Published in final edited form as:

J Proteomics. 2014 January 31; 97: 48–61. doi:10.1016/j.jprot.2013.04.021.

Moving from unsequenced to sequenced genome: Reanalysis of the proteome of *Leishmania donovani*[☆]

Raja Sekhar Nirujogi^{a,b,1}, Harsh Pawar^{a,c,1}, Santosh Renuse^{a,d,1}, Praveen Kumar^a, Sandip Chavan^{a,e}, Gajanan Sathe^{a,e}, Jyoti Sharma^{a,e}, Sweta Khobragade^f, Janhavee Pande^f, Bhakti Modak^f, T.S. Keshava Prasad^{a,b,e}, H.C. Harsha^a, Milind S. Patole^{f,*}, and Akhilesh Pandey^{g,h,i,j,**}

^aInstitute of Bioinformatics, International Technology Park, Bangalore 560066, India

^bBioinformatics Centre, School of Life Sciences, Pondicherry University, Puducherry 605014, India

^cRajiv Gandhi University of Health Sciences, Bangalore 560041, India

^dDepartment of Biotechnology, Amrita Vishwa Vidyapeetham, Kollam 690525, India

^eManipal University, Madhav Nagar, Manipal 576104, India

^fNational Centre for Cell Sciences, Pune 411007, India

^gMcKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore 21205, MD, USA

^hDepartment of Biological Chemistry, Johns Hopkins University School of Medicine, Baltimore 21205, MD, USA

ⁱDepartment of Oncology, Johns Hopkins University School of Medicine, Baltimore 21205, MD, USA

^jDepartment of Pathology, Johns Hopkins University School of Medicine, Baltimore 21205, MD, USA

Abstract

The kinetoplastid protozoan parasite, *Leishmania donovani*, is the causative agent of kala azar or visceral leishmaniasis. Kala azar is a severe form of leishmaniasis that is fatal in the majority of untreated cases. Studies on proteomic analysis of *L. donovani* thus far have been carried out using homology-based identification based on related *Leishmania* species (*L. infantum*, *L. major* and *L. braziliensis*) whose genomes have been sequenced. Recently, the genome of *L. donovani* was fully sequenced and the data became publicly available. We took advantage of the availability of its

[☆]This article is part of a Special Issue entitled: Trends in Microbial Proteomics.

^{*}Corresponding author. Fax: +91 2025692259. ^{**}Correspondence to: A. Pandey, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA. Fax: +1 410 502 7544. patole@nccs.res.in (M.S. Patole), pandey@jhmi.edu (A. Pandey).

¹Authors have contributed equally.

Conflict of interest

All authors declare no conflict of interest.

genomic sequence to carry out a more accurate proteogenomic analysis of *L. donovani* proteome using our previously generated dataset. This resulted in identification of 17,504 unique peptides upon database-dependent search against the annotated proteins in *L. donovani*. These peptides were assigned to 3999 unique proteins in *L. donovani*. 2296 proteins were identified in both the life stages of *L. donovani*, while 613 and 1090 proteins were identified only from amastigote and promastigote stages, respectively. The proteomic data was also searched against six-frame translated *L. donovani* genome, which led to 255 genome search-specific peptides (GSSPs) resulting in identification of 20 novel genes and correction of 40 existing gene models in *L. donovani*.

Biological significance—*Leishmania donovani* genome sequencing was recently completed, which permitted us to use a proteogenomic approach to map its proteome and to carry out annotation of its genome. This resulted in mapping of 50% (3999 proteins) of *L. donovani* proteome. Our study identified 20 novel genes previously not predicted from the *L. donovani* genome in addition to correcting annotations of 40 existing gene models. The identified proteins may help in better understanding of stage-specific protein expression profiles in *L. donovani* and to identify novel stage-specific drug targets in *L. donovani* which could be used in the treatment of leishmaniasis.

Keywords

Genome annotation; Digenic parasite; Intracellular pathogen; High resolution mass spectrometry

1. Introduction

Leishmaniasis refers to a broad range of diseases caused by protozoan parasite of the genus *Leishmania* and is the second most common cause of mortality amongst various tropical infections [1]. *Leishmania* is a kinetoplastid protozoan and digenic parasite of blood feeding insects and vertebrates including humans [2]. The promastigote motile form of the parasite resides in the gut of the phlebotomine sand fly. It infects the vertebrate host when the female sand fly takes a blood meal and, once inside the host, the promastigotes are engulfed by circulatory macrophages by phagocytosis. The parasite is transformed into a non-motile amastigote within the macrophage phagolysosome. The spectrum of clinical manifestations caused by the *Leishmania* species depends on several factors including species of infecting parasite and host factors such as immunological response resulting in cutaneous, mucocutaneous, or visceral leishmaniasis [3].

Visceral leishmaniasis, also known as Kala azar, is one of the most challenging infectious ailments and is fatal if untreated. It is caused by a complex of species comprising *L. donovani*, *L. infantum* and *L. chagasi*. *L. donovani* is predominantly found in the Indian subcontinent, Asia, and Africa while *L. infantum* or *L. chagasi* exist in the Mediterranean region, Southwest and Central Asia and South America [4]. In subclinical visceral leishmaniasis, which is newly acquired, there are no indications of disease and, as it progresses, it becomes oligosymptomatic. If not treated, it becomes fully established with hepatosplenomegaly and associated anemia, leukopenia or thrombocytopenia and hypergammaglobulinemia leading to vulnerability to secondary infections and finally death. Post Kala azar dermal leishmaniasis is a complication characterized by a macular,

maculopapular and nodular rash in patients who have recovered from visceral leishmaniasis. Increasing HIV-visceral leishmaniasis co-infection worldwide is also a cause for concern [5].

The *L. donovani* genome sequencing was recently completed and provided insights into the organization of genes in this protozoan parasite. The *L. donovani* genome has 36 chromosomes comprising 32.4 million base pairs and 8195 annotated genes [6]. The other old world *Leishmania* species – *L. major* and *L. infantum* – have 36 chromosomes each [7,8]. However, the new world *Leishmania* species – *L. braziliensis* and *L. mexicana* – have 35 and 34 chromosomes [9], respectively which can be explained by fusion of chromosomes [10]. The details of the sequenced *Leishmania* species till date are provided in Table 1.

Searching tandem mass spectrometry (MS/MS) data against six-frame translated nucleotide sequences from a genome provides the most direct evidence for protein coding genes. Thus, the use of proteomic data to annotate sequenced genomes is a complementary strategy to conventional genome annotation [11–13]. One of the first approaches to search mass spectrometry data against a six-frame translated nucleotide database was described by Yates et. al. [14] and Choudhary et. al. suggested the use of EST databases along with genome database [15]. This proteogenomic approach has been used to identify and correct annotated genes in *Homo sapiens* [16–19], *Anopheles gambiae* [20,21], *Mycobacterium tuberculosis* [22], *Candida glabrata* [23], *Plasmodium falciparum* [24], *Toxoplasma gondii* [25], *Schmidtea mediterranea* [26], *Yersinia pestis* [27], *Pristionchus pacificus* [28], *Deinococcus deserti* [29] and many other organisms. This approach provides useful information regarding the existence of novel genes, translational start sites from N-terminally acetylated peptides, frame shifts, splice variants, truncation and extension of existing proteins. This approach is mostly useful in scenarios where the genome sequence is available for the organism under study.

Non-availability of genome sequence can impede identification of proteins owing to lack of the most appropriate database against which the MS/MS spectra must be searched. However, this limitation can be overcome by using protein and genome sequence data from taxonomically related species, if they are available. Using this comparative proteogenomics approach, we had previously mapped the proteome of *L. donovani* [30] using protein data from three related *Leishmania* species (*L. infantum*, *L. major* and *L. braziliensis*). This proteomic analysis resulted in identification of 22,322 unique peptides which mapped to 9431 proteins from the three related *Leishmania* species. Out of these, a majority of peptides (12,141) mapped to 3711 proteins in *L. infantum*, which is taxonomically the closest species to *L. donovani*. We identified 1387 proteins to be expressed in both amastigote and promastigote life stages of *L. donovani*. However, 1423 proteins were identified in amastigotes and 901 proteins were identified in promastigotes.

In the present study, we used the mass spectrometry data from our previous proteomic analysis of *L. donovani* and searched it against the newly annotated protein and six-frame translation of the recently sequenced *L. donovani* genome. This proteogenomic analysis resulted in identification of 3999 proteins in *L. donovani*, accounting for 50% of the annotated *L. donovani* proteome. In addition, this mass spectrometry data was searched

against a six-frame translated genome of *L. donovani* resulting in the identification of 255 GSSPs which led to identification of 20 novel genes as well as 14 N-terminal and 26 C-terminal extensions of existing gene models in *L. donovani*.

2. Materials and methods

2.1. Leishmania donovani sample preparation, fractionation and LC–MS/MS analysis

We used axenic promastigote and amastigote stages of Dd8 strain of *L. donovani* (MHOM/IND/80/Dd8) in our previous comparative proteogenomic analysis. We fractionated cell lysates from both the life stages by SDS-PAGE followed by in-gel trypsin digestion. Similarly, we carried out in-solution digestion of cell lysates followed by strong cationic exchange (SCX) chromatography. A total of 112 fractions obtained from 2 different fractionation techniques were analyzed by LC–MS/MS using LTQ-Orbitrap Velos ETD mass spectrometer. MS was acquired with FT analyzer at the resolving power of 60,000 at 400 m/z. MS/MS was carried out in HCD mode with resolving power of 15,000 at 400 m/z. The lock mass option was enabled for accurate mass measurements.

2.2. Mass spectrometry data analysis

A non-redundant (nr) protein database of *BPK282A1* strain of *L. donovani* (n = 8032) database available from NCBI (<http://www.ncbi.nlm.nih.gov>) as of February 20, 2012 was used. Sequest and Mascot (version 2.2) search engines were used to search *L. donovani* mass spectrometry data and searches were submitted through Proteome Discoverer software (version 1.3). The search parameters used were as follows: a) trypsin as a proteolytic enzyme (with up to one missed cleavage); b) peptide mass error tolerance of 20 ppm; c) fragment mass error tolerance of 0.1 Da; d) carbamidomethylation of cysteine as fixed modification and oxidation of methionine, deamidation of asparagine, glutamine and acetylation of protein N-termini were included as variable modifications. Sequest and Mascot peptide data were extracted with 1% false discovery rate (FDR) threshold. Unique peptide data obtained from both Sequest and Mascot search algorithms after protein database dependent searches were used for further analysis.

2.3. Proteogenomic data analysis

The genome sequence of *BPK282A1* strain of *L. donovani* was downloaded from NCBI and a six-frame translated database (n = 1,096,898) was created using in-house python scripts. The six-frame translated genome database was created by masking the gaps in the genome. The translation was truncated at a codon penultimate to the gap in the genome and the first codon after the gap was treated as a start codon for a new translated protein entry. This strategy was used since it is unbiased and does not rely on pre-computed gene models. Most of the commonly occurring contaminants such as trypsin, keratins and albumin were added to both the protein and six-frame translated genome databases that were used for MS/MS ion search. Searches were submitted through Proteome Discoverer software (Thermo Scientific, version 1.3) to Sequest and Mascot (version 2.2) search engines as described above.

The *L. donovani* six-frame translated genome database searches resulted in identification of unique peptide data which includes GSSPs. This unique list of GSSPs was generated by

comparing the unique peptide data with the protein database search results. Peptides that do not map to the protein database but map uniquely to the six-frame translated genome database were considered as GSSPs. The genomic regions to which these GSSPs map were further analyzed to identify novel genes or corrections to existing annotations (N- and C-terminal extensions). GSSPs were categorized either as 1) mapping intergenic regions or 2) overlapping annotated genes in *BPK282A1* strain of *L. donovani*. Additionally, we used two different gene prediction programs – *FgeneSH* and *GeneMark* – to identify alternative gene models. Hence, novel genes or extensions to the existing *L. donovani* gene models obtained using GSSP evidence and gene prediction tools were further checked for their conservation across members of the genus *Leishmania*.

2.4. Bioinformatics analysis

Proteins identified in our current analysis were categorized into groups based on their primary subcellular localization e.g. cytoplasm, nucleus, extracellular. In addition to this, the proteins were also classified based on biological process (e.g., cell cycle proteins). These analyses were performed in compliance with Gene Ontology (GO) standards.

2.5. Proteomic data availability

To make our observations publicly available and accessible to other researchers, the complete set of raw mass spectrometry data (.raw files) generated from this study has been made available through the Tranche server (<http://proteomecommons.org/tranche>). The Tranche Hash is:

t+0N3q+IYbj+6iSsAkkVwPow0FVKpJV1YIsHERg9DDwPc
EtPz2T7LE2mBpK8yxH0ux3cpz2zN/zRTA2xTJy8TobE22cAAA
AAAAA4fw==.

3. Results and discussion

3.1. Summary of proteomic data

The aim of this study was to carry out a detailed genome annotation of *L. donovani* using the recently sequenced genome sequence data. The strategy used for proteomic profiling and genome annotation of *L. donovani* is outlined in Fig. 1. In our previous study, we had used 2 different life stages of *L. donovani* i.e. promastigote and amastigote stages to map the proteome of *L. donovani* using a comparative proteogenomic approach. The 112 LC–MS/MS data files that were acquired in the previous study were used for the current analysis. This dataset included ~1.5 million MS/MS spectra that were searched using Sequest and Mascot search algorithms using the recently available *L. donovani* protein and six-frame translated genome database. In all, 203,342 peptide spectrum matches (PSMs) that led to 17,759 unique peptides at a 1% FDR cutoff were identified from both protein and six-frame translated genome database searches. Supplementary Tables 1 and 2 provide a complete list of proteins and peptides identified in protein database searches.

3.2. Confirmation of annotated proteins in *L. donovani* from protein database searches

The mass spectrometry data was searched against non-redundant (nr) protein database of *BPK282A1* strain of *L. donovani*. This resulted in identification of 17,504 unique peptides which mapped to 3999 proteins thus confirming the existence of corresponding protein annotations in *L. donovani* which accounted for 50% of *L. donovani* proteome. The peptide data from our current (17,504 peptides) and previous (22,322 peptides) study was compared for identification of orthologs and paralogs in different *Leishmania* species. We identified many peptides which are shared between 2 or more *Leishmania* species this shows the homology between orthologs from different *Leishmania* species. *L. infantum* is taxonomically the closest species to *L. donovani* and majority of our peptides (12,177) from *L. donovani* mapped to *L. infantum* thus validating their phylogenetic relationship.

3.3. Confirming start sites of known proteins

Identifying correct translational start sites (TSS) is a challenging task for predicted transcripts. Most of the methods used for TSS determination depend on annotating the longest open reading frame in a given nucleotide sequence [31]. Thus, additional methods such as homology-based bioinformatics approach and mass spectrometry-based approaches have been used to determine N-terminal acetylation of proteins [17,32,33]. N-terminal acetylation occurs in a majority of eukaryotic proteins and it is catalyzed by N-acetyltransferases (NAT) which transfer acetyl groups from acetyl-CoA to the intact N-termini or to one of the internal amino acids after the initiator methionine alone or additional residues are cleaved. There are 5 groups of N-terminal acetyl transferases (A to E) which carry out N-terminal acetylation of majority of eukaryotic proteins [34,35]. Confirmation of exact TSS based on N-terminal acetylation can be determined by using mass spectrometry-based approaches [36,37]. N-terminally acetylated peptides can be determined by searching mass spectrometry data using N-terminal acetylation as a modification during database searches. In our present analysis we identified 113 N-terminally acetylated peptides and among these the majority had methionine (43%) residue acetylated. In addition to this, alanine (17%) and serine (30%) residues were also found to be modified. The complete list of N-terminally acetylated peptides and the corresponding proteins is shown in Supplementary Table 2.

3.4. Proteogenomic analysis of *L. donovani*

3.4.1. Identification of novel genes in *L. donovani*—GSSPs identified in six-frame translated genome database searches were categorized as intergenic if they mapped to a particular region in the genome with no annotated protein coding genes. Such matches suggest the presence of novel protein coding genes which are not annotated in *L. donovani* genome. In the present study, we identified 113 peptides mapping to the intergenic region which in turn resulted in identification of 20 novel protein coding genes in *L. donovani*. A complete list of GSSPs with the genome co-ordinates which resulted in identification of novel genes in *L. donovani* is provided in Supplementary Table 3.

An illustrative example of a novel gene identified in *L. donovani* is where a novel gene coded for a flagellar radial spoke protein. This protein forms a part of the flagellar apparatus where it plays an important role in controlling flagellar movements. We identified 4 unique

GSSPs mapping to intergenic region on chromosome 29 on the negative strand. Upon gene prediction analysis, we identified a 702 amino acid long protein coding gene in *L. donovani* (Fig. 2). An ortholog for flagellar radial spoke protein exists in the closely related species *L. infantum* which is coded by gene *LinJ29_0690* and it codes for 702 amino acids.

Additionally, there is a truncated flagellar radial spoke protein coding gene (*LDBPK_290690*) on chromosome 29 on the negative strand which codes for 36 amino acid long truncated protein. This indicated a wrong annotation of flagellar radial spoke protein coding gene and the full length gene was annotated based on our analysis. A pair wise alignment of novel flagellar radial spoke protein identified in *L. donovani* and its ortholog, *LINJ29_0690*, in *L. infantum* (Supplementary Fig. 1) shows that the full length novel flagellar radial spoke protein (FgeneSH) identified in our study is highly conserved when compared to its *L. infantum* ortholog (*LINJ29_0690*).

3.4.2. Identification of N-terminal extensions in *L. donovani*—We also refined existing gene models in *L. donovani*. In the present study, we identified 26 GSSPs mapping to the 5' boundary of existing genes which resulted in N-terminal extension of 14 proteins in *L. donovani*. As an example, we observed an N-terminal extension of an existing hypothetical protein-coding gene (*LDBPK_261010*), which was predicted to encode a 357 amino acid long protein. Based on three GSSPs mapping to the 5' region upstream of this gene and additional gene prediction analysis, we were able to annotate a longer protein product of 779 amino acids which contained the 3 GSSPs. This longer protein had a conserved ortholog in *L. major*, which was encoded by the gene *LmjF_26_1030*. This ortholog codes for a longer protein product of 779 amino acids (Fig. 3). A pair wise alignment of truncated *L. donovani* protein *LDBPK_261010*, *L. major* ortholog *LMJF_26_1030* and the full length N-terminally extended *L. donovani* protein (FgeneSH) (Supplementary Fig. 2) shows that the N-terminally extended *L. donovani* protein identified in our study is highly conserved.

3.4.3. Identification of C-terminal extension in *L. donovani*—In the present study, we identified 116 peptides mapping to the 3' boundaries of existing genes which resulted in C-terminal extension of 26 proteins in *L. donovani*. As an illustration for C-terminal extension, we identified a heat shock protein 70 (Hsp70) coded by the gene *LDBPK_302540* in *L. donovani*. It codes for a small truncated protein product (86 amino acids) which is a result of gap due to series of nucleotides (724 Ns) downstream of gene *LDBPK_302540* where no nucleotide could be assigned. The full length Hsp70 protein (502 amino acids) is encoded by the gene *LDBPK_302480* and is localized to the mitochondria. However, we have identified 11 unique GSSPs which map downstream of the truncated *LDBPK_302540* gene. Upon further bioinformatics analysis, we determined that these peptides are well conserved in HSP orthologs from *L. infantum* (*LINJ_30_2530*) and *L. major* (*LMJF_30_2550*) which are of 661 amino acids. This resulted in C-terminal extension of Hsp70 (*LDBPK_302540*) by 575 amino acids (Fig. 4). A pair wise alignment of truncated *L. donovani* heat shock proteins (*LDBPK_302540* and *LDBPK_302480*) and full length orthologs from *L. infantum* (*LINJ30_2870*) and *L. major* (*LMJF_30_2550*) shows that the C-terminally extended regions identified in our study is highly conserved (Supplementary Fig. 3).

In our previous comparative proteogenomic study we had identified 13 N-terminal extensions in *L. donovani*. We have reanalyzed these previously identified extensions in *L. donovani* with the availability of genome sequence data and we found that only 2 out of the 13 N-terminal extensions are incorporated in the current genome annotation of *L. donovani*. This shows the utility of proteogenomic approaches in annotating genomes. Our current proteogenomic analysis of *L. donovani* using sequenced genome data identified many genes that were erroneously annotated as truncated forms. These truncated forms were a result of gaps arising from the non-sequenced flanking regions with no nucleotide sequence. These truncated genes are frequently referred to in the current genome annotation of *L. donovani* as “partial” and there are many such truncated partial protein coding genes in *L. donovani*. However, our proteogenomic analysis of *L. donovani* identified peptides mapping upstream or downstream of the gap which resulted in N- or C-terminal extension of these protein coding genes. A complete list of GSSPs along with the genome coordinates which resulted in identification of N- and C-terminal extensions to the existing gene models in *L. donovani* is provided in Supplementary Table 4.

3.5. Expression of proteins in promastigote and amastigote stages of *L. donovani*

The promastigote stage represents the insect stage and is well suited for survival in the gut of sandflies. Amastigotes on the other hand are adapted to survive in the highly acidic environment inside the phagosome of infected macrophage. A different repertoire of proteomes thus needs to be expressed in order for the parasite to survive in different environments (i.e. gut of sandflies or inside the infected macrophages). Previous transcriptomic [38–42] and proteomic analyses from other groups has provided a set of differentially expressed genes/proteins in these 2 different life stages of *Leishmania* [43–49]. Many of these studies including our proteogenomics made use of axenic amastigotes which are generated by differentiating promastigotes by culturing them at 37 °C and pH 6 in 5% CO₂ which mimics the intracellular environment inside the macrophages. In a study by Lahav et al. the differentiation from promastigote to amastigote was studied in the context of mRNA and protein levels using a transcriptomic and differential proteomic approach. Various time points during the differentiation of promastigote to amastigote were sampled and it was found that there is a substantial change at mRNA level during early differentiation. However, during the later stages of the differentiation process, translation and post-translational regulation of protein abundance was shown to play an important role [50].

It is debated that the true proteome of intracellular amastigote is different from the axenic amastigotes which are frequently used in proteomic analysis due to the ease of culturing and scaling up of axenic amastigote cultures and frequently these axenic forms are referred to as amastigote-like to differentiate them from true intracellular amastigotes [51,52]. In a recent comparative proteomic study by Biyani et al., intracellular amastigotes and axenic promastigotes from three different clones were used for analysis. This study resulted in identification of differential set of proteins between the two life stages and many of the differentially expressed proteins were conserved across the different clones. In this study, they identified 205 proteins in total and out of these 29 proteins were upregulated and 36 were downregulated in amastigotes compared to promastigotes [53].

In our present proteogenomic analysis, 613 proteins were identified in the amastigote stage and 1090 proteins were identified in the promastigote stage of *L. donovani*. In contrast 2296 proteins were identified in both the life stages of *L. donovani*. The distribution of identified proteins between the two life stages of *L. donovani* is shown in Fig. 5. We identified various classes of proteins expressed in both life stages including proteins involved in cell motility, cytoskeleton, energy metabolism and cellular signaling. In addition to this, many of the factors involved in survival and virulence were also identified in this study.

3.6. Identification of virulence factors in *L. donovani*

We identified various virulence factors known to be associated with infection process. This included previously known virulence factors such as *Leishmania* homolog of activated C kinase receptor (LACK) which is encoded by the gene *LDBPK_282970* and is known to induce suppression of CD4+ T cells upon infection of the host and it may play a crucial role in parasite infection [54,55]. Prohibitins are GPI anchored proteins which are shown to be localized in flagellar and aflagellar pole of the *L. donovani* promastigote. Prohibitins encoded by the gene *LDBPK_161710* in *L. donovani*. The aflagellar pole is involved in invasion process of macrophage and it involves binding of prohibitin to its corresponding surface HSP70 receptor on macrophage. Hence, it seems to play an important role in host–parasite interaction [56]. Magnesium transporter 2 (MGT2) is a cation transporter encoded by the gene *LDBPK_251130* and it is involved in the transport of magnesium in intracellular amastigotes engulfed in phagolysosome of macrophages. *MGT2* single allele knockout study showed drastic loss of virulence and it also reduced ability of *Leishmania* to survive inside acidified phagolysosomes. This was attributed to the inability of promastigote to transform completely to amastigote which is better suited to survive inside the harsh environment of phagolysosome [57].

Sphingolipids are components of lipid rafts and are involved in cellular signaling. The de novo synthesis of sphingolipids involves serine palmitoyltransferase which is a crucial enzyme in the biosynthetic pathway and is coded by the gene *LDBPK_350320* in *L. donovani*. Mutation of *spt* gene has been shown to result in inability of parasites to differentiate into infective metacyclic forms. Additionally, the undifferentiated parasites die due to accumulation of vesicles [58] and there is defect in membrane trafficking in extracellular *Leishmania*. It was also seen that the parasites are defective in infecting host macrophages and this was attributed to defect in the localization of GP63 (leishmanolysin) and GPI mediated association GP63 to the lipid rafts [59]. In our study, we identified many other virulence factors such as phosphoglycan beta 1,3 galactosyltransferase, lipophosphoglycan biosynthetic enzyme and surface antigen-like protein which were also identified in our previous comparative proteogenomic analysis of *L. donovani*. These virulence factors are expressed at a basal level in axenic *L. donovani*.

3.7. Identification of survival factors in *L. donovani*

We identified many ATP binding cassette (ABC) class of proteins in our current analysis of *L. donovani*. ABC proteins have been shown to be associated with drug resistance in *Leishmania*. ABC proteins characteristically have transmembrane domains that are hydrophobic and form a channel for passage of molecules. In addition to this, they have

nucleotide binding domain which bind to ATP and bring about hydrolysis of ATP to drive the process of pumping drugs or other metabolites outside the cell [60,61]. We identified a classic ABC transporter P-glycoprotein (*LDBPK_341060*) in *L. donovani*. P-glycoprotein or Pgp was identified initially as an ABC gene in the genus *Leishmania* and it was shown to be a part of the H-circle DNA which is frequently found to be amplified in drug resistant *Leishmania* [62,63]. Initially P-glycoprotein was shown to induce methotrexate resistance in *Leishmania tarentolae* [62] and *L. major* [64]. It was shown later that P-glycoprotein could confer arsenate resistance in different *Leishmania* species upon transfection [65] but the amount of resistance to the drugs varied with type of species [66]. In addition to this, the gene amplification of *pgpA* gene was shown not to be directly related to the level of resistance that it confers to *Leishmania* [65]. However, null mutants of *pgpA* gene can result in increased sensitivity to oxyanions such as arsenate and antimony in *L. tarentolae*, but the PgpA was shown to confer only low level of protection against oxyanions upon *pgpA* gene transfection in mutant cells and it is not the major transporter of oxyanions [67].

We identified another member of the ABC family multidrug resistance protein in *L. donovani* which is encoded by the gene *LDBPK_282660*. This MRP like protein is shown to be involved in active transport of multiple drugs including amphotericin B and sodium stibogluconate [68]. This ABC protein was shown to be 2-fold upregulated in antimony resistant *L. donovani* clinical isolate [69]. In addition to these, we identified many other ABC proteins in our current analysis of *L. donovani* and the partial list of the identified ABC proteins is shown in Table 2. We also identified additional proteins which play a crucial role in conjunction with ABC proteins such as PgpA in efflux of various metabolites including drugs. We identified γ -glutamylcysteine synthase (γ -GCS) and ornithine decarboxylase (ODC) which are encoded by genes *LDBPK_181660* and *LDBPK_120100* in *L. donovani*, respectively. These proteins are crucial enzymes involved in synthesis of glutathione and polyamine spermidine. γ -glutamylcysteine synthase (γ -GCS) was shown to be upregulated in drug resistant *Leishmania* [70]. Both these molecules are precursors for synthesis of low molecular weight thiols such as trypanothione [71] which play important role in preventing oxidative damage and maintaining intracellular reducing environment [72]. Trypanothione is synthesized by trypanothione synthetase [73] which we have identified in *L. donovani* and is coded by the gene *LDBPK_271770*. This trypanothione is conjugated by a transferase to heavy metals/drugs and is recognized by PgpA and it assists in the active extrusion of trypanothione conjugated molecules outside the cells [74].

3.8. Analysis of secreted and membrane proteins from *L. donovani*

In our current study, we carried out bioinformatics analysis to identify membrane and secreted proteins in *L. donovani*. Of the 3999 proteins identified in *L. donovani* 202 proteins contained only signal peptides (SP) and 282 proteins contained a transmembrane domain. In addition, 71 proteins contained both a transmembrane domain and a signal peptide (Fig. 6A). Identifying secreted and transmembrane proteins in *L. donovani* can provide valuable insights into the function of these proteins in relation to virulence and pathogenesis. Amastin which is encoded by gene *LDBPK_281510* is known to adjust the intracytoplasmic pH of macrophage engulfed amastigotes in the phagolysosome by pumping ions across the membrane [75]. It is predominantly expressed in the amastigote life stage upon exposure to

acidic pH [76]. The amastin gene family is highly diversified in the genus *Leishmania* [77]. Tuzins are a class of surface proteins identified to be part of amastin family and it has been studied in context of related kinetoplastid *Trypanosoma cruzi*, where its expression was shown to be post-transcriptionally suppressed resulting in low levels of tuzin [78]. Cysteine peptidase B was shown to inhibit IL-12 production from macrophages stimulated with bacterial lipopolysaccharide (LPS). It was shown to bring about degradation of IKappaB alpha and beta as well as NF-kappaB. Hence, *Leishmania* is able to inhibit NF-KappaB signaling and in turn IL-12 production [79]. The partial list of secreted or transmembrane molecules identified in this study is shown in Table 3.

3.9. Gene ontology and bioinformatics analysis

Gene ontology analysis was carried out to categorize protein based on biological process and localization. The proteins identified in the current proteomic analysis of *L. donovani* were classified based on biological process (e.g. cell communication and signaling). This GO analysis resulted in 1761 proteins (44%) being grouped into one of the biological functions (Fig. 6B). In addition to this, proteins were categorized based on primary subcellular localization e.g. nucleus, cytoplasm and cytoskeleton; 1020 proteins (26%) were assigned one of the subcellular locations (Fig. 6C). In our current study, we also analyzed biological domains and motifs for the proteins identified in *L. donovani* and the details can be found in Supplementary Table 1. In addition to this, *L. donovani* has many genes that are functionally uncharacterized and do not have sequence similarity with other known genes in protozoa. We identified 1578 that are annotated as hypothetical proteins in recently sequenced *L. donovani*. Many of these proteins could be interesting targets for drug and vaccine development.

4. Conclusions

In our proteogenomic analysis reported here, we have cataloged the largest set of proteins from the two life stages of *L. donovani*. This illustrates the importance of high-resolution mass spectrometry derived data in proteomic profiling of *L. donovani*. Many of the proteins identified in our current analysis using sequenced *L. donovani* genome and protein data have not been shown to exist previously. However, many proteomic studies in the past, including our previous comparative proteogenomic study, had identified many proteins based on homology with the closest taxonomically related *Leishmania* species i.e. *L. infantum*. Significantly, we identified 255 GSSPs from six-frame translated genome searches of *L. donovani* which led to correction of 40 gene models (14 N-and 26 C-terminal extensions). We also identified 20 novel genes in the recently sequenced *L. donovani* genome based on genome search-specific peptide evidence. This shows that annotating exact protein-coding genes is a challenging task and proteogenomic approach can help not only to identify novel genes but also refine gene models of existing genes.

Supplementary data to this article can be found online at <http://dx.doi.org/10.1016/j.jprot.2013.04.021>.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the Department of Biotechnology (DBT), Government of India, for research support to the Institute of Bioinformatics. We thank Agilent Technologies for access to instrumentation. T. S. Keshava Prasad is supported by a research grant on “Development of Infrastructure and a Computational Framework for Analysis of Proteomic Data” from DBT. H. C. Harsha is a Wellcome Trust/DBT India Alliance Early Career Fellow. Raja Sekhar Nirujogi, Harsh Pawar, Sweta Khobragade, and Jyoti Sharma are recipients of Senior Research Fellowship and Gajanan Sathe, Sandip Chavan are recipients of Junior Research Fellowship from the Council of Scientific and Industrial Research (CSIR), Government of India. Santosh Renuse is a recipient of Senior Research Fellowship from University Grants Commission (UGC), Government of India. We would like to thank Kiran N. Mahale from National Centre for Cell Sciences, Pune for his technical input.

Abbreviations

GSSPs	Genome search specific peptides
FDR	False discovery rate
PSM	Peptide spectrum matches
HCD	High energy collision induced dissociation

References

1. WHO. Leishmaniasis: burden of disease, surveillance e control, epidemics, access to medicines, information resources. World Health Organization; 2010.
2. Killick-Kendrick R. Phlebotomine vectors of the leishmaniasis: a review. *Med Vet Entomol*. 1990; 4:1–24. [PubMed: 2132963]
3. Murray HW, Berman JD, Davies CR, Saravia NG. Advances in leishmaniasis. *Lancet*. 2005; 366:1561–77. [PubMed: 16257344]
4. Berman J. Visceral leishmaniasis in the New World & Africa. *Indian J Med Res*. 2006; 123:289–94. [PubMed: 16778311]
5. Alvar J, Aparicio P, Aseffa A, Den Boer M, Canavate C, Dedet JP, et al. The relationship between leishmaniasis and AIDS: the second 10 years. *Clin Microbiol Rev*. 2008; 21:334–59. [table of contents]. [PubMed: 18400800]
6. Downing T, Imamura H, Decuypere S, Clark TG, Coombs GH, Cotton JA, et al. Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance. *Genome Res*. 2011; 21:2143–56. [PubMed: 22038251]
7. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, et al. The genome of the kinetoplastid parasite, *Leishmania major*. *Science*. 2005; 309:436–42. [PubMed: 16020728]
8. Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail MA, et al. Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. *Nat Genet*. 2007; 39:839–47. [PubMed: 17572675]
9. Rogers MB, Hilley JD, Dickens NJ, Wilkes J, Bates PA, Depledge DP, et al. Chromosome and gene copy number variation allow major structural change between species and strains of *Leishmania*. *Genome Res*. 2011; 21:2129–42. [PubMed: 22038252]
10. Dujardin JC. Structure, dynamics and function of *Leishmania* genome: resolving the puzzle of infection, genetics and evolution? *Infect Genet Evol*. 2009; 9:290–7. [PubMed: 19100863]
11. Mann M, Pandey A. Use of mass spectrometry-derived data to annotate nucleotide and protein sequence databases. *Trends Biochem Sci*. 2001; 26:54–61. [PubMed: 11165518]

12. Pandey A, Mann M. Proteomics to study genes and genomes. *Nature*. 2000; 405:837–46. [PubMed: 10866210]
13. Pandey A, Lewitter F. Nucleotide sequence databases: a gold mine for biologists. *Trends Biochem Sci*. 1999; 24:276–80. [PubMed: 10390617]
14. Yates III JR, Eng JK, McCormack AL. Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal Chem*. 1995; 67:3202–10. [PubMed: 8686885]
15. Choudhary JS, Blackstock WP, Creasy DM, Cottrell JS. Interrogating the human genome using uninterpreted mass spectrometry data. *Proteomics*. 2001; 1:651–67. [PubMed: 11678035]
16. Menon R, Zhang Q, Zhang Y, Fermin D, Bardeesy N, DePinho RA, et al. Identification of novel alternative splice isoforms of circulating proteins in a mouse model of human pancreatic cancer. *Cancer Res*. 2009; 69:300–9. [PubMed: 19118015]
17. Molina H, Bunkenborg J, Reddy GH, Muthusamy B, Scheel PJ, Pandey A. A proteomic analysis of human hemodialysis fluid. *Mol Cell Proteomics*. 2005; 4:637–50. [PubMed: 15723814]
18. Desiere F, Deutsch EW, Nesvizhskii AI, Mallick P, King NL, Eng JK, et al. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol*. 2005; 6:R9. [PubMed: 15642101]
19. Fermin D, Allen BB, Blackwell TW, Menon R, Adamski M, Xu Y, et al. Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *Genome Biol*. 2006; 7:R35. [PubMed: 16646984]
20. Chaerkady R, Kelkar DS, Muthusamy B, Kandasamy K, Dwivedi SB, Sahasrabudhe NA, et al. A proteogenomic analysis of *Anopheles gambiae* using high-resolution Fourier transform mass spectrometry. *Genome Res*. 2011; 21:1872–1881. [PubMed: 21795387]
21. Kalume DE, Peri S, Reddy R, Zhong J, Okulate M, Kumar N, et al. Genome annotation of *Anopheles gambiae* using mass spectrometry-derived data. *BMC Genomics*. 2005; 6:128. [PubMed: 16171517]
22. Kelkar DS, Kumar D, Kumar P, Balakrishnan L, Muthusamy B, Yadav AK, et al. Proteogenomic analysis of *Mycobacterium tuberculosis* by high resolution mass spectrometry. *Mol Cell Proteomics*. 2011; 10:1–12.
23. Prasad TS, Harsha HC, Keerthikumar S, Sekhar NR, Selvan LD, Kumar P, et al. Proteogenomic analysis of *Candida glabrata* using high resolution mass spectrometry. *J Proteome Res*. 2012; 11:247–60. [PubMed: 22129275]
24. Lasonder E, Ishihama Y, Andersen JS, Vermunt AM, Pain A, Sauerwein RW, et al. Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature*. 2002; 419:537–42. [PubMed: 12368870]
25. Xia D, Sanderson SJ, Jones AR, Prieto JH, Yates JR, Bromley E, et al. The proteome of *Toxoplasma gondii*: integration with the genome provides novel insights into gene expression and annotation. *Genome Biol*. 2008; 9:R116. [PubMed: 18644147]
26. Bocchinfuso DG, Taylor P, Ross E, Ignatchenko A, Ignatchenko V, Kislinger T, et al. Proteomic profiling of the planarian *Schmidtea mediterranea* and its mucous reveals similarities with human secretions and those predicted for parasitic flatworms. *Mol Cell Proteomics*. 2012; 11:681–91. [PubMed: 22653920]
27. Payne SH, Huang ST, Pieper R. A proteogenomic update to *Yersinia*: enhancing genome annotation. *BMC Genomics*. 2010; 11:460. [PubMed: 20687929]
28. Borchert N, Dieterich C, Krug K, Schutz W, Jung S, Nordheim A, et al. Proteogenomics of *Pristionchus pacificus* reveals distinct proteome structure of nematode models. *Genome Res*. 2010; 20:837–46. [PubMed: 20237107]
29. Baudet M, Ortet P, Gaillard JC, Fernandez B, Guerin P, Enjalbal C, et al. Proteomics-based refinement of *Deinococcus deserti* genome annotation reveals an unwonted use of non-canonical translation initiation codons. *Mol Cell Proteomics*. 2010; 9:415–26. [PubMed: 19875382]
30. Pawar H, Sahasrabudhe NA, Renuse S, Keerthikumar S, Sharma J, Kumar GS, et al. A proteogenomic approach to map the proteome of an unsequenced pathogen – *Leishmania donovani*. *Proteomics*. 2012; 12:832–44. [PubMed: 22539434]

31. Peri S, Pandey A. A reassessment of the translation initiation codon in vertebrates. *Trends Genet.* 2001; 17:685–7. [PubMed: 11718907]
32. Gevaert K, Goethals M, Martens L, Van Damme J, Staes A, Thomas GR, et al. Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nat Biotechnol.* 2003; 21:566–9. [PubMed: 12665801]
33. Goetze S, Qeli E, Mosimann C, Staes A, Gerrits B, Roschitzki B, et al. Identification and functional characterization of N-terminally acetylated proteins in *Drosophila melanogaster*. *PLoS Biol.* 2009; 7:e1000236. [PubMed: 19885390]
34. Arnesen T, Van Damme P, Polevoda B, Helsens K, Evjenth R, Colaert N, et al. Proteomics analyses reveal the evolutionary conservation and divergence of N-terminal acetyltransferases from yeast and humans. *Proc Natl Acad Sci U S A.* 2009; 106:8157–62. [PubMed: 19420222]
35. Hollebeke J, Van Damme P, Gevaert K. N-terminal acetylation and other functions of Nalpha-acetyltransferases. *Biol Chem.* 2012; 393:291–8. [PubMed: 22718636]
36. Mischerikow N, Heck AJ. Targeted large-scale analysis of protein acetylation. *Proteomics.* 2011; 11:571–89. [PubMed: 21246731]
37. Helbig AO, Gauci S, Raijmakers R, van Breukelen B, Slijper M, Mohammed S, et al. Profiling of N-acetylated protein termini provides in-depth insights into the N-terminal nature of the proteome. *Mol Cell Proteomics.* 2010; 9:928–39. [PubMed: 20061308]
38. Holzer TR, McMaster WR, Forney JD. Expression profiling by whole-genome interspecies microarray hybridization reveals differential gene expression in procyclic promastigotes, lesion-derived amastigotes, and axenic amastigotes in *Leishmania mexicana*. *Mol Biochem Parasitol.* 2006; 146:198–218. [PubMed: 16430978]
39. Saxena A, Lahav T, Holland N, Aggarwal G, Anupama A, Huang Y, et al. Analysis of the *Leishmania donovani* transcriptome reveals an ordered progression of transient and permanent changes in gene expression during differentiation. *Mol Biochem Parasitol.* 2007; 152:53–65. [PubMed: 17204342]
40. Srividya G, Duncan R, Sharma P, Raju BV, Nakhasi HL, Salotra P. Transcriptome analysis during the process of in vitro differentiation of *Leishmania donovani* using genomic microarrays. *Parasitology.* 2007; 134:1527–39. [PubMed: 17553180]
41. Li Q, Zhao Y, Ni B, Yao C, Zhou Y, Xu W, et al. Comparison of the expression profiles of promastigotes and axenic amastigotes in *Leishmania donovani* using serial analysis of gene expression. *Parasitol Res.* 2008; 103:821–8. [PubMed: 18568446]
42. Alcolea PJ, Alonso A, Gomez MJ, Sanchez-Gorostiaga A, Moreno-Paz M, Gonzalez-Pastor E, et al. Temperature increase prevails over acidification in gene expression modulation of amastigote differentiation in *Leishmania infantum*. *BMC Genomics.* 2010; 11:31. [PubMed: 20074347]
43. El Fakhry Y, Ouellette M, Papadopoulou B. A proteomic approach to identify developmentally regulated proteins in *Leishmania infantum*. *Proteomics.* 2002; 2:1007–17. [PubMed: 12203896]
44. Bente M, Harder S, Wiesgigl M, Heukeshoven J, Gelhaus C, Krause E, et al. Developmentally induced changes of the proteome in the protozoan parasite *Leishmania donovani*. *Proteomics.* 2003; 3:1811–29. [PubMed: 12973740]
45. Walker J, Vasquez JJ, Gomez MA, Drummelsmith J, Burchmore R, Girard I, et al. Identification of developmentally-regulated proteins in *Leishmania panamensis* by proteome profiling of promastigotes and axenic amastigotes. *Mol Biochem Parasitol.* 2006; 147:64–73. [PubMed: 16530278]
46. McNicoll F, Drummelsmith J, Muller M, Madore E, Boilard N, Ouellette M, et al. A combined proteomic and transcriptomic approach to the study of stage differentiation in *Leishmania infantum*. *Proteomics.* 2006; 6:3567–81. [PubMed: 16705753]
47. Leifso K, Cohen-Freue G, Dogra N, Murray A, McMaster WR. Genomic and proteomic expression analysis of *Leishmania* promastigote and amastigote life stages: the *Leishmania* genome is constitutively expressed. *Mol Biochem Parasitol.* 2007; 152:35–46. [PubMed: 17188763]
48. Paape D, Lippuner C, Schmid M, Ackermann R, Barrios-Llerena ME, Zimny-Arndt U, et al. Transgenic, fluorescent *Leishmania mexicana* allow direct analysis of the proteome of intracellular amastigotes. *Mol Cell Proteomics.* 2008; 7:1688–701. [PubMed: 18474515]

49. Paape D, Barrios-Llerena ME, Le Bihan T, Mackay L, Aebischer T. Gel free analysis of the proteome of intracellular *Leishmania mexicana*. Mol Biochem Parasitol. 2010; 169:108–14. [PubMed: 19900490]
50. Lahav T, Sivam D, Volpin H, Ronen M, Tsigankov P, Green A, et al. Multiple levels of gene regulation mediate differentiation of the intracellular pathogen *Leishmania*. Faseb J. 2011; 25:515–25. [PubMed: 20952481]
51. Somanna A, Mundodi V, Gedamu L. In vitro cultivation and characterization of *Leishmania chagasi* amastigote-like forms. Acta Trop. 2002; 83:37–42. [PubMed: 12062791]
52. Gupta N, Goyal N, Rastogi AK. In vitro cultivation and characterization of axenic amastigotes of *Leishmania*. Trends Parasitol. 2001; 17:150–3. [PubMed: 11286801]
53. Biyani N, Madhubala R. Quantitative proteomic profiling of the promastigotes and the intracellular amastigotes of *Leishmania donovani* isolates identifies novel proteins having a role in *Leishmania* differentiation and intracellular survival. Biochim Biophys Acta. 2012; 1824:1342–50. [PubMed: 22850196]
54. Kelly BL, Stetson DB, Locksley RM. *Leishmania major* LACK antigen is required for efficient vertebrate parasitization. J Exp Med. 2003; 198:1689–98. [PubMed: 14657221]
55. Locksley RM, Pingel S, Lacy D, Wakil AE, Bix M, Fowell DJ. Susceptibility to infectious diseases: *Leishmania* as a paradigm. J Infect Dis. 1999; 179(Suppl. 2):S305–8. [PubMed: 10081500]
56. Jain R, Ghoshal A, Mandal C, Shaha C. *Leishmania* cell surface prohibitin: role in host-parasite interaction. Cell Microbiol. 2010; 12:432–52. [PubMed: 19888987]
57. Zhu Y, Davis A, Smith BJ, Curtis J, Handman E. *Leishmania major* CorA-like magnesium transporters play a critical role in parasite development and virulence. Int J Parasitol. 2009; 39:713–23. [PubMed: 19136005]
58. Zhang K, Showalter M, Revollo J, Hsu FF, Turk J, Beverley SM. Sphingolipids are essential for differentiation but not growth in *Leishmania*. EMBO J. 2003; 22:6016–26. [PubMed: 14609948]
59. Denny PW, Goulding D, Ferguson MA, Smith DF. Sphingolipid-free *Leishmania* are defective in membrane trafficking, differentiation and infectivity. Mol Microbiol. 2004; 52:313–27. [PubMed: 15066023]
60. Perez-Victoria JM, Parodi-Talice A, Torres C, Gamarro F, Castanys S. ABC transporters in the protozoan parasite *Leishmania*. Int Microbiol. 2001; 4:159–66. [PubMed: 11820434]
61. Ouellette M, Legare D, Papadopolou B. Microbial multidrug-resistance ABC transporters. Trends Microbiol. 1994; 2:407–11. [PubMed: 7850210]
62. Ouellette M, Fase-Fowler F, Borst P. The amplified H circle of methotrexate-resistant *Leishmania tarentolae* contains a novel P-glycoprotein gene. EMBO J. 1990; 9:1027–33. [PubMed: 1969794]
63. Ouellette M, Hettema E, Wust D, Fase-Fowler F, Borst P. Direct and inverted DNA repeats associated with P-glycoprotein gene amplification in drug resistant *Leishmania*. EMBO J. 1991; 10:1009–16. [PubMed: 1672636]
64. Callahan HL, Beverley SM. Heavy metal resistance: a new role for P-glycoproteins in *Leishmania*. J Biol Chem. 1991; 266:18427–30. [PubMed: 1680861]
65. Papadopolou B, Roy G, Dey S, Rosen BP, Ouellette M. Contribution of the *Leishmania* P-glycoprotein-related gene ltpgpA to oxyanion resistance. J Biol Chem. 1994; 269:11980–6. [PubMed: 7909316]
66. Legare D, Papadopolou B, Roy G, Mukhopadhyay R, Haimeur A, Dey S, et al. Efflux systems and increased trypanothione levels in arsenite-resistant *Leishmania*. Exp Parasitol. 1997; 87:275–82. [PubMed: 9371094]
67. Papadopolou B, Roy G, Dey S, Rosen BP, Olivier M, Ouellette M. Gene disruption of the P-glycoprotein related gene pgpa of *Leishmania tarentolae*. Biochem Biophys Res Commun. 1996; 224:772–8. [PubMed: 8713121]
68. Mandal G, Sarkar A, Saha P, Singh N, Sundar S, Chatterjee M. Functionality of drug efflux pumps in antimonial resistant *Leishmania donovani* field isolates. Indian J Biochem Biophys. 2009; 46:86–92. [PubMed: 19374259]

69. Singh N, Almeida R, Kothari H, Kumar P, Mandal G, Chatterjee M, et al. Differential gene expression analysis in antimony-unresponsive Indian kala azar (visceral leishmaniasis) clinical isolates by DNA microarray. *Parasitology*. 2007; 134:777–87. [PubMed: 17306059]
70. Guimond C, Trudel N, Brochu C, Marquis N, El Fadili A, Peytavi R, et al. Modulation of gene expression in *Leishmania* drug resistant mutants as determined by targeted DNA microarrays. *Nucleic Acids Res*. 2003; 31:5886–96. [PubMed: 14530437]
71. Augustyns K, Amssoms K, Yamani A, Rajan PK, Haemers A. Trypanothione as a target in the design of antitrypanosomal and antileishmanial agents. *Curr Pharm Des*. 2001; 7:1117–41. [PubMed: 11472257]
72. Romao PR, Fonseca SG, Hothersall JS, Noronha-Dutra AA, Ferreira SH, Cunha FQ. Glutathione protects macrophages and *Leishmania major* against nitric oxide-mediated cytotoxicity. *Parasitology*. 1999; 118(Pt 6):559–66. [PubMed: 10406034]
73. Saudagar P, Dubey VK. Cloning, expression, characterization and inhibition studies on trypanothione synthetase, a drug target enzyme, from *Leishmania donovani*. *Biol Chem*. 2011; 392:1113–22. [PubMed: 22050226]
74. Grondin K, Haimet A, Mukhopadhyay R, Rosen BP, Ouellette M. Co-amplification of the gamma-glutamylcysteine synthetase gene gsh1 and of the ABC transporter gene pgpA in arsenite-resistant *Leishmania tarentolae*. *EMBO J*. 1997; 16:3057–65. [PubMed: 9214623]
75. Rochette A, McNicoll F, Girard J, Breton M, Leblanc E, Bergeron MG, et al. Characterization and developmental gene regulation of a large gene family encoding amastin surface proteins in *Leishmania* spp. *Mol Biochem Parasitol*. 2005; 140:205–20. [PubMed: 15760660]
76. Wu Y, El Fakhry Y, Sereno D, Tamar S, Papadopolou B. A new developmentally regulated gene family in *Leishmania* amastigotes encoding a homolog of amastin surface proteins. *Mol Biochem Parasitol*. 2000; 110:345–57. [PubMed: 11071288]
77. Jackson AP. The evolution of amastin surface glycoproteins in trypanosomatid parasites. *Mol Biol Evol*. 2010; 27:33–45. [PubMed: 19748930]
78. Teixeira SM, Kirchhoff LV, Donelson JE. *Trypanosoma cruzi*: suppression of tuzin gene expression by its 5'-UTR and spliced leader addition site. *Exp Parasitol*. 1999; 93:143–51. [PubMed: 10529356]
79. Cameron P, McGachy A, Anderson M, Paul A, Coombs GH, Mottram JC, et al. Inhibition of lipopolysaccharide-induced macrophage IL-12 production by *Leishmania mexicana* amastigotes: the role of cysteine peptidases and the NF-kappaB signaling pathway. *J Immunol*. 2004; 173:3297–304. [PubMed: 15322192]
80. Padilla A, Noiva R, Lee N, Mohan KV, Nakhasi HL, Debrabant A. An atypical protein disulfide isomerase from the protozoan parasite *Leishmania* containing a single thioredoxin-like domain. *J Biol Chem*. 2003; 278:1872–8. [PubMed: 12427741]
81. Ben Achour Y, Chenik M, Louzir H, Dellagi K. Identification of a disulfide isomerase protein of *Leishmania major* as a putative virulence factor. *Infect Immun*. 2002; 70:3576–85. [PubMed: 12065498]

Fig. 1.

Proteogenomic workflow for analysis of the *Leishmania donovani* proteome. Mass spectrometry data from *L. donovani* promastigote and amastigote was searched against *L. donovani* protein database. In addition, a six-frame translated *L. donovani* genome database was searched to identify unique genome search-specific peptides (GSSPs) which should not map to protein database. These GSSPs were further analyzed resulting in identification of novel genes and corrections to the existing gene models in *L. donovani*.

Fig. 2.

Identification of a novel protein coding gene based on peptides mapping to an intergenic region in *L. donovani*. (A) Chromosomal position of the truncated flagellar radial spoke protein coding gene LDBPK_290690 (blue box) which is a part of the current genome annotation in *L. donovani*. Four intergenic peptides (marked as red bars) mapped to a region where no known gene was predicted in *L. donovani*. Gene prediction analysis of this region resulted in identification of full length flagellar radial protein coding gene (white box) on chromosome 29 in *L. donovani*. Upon further bioinformatics analysis a corresponding conserved ortholog was identified in *L. infantum* LINJ_29_0690. (B) A pair wise sequence alignment of full length ortholog from *L. infantum* LINJ_29_0690 and novel predicted protein identified in *FgeneSH* gene prediction analysis of *L. donovani*. Red color sequences denote the GSSPs identified in *L. donovani* which support the existence of this novel protein coding gene in *L. donovani* which codes for flagellar radial spoke protein. (C) A representative MS/MS spectrum for identification of genome search specific peptide LPDVSPHHITVAR is shown.

Fig. 3.

Identification of N-terminal extended protein in *L. donovani*. (A) Three peptides (marked as red bars) mapped to the upstream region of *L. donovani* gene LDBPK_261010 (red box) and this codes for short protein product of 367 amino acids (blue box). However, upon gene prediction analysis the presence of a much longer protein coding gene extending N terminal of LDBPK_261010 gene was identified (white box). Ortholog LMJF_26_1030 codes for 779 amino acid long protein (blue box) in *L. major* and this supports the N terminal extension. (B) A pair wise sequence alignment of full length ortholog (LMJ_26_1030) from *L. major* and the truncated *L. donovani* protein (LDBPK_261010) as well as full length predicted protein identified in *FgeneSH* gene prediction analysis. Red color sequences denote the

GSSPs identified in *L. donovani* which support N-terminal extension and correction of *L. donovani* gene LDBPK_261010 and its conservation across the other ortholog in *L. major*. (C) A representative MS/MS spectrum for identification of genome search specific peptide GIYEALQAFELDR is shown.

Fig. 4.

Identification of C-terminal extended protein in *L. donovani*. (A) Eleven peptides (marked as red bars) mapped to the downstream region of the truncated *L. donovani* Hsp70 gene *LDBPK_302540* which codes small truncated protein of 86 amino acids (blue box) which is a part of the current genome annotation in *L. donovani*. However, upon gene prediction the presence of a much longer gene extending C terminal of the current gene model (white box) was identified in our analysis. Orthologs *LINJ_30_2530* and *LMJF_30_2550* genes in *L. infantum* and *L. major* respectively support (blue box) the C terminal extension. (B) Full length heat shock protein 70 (Hsp70) protein sequence (*LINJ_30_2530*) and the blue colored sequence represents conserved truncated Hsp70 protein *LDBPK_302540* in *L.*

donovani showing complete homology with *L. infantum* ortholog. The remaining black colored extended sequence in the *L. infantum* ortholog contains the eleven GSSPs (red colored) which are conserved in *L. donovani*. (C) A representative MS/MS spectrum for identification of genome search specific peptide NNAETQLTTAER is shown.

Fig. 5. Distribution of proteins identified in the 2 life stages of *L. donovani*. Distribution of proteins identified in *L. donovani* between the 2 different life stages i.e. promastigote (insect stage) and amastigote (human stage).

Fig. 6.

Feature of proteins identified in *L. donovani*. (A) Distribution of membrane and secreted proteins identified in *L. donovani*. (B) Distribution of *L. donovani* proteins based on biological function. (C) Distribution of *L. donovani* proteins based on subcellular localization. A great percentage of proteins remain unclassified.

Table 1

Summary of *Leishmania* species whose genomes have been sequenced.

	Organism	Year	Genome size	Chromosomes	Protein coding genes
1.	<i>Leishmania major</i>	2004	32.8 Mbp	36	8311
2.	<i>Leishmania infantum</i>	2007	32.1 Mbp	36	8154
3.	<i>Leishmania braziliensis</i>	2007	30 Mbp	35	8153
4.	<i>Leishmania mexicana</i>	2011	32 Mbp	34	8250
5.	<i>Leishmania donovani</i>	2011	32.4 Mbp	36	8195

Table 2List of ATP binding cassette proteins identified in *L. donovani*.

	Gene	Protein	Unique peptides	Life stage
1	<i>LDBPK_290640</i>	ATP-binding cassette protein subfamily A, member 10	15	Amastigote and promastigote
2	<i>LDBPK_210770</i>	ATP-binding cassette protein subfamily E, member 1	14	Amastigote and promastigote
3	<i>LDBPK_250540</i>	ATP-binding cassette protein subfamily B, member 1	13	Amastigote and promastigote
4	<i>LDBPK_341060</i>	P-glycoprotein	11	Amastigote and promastigote
5	<i>LDBPK_190800</i>	ATP-binding cassette protein subfamily F, member 2	9	Amastigote and promastigote
6	<i>LDBPK_330340</i>	ATP-binding cassette protein subfamily F, member 3	8	Amastigote and promastigote
7	<i>LDBPK_150950</i>	ATP-binding cassette protein subfamily G, member 4	6	Promastigote
8	<i>LDBPK_282660</i>	MRP protein-like protein	4	Amastigote and promastigote

Table 3List of transmembrane and secreted proteins identified in *L. donovani*.

	Gene	Protein	Biological role	Life stage
1	<i>LDBPK_362720</i>	Membrane-bound acid phosphatase	It is a membrane bound acid phosphate involved in protection against oxidative stress to intracellular <i>Leishmania</i>	Amastigote and promastigote
2	<i>LDBPK_150370</i>	Ecotin	It is an inhibitor of host serine peptidases. It modulates <i>Leishmania</i> infectivity and phagocytic activity by inhibiting macrophage serine peptidases [79]. It is also involved in normal development of <i>Leishmania</i> in infected macrophages.	Amastigote and promastigote
3	<i>LDBPK_010660</i>	HSP70-like protein	Stress response protein that has been shown to play crucial role not only in intracellular survival of parasite but also during infection process of host macrophages	Promastigote
4	<i>LDBPK_367280</i>	Protein disulfide isomerase	It is thioredoxin domain containing protein involved in formation of disulfide bonds in newly synthesized proteins [80,81].	Amastigote and promastigote
5	<i>LDBPK_080740</i>	Tuzin	It is a surface protein	Promastigote
6	<i>LDBPK_070600</i>	Cysteine peptidase B	Cysteine peptidase B was shown to inhibit IL-12 production from macrophages stimulated with bacterial lipopolysaccharide (LPS) by degrading IKappaB alpha and beta as well as NF-kappaB.	Amastigote and promastigote
7	<i>LDBPK_281310</i>	Glucose-regulated protein 78	It is a member of HSP 70 family and it is localized to the endoplasmic reticulum.	Amastigote and promastigote
8	<i>LDBPK_291230</i>	Tryparedoxin-like	It is a oxidoreductase involved in maintaining reducing environment within <i>Leishmania</i>	Amastigote and promastigote