJMNGS024B09.b.4107 | N/A            |                      | 198/658 (30%)  |
|            |                                         | hsa:1666          | Q16698            | 1W6U           | Structure solved     |                |
|            |                                         | hsa:26063         | Q0NU11            | 4FC6           | Structure solved     |                |
| 1.3.1.74   | 2-alkenal reductase                     | *L. amazonensis*  | LAJMNGS036G08.b.6014 | N/A            | 4GBY                 | 139/482 (29%)  |
|            |                                         | *H. sapiens*      | hsa:22949         | Q14914         | 12SV (+)             | Structure solved |
| 2.7.4.2    | Phosphomevalonate kinase                 | *L. amazonensis*  | LAJMNGS005E09.b.95 | N/A            | N/A                  | N/A            |
|            |                                         | *H. sapiens*      | hsa:10654         | Q15126         | 3CH4                 |                |
|            |                                         |                    | Q8FGV9            |                | Structure solved     |                |
| 3.11.1.2   | Exodeoxyribonuclease III (Predicted NISE)| *L. amazonensis*  | LAJMNGS001G08.b.166 | N/A            | N/A                  | N/A            |
|            |                                         | *H. sapiens*      | hsa:5810          | O60671         | 3G65 (+)             | Structure solved |
|            |                                         |                    | hsa:5883          | Q99638         | 3GGR (+)             | Structure solved |
|            |                                         |                    | hsa:11219         | Q9BS50         | 1Y97                 | Structure solved |
|            |                                         |                    | hsa:11277         | Q9NSU2         | 3U6F                 | 178/304 (59%)  |
|            |                                         |                    |                    | Q5TZT0         |                      |                |
| 5.3.3.2    | Isopentenyl-diphosphate Delta-isomerase  | *L. amazonensis*  | LAJMNGS034G09.b.5743 | N/A            | 2ZRU                 | 118/352 (34%)  |
|            |                                         | *H. sapiens*      | hsa:91734         | Q9BX5         | 2PNN                 | Structure solved |
|            |                                         |                    | hsa:3422          | Q13907         | 2ICJ                 | Structure solved |

Notes: (*) The sequences IDs from *H. sapiens* are from KEGG database. (**) The (+) signal on "PDB [Best hit]" column represent that there are more structures solved for this sequence.

---

**Table 6.** Intragenomic NISEs, their official enzyme names, sequences IDs, PDB structures identified and the identity for each sequence.

| EC         | ENZYME NAME (OFFICIAL)                  | ORIGINAL ANNOTATION                  | SEQUENCES IDs: *L. amazonensis* (*) | PDB [BEST HIT] | IDENTITY [PDB] |
|------------|----------------------------------------|--------------------------------------|--------------------------------------|----------------|----------------|
| 4.2.1.1    | Carbonate dehydratase                  | carbonic anhydrase-like protein       | LAJMNGS019E05.b.3366                | 4G7A           | 53/164 (32%)   |
|            | Carbonate dehydratase                  | carbonic anhydrase family protein, putative | LAJMNGS035D05.b.5816               | 116O           | 97/229 (42%)   |
| 4.2.99.18  | DNA-(apurinic or apyrimidinic site) lyase| endonuclease III, putative            | LAJMNGS002A05.b.218 (2)             | 1P59           | 66/194 (34%)   |
|            | DNA-(apurinic or apyrimidinic site) lyase| endonuclease/exonuclease protein-like protein | LAJMNGS041H02.b.6678 (2)             | 2ISI           | 37/106 (35%)   |
| 5.4.2.1    | Phosphoglycerate mutase                | phosphoglycerate mutase protein, putative | LAJMNGS013E01.b.2299                | 4J5            | 45/152 (30%)   |
|            | Phosphoglycerate mutase                | 2,3-bisphosphoglycerate-independent phosphoglycerate mutase, 2,3-bisphosphoglycerate-independent phosphoglyceratemutase | LAJMNGS025H05.b.4375 (2) | 3IGY | 497/552 (90%) |

Notes: (*) The numbers between parenthesis on "Sequences IDs: *L. amazonensis*" column, represent the number of copies of this enzyme.
to kinetoplastids, which are expressed specifically in the amastigote stage of the parasite. Among the 14 amastin copies found in *L. amazonensis*, 10 of them are found in orthologous genes shared with the other 5 analyzed species (*L. braziliensis*, *L. infantum*, *L. major*, *L. mexicana*, and *L. donovani*), while only 1 copy is shared between Leishmania subgenera. *L. mexicana* shares two amastin genes exclusively with *L. amazonensis*, which has a total of 28 genes of this family. These results are expected because amastin family has four subfamilies, among which we found some copies more conserved and other more divergent, explaining the fact that we found some copies shared among LCP, or only within the mexicana complex, that may be more specific subfamilies.

Figure 6. Phylogenomics analysis tree for all six *Leishmania* species (in red) and for other 28 protozoa species, inferred by Maximum Likelihood with 1,000 bootstrap replicates, based on thirty-one universal orthologous (UO) genes. Name and legend of the 34 species: *Angomonas deanei* (A deanei), *Strigomonas culicis* (S culicis), *Leishmania amazonensis* (L amazonensis), *Leishmania braziliensis* (L braziliensis), *Leishmania donovani* (L donovani), *Leishmania infantum* (L infantum), *Leishmania major* (L major), *Leishmania mexicana* (L mexicana), *Trypanosoma brucei* (T brucei), *Trypanosoma cruzi* (T cruzi), *Trypanosoma vivax* (T vivax), *Giardia lamblia* (G lamblia), *Naegleria gruberi* (N gruberi), *Dictyosteliida* spp.: *Dictyostelium discoideum* and *Polysphondylium pallidum*. *Trichomonas vaginalis* (T vaginalis), *Entamoeba* spp.: *Entamoeba dispar*, *Entamoeba histolytica* and *Entamoeba invadens*. *Tetrahymena thermophila* (T thermophila), *Plasmodium* spp.: *Plasmodium berghei*, *Plasmodium cynomolgi*, *Plasmodium falciparum*, *Plasmodium knowlesi* and *Plasmodium vivax*. *Coccids* spp.: *Cryptosporidium muris*, *Neospora caninum* and *Toxoplasma gondii*. *Piroplasmids* spp.: *Babesia bovis*, *Babesia equi*, *Babesia microti*, *Theileria annulata*, *Theileria orientalis* and *Theileria parva*. 

Despite the fact that *L. amazonensis* and *L. mexicana* belong to the same complex, they show differentiated epidemiology. It is interesting to note the presence of an amastin that could be used as a marker for the VL, shared only by these two species. It is known that *L. amazonensis* rarely causes VL, while *L. mexicana* can visceralize. Rogers and collaborators found a unique gene in *L. mexicana* that encodes a protein of unknown function that contains a predicted kelch actin binding domain (Pfam: PF01344). In our work, we found a hypothetical protein shared only by *L. amazonensis* and *L. mexicana*, which also contains a predicted kelch actin binding domain, with the same Pfam (PF01344) mapped, reinforcing the proximity between these two species, since these proteins were not found in the remaining four *Leishmania* species analyzed in this study.

*L. braziliensis* presented the highest number of paralogous genes (15), similar to the results of Peacock and colleagues and Rogers and colleagues. Genes related to telomerase
activity and transposons were found such as TATE DNA transposon, SLACS-like gene retrotransposon element, which we know are unique compared with the other five species of *Leishmania* examined, including the recently sequenced *L. amazonensis* presented in our study. Another notable difference is that *L. braziliensis* contains a functional putative RNAi pathway, absent in *L. major*, *L. tarentolae*, and *L. amazonensis* (Fig. 11). We also found some highly divergent copies of surface protein in *L. braziliensis*, not shared with the other *Leishmania* species analyzed, such as GP63, amastin, and surface antigen-like protein, corroborating previous studies. It is known that GP63 protein is involved in *Leishmania* virulence, and its function is host cell binding, conferring parasite protection from complement-mediated lysis. Interestingly, some studies showed that GP63 is under positive selection, and this incentive for changes may contribute to the functional variations of GP63 protease. It has been also described that GP63 is encoded by repeated gene cluster that seems to be enlarged fourfold in *L. braziliensis* compared with the Old World *Leishmania*. *L. braziliensis* has 39 genes encoding GP63, while in *L. amazonensis* and *L. mexicana* only 7 genes were found. Curiously, even adding the previously published proteome of *L. donovani* and the newly generated *L. amazonensis*, unique genes in *L. braziliensis* remained, although *L. amazonensis* and *L. braziliensis* have a similar geographical distribution. Only the distribution is similar in these species, once they have different vectors, have different clinical manifestations, and belong to different subgenera. This corroborates the similarity results between studies, besides being the most divergent species in these studies.

Figure 7. Comparative analysis of species *Leishmania* using orthologous and paralogous protein groups generated by OrthoMCL. The colors represent the number of protein shared between the species. blue (intern paralogous into specie); green orthologous groups between 2 species (*L. amazonensis* and *L. mexicana*; 18; *L. amazonensis* and *L. donovani*: 15; *L. amazonensis* and *L. braziliensis*; 9; *L. amazonensis* and *L. major*: 4; *L. amazonensis* and *L. infantum*: 1); and red: 7026 orthologous groups shared between all six *Leishmania* species. Orthologous groups shared between 3, 4 and 5 species are yellow.

Table 7. RNAi pathway related sequences in *L. amazonensis*.

| Piwi-AGO     | LAJMNGS037G03.b.6124 |
|--------------|---------------------|
| Tudor        | LAJMNGS051A10.b.7989 |
| ERI-1        | LAJMNGS009D01.b.1653 |
|              | LAJMNGS023D01.b.3956 |
|              | LAJMNGS034E11.b.5717 |
|              | LAJMNGS035F02.b.5853 |
| DEAD-Box RNA helicase | LAJMNGS002E10.b.336 |
|              | LAJMNGS005H09.b.1043 |
|              | LAJMNGS016A07.b.2785 |
|              | LAJMNGS018H11.b.3270 |
|              | LAJMNGS021E12.b.3675 |
|              | LAJMNGS024C05.b.4124 |
|              | LAJMNGS042E06.b.6755 |
|              | LAJMNGS045F10.b.7200 |
|              | LAJMNGS046A05.b.7244 |
| Ribonuclease III/Dicer | LAJMNGS020H09.b.3587 |
|              | LAJMNGS021A10.b.3613 |
divergent inside the genus. Since no comparative studies were carried out between L. infantum and L. donovani with the aim of identifying unique genes, the number of unique L. infantum genes (26) may be underestimated. However, in our study the analysis of these two closely related species showed that they share 25 orthologs, and only one highly divergent paralog, amastin like, was found in L. infantum. The fact that L. infantum belongs to the same complex of L. donovani and they share 7,619 orthologous groups (93.5% of the L. infantum proteome), while with L. braziliensis it shares 7,401 (90.8% of its proteome) orthologs, could explain this scenario. Similarly, in L. amazonensis, only one inparalog found is probably because it belongs to the same L. mexicana complex and is very close to this species. For example, L. amazonensis and L. mexicana share several orthologous, 7,380 in total (85.8%), whereas L. amazonensis and L. braziliensis share 7,162 groups (83.3% of L.amazonensis proteome). As expected, we noted that the closer species have greater number of genes shared by them. This is especially true when comparing species within the same complex that shows the higher number of shared genes. Some large gene families present in L. amazonensis may have only one conserved domain in common. The remaining of their sequences is so divergent that subfamilies or classes are identified. As an example, we can mention amastin, which is found in all the six Leishmania species analyzed that present the signature C-[IVLYF]-[TS]-[LF]-[WF]-G-[KRQ]-X-[DENT]-C; however, some amastin genes are so divergent that they can be classified into four subfamilies or classes: α, β, γ, and δ.

**Phylogenomics**
Mauricio et al. used the gene mspC3 as a marker to reconstruct a phylogeny of species of Leishmania subgenus, and the results were very close to the ones found in this study, keeping L. infantum and L. donovani in the same branch, with L. major and L. mexicana more distant, which was expected since these species, L. infantum and L. donovani, belong to the same complex. However, in this study of Mauricio and colleagues, it was not possible to observe the separation of the subgenus Leishmania in New and Old World species. Differently, Simpson and

---

**Figure 8.** In green area, a total of 2,483 L. amazonensis proteins identified by both Conserved Domains Database (RpsBlast-CDD) and Protein Families Database (HMMER-PFam). In lateral tables we visualize most frequent Families (Pfam) and Domains (CDD). 269 proteins were identified only by Protein Families Database, and inside yellow area we show 10 most frequent families found by Pfam. A total of 2,634 proteins were identified only by Conserved Domains Database (RpsBlast-CDD) and Protein Families (Pfam) databases. 269 proteins were identified only by CDD 2,634 proteins identified by both Conserved Domains Database (RpsBlast-CDD) and Protein Families (Pfam) databases. In blue area the 10 most common domains assigned by CDD in L. amazonensis.
colleagues achieved a reasonable and consistent separation of the subgenera Viannia and Leishmania using HSP70 genes, as well as the division of the Old and New World Leishmania species inside Leishmania subgenus. Mauricio and colleagues using the GP63 gene, which is a multicopy gene under positive selection, achieved a good separation between the subgenus and could identify those species originating from the New or Old World, although depending on which copy of the gene is used for classification, the results may be distinct.

L. donovani complex formation (L. donovani and L. infantum) was constant in the majority of studies, as well as the formation of L. mexicana complex, besides the correct separation of subgenus, and within the Leishmania subgenus the separation between Old and New World species.

Fraga and colleagues used the HSP20 gene to separate the Old/New World Leishmania subgenus, with a bootstrap of 89. However, when HSP20 and HSP70 genes were concatenated, the bootstrap support value of this separation improved up to 99 and 100 to support the division of the subgenus. This example demonstrates the advantage of concatenating genes to infer phylogenomic-based species trees. Our approach of species tree by using 31 UO genes and phylogenomic-based approach was robust showing a bootstrap support of 100 for the Kinetoplastida clade.

This analysis showed the expected separation of this genus for all six species analyzed: L. (V) braziliensis as outgroup, and L. (L.) infantum and L. (L.) donovani very close, reflecting the complex formed by them. As expected, this complex is closer to L. (L.) major recapitulating phylogeny of the Old World species inside Leishmania subgenus. Nevertheless, L. mexicana and L. amazonensis are placed together on the same clade, reflecting the Mexicana complex, corroborating classical phylogeny. It should be noted that the bootstrap values were higher than that observed in other works (bootstrap value 100), although the taxonomic position of these species remained mostly the same.

Although the L. amazonensis taxonomic position is already known, the phylogenomic species tree obtained using 31 UO genes proved to be a good approach for robust species tree inference using multiple genes, and also a good option to avoid the bias of extrapolating single-gene phylogenies. Another interesting point was that our phylogenomic tree recapitulated the Kinetoplastida monophyly and its correct separation.

Intergenomic and Intragenomic NISE as Possible Drug Targets

This work identified a set of NISE (also known as analogous enzymes) between L. amazonensis and H. sapiens (Supple-
proteome (intragenomic NISE) could provide further information about other enzymatic activities is scarce, particularly when considering trypanosomatids. An example is 2-alkenal reductase (EC: 1.3.1.74). A defensive role has been shown for this enzyme in some plants, apparently by protecting them from oxidative stress by catalyzing the reduction of reactive carbonyls, but no information about its biological role has been found for trypanosomatids.

On the other hand, the identification of NISEs inside L. amazonensis proteome (intragenomic NISE) could provide new insights about alternative biochemical pathways and the meaning of functional redundancy inside a genome. Among the NISEs inside L. amazonensis proteome (Tables 6 and S3), carbonate dehydratase (EC 4.2.1.1), DNA-(apurinic or apyrimidinic site) lyase (EC 4.2.99.18), and phosphoglycerate mutase (PGAM) (EC 5.4.2.1) could be proposed as potential drug targets. Carbonate dehydratase catalyzes the interconversion of CO$_2$ and HCO$_3^-$: This enzymatic function is present in animals, plants, yeast, archaea, bacteria, and parasites.

Studies have proposed this enzyme as a candidate drug target in P. falciparum since the inhibition of this enzyme affects the pathway of pyrimidine biosynthesis, DNA-(apurinic or apyrimidinic site) lyase is involved in the repair of abasic sites caused by oxidative stress and external agents (chemical or physical), with spontaneous hydrolysis resulting in purine or pyrimidine loss. PGAM catalyzes the interconversion of 2-phosphoglycerate (2PG) and 3-phosphoglycerate (3PG) in the glycolytic and gluconeogenic pathways. PGAM was structurally characterized in L. mexicana, and has been proposed as a possible drug target, since the enzymatic form in the parasite is structurally different from the host and has different properties, an earlier example of analogy found by an experimental approach.

An integrative approach will be employed in the future to obtain a broader understanding of the biological role of the intergenomic and intragenomic NISE detected in this work.
Figure 11. Structural comparison of selected intergenomic NISE cases between *L. amazonensis* and Human. Top panel (EC 1.1.1.2): A LAJMNGS050H11.b.7960 “putative NADP-dependent alcohol dehydrogenase” from *L. amazonensis* (A) and human Aldehyde reductase (PDB 2ALR) (B). Middle panel (EC 1.3.1.34): LAJMNGS010C07.b.1806 “putative 2,4-dienoyl-coa reductase FADH1” from *L. amazonensis* (C) and human mitochondrial 2,4-dienoyl-CoA reductase (PDB 1W6U) (D). Bottom panel (EC 5.3.3.2): LAJMNGS034G09.b.5743 “putative isopentenyl-diphosphate delta-isomerase” from *L. amazonensis* (E) and human Isopentenyl-diphosphate Delta-isomerase (PDB 2ICK) (F). Models for all proteins are presented as ribbons. Parasite proteins are colored by secondary structure and presented superposed on their templates (gray ribbons) used in homology modeling. Human analogs are colored by secondary structure, except for 1W6U, which is colored by chain and presented superposed on the peroxisomal isoform (PDB ) shown as gray ribbons. The insets show details of the proposed catalytic residues and co-factors for each analogous enzyme. Residues colored blue belong to the parasite enzymes while residues from human analogs are color-coded by atom type.
RNAi Machinery

One of the first organisms where functional RNAi pathway was described was *T. brucei*. Since then, several trypanosomatids were subject to RNAi characterization through direct analysis or genome sequencing. RNA silencing pathways play critical roles in gene regulation, virus infection, and transposon control. RNAi is mediated by small interfering RNAs (siRNAs), which are liberated from double-stranded (ds) RNA precursors by Dicer and guide the RISC to degenerate sequence–specific mRNA targets. Phylogenetic analysis suggests the presence of the RNAi pathway in the last common ancestor of eukaryotes with putative important role in defense responses against genomic parasites such as transposable elements and viruses.

RNAi pathway-related genes present in different trypanosomatids were used to identify orthologous genes in *L. amazonensis* genome (Table 7). A key step of RNAi pathway is the Dicer activity, which converts dsRNA into siRNA. Dicer was identified in *T. brucei* (Tb927.8.2370) and a protein with a similar architecture domain, bearing the two RNAse III-like domains, and was characterized in *L. braziliensis* (LbrM23_V2.0390). Such proteins seem to be missing in trypanosomatids that lack a functional RNAi pathway, such as *T. cruzi* and *L. major*. Genomic analysis of *L. infantum*, *L. braziliensis*, and *L. major* has demonstrated the presence of Dicer only in *L. braziliensis* and, otherwise, shows synteny for the other Leishmania species. We were unable to detect Dicer in *L. amazonensis* genome or in any sequence bearing the characteristic Rnc (dsRNA-specific ribonuclease) domain of *L. braziliensis* putative Dicer gene. Since Dicer activity might be performed by a combination of different proteins bearing typical RNAi domains such as DEAD-box RNA helicase and Ribonuclease III, such domains were subject of analysis in *L. amazonensis* genome data set. Nine DEAD/H box RNA helicase and two Ribonuclease III were identified in *L. amazonensis*, with putative relationship to RNAi pathway (Table 7). Although Dicer was not identified, some Dicer-related genes were characterized. ERI proteins are another important components of RNAi pathway involved in the formation of the ERI/DICER complex. We were able to identify four ERI sequences in *L. amazonensis* genome data set (LAJMNGS009D01.b.1653, LAJMNGS023D01.b.3956, LAJMNGS034E11.b.5717, and LAJMNGS035F02.b.5853) (Table 7). Two genes of the RISC (a major effector complex of the RNAi pathway) were also identified: tudor and piwi (argonaute family) (Table 7). Several argonaute family genes have been described in trypanosomatids. In *T. brucei*, two argonaute-like genes were identified (TbAGO1 and TbPW1). Both forms are expressed in the procyclic culture stages but only TbAGO1 is involved in RNAi. Several previous data have demonstrated the presence of RNAi key genes argonaute and/or Dicer in *Leishmania* subgenus *Viannia* (*L. braziliensis*, *L. guyanensis*, and *L. panamensis*) but not in the subgenus *Leishmania* (*L. mexicana*, *L. major*, and *L. donovani*).

Here we describe the first evidence through genomic analysis of RNAi pathway absence in *L. amazonensis*. So far, experimental evidences pointed out the absence of a functional RNAi pathway in the whole subgenus *Leishmania*, corroborated by the analysis of *L. amazonensis* genome data set. Several arguments have been elegantly raised by Lye et al. in an attempt to understand this phenomena; they describe the viral infections, genome plasticity, and phenotype selection as the major players of RNAi lost event. The identified sequences related to RNAi pathway in *L. amazonensis* might reflect the remains of an erstwhile ancient functional RNAi pathway. Hypothetically, the remaining functional genes might be present because of an association with different pathways required for parasite survival. It might be the case of ERI sequences where its dual role in rRNA processing and RNA may have prevented its loss.

Comparative genome analysis shows that, most likely, the last common eukaryote possesses two copies of argonaute-related genes, suggesting the presence of two distinct silencing machineries. The argonaute-like proteins had possibly been involved in transcriptional regulation by targeting RNA in cytoplasm, while piwi-like proteins would act in nucleus targeting transposons. In contrast to most eukaryotes, in which the argonaute duplication followed by functional diversification is common, trypanosomatids have no more than one copy of each argonaute-like genes (ago and piwi) per species. Indeed, trypanosomatids with functional RNAi (*T. brucei, L. braziliensis, L. guyanensis*) have both genes; however, species with non functional RNAi pathway (*T. cruzi, L. amazonensis, L. major, L. mexicana*) possesses only the piwi version of the argonaute family. The main difference in the protein domain architecture between the two argonaute families is the lack of a PAZ domain in piwi-like proteins. The PAZ domain consists of two subdomains, with a oligonucleotide/oligosaccharide binding region which is responsible for 3’ ends ssRNA recognition typically found in 3’ overhangs of the siRNAs. In early work on RNAi characterization in
trypanosomatids, two argonaute–like genes were identified in *T. brucei* termed TbAGO1 and TbPW1. After functional analysis, the authors showed that TbAGO1, but not TbPW1, was involved in RNAi. *L. amazonensis* does not have the ago-like gene, and the piwi-gene (LaPW1) is homologous to TbPW1, with orthologs group in subgenus *Leishmania*. Recently, Padmanabhan et al. identified putative functions for piwi-like gene in *L. infantum* and *L. major*. Similar to *T. brucei*, *Leishmania* piwi-like protein is neither related to RNAi pathway nor to siRNA biogenesis. Piwi-like gene is expressed in both parasite forms, but piwi mutation affects the amastigote infection delaying the pathology and increasing apoptosis susceptibility. The authors raised the hypothesis about piwi-like protein role: located in the parasite single mitochondrion, it might act as an apoptotic sensor.

The absence of post-transcriptional control of the RNAi might help to explain also the differences observed among the *Leishmania* and *Viannia* subgenera related with pathogenicity in mammalian host, insect vector relationship, and distinct surface glyocalyx structure.

**Synteny Analysis**

Comparisons of the *L. mexicana* and *L. amazonensis* genomes revealed that more than 99.5% of the genes were syntenic, as expected, since these two species were very close and belong to the same complex. In fact, previous studies have described that the closer the species, the higher the degree of synteny between their genomes. For instance, the work of El-Sayed et al. showed that 94% of the genes (6,200 genes) that are conserved among the TriTryps are also syntenic, in spite of the fact that these three species are not as close as *Leishmania* spp. In another study, the analysis of the three related *Leishmania* species (*L. major*, *L. infantum* and *L. braziliensis*) revealed that more than 99% of the genes are syntenic among these species. Likewise, the comparison of *L. tarentolae* to the three sequenced pathogenic *Leishmania* species showed that these four species are highly syntenic.

**Conclusions**

The *L. amazonensis* genome assembly resulted in approximately 29 million base pairs. The smallest contig had 96 bases and the largest 141,211 bases. The annotation resulted in 8,802 CDS, where the largest coding regions had 19,872 bases and the smallest 66 bases, with a median and mean value of 1,637 and 1,209 bp, respectively. Of these *L. amazonensis* CDS, 63.1% (5,554/8,802) were annotated as “hypothetic protein” and 79.71% (7,016/8,802) were grouped into *Leishmania* spp. core proteome. Our work is the first to propose a *Leishmania* spp. core proteome using the six sequenced human-pathogenic *Leishmania*. Generally, the following housekeeping proteins were found within *Leishmania* spp. core proteome (LCP): 40S ribosomal protein S16, RNA helicase, protein kinase, dynein heavy chain, activated protein kinase c receptor (LACK), ABC transporter, calpain-like cysteine peptidase, and DNA primase. These LCP genes can be potentially explored as molecular markers, either for diagnosis or for the genotyping of *Leishmania* populations. Furthermore, some genes related to membrane surface were found: GP63, amastin, and tuzin. *L. amazonensis* and *L. mexicana* showed the largest number of specific shared orthologs (18), most of them without a defined function. However, divergent amastin-like protein and viscerotropic leishmaniasis antigen were found as an ortholog only between these two species, and it may be possible to use these as a complex marker. The specific *L. amazonensis* and *L. mexicana* orthologs are potential specific “mexicana complex” markers, since they are unique to these species. The orphans genes found can possibly be explored as markers for species-specific diagnosis, once they are uniquely present in these species. Our original phylogenetic tree confirmed the position of *L. amazonensis* as closer to *L. mexicana* and belonging to the New World *Leishmania* subgenus. In addition, RNAi pathway in *L. amazonensis* is likely to be not functional since key genes are missing in its genome. Finally, we present new information regarding the NISE search in *L. amazonensis* genome. The NISE search resulted in 25 potential analogous between *L. amazonensis* and *H. sapiens*. Also, 31 potential analogous were found in *L. amazonensis* protein sequences. Five out of the six main EC classes showed potential NISEs: Oxidoreductases (EC 1), Transferases (EC 2), Hydrolases (EC 3), Lyases (EC 4), and Isomerases (EC 5). These NISE findings are new and represent potential drug targets because analogous proteins perform the same function using different proteins and 3D structures. In other words, an analogous protein in *L. amazonensis* can be silenced without affecting the host.

**Acknowledgments**

We would like to thank Prof. Jeffrey Shaw, Dr. Aldina Barral, Dr. Elisa Cupolillo, Prof. Elvira Saraiva, Prof. Bartira Rossi Bergmann, and Prof. Salvatore Giovanni de Simone for participation in the choice of the *L. amazonensis* strain for this study.

**Author Contributions**

Generated and validated the genome sequence: NJD, JCM. Assembled the genome: JR. Analyzed the data: DAT, GLN, RJ, JL, ASRD, MRG, LMP, DRL, PHS, AP, FPS. Wrote the first draft of the manuscript: DAT, GLN. Contributed to the writing of the manuscript: MRG, LMP, AP, HLMG, CMP, ECG, FPS. Jointly developed the structure and arguments for the paper: DAT, HLMG, ABM, FPS, AMRD. Made critical revisions and approved final version: DAT, GLN, RJ, JL, ASRD, MRG, LMP, DRL, PHS, HLMG, ABM, JR, AP, FPS, CMP, ECG, AMRD, NJD, JCM. All authors reviewed and approved the final manuscript.

**Supplementary Material**

Table S1. OrthoMCL identifier for each LCP ortholog, identifier for protein into LCP Ortholog, protein function
and species name for each sequence. Complete table of the orthologs shared only between *L. amazonensis* and the other five *Leishmania* species.

### Table S2. Intragenomic NISEs, original annotation and official enzyme names, fold and superfamilies classification based on SUPERFAMILY database and the function of each enzyme.

### Table S3. Intragenomic NISEs, original annotation and official enzyme names, fold and superfamilies classification based on SUPERFAMILY database and the function of each enzyme.

### Table S4. Total of enzymatic activities with NISE cases detected by our methodology.

### Table S5. Intragenomic NISE as potential drug target searched among three drug target databases (TDR target, TTD, DrugBank).

### REFERENCES

1. Choi J, El-Sayed NM. Functional genomics of trypanosomatids. *Parasite Immunol.* 2012;34(2):3–72.
2. Desjeux P. *Leishmaniasis*. Public health aspects and control. *Clin Dermatol.* 1996;14(5):417–23.
3. WHO. *World Health Statistics*. 2010. Available at http://www.who.int/whosis/whostat/2010/en/. Accessed April 8, 2013.
4. Passos VM, Fernandes O, Luceda PA, et al. *Leishmania (Viannia) braziliensis* in the predominant species infecting patients with American cutaneous leishmaniasis in the State of Minas Gerais, Southeast Brazil. *Acta Trop.* 1999;72(3):251–8.
5. TDR. *Leishmaniasis*. 2013. Available at http://www.who.int/trd/diseases-topics/leishmaniasis/en/index.html. Accessed April 11, 2013.
6. Barrau A, Pont-Sampayo D, Grimaldi Júnior G, et al. *Leishmania* in Bahia, Brazil: evidence that *Leishmania amazonensis* produces a wide spectrum of clinical disease. *Am J Trop Med Hyg.* 1991;44(5):536–46.
7. Grimaldi G, McMahan-Pratt D. *Leishmania* and its etiologic agents in the New World: an overview. *Prog Clin Parasitol*. 1991;27:113–118.
8. Weige K, Saravia NG. Natural history, clinical evolution, and the host-parasite interaction in New World cutaneous leishmaniasis. *Clin Dermatol*. 1996;14(5):433–50.
9. Chappuis F, Sundar S, Hailu A, et al. Visceral leishmaniasis: what are the needs for diagnosis, treatment and control? *Nat Rev Microbiol*. 2007;5(11):873–82.
10. Akloov OE, Khachemoune A, Hassan T. Clinical manifestations and classification of Old World cutaneous leishmaniasis. *Int J Dermatol*. 2007;46(2):132–42.
11. Desjeux P. *Leishmaniasis*: current situation and new perspectives. *Comp Immunol Microbiol Infect Dis*. 2004;27(5):305–18.
12. Marsden PD. *Mucosal leishmaniasis* (*espundia* Espumol, 1911). *Tracts R Soc Med*. 1986:80(6):859–76.
13. Cupollo SMN, Souza CSF, Abreu-Silva AL, Calabrese KS. *Goncalves da Passos* et al. *Leishmania (Viannia) braziliensis* genome by shotgun sequencing. *J Comp Pathol*. 1996;14(5):417–23.

22. Real F, Vidal RO, Carrazzolle MF, et al. The genome sequence of *Leishmania (Leishmania) amazonensis*: functional annotation and extended analysis of gene models. *DNA Res.* 2013.
23. El-Sayed NM, Myler PJ, Blandin G, et al. Comparative genomics of trypanosomatid parasite proteomes. *Science*. 2005;309(5733):404–9.
24. Teixeira SM, de Paiva RMC, Kangasuo-Marcolino MM, Darocho WA. *Trypanosoma* comparative genomics: contributions to the study of parasite biology and different parasitic diseases. *Genet Mol Biol*. 2012;35(1):1–17.
25. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* 1998;8(3):186–94.
26. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* 1998;8(3):175–85.
27. Zerboino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 2008;18(5):821–9.
28. Gläuber Wagner, Rodrigo Jardim, Diogo A Tschoeke, et al. STINGRAY: system for integrated genomic resources and analysis. *BMCR Research Notes*. 2014;17:122.
29. Dávila AM, Lorenzini DM, Mendes PN, et al. GARS: genomic analysis resources for sequence annotation. *Bioinformatics*. 2005;21(23):4302–3.
30. Atrshul SF, Maddie TD, Shaaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997;25(17):3389–402.
31. Bateman A, Birney E, Cerutti L, et al. The Pfam protein families database. *Nucleic Acids Res.* 2002;30(1):71–80.
32. Punta M, Coggill PC, Eberhardt RY, et al. The Pfam protein families database. *Nucleic Acids Res.* 2012;40(Database issue):D290–301.
33. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* 2013;41(12):e121.
34. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 2000;25(1):25–9.
35. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 2003;13(9):2178–87.
36. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2013 ISBN 3-900051-07-0.
37. Ocaña KACS, Dávila AMR. Phylogenomics-based reconstruction of protozoan species tree. *Evol Bioinform Online*. 2011;7:107–21.
38. Ciccacci FD, Doerk T, van Mering C, Creevey CJ, Snel B, Bork P. Toward automatic reconstruction of a highly resolved tree of life. *Science*. 2006;311(5765):1283–7.
39. Dávila AM, Mendes PN, Wagner G, et al. *Protodba*: dynamic visualization and exploration of protozoan genomes. *Nucleic Acids Res.* 2008;36(Database issue):D547–52.
40. Katoh K, Kuma K, Tokh H, Miyata T. MAFFT version 6: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* 2005;33(2):518–525.
41. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol*. 2011;28(10):2731–9.
42. Jones DT, Taylor WR, Thornton JM. The rapid generation of multiple alignment databases from protein sequences. *Comput Appl Biosci*. 1992;8:275–82.
43. Ores TD, Guimaraes ACR, Degreame WV, de Miranda AB. *AnEnPi*: identification and annotation of analogous enzymes. *MBioinformatics*. 2008;9:544.
44. Arves-Ferreira M, Guimaraes ACR, Capriles PV, Dardenne LE, Degreame WV. A new approach for potential drug target discovery through in silico metabolic pathway analysis using *Trypanosoma cruzi* genome information. *Mem Inst Oswaldo Cruz*. 2009;104(1):109–10.
45. Gomes MR, Guimaraes ACR, de Miranda AB. Specific and nonhomologous isofunctional enzymes of the genetic information processing pathways as potential therapeutic targets for trypanosomatid parasites. *Enzyme Res.* 2011;2013:543912.
46. Barrett AJ. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme Nomenclature. Recommendations 1992. *Supplement 4: corrections and additions* (1997). *Eur J Biochem*. 1997;250(1):1–6.
47. Kaneshia M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 2012;40(Database issue):D69–72.
48. Capriles PVSZ, Guimaraes ACR, Otro TD, Miranda AB, Dardenne LE, Degreame WV. Structural modelling and comparative analysis of homologous, analogous and specific proteins from *Trypanosoma cruzi* versus *Homo sapiens* putative drug targets for Chagas disease treatment. *BMG Genomics*. 2010;11:610.
49. Andreeva A, Howorth D, Chandonia JM, et al. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.* 2008;36(Database issue):D419–25.
50. Wilson D, Pethica R, Zhou Y, et al. SUPERFAMILY – sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.* 2009;37(Database issue):D380–6.
51. Sali A, Blindell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 1993;234(3):779–81.
52. Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. Trends Genet. 2000;16(6):276–7.
53. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science. 1997;278(5331):631–7.
54. Tatusov RL, Fedorova ND, Jackson JD, et al. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 2003;4:41.
55. Logan-Khumber PJ, De Silva N, Boehme U, et al. GeneDB – an annotation framework for pathogenic and non-pathogenic microorganisms. Nucleic Acids Res. 2012;40(Database issue):D398–108.
56. Assella S, Krane TM, Otto TD, Newbold C, Berriman M. ABCAS: algo-rithm-based automatic construction of assembled sequences. Bioinformatics. 2009;25(15):1968–9.
57. Carver T, Rutherford KM, Berriman M, Rajandream M-A, Burrell BG, Parkhill J. ACv3: the articulated comparative view. Nucleic Acids Res. 2005;33(Database issue):D422–3.
58. Saurin W, Hofnung M, Dassa E. Getting in or out: early segregation between importers and exporters in the evolution of ATP-binding cassette (ABC) transporters. J Mol Evol. 1999;48(1):21–44.
59. Sauvage V, Aubert D, Escotte-Bianet S, Viliena I. The role of ATP-binding cassette (ABC) proteins in protozoan parasites. Mol Biochem Parasitol. 2009;167(2):81–94.
60. Leprohon P, Légaré D, Giraud L, Papadopoulou B, Ouellette M. Modulation of Leishmania ABC protein gene expression through life stages and among drug-resistant parasites. Eukaryot Cell. 2006;5(10):1713–25.
61. Torres DC, Adawi V, Ribeiro-Alves M, et al. Targeted gene expression profiling in Leishmania braziliensis and Leishmania guyanensis parasites isolated from Brazilian patients with different antimonial treatment outcomes. Infect Genet Evol. 2010;10(6):727–33.
62. Mottram JC, Coombs GH, Alexander J. Cysteine peptides as virulence factors of Leishmania. Curr Opin Microbiol. 2004;7(6):375–8.
63. Erisch K, Barraclough T, Gull K. Evolutionary relationships and protein domain architecture in an expanded calpain superfamily in kinetoplastid parasites. J Mol Evol. 2005;61(6):742–57.
64. d’Avila-Leyva CM, Marinho FA, Santos LO, Martins JL, Santos ALS, Bezerra MH. Antileishmanial activity of MZ1, 28170, a potent calpain inhibitor. Int J Antimicrob Agents. 2006;28(2):138–42.
65. Raymond F, Boisvert S, Roy G, et al. The ABCAS: algorithm-based automatic construction of assembled sequences. Bioinformatics. 2005;21(10):1222–3.
66. Parke E, Kumar Shakyawar S, Saudagar P, Kumar Dubey V, Sundar D. Epitopic mapping of Leishmania major glycoprotein 63 using auranofin reveals a dual mechanism of enzyme inhibition. Aminos Acids. 2012;40(3):1131–47.
67. Kaslow DC, Adawi V, Ribeiro-Alves M, et al. A gold-containing drug against parasitic and bacterial pathogens. Trans R Soc Trop Med Hyg. 2006;100(7):79–82.
68. Mestra L, Lopez L, Robledo SM, Musкус CE, Nicholls RS, Vélez ID. Transmission-transmitted visceral leishmaniasis caused by Leishmania (Leishmania) mexicana in an immunocompromised patient: a case report. Transfusion. 2011;51(9):1919–23.
69. Yao C, Donelson JE, Wilson ME. The major surface protease (MSP or GP63) of Trypanosoma cruzi: antigenic properties and the importance of the protease domain in the evaulation of a vaccine. Leishmania 2003;23(4):249–56.
70. Vidal AE, Harkiolaki M, Gallego C, et al. Crystal structure and DNA repair activities of the AP endonuclease of Leishmania major. J Mol Biol. 2007;373(4):827–38.
71. Nowicki MW, Kuwaprasit B, McNae JW, et al. Crystal structures of Leishmania mexicana phosphoglycerate mutase suggest a one-metal mechanism and a new enzyme subclass. J Mol Biol. 2009;394(3):535–43.
72. Ngj H, Tschudi C, Gull K, Ullu E. Double-stranded RNA induces mRNA degradation in Trypanosoma brucei. Proc Natl Acad Sci U S A. 1998;95(25):14687–92.
73. Robinson KA, Beverley SM. Improvements in transfection efficiency and tests of the ERI/DICER complex in endogenous RNA interference and sperm development pathway in Trypanosoma brucei. Infect Genet Evol. 2003;128(2):217–28.
74. Lye LF, Owens K, Shi H, et al. Retention and loss of RNA interference pathways in trypanomatid protozoans. PLoS Pathog. 2010;6(10):e1001161.
75. Cerutti H, Cassa-Mollano JA. On the origin and functions of RNA-mediated silencing: from protists to man. Curr Genet. 2006;50(2):81–99.
76. Shi H, Tschudi C, Ullu E. An unusual Dicer-like protein fuels the RNA interference pathway in Trypanosoma brucei. RNA. 2006;12(2):2063–72.
77. Pavesic DM, Lachowicz J, Duchaine TF, Smith HE, Kennedy S. Requirement for the ER1/DICER complex in endogenous RNA interference and sperm development in Caenorhabditis elegans. Genetics. 2009;183(4):1283–95.
78. Durand-Dubief M, Bastin P, Thégo, et al. An argonaute protein required for RNA interference, is involved in mitosis and chromosome segregation in Trypanosoma brucei. Nat Genet. 2002;31(4):217–28.
79. Lexer M, Genin H, Beissbarth T, Wernery U. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 2003;4:41.
70. Grover A, Kwon YS, Kawamura K, Murray DW, Sundar D. Tumor suppressor activity of an RNA virus from the parasite Leishmania. Proc Natl Acad Sci U S A. 1989;86(15):5799–82.
71. Lunel P, Dallaire F, Callahan HA, Litaker RW, Roger AJ. Early evolution within kinetoplastids (euglenozoan), and the late emergence of trypanosomatids. Mol Biol Evol. 2004;21(5):4029–39.
72. Cunningham ML, Fairlamb AH. Trypanothione reductase from the parasite Leishmania. Purification, characterisation and inhibition by trivalent ars lamines. Eur J Biochem. 1995;230(2):460–8.
73. Widmer G, Comeau AM, Fungong DB, Wirth DF, Patterson JL. Characterization of a RNA virus from the parasite Leishmania. Proc Natl Acad Sci U S A. 2006;103(42):15785–80.
74. Fraj A, Montalvo AM, Youk D, Amezua G, Maes I, Dujardin J-C, Requena JM. Species discrimination according to the Leishmania heat-shock protein 20 gene. Infect Genet Evol. 2013;18:229–37.
75. Ilari A, Baioo F, Messiri L, et al. A gold-containing drug against parasitic polyamine metabolism: the X-ray structure of trypanothione reductase from Leishmania donovani reveals a dual mechanism of enzyme inhibition. Aminos Acids. 2012;40(3):803–11.
76. Lindahl T, Demple B, Robins P. Suicide inactivation of the E. coli 06-methylguanine-DNA methyltransferase. EMBO J. 1982;11(1):1359–63.
77. Sultana R, McNeill DR, Abbas B, et al. Synthetic lethal targeting of DNA double-strand break repair cells by human apurinic/apyrimidinic endonucl ease inhibitors. Int J Cancer. 2012;131(10):2433–44.
78. Doun SS, Burgner JW, Briggs SD, Rodwell WW. Enterococcus faecalis phospho-mevalonate kinase. Protein Sci. 2005;14(5):1134–9.
79. Liebl P, Kishimoto T, Kallenbach R, et al. The ABCAs: algorithm-based automatic construction of assembled sequences. Bioinformatics. 2005;21(10):1222–3.
80. Mauricio IL, Howard MK, Storchard JR, Miles MA. Genomic diversity in the Leishmania donovani complex. Parasitology. 1999;119(3 pt 3):217–46.
81. Simpson AGB, Gill EE, Callahan HA, Litaker RW, Roger AJ. Early evolution within kinetoplastids (euglenozoan), and the late emergence of trypanosomatids. Mol Biol Evol. 2005;22(4):609–16.
82. Cunningham ML, Fairlamb AH. Trypanothione reductase from Leishmania donovani. Purification, characterisation and inhibition by trivalent ars lamines. Eur J Biochem. 1995;230(2):460–8.