# RNA-seq generates new insights into *Leishmania* differentiation

Carolyn A Paisie

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2017

Committee:

Peter J Myler

David R Crosslin

Program Authorized to Offer Degree:

Biomedical and Health Informatics

University of Washington

## Abstract

RNA-seq generates new insights into *Leishmania* differentiation

Carolyn A Paisie

Chair of the Supervisory Committee:

Peter J Myler, PhD

Professor & Director of Core Services, Center for Infectious Disease Research

Director of Seattle Structural Genomics Center for Infectious Disease (SSGCID)

Affiliate Professor, Biomedical Informatics and Medical Education

*Leishmania donovani,* an intracellular parasitic trypanosomatid, causes kala-azar, a fatal form of visceral leishmaniasis in humans. Infection occurs through a cycle whereby parasites (promastigote stage) living in the midguts of female sand flies are transferred to the host via a bite from an infected female sand fly, are phagocytosed by human macrophages, and are then transferred to phagolysosomes of human macrophages (amastigote stage).

With the large increase in data that is generated by large-scale next-generation sequencing experiments, we embarked upon a systematic organization of the Myler lab next-generation

sequencing data. This was necessary as using spreadsheets to track data had become impractical due to the long-term nature of several of the experiments, in addition to the turnover in lab personnel. We developed standard terminology and nomenclature, both key for ensuring consistency of organization among the various personnel involved in these experiments (biologists, bioinformaticians, and collaborators). We also developed two different data organization systems: one for the organization of raw data and one for the organization of analyzed data. Finally, we created a webpage to document the Myler lab data organization system and serve as a resource for those who are storing or searching for the data.

Previous studies have demonstrated that *L. donovani* differentiation is regulated by changes in gene expression. Thus, we performed high throughput spliced leader RNA-sequencing to elucidate changes in transcript abundance for all cellular mRNAs during *Leishmania donovani* differentiation from promastigotes into amastigotes. Analyses revealed 534 statistically significant (p-value $< 0.05$ and mean $\log_2$ fold-change $\geq 1$ or $\leq -1$) genes and *K*-medians clustering of these genes revealed at least 6 different gene expression patterns (up early, late, or transiently; down early, late, or transiently). We also identified genes which encoded proteins (*e.g.* putative paraflagellar rod protein 1D, glucose transporter 2) that were expected or likely to be differentially expressed during promastigote-to-amastigote differentiation due to the morphological and environmental changes the parasite experiences during this process. In addition, it appears that the technique we have employed, spliced leader RNA-sequencing, allows us to detect gene expression changes in specific members of gene families, in contrast to the microarray studies we previously undertook, as well as providing insight into how post-transcriptional and post-translational regulation may have a role in mRNA expression changes

during differentiation.  Finally, this technique provides additional information about the sequences present in the 5’ untranslated regions of genes that can be used to improve genome annotation and may have a role in the generation of alternative transcripts.

## TABLE OF CONTENTS

**List of Figures**



**List of Appendices**

# Chapter 1. Introduction

## 1.1 Infectious disease

Infectious diseases are one of the main causes of human mortality worldwide[1] and a subset are considered neglected[2]. The neglected diseases consist of 13 parasitic (helminthic and protozoan) and bacterial infections, which along with dengue are the diseases with the highest burden; 20 additional infections include those caused by fungus, viruses, and ectoparasites[2,3]. One way to control these diseases is through the use of preventative chemotherapy, with worldwide economic benefits having been seen as early as the 1900s[4]. Preventative treatment is also cost effective[5] which is key because the financial burdens of these diseases is immense. For example, the cost of treatment for leishmaniasis may equal a family's yearly income and may lead to selling of assets (e.g. livestock, land)[6]. Thus, developing a better understanding of these diseases, and the agents that cause them, will lead to global health improvements.

## 1.2 Leishmaniasis

Leishmaniasis is a group of parasitic diseases ranging from skin lesions which can heal without treatment (cutaneous leishmaniasis [CL]) to mucosal lesions[7] to fatal visceral disease (visceral leishmaniasis [VL]) and is caused by the trypanosomatid protozoan parasites of the *Leishmania* genus[8,9]. As a group, the trypanosomatidae are unicellular and flagellated, whose members include many important human and animal pathogens[9]. *Leishmania* parasites are found worldwide, with certain species (e.g. *Leishmania major* [*L. major*]) being more prevalent in the old world and other species (e.g. *Leishmania infantum* [*L. infantum*], *Leishmania braziliensis* [*L. braziliensis*]) being more prevalent in the new world[9].

There are more than 397 million people worldwide at risk of contracting these diseases, with an estimated 12 million infected people[10]. Recent estimates of the global disease burden caused by leishmaniasis suggest that 0.4 million cases of VL and 1.2 million cases of CL occur every year in 98 countries and three territories in which leishmaniasis is endemic[7]. However, 90% of all worldwide cases of VL occur in only six countries (India, Bangladesh, Sudan, South Sudan, Ethiopia, and Brazil) and between 70 and 75% of all worldwide cases of CL occur in only ten countries (Afghanistan, Algeria, Colombia, Brazil, Iran, Syria, Ethiopia, North Sudan, Costa Rica, and Peru)[7]. More than 20 species of *Leishmania* are capable of infecting humans and causing leishmaniasis, however, the *Leishmania donovani* (*L. donovani*) complex, which includes *L. infantum*, is responsible for most cases of VL worldwide[11,12]. The *Leishmania mexicana* (*L. mexicana*) and *L. braziliensis* complexes are responsible for the majority of CL in the new world *vs L. major* and *Leishmania tropica* (*L. tropica*) in the old world[13].

VL and CL are distinct subtypes of leishmaniasis with different symptoms and effects. VL, also known as kala-azar[14], is present throughout both the old and new world; there are approximately 50,000 deaths each year from this disease[15]. VL infection results in fever, substantial weight loss, anemia, and swelling of the liver and spleen[16,17]. CL, both local and diffuse forms, is characterized by the formation of skin ulcers; these ulcers may be present for as short as 3-5 months or as long as 15-20 years[18]. Diffuse CL may also be characterized by the presence of fever and overall poor condition of the patient[18]. Diagnosis of leishmaniasis is not straightforward as the currently available tests do not provide a 100% specificity and sensitivity[19]. One method for diagnosis involves the microscopic identification of amastigotes in stained patient smears from skin lesions, the liver, spleen, lymph nodes, or bone marrow[20,21].

Other methods include polymerase chain reaction (PCR)[21–24], the indirect fluorescent antibody test, the enzyme linked immunosorbent assay test (more useful for VL diagnosis), and leishmanin skin test (for CL)[25,26]. Although there are drugs, such as pentavalent antimony[27], amphotericin B[27], miltefosine[27], and paromomycin[28–31], available for the treatment of leishmaniasis, they are toxic, costly, and/or require long-term treatment[32,33]. Additionally, these drugs may have varying efficacies in different regions and, despite treatment, leishmaniasis may recur[17,34]. Thus, there is a need for a better understanding of the basic biology of this disease to develop new methods for its prevention and/or treatment.

The majority of cases of leishmaniasis are zoonoses, which affect poor people who live in rural and natural areas that contain a variety of domestic and wild reservoir hosts as well as sand fly vectors that effectively maintain the cycle of infection[35]. To control the spread of disease, it is necessary to consider preventative strategies that do not focus solely on human and hosts and insect vectors but also include the environment in which the latter are found[36]. Currently, practices rely mainly on the early diagnosis and treatment of human disease, control of the disease vectors, and, occasionally, management of reservoir hosts (*e.g.* treatment, elimination)[36]. Control of the sand fly vector is typically done by residual spraying indoors; however, there is concern regarding the development of resistance to dichlorodiphenyltrichloroethane (DDT), particularly in highly endemic areas where spraying is difficult to sustain[33,37,38]. Additionally, no human vaccine against leishmaniasis is currently available, with most candidate vaccines in early research and development, although some have entered clinical trials[38–40]. Thus it is important to develop strategies for the prevention of the disease and a clear and comprehensive understanding of host-pathogen-vector interactions, particularly at the molecular level, is vital[41]. Potentially,

this will allow for the development of new measures that focus on ways to control the reservoir hosts.

### 1.3 *Leishmania* lifecycle

The lifecycle of *Leishmania* includes two main stages: 1) promastigote forms that exist in the sand fly alimentary tract; and 2) amastigote forms that replicate and survive in mammalian macrophage phagolysosomes[42]. See Figure 1 for an overview of the *Leishmania* lifecycle[43]. *Leishmania* parasites are transmitted from their sand fly host to humans, as well as other animal reservoirs, via the bite of an infected sand fly[8]. *Leishmania* replicate within the midgut of the sand fly as flagellated procyclic promastigotes and are not infective at this stage[8]. These procyclics then enter a stationary phase in which they differentiate and enter the metacyclic stage; at this stage, the parasites have undergone adaptive changes which allow them to be transmitted to a mammalian host[44]. These infective metacyclics, once present in the host, are phagocytosed by macrophages where they undergo yet another transformation into non-motile, intracellular, replicative amastigotes capable of surviving in the harsh phagolysosome environment[8].

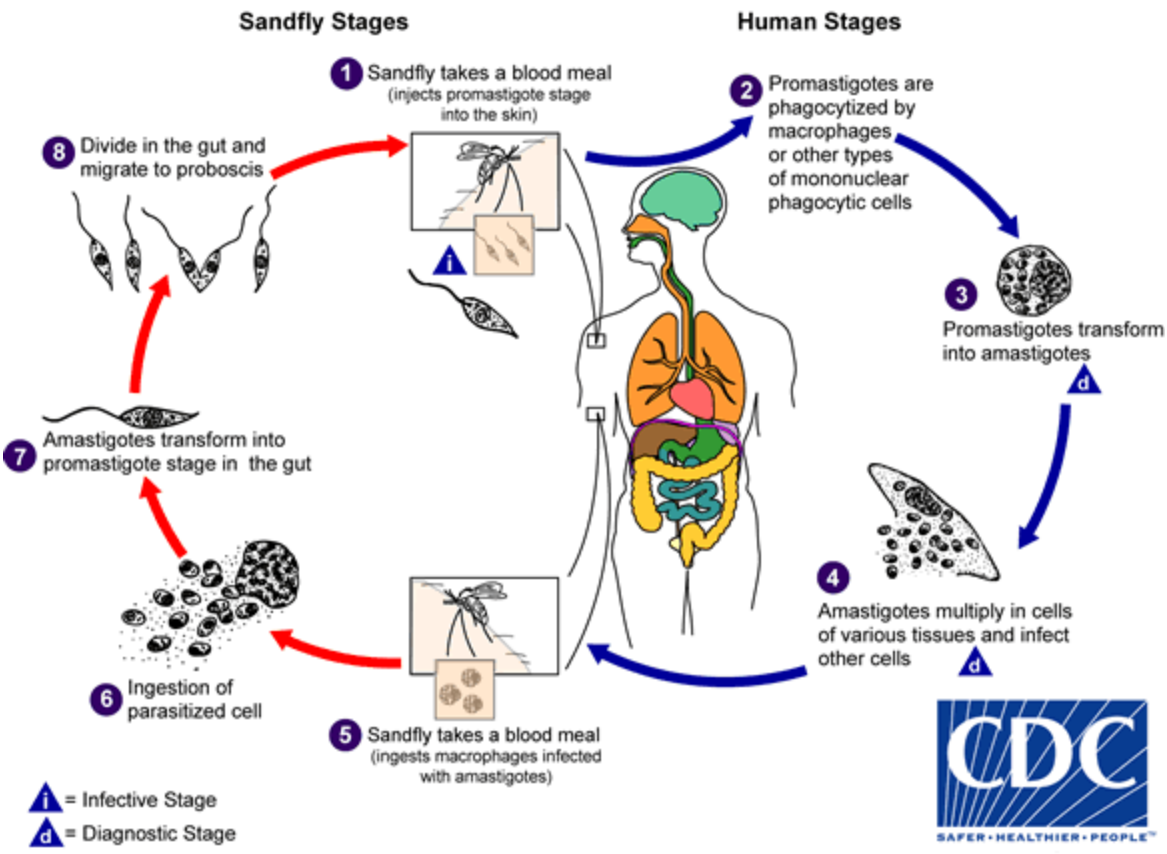


**Figure 1 – *Leishmania* lifecycle[43] in both the sand fly vector and human host**

During the course of differentiation, the parasite's morphology changes such that it becomes ovoid rather than elongated and it loses its ability to be motile[42]. Additionally there are changes in the cell surface composition which include the loss of the lipophosphoglycan coat and the upregulation of amastin surface proteins[45]. There is also a change in the source of carbon from glucose and proline to the beta-oxidation of fatty acids and an increase in the use of amino acids[46,47]. These changes allow the parasite to survive while living in either the more permissive sand fly gut environment or the harsh macrophage phagolysosome environment.

### 1.4 Genomics

With the discovery of the double helix in 1953 and the first DNA sequencing performed in 1968, the modern-day sequencing era began in 1977 when Maxim and Gilbert developed the chemical method and Sanger, Nicklen, and Coulson developed the dideoxy method[48]. Prior to 1995, only viral and organelle genomes were able to be completely sequenced[48]. The first free-living organism sequenced was *H. influenzae* in 1995[49]; shortly followed by eukaryotic, eubacterial, and archaebacterial genomes[48]. This marked the first time whole genome shotgun assembly was seen and this now commonly used method involves randomly fragmenting a genome and then performing computational reassembly[49]. In 2005, array-based pyrosequencing was introduced[50], followed quickly by sequencing-by-synthesis and sequencing-by-ligation, which led to the explosion in the generation of DNA sequencing data[51]; also contributing was the 1000-fold drop in the costs of sequencing since 1990[52].

In 1990 the US Human Genome Project (HGP) was presented to Congress with the proposal that called for a 15 year timeline at a cost of $3 billion[48]. In addition to sequencing the human genome, this proposal also included mapping, and, in certain cases, sequencing, model organisms such as *E. coli* (bacteria), *S. cerevisiae* (yeast), *C. elegans* (worm), *D. melanogaster* (fruit fly), and *M. domesticus* (mouse)[48]. Milestones in the HGP included a detailed genetic map including 5840 mapped loci in 1994, the first completely sequenced chromosome (chromosome 22)[53] in 1999, and draft human genome sequences from Celera[54] and the public project[55] that were published in 2001. NGS has come a long way since 1990, and a number of large-scale projects have been undertaken since the completion of the HGP, including the 1000 Genomes Project[56] and the ENCODE project[57]. With the improvements to sequencing techniques over the

years, it is now possible to sequence very large molecules (> 200 kilobases) which has resulted in a very large amount of data that requires computational analysis[48]. Today, the sequence of the human genome, in combination with many major pathogens, is impacting human health *via* the diagnosis, treatment, and prevention of diseases[48]. Additionally, these sequences have been used to identify possible drug targets[58,59] and vaccine candidates[60,61].

The genomics era has come to infectious disease research, where scientists are still catching up with the large amounts of data produced and incorporating these approaches into their research[49]. There are now more than 38,000 genome sequences available in public databases, including the genomes of numerous human pathogens, which act as a valuable resource for infectious disease research in areas such as genetic diversity, pathogenesis, evolution, detection, and treatment[62]. Additionally, next-generation sequencing (NGS) is now beginning to be used in clinical diagnosis *via* detection of pathogens in patient samples or isolates[63]. However, sequencing is often just the beginning of developing an understanding of the survival of pathogens and the mechanisms of disease[49]. It is likely that both bench research and *in silico* analyses will be necessary to truly understand diseases such as leishmaniasis and to determine strategies for combating them.

Within the field of NGS, there are many different techniques. One of these techniques is RNA-sequencing (RNA-seq), which can be used for the profiling of transcriptomes and genomes[64–67], and offers an alternative to the use of microarrays for large-scale gene expression studies[68]. Use of RNA-seq has allowed for avoidance of some of the technical issues (*e.g.* nonspecific hybridization, dynamic range) associated with microarrays[68]. This means that it is possible to

identify more differentially expressed genes, with higher fold-change, *via* RNA-seq studies *vs.* microarray studies, as well as performing these studies in organisms that lack a complete reference genome[68]. This may be particularly applicable in the realm of infectious disease research as the available genome(s) for some pathogen species, although complete, may contain errors as well as for those pathogens that lack a complete genome.

### 1.5 *Leishmania* biology

Compared to other organisms, the genome of *Leishmania* is fairly small (~36 megabases) and the majority of chromosomes are diploid [69]. The genomes of Old World *Leishmania* species contain 36 chromosomes as compared to New World *Leishmania* species that may contain 34 chromosomes (*e.g. L. mexicana*) or 35 chromosomes (*e.g. L. braziliensis*)[70].

In contrast to the genomes of other organisms, the genomes of *Leishmania* contain long polycistronic gene clusters, have a high gene density, and almost completely lack introns[71]. *Leishmania* also contain a 39-nucleotide spliced leader (SL) RNA which defines the 5' end of all nuclear-encoded mRNAs[72] and SL RNA transcription *via* RNA polymerase II is driven by two upstream elements[73-77]. This is responsible for a large amount of a cell's transcriptional activity[78]; polycistronic transcription results in the synthesis of pre-mRNAs in stretches of thousands of bases to produce monocistronic mRNAs[8]. This polycistronic transcription utilizes transcription initiation sites that typically occur at divergent strand switch regions[79], regions that involve polycistronic transcription units whose origins are in opposite directions on opposing DNA strands[80–82]. Post-transcriptional processing involves the simultaneous occurrence of *trans*-splicing and polyadenylation with the location of the SL acceptor site of the downstream

gene dictating the location of polyadenylation of the upstream gene[83–85]. Previous studies have suggested that polyadenylation sites are found 500-600 nucleotides upstream of the coupled *trans*-splicing acceptor site[83].

It has been suggested that the fact that the genome of *Leishmania* is organized in polycistronic transcription units, rather than individual gene transcription units where RNA polymerase II initiates transcription[86], accounts for the almost complete lack of transcriptional regulation of gene expression[87], as well as making the genomes of species causing CL and VL so similar[71]. It is likely, because of these differences, that transcriptional regulation of mRNA levels is drastically reduced and that post-transcriptional regulation mechanisms are more important factors in changes in gene expression levels during differentiation[8]. Interestingly, 3' untranslated regions (UTRs) of transcripts have previously been shown to have a role in differentiation *via* a role in the control of mRNA stability[45,88-92] and translation[104-107]. More specifically, conserved sequences within the 3' UTRs of a number of *Leishmania* transcripts have previously been suggested to control translation in a stage-specific manner[104,107,108]. Although it is known that *Leishmania* parasites undergo morphological changes (*e.g.* changes in size or shape of organelles) and changes in the components of the cell surface throughout differentiation[99–101], much less is known about how transcriptional changes occur at the global level[79].

Protein levels are regulated by alternative events such as mRNA processing, control of mRNA stability or translation, or post-translational modifications (PTMs)[102]. PTMs (*e.g.* glycosylation, phosphorylation, methylation, acetylation) have been shown to have a role in protein function regulation in other species but there is limited data on the role of such modifications in the

control of gene expression during *Leishmania* differentiation[103]. However, the limited data does support the role of PTMs in regulating gene expression during differentiation[103]. An axenic culture system has been developed such that promastigotes are moved from an environment that mimics that of a sand fly (26°C, pH 7.4) to one that mimics that of a mammalian macrophage phagolysosome (37°C, pH 5.5., 5% $CO_2$)[15,104,105]. Barak *et al.* utilized this system to perform an in-depth examination of *L. donovani* differentiation; their results suggested that differentiation consists of four phases, based on cell morphology: 1) phase I: 0-5 hours (hr) after differentiation signal and when signals are perceived; 2) phase II: 5-10 hr and when parasites stop moving and begin to aggregate; 3) phase III: 10-24 hr and when cells change morphologically into amastigote-shaped cells; 4) phase IV: 24-120 hr and when amastigotes mature[106]. There are transient changes in both protein and mRNA abundance during differentiation[102]; in phases I and II, some variations in mRNA abundance correlate with temporary changes in SL RNA abundance[107]. Previous studies have also demonstrated that 10-15 hr after the start of differentiation the majority of changes in protein abundance have occurred; very limited numbers of proteins have significant changes in expression during the initial phase of differentiation[103].

### 1.6 *Leishmania* genomics

The Leishmania Genome Network initiative was established in 1994 in Rio de Janeiro, Brazil with the goal of determining whole genome sequences of several important species of *Leishmania* that were infecting humans[41]. The first complete genome sequence of *L. major* (CL) was published in 2005[108], followed by sequences for *L. infantum* (VL) and *L. braziliensis* (mucocutaneous leishmaniasis)[109]. More recently, genome sequences for *L. mexicana* (CL)[110], *L. donovani* (VL)[111], and *Leishmania amazonensis* (*L. amazonensis*) (CL)[112] have been completed.

RNA-seq studies of the differentiation of promastigotes to amastigotes in *Leishmania* species often utilize the *in vitro* axenic culture system, described above, that can generate sufficient quantities of starting material. This system has been utilized by a number of different groups to study *Leishmania* biology in general as well as to examine global mRNA and/or protein expression throughout differentiation and during development without necessitating the inclusion of material from host cells[95,113–118].

Previous studies have characterized a few differentially expressed genes during differentiation[98,119–125]; of note, compared to the approximately 10% of genes whose protein abundance appears to change throughout the parasite life-cycle[95,126-128], only approximately 1-3.5% of genes have changes in mRNA abundance[8,113,129]. Previous *Leishmania* microarray studies have demonstrated that gene expression fold-change is relatively modest (approximately 2-10 fold) which suggests smaller changes in mRNA abundance are important for gene regulation, perhaps as a consequence of the relatively stable and predictable environment that is provided by mammalian or sand fly hosts[8]. Microarray studies have also contributed to the assessment of differential gene expression genome-wide[113,114,130]. Additional studies in *L. donovani* have revealed a large number of genes whose expression is up- or down-regulated transiently during differentiation; these genes are expressed at different times than genes whose expression change is fixed[115]. This suggests that, at least in *L. donovani*, changes in gene expression following exposure of promastigotes to the differentiation signal occur in a specific order and ultimately lead to morphological and physiological changes that allow for the survival and growth of amastigotes in macrophage phagolysosomes[115].

Current research involving –omics (*e.g.* genomics, proteomics) technologies is becoming increasingly important for understanding disease pathogenesis in humans and potential drug-resistance mechanisms to currently available antileishmanial medicines; these studies often focus on aspects of disease phenotype[131], mechanism of action of current medications[132], and parasite biology[133]. In looking specifically at protein-coding genes, a fairly small number of *Leishmania*-specific genes have been identified and the majority of these genes encode proteins with no known function[110]. Studies have also found that most *Leishmania* genes are expressed constitutively throughout parasite differentiation from promastigotes to amastigotes[134] while post-transcriptional mechanisms, such as those responsible for mRNA levels, translation rates, and protein stability, are hypothesized to be key for protein abundance regulation[134]. Analyses of whole genome sequencing data sets from *Leishmania* species responsible for causing CL and VL provide a means by which host-parasite interactions could be studied to identify important mechanisms responsible for different types of infections[41]. That is, an analysis of differentiation, and the ensuing changes in gene expression, is important for understanding why certain species of *Leishmania* stay at the sand fly bite site and cause CL while others migrate to internal organs and cause VL[135]. Additionally, these data sets provide a means by which to study parasite gene expression during differentiation in vertebrate hosts[41].

An understanding of gene expression changes during the course of *Leishmania* differentiation is key, not only to understand how the process of differentiation occurs in each of the different hosts, but also for the effective treatment and cure of leishmaniasis[8]. Additionally, an understanding of these changes was one of the main forces behind genome-wide efforts to

identify new drug and vaccine candidates[136,137]. Possible candidates include genes that are expressed in parasitic life stages that are capable of invading and surviving within the vertebrate host[136]. A potential area for further investigation involves genes that are specifically expressed in amastigotes as these genes are likely vital for survival in mammalian hosts[138] as previous studies have demonstrated the effectiveness of identifying key survival pathways *via* the identification of genes whose expression is stage-specific[8].

### 1.7 Data Storage

Challenges arising as a result of embracing NGS approaches include data storage, analysis, and computational requirements; direct data storage costs are also a consideration[52]. The main contributor to the storage requirement is "unprocessed" data; however, data volume is not as much of an issue as constantly growing sequencing capacity[52]. This is because, in the absence of lower storage costs, there either must be a decrease in the data that are stored or an increase in budgets for storage[52]. Possible approaches to address storage needs include: 1) compressing stored data; 2) adding storage; and 3) throwing away some data[52]. It is likely that some combination of these three approaches will be the most efficient and cost-effective way to store data in the future. Arguments for not storing all sequence data include just storing the actual DNA sample, throwing out older data or data from samples that can be re-done, and storing only the output from data analysis, rather than the raw data itself[52]. When considering discarding data, one idea for consideration revolves around unaligned reads; while this may be a valid method for reducing the amount of data stored, there is still debate in the field regarding how much of this data should be discarded[52]. One consideration is that, during the course of projects spanning several years, it will be necessary to be able to easily access the data, particularly for

sub-projects that stem from the original proejct[52]. The decrease in storage costs also factors in, with the bulk of the cost of storing data coming early in the project[52]. However, biologists need to consider the effects of constantly improving DNA sequencing technologies; better methods for data compression need to be developed to ensure there is a balance between data generation and storage[52].

### 1.8 RNA-sequencing data analysis

One of the challenges in today's genomics era is the analysis of NGS data[139]. This stems from the multiple steps that are involved, from the processing of raw reads to performing the read mapping and counting to determining differential gene expression, and the number of programs that exist to perform each of these steps. More specifically, RNA-seq data analyses usually include the following: 1) quality check and pre-processing of raw sequencing reads; 2) mapping reads to a reference genome or transcriptome; 3) counting reads mapped to genes or transcripts; 4) differential gene expression analysis; and 5) biological interpretation[68,140,141]. Although numerous algorithms[142] exist, there is not a consensus of which programs are best to ensure accurate read mapping, gene quantification, and data normalization[68]. One consideration at the level of read mapping is the quality of the reference genome; the presence of assembly errors, repetitive regions, and/or assembly gaps all are associated with problems in correctly mapping reads[68]. Additionally, a list of differentially expressed genes only provides a starting point for understanding the biological implications such as the role of these genes in disease or molecular mechanisms[68]. Thus, there is a need for communication between the biologists who generate the RNA-seq libraries and the bioinformaticians who analyze the data to return biologically meaningful results.

Looking more specifically at the algorithms used for data analyses, the selection of which programs to use often varies, depending on the design of the RNA-seq experiment. The methods that underlie these algorithms also play a role in the decision of which ones to use. For quality control checks on raw RNA-seq data, the FastQC[143] program is often used to give an overview of any potential problems (*e.g.* low-quality bases, presence of adapter sequences) that should be addressed prior to conducting further data analysis[68]. For performing read mapping, several algorithms have been designed in the last several years: TopHat2[144]; STAR[145,146]; GSNAP[147]; OSA[148]; MapSplice[149]; Bowtie 2[150]. One difference in the read mapping algorithms stems from what each program is optimized to detect; for example, GSNAP is ideal for detecting single nucleotide polymorphisms or insertions/deletions while STAR is ideal for mapping long reads[151]. Multiple programs also exist for read counting, including the commonly used Cufflinks[152], featureCounts[153], HTSeq[154], and RSEM[155]. One major difference is the method upon which these algorithms are based; they can be split into either transcript-based approaches (*e.g.* RSEM[155]) or union-exon-based approaches (*e.g.* featureCounts[153]). Union-exon-based methods merge all overlapping exons from the same gene into union exons; a read for a gene is counted if it sufficiently overlaps any of the associated union exons and can be assigned to a gene with much higher confidence[68]. Thus, union-exon-based approaches for counting reads is frequently used for RNA-seq studies[156]. Numerous algorithms (DESeq[157], edgeR[158,159], GENE-Counter[160], NOISeq[161], NBPSeq[162], and Cuffdiff2[163]) exist for performing differential gene expression analysis for RNA-seq data sets. However, there is not a consensus in the field regarding study design and the choice of software for performing these analyses on RNA-seq data sets. A recently published study by Zhang *et al.* compared the use of DESeq, edgeR, and Cuffdiff2 for

performing differential expression analyses and found that one program was not best in all types of differential expression analyses[164]. edgeR may be slightly better than the other two programs for determining true positives and Cuffdiff2 is not recommended for differential expression analyses at the gene level, especially for samples with low sequencing depth[164].

Data normalization is a key step in the data analysis process not only because of its role in accurately determining gene expression but also because of its role in subsequent downstream analysis[68]. One consideration centers around library sizes (total number of mapped reads), which are usually different for each sample, and thus do not allow for direct comparison between samples[68]. One possible way to address this issue involves rescaling total read counts, but this is complicated by the fact that RNA-seq read counts reflect the relative abundances of genes in a sample which means that it is dependent on gene expression level and length as well as the RNA population composition[68]. This means that, if a few genes are very highly expressed, they may account for a large number of the total sample reads and thus prevent reads from other genes from being counted[68].

Accounting for technical limitations is also an important part of the interpretation of analyzed RNA-seq data. When RNA-seq was first developed, the library generation protocol did not retain strand information for individual transcripts, a key shortcoming[68]. However, more recent updates to the protocols have allowed for the development of stranded or strand-specific protocols which retain this information[165]. Previous studies have found that the data using these updated protocols is more reliable as it allows for the correct determination of expression of reads found on overlapping genes and antisense RNA[166]. Another consideration when interpreting analyzed RNA-seq data stems from the potential presence of bias due to preparing or

fragmenting RNA or library construction[68]. For example, bias can be introduced as a consequence of variability in sequencing depth across the transcriptome or genome due to certain sites being more amendable to fragmentation, variability in priming, and the effect of nucleotide composition on tags[167,168].

## Chapter 2. Data organization

### 2.1 Introduction

Data management and organization has often been overlooked in the life sciences and was often seen as something that was not particularly important or applicable for data analysis[169]. However, with the advent of large-scale NGS experiments, and the large amount of data they generate, as discussed above, data organization has become an important part of the NGS analysis pipeline. A consideration when developing a data organization system is determining who, on the research team, needs to be able to access the data and what individual member's needs are, in order to most efficiently and effective store, manage, index, and integrate the data[169]. In developing a data organization system, spreadsheets are often used in order to track sample information and data files[170]; however, there is often a lack of standardization which means that different lab members may store different information or use a different format. This can make it difficult when reconciling multiple spreadsheets to see what data exists and can also lead to mistakes and redundancies[170].

Data organization is very important in research labs where there is often turnover in personnel as well as collaborations with other groups. This is especially relevant with NGS projects that may span multiple years and involve the transfer of data from the original investigator to his/her successor[169]. If a system is not in place for the organization of the raw data, analyzed data, and metadata, it can become very hard for co-workers or collaborators to understand what the data is and what the final results mean[171]; this is particularly true when combining data from multiple NGS experiments that may have been done at different times by different people. Furthermore, having well-organized data is important when publishing a manuscript – it is crucial to know

exactly what data was used and how the analysis was done so that the raw data files and analysis details can be shared with the larger scientific community[171].

Conducting biological research in today's –omics era has led to a various challenges associated with different aspects of the data – one issue is related to the interpretation and level of data integration[172–179], with consideration of what constitutes data ranging from genomic data to protein-protein interactions[180]. One way to define biological data is to consider it to be molecular sequence information and bench experimental information from genome and gene product analyses[68]. Defining what constitutes biological data *vs.* analyzed data is important in the context of data organization and management because a system that makes sense for biological data may not make sense for analyzed data. For example, it may make sense to organize raw data directly downloaded from the sequencing machine into separate folders based on when the sequencing was completed. However, a project may contain multiple libraries sequenced on different days and thus it may make sense to store the analyzed data using a series of folders organized by projects.

Data management, storage, retrieval, analysis, and interpretation lie at the center of any biological research study, particularly with the advent of NGS[180], and a consideration is whether standards exist for such processes[180]. If not, it is important to design and implement such standards, where standards may be defined as certain terminology or structure to represent biological entities and biological entities are defined as all types of biological information[180]. The use of standards allow for re-use of data, in addition to making it easier to share data[180].

One consideration when organizing a large volume of NGS data revolves around how to be efficient in managing various types of files (*e.g.* BAM, count) in order to obtain information that is both relevant and accurate[181]. This is especially important because of the number of researchers that are typically involved in NGS projects who need to be able to access the data and quickly determine the key information (*e.g.* program used for alignment, the genome used for alignment)[182].

**2.2 Myler lab NGS data organization**

In the context of organizing the Myler lab NGS data, we first defined terminology to apply to our data. This was necessary because we decided upon one system to store pre-analyzed data (see Figure 2) and another system to store analyzed data (see Figure 3). Additionally, we created a Confluence[183] webpage, accessible to Myler lab members and collaborators, as a means of sharing the organization system and standard terminology (discussed below) we developed.

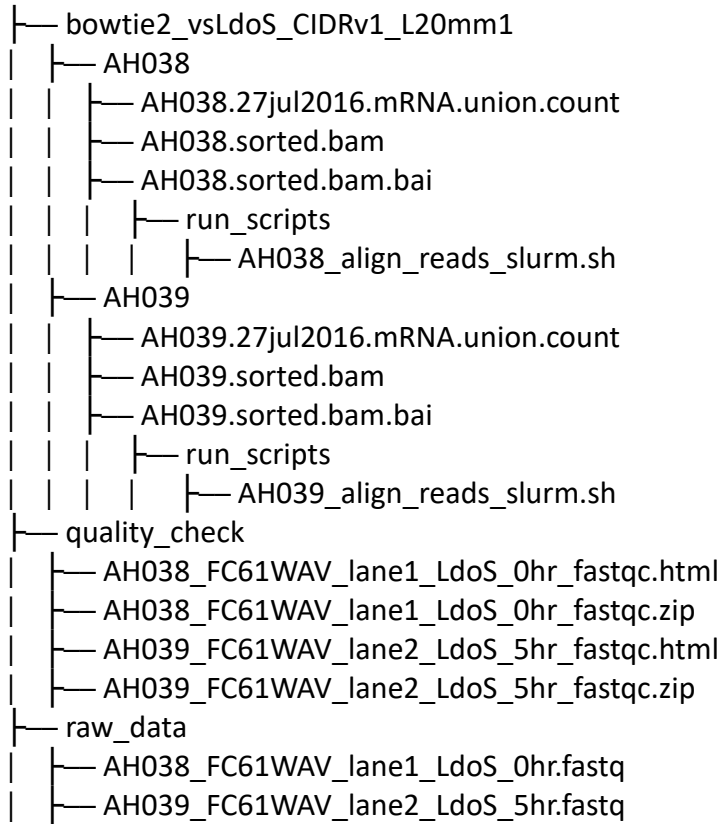


**Figure 2 – Example of folder system for pre-analyzed data**

Pre-analyzed data is considered to be raw sequence files (fastq files), files generated following alignment (sorted and indexed BAM files), and files generated following read counting (count files). Although SAM and BAM files are generated during the alignment process, and would be considered pre-analyzed data, we decided only to keep the sorted and indexed BAM files to minimize the amount of data stored. Analyzed data is considered to be all other files (*e.g.* spreadsheets containing normalized read counts generated by edgeR, multiple experiment viewer (MeV)[184] analysis files). We then defined standard terminology to use when naming folders or files; for example, including the name of the alignment program as part of the folder name where sorted and indexed BAM files are stored. Additionally, standards included the use of capitalization and punctuation to ensure uniformity and to follow already established conventions (*e.g.* abbreviating *L. donovani* as Ldo).

In this system, all pre-analyzed data from libraries sequenced on the same flowcell are organized into a main folder named as follows: yyyymmdd_Flowcell_lastnameofPI_organism_experimentdescription_experimenttype (*e.g.* 20160317_HTNF3BGXX_Zilberstein_LdoS_LdoSdiff_RNASeq). Sub-folders are then used as the main locations for specific types of data: 1) folder for raw data (raw_data); 2) folder for quality check files (quality_check); and 3) folder(s) for alignment and count files (alignmentprogram_genome_version). Within the alignmentprogram_genome_version folder, there are sub-folders for each specific library; the library sub-folders contain the sorted and indexed BAM files and count files as well as another folder (run_scripts) containing the scripts used to generate the data. In the quality_check folder, the output of the FastQC program (described above) is stored. Finally, the raw_data folder contains all the raw data as fastq files. We decided upon this system as it allows for all lab members (both current and future), as well as collaborators, to understand where to find specific data files and to determine how the alignment was done and what genome and version were used.

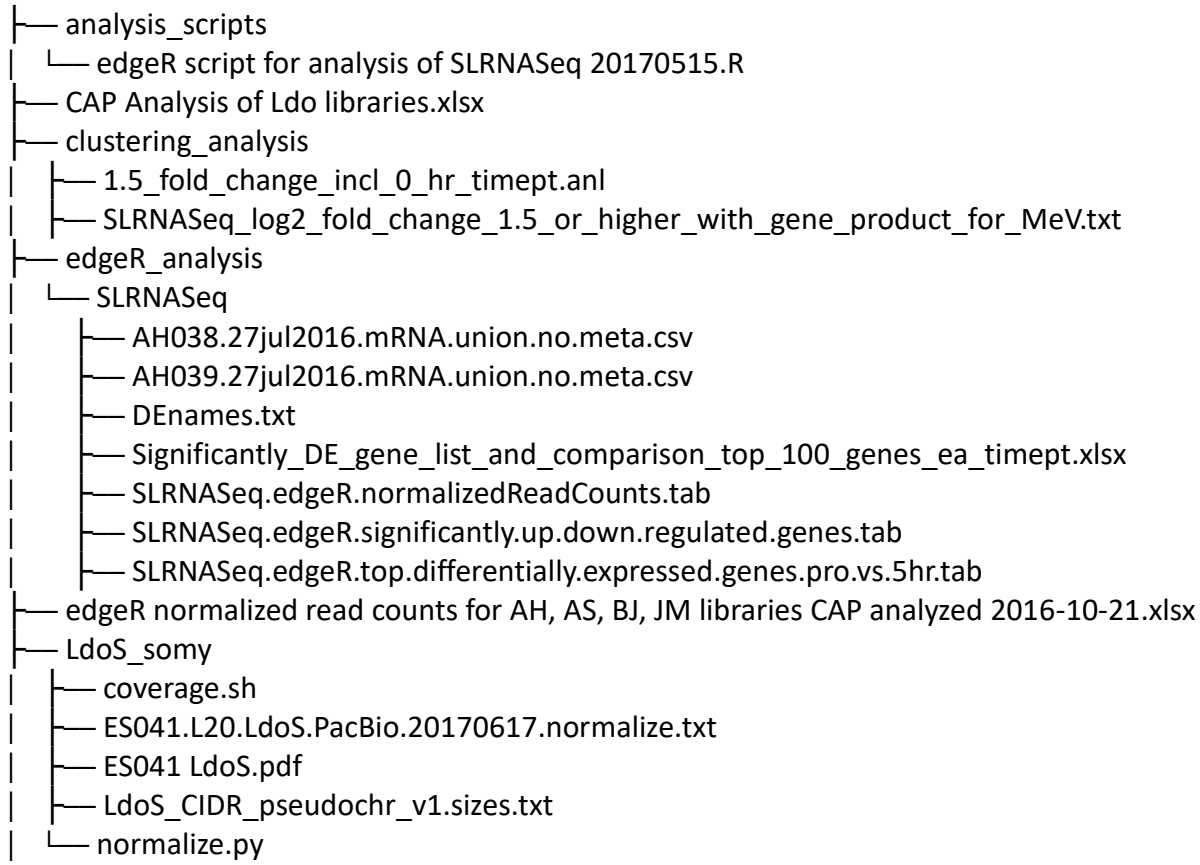


**Figure 3 – Example of folder system for analyzed data**

In the analyzed data system, we use top level folders organized based on project type, which allows data from multiple different NGS experiments to be combined. Files stored in the top-level folder contain analyzed data that are suitable for distribution (*e.g.* shared with collaborators, used in preparing a manuscript for publication). There are then sub-folders named based on the type of analysis done that contain intermediate analyzed data files; these may be files that contain data only for a single library (*e.g.* SLRNASeq.edgeR.top.differentially.expressed.genes.pro.vs.5hr.tab that contains edgeR data for a single library) or files that are used for input into a program (e.g. SLRNASeq_log2_fold_change_1_or_higher_with_gene_product_for_MeV.txt that is used for clustering analysis in MeV). This system was implemented as it allows the analyzed data to be

easily shared, either with collaborators or other lab members, as well as making it easy to re-analyze data.

In summary, we have described here a system for the organization and management of NGS data in the Myler lab. We started the process by first deciding how to efficiently store the data and then developing standard terminology and folder and file naming system. This allows all lab members and collaborators to quickly and effectively find the necessary data. We have also created a webpage to document this system.

## Chapter 3. RNA-seq provides new insight into transcriptome dynamics during promastigote-to-amastigote differentiation of *Leishmania donovani*

### 3.1 Introduction

*Trypanosomatidae* is a family of ancient single-cell eukaryotes whose protein coding genes lack individual RNA polymerase II promoters. Instead, they are organized in clusters of tens to hundreds of genes that are transcribed into polycistronic preRNAs from only a small number of promoters[185]. mRNA maturation occurs *via* concomitant *trans*-splicing of SL RNA 5' of each coding sequence (CDS) and polyadenylation of the 3' end of its upstream neighbor[186]. Organisms that cannot regulate transcription rely on post-transcriptional and translational processes to regulate gene expression[187]. *L. donovani* is a parasitic trypanosomatid that causes kala-azar, a fatal form of VL in humans. These organisms cycle between the midgut of female sand flies, where they reside as extracellular flagellated promastigotes, and phagolysosomes of human macrophages, where they live as aflagellated amastigotes[188,189]. Promastigote-to-amastigote differentiation can be mimicked in host-free axenic culture by shifting promastigotes to a lysosome-like environment (pH 5.5 and 37ºC in 5% CO2). Our laboratory has used a number of biochemical and genomic approaches to characterize *L. donovani* differentiation, establishing that it is regulated by a specific program of changes in gene expression[102,118,190].

The dynamics of mRNA and protein abundance during *Leishmania* differentiation indicate that gene expression is regulated by changes in mRNA levels at early (0-10 hr) time points, while translational and post-translational regulation is more important later (24-120 hr) in the process[102]. For example, many enzymes of the intermediary metabolism are up-regulated in protein abundance, even though mRNA levels remain unchanged throughout differentiation.

Here, we have employed high throughput RNA-seq to elucidate changes in transcript abundance for all cellular mRNAs during *L. donovani* differentiation from promastigotes into amastigotes.

## 3.2 Materials and methods

### *3.2.1 Cell culture and sampling*

*L. donovani* promastigotes and amastigotes were cultured as described previously[102,191]; RNA was extracted as described previously for biological replicates 1 and 2[102]. RNA was extracted for biological replicate 3 using TRI Reagent (Molecular Research Center, Inc), following the standard protocol with the following modification: 50 mL of cell culture per time point was used.

### *3.2.2 Preparation and sequencing of SL libraries*

SL RNA-seq libraries were prepared and sequenced for biological replicates 1 and 2 as described previously[192]. SL RNA-seq libraries were prepared for biological replicate 3 as follows. Briefly, poly (A) RNA was purified using the Dynabead mRNA purification kit (ThermoFischer). One hundred ng mRNA was annealed to the first strand primer (20 μM Random New Primer) and mixed with 10 mM RNA grade dNTPs, incubated at 65°C for 5 minutes (min), and then cooled to 4°C. Samples were then mixed with Superscript III, Superscript III First Strand Buffer, and 0.1M DTT and then incubated at 25°C for 5 min, 50°C for 60 min, and 70°C for 15 min. Samples were then treated with RNAse H and incubated at 37°C for 20 min. Ampure XP purification was performed by adding Ampure XP beads to each sample, incubating at room temperature for 10 min, and placing on the magnet until supernatant was clear. Beads were washed twice with freshly prepared 80% EtOH and dried. Sample was eluted from beads with 10 mM Tris, pH 8. Second strand synthesis was performed by adding 10X NEBbuffer 2, 10 μM SL

primer, and nuclease free water. Samples were heated to 95°C and allowed to cool to room temperature for 30 min to bind SL primer. To create the second strand product, NEB Klenow Fragment, and 10 mM dNTPs were added to each sample and incubated at 37°C for 60 min. The Qiaquick PCR purification kit (Qiagen) was used to purify the second strand product. Product was amplified using KAPA Hifi DNA polymerase Mastermix, 10 μM P5 primer, 10 μM P7 primer, and water in a final volume of 50 μl. The following amplification parameters were used: initial denaturation at 95°C for 3 min; followed by 2 cycles of 94°C for 2 min, 40°C for 2 min, and 72°C for 1 min; 10 cycles of 94°C for 10 seconds (sec), 60°C for 30 sec, and 72°C for 1 min; final extension at 72°C for 5 min; cooling to 4°C. PCR products were purified using PCR purification kit (Qiagen) and sequenced on the NextSeq 500 using the SL Sequencing Primer. Primer sequences were as follows:

Random new:

CTCTTCCGATCTNNNNNNN

SL new:

TCAGTTTCTGTA

P5:

AATGATACGGCGACCACCGAGATCTCACTCTTTCCCTACATCAGTTTCTGTAC

P7 (with Illumina index 2 underlined):

CAAGCAGAAGACGGCATACGAGAT<u>ACATCG</u>GTGACTGGAGTTCAGACGTGTGCTCT

TCCGATCT

SL sequencing primer:

CACCGAGATCTCACTCTTTCCCTA CATCAGTTTCTGTACTTTATTG

*3.2.3 RNA-seq data processing and analysis*

Sequencing read quality was verified using FastQC v0.11.5[143]. Reads were aligned using bowtie2-2.1.0[150] to the *L. donovani* 1S genome using standard mapping parameters. SAM files were converted to BAM files and mapping statistics were determined using samtools-0.1.19[193]. Read counts-per-gene were generated using HTseq[154]; as the libraries sequenced were SL libraries, an annotation file that includes 5' and 3' UTRs was required for HTseq.

Differential gene expression analysis was performed as described previously[192]; RNA genes were excluded from all analyses and replicate 2 2.5 and 15 hr time points were excluded from clustering analysis. The likelihood ratio test, implemented as part of our differential gene expression analysis using the edgeR Bioconductor package[158,159], was used to determine the p-value for the determination of statistical significance (mean $\log_2$ fold-change $\geq 1$ or $\leq -1$ and p-value $< 0.05$). $\text{Log}_2$ fold-changes were calculated *via* comparison of each time point to the 0 hr time point. Correlation coefficients for mean of trimmed mean of M-values (TMM) normalized read counts were calculated using Microsoft Excel.

*3.2.4 Cluster analysis*

*K*-medians clustering[194] was performed using TMeV, part of the TM4 software package[195]. These analyses utilized the mean $\log_2$ fold-changes for mRNA abundance at the following

differentiation time points (5, 7.5, 10, 24, 120 hr) for those mRNAs with a p-value < 0.05 (calculated using the likelihood ratio test as implemented in edgeR as described above) and a mean $\log_2$ fold-change ≥ 1 or ≤ -1. The following parameters were used: 1) cluster genes into 9 clusters by *K* medians for 50 iterations using the Euclidean distance metric; 2) hierarchical clustering of each *K* medians clustering cluster using Gene Trees with Gene Leaf Order optimization using the Euclidean distance metric and complete linkage clustering.

### 3.3 Results

We have previously analyzed RNA samples from eight different time points during *L. donovani* promastigote-to-amastigote differentiation[102,115]. We have now used RNA-seq on the same samples to further analyze differentiation-related changes in mRNA abundance. SL RNA-seq libraries were made from eight time points (0, 2.5, 5, 7.5, 10, 15, 24, and 120 hr) with replicate 1 consisting of libraries for 0, 5, 10, 24, and 120 hr, replicate 2 consisting of libraries for 0, 2.5, 5, 7.5, 10, 15, 24, and 120 hr, and replicate 3 consisting of libraries for 0, 7.5, 24, and 120 hr. Each library contained 3-32 million reads, of which 56-98% could be aligned to the *L. donovani* genome (Appendix 1). To correct for differences in library size, we utilized TMM normalization in our edgeR analysis.

In order to examine overall changes in mRNA abundance during promastigote-to-amastigote differentiation, we compared the mean number of TMM normalized SL reads associated with each gene in the 0 hr library with each of the other seven libraries; see Figure 4. These comparisons revealed that there appears to be little change in the mean of TMM normalized read counts at different time points as compared to the 0 hr time point. There is good correlation ($r^2$ >

0.5) for all time points except for 2.5 and 5 hr. It is possible that this is a result of only having a single replicate for each of these time points; this is supported by the much higher correlation ($r^2$ = 0.84) for the 120 hr time point which has three replicates and for the 7.5 hr time point ($r^2$ = 0.87) which has two replicates.

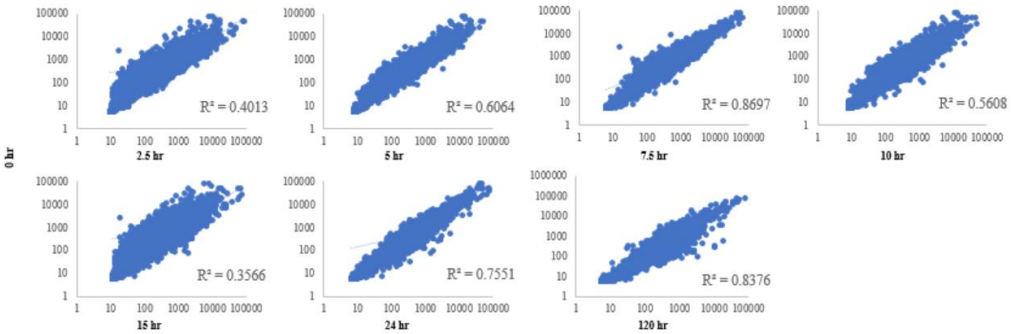


**Figure 4 – Mean of TMM normalized read counts of all mRNAs. The y-axis is the mean of TMM normalized read counts for the 0 hr time point and the x-axis is the mean of TMM normalized read counts for indicated time points.**

To more closely examine genes whose expression changes during differentiation, we first determined those genes that had either a mean $\log_2$ 1-fold increase or decrease at each time point as compared to 0 hr (see figure 5). Comparison of 0 and 5 hr samples revealed 133 genes with a mean $\log_2 \geq$ 1-fold increase and 879 genes with a mean $\log_2 \geq$ 1-fold decrease; of these genes, 39 are statistically significant ($p < 0.05$ calculated using the likelihood ratio test as implemented in edgeR as described above and mean $\log_2$ fold-change $\geq 1$ or $\leq -1$). At later time points, there is a large increase in the number of genes with a mean $\log_2 \geq$ 1-fold-change; at 120 hr (mature amastigotes) 631 genes showed mRNA levels with a mean $\log_2$ fold-change that increased at least 1-fold and 639 genes showed mRNA levels with a mean $\log_2$ fold-change that decreased at least 1-fold (see figure 5); 399 genes at 120 hr are statistically significant ($p < 0.05$ and mean

$\log_2$ fold-change ≥ 1 or ≤ -1) (see figure 5). Overall, 4753 mRNAs have a mean $\log_2$ fold-change that varies by 1-fold or more in at least one time point.

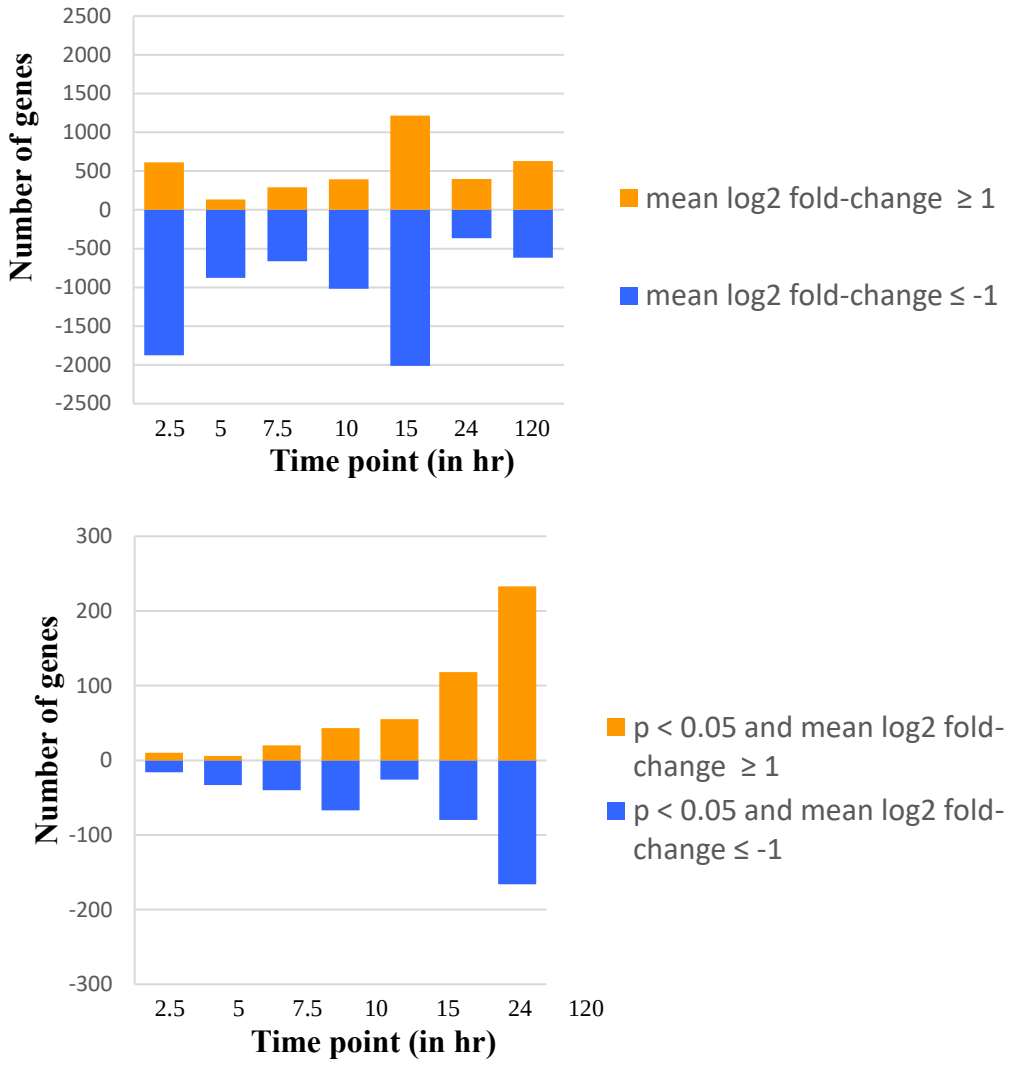


**Figure 5 – Number of genes with mean $\log_2$ fold-change ≥ 1 or ≤ -1 at each time point and statistically significant (p-value < 0.05 and mean $\log_2$ fold-change ≥ 1 or ≤ -1) genes at each time point**

To more closely examine mRNA expression patterns during differentiation, we performed clustering analysis on the set of 534 statistically significant (p-value < 0.05 and mean $\log_2$ fold-change ≥ 1 or ≤ -1) genes. Two popular methods for performing clustering include *K*-means and

*K*-medians clustering; the main variation between these two methods involved the use of either the mean or median as the center of individual clusters[196]. *K*-medians clustering involves separating data into *k* clusters, utilizing the median as the center of the cluster, and attempts to minimize the distance between the center and all other data points within the cluster[196]. As our RNA-seq data set initially consisted of ~8600 genes (excluding RNA genes), we decided to perform *K*-medians clustering on a subset of these genes; including all genes was unlikely to yield meaningful results due to the noisiness of the data. We initially considered subsetting the genes to include all genes with at least a mean $\log_2$ 1-fold increase or decrease in at least one time point; however, to minimize false positives, we decided to include the additional criterion that genes have a p-value < 0.05 as well as exclude the 2.5 and 15 hr time points. Thus, we used a subset of 534 statistically significant (p-value < 0.05 and mean $\log_2$ fold-change $\geq 1$ or $\leq -1$) genes.

Prior to performing *K*-medians clustering, we considered possible gene expression patterns that could be represented in our data set: 1) genes whose expression increases early in differentiation; 2) genes whose expression increases late in differentiation; 3) genes whose expression increases transiently during differentiation; 4) genes whose expression decreases early during differentiation; 5) genes whose expression decreases late during differentiation; and 6) genes whose expression decreases transiently during differentiation. We then performed initial *K*-medians clustering of statistically significant ($p < 0.05$ and mean $\log_2$ fold-change $\geq 1$ or $\leq -1$) genes (Appendix 2) using either 5, 6, 8, 9, 10, or 12 clusters (data not shown for clusters of sizes 5, 6, 8, 10, and 12). However, we found that using 10 or 12 clusters resulted in several clusters only containing a small number (e.g. 5 or 6) of genes and one or two clusters containing a very

large number (e.g. 175 or 200) of genes (data not shown). When we visually inspected the gene expression patterns in these 10 or 12 clusters, we also found that multiple clusters often had the same expression pattern (e.g. genes had increased expression early in differentiation), which suggested that the genes should likely be part of the same cluster. When we considered using 5, 6, 8, or 9 clusters, we found that 8 clusters resulted in multiple clusters with the same expression pattern but clustered all transiently expressed genes together (data not shown) while 5 or 6 clusters resulted in clusters with mixed expression patterns that were also somewhat noisy (data not shown). Therefore, we selected 9 as the number of *K*-medians clusters to use in our analysis of the 534 statistically significant (p-value $< 0.05$ and mean $\log_2$ fold-change $\geq 1$ or $\leq -1$) genes.

*K*-medians clustering of statistically significant ($p < 0.05$ and mean $\log_s$ fold-change $\geq 1$ or $\leq -1$) genes (Appendix 2) revealed at least 6 different patterns of changes in mRNA abundance (figure 6): up early (cluster 6), up late (clusters 3, 4, 5), up transiently (cluster 9), down early (cluster 1), down late (cluster 2), and down transiently (clusters 7, 8). The magnitude of most changes in these statistically significant ($p < 0.05$ and mean $\log_s$ fold-change $\geq 1$ or $\leq -1$) genes is relatively modest, with only 26 genes showing mean $\log_2 > 3$-fold increase in mRNA levels and 32 genes showing mean $\log_2 > 3$-fold decrease in at least one time point; in amastigotes, there are 22 genes showing mean $\log_2 > 3$-fold increase in mRNA levels and 5 genes showing mean $\log_2 > 3$-fold decrease in mRNA levels.

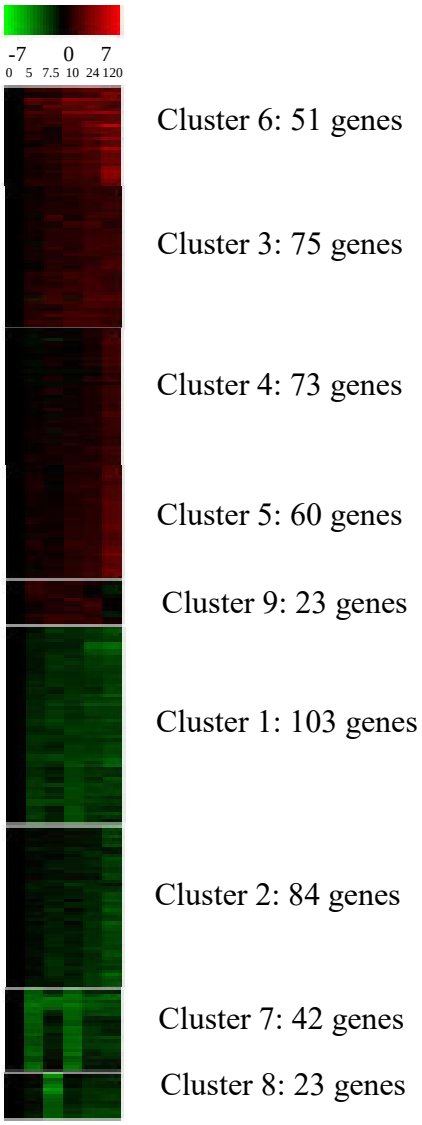


**Figure 6 – *K*-medians clustering analysis of mRNA expression during *Leishmania* differentiation**

We then examined the gene products encoded by genes which are up-regulated (early [cluster 6], late [clusters 3, 4, 5], or transiently [cluster 9]) during differentiation in more detail (see Appendix 2). Cluster 6 contains a number of genes encoding putative amastin surface glycoprotein, amastin-like protein, and hypothetical proteins as well as three genes encoding A2 protein. The *A2* gene family is amastigote-specifc[115] and as such it is not surprising that is up-regulated early during differentiation. In clusters 3, 4, and 5, there were several putative amastin

or amastin-like surface proteins, as well as genes encoding amastin-like proteins. Additional genes include those encoding a putative kinesin, a couple of putative transporters (e.g. ABC transporter), and several hypothetical proteins. This corresponds well with prior studies demonstrating that several genes encoding amastin proteins, as well as a putative kinesin, and transporters are up-regulated late in differentiation[102,115]. Genes that are up-regulated transiently encode several hypothetical conserved proteins. .

We then continued our analysis of by examining gene products encoded by genes which are down-regulated (early [cluster 1], late [cluster 2], and transiently [clusters 7, 8]) during differentiation (see Appendix 2). Genes that are down-regulated early include those that encode a putative paraflagellar rod component, putative paraflagellar rod protein 1D, and glucose transporters; several hypothetical proteins are also encoded. It is not unexpected that these mRNAs are down-regulated early in differentiation as it is amastigotes, not promastigotes, that are found in an environment rich in fatty acids and as such do not use glucose as a nutrient source[197]. Genes encoding paraflagellar rod protein 2C, in addition to several hypothetical proteins, are down-regulated late in differentiation. Downregulation of paraflagellar rod proteins and paraflagellar rod component protein is as expected because of the morphological changes that promastigotes undergo during differentiation into aflagellated amastigotes[115] and agrees with our previous microarray study[102]. Transiently down-regulated genes encode proteins such as a putative dynein heavy chain as well as several hypothetical proteins. Previous studies have demonstrated that dynein heavy chain is transiently down-regulated[115].

### 3.4 Discussion

Previous studies examining transcriptomic and proteomic data sets have provided insight into the regulation and coordination of promastigote-to-amastigote differentiation. We utilized a high throughput RNA-seq approach to further elucidate this process and more specifically examine changes in gene expression. Our results revealed that there appears to be little difference in overall gene expression levels between 0 hr and other time points during differentiation and that there are a modest number (1994) of mRNAs with a mean $\log_2$ fold-change that increased or decreased by at least 1.5-fold in at least one time point during the differentiation process. This suggests that changes in mRNA abundance level may not be as important as other mechanisms of regulation (e.g. post-transcriptional or translational) that contribute to changes in protein expression observed during differentiation. We also identified a number of statistically significant ($p < 0.05$ and mean $\log_s$ fold-change $\geq 1$ or $\leq -1$) differentially expressed genes at different time points during differentiation, with most of these genes being observed at the 120 hr time point. However, statistical significance ($p < 0.05$ and mean $\log_s$ fold-change $\geq 1$ or $\leq -1$) is not always indicative of biological significance nor relevance. For example, a gene may have a mean of TMM normalized read counts of 10 in amastigotes and a mean of TMM normalized read counts of 2 in promastigotes, thus resulting in a $\log_2$ fold-change of 5. However, it is unlikely that this gene is biologically significant and/or relevant since the mean of TMM normalized read counts is so low.

In considering our current SL RNA-seq results in the context of our previous microarray results, we find that there are numerous similarities in genes whose mRNA expression levels change during differentiation (e.g. paraflagellar rod protein 2C, dynein heavy chain, transporters, heat shock proteins)[102,115]. However, in comparing our SL RNA-seq results with our prior microarray

results, it is important to understand a fundamental difference between SL RNA-seq and microarray: SL RNA-seq can be used to detect changes in the 5' UTRs of mRNAs, and thus differentiate between members of a gene family, while microarray cannot because probes are designed to hybridize to the mRNA CDS which remains the same between gene family members. Thus, a possible explanation for why a gene whose expression is decreased in SL RNA-seq and increased in the microarray (e.g. a gene encoding eukaryotic translation initiation factor 3 subunit b) may be that this gene is actually part of a gene family and it is this specific mRNA whose expression level is decreasing as compared to other gene family members whose mRNA expression levels are increasing; the converse may explain why a gene has increased expression in SL RNA-seq and decreased expression in the microarray (e.g. a gene encoding a putative major facilitator superfamily protein). Therefore, SL RNA-seq, in combination with proteomic studies, may more clearly elucidate the relationship between mRNA and protein expression levels during promastigote-to-amastigote differentiation and may provide additional insights into possible regulation of mRNAs and/or proteins.

It is possible that genes whose expression we might expect to change during differentiation (e.g. genes encoding proteins that are involved in transcription or translation or amino acid transport) do not have p-values that meet the criteria for statistical significance (p-value < 0.05). However, even if mRNA levels do not change in a statistically significant ($p < 0.05$ and mean $\log_s$ fold-change $\geq 1$ or $\leq -1$) manner, it does not mean that protein expression does not change. Indeed, we have previously shown that there is poor correlation (~20-30%) between mRNA levels and protein expression in a set of 902 genes for which both mRNA and protein expression data were available[102]. These findings suggest it is possible for mRNA expression to remain the same

during promastigote-to-amastigote differentiation while the associated protein may show an increase or decrease in expression and *vice versa*; potential mechanisms for such changes include post-transcriptional and translational regulation of mRNAs and proteins.

There were numerous genes in our RNA-seq data set that had a mean $\log_2$ fold-change that was less than 1.5-fold for all time points, thus suggesting there is little to no change in RNA expression during differentiation. However, it may be the case that some of these genes may be regulated post-transcriptionally. In this scenario, there may be stabilization of the mRNA throughout the course of differentiation, thus making it appear that there is no change in mRNA expression in amastigotes *vs.* promastigotes. However, such mRNA stabilization may result in increased protein expression, thus explaining the discordance between mRNA and protein expression levels. This is supported by previous studies that have demonstrated that some metabolic enzymes have increased protein levels despite the fact that the corresponding mRNA levels remain the same during differentiation[102], which suggests possible post-transcriptional, rather than post-translational, regulation. It has been shown that alternative *trans*-splicing can occur in *Trypanosoma brucei*, resulting in changes in 5' UTR length[198]; it is possible that this can generate proteins with shorter N-terminal sequences that consequently have altered sub-cellular localization[198,199]. Thus, post-transcriptional regulation may also be involved in the control of protein levels during differentiation *via* the production of alternative transcripts. Alternatively, genes whose expression does not appear to change during the course of differentiation may encode proteins that are post-translationally regulated via PTMs such as phosphorylation or glycosylation which may result in protein degradation or stabilization, thus resulting in discordant mRNA and protein expression levels. Therefore, further study is

necessary to tease apart the roles for the different types of regulation, as well as modifications, which may play a role in regulating both mRNA and protein levels during differentiation, not just in *L. donovani* but also in other *Leishmania* species.

## Chapter 4. Conclusions and future directions

## 4.1 Conclusions

Data organization is crucial for conducting large-scale NGS experiments, particularly with the improvements that have been made in sequencing technology that have resulted in evermore data being generated and projects that span multiple years. We have developed standard terminology and naming conventions and implemented a data organization system for the Myler lab that takes into the accounts the needs of different users, spanning biologists to bioinformaticians to collaborators.

Previous studies have been undertaken, utilizing microarrays[102,115] and proteomics[102,191], to study *Leishmania* differentiation and analyze mRNA and protein expression changes that occur. We performed an RNA-seq analysis of *L. donovani* differentiation utilizing 8 time points that revealed a number of different genes that have at least a mean $\log_2$ 1-fold increase or decrease at one time point during differentiation as well as a number of genes that are statistically significantly (p-value < 0.05 and mean $\log_2$ fold-change ≥ 1-fold or ≤ -1-fold) differentially expressed at each time point compared to promastigotes with amastigotes having the most statistically significantly (p-value < 0.05 and mean $\log_2$ fold-change ≥ 1-fold or ≤ -1-fold) differentially expressed genes. We also performed clustering analysis on a subset of 534 genes that are statistically significantly (p-value < 0.05 and mean $\log_2$ fold-change ≥ 1-fold or ≤ -1-fold) differentially expressed; we found several genes whose mRNA expression changed during differentiation as expected, based either on prior studies[102,115] or known changes in parasite morphology and environment[197].

### 4.2 Future directions

Data from SL RNA-seq experiments, such as the one described here, can also be utilized for the mapping of SL sites and SL site usage[200]. This is important as it provides a means for more precisely determining the boundaries of the 5' UTRs of genes and hence results in improved genome annotations. Determination of SL site usage may also reveal changes in the location of the SL site used which could contribute to mRNA turnover[200] and may explain differences in mRNA levels during promastigote-to-amastigote differentiation. Additionally, it is possible that alterations in the size of the 5' UTR of a gene may result in alternative *trans*-splicing that can then produce alternative transcripts as discussed above.

## WORKS CITED

1. Lozano R, Naghavi M, Foreman K, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet (London, England)*. 2012;380(9859):2095-2128. doi:10.1016/S0140-6736(12)61728-0.
2. Hotez PJ, Molyneux DH, Fenwick A, et al. Control of neglected tropical diseases. *N Engl J Med*. 2007;357(10):1018-1027. doi:10.1056/NEJMra064142.
3. Hotez PJ, Molyneux DH, Fenwick A, Ottesen E, Ehrlich Sachs S, Sachs JD. Incorporating a rapid-impact package for neglected tropical diseases with programs for HIV/AIDS, tuberculosis, and malaria. *PLoS Med*. 2006;3(5):e102. doi:10.1371/journal.pmed.0030102.
4. Bleakley H. Disease and Development: Evidence from Hookworm Eradication in the American South. *Q J Econ*. 2007;122(1):73-117. doi:10.1162/qjec.121.1.73.
5. Hotez PJ, Fenwick A, Savioli L, Molyneux DH. Rescuing the bottom billion through control of neglected tropical diseases. *Lancet (London, England)*. 2009;373(9674):1570-1575. doi:10.1016/S0140-6736(09)60233-6.
6. Adhikari SR, Maskay NM. The economic burden of Kala-azar in households of the Danusha and Mahottari districts of Nepal. *Acta Trop*. 2003;88(1):1-2. http://www.ncbi.nlm.nih.gov/pubmed/12943969. Accessed July 8, 2017.
7. Alvar J, Vélez ID, Bern C, et al. Leishmaniasis Worldwide and Global Estimates of Its Incidence. Kirk M, ed. *PLoS One*. 2012;7(5):e35671. doi:10.1371/journal.pone.0035671.
8. Akopyants NS, Matlib RS, Bukanova EN, et al. Expression profiling using random genomic DNA microarrays identifies differentially expressed genes associated with three major developmental stages of the protozoan parasite Leishmania major. *Mol Biochem Parasitol*. 2004;136(1):71-86. doi:10.1016/j.molbiopara.2004.03.002.
9. Aslett M, Aurrecoechea C, Berriman M, et al. TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Res*. 2010;38(Database):D457-D462. doi:10.1093/nar/gkp851.
10. WHO | Magnitude of the problem. *WHO*. 2014. http://www.who.int/leishmaniasis/burden/magnitude/burden_magnitude/en/. Accessed July 8, 2017.
11. Murray L, Woolgar M, Murray J, Cooper P. Self-exclusion from health care in women at high risk for postpartum depression. *J Public Health Med*. 2003;25(2):131-137. http://www.ncbi.nlm.nih.gov/pubmed/12848402. Accessed January 22, 2016.
12. McCall L-I, Zhang W-W, Matlashewski G. Determinants for the Development of Visceral Leishmaniasis Disease. Chitnis CE, ed. *PLoS Pathog*. 2013;9(1):e1003053. doi:10.1371/journal.ppat.1003053.
13. de Vasquez AM, Saenz RE, Petersen JL, Christensen HA, Johnson CM. Leishmania mexicana complex: human infections in the Republic of Panama. *Am J Trop Med Hyg*. 1990;43(6):619-622. http://www.ncbi.nlm.nih.gov/pubmed/2267966. Accessed July 8, 2017.
14. Rochette A, Raymond F, Corbeil J, Ouellette M, Papadopoulou B. Whole-genome comparative RNA expression profiling of axenic and intracellular amastigote forms of Leishmania infantum. *Mol Biochem Parasitol*. 2009;165(1):32-47. doi:10.1016/j.molbiopara.2008.12.012.
15. Goldman-Pinkovich A, Balno C, Strasser R, et al. An Arginine Deprivation Response Pathway Is Induced in Leishmania during Macrophage Invasion. Muller I, ed. *PLOS Pathog*. 2016;12(4):e1005494. doi:10.1371/journal.ppat.1005494.

16. Herwaldt BL. Leishmaniasis. *Lancet (London, England)*. 1999;354(9185):1191-1199. doi:10.1016/S0140-6736(98)10178-2.
17. Berman JD. Human leishmaniasis: clinical, diagnostic, and chemotherapeutic developments in the last 10 years. *Clin Infect Dis*. 1997;24(4):684-703. http://www.ncbi.nlm.nih.gov/pubmed/9145744. Accessed July 8, 2017.
18. Torres-Guerrero E, Quintanilla-Cedillo MR, Ruiz-Esmenjaud J, Arenas R. Leishmaniasis: a review. *F1000Research*. 2017;6:750. doi:10.12688/f1000research.11120.1.
19. Wijerathna T, Gunathilaka N, Gunawardana K, Rodrigo W. Potential Challenges of Controlling Leishmaniasis in Sri Lanka at a Disease Outbreak. *Biomed Res Int*. 2017;2017:1-9. doi:10.1155/2017/6931497.
20. Nawaratna SSK, Weilgama DJ, Wijekoon CJ, Dissanayake M, Rajapaksha K. Cutaneous leishmaniasis, Sri Lanka. *Emerg Infect Dis*. 2007;13(7):1068-1070. doi:10.3201/eid1307.070227.
21. Barrett MP, Croft SL. Management of trypanosomiasis and leishmaniasis. *Br Med Bull*. 2012;104(1):175-196. doi:10.1093/bmb/lds031.
22. G D, Z G, K S. A Review on Biology, Epidemiology and Public Health Significance of Leishmaniasis. *J Bacteriol Parasitol*. 2013;4(2). doi:10.4172/2155-9597.1000166.
23. Harms G, Schönian G, Feldmeier H. Leishmaniasis in Germany. *Emerg Infect Dis*. 2003;9(7):872-875. doi:10.3201/eid0907.030023.
24. Maleki-Ravasan N, Oshaghi M, Javadian E, Rassi Y, Sadraei J, Mohtarami F. Blood Meal Identification in Field-Captured Sand flies: Comparison of PCR-RFLP and ELISA Assays. *Iran J Arthropod Borne Dis*. 2009;3(1):8-18. http://www.ncbi.nlm.nih.gov/pubmed/22808367. Accessed July 8, 2017.
25. Singh S. New developments in diagnosis of leishmaniasis. *Indian J Med Res*. 2006;123(3):311-330. http://www.ncbi.nlm.nih.gov/pubmed/16778313. Accessed July 8, 2017.
26. Musa AM, Khalil EAG, Raheem MA, et al. The natural history of Sudanese post-kala-azar dermal leishmaniasis: clinical, immunological and prognostic features. *Ann Trop Med Parasitol*. 2002;96(8):765-772. doi:10.1179/000349802125002211.
27. Matlashewski G, Arana B, Kroeger A, et al. Visceral leishmaniasis: elimination with existing interventions. *Lancet Infect Dis*. 2011;11(4):322-325. doi:10.1016/S1473-3099(10)70320-0.
28. Croft SL, Coombs GH. Leishmaniasis--current chemotherapy and recent advances in the search for novel drugs. *Trends Parasitol*. 2003;19(11):502-508. http://www.ncbi.nlm.nih.gov/pubmed/14580961. Accessed July 8, 2017.
29. Sundar S. Drug resistance in Indian visceral leishmaniasis. *Trop Med Int Health*. 2001;6(11):849-854. http://www.ncbi.nlm.nih.gov/pubmed/11703838. Accessed July 8, 2017.
30. Fong D, Chan MM, Rodriguez R, Gately LJ, Berman JD, Grogl M. Paromomycin resistance in Leishmania tropica: lack of correlation with mutation in the small subunit ribosomal RNA gene. *Am J Trop Med Hyg*. 1994;51(6):758-766. http://www.ncbi.nlm.nih.gov/pubmed/7810808. Accessed July 8, 2017.
31. Jhingran A, Chawla B, Saxena S, Barrett MP, Madhubala R. Paromomycin: Uptake and resistance in Leishmania donovani. *Mol Biochem Parasitol*. 2009;164(2):111-117. doi:10.1016/j.molbiopara.2008.12.007.
32. Gillespie PM, Beaumier CM, Strych U, Hayward T, Hotez PJ, Bottazzi ME. Status of

vaccine research and development of vaccines for leishmaniasis. *Vaccine*. 2016;34(26):2992-2995. doi:10.1016/j.vaccine.2015.12.071.
33. Beaumier CM, Gillespie PM, Hotez PJ, Bottazzi ME. New vaccines for neglected parasitic diseases and dengue. *Transl Res*. 2013;162(3):144-155. doi:10.1016/j.trsl.2013.03.006.
34. de Menezes JPB, Guedes CES, Petersen AL de OA, Fraga DBM, Veras PST. Advances in Development of New Treatment for Leishmaniasis. *Biomed Res Int*. 2015;2015:1-11. doi:10.1155/2015/815023.
35. Ready PD. Biology of Phlebotomine Sand Flies as Vectors of Disease Agents. *Annu Rev Entomol*. 2013;58(1):227-250. doi:10.1146/annurev-ento-120811-153557.
36. Otranto D, Dantas-Torres F. The prevention of canine leishmaniasis and its impact on public health. *Trends Parasitol*. 2013;29(7):339-345. doi:10.1016/j.pt.2013.05.003.
37. Coleman M, Foster GM, Deb R, et al. DDT-based indoor residual spraying suboptimal for visceral leishmaniasis elimination in India. *Proc Natl Acad Sci U S A*. 2015;112(28):8573-8578. doi:10.1073/pnas.1507782112.
38. Duthie MS, Raman VS, Piazza FM, Reed SG. The development and clinical evaluation of second-generation leishmaniasis vaccines. *Vaccine*. 2012;30(2):134-141. doi:10.1016/j.vaccine.2011.11.005.
39. Nagill R, Kaur S. Vaccine candidates for leishmaniasis: a review. *Int Immunopharmacol*. 2011;11(10):1464-1488. doi:10.1016/j.intimp.2011.05.008.
40. Rezvan H, Moafi M. An overview on Leishmania vaccines: A narrative review article. *Vet Res forum an Int Q J*. 2015;6(1):1-7. http://www.ncbi.nlm.nih.gov/pubmed/25992245. Accessed July 8, 2017.
41. Cantacessi C, Dantas-Torres F, Nolan MJ, Otranto D. The past, present, and future of Leishmania genomics and transcriptomics. *Trends Parasitol*. 2015;31(3):100-108. doi:10.1016/j.pt.2014.12.012.
42. Fiebig M, Kelly S, Gluenz E. Comparative Life Cycle Transcriptomics Revises Leishmania mexicana Genome Annotation and Links a Chromosome Duplication with Parasitism of Vertebrates. Myler PJ, ed. *PLoS Pathog*. 2015;11(10):e1005186. doi:10.1371/journal.ppat.1005186.
43. Prevention C-C for DC and. CDC - Leishmaniasis - Biology. https://www.cdc.gov/parasites/leishmaniasis/biology.html. Accessed June 26, 2017.
44. Sacks DL. Metacyclogenesis in Leishmania promastigotes. *Exp Parasitol*. 1989;69(1):100-103. http://www.ncbi.nlm.nih.gov/pubmed/2659372. Accessed July 8, 2017.
45. Wu Y, El Fakhry Y, Sereno D, Tamar S, Papadopoulou B. A new developmentally regulated gene family in Leishmania amastigotes encoding a homolog of amastin surface proteins. *Mol Biochem Parasitol*. 2000;110(2):345-357. http://www.ncbi.nlm.nih.gov/pubmed/11071288. Accessed July 9, 2017.
46. Opperdoes FR, Coombs GH. Metabolism of Leishmania: proven and predicted. *Trends Parasitol*. 2007;23(4):149-158. doi:10.1016/j.pt.2007.02.004.
47. McConville MJ, Naderer T. Metabolic Pathways Required for the Intracellular Survival of *Leishmania*. *Annu Rev Microbiol*. 2011;65(1):543-561. doi:10.1146/annurev-micro-090110-102913.
48. Hutchison CA. DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res*. 2007;35(18):6227-6237. doi:10.1093/nar/gkm688.

49. Krol A. The Genomics of Infectious Diseases - Bio-IT World. Bio-IT World. http://www.bio-itworld.com/2014/8/7/genomics-infectious-diseases.html. Published 2014. Accessed June 28, 2017.
50. Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. 2005;437(7057):376-380. doi:10.1038/nature03959.
51. Stein LD. The case for cloud computing in genome informatics. *Genome Biol*. 2010;11(5):207. doi:10.1186/gb-2010-11-5-207.
52. Hsi-Yang Fritz M, Leinonen R, Cochrane G, Birney E. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res*. 2011;21(5):734-740. doi:10.1101/gr.114819.110.
53. Dunham I, Hunt AR, Collins JE, et al. The DNA sequence of human chromosome 22. *Nature*. 1999;402(6761):489-495. doi:10.1038/990031.
54. Venter JC, Adams MD, Myers EW, et al. The Sequence of the Human Genome. *Science (80- )*. 2001;291(5507):1304-1351. doi:10.1126/science.1058040.
55. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;409(6822):860-921. doi:10.1038/35057062.
56. 1000 Genomes Project Consortium RM, Abecasis GR, Altshuler D, et al. A map of human genome variation from population-scale sequencing. *Nature*. 2010;467(7319):1061-1073. doi:10.1038/nature09534.
57. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*. 2004;306(5696):636-640. doi:10.1126/science.1105136.
58. Drews J. Drug discovery: a historical perspective. *Science*. 2000;287(5460):1960-1964. http://www.ncbi.nlm.nih.gov/pubmed/10720314. Accessed July 8, 2017.
59. Hopkins AL, Groom CR. The druggable genome. *Nat Rev Drug Discov*. 2002;1(9):727-730. doi:10.1038/nrd892.
60. Wizemann TM, Heinrichs JH, Adamou JE, et al. Use of a whole genome approach to identify vaccine molecules affording protection against Streptococcus pneumoniae infection. *Infect Immun*. 2001;69(3):1593-1598. doi:10.1128/IAI.69.3.1593-1598.2001.
61. Adu-Bobie J, Capecchi B, Serruto D, Rappuoli R, Pizza M. Two years into reverse vaccinology. *Vaccine*. 2003;21(7-8):605-610. http://www.ncbi.nlm.nih.gov/pubmed/12531326. Accessed July 8, 2017.
62. Fournier P-E, Dubourg G, Raoult D, et al. Clinical detection and characterization of bacterial pathogens in the genomics era. *Genome Med*. 2014;6(11):114. doi:10.1186/s13073-014-0114-2.
63. Lefterova MI, Suarez CJ, Banaei N, Pinsky BA. Next-Generation Sequencing for Infectious Disease Diagnosis and Management: A Report of the Association for Molecular Pathology. *J Mol Diagn*. 2015;17(6):623-634. doi:10.1016/j.jmoldx.2015.07.004.
64. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*. 2008;5(7):621-628. doi:10.1038/nmeth.1226.
65. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10(1):57-63. doi:10.1038/nrg2484.
66. Costa V, Angelini C, De Feis I, Ciccodicola A. Uncovering the complexity of transcriptomes with RNA-Seq. *J Biomed Biotechnol*. 2010;2010:853916. doi:10.1155/2010/853916.
67. Mutz K-O, Heilkenbrinker A, Lönne M, Walter J-G, Stahl F. Transcriptome analysis

using next-generation sequencing. *Curr Opin Biotechnol*. 2013;24(1):22-30. doi:10.1016/j.copbio.2012.09.004.
68. Abdurakhmonov IY. *Bioinformatics : Updated Features and Applications.* INTECH; 2016.
69. Martínez-Calvillo S, Stuart K, Myler PJ. Ploidy changes associated with disruption of two adjacent genes on Leishmania major chromosome 1. *Int J Parasitol*. 2005;35(4):419-429. doi:10.1016/j.ijpara.2004.12.014.
70. Britto C, Ravel C, Bastien P, et al. Conserved linkage groups associated with large-scale chromosomal rearrangements between Old World and New World Leishmania genomes. *Gene*. 1998;222(1):107-117. http://www.ncbi.nlm.nih.gov/pubmed/9813266. Accessed July 9, 2017.
71. Gruber A, Durham AM, Huynh C, Portillo HA del, eds. *Bioinformatics in Tropical Disease Research: A Practical and Case-Study Approach [Internet]*. Bethesda: National Center for Biotechnology Information (US); 2008. https://www.ncbi.nlm.nih.gov/books/NBK6830/#!po=100.000. Accessed July 8, 2017.
72. Clayton CE. Life without transcriptional control? From fly to man and back again. *EMBO J*. 2002;21(8):1881-1888. doi:10.1093/emboj/21.8.1881.
73. Yu MC, Sturm NR, Saito RM, Roberts TG, Campbell DA. Single nucleotide resolution of promoter activity and protein binding for the Leishmania tarentolae spliced leader RNA gene. *Mol Biochem Parasitol*. 1998;94(2):265-281. http://www.ncbi.nlm.nih.gov/pubmed/9747976. Accessed July 9, 2017.
74. Campbell DA, Thomas S, Sturm NR. Transcription in kinetoplastid protozoa: why be normal? *Microbes Infect*. 2003;5(13):1231-1240. http://www.ncbi.nlm.nih.gov/pubmed/14623019. Accessed July 9, 2017.
75. Bindereif A. *RNA Metabolism in Trypanosomes*. Springer; 2012.
76. Palenchar JB, Bellofatto V. Gene transcription in trypanosomes. *Mol Biochem Parasitol*. 2006;146(2):135-141. doi:10.1016/j.molbiopara.2005.12.008.
77. Martínez-Calvillo S, Vizuet-de-Rueda JC, Florencio-Martínez LE, Manning-Cela RG, Figueroa-Angulo EE. Gene expression in trypanosomatid parasites. *J Biomed Biotechnol*. 2010;2010:525241. doi:10.1155/2010/525241.
78. Haanstra JR, Stewart M, Luu V-D, et al. Control and regulation of gene expression: quantitative analysis of the expression of phosphoglycerate kinase in bloodstream form Trypanosoma brucei. *J Biol Chem*. 2008;283(5):2495-2507. doi:10.1074/jbc.M705782200.
79. Dillon LA, Okrah K, Hughitt VK, et al. Transcriptomic profiling of gene expression and RNA processing during Leishmania major differentiation. *Nucleic Acids Res*. 2015;43(14):6799-6813. doi:10.1093/nar/gkv656.
80. Martínez-Calvillo S, Yan S, Nguyen D, Fox M, Stuart K, Myler PJ. Transcription of Leishmania major Friedlin chromosome 1 initiates in both directions within a single region. *Mol Cell*. 2003;11(5):1291-1299. http://www.ncbi.nlm.nih.gov/pubmed/12769852. Accessed July 8, 2017.
81. McDonagh PD, Myler PJ, Stuart K. The unusual gene organization of Leishmania major chromosome 1 may reflect novel transcription processes. *Nucleic Acids Res*. 2000;28(14):2800-2803. http://www.ncbi.nlm.nih.gov/pubmed/10908338. Accessed July 8, 2017.
82. Myler PJ, Audleman L, deVos T, et al. Leishmania major Friedlin chromosome 1 has an

unusual distribution of protein-coding genes. *Proc Natl Acad Sci U S A*. 1999;96(6):2902-2906. http://www.ncbi.nlm.nih.gov/pubmed/10077609. Accessed July 8, 2017.
83. LeBowitz JH, Smith HQ, Rusche L, Beverley SM. Coupling of poly(A) site selection and trans-splicing in Leishmania. *Genes Dev*. 1993;7(6):996-1007. http://www.ncbi.nlm.nih.gov/pubmed/8504937. Accessed July 8, 2017.
84. Matthews KR, Tschudi C, Ullu E. A common pyrimidine-rich motif governs trans-splicing and polyadenylation of tubulin polycistronic pre-mRNA in trypanosomes. *Genes Dev*. 1994;8(4):491-501. http://www.ncbi.nlm.nih.gov/pubmed/7907303. Accessed July 8, 2017.
85. Ullu E, Matthews KR, Tschudi C. Temporal order of RNA-processing reactions in trypanosomes: rapid trans splicing precedes polyadenylation of newly synthesized tubulin transcripts. *Mol Cell Biol*. 1993;13(1):720-725. http://www.ncbi.nlm.nih.gov/pubmed/8417363. Accessed July 8, 2017.
86. Rastrojo A, Carrasco-Ramiro F, Martín D, et al. The transcriptome of Leishmania major in the axenic promastigote stage: transcript annotation and relative expression levels by RNA-seq. *BMC Genomics*. 2013;14(1):223. doi:10.1186/1471-2164-14-223.
87. Soysa R, Carter NS, Yates PA. A dual luciferase system for analysis of post-transcriptional regulation of gene expression in Leishmania. *Mol Biochem Parasitol*. 2014;195(1):1-5. doi:10.1016/j.molbiopara.2014.05.002.
88. D'Orso I, De Gaudenzi JG, Frasch ACC. RNA-binding proteins and mRNA turnover in trypanosomes. *Trends Parasitol*. 2003;19(4):151-155. http://www.ncbi.nlm.nih.gov/pubmed/12689640. Accessed July 9, 2017.
89. Furger A, Schürch N, Kurath U, Roditi I. Elements in the 3' untranslated region of procyclin mRNA regulate expression in insect forms of Trypanosoma brucei by modulating RNA stability and translation. *Mol Cell Biol*. 1997;17(8):4372-4380. http://www.ncbi.nlm.nih.gov/pubmed/9234695. Accessed July 9, 2017.
90. Myung KS, Beetham JK, Wilson ME, Donelson JE. Comparison of the post-transcriptional regulation of the mRNAs for the surface proteins PSA (GP46) and MSP (GP63) of Leishmania chagasi. *J Biol Chem*. 2002;277(19):16489-16497. doi:10.1074/jbc.M200174200.
91. Aly R, Argaman M, Halman S, Shapira M. A regulatory role for the 5' and 3' untranslated regions in differential expression of hsp83 in Leishmania. *Nucleic Acids Res*. 1994;22(15):2922-2929. http://www.ncbi.nlm.nih.gov/pubmed/8065903. Accessed July 9, 2017.
92. Charest H, Zhang WW, Matlashewski G. The developmental expression of Leishmania donovani A2 amastigote-specific genes is post-transcriptionally mediated and involves elements located in the 3'-untranslated region. *J Biol Chem*. 1996;271(29):17081-17090. http://www.ncbi.nlm.nih.gov/pubmed/8663340. Accessed July 9, 2017.
93. Boucher N, Wu Y, Dumas C, et al. A common mechanism of stage-regulated gene expression in Leishmania mediated by a conserved 3'-untranslated region element. *J Biol Chem*. 2002;277(22):19511-19520. doi:10.1074/jbc.M200500200.
94. Larreta R, Soto M, Quijada L, et al. The expression of HSP83 genes in Leishmania infantum is affected by temperature and by stage-differentiation and is regulated at the levels of mRNA stability and translation. *BMC Mol Biol*. 2004;5(1):3. doi:10.1186/1471-2199-5-3.
95. McNicoll F, Drummelsmith J, Müller M, et al. A combined proteomic and transcriptomic

approach to the study of stage differentiation in Leishmania infantum. *Proteomics*. 2006;6(12):3567-3581. doi:10.1002/pmic.200500853.
96. Zilka A, Garlapati S, Dahan E, Yaolsky V, Shapira M. Developmental regulation of heat shock protein 83 in Leishmania. 3' processing and mRNA stability control transcript abundance, and translation id directed by a determinant in the 3'-untranslated region. *J Biol Chem*. 2001;276(51):47922-47929. doi:10.1074/jbc.M108271200.
97. McNicoll F, Müller M, Cloutier S, et al. Distinct 3'-untranslated region elements regulate stage-specific mRNA accumulation and translation in Leishmania. *J Biol Chem*. 2005;280(42):35238-35246. doi:10.1074/jbc.M507511200.
98. Rochette A, McNicoll F, Girard J, et al. Characterization and developmental gene regulation of a large gene family encoding amastin surface proteins in Leishmania spp. *Mol Biochem Parasitol*. 2005;140(2):205-220. doi:10.1016/j.molbiopara.2005.01.006.
99. Ambit A, Woods KL, Cull B, Coombs GH, Mottram JC. Morphological Events during the Cell Cycle of Leishmania major. *Eukaryot Cell*. 2011;10(11):1429-1438. doi:10.1128/EC.05118-11.
100. Beverley SM, Turco SJ. Lipophosphoglycan (LPG) and the identification of virulence genes in the protozoan parasite Leishmania. *Trends Microbiol*. 1998;6(1):35-40. doi:10.1016/S0966-842X(97)01180-3.
101. Wheeler RJ, Gluenz E, Gull K. The cell cycle of Leishmania: morphogenetic events and their implications for parasite biology. *Mol Microbiol*. 2011;79(3):647-662. doi:10.1111/j.1365-2958.2010.07479.x.
102. Lahav T, Sivam D, Volpin H, et al. Multiple levels of gene regulation mediate differentiation of the intracellular pathogen Leishmania. *FASEB J*. 2011;25(2):515-525. doi:10.1096/fj.10-157529.
103. Rosenzweig D, Smith D, Myler PJ, Olafson RW, Zilberstein D. Post-translational modification of cellular proteins during Leishmania donovani differentiation. *Proteomics*. 2008;8(9):1843-1850. doi:10.1002/pmic.200701043.
104. Saar Y, Ransford A, Waldman E, et al. Characterization of developmentally-regulated activities in axenic amastigotes of Leishmania donovani. *Mol Biochem Parasitol*. 1998;95(1):9-20. http://www.ncbi.nlm.nih.gov/pubmed/9763285. Accessed July 8, 2017.
105. Bates PA, Robertson CD, Tetley L, Coombs GH. Axenic cultivation and characterization of Leishmania mexicana amastigote-like forms. *Parasitology*. 1992;105 ( Pt 2):193-202. http://www.ncbi.nlm.nih.gov/pubmed/1454417. Accessed July 8, 2017.
106. Barak E, Amin-Spector S, Gerliak E, Goyard S, Holland N, Zilberstein D. Differentiation of Leishmania donovani in host-free system: analysis of signal perception and response. *Mol Biochem Parasitol*. 2005;141(1):99-108. doi:10.1016/j.molbiopara.2005.02.004.
107. Murray HW, Berman JD, Davies CR, Saravia NG. Advances in leishmaniasis. *Lancet (London, England)*. 2005;366(9496):1561-1577. doi:10.1016/S0140-6736(05)67629-5.
108. Ivens AC, Peacock CS, Worthey E a, et al. The genome of the kinetoplastid parasite, Leishmania major. *Science*. 2005;309(5733):436-442. doi:10.1126/science.1112680.
109. Peacock CS, Seeger K, Harris D, et al. Comparative genomic analysis of three Leishmania species that cause diverse human disease. *Nat Genet*. 2007;39(7):839-847. doi:10.1038/ng2053.
110. Rogers MB, Hilley JD, Dickens NJ, et al. Chromosome and gene copy number variation allow major structural change between species and strains of Leishmania. *Genome Res*. 2011;21(12):2129-2142. doi:10.1101/gr.122945.111.

111. Downing T, Imamura H, Decuypere S, 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(12):2143-2156. doi:10.1101/gr.123430.111.
112. Real F, Vidal RO, Carazzolle MF, et al. The genome sequence of Leishmania (Leishmania) amazonensis: functional annotation and extended analysis of gene models. *DNA Res*. 2013;20(6):567-581. doi:10.1093/dnares/dst031.
113. 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(2):198-218. doi:10.1016/j.molbiopara.2005.12.009.
114. 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(1):35-46. doi:10.1016/j.molbiopara.2006.11.009.
115. Saxena A, Lahav T, Holland N, 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(1):53-65. doi:10.1016/j.molbiopara.2006.11.011.
116. Srividya G, Duncan R, Sharma P, Raju BVS, Nakhasi HL, Salotra P. Transcriptome analysis during the process of in vitro differentiation of Leishmania donovani using genomic microarrays. *Parasitology*. 2007;134(Pt 11):1527-1539. doi:10.1017/S003118200700296X.
117. Walker J, Vasquez J-J, Gomez MA, et al. Identification of developmentally-regulated proteins in Leishmania panamensis by proteome profiling of promastigotes and axenic amastigotes. *Mol Biochem Parasitol*. 2006;147(1):64-73. doi:10.1016/j.molbiopara.2006.01.008.
118. Rosenzweig D, Smith D, Opperdoes F, Stern S, Olafson RW, Zilberstein D. Retooling Leishmania metabolism: from sand fly gut to human macrophage. *FASEB J*. 2008;22(2):590-602. doi:10.1096/fj.07-9254com.
119. Charest H, Matlashewski G. Developmental gene expression in Leishmania donovani: differential cloning and analysis of an amastigote-stage-specific gene. *Mol Cell Biol*. 1994;14(5):2975-2984. http://www.ncbi.nlm.nih.gov/pubmed/7545921. Accessed July 8, 2017.
120. Argaman M, Aly R, Shapira M. Expression of heat shock protein 83 in Leishmania is regulated post-transcriptionally. *Mol Biochem Parasitol*. 1994;64(1):95-110. http://www.ncbi.nlm.nih.gov/pubmed/8078527. Accessed July 8, 2017.
121. Hübel A, Krobitsch S, Hörauf A, Clos J. Leishmania major Hsp100 is required chiefly in the mammalian stage of the parasite. *Mol Cell Biol*. 1997;17(10):5987-5995. http://www.ncbi.nlm.nih.gov/pubmed/9315657. Accessed July 8, 2017.
122. Burchmore RJ, Landfear SM. Differential regulation of multiple glucose transporter genes in Leishmania mexicana. *J Biol Chem*. 1998;273(44):29118-29126. http://www.ncbi.nlm.nih.gov/pubmed/9786920. Accessed July 8, 2017.
123. Handman E, Osborn AH, Symons F, van Driel R, Cappai R. The Leishmania promastigote surface antigen 2 complex is differentially expressed during the parasite life cycle. *Mol*

*Biochem Parasitol*. 1995;74(2):189-200. http://www.ncbi.nlm.nih.gov/pubmed/8719160. Accessed July 8, 2017.
124. Barr SD, Gedamu L. Cloning and characterization of three differentially expressed peroxidoxin genes from Leishmania chagasi. Evidence for an enzymatic detoxification of hydroxyl radicals. *J Biol Chem*. 2001;276(36):34279-34287. doi:10.1074/jbc.M104406200.
125. Moore LL, Santrich C, LeBowitz JH. Stage-specific expression of the Leishmania mexicana paraflagellar rod protein PFR-2. *Mol Biochem Parasitol*. 1996;80(2):125-135. http://www.ncbi.nlm.nih.gov/pubmed/8892290. Accessed July 8, 2017.
126. Bente M, Harder S, Wiesgigl M, et al. Developmentally induced changes of the proteome in the protozoan parasite Leishmania donovani. *Proteomics*. 2003;3(9):1811-1829. doi:10.1002/pmic.200300462.
127. El Fakhry Y, Ouellette M, Papadopoulou B. A proteomic approach to identify developmentally regulated proteins in Leishmania infantum. *Proteomics*. 2002;2(8):1007-1017. doi:10.1002/1615-9861(200208)2:8<1007::AID-PROT1007>3.0.CO;2-G.
128. Nugent PG, Karsani SA, Wait R, Tempero J, Smith DF. Proteomic analysis of Leishmania mexicana differentiation. *Mol Biochem Parasitol*. 2004;136(1):51-62. doi:10.1016/j.molbiopara.2004.02.009.
129. Diehl S, Diehl F, El-Sayed NM, Clayton C, Hoheisel JD. Analysis of stage-specific gene expression in the bloodstream and the procyclic form of Trypanosoma brucei using a genomic DNA-microarray. *Mol Biochem Parasitol*. 2002;123(2):115-123. http://www.ncbi.nlm.nih.gov/pubmed/12270627. Accessed July 9, 2017.
130. Rochette A, Raymond F, Ubeda J-M, et al. Genome-wide gene expression profiling analysis of Leishmania major and Leishmania infantum developmental stages reveals substantial differences between the two species. *BMC Genomics*. 2008;9(1):255. doi:10.1186/1471-2164-9-255.
131. McCall L-I, McKerrow JH. Determinants of disease phenotype in trypanosomatid parasites. *Trends Parasitol*. 2014;30(7):342-349. doi:10.1016/j.pt.2014.05.001.
132. Kaur G, Rajput B. Comparative analysis of the omics technologies used to study antimonial, amphotericin B, and pentamidine resistance in leishmania. *J Parasitol Res*. 2014;2014:726328. doi:10.1155/2014/726328.
133. Pawar H, Renuse S, Khobragade SN, et al. Neglected tropical diseases and omics science: proteogenomics analysis of the promastigote stage of Leishmania major parasite. *OMICS*. 2014;18(8):499-512. doi:10.1089/omi.2013.0159.
134. Requena JM. Lights and shadows on gene organization and regulation of gene expression in Leishmania. *Front Biosci (Landmark Ed*. 2011;16:2069-2085. http://www.ncbi.nlm.nih.gov/pubmed/21622163. Accessed July 8, 2017.
135. Zhang WW, Ramasamy G, McCall L-I, et al. Genetic Analysis of Leishmania donovani Tropism Using a Naturally Attenuated Cutaneous Strain. *PLoS Pathog*. 2014;10(7):e1004244. doi:10.1371/journal.ppat.1004244.
136. Almeida R, Gilmartin BJ, McCann SH, et al. Expression profiling of the Leishmania life cycle: cDNA arrays identify developmentally regulated genes present but not annotated in the genome. *Mol Biochem Parasitol*. 2004;136(1):87-100. doi:10.1016/j.molbiopara.2004.03.004.
137. Duncan RC, Salotra P, Goyal N, Akopyants NS, Beverley SM, Nakhasi HL. The application of gene expression microarray technology to kinetoplastid research. *Curr Mol*

*Med*. 2004;4(6):611-621. http://www.ncbi.nlm.nih.gov/pubmed/15357212. Accessed July 9, 2017.
138. Zhang W-W, Mendez S, Ghosh A, et al. Comparison of the A2 gene locus in Leishmania donovani and Leishmania major and its control over cutaneous infection. *J Biol Chem*. 2003;278(37):35508-35515. doi:10.1074/jbc.M305030200.
139. Hardwick SA, Deveson IW, Mercer TR. Reference standards for next-generation sequencing. *Nat Rev Genet*. June 2017. doi:10.1038/nrg.2017.44.
140. Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Methods*. 2011;8(6):469-477. doi:10.1038/nmeth.1613.
141. Capobianco E. RNA-Seq Data: A Complexity Journey. *Comput Struct Biotechnol J*. 2014;11(19):123-130. doi:10.1016/j.csbj.2014.09.004.
142. Han Y, Gao S, Muegge K, Zhang W, Zhou B. Advanced Applications of RNA Sequencing and Challenges. *Bioinform Biol Insights*. 2015;9(Suppl 1):29-46. doi:10.4137/BBI.S28991.
143. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed July 8, 2017.
144. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. 2013;14(4):R36. doi:10.1186/gb-2013-14-4-r36.
145. Dobin A, Gingeras TR. Mapping RNA-seq Reads with STAR. *Curr Protoc Bioinforma*. 2015;51:11.14.1-19. doi:10.1002/0471250953.bi1114s51.
146. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635.
147. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*. 2010;26(7):873-881. doi:10.1093/bioinformatics/btq057.
148. Hu J, Ge H, Newman M, Liu K. OSA: a fast and accurate alignment tool for RNA-Seq. *Bioinformatics*. 2012;28(14):1933-1934. doi:10.1093/bioinformatics/bts294.
149. Wang K, Singh D, Zeng Z, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res*. 2010;38(18):e178. doi:10.1093/nar/gkq622.
150. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9(4):357-359. doi:10.1038/nmeth.1923.
151. Conesa A, Madrigal P, Tarazona S, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol*. 2016;17(1):13. doi:10.1186/s13059-016-0881-8.
152. Trapnell C, Williams BA, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*. 2010;28(5):511-515. doi:10.1038/nbt.1621.
153. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 2014;30(7):923-930. doi:10.1093/bioinformatics/btt656.
154. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*. 2015;31(2):166-169. doi:10.1093/bioinformatics/btu638.
155. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12(1):323. doi:10.1186/1471-2105-12-323.
156. Zhao S, Xi L, Zhang B. Union Exon Based Approach for RNA-Seq Gene Quantification:

To Be or Not to Be? Jordan IK, ed. *PLoS One*. 2015;10(11):e0141910. doi:10.1371/journal.pone.0141910.
157. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010;11(10):R106. doi:10.1186/gb-2010-11-10-r106.
158. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. 2010;11(3):R25. doi:10.1186/gb-2010-11-3-r25.
159. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139-140. doi:10.1093/bioinformatics/btp616.
160. Cumbie JS, Kimbrel JA, Di Y, et al. GENE-counter: a computational pipeline for the analysis of RNA-Seq data for gene expression differences. Freitag M, ed. *PLoS One*. 2011;6(10):e25279. doi:10.1371/journal.pone.0025279.
161. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA-seq: a matter of depth. *Genome Res*. 2011;21(12):2213-2223. doi:10.1101/gr.124321.111.
162. Di Y, Schafer D, Cumbie J. The NBP negative binomial model for assessing differential gene expression from RNA-Seq. *Stat Appl Genet Mol Biol*. 2011;10(1). https://www.degruyter.com/view/j/sagmb.2011.10.issue-1/sagmb.2011.10.1.1637/sagmb.2011.10.1.1637.xml. Accessed July 9, 2017.
163. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol*. 2012;31(1):46-53. doi:10.1038/nbt.2450.
164. Zhang ZH, Jhaveri DJ, Marshall VM, et al. A Comparative Study of Techniques for Differential Expression Analysis on RNA-Seq Data. Provero P, ed. *PLoS One*. 2014;9(8):e103207. doi:10.1371/journal.pone.0103207.
165. Levin JZ, Yassour M, Adiconis X, et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods*. 2010;7(9):709-715. doi:10.1038/nmeth.1491.
166. Mills JD, Kawahara Y, Janitz M. Strand-Specific RNA-Seq Provides Greater Resolution of Transcriptome Profiling. *Curr Genomics*. 2013;14(3):173-181. doi:10.2174/1389202911314030003.
167. Hansen KD, Brenner SE, Dudoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res*. 2010;38(12):e131. doi:10.1093/nar/gkq224.
168. Risso D, Schwartz K, Sherlock G, Dudoit S. GC-Content Normalization for RNA-Seq Data. *BMC Bioinformatics*. 2011;12(1):480. doi:10.1186/1471-2105-12-480.
169. Bauch A, Adamczyk I, Buczek P, et al. openBIS: a flexible framework for managing and analyzing complex data in biology research. *BMC Bioinformatics*. 2011;12(1):468. doi:10.1186/1471-2105-12-468.
170. Calabria A, Spinozzi G, Benedicenti F, Tenderini E, Montini E. adLIMS: a customized open source software that allows bridging clinical and basic molecular research studies. *BMC Bioinformatics*. 2015;16(Suppl 9):S5. doi:10.1186/1471-2105-16-S9-S5.
171. Bianchi V, Ceol A, Ogier AGE, et al. Integrated Systems for NGS Data Management and Analysis: Open Issues and Available Solutions. *Front Genet*. 2016;7:75. doi:10.3389/fgene.2016.00075.
172. Jansen R, Yu H, Greenbaum D, et al. A Bayesian Networks Approach for Predicting

Protein-Protein Interactions from Genomic Data. *Science (80- )*. 2003;302(5644):449-453. doi:10.1126/science.1087361.
173. Hwang D, Rust AG, Ramsey S, et al. A data integration methodology for systems biology. *Proc Natl Acad Sci*. 2005;102(48):17296-17301. doi:10.1073/pnas.0508647102.
174. Myers CL, Troyanskaya OG. Context-sensitive data integration and prediction of biological networks. *Bioinformatics*. 2007;23(17):2322-2330. doi:10.1093/bioinformatics/btm332.
175. Chung SY, Wong L. Kleisli: a new tool for data integration in biology. *Trends Biotechnol*. 1999;17(9):351-355. http://www.ncbi.nlm.nih.gov/pubmed/10461180. Accessed July 18, 2017.
176. Letunic I, Copley RR, Schmidt S, et al. SMART 4.0: towards genomic data integration. *Nucleic Acids Res*. 2004;32(Database issue):D142-4. doi:10.1093/nar/gkh088.
177. von Mering C, Jensen LJ, Kuhn M, et al. STRING 7--recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res*. 2007;35(Database issue):D358-62. doi:10.1093/nar/gkl825.
178. Cheung K-H, Yip KY, Smith A, Deknikker R, Masiar A, Gerstein M. YeastHub: a semantic web use case for integrating data in the life sciences domain. *Bioinformatics*. 2005;21 Suppl 1(Suppl 1):i85-96. doi:10.1093/bioinformatics/bti1026.
179. Goldovsky L, Janssen P, Ahrén D, et al. CoGenT++: an extensive and extensible data environment for computational genomics. *Bioinformatics*. 2005;21(19):3806-3810. doi:10.1093/bioinformatics/bti579.
180. Lapatas V, Stefanidakis M, Jimenez RC, Via A, Schneider MV. Data integration in biological research: an overview. *J Biol Res (Thessalonike, Greece)*. 2015;22(1):9. doi:10.1186/s40709-015-0032-5.
181. Merelli I, Pérez-Sánchez H, Gesing S, D'Agostino D. Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives. *Biomed Res Int*. 2014;2014:134023. doi:10.1155/2014/134023.
182. de Brevern AG, Meyniel J-P, Fairhead C, Neuvéglise C, Malpertuy A. Trends in IT Innovation to Build a Next Generation Bioinformatics Solution to Manage and Analyse Biological Big Data Produced by NGS Technologies. *Biomed Res Int*. 2015;2015:1-15. doi:10.1155/2015/904541.
183. Get started - Atlassian Documentation. https://confluence.atlassian.com/doc/get-started-777010817.html. Accessed July 10, 2017.
184. Howe EA, Sinha R, Schlauch D, Quackenbush J. RNA-Seq analysis in MeV. *Bioinformatics*. 2011;27(22):3209-3210. doi:10.1093/bioinformatics/btr490.
185. Thomas S, Green A, Sturm NR, Campbell D a, Myler PJ. Histone acetylations mark origins of polycistronic transcription in Leishmania major. *BMC Genomics*. 2009;10:152. doi:10.1186/1471-2164-10-152.
186. Michaeli S. Trans-splicing in trypanosomes: machinery and its impact on the parasite transcriptome. *Future Microbiol*. 2011;6(4):459-474. doi:10.2217/fmb.11.20.
187. Clayton C, Shapira M. Post-transcriptional regulation of gene expression in trypanosomes and leishmanias. *Mol Biochem Parasitol*. 2007;156(2):93-101. doi:10.1016/j.molbiopara.2007.07.007.
188. Chang KP, Dwyer DM. Multiplication of a human parasite (Leishmania donovani) in phagolysosomes of hamster macrophages in vitro. *Science*. 1976;193(4254):678-680. http://www.ncbi.nlm.nih.gov/pubmed/948742. Accessed October 18, 2016.

189. Herwaldt BL. Leishmaniasis. *Lancet (London, England)*. 1999;354(9185):1191-1199. doi:10.1016/S0140-6736(98)10178-2.
190. Tsigankov P, Gherardini PF, Helmer-Citterich M, Zilberstein D. What has proteomics taught us about Leishmania development? *Parasitology*. 2012;139(9):1146-1157. doi:10.1017/S0031182012000157.
191. Tsigankov P, Gherardini PF, Helmer-Citterich M, Späth GF, Myler PJ, Zilberstein D. Regulation dynamics of Leishmania differentiation: deconvoluting signals and identifying phosphorylation trends. *Mol Cell Proteomics*. 2014;13(7):1787-1799. doi:10.1074/mcp.M114.037705.
192. Haydock A, Terrao M, Sekar A, Ramasamy G, Baugh L, Myler PJ. RNA-seq approaches for determining mRNA abundance in Leishmania. *Methods Mol Biol*. 2015;1201:207-219. doi:10.1007/978-1-4939-1438-8_12.
193. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352.
194. Soukas A, Cohen P, Socci ND, Friedman JM. Leptin-specific patterns of gene expression in white adipose tissue. *Genes Dev*. 2000;14(8):963-980. http://www.ncbi.nlm.nih.gov/pubmed/10783168. Accessed July 31, 2017.
195. Saeed AI, Sharov V, White J, et al. TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*. 2003;34(2):374-378. http://www.ncbi.nlm.nih.gov/pubmed/12613259. Accessed July 31, 2017.
196. Whelan C, Harrell G, Wang J. Understanding the K-Medians Problem. *Proc Int Conf Sci Comput*. 2015:219-222.
197. Tsigankov P, Gherardini PF, Helmer-Citterich M, Späth GF, Myler PJ, Zilberstein D. Regulation dynamics of Leishmania differentiation: deconvoluting signals and identifying phosphorylation trends. *Mol Cell Proteomics*. 2014;13(7):1787-1799. doi:10.1074/mcp.M114.037705.
198. Nilsson D, Gunasekera K, Mani J, et al. Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of Trypanosoma brucei. Parsons M, ed. *PLoS Pathog*. 2010;6(8):e1001037. doi:10.1371/journal.ppat.1001037.
199. Rettig J, Wang Y, Schneider A, Ochsenreiter T. Dual targeting of isoleucyl-tRNA synthetase in Trypanosoma brucei is mediated through alternative trans-splicing. *Nucleic Acids Res*. 2012;40(3):1299-1306. doi:10.1093/nar/gkr794.
200. Cuypers B, Domagalska MA, Meysman P, et al. Multiplexed Spliced-Leader Sequencing: A high-throughput, selective method for RNA-seq in Trypanosomatids. *Sci Rep*. 2017;7(1):3725. doi:10.1038/s41598-017-03987-0.

**Appendix 1**

| Library ID | Library Size | % alignment to *Ldo* genome |
|---|---|---|

| | | |
|---|---|---|
| **AH038** | 12741826 | 70.78 |
| **AH039** | 12881673 | 75.49 |
| **AH040** | 12638373 | 72.54 |
| **AH041** | 12593482 | 76.58 |
| **AH042** | 12587670 | 76.99 |
| **AS020** | 4642595 | 57.3 |
| **AS021** | 5780687 | 56.08 |
| **AS022** | 4048873 | 57.85 |
| **AS023** | 4629259 | 57.85 |
| **AS024** | 4160035 | 57.32 |
| **AS025** | 4692891 | 56.57 |
| **AS026** | 3029891 | 57.93 |
| **AS027** | 4642293 | 56.96 |
| **JM033** | 23726555 | 97.91 |
| **JM034** | 32349737 | 98.2 |
| **JM035** | 20789661 | 97.96 |
| **JM036** | 22138837 | 98.1 |

**Appendix 2**

| LdoS Gene ID | LinJ Gene ID | Gene Product | *K*-medians Cluster |
|---|---|---|---|
| LDOS_000008100 | LinJ.01.0320 | hypothetical protein, conserved | 1 |
| LDOS_000014500 | LinJ.02.0070 | hypothetical protein, conserved | 1 |
| LDOS_000016800 | LinJ.02.0280 | Soluble NSF attachment protein, SNAP, putative | 1 |
| LDOS_000023100 | LinJ.03.0190 | delta-1-pyrroline-5-carboxylate dehydrogenase, putative | 1 |
| LDOS_000023400 | LinJ.03.0220 | long-chain-fatty-acid-CoA ligase, putative | 1 |
| LDOS_000024700 | LinJ.03.0350 | GPR1/FUN34/yaaH family, putative | 1 |
| LDOS_000034200 | LinJ.04.0300 | beta-fructofuranosidase, putative | 1 |
| LDOS_000034300 | LinJ.04.0310 | beta-fructofuranosidase, putative | 1 |
| LDOS_000036800 | LinJ.04.0570 | spermidine synthase | 1 |
| LDOS_000053500 | LinJ.05.0960 | metallo-peptidase, Clan M-, Family M49 | 1 |
| LDOS_000067500 | LinJ.06.0350 | NAD (P )-dependent steroid dehydrogenase protein, putative | 1 |
| LDOS_000071300 | LinJ.06.0730 | C2 domain/Ankyrin repeats (3 copies )/Ankyrin repeat, putative | 1 |
| LDOS_000077600 | LinJ.06.1340 | protoporphyrinogen oxidase-like protein | 1 |
| LDOS_000083200 | LinJ.07.0150 | acyl-CoA dehydrogenase, mitochondrial precursor, putative | 1 |
| LDOS_000088300 | LinJ.07.1020 | splicing factor ptsr1-like protein | 1 |
| LDOS_000095000 | LinJ.08.0320 | Isochorismatase family, putative | 1 |
| LDOS_000104300 | LinJ.08.1190 | hypothetical protein, conserved | 1 |
| LDOS_000105200 | LinJ.08.1190 | hypothetical protein | 1 |
| LDOS_000122000 | LinJ.09.1390 | paraflagellar rod component, putative | 1 |
| LDOS_000133700 | LinJ.10.0760 | amino acid permease 24, putative | 1 |
| LDOS_000165000 | LinJ.12.0665 | hypothetical protein, conserved | 1 |
| LDOS_000165200 | LinJ.12.0671 | hypothetical protein, conserved | 1 |
| LDOS_000169000 | LinJ.12.0850 | arginine N-methyltransferase-like protein | 1 |
| LDOS_000180800 | LinJ.13.1000 | leucyl-tRNA synthetase, putative | 1 |
| LDOS_000186900 | LinJ.13.1400 | chaperonin TCP20, putative | 1 |
| LDOS_000194400 | LinJ.14.0670 | fatty acid elongase, putative | 1 |
| LDOS_000194600 | LinJ.14.0700 | fatty acid elongase, putative | 1 |

| | | | |
|---|---|---|---|
| LDOS_000203500 | LinJ.14.1580 | glutathione-S-transferase/glutaredoxin, putative | 1 |
| LDOS_000219300 | LinJ.15.1500 | proliferative cell nuclear antigen (PCNA ), putative | 1 |
| LDOS_000236200 | LinJ.16.1510 | paraflagellar rod protein 1D, putative | 1 |
| LDOS_000236300 | LinJ.16.1510 | paraflagellar rod protein 1D, putative | 1 |
| LDOS_000236400 | LinJ.16.1510 | paraflagellar rod protein 1D, putative | 1 |
| LDOS_000236500 | LinJ.16.1510 | paraflagellar rod protein 1D, putative | 1 |
| LDOS_000239200 | LinJ.17.0010 | hypothetical protein, conserved | 1 |
| LDOS_000243200 | LinJ.17.0310 | hypothetical protein, conserved | 1 |
| LDOS_000258900 | LinJ.18.0220 | RNA-binding protein 29, putative | 1 |
| LDOS_000279100 | LinJ.19.0520 | hypothetical protein, conserved | 1 |
| LDOS_000284200 | LinJ.19.0920 | peptidyl-prolyl cis-trans isomerase, macrophage infectivity potentiator precursor, putative | 1 |
| LDOS_000307200 | N/A | N/A | 1 |
| LDOS_000311300 | LinJ.21.2140 | ATP synthase F1 subunit gamma protein, putative | 1 |
| LDOS_000314400 | LinJ.21.1830 | hypothetical protein, conserved | 1 |
| LDOS_000315400 | LinJ.21.1720 | hypothetical protein, conserved | 1 |
| LDOS_000329100 | LinJ.21.0490 | hypothetical protein, conserved | 1 |
| LDOS_000329600 | LinJ.21.0440 | Protein of unknown function (DUF667 ), putative | 1 |
| LDOS_000331000 | LinJ.21.0300 | hexokinase, putative | 1 |
| LDOS_000349400 | LinJ.22.1260 | centrin-4, putative | 1 |
| LDOS_000358100 | LinJ.23.0580 | acetyl-CoA synthetase, putative | 1 |
| LDOS_000368300 | LinJ.23.1460 | T-complex protein 1, gamma subunit, putative | 1 |
| LDOS_000371000 | LinJ.23.1730 | hypothetical protein, conserved | 1 |
| LDOS_000376300 | LinJ.24.0270 | Dynein intermediate chain 1, axonemal | 1 |
| LDOS_000381600 | LinJ.24.0790 | malic enzyme, putative | 1 |
| LDOS_000386200 | LinJ.24.1240 | translation factor sui1, putative | 1 |
| LDOS_000390000 | LinJ.24.1630 | hypothetical protein, conserved | 1 |
| LDOS_000411200 | LinJ.25.1210 | ATP synthase subunit beta, mitochondrial, putative | 1 |

| | | | |
|---|---|---|---|
| LDOS_000411300 | LinJ.25.1210 | ATP synthase subunit beta, mitochondrial, putative | 1 |
| LDOS_000421500 | LinJ.25.2230 | succinyl-CoA synthetase alpha subunit, putative | 1 |
| LDOS_000441000 | LinJ.26.1550 | hypothetical protein | 1 |
| LDOS_000448900 | LinJ.26.2320 | hypothetical protein, conserved | 1 |
| LDOS_000460000 | LinJ.27.2680 | amino acid transporter aATP11, putative | 1 |
| LDOS_000471500 | LinJ.27.1610 | eRF1 domain 1/eRF1 domain 2/eRF1 domain 3, putative | 1 |
| LDOS_000473400 | LinJ.27.1800 | hypothetical protein, conserved | 1 |
| LDOS_000490700 | LinJ.28.0900 | RNA binding protein rbp16, putative | 1 |
| LDOS_000509100 | LinJ.28.2670 | hypothetical protein, conserved | 1 |
| LDOS_000510500 | LinJ.28.2810 | IQ calmodulin-binding motif containing protein, putative | 1 |
| LDOS_000522100 | LinJ.29.0640 | ATP-binding cassette protein subfamily A, member 10, putative | 1 |
| LDOS_000528500 | LinJ.29.1260 | hypothetical protein, conserved | 1 |
| LDOS_000532800 | LinJ.29.1600 | Nodulin-like, putative | 1 |
| LDOS_000535500 | LinJ.29.1890 | paraflagellar rod protein 1D, putative | 1 |
| LDOS_000535600 | LinJ.29.1890 | paraflagellar rod protein 1D, putative | 1 |
| LDOS_000535700 | LinJ.29.1890 | paraflagellar rod protein 1D, putative | 1 |
| LDOS_000535800 | LinJ.29.1890 | paraflagellar rod protein 1D, putative | 1 |
| LDOS_000535900 | LinJ.29.1890 | paraflagellar rod protein 1D, putative | 1 |
| LDOS_000574700 | LinJ.30.2540 | heat shock 70-related protein 1, mitochondrial precursor, putative | 1 |
| LDOS_000576100 | LinJ.30.2610 | hypothetical protein, conserved | 1 |
| LDOS_000598700 | LinJ.31.0940 | hypothetical protein, conserved | 1 |
| LDOS_000629400 | LinJ.32.0500 | hypothetical protein, conserved | 1 |
| LDOS_000675500 | LinJ.33.0660 | paraflagellar rod component, putative | 1 |
| LDOS_000681500 | LinJ.33.1130 | hypothetical protein, conserved | 1 |
| LDOS_000688000 | LinJ.33.1730 | cyclophilin 4, putative | 1 |
| LDOS_000692800 | LinJ.33.2180 | hypothetical protein, conserved | 1 |

| | | | |
|---|---|---|---|
| LDOS_000697900 | LinJ.33.2680 | isocitrate dehydrogenase, putative | 1 |
| LDOS_000705400 | LinJ.33.3390 | h1 histone-like protein | 1 |
| LDOS_000709400 | LinJ.34.0370 | eukaryotic translation initiation factor 5, putative | 1 |
| LDOS_000716100 | LinJ.34.0890 | elongation factor 1-beta | 1 |
| LDOS_000716300 | LinJ.34.0890 | elongation factor 1-beta | 1 |
| LDOS_000725300 | LinJ.34.1630 | p25-alpha, putative | 1 |
| LDOS_000736800 | LinJ.34.2530 | RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain )/RNA recognition motif (a.k.a. RRM, RBD, or RNP domain ), putative | 1 |
| LDOS_000778500 | LinJ.35.2200 | adenine aminohydrolase | 1 |
| LDOS_000797300 | LinJ.35.4060 | protein kinase A catalytic subunit isoform 1 | 1 |
| LDOS_000813900 | LinJ.36.0270 | eukaryotic translation initiation factor 3 subunit L, putative | 1 |
| LDOS_000855200 | LinJ.36.4180 | hslVU complex proteolytic subunit, threonine peptidase, Clan T (1 ), family T1B | 1 |
| LDOS_000857900 | LinJ.36.4440 | paraflagellar rod component, putative | 1 |
| LDOS_000863400 | LinJ.36.4990 | hypothetical protein, conserved | 1 |
| LDOS_000863600 | LinJ.36.5010 | paraflagellar rod component, putative | 1 |
| LDOS_000866600 | LinJ.36.5310 | hypothetical protein, conserved | 1 |
| LDOS_000874600 | LinJ.36.6100 | kinetoplast-associated protein, putative | 1 |
| LDOS_000875300 | LinJ.36.6170 | Haloacid dehalogenase-like hydrolase, putative | 1 |
| LDOS_000879100 | LinJ.36.6550 | glucose transporter 2 | 1 |
| LDOS_000879200 | LinJ.36.6550 | glucose transporter 2 | 1 |
| LDOS_000879300 | LinJ.36.6550 | glucose transporter 2 | 1 |
| LDOS_000879400 | LinJ.36.6550 | glucose transporter 2 | 1 |
| LDOS_000884600 | LinJ.36.6950 | hypothetical protein, conserved | 1 |
| LDOS_000888500 | LinJ.36.7320 | eukaryotic translation initiation factor 3 subunit 8, putative | 1 |
| LDOS_000006700 | LinJ.01.0180 | CLC-type chloride channel, putative | 2 |
| LDOS_000015900 | LinJ.02.0190 | phosphoglycan beta 1, 2 arabinosyltransferase | 2 |
| LDOS_000029600 | LinJ.31.3360 | hypothetical protein, conserved | 2 |

| | | | |
|---|---|---|---|
| LDOS_000048800 | LinJ.05.0510 | ATP synthase F1, alpha subunit, putative | 2 |
| LDOS_000048900 | LinJ.05.0510 | ATP synthase F1, alpha subunit, putative | 2 |
| LDOS_000049000 | LinJ.05.0510 | ATP synthase F1, alpha subunit, putative | 2 |
| LDOS_000054700 | LinJ.05.1070 | Legume-like lectin family, putative | 2 |
| LDOS_000077500 | LinJ.06.1330 | coproporphyrinogen III oxidase | 2 |
| LDOS_000116500 | LinJ.09.0980 | calmodulin, putative | 2 |
| LDOS_000116700 | LinJ.09.0980 | calmodulin, putative | 2 |
| LDOS_000150300 | LinJ.11.0820 | hypothetical protein, conserved | 2 |
| LDOS_000158000 | LinJ.12.0012 | hypothetical protein, conserved | 2 |
| LDOS_000167300 | LinJ.12.0710 | DnaJ domain containing protein, putative | 2 |
| LDOS_000171200 | LinJ.13.0140 | hypothetical protein, conserved | 2 |
| LDOS_000179900 | LinJ.13.0910 | Cryptococcal mannosyltransferase 1, putative | 2 |
| LDOS_000181400 | LinJ.13.1060 | calmodulin, putative | 2 |
| LDOS_000197600 | LinJ.14.0990 | tc40 antigen-like | 2 |
| LDOS_000198200 | LinJ.14.1050 | ADP/ATP mitochondrial carrier-like protein | 2 |
| LDOS_000206100 | LinJ.15.0250 | hypothetical protein, conserved | 2 |
| LDOS_000216900 | LinJ.15.1260 | nucleoside transporter 1, putative | 2 |
| LDOS_000219600 | LinJ.15.1530 | ribosomal protein S6, putative | 2 |
| LDOS_000234900 | LinJ.16.1390 | cytochrome c, putative | 2 |
| LDOS_000235000 | LinJ.16.1390 | cytochrome c, putative | 2 |
| LDOS_000243100 | LinJ.17.0300 | cystathionine beta-synthase | 2 |
| LDOS_000250200 | LinJ.17.0970 | META domain containing protein | 2 |
| LDOS_000274400 | LinJ.19.0040 | histone H2B | 2 |
| LDOS_000287600 | LinJ.19.1270 | SPFH domain / Band 7 family, putative | 2 |
| LDOS_000289200 | LinJ.19.1400 | hypothetical protein, conserved | 2 |
| LDOS_000292800 | LinJ.20.0060 | anti-silencing protein asf 1-like protein | 2 |
| LDOS_000303900 | LinJ.20.1210 | cysteine peptidase, Clan CA, family C2, putative | 2 |
| LDOS_000330900 | LinJ.21.0310 | hexokinase, putative | 2 |
| LDOS_000335200 | LinJ.22.0010 | hypothetical protein, conserved | 2 |
| LDOS_000336900 | LinJ.22.0180 | hypothetical protein, conserved | 2 |

| | | | |
|---|---|---|---|
| LDOS_000337300 | LinJ.22.0220 | Tetrahydrofolate dehydrogenase/cyclohydrolase, catalytic domain/Tetrahydrofolate dehydrogenase/cyclohydrolase, NAD (P )-binding domain containing protein, putative | 2 |
| LDOS_000346700 | LinJ.22.0970 | Protein kinase domain containing protein, putative | 2 |
| LDOS_000362000 | LinJ.23.0860 | 3-ketoacyl-CoA thiolase, putative | 2 |
| LDOS_000362200 | LinJ.23.0880 | acetyl-CoA carboxylase, putative | 2 |
| LDOS_000365500 | LinJ.23.1160 | hypothetical protein, conserved | 2 |
| LDOS_000365800 | LinJ.23.1190 | hypothetical protein, conserved | 2 |
| LDOS_000368000 | LinJ.23.1430 | membrane-bound acid phosphatase 2 | 2 |
| LDOS_000369400 | LinJ.23.1560 | lathosterol oxidase-like protein | 2 |
| LDOS_000372000 | LinJ.23.1830 | hypothetical protein, conserved | 2 |
| LDOS_000390800 | LinJ.24.1710 | NAD dependent epimerase/dehydratase family/Male sterility protein, putative | 2 |
| LDOS_000394900 | LinJ.24.2060 | STOP axonemal protein, putative | 2 |
| LDOS_000402300 | LinJ.25.0360 | Protein of unknown function (DUF1308 ), putative | 2 |
| LDOS_000410600 | LinJ.25.1160 | aldehyde dehydrogenase, mitochondrial precursor | 2 |
| LDOS_000417000 | LinJ.25.1790 | pyruvate dehydrogenase E1 beta subunit, putative | 2 |
| LDOS_000420100 | LinJ.25.2090 | 2, 4-dihydroxyhept-2-ene-1,7-dioic acid aldolase, putative | 2 |
| LDOS_000462400 | LinJ.27.0740 | 2-oxoglutarate dehydrogenase subunit, putative | 2 |
| LDOS_000471600 | LinJ.27.1630 | hypothetical protein, conserved | 2 |
| LDOS_000472500 | LinJ.27.2500 | glycosomal phosphoenolpyruvate carboxykinase, putative | 2 |
| LDOS_000476100 | LinJ.27.2020 | RNA-binding protein, putative | 2 |
| LDOS_000483200 | LinJ.28.0140 | pantothenate kinase subunit, putative | 2 |
| LDOS_000504400 | LinJ.28.2200 | DNA-directed RNA polymerase-like protein | 2 |

| | | | |
|---|---|---|---|
| LDOS_000549000 | LinJ.30.0120 | alkyldihydroxyacetonephosphate synthase | 2 |
| LDOS_000592800 | LinJ.31.0450 | cytoskeleton-associated protein CAP5.5, putative | 2 |
| LDOS_000614600 | LinJ.31.2380 | 3 ' -nucleotidase/nuclease precursor, putative | 2 |
| LDOS_000621900 | LinJ.31.3080 | acetyl-CoA carboxylase, putative | 2 |
| LDOS_000623700 | LinJ.31.3260 | methylcrotonoyl-coa carboxylase biotinylated subunitprotein-like protein | 2 |
| LDOS_000624600 | LinJ.32.0020 | nuclear segregation protein, putative | 2 |
| LDOS_000636700 | LinJ.32.1220 | long chain polyunsaturated fatty acid elongation enzyme-like protein | 2 |
| LDOS_000660000 | LinJ.32.3510 | dihydrolipoamide dehydrogenase, putative | 2 |
| LDOS_000678100 | LinJ.33.0860 | beta tubulin | 2 |
| LDOS_000682300 | LinJ.33.1200 | hypothetical protein, conserved | 2 |
| LDOS_000689800 | LinJ.33.1910 | hypothetical protein, conserved | 2 |
| LDOS_000696800 | LinJ.33.2570 | hypothetical protein, conserved | 2 |
| LDOS_000709700 | LinJ.34.0400 | hypothetical protein, conserved | 2 |
| LDOS_000718100 | LinJ.34.1040 | amastin-like protein | 2 |
| LDOS_000724100 | LinJ.34.4360 | d-isomer specific 2-hydroxyacid dehydrogenase- protein | 2 |
| LDOS_000762900 | LinJ.35.0650 | beta-fructofuranosidase, putative | 2 |
| LDOS_000766400 | LinJ.35.1000 | aldose 1-epimerase, putative | 2 |
| LDOS_000771000 | LinJ.35.1450 | 60S ribosomal protein L2, putative | 2 |
| LDOS_000813000 | LinJ.36.0210 | elongation factor 2 | 2 |
| LDOS_000813100 | LinJ.36.0210 | elongation factor 2 | 2 |
| LDOS_000813200 | LinJ.36.0210 | elongation factor 2 | 2 |
| LDOS_000813300 | LinJ.36.0210 | elongation factor 2 | 2 |
| LDOS_000826000 | LinJ.36.1420 | Transitional endoplasmic reticulum ATPase, putative | 2 |
| LDOS_000837000 | LinJ.36.2490 | tyrosine aminotransferase | 2 |
| LDOS_000840700 | LinJ.36.2790 | dihydrolipoamide acetyltransferase precursor, putative | 2 |
| LDOS_000860200 | LinJ.36.4680 | hypothetical protein, conserved | 2 |
| LDOS_000881900 | LinJ.36.6740 | tartrate-sensitive acid phosphatase acp-3.2, putative | 2 |

| | | | |
|---|---|---|---|
| LDOS_000882200 | LinJ.36.6770 | histidine secretory acid phosphatase, putative | 2 |
| LDOS_000882500 | LinJ.36.6770 | histidine secretory acid phosphatase, putative | 2 |
| LDOS_000882800 | LinJ.36.6770 | histidine secretory acid phosphatase, putative | 2 |
| LDOS_000005400 | LinJ.01.0050 | carboxylase, putative | 3 |
| LDOS_000024100 | LinJ.03.0290 | hypothetical protein, conserved | 3 |
| LDOS_000074100 | LinJ.06.1010 | Leucine rich repeat/Leucine Rich Repeat, putative | 3 |
| LDOS_000077800 | LinJ.06.1360 | CLN3 protein, putative | 3 |
| LDOS_000080700 | LinJ.07.0430 | acetylornithine deacetylase-like protein | 3 |
| LDOS_000088200 | LinJ.07.1010 | hypothetical protein, conserved | 3 |
| LDOS_000089200 | LinJ.07.1100 | hypothetical protein, conserved | 3 |
| LDOS_000104500 | LinJ.08.1220 | Amidinotransferase, putative | 3 |
| LDOS_000111600 | LinJ.09.0501 | hypothetical protein, conserved | 3 |
| LDOS_000166900 | LinJ.12.0667 | hypothetical protein, conserved | 3 |
| LDOS_000171900 | LinJ.13.0210 | Triglyceride lipase, putative | 3 |
| LDOS_000190700 | LinJ.14.0330 | Transmembrane amino acid transporter protein, putative | 3 |
| LDOS_000224800 | LinJ.16.0410 | hypothetical protein, conserved | 3 |
| LDOS_000230400 | LinJ.16.0920 | EF-hand domain pair/EF-hand domain containing protein, putative | 3 |
| LDOS_000235400 | LinJ.16.1430 | Amidase, putative | 3 |
| LDOS_000238700 | LinJ.16.1730 | Putative snoRNA binding domain containing protein, putative | 3 |
| LDOS_000257100 | LinJ.18.0040 | major facilitator superfamily protein (MFS ), putative | 3 |
| LDOS_000271200 | N/A | N/A | 3 |
| LDOS_000277500 | LinJ.19.0360 | protein kinase, putative | 3 |
| LDOS_000278100 | LinJ.19.0420 | DNA-directed RNA polymerase III subunit C11, putative | 3 |
| LDOS_000284400 | LinJ.19.0940 | 4-coumarate:coa ligase-like protein | 3 |
| LDOS_000305600 | LinJ.20.1370 | kinase-like protein | 3 |
| LDOS_000316700 | N/A | N/A | 3 |
| LDOS_000323300 | LinJ.21.1040 | hypothetical protein, conserved | 3 |
| LDOS_000329200 | LinJ.21.0480 | NADPH:adrenodoxin oxidoreductase, mitochondrial, putative | 3 |

| | | | |
|---|---|---|---|
| LDOS_000367000 | LinJ.23.1340 | hypothetical protein, conserved | 3 |
| LDOS_000386800 | LinJ.24.1300 | amastin-like surface protein-like protein | 3 |
| LDOS_000399300 | LinJ.25.0070 | hypothetical protein, conserved | 3 |
| LDOS_000402700 | LinJ.25.0400 | hypothetical protein, conserved | 3 |
| LDOS_000411500 | LinJ.25.1230 | modification methylase-like protein | 3 |
| LDOS_000424000 | LinJ.25.2480 | adenylate kinase, putative | 3 |
| LDOS_000435400 | LinJ.26.0990 | hypothetical protein, conserved | 3 |
| LDOS_000444000 | LinJ.26.1850 | engulfment and cell motility domain 2, putative | 3 |
| LDOS_000465700 | LinJ.27.1080 | RING-H2 zinc finger/Anaphase-promoting complex subunit 11 RING-H2 finger/Ring finger domain containing protein, putative | 3 |
| LDOS_000472100 | LinJ.16.0880 | hypothetical protein | 3 |
| LDOS_000483300 | LinJ.28.0150 | Major Facilitator Superfamily, putative | 3 |
| LDOS_000492500 | LinJ.28.1080 | Serine incorporator (Serinc ), putative | 3 |
| LDOS_000499600 | N/A | N/A | 3 |
| LDOS_000506500 | LinJ.28.2420 | glycosomal membrane protein-like protein | 3 |
| LDOS_000508300 | LinJ.28.2590 | hypothetical protein, conserved | 3 |
| LDOS_000514300 | LinJ.28.3170 | hypothetical protein, conserved | 3 |
| LDOS_000514400 | LinJ.28.3180 | hypothetical protein, conserved | 3 |
| LDOS_000532900 | LinJ.29.1610 | Nodulin-like/Major Facilitator Superfamily, putative | 3 |
| LDOS_000563100 | LinJ.30.1500 | hypothetical protein, conserved | 3 |
| LDOS_000584200 | LinJ.30.3420 | hypothetical protein, conserved | 3 |
| LDOS_000593000 | LinJ.31.0460 | Amastin surface glycoprotein, putative | 3 |
| LDOS_000593200 | LinJ.31.0460 | Amastin surface glycoprotein, putative | 3 |
| LDOS_000593300 | LinJ.31.0460 | Amastin surface glycoprotein, putative | 3 |
| LDOS_000593500 | LinJ.34.1030 | Amastin surface glycoprotein, putative | 3 |
| LDOS_000597900 | LinJ.31.0870 | lipase precursor-like protein | 3 |
| LDOS_000598600 | LinJ.31.0930 | hypothetical protein, conserved | 3 |
| LDOS_000602700 | LinJ.31.1260 | hypothetical protein, conserved | 3 |
| LDOS_000612200 | LinJ.31.2140 | hypothetical protein, conserved | 3 |

| | | | |
|---|---|---|---|
| LDOS_000612400 | LinJ.31.2150 | hypothetical protein, conserved | 3 |
| LDOS_000616300 | LinJ.31.2560 | uridine kinase-like protein | 3 |
| LDOS_000618100 | LinJ.31.2730 | hypothetical protein, conserved | 3 |
| LDOS_000652100 | LinJ.32.2730 | hypothetical protein, conserved | 3 |
| LDOS_000662600 | LinJ.32.3760 | hypothetical protein, conserved | 3 |
| LDOS_000690900 | LinJ.33.2010 | vesicular protein trafficking mediator, putative | 3 |
| LDOS_000712800 | LinJ.34.0530 | phosphoglycan beta 1, 2 arabinosyltransferase, (SCA like ) | 3 |
| LDOS_000734800 | LinJ.34.2330 | hypothetical protein | 3 |
| LDOS_000744200 | LinJ.34.3110 | cytochrome p450-like protein | 3 |
| LDOS_000749100 | LinJ.34.3560 | actin-like protein, putative | 3 |
| LDOS_000753300 | LinJ.34.3980 | hypothetical protein, conserved | 3 |
| LDOS_000755200 | LinJ.34.4260 | NADH-ubiquinone oxidoreductase complex I subunit, putative | 3 |
| LDOS_000755500 | LinJ.34.4160 | phosphatidylinositol 3-kinase (tor2 ), putative | 3 |
| LDOS_000757300 | LinJ.35.0080 | Trk system potassium uptake protein | 3 |
| LDOS_000785200 | LinJ.35.2870 | Major Facilitator Superfamily/Sugar (and other ) transporter, putative | 3 |
| LDOS_000799400 | LinJ.35.4270 | hypothetical protein, conserved | 3 |
| LDOS_000807400 | LinJ.35.5350 | hypothetical protein, conserved | 3 |
| LDOS_000824700 | LinJ.36.1330 | hypothetical protein | 3 |
| LDOS_000826800 | LinJ.36.1500 | N-terminal region of Chorein, a TM vesicle-mediated sorter, putative | 3 |
| LDOS_000841600 | LinJ.36.2880 | hypothetical protein, conserved | 3 |
| LDOS_000861000 | LinJ.36.4760 | elongation factor-2 kinase-like protein | 3 |
| LDOS_000871600 | LinJ.36.5800 | aminopeptidase P1, putative | 3 |
| LDOS_000012100 | LinJ.01.0720 | PLAC8 family, putative | 4 |
| LDOS_000024800 | LinJ.03.0360 | GPR1/FUN34/yaaH family, putative | 4 |
| LDOS_000036700 | LinJ.04.0560 | DENN (AEX-3 ) domain containing protein, putative | 4 |
| LDOS_000048200 | LinJ.05.0440 | hypothetical protein, conserved | 4 |
| LDOS_000063900 | LinJ.06.0020 | hypothetical protein, conserved | 4 |
| LDOS_000086600 | LinJ.07.0850 | Protein kinase domain containing protein, putative | 4 |
| LDOS_000099800 | LinJ.08.0760 | amastin-like protein | 4 |

| | | | |
|---|---|---|---|
| LDOS_000100000 | LinJ.08.0760 | amastin-like protein | 4 |
| LDOS_000107200 | LinJ.09.0150 | kinesin, putative | 4 |
| LDOS_000127200 | LinJ.10.0250 | Amastin surface glycoprotein, putative | 4 |
| LDOS_000144300 | LinJ.11.0320 | DNA repair and recombination helicase protein PIF2, putative | 4 |
| LDOS_000144500 | LinJ.11.0340 | hypothetical protein, conserved | 4 |
| LDOS_000167500 | LinJ.30.1440 | hypothetical protein | 4 |
| LDOS_000201500 | LinJ.14.1370 | hypothetical protein, conserved | 4 |
| LDOS_000205000 | LinJ.15.0140 | zinc finger (CCCH type ) protein, putative | 4 |
| LDOS_000210000 | LinJ.15.0570 | hypothetical protein, conserved | 4 |
| LDOS_000221700 | LinJ.16.0100 | hypothetical protein, conserved | 4 |
| LDOS_000222100 | LinJ.16.0140 | hypothetical protein, conserved | 4 |
| LDOS_000225800 | LinJ.16.0510 | hypothetical protein, conserved | 4 |
| LDOS_000230200 | LinJ.16.0900 | hypothetical protein, conserved | 4 |
| LDOS_000243800 | LinJ.17.0370 | hypothetical protein, conserved | 4 |
| LDOS_000300300 | LinJ.20.0830 | hypothetical protein, conserved | 4 |
| LDOS_000303000 | LinJ.20.1100 | WD40 repeat-containing protein | 4 |
| LDOS_000323000 | LinJ.21.1070 | ABC transporter, putative | 4 |
| LDOS_000333800 | LinJ.31.3330 | phosphoglycan beta 1, 3 galactosyltransferase 4 | 4 |
| LDOS_000346200 | LinJ.22.0920 | Zinc finger, C3HC4 type (RING finger ) containing protein, putative | 4 |
| LDOS_000351600 | LinJ.22.1460 | hypothetical protein, conserved | 4 |
| LDOS_000353600 | LinJ.23.0130 | Concanavalin A-like lectin/glucanases superfamily/Beige/BEACH domain containing protein, putative | 4 |
| LDOS_000356800 | LinJ.23.0460 | hypothetical protein, conserved | 4 |
| LDOS_000379600 | N/A | N/A | 4 |
| LDOS_000383700 | LinJ.24.1000 | hypothetical protein, conserved | 4 |
| LDOS_000388000 | LinJ.24.1410 | histone deacetylase, putative | 4 |
| LDOS_000416800 | LinJ.25.1770 | Leucine-rich repeat/Leucine Rich repeats (2 copies ), putative | 4 |
| LDOS_000422500 | LinJ.25.2330 | glycosome import protein, putative | 4 |
| LDOS_000439200 | LinJ.26.1370 | hypothetical protein, conserved | 4 |
| LDOS_000468200 | N/A | N/A | 4 |
| LDOS_000473700 | LinJ.27.1830 | hypothetical protein, conserved | 4 |

| | | | |
|---|---|---|---|
| LDOS_000479600 | LinJ.27.2300 | vesicle-associated membrane protein | 4 |
| LDOS_000490600 | LinJ.28.0890 | hypothetical protein, conserved | 4 |
| LDOS_000502200 | N/A | N/A | 4 |
| LDOS_000508700 | LinJ.28.2630 | Putative serine esterase (DUF676 ), putative | 4 |
| LDOS_000509600 | LinJ.28.2720 | DNA repair and recombination helicase protein PIF3, putative | 4 |
| LDOS_000514000 | LinJ.28.3140 | glutamate dehydrogenase, putative | 4 |
| LDOS_000526000 | LinJ.29.1020 | A-1 protein, putative | 4 |
| LDOS_000561900 | LinJ.30.1410 | GTPase activating protein, putative | 4 |
| LDOS_000573800 | LinJ.30.2390 | hypothetical protein, conserved | 4 |
| LDOS_000577700 | LinJ.30.2770 | hypothetical protein, conserved | 4 |
| LDOS_000578000 | LinJ.30.2800 | Eukaryotic protein of unknown function (DUF872 ), putative | 4 |
| LDOS_000593600 | LinJ.31.0480 | calpain-like cysteine peptidase, putative | 4 |
| LDOS_000603200 | LinJ.31.1310 | pentamidine resistance protein 1 | 4 |
| LDOS_000603900 | LinJ.31.1380 | hypothetical protein, conserved | 4 |
| LDOS_000611700 | LinJ.31.2080 | hypothetical protein, conserved | 4 |
| LDOS_000613900 | LinJ.31.2310 | hypothetical protein, conserved | 4 |
| LDOS_000615800 | LinJ.31.2500 | hypothetical protein, conserved | 4 |
| LDOS_000624400 | LinJ.31.3330 | phosphoglycan beta 1, 3 galactosyltransferase 4 | 4 |
| LDOS_000638900 | LinJ.32.1440 | hypothetical protein, conserved | 4 |
| LDOS_000650700 | N/A | N/A | 4 |
| LDOS_000662000 | LinJ.32.3700 | hypothetical protein, conserved | 4 |
| LDOS_000673300 | LinJ.33.0440 | hypothetical protein, conserved | 4 |
| LDOS_000711400 | LinJ.34.4370 | Amastin surface glycoprotein, putative | 4 |
| LDOS_000712000 | LinJ.34.4370 | Amastin surface glycoprotein, putative | 4 |
| LDOS_000727900 | LinJ.34.1730 | amastin-like surface protein, putative | 4 |
| LDOS_000750900 | LinJ.34.3740 | expression site-associated protein 5 (ESAG5 ), putative | 4 |
| LDOS_000756500 | LinJ.31.3330 | phosphoglycan beta 1, 3 galactosyltransferase 4 | 4 |
| LDOS_000782900 | LinJ.35.2640 | hypothetical protein, conserved | 4 |
| LDOS_000811200 | LinJ.21.0010 | phosphoglycan beta 1, 3 galactosyltransferase 4 | 4 |

| | | | |
|---|---|---|---|
| LDOS_000815400 | N/A | N/A | 4 |
| LDOS_000832900 | LinJ.36.2100 | hypothetical protein, conserved | 4 |
| LDOS_000837600 | LinJ.36.2520 | sterol 24-c-methyltransferase, putative | 4 |
| LDOS_000838200 | LinJ.36.2580 | hypothetical protein, conserved | 4 |
| LDOS_000843200 | LinJ.36.3040 | ATP-binding cassette protein subfamily G, member 6, putative | 4 |
| LDOS_000848200 | LinJ.36.3500 | Protein of unknown function (DUF3595 ), putative | 4 |
| LDOS_000888700 | LinJ.36.7330 | U1A small nuclear ribonucleoprotein, putative | 4 |
| LDOS_000013600 | LinJ.25.2570 | phosphoglycan beta 1, 3 galactosyltransferase 4 | 5 |
| LDOS_000064600 | LinJ.06.0080 | ATP-binding cassette protein subfamily G, member 1, putative | 5 |
| LDOS_000088500 | LinJ.07.1060 | hypothetical protein, conserved | 5 |
| LDOS_000157800 | LinJ.12.0009 | hypothetical protein, conserved | 5 |
| LDOS_000180900 | LinJ.13.1010 | hypothetical protein, conserved | 5 |
| LDOS_000235700 | LinJ.16.1460 | OTU-like cysteine protease, putative | 5 |
| LDOS_000263900 | LinJ.18.0700 | citrate synthase, putative | 5 |
| LDOS_000336300 | LinJ.22.0120 | phosphoinositide phosphatase | 5 |
| LDOS_000342600 | LinJ.22.0570 | hypothetical protein, conserved | 5 |
| LDOS_000353900 | LinJ.23.0160 | Alpha/beta hydrolase family/alpha/beta hydrolase fold/Serine hydrolase, putative | 5 |
| LDOS_000354800 | LinJ.23.0300 | Arginosuccinate synthase, putative | 5 |
| LDOS_000355200 | LinJ.23.0300 | argininosuccinate synthase, putative | 5 |
| LDOS_000430400 | LinJ.26.0500 | hypothetical protein | 5 |
| LDOS_000493900 | LinJ.28.1220 | hypothetical protein, conserved | 5 |
| LDOS_000588300 | LinJ.31.0030 | Aquaglyceroporin 1 | 5 |
| LDOS_000592900 | LinJ.31.0460 | Amastin surface glycoprotein, putative | 5 |
| LDOS_000597500 | LinJ.31.0830 | hypothetical protein, conserved | 5 |
| LDOS_000603600 | LinJ.31.1350 | hypothetical protein, conserved | 5 |
| LDOS_000603700 | LinJ.31.1360 | Cold-shock DNA-binding domain containing protein, putative | 5 |
| LDOS_000606700 | LinJ.31.1630 | Leucine Rich repeat, putative | 5 |
| LDOS_000608300 | LinJ.31.1780 | hypothetical protein, conserved | 5 |

| | | | |
|---|---|---|---|
| LDOS_000634600 | LinJ.32.1010 | ubiquitin-conjugating enzyme protein, putative | 5 |
| LDOS_000640700 | LinJ.32.1630 | PSP1 C-terminal conserved region containing protein, putative | 5 |
| LDOS_000647400 | LinJ.32.2320 | hypothetical protein, conserved | 5 |
| LDOS_000660900 | LinJ.32.3600 | Alpha/beta hydrolase family, putative | 5 |
| LDOS_000700400 | LinJ.33.2920 | Concanavalin A-like lectin/glucanases superfamily/Beige/BEACH domain containing protein, putative | 5 |
| LDOS_000700900 | LinJ.33.2970 | cysteine peptidase, Clan CA, family C51, putative | 5 |
| LDOS_000711300 | LinJ.34.1670 | Amastin surface glycoprotein, putative | 5 |
| LDOS_000711500 | LinJ.34.4370 | Amastin surface glycoprotein, putative | 5 |
| LDOS_000712100 | LinJ.34.4370 | Amastin surface glycoprotein, putative | 5 |
| LDOS_000712200 | LinJ.34.4370 | Amastin surface glycoprotein, putative | 5 |
| LDOS_000726000 | LinJ.34.1700 | Amastin surface glycoprotein, putative | 5 |
| LDOS_000726100 | LinJ.34.1730 | Amastin surface glycoprotein, putative | 5 |
| LDOS_000726200 | LinJ.34.1730 | amastin-like surface protein, putative | 5 |
| LDOS_000726300 | LinJ.34.1730 | amastin-like surface protein, putative | 5 |
| LDOS_000726400 | LinJ.34.1730 | amastin-like surface protein, putative | 5 |
| LDOS_000726500 | LinJ.34.1730 | Amastin surface glycoprotein, putative | 5 |
| LDOS_000726600 | LinJ.34.1730 | amastin-like surface protein, putative | 5 |
| LDOS_000726700 | LinJ.34.1710 | amastin-like surface protein, putative | 5 |
| LDOS_000726800 | LinJ.34.1710 | amastin-like surface protein, putative | 5 |
| LDOS_000726900 | LinJ.34.1700 | amastin-like surface protein, putative | 5 |

| | | | |
|---|---|---|---|
| LDOS_000727000 | LinJ.34.1700 | amastin-like surface protein, putative | 5 |
| LDOS_000727200 | LinJ.34.1690 | Amastin surface glycoprotein, putative | 5 |
| LDOS_000727300 | LinJ.34.1710 | Amastin surface glycoprotein, putative | 5 |
| LDOS_000727400 | LinJ.34.1680 | Amastin surface glycoprotein, putative | 5 |
| LDOS_000727500 | LinJ.34.1700 | amastin-like surface protein, putative | 5 |
| LDOS_000727700 | LinJ.34.1700 | amastin-like surface protein, putative | 5 |
| LDOS_000727800 | LinJ.34.1700 | amastin-like surface protein, putative | 5 |
| LDOS_000728000 | LinJ.34.1700 | amastin-like surface protein, putative | 5 |
| LDOS_000728100 | LinJ.34.1700 | Amastin surface glycoprotein, putative | 5 |
| LDOS_000728200 | LinJ.34.1700 | Amastin surface glycoprotein, putative | 5 |
| LDOS_000728300 | LinJ.34.1710 | Amastin surface glycoprotein, putative | 5 |
| LDOS_000728400 | LinJ.34.1700 | amastin-like surface protein, putative | 5 |
| LDOS_000728500 | LinJ.34.1710 | Amastin surface glycoprotein, putative | 5 |
| LDOS_000728600 | LinJ.34.1710 | amastin-like surface protein, putative | 5 |
| LDOS_000738100 | LinJ.34.4340 | Amastin surface glycoprotein, putative | 5 |
| LDOS_000738200 | LinJ.34.2650 | Amastin surface glycoprotein, putative | 5 |
| LDOS_000738300 | LinJ.34.2650 | Amastin surface glycoprotein, putative | 5 |
| LDOS_000738700 | LinJ.34.4350 | Amastin surface glycoprotein, putative | 5 |
| LDOS_000838500 | LinJ.36.2610 | hypothetical protein, conserved | 5 |
| LDOS_000015300 | LinJ.02.0140 | phosphoglycan beta 1, 3 galactosyltransferase | 6 |
| LDOS_000098500 | LinJ.08.0650 | Amastin surface glycoprotein, putative | 6 |
| LDOS_000098800 | LinJ.08.1320 | amastin-like protein | 6 |
| LDOS_000098900 | LinJ.08.1320 | amastin-like protein | 6 |

| | | | |
|---|---|---|---|
| LDOS_000099000 | LinJ.08.1320 | amastin-like protein | 6 |
| LDOS_000099100 | LinJ.08.1330 | Amastin surface glycoprotein, putative | 6 |
| LDOS_000099200 | LinJ.08.1330 | Amastin surface glycoprotein, putative | 6 |
| LDOS_000099300 | LinJ.08.1330 | Amastin surface glycoprotein, putative | 6 |
| LDOS_000099400 | LinJ.08.0690 | amastin-like protein | 6 |
| LDOS_000162300 | LinJ.12.0460 | hypothetical protein, conserved | 6 |
| LDOS_000183600 | LinJ.13.1260 | hypothetical protein, conserved | 6 |
| LDOS_000216800 | LinJ.15.1230 | nucleoside transporter 1, putative | 6 |
| LDOS_000239400 | LinJ.17.0020 | Protein of unknown function (DUF3522 ), putative | 6 |
| LDOS_000251600 | LinJ.17.1110 | hydrolase, alpha/beta fold family-like protein | 6 |
| LDOS_000341500 | LinJ.22.0670 | A2 protein | 6 |
| LDOS_000341700 | LinJ.22.0670 | A2 protein | 6 |
| LDOS_000341900 | LinJ.22.0670 | A2 protein | 6 |
| LDOS_000343600 | LinJ.22.0660 | 5 ' a2rel-related protein | 6 |
| LDOS_000378400 | LinJ.24.0490 | Kelch motif/Galactose oxidase, central domain containing protein, putative | 6 |
| LDOS_000472400 | LinJ.27.2500 | glycosomal phosphoenolpyruvate carboxykinase, putative | 6 |
| LDOS_000492400 | LinJ.28.1070 | P27 protein, putative | 6 |
| LDOS_000528000 | LinJ.29.1210 | PSP1 C-terminal conserved region containing protein, putative | 6 |
| LDOS_000530400 | LinJ.29.1450 | hypothetical protein | 6 |
| LDOS_000530800 | LinJ.29.3010 | Amastin surface glycoprotein, putative | 6 |
| LDOS_000531000 | LinJ.29.3000 | Amastin surface glycoprotein, putative | 6 |
| LDOS_000537300 | LinJ.29.2010 | DnaJ domain containing protein, putative | 6 |
| LDOS_000559200 | LinJ.30.1140 | IQ calmodulin-binding motif containing protein, putative | 6 |
| LDOS_000571900 | LinJ.30.2200 | RNA-binding protein, putative | 6 |
| LDOS_000573300 | LinJ.30.2340 | hypothetical protein, conserved | 6 |
| LDOS_000574300 | LinJ.30.2440 | hypothetical protein, conserved | 6 |

| | | | |
|---|---|---|---|
| LDOS_000593100 | LinJ.31.0460 | Amastin surface glycoprotein, putative | 6 |
| LDOS_000593400 | LinJ.31.0460 | Amastin surface glycoprotein, putative | 6 |
| LDOS_000593700 | LinJ.31.0490 | hypothetical protein, conserved | 6 |
| LDOS_000616200 | LinJ.31.2540 | lipase, putative | 6 |
| LDOS_000706400 | LinJ.34.0070 | ascorbate peroxidase | 6 |
| LDOS_000711100 | LinJ.34.1670 | Amastin surface glycoprotein, putative | 6 |
| LDOS_000711200 | LinJ.34.1670 | Amastin surface glycoprotein, putative | 6 |
| LDOS_000711600 | LinJ.34.4370 | Amastin surface glycoprotein, putative | 6 |
| LDOS_000711700 | LinJ.34.4370 | Amastin surface glycoprotein, putative | 6 |
| LDOS_000711800 | LinJ.34.4370 | Amastin surface glycoprotein, putative | 6 |
| LDOS_000711900 | LinJ.34.4370 | Amastin surface glycoprotein, putative | 6 |
| LDOS_000712300 | LinJ.34.4370 | Amastin surface glycoprotein, putative | 6 |
| LDOS_000712700 | LinJ.34.4370 | Amastin surface glycoprotein, putative | 6 |
| LDOS_000725800 | LinJ.34.1680 | Amastin surface glycoprotein, putative | 6 |
| LDOS_000738400 | LinJ.34.2650 | Amastin surface glycoprotein, putative | 6 |
| LDOS_000738500 | LinJ.34.4350 | Amastin surface glycoprotein, putative | 6 |
| LDOS_000738600 | LinJ.34.4350 | Amastin surface glycoprotein, putative | 6 |
| LDOS_000747800 | LinJ.34.3430 | hypothetical protein, conserved | 6 |
| LDOS_000751200 | LinJ.34.3770 | hypothetical protein | 6 |
| LDOS_000817500 | LinJ.36.0630 | WD domain, G-beta repeat, putative | 6 |
| LDOS_000821200 | N/A | N/A | 6 |
| LDOS_000033800 | LinJ.04.0260 | Mitochondrial 39-S ribosomal protein L47 (MRP-L47 ), putative | 7 |
| LDOS_000047500 | LinJ.05.0370 | hypothetical protein, conserved | 7 |
| LDOS_000077000 | LinJ.06.1290 | FAD binding domain containing protein, putative | 7 |
| LDOS_000077400 | LinJ.06.1320 | pteridine transporter, putative | 7 |

| | | | |
|---|---|---|---|
| LDOS_000077700 | LinJ.06.1350 | Cytochrome b5-like Heme/Steroid binding domain containing protein, putative | 7 |
| LDOS_000109400 | LinJ.09.0420 | integral membrane transport protein, putative | 7 |
| LDOS_000121200 | LinJ.09.1310 | hypothetical protein, conserved | 7 |
| LDOS_000223000 | LinJ.16.0230 | hypothetical protein, conserved | 7 |
| LDOS_000259300 | LinJ.18.0260 | hypothetical protein, conserved | 7 |
| LDOS_000265000 | LinJ.18.0810 | ethanolamine phosphotransferase, putative | 7 |
| LDOS_000276000 | LinJ.19.0200 | ADP, ATP carrier protein 1, mitochondrial precursor, putative | 7 |
| LDOS_000315500 | LinJ.21.1710 | hypothetical protein, conserved | 7 |
| LDOS_000333700 | LinJ.36.0020 | histone H4 | 7 |
| LDOS_000334000 | LinJ.22.0002 | Cytochrome b5-like Heme/Steroid binding domain containing protein, putative | 7 |
| LDOS_000354100 | LinJ.23.0180 | hypothetical protein, conserved | 7 |
| LDOS_000389400 | LinJ.24.1570 | translationally controlled tumor protein (TCTP), putative | 7 |
| LDOS_000389500 | LinJ.24.1570 | translationally controlled tumor protein (TCTP), putative | 7 |
| LDOS_000397400 | LinJ.24.2320 | hypothetical protein, conserved | 7 |
| LDOS_000408500 | LinJ.25.0960 | Ankyrin repeats (3 copies )/Ankyrin repeats (many copies )/Ankyrin repeat, putative | 7 |
| LDOS_000432900 | LinJ.26.0730 | RNA recognition motif (a.k.a. RRM, RBD, or RNP domain)/RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain ), putative | 7 |
| LDOS_000469500 | LinJ.27.1440 | GNAT acetyltransferase, putative | 7 |
| LDOS_000497800 | LinJ.28.1610 | Protein of unknown function (DUF423 ), putative | 7 |
| LDOS_000507000 | LinJ.28.2470 | hypothetical protein, conserved | 7 |
| LDOS_000525000 | LinJ.29.0930 | Cytochrome b5-like Heme/Steroid binding domain containing protein, putative | 7 |
| LDOS_000525700 | LinJ.29.0990 | signal peptide peptidase, putative | 7 |
| LDOS_000556800 | LinJ.30.0910 | 50S ribosome-binding GTPase, putative | 7 |

| | | | |
|---|---|---|---|
| LDOS_000566400 | LinJ.30.1680 | CRAL/TRIO, N-terminal domain/CRAL/TRIO domain/Divergent CRAL/TRIO domain containing protein, putative | 7 |
| LDOS_000578700 | LinJ.30.2870 | hypothetical protein, conserved | 7 |
| LDOS_000602000 | LinJ.31.1210 | Cytochrome b5-like Heme/Steroid binding domain containing protein, putative | 7 |
| LDOS_000623000 | LinJ.31.3190 | iron/zinc transporter protein-like protein | 7 |
| LDOS_000649900 | N/A | N/A | 7 |
| LDOS_000715100 | LinJ.34.0770 | Stage II sporulation protein E (SpoIIE ), putative | 7 |
| LDOS_000757400 | LinJ.35.0090 | hypothetical protein, conserved | 7 |
| LDOS_000764100 | LinJ.35.0790 | rRNA dimethyltransferase, putative | 7 |
| LDOS_000810600 | LinJ.35.5310 | Dynein heavy chain and region D6 of dynein motor/Ankyrin repeats (3 copies )/Ankyrin repeat, putative | 7 |
| LDOS_000811300 | LinJ.36.0020 | histone H4 | 7 |
| LDOS_000822700 | LinJ.36.1160 | ribosomal protein L24, putative | 7 |
| LDOS_000823000 | LinJ.36.1160 | ribosomal protein L24, putative | 7 |
| LDOS_000827500 | LinJ.36.1570 | hypothetical protein, conserved | 7 |
| LDOS_000833300 | LinJ.36.2120 | hypothetical protein, conserved | 7 |
| LDOS_000852600 | LinJ.36.3930 | 60S ribosomal protein L34, putative | 7 |
| LDOS_000854000 | LinJ.36.4070 | eukaryotic translation initiation factor 3 subunit 2, putative | 7 |
| LDOS_000142800 | LinJ.11.0180 | cytoplasmic translation machinery associated protein, putative | 8 |
| LDOS_000147700 | LinJ.11.0490 | cytochrome b5, putative | 8 |
| LDOS_000201800 | LinJ.14.1400 | serine hydroxymethyltransferase (SHMT-L ) | 8 |
| LDOS_000305100 | LinJ.20.1320 | small myristoylated protein 4, putative | 8 |
| LDOS_000362700 | LinJ.23.0930 | mitochondrial RNA binding protein, putative | 8 |
| LDOS_000458400 | LinJ.27.2520 | calpain-like cysteine peptidase, putative | 8 |
| LDOS_000459900 | LinJ.27.2670 | hypothetical protein, conserved | 8 |

| | | | |
|---|---|---|---|
| LDOS_000466300 | LinJ.27.1150 | T-complex protein 1, beta subunit, putative | 8 |
| LDOS_000493200 | LinJ.28.1150 | hypothetical protein, conserved | 8 |
| LDOS_000504600 | LinJ.28.2220 | DEAD-box ATP-dependent RNA helicase, mitochondrial | 8 |
| LDOS_000505600 | LinJ.28.2320 | Mitochondrial import receptor subunit ATOM69, putative | 8 |
| LDOS_000543200 | LinJ.29.2550 | 3 ' 5 ' -cyclic nucleotide phosphodiesterase, putative | 8 |
| LDOS_000567700 | LinJ.30.1810 | Flagellar-associated PapD-like/Zeta toxin, putative | 8 |
| LDOS_000622900 | LinJ.31.3190 | iron/zinc transporter protein-like protein | 8 |
| LDOS_000659600 | LinJ.32.3470 | chaperonin alpha subunit, putative | 8 |
| LDOS_000664900 | LinJ.32.3910 | enolase, putative | 8 |
| LDOS_000690400 | LinJ.33.1960 | glycosomal transporter (GAT2 ), putative | 8 |
| LDOS_000716400 | LinJ.34.0900 | serine/threonine-protein phosphatase PP1, putative | 8 |
| LDOS_000754700 | LinJ.34.4120 | nucleolar protein family a, putative | 8 |
| LDOS_000764700 | LinJ.35.0840 | aspartate aminotransferase, putative | 8 |
| LDOS_000798700 | LinJ.35.4200 | poly (A )-binding protein 2 | 8 |
| LDOS_000833400 | LinJ.36.2130 | chaperonin HSP60, mitochondrial precursor | 8 |
| LDOS_000845100 | LinJ.36.3220 | fibrillarin | 8 |
| LDOS_000087200 | LinJ.07.0910 | flavoprotein subunit-like protein | 9 |
| LDOS_000130800 | LinJ.10.0520 | GP63, leishmanolysin | 9 |
| LDOS_000131800 | LinJ.10.0590 | hypothetical protein, conserved | 9 |
| LDOS_000153100 | N/A | N/A | 9 |
| LDOS_000189100 | LinJ.14.0180 | carboxypeptidase, putative | 9 |
| LDOS_000227200 | LinJ.16.0620 | hypothetical protein, conserved | 9 |
| LDOS_000273800 | N/A | N/A | 9 |
| LDOS_000518700 | LinJ.29.0290 | D-lactate dehydrogenase-like protein | 9 |
| LDOS_000567500 | LinJ.30.1790 | DNAJ domain protein, putative | 9 |
| LDOS_000597100 | LinJ.31.0790 | hypothetical protein, conserved | 9 |
| LDOS_000602100 | LinJ.31.1220 | Cytochrome b5-like Heme/Steroid binding domain containing protein, putative | 9 |
| LDOS_000671000 | LinJ.33.0370 | heat shock protein 83 | 9 |

| | | | |
|---|---|---|---|
| LDOS_000671100 | LinJ.33.0370 | heat shock protein 83 | 9 |
| LDOS_000671200 | LinJ.33.0370 | heat shock protein 83 | 9 |
| LDOS_000672200 | LinJ.33.0370 | hypothetical protein | 9 |
| LDOS_000672300 | LinJ.33.0370 | heat shock protein 83 | 9 |
| LDOS_000672400 | LinJ.33.0370 | heat shock protein 83 | 9 |
| LDOS_000672500 | LinJ.33.0370 | heat shock protein 83 | 9 |
| LDOS_000766300 | N/A | N/A | 9 |
| LDOS_000800300 | LinJ.35.4360 | hypothetical protein, conserved | 9 |
| LDOS_000807800 | LinJ.35.5390 | hypothetical protein, conserved | 9 |
| LDOS_000841800 | LinJ.36.2900 | Nodulin-like/Major Facilitator Superfamily, putative | 9 |
| LDOS_000853800 | LinJ.36.4050 | similar to leishmania major. l411.4-like protein | 9 |