UNIVERSIDADE FEDERAL DE MINAS GERAIS

INSTITUTO DE CIÊNCIAS BIOLÓGICAS

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

PhD Thesis

Comparative genomics, phylogenomics and population genomics of New World *Leishmania*

By:

Hugo Oswaldo Valdivia Rodríguez

Under the supervision of:

Dr. Daniella Castanheira Bartholomeu

Belo Horizonte
2016

Hugo Oswaldo Valdivia Rodríguez

Comparative genomics, phylogenomics and population genomics of New World *Leishmania*

Thesis presented in fulfillment of the requirements for the degree of Doctor of Philosophy in Bioinformatics of the Universidade Federal de Minas Gerais.

Advisor: Dr. Daniella Castanheira Bartholomeu

Belo Horizonte
2016

ACKNOWLEDGEMENTS

I would like to thank the people that contributed to this thesis with their expertise, support and commitment:

I would like to express my gratitude to Dr. Daniella Castanheira Bartholomeu, for her advice and mentoring during the execution of this work and her unconditional and disinterested support in my career and professional development, but overall for her friendship.

To Dr. Andres G. Lescano, my friend and mentor, for his constant guidance, support and for sharing with me his academic and non-academic experience.

To my friends at the Laboratory of Parasite Genomics and Immunology (LIGP). In special, Michelle, Sebastiao, Fernando, Natalia, Mariana, Ana Leao and Gaby for all the good times, chit-chat and procrastination. But also for their advice, pertinent discussion, motivation and a special recognition for Mariana for all her patience guiding me during wet-lab experiments.

This thesis goas for all my family starting with my parents (Carlos and Eveling), grandparents (Esther, Carmen, Hugo) aunts and uncles (Miriam, Grisel, Carlos), brother and sisters (Angela, Carlos, Daniela) for all the patience and sacrifice during all these years and specially for their motivation, encouraging me to fight for my dreams regardless of how far and impossible they might seem to be.

To my friends Fabricio, Carlos, Ronald and David for all their advice, support and patience. Helping me to deal with the distance and the “saudades”.

COLLABORATORS

Finally, I would like to thank all the people who directly contribute in the different projects part of this thesis, hoping to continue and nurture our collaborations in the fight against pathogenic diseases.

Dr. Stephen Beverley
Dr. Christian Baldeviano
Dr. Toni Gabaldon
Dr. Robert Gerbassi
Dr. Guilherme Oliveira
Dr. Gabriela F. Luiz
Dr. Rodrigo Baptista
Dr. Larissa Silva
Dr. Bruno M. Roatt
Dr. Ricardo Toshio Fujiwara
Dr. Celia Gontijo
Dr. Alexandre B. Reis
Dr. James Cotton
Dr. Mathew Berriman
MSc. Joao L. Reis
Ms. Laila Almeida

INDEX

List of Figures

List of Tables

Abbreviations

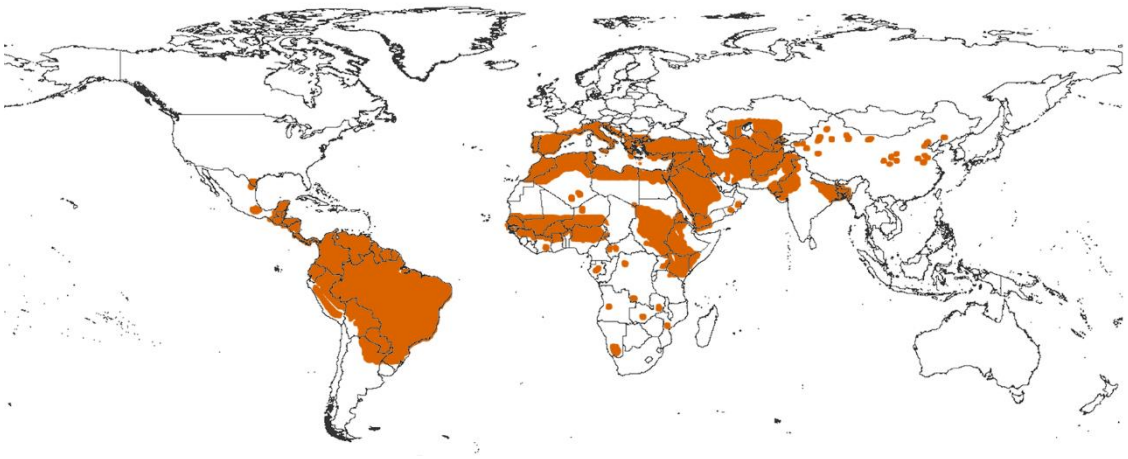
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
CL	Cutaneous leishmaniasis
CCN	Chromosome copy number
CN	Copy number
CNV	Copy number variations
DCL	Diffuse cutaneous leishmaniasis
DT	Decision Theory
HC	Hierarchical clustering
HCN	Haploid copy number
INMET	Instituto Nacional de Meteorologia
MAF	Minimum allele frequency
MCMC	Markov Chain Monte Carlo
ML	Mucosal leishmaniasis
PCA	Principal component analysis
SNP	Single Nucleotide Polymorphisms
TL	Tegumentary leishmaniasis
VL	Visceral leishmaniasis

PREFACE

Epidemiological context of the leishmaniasis

Leishmaniasis is a complex parasitic disease with diverse clinical manifestations and epidemiology that is caused by protozoan belonging to the genus *Leishmania*. These parasites are transmitted to the mammalian host by the bite of infected phlebotomine *Lutzomyia* sand flies in the New World and *Phlebotomus* in the Old World.

The leishmaniasis is spread in more than 98 countries worldwide putting 350 million people at risk of infection and causing more than 1.5 new million cases per year (Murray *et al.*, 2005; Alvar *et al.*, 2012) **(Preface: Figure 1)**. Currently, this infection is considered as an emergent and re-emergent disease and there is increased concern about its progressive adaptation into urban environments, the effects of human migration, climate change and co-infection with other diseases (Desjeux, 2004).



Preface: Figure 1

Geographic distribution of the leishmaniasis. Source: World Health Organization, October 2010

An important characteristic of the leishmaniasis is its wide spectrum of clinical manifestations that have been classified into tegumentary leishmaniasis (TL), mucosal leishmaniasis (ML) and visceral leishmaniasis (VL) (Murray *et al.*, 2005; David e Craft, 2009) **(Preface: Figure 2)**. These distinct diseases can be caused by up to 20 different *Leishmania* species.

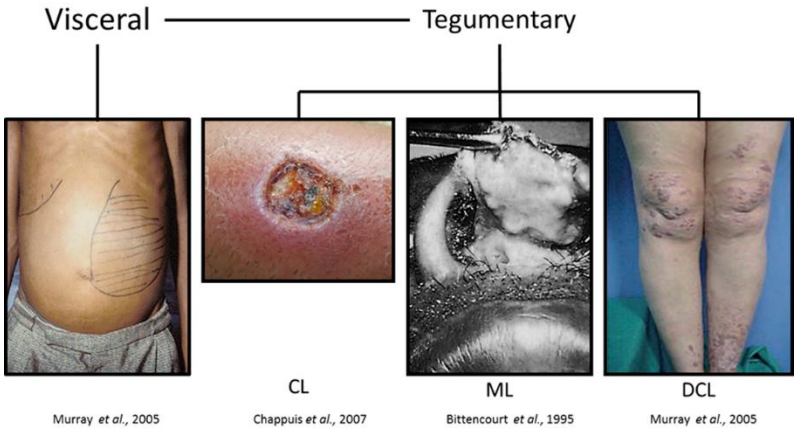
Epidemiological data show that six countries hold around 90% of all VL in the world (India, Bangladesh, Sudan, South Sudan, Brazil and Ethiopia), whereas 75% of all TL cases are reported in Afghanistan,

Algeria, Colombia, Brazil, Iran, Syria, Ethiopia, North Sudan, Costa Rica and Peru (Desjeux, 2004; Alvar *et al.*, 2012) constituting an important public health problem in these areas.

TL can be sub classified into three distinct subtypes that are referred as cutaneous leishmaniasis (CL), mucosal leishmaniasis (ML) and diffuse cutaneous leishmaniasis (DCL). TL is widely distributed in the Old and New World and is characterized by a progressive destruction of tissues with the appearance of papules and ulcerative lesions with keratotic plates near the sand fly bite (Murray *et al.*, 2005). ML usually appears after a first cutaneous lesion and affects the nares and pharyngeal cavity causing important respiratory complications and disfigurement (Goto e Lindoso, 2010).

Diffuse disease has been associated with *L. (L.) amazonensis* and *L. (L.) aethiopica* infection in the New and Old World and results in disseminated non-ulcerative lesions on the patient's body (Akuffo *et al.*, 1997; Sinha *et al.*, 2008).

Visceral disease is the most important form of leishmaniasis and can be deadly if remains untreated. The range of symptoms of VL can vary from subclinical to fully established infection leading to fever, general weakness, weight loss and more complex symptoms like multi-organ failure (hepatomegaly or splenomegaly (Murray *et al.*, 2005; Banuls *et al.*, 2007). The two most important etiological agents of VL are *L. (L.) donovani* and *L. (L.) infantum* that are present in the Americas, Europe, Asia and Africa (Murray *et al.*, 2005; Alvar *et al.*, 2012).



Preface: Figure 2

Clinical manifestations of the leishmaniasis. CL: cutaneous leishmaniasis; ML: mucosal leishmaniasis; DCL: diffuse cutaneous leishmaniasis

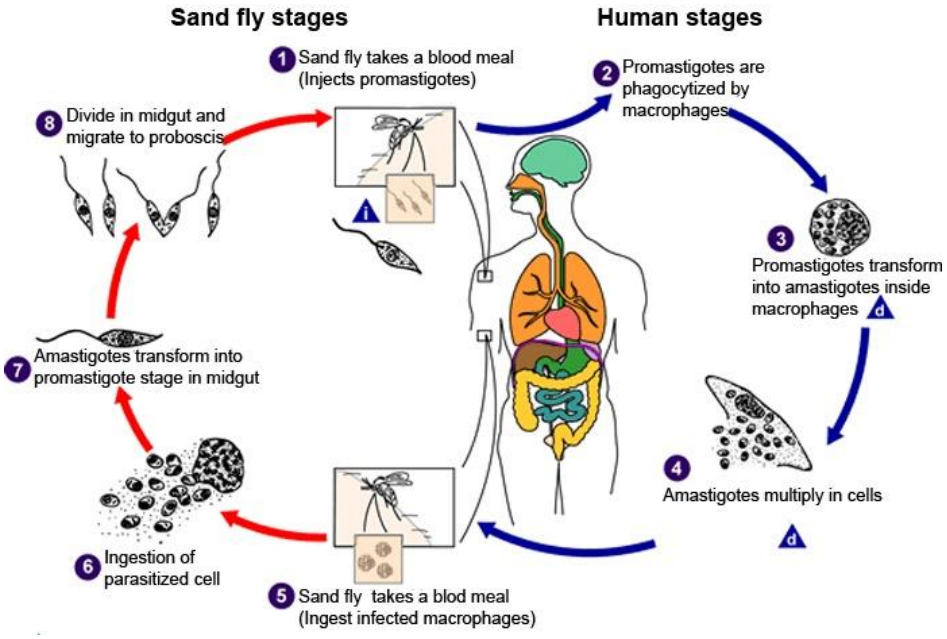
The life cycle of Leishmania

Leishmania is a digenetic protozoan that develops part of its lifecycle on an invertebrate female phlebotomine and the other on a vertebrate host **(Preface: Figure 3)**. This characteristic of the lifecycle exposes the parasite to two distinct environments within their hosts and results in two different parasite forms. On the sand fly, the parasite develops into a flagellated motile extracellular form called promastigote whereas in the vertebrate host it presents a non-motile intracellular stage named amastigote (Banuls *et al.*, 2007).

The lifecycle begins when a non-infected female sand fly ingests *Leishmania* amastigotes from an infected reservoir. While more than 500 different sand fly species have been reported in South America, only 30 have been incriminated as putative vectors (Killick-Kendrick, 1990; Kato *et al.*, 2010). Inside the sand fly tract, amastigotes rapidly differentiate into promastigotes (up to 24 hours after blood is ingested) and start to multiply by binary fission.

Many authors consider up to five different promastigote forms inside the invertebrate host: procyclic promastigotes, nectomonads, haptomonads, paramastigotes and metacyclic promastigotes (Kato *et al.*, 2010). The latter is considered the infective form for the vertebrate host.

The lifecycle continues when an infected female feeds on a non-infected vertebrate host and regurgitates metacyclic promastigotes. These promastigotes are rapidly phagocytized by the host macrophages and inside parasitophorous vacuoles they start to develop into amastigotes that are adapted to the acidic pH conditions of the macrophage (Banuls *et al.*, 2007). Amastigotes start to multiplicate by binary fission leading to a large population of parasites that eventually burst the cell and infect other macrophages. Eventually a non-infected female sand fly bites an infected host and the cycle starts again.



Preface: Figure 3

The life cycle of *Leishmania*. Source: Centers of Disease Control and Prevention.

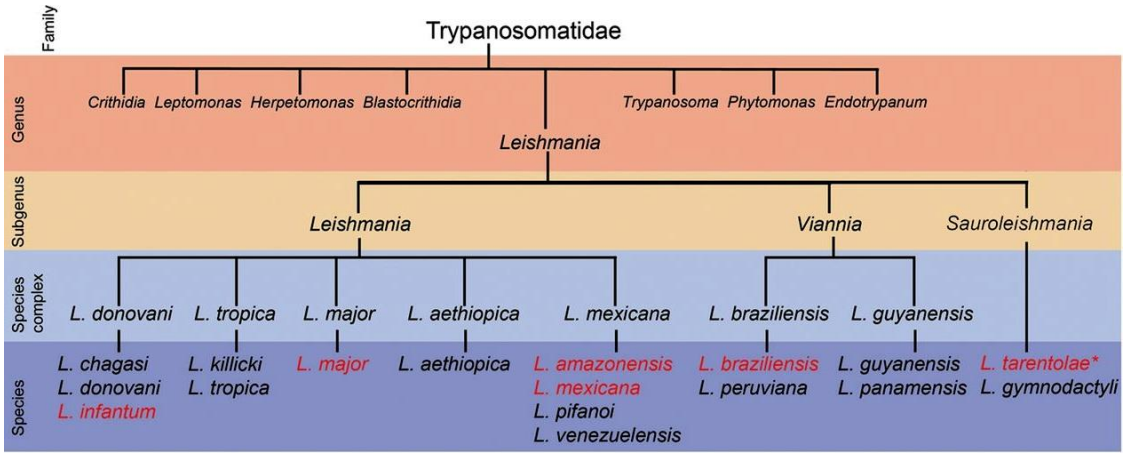
Taxonomy of *Leishmania*

The *Leishmania* genus belongs to the *Trypanosomatidae* family sharing its taxonomic position with other genus of public health importance like *Trypanosoma*. The *Leishmania* genus is comprised of more than 35 different species, of which at least 20 are pathogenic to humans and are classified in three subgenera (*Leishmania, Viannia* and *Sauroleishmania*) **(Preface: Figure 4)**. The classification of these species has been debated for a long time in the scientific community and is not fully resolved. *Leishmania* was first classified based on ecobiological criteria including vectors, geographic distribution and clinical manifestations (Lainson e Shaw, 1979; Lainson *et al.*, 1987; Desjeux, 2004; Murray *et al.*, 2005; Goto e Lindoso, 2010). In this sense, Lainson and Shaw identified distinct patterns in the development of some species inside the sand fly tract and classified them based on their location as suprapylaria, peripylaria and hypopylaria (Killick-Kendrick *et al.*, 1977; Lainson *et al.*, 1977; Lainson *et al.*, 1979).

Based on these criteria, the two most important subgenus (*Viannia* and *Leishmania*) present distinct sites of development. The *Viannia* subgenus that includes species exclusively from the New World is

characterized by a phase of development at the hindgut, multiplication at the midgut and subsequent invasion of the foregut (peripylaria) (Lainson *et al.*, 1979) . The *Leishmania* subgenus belong to a common monophyletic lineage that includes New World and Old World species with a phase of intraluminal development in the midgut (suprapylaria)(Lainson *et al.*, 1977).

These extrinsic classification criteria were later complemented with the use of biochemical and molecular tools includying the use of multilocus enzyme electrophoresis to distinguish species complexes within both subgenera (Rioux *et al.*, 1990) and the detection of polymorphisms in the kinetoplast DNA and nuclear genes among species (Tsukayama *et al.*, 2009; Van Der Auwera *et al.*, 2014).



Preface: Figure 4

Classification of *Leishmania.* Adapted from: Real et al., 2013 (Real *et al.*, 2013)

Genomics of *Leishmania*

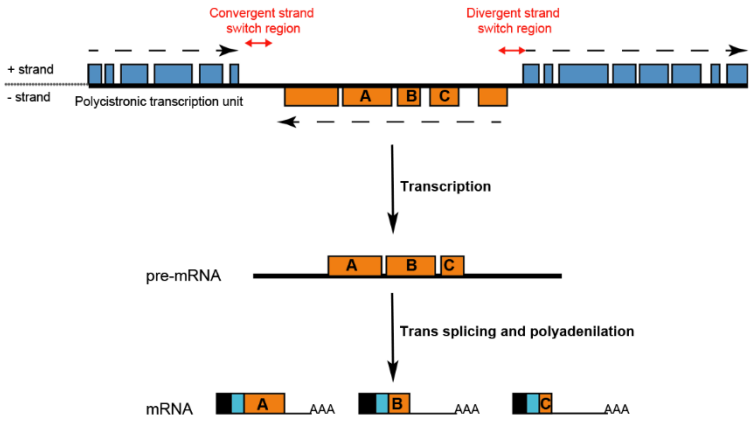
The availability of genomic sequences of several Trypanosomatid species has greatly contributed to a better understanding of the biology of these parasites.

Leishmania genome endeavors have resulted in the sequencing of several species of *Leishmania* including *L. (L.) major* Friedlin (Ivens *et al.*, 2005), *L. (L.) infantum* JPCM5, *L. (L.) braziliensis* M2904 (Peacock *et al.*, 2007), *L. (L.) amazonensis* M2269 (Real *et al.*, 2013) and several other draft assemblies that are available to the scientific community through the Tritryp database (Aslett *et al.*, 2010).

The genomes of *Leishmania* differ from other eukaryotic organisms in terms of their organization and transcription. Unlike most eukaryotes, there are very few introns and genes are grouped into bidirectional gene clusters that can comprise hundreds of genes located at the same DNA strand and are separated by strand switch regions (Myler *et al.*, 1999; Ivens *et al.*, 2005)**(Preface: Figure 5)**. Some of the multigene families are distributed in tandem gene arrays (Myler *et al.*, 1999).

Transcription in these organisms is polycistronic in a similar way as operons in prokaryotes. The polycistrons are post-transcriptionally processed through polyadenylation and trans-splicing of the spliced-leader sequence containing CAP to the 3' and 5' end the messenger RNA, respectively (Peacock *et al.*, 2007).

The analysis of *Leishmania* genomes have revealed important similarities as well as structural differences among species. These studies have shown a high degree of synteny across species despite a breach of 36-46 million years of divergence between New World and Old World species (Lukes *et al.*, 2007). Importantly, approximately 200 genes have been identified with differential distributions across *L. (L.) major, L. (L.) infantum,* and *L. (V.) braziliensis* (Peacock *et al.*, 2007).



Preface: Figure 5

Genomic organization and transcription in *Leishmania*.

The genomes of *L. (L.) donovani, L. (L.) major* and *L. (L.) infantum* are organized into 36 chromosomes (Wincker *et al.*, 1996) whereas New World species *L. (V.) braziliensis* and *L. (L.) mexicana* present of 35 and 34 chromosomes respectively, due to fusion events between chromosomes 20 and 34 in *L. (V.)*

braziliensis and chromosomes 8 and 29 and 20 and 36 in *L. (L.) mexicana* (Britto *et al.*, 1998; Peacock *et al.*, 2007).

Recently, gene duplication and tandem gene arrays have been proposed as a form to increase gene expression in the absence of a regulatory transcriptional mechanism with numerous gene amplifications of species-specific genes (Rogers *et al.*, 2011). Analyses of chromosomal content from different cells within the same isolate have led to conclude that *Leishmania* presents a mosaic structure; this is referred as neighboring cells within the same isolate presenting different chromosomal content (Sterkers *et al.*, 2012).

Increased gene copy number (CN) due to this mosaic composition may contribute to gene expression modulation in response to environmental changes (Rogers *et al.*, 2011; Sterkers *et al.*, 2012), although this needs to be further investigated. As a consequence, mosaic aneuploidy appears to be unstable within the same isolate as has been shown to occur in the *L. major* Friedlin strain (Sterkers *et al.*, 2012).

JUSTIFICATION

The leishmaniasis constitute a major health problem worldwide with an increase in its incidence and distribution with more than 98 endemic countries and 1.5 million documented cases per year. Clinical manifestations of this disease are diverse and can lead to disfigurement, functional impairment and even death. In this line, it is known that this wide range of symptoms are mainly associated with the infecting *Leishmania* species and the host immune response.

However, genomics studies have shown few inter-species differences in terms of gene content, a high degree of genomic conservation and synteny in these parasites. This evidence suggests the presence of other parasite related factors that might contribute for the distinct disease outcomes.

A better understanding of such aspects might be useful for identifying new targets for vaccines and drug development as well as to assess the structure of the parasite population in endemic foci with the ultimate goal to improve and design better strategies for controlling this devastating disease.

GENERAL OBJECTIVE

The overall goal of this research is to identify specific feature of the genomes of *Leishmania* that may contribute to the distinct clinical presentation, mediate host parasite interaction and assess the population structure in endemic foci with a special interest in New World *Leishmania* species.

For this purpose, we employed a comparative genomics approach (Chapter 1), a phylogenomics method (Chapter 2) and a population genomics study (Chapter 3).

CHAPTER I: Comparative genomic analysis of Leishmania *(Viannia) peruviana* and *Leishmania (Viannia) braziliensis*

Justification

The *L. (V.) braziliensis* complex is among the most important *Leishmania* groups in the New World (Mimori *et al.*, 1989). This complex comprises two closely related species (*L (V.). peruviana* and *L. (V.) braziliensis*) whose status as distinct species have been debated (Fraga *et al.*, 2013).

Phylogeny studies on *L. (V.) peruviana* have given contradictory results regarding its position as a distinct species or as a sub-species of *L. (V.) braziliensis*, whereas studies using multi-locus enzyme electrophoresis have treated them as closely related but still different species (Arana *et al.*, 1990; Banuls *et al.*, 2000).

In this study we have conducted a comparative genomics analysis of two *L. (V.) peruviana* isolates against the closely related *L. (V.) braziliensis* reference strain M2904 in order to better understand species-specific variation that could influence different disease phenotypes that are typical of this complex and shed lights into its phylogenetic relationship.

The results of this study were published in BMC Genomics: DOI: 10.1186/s12864-015-1928-z.

Objectives

General Objective

The main objective of this chapter was to identify differences in the genomes of *L. (V.) braziliensis* and *L. (V.) peruviana* and determine their phylogenetical status.

Specific Objectives

To accomplish our main objective we worked on the following specific aims:

- Assembly the *L. (V.) peruviana* genome and transfer functional annotations from the *L. (V.) braziliensis* M2904 reference genome.
- Analyze variants in terms of Single Nucleotide Polymorphisms (SNPs) and Insertions and Deletions (INDELS) and assess their effects on coding sequences (CDS).

- Determine chromosome copy number variations (CNV) between *L. (V.) peruviana* and *L. (V.) braziliensis*.
- Characterize tandem duplicated gene arrays and dispersed duplicated genes.

Valdivia et al. BMC Genomics (2015) 16:715
DOI 10.1186/s12864-015-1928-z

BMC Genomics

RESEARCH ARTICLE Open Access

Comparative genomic analysis of *Leishmania (Viannia) peruviana* and *Leishmania (Viannia) braziliensis*

Hugo O. Valdivia[1,2], João L. Reis-Cunha[1], Gabriela F. Rodrigues-Luiz[1], Rodrigo P. Baptista[1], G. Christian Baldeviano[2], Robert V. Gerbasi[2], Deborah E. Dobson[4], Francine Pratlong[5], Patrick Bastien[5], Andrés G. Lescano[2,3], Stephen M. Beverley[4] and Daniella C. Bartholomeu[1*]

Abstract

Background: The *Leishmania (Viannia) braziliensis* complex is responsible for most cases of New World tegumentary leishmaniasis. This complex includes two closely related species but with different geographic distribution and disease phenotypes, *L. (V.) peruviana* and *L. (V.) braziliensis*. However, the genetic basis of these differences is not well understood and the status of *L. (V.) peruviana* as distinct species has been questioned by some.
Here we sequenced the genomes of two *L. (V.) peruviana* isolates (LEM1537 and PAB-4377) using Illumina high throughput sequencing and performed comparative analyses against the *L. (V.) braziliensis* M2904 reference genome. Comparisons were focused on the detection of Single Nucleotide Polymorphisms (SNPs), insertions and deletions (INDELs), aneuploidy and gene copy number variations.

Results: We found 94,070 variants shared by both *L. (V.) peruviana* isolates (144,079 in PAB-4377 and 136,946 in LEM1537) against the *L. (V.) braziliensis* M2904 reference genome while only 26,853 variants separated both *L. (V.) peruviana* genomes. Analysis in coding sequences detected 26,750 SNPs and 1,513 indels shared by both *L. (V.) peruviana* isolates against *L. (V.) braziliensis* M2904 and revealed two *L. (V.) braziliensis* pseudogenes that are likely to have coding potential in *L. (V.) peruviana*. Chromosomal read density and allele frequency profiling showed a heterogeneous pattern of aneuploidy with an overall disomic tendency in both *L. (V.) peruviana* isolates, in contrast with a trisomic pattern in the *L. (V.) braziliensis* M2904 reference.
Read depth analysis allowed us to detect more than 368 gene expansions and 14 expanded gene arrays in *L. (V.) peruviana*, and the likely absence of expanded amastin gene arrays.

Conclusions: The greater numbers of interspecific SNP/indel differences between *L. (V.) peruviana* and *L. (V.) braziliensis* and the presence of different gene and chromosome copy number variations support the classification of both organisms as closely related but distinct species.
The extensive nucleotide polymorphisms and differences in gene and chromosome copy numbers in *L. (V.) peruviana* suggests the possibility that these may contribute to some of the unique features of its biology, including a lower pathology and lack of mucosal development.

* Correspondence: daniella@icb.ufmg.br
[1]Laboratório de Imunologia e Genômica de Parasitos, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
Full list of author information is available at the end of the article

Background

Leishmaniasis is a neglected tropical disease caused by a group of digenetic protozoan belonging to the genus *Leishmania*. It is transmitted by the bite of an infected female phlebotomine sand fly belonging to the genus *Lutzomyia* in the New World and *Phlebotomus* in the Old World [1]. Leishmaniasis is endemic in 98 countries and causes more than 1.5 million cases per year with more than 350 million people at risk [2, 3].

Leishmaniasis presents a wide spectrum of clinical manifestations that ranges from cutaneous leishmaniasis (CL) that affects tissues near the sand fly bite to mucosal leishmaniasis (ML) that is characterized by a progressive ulceration at the nares and nasal septum to the lethal visceral leishmaniasis (VL) that disseminates to visceral organs causing hepatomegaly, splenomegaly and even death [3, 4].

The *L. (V.) braziliensis* complex is one of the most important *Leishmania* group in the New World [5]. It comprises two closely related species (*L. (V.) peruviana* and *L. (V.) braziliensis*) [6], although there is some controversy regarding their status as distinct species [6]. As currently defined, *L.(V.) peruviana* is an endemic species in Peru with a limited distribution range within the Andean and inter-Andean valleys with some narrow areas of sympatry with *L. (V.) braziliensis* [7, 8].

L. (V.) peruviana causes CL and has been isolated from peridomestic mammals including dogs, mice and opossums, revealing its zoonotic status [9]. *L. (V.) braziliensis* is widely distributed in South America, although primarily in the Amazon Basin, and is referred as an anthropozoonosis [10]. *L. (V.) braziliensis* infections have a substantially higher potential to manifest as ML than any other new world leishmaniasis species, including *L. (V.) peruviana* [3, 11]. However, the parasite genetic factors that contribute to the differences in the pathogenesis of these two species are not well known.

Next generation sequencing has provided several advantages for characterizing species-specific traits across the genomes of several organisms. In *Leishmania* it has allowed to rapidly and comprehensively analyze a wide range of mutation types, including gene copy number variations (CNV) and aneuploidy [12]. Recently, CNV and expansions in tandem gene arrays have been proposed as a mechanism to increase gene expression with numerous species-specific gene amplifications [12]. These studies have suggested that extensive variation among duplicated tandem gene arrays plays a role in higher expression of their products and a diversification process in amplified genes [13]. Moreover, analysis of the chromosomal content from different cells within the same isolate have led to conclude that *Leishmania* presents a mosaic structure that may contribute to gene expression changes in response to environmental alteration modulating parasite phenotypes [12, 14].

In this study, we have conducted a comparative genomics analysis of two L. *(V.) peruviana* isolates against the reference genome M2904 of *L. (V.) braziliensis*. Comparative assessments have shown important differences in chromosome and gene copy number between both species. These analyses may serve to improve our understanding of parasite variation between these two closely related species that could be linked to their different disease phenotypes and to provide further insights into their status as distinct species.

Results and Discussion

Genome assembly

We used a combined *de novo* and reference based assembly approach (Baptista et al. in preparation) to generate a draft genome for each strain. *L. (V.) peruviana* mapped reads showed an overall 92.51 % mapping rate for PAB-4377 and 95.87 % for LEM1537 against *L. (V.) braziliensis*. Median genome coverage estimated from mapped reads into 6,899 single copy genes was of 59.1 and 35.0 for PAB-4377 and LEM1537, respectively.

The *L. (V.) peruviana* assemblies resulted in 28.51 and 25.27 megabases that were generated from 11,504 and 29,816 contigs in PAB-4377 and LEM1537, respectively. The resulting ordered assemblies consisted of 37 pseudochromosomes, due to the split of chromosome 20 in the *L. (V.) braziliensis* M2904 reference genome (LbrM.20.1 and LbrM.20.2) and a pseudo-chromosome containing unordered scaffolds (Chromosome 0).

The overall identity between *L. (V.) braziliensis* and *L. (V.) peruviana* calculated with MUMmer [15] confirmed the close relationship between *L. (V.) braziliensis* and *L. (V.) peruviana* (identity of 87.58 % for PAB-4377 and 77.1 % for LEM1537), and a closer relationship between the two *L. (V.) peruviana* isolates (99 %).

SNP and Indel comparisons

Variants were identified following filtering for quality, read depth and haplotype score as described in the methods.

Comparisons identified 144,079 and 136,946 variants between *L. (V.) braziliensis* and *L. (V.) peruviana* PAB-4377 (115,851 SNPs and 28,228 Indels) and *L. (V.) peruviana* LEM1537 (108,826 SNPs and 28,120 Indels), respectively. Of these; 94,070 variants were shared between the two *L. (V.) peruviana* isolates. In contrast, the two *L. (V.) peruviana* isolates showed fewer variants among them (26,853). This finding is consistent with the high similarity obtained with MUMmer3 between both *L. (V.) peruviana* isolates and the greater difference with *L. (V.) braziliensis*.

Our results show that there is significant genetic differentiation between *L. (V.) braziliensis* and *L. (V.) peruviana* while intra *L. (V.) peruviana* variation is substantially lower. For comparison, a previous comparative study between *L. (L.) infantum* and *L. (L.) donovani* reference genomes found that 156,274 nucleotide changes differentiate between these

closely related species [16], comparable to what we describe here for *L. (V.) braziliensis* and *L. (V.) peruviana*.

We then focused on the 94,070 variants from *L. (V.) braziliensis* that were shared by the two *L. (V.) peruviana* lines. Of these; 26,750 SNPs were located in 6,114 coding DNA sequences (CDS) (Additional file 1: Table S1). Of these, 14,244 SNPs (53.24 %) were synonymous mutations and 12,462 (46.59 %) were non-synonymous mutations. Additionally, eight SNPs mutating the annotated start codon (0.03 %) and 36 mutating the annotated stop codon were found (0.13 %). Most genes with high counts of SNP are hypothetical proteins, kinases and trafficking proteins stressing the need to characterize the function of these variable proteins (Table 1).

Variant calls for indels shared by both isolates detected 1,513 sites distributed in 408 CDS (Additional file 1: Table S2). Of these, 1,014 (67.0 %) were codon deletions, 146 (9.6 %) were insertions, 351 (23.2 %) frameshifts and two stop codons (0.1 %) were gained. Genes with most bases affected by indels include hypothetical proteins, kinesins and a lysine transport protein (Table 2).

Analysis of potential diagnosis targets that could accurately differentiate *L. (V.) peruviana* from *L. (V.) braziliensis* resulted in 270 genes with high SNP density regions between both species (Additional file 2, Additional file 1: Table S3). While most of these genes are hypothetical proteins, they could serve to design better molecular diagnosis tools to discriminate between these closely related species.

Two *L. (V.) braziliensis* pseudogenes (LbrM.04.0060, LbrM.28.2130) appeared to be functional in *L. (V.) peruviana*. LbrM.28.2130 codes for an X-pro, dipeptidyl-peptidase, serine peptidase and has orthologs in other *Leishmania* species from the Old and New World suggesting that it could be functional in *L. (V.) peruviana*. Peptidases have an important role in parasite survival, invasion, metabolism and host-parasite interaction [17], highlighting the importance of confirming coding function of this potential gene. LbrM.04.0060 codes for a putative pteridine transporter and shares 84 % identity with a folate/biopterin in *L. infantum*. It has been shown that *Leishmania* are pteridine auxotrophs and rely on a network of folate and biopterin transporters. Pteridine levels have a strong influence on metacyclogenesis in *L. (L.) major* [18].

Chromosome copy number variation

Chromosome numbers were estimated by the average read density to each chromosome, and normalized to an assumed overall genome ploidy of 2n. Normalized chromosome copy number clustered around "disomy" although with significant departures from non-integral values evident for some chromosomes (Fig. 1). This finding is particularly important since the *L. (V.) braziliensis* M2904 strain is mostly trisomic [12].

The most pronounced departure from disomy occurred in chromosome 31, which presented a read depth between tetrasomy to hexasomy in PAB 4377 and trisomy in LEM1537 (Fig. 1, Additional files 3 and 4). In both isolates, read depth was evenly distributed along the entire sequence of Chr31, arguing against region-specific amplification (Fig. 2).

In both samples, chromosomes 1–5 and 7 appear to be closer to monosomy. This characteristic has also been estimated for chromosomes 1 and 3 of *L. (L.) mexicana* [12]. Interestingly, the pattern of aneuploidy involving chromosomes 8, 11, 20 and 22 in LEM1537 and 35 in PAB-4377 is different from the median ploidy of the rest of the chromosomes in both samples. These chromosomes appear to have intermediate read depth between disomic and trisomic profiles, suggesting a mosaic ploidy within the cell population (Fig. 1).

It has been suggested that mosaic aneuploidy could be a mechanism of rapid parasite adaptation in response to environmental changes within its host [14] and it has been

Table 1 Top ten high SNP count genes in two *L. (V.) peruviana* isolates

Gene ID	Annotation	Number of SNP	Gene length	CN PAB	LEM
LbrM.33.3060	Hypothetical proteins	135	14,943	0.97	0.87
LbrM.30.2340		83	11,340	0.98	0.75
LbrM.34.5330		82	19,875	1.07	1.25
LbrM.16.0180		69	13,302	0.82	0.79
LbrM.35.1580		68	16,767	1.01	0.76
LbrM.14.0770		63	12,570	0.68	0.39
LbrM.35.3160		43	12,582	1.05	0.98
LbrM.30.2160	Endosomal trafficking protein RME-8, putative	40	7335	1.20	1.19
LbrM.02.0130	Phosphatidylinositol kinase related protein, putative	39	14,775	0.51	0.47
LbrM.30.1620	protein kinase, putative	38	5112	1.30	1.23

Top ten genes showing high SNP differences in *L. (V.) peruviana* compared with *L. (V.) braziliensis* orthologs. Number of SNP and gene length are presented in nucleotides. Copy number (CN) estimated for the haploid genome of PAB-4377 (PAB) and LEM-1537 (LEM)

Table 2 Top ten high INDEL count genes in L. (V.) peruviana

Gene ID	Annotation	Affected nucleotides	Gene length	CN PAB LEM	
LbrM.17.0390	Hypothetical proteins	57	3480	0.63	0.52
LbrM.21.1080		42	2895	0.76	0.56
LbrM.15.1180	Nucleoside transporter 1, putative	28	1848	1.34	1.33
LbrM.34.2710	Hypothetical protein	24	2133	1.43	1.6
LbrM.14.0785	Kinesin, putative	21	957	0.75	1.02
LbrM.31.1470	Hypothetical proteins	21	4089	0.82	0.66
LbrM.32.3450		21	2469	0.74	0.54
LbrM.33.2950		21	3582	0.91	0.77
LbrM.07.1050	RNA binding protein-like protein	19	1377	1.02	1.21
LbrM.25.1000	Hypothetical proteins	18	19,518	0.56	0.33

Top ten high indel count genes in *L. (V.) peruviana* compared with *L. (V.) braziliensis* orthologs. Gene length is presented in nucleotides. Copy number (CN) estimated for the haploid genome of PAB-4377 (PAB) and LEM-1537 (LEM)

shown to occur in closely related strains [16]. However, its origin in *Leishmania* remains to be investigated [16].

A second approach for assessing chromosome number is based upon allele frequencies. For disomic chromosomes, heterozygous SNPs should cluster around 50 %, while trisomic chromosomes should show two peaks at 33 and 67 % and tetrasomics at 25, 50 and 75 %, [12].

Allele frequency counts for each predicted heterozygous SNP further confirmed the overall disomic tendency (Fig. 3) and the highly heterogeneous structure within the cell populations with chromosomes presenting mixtures of allele profiles (Additional files 5 and 6).

Chromosomes with discordance between read depth analysis and their allele frequencies included chromosome 5, 7, 13, 17 and 19 that presented a tetrasomic or a mixture of trisomic and disomic patterns in PAB-4377 (Additional file 5).

In LEM1537, chromosomes 6 and 9 did not have a marked allele frequency pattern and chromosome 11, 14 and 25 presented discordance between read depth and allele frequencies (Additional file 6). Additionally, chromosomes 22, 23, 28 and 34 presented mixtures of disomy and monosomy that corresponded with their estimated read depth (Additional file 6, Fig. 1).

Discordance between allele frequency and read depth may be explained by cells presenting a high variation in their ploidies due to chromosome mosaism as has been previously suggested [12].

Interestingly, chromosome 31 that has been identified as supernumerary in both isolates appears to have disomic pattern (Additional files 5 and 6). This chromosome has been previously described as supernumerary in all *Leishmania* species [12]. It may be possible that this chromosome accumulates mutations in disomic alleles

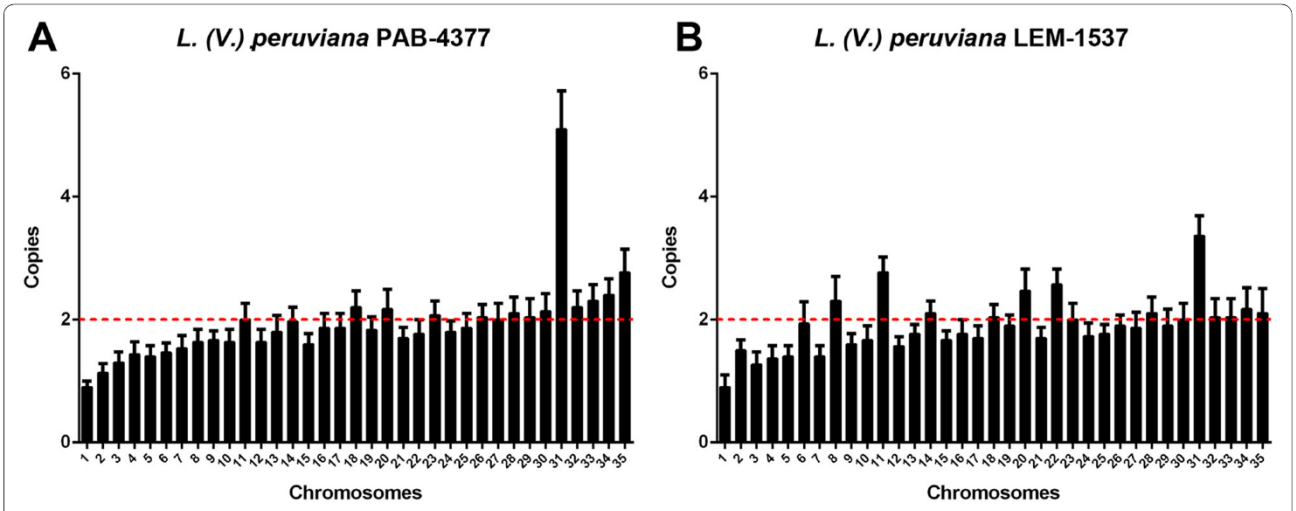


Fig. 1 Chromosome copy number in *L. (V). peruviana*. Chromosome copy number variation in *L. (V.) peruviana*. **a** PAB-4377; **b** LEM-1537. Boxes represent the estimated copy number for each chromosome and standard deviation from the three methods. Mean genome ploidy is indicated by a dotted red line

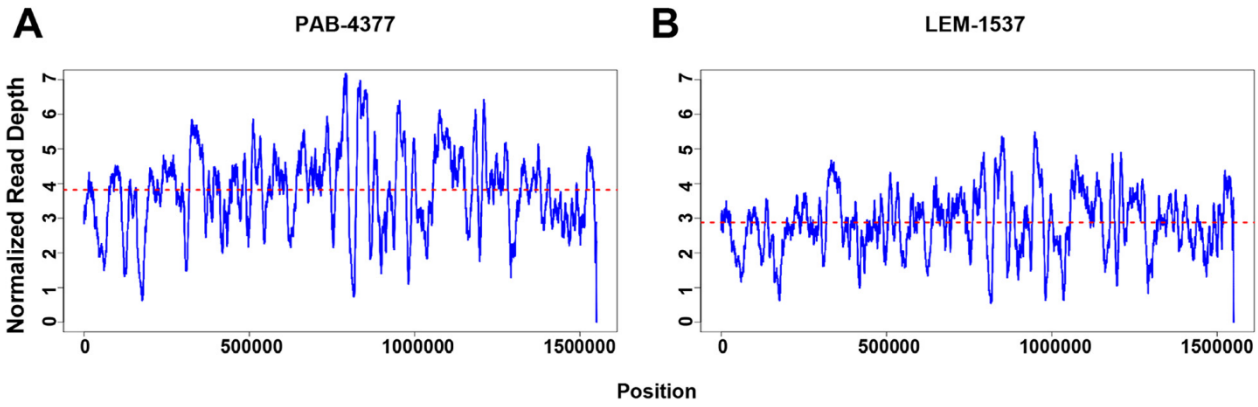


Fig. 2 Chromosome 31 normalized read depth. Normalized read depth for supernumerary chromosome 31. **a** PAB-4377; **b** LEM-1537. Estimated ploidy indicated by a dotted red line. Blue lines represent normalized read depth for each position at the chromosome

as has been reported in other chromosomes with the same pattern in *L. (L.) mexicana* [12].

Ontology analysis in the supernumerary chromosome 31 showed that this chromosome is enriched in genes involved in iron metabolism and other related molecular functions (Table 3, Additional file 1: Table S4). Iron sulfur proteins (Fe-S) are crucial for life since they mediate oxidation-reduction reactions during mitochondrial electron transport and are involved in the synthesis of amino acids, biotin and lipoic [19].

Biosynthesis of Fe-S proteins is highly dependent on iron regulation in the cell [20]. Interestingly, ferrous iron transporters located in chromosome 31 have been described in *Leishmania* and they appear to be important for growth, replication and pathology, further stressing this connection [21, 22].

A sustained copy number increase in chromosome 31 among all *Leishmania* species [12] could serve as a mechanism to facilitate iron uptake and increase gene dosage of Fe-S proteins in an oxidative stressed environment.

Gene copy number variation

Expanded tandem gene arrays and dispersed genes appear to be a major source of inter and intra-species variation in *Leishmania* [12]. The tandem duplicated gene arrays analysis showed a total of 20 and 26 expanded arrays in PAB-4377 (Fig. 4a) and LEM1537 (Fig. 4b), respectively, relative to the *L. (V.) braziliensis* reference genome (Additional file 1: Table S5).

In both samples, 14 tandem arrays were shared showing that gene array expansions may vary across strains from the same species (Additional file 1: Table S5). The most expanded gene arrays in both isolates belonged to a group of TATE DNA transposons (OG5_128620), NADH-dependent reductases (OG5_128620), heat shock protein 83 (OG5_126623) and hypothetical proteins among others (Additional file 1: Table S5).

The same analysis in *L. (V.) braziliensis* M2904 resulted in 18 tandem gene arrays from which only three arrays were shared with *L. (V.) peruviana* (Additional file 1: Table S6). Interestingly, amastin surface protein arrays

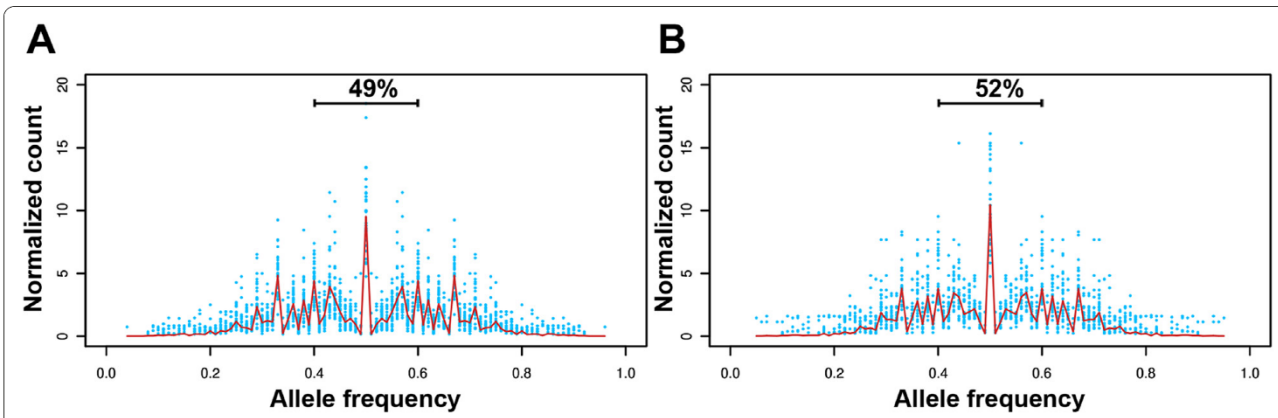


Fig. 3 Normalized allele frequency distribution. Normalized allele frequency counts for *L. (V.) peruviana*. **a** PAB-4377; **b** LEM-1537. Blue dots show normalized counts at heterozygous positions for all disomic chromosomes. Mean count at each allele frequency is indicated by a red line. Cumulative percentage between 0.4 and 06 heterozygous frequencies support disomic tendency of most chromosomes

Table 3 Ontology analysis for chromosome 31

Go ID	Description	Corrected *p*-value
51,537	2 iron, 2 sulfur cluster binding	1.08E-03
9055	electron carrier activity	1.85E-02
4198	calcium-dependent cysteine-type endopeptidase activity	1.85E-02
51,536	iron-sulfur cluster binding	1.85E-02
51,540	metal cluster binding	1.85E-02
4148	dihydrolipoyl dehydrogenase activity	1.85E-02
4197	cysteine-type endopeptidase activity	3.81E-02
8234	cysteine-type peptidase activity	4.94E-02

that are present in *L. (V.) braziliensis* seems to be not expanded in *L. (V.) peruviana*.

Amastins have been shown to be highly expressed in the amastigote life stage and appear to mediate host-parasite interactions allowing infection and survival [23]. While the effect of this variation remains to be confirmed, these differences may be related with different host interactions in both species.

We found 398 and 942 dispersed duplicated genes in PAB-4377 and LEM1537 with 360 expansions in common (Fig. 4c, d, Additional file 1: Table S7 and S8). Most expanded genes include thioredoxins, NADH-dependent fumarate reductases and several hypothetical proteins.

We did not detect an increase in copy number in GP63 genes in *L. (V.) peruviana* as has been previously shown in *L. (V.) braziliensis* [12] reinforcing a previous finding of GP63 copy number differences between these species [24].

The zinc-metalloprotease GP63 stands out as a major virulence factor in *Leishmania* presenting different roles in the vector and mammal host that aim to protect parasites from host immune responses and promote infection [25]. Therefore, deletion of some GP63 genes in *L. (V.) peruviana* could affect parasite-host interactions and influence its distribution and clinical manifestation with lack of mucosal development.

The marked intra-species difference in dispersed duplicated genes shows that extensive variation in gene copy number can occur between isolates belonging to the same species and supports the hypothesis that chromosome and gene CNV act as a mechanism of rapid parasite adaptation [12, 26].

Conclusions

Extensive chromosomal and gene copy number variations have been described in *Leishmania* and were proposed as a mechanism of rapid parasite adaptation to different environments and pressures in the host. Our study shows that there are major differences regarding gene copy number variations and aneuploidy even between closely related *Leishmania* species.

Although highly similar to *L. (V.) braziliensis*, *L. (V.) peruviana* presents a different set of expanded gene arrays that can result in different expression profiles between both species. Moreover, high SNP and indel counts as well as extensive variation in chromosome and gene copy numbers between *L. (V.) peruviana* and *L. (V.) braziliensis* support maintaining the classification of both organisms as closely related but distinct species.

Further analysis including a greater number of *L. (V.) peruviana* and *L. (V.) braziliensis* isolates and the use of transcriptomic data are needed to assess if these differences are conserved across isolates of *L. (V.) peruviana* and reveal how tandem gene arrays and CNV affect genome expression.

Methods

Parasite isolates and sequencing

L. (V.) peruviana isolate PAB-4377 was kindly provided by the U.S. Naval Medical Research Unit No. 6 (NAMRU-6) and the LEM1537 (MHOM/PE/84/LC39) isolate was obtained from the Montepellier reference center.

PAB-4377 was confirmed as *L. (V.) peruviana* by Multilocus Enzyme Electrophoresis (MLEE) and sequencing of the Manose Phosphate Isomerase and 6-phosphogluconate dehydrogenase genes. LEM1537 is a *L. (V.) peruviana* reference strain (MHOM/PE/84/LC39) and has been widely characterized by MLEE.

Libraries consisting of 350 bp fragments were obtained and 100 bp paired end reads were generated at the Genome Technology Access Center (GTAC) at Washington University in St. Louis by Illumina HiSeq 2000. The version 6 of the *L. (V.) braziliensis* M2904 genome was obtained from the Tritryp database (http://tritrypdb.org/) to serve as a reference for comparative analysis.

Genome assembly and annotation

L. (V.) peruviana reads were filtered by quality using Trimmomatic [27] with a minimum base quality cutoff of 30, leading and trailing base qualities of 28, five bases sliding window with minimum per base average quality of 20 and a minimum read length of 70 bp.

A combined *De novo* and reference based assembly approach (Baptista et al., in preparation) was used to generate a draft assembly for each sample. Briefly, *De Novo* assemblies were generated using the Velvet optimizer perl script under Velvet version 1.2.10 [28]. Draft assemblies were extended by iterative mapping using IMAGE [29] and corrected using iCORN2 [30].

For reference-based assembly, reads from each sample were mapped against the *L. (V.) braziliensis* M2904 genome using Bowtie2 [31]. Redundant reads were removed and a reference-based sequence was generated using SAMtools Mpileup and vcfutils [32] using base quality scores greater or equal than 40, mapping quality scores

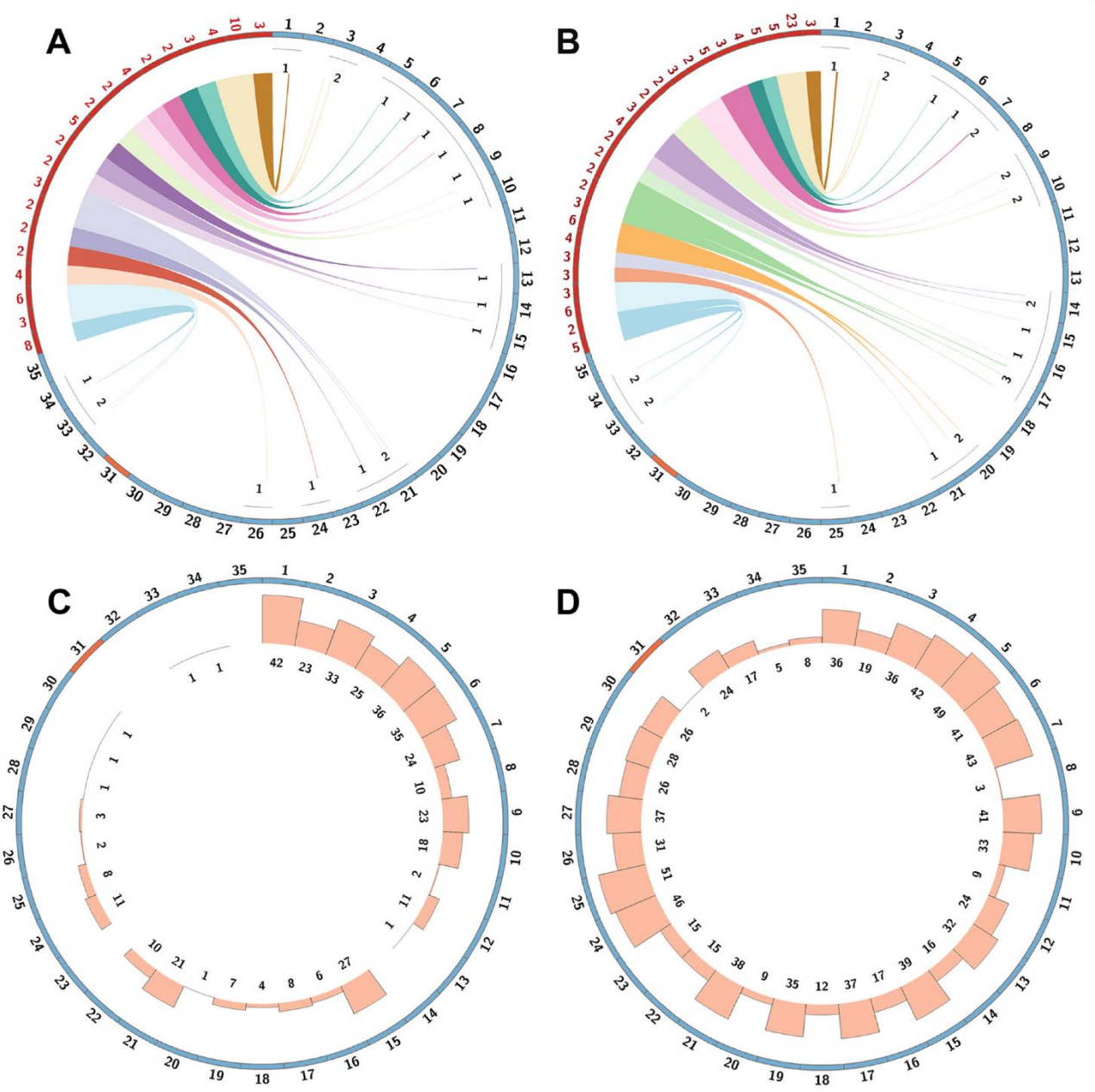


Fig. 4 Gene copy number variations in *L. (V.) peruviana*. Mapping of expanded genes in both *L. (V.) peruviana* isolates. **a**, **b** Tandem duplicated gene arrays in PAB-4377 and LEM-1537, respectively. Outer circle shows gene arrays as red with numbers indicating the calculated array expansion. Disomic chromosomes are shown in blue and supernumerary chromosomes in orange with outer numbers describing each chromosome. Colored lines map the location of duplicated arrays in their respective chromosomes. **c**, **d** dispersed duplicated genes in PAB-4377 and LEM-1537, respectively. Histogram and numbers represents the total number of gene expansions in each chromosome

greater or equal than 25, coverage greater or equal than 10 reads and less than twice the median genome coverage.

De Novo and referenced based sequences of each sample were combined using the ZORRO hybrid assembler as previously described [33]. The final hybrid assemblies were furthered extended and corrected with IMAGE and iCORN and contigs were scaffolded with SSPACE [34]. Scaffolds were aligned and orientated into pseudochromosomes with ABACAS [35] using the *L. (V.) braziliensis* M2904 genome as a reference sequence.

MUMmer3 [15] was used to calculate similarity between the assembled *L. (V.) peruviana* genomes and the reference *L. (V.) braziliensis*. Briefly, identity scores and number of bases from best local alignments among

assembled and reference genomes were retrieved and normalized with the total number of bases in the draft genome in order to compute a global identity score.

Read and assembly files are available through the European Nucleotide Archive under the project number PRJEB7263.

SNP and pseudogene analysis

To detect SNPs between *L. (V.) peruviana* and *L. (V.) braziliensis* and determine their potential effects on coding sequences, *L. (V.) peruviana* reads were mapped onto the *L. (V.) braziliensis* M2904 reference genome using Bowtie2 and analyzed using the recommended parameters of GATK [36]. Briefly, mapped bam files were filtered for redundant reads and local realignment was performed around indels in order to remove potential mapping artifacts. SNPs were called using the haplotype caller module and raw variants were filtered using GATK's variant quality score recalibration selecting sites with a minimum raw coverage of 10, Root Mean Square mapping quality lower than 40, quality by depth greater than 2 and haplotype score greater than 13. The same method was employed to call variants between both *L. (V.) peruviana* isolates.

To analyze the effects of SNPs in coding regions of the *L. (V.) peruviana* genome, we filtered variant calls of PAB-4377 and LEM1537 selecting only SNPs shared by both isolates to limit the potential impact of within-species SNP variability and minimize incorrect SNP calling. The combined variant called was used as input for SnpEff [37] to annotate and predict the effects of variants of genes.

To find potential targets sequences to accurately discriminate *L. (V.) peruviana* from *L. (V.) braziliensis* we employed a custom Perl script to screen the genes with variant calls. These genes were analyzed using a sliding window of 1000 nucleotides to report the region with the highest SNP density and the number of SNP that it presented. Genes with significant SNP calls were detected using the ROUT test under Graph Pad Prism V5 [38]

We downloaded *L. (V.) braziliensis* pseudogenes from the Tritryp database and compared them against *L. (V.) peruviana* to detect potential pseudogenes that remained functional in *L. (V.) peruviana*. Briefly, *L. (V.) peruviana* amino acid fasta sequences were generated using SAMtools Mpileup and translated into amino acids for sequence alignment against *L. (V.) braziliensis* pseudogenes in ClustalΩ [39].

Allele frequency distribution

Allele frequencies for PAB-4377 and LEM1537 assemblies were obtained from filtered SAMtools Mpileup results as described elsewhere [12]. Briefly, the proportion of reads mapping to each heterozygous site under the total mapped reads for the site was estimated. Allele frequencies were categorized from 0.1 to 1.0 and normalized by the sum of all allele frequencies for the chromosome. Allele frequencies distributions were plotted in R and plots from chromosomes sharing the same pattern were combined.

Chromosome and tandem gene array analysis

To analyze chromosome copy number, we combined three different approaches based on the assumption that the overall chromosome organization is similar between *L. (V.) braziliensis* and *L. (V.) peruviana*. First, OrthoMCL was used to select single copy genes from the proteomes of *L. (V.) braziliensis*, *L. (L.) mexicana* and *L. (L.) major*, *L. (L.) infantum*, *L. (L.) donovani* and *L. (Sauroleishmania) tarentolae* (Additional file 1: Table S9).

This group of single copy genes was used to normalize read mapping counts for each position along the chromosome in order to calculate haploid copy number. Second, the number of reads mapping to the whole chromosome was counted and normalized by the median number of mapped reads to the whole genome. Third, we normalized FPKM (Fragments Per Kilobase per Million fragments mapped reads) values for each chromosome by the median FPKM of the whole genome. We plotted the mean and standard deviations from the three approaches using Graph Pad Prism V5.

We normalized haploid copy numbers with the average chromosome ploidy calculated from the allele frequency analysis to estimate chromosome ploidy. Plots for each chromosome were generated in R using a sliding window of 10 kilo bases.

Gene Ontology codes that were significantly overrepresented in the genes of supernumerary chromosomes were detected using the hypergeometric distribution analysis in BiNGO [40] with Benjamini and Hochberg false discovery rate correction.

We defined tandem gene arrays as groups of genes that are located contiguously in a chromosome and that share a homology relationship. Dispersed gene duplications are defined as genes that are duplicated and do not belong to any tandem array.

Dispersed and tandem gene duplications were identified using a combination of Bowtie2 and Cufflinks2 [41]. Briefly, mapped reads against *L. (V.) braziliensis* M2904 and a coding sequence (CDS) GFF file were used as input for Cufflinks2 to determine FPKM for each CDS and chromosome. Haploid copy number for each CDS was estimated by a proportion of their respective FPKM and the median FPKM of all CDS in the respective chromosome. We employed OrthoMCL [42] to identify homology relationships in mapped CDS and the mean haploid copy number was estimated for each array as reported by Rogers [12]. Gene duplications were defined as those greater than a cutoff of 1.85 for the haploid number computed by our analysis [12].

We employed this same approach to detect expanded gene arrays in the *L. (V.) braziliensis* genome using reads from the M2904 reference strain.

Additional files

Additional file 1: Supplementary tables. (XLSX 1072 kb)

Additional file 2: Top five high SNP density genes. (TIFF 448 kb)

Additional file 3: Normalized read depth for PAB-4377 chromosomes. (TIFF 2222 kb)

Additional file 4: Normalized read depth for LEM-1537 chromosomes. (TIFF 2449 kb)

Additional file 5: Normalized allele frequency distributions for PAB-4377 chromosomes. (TIFF 443 kb)

Additional file 6: Normalized allele frequency distributions for LEM-1537 chromosomes. (TIFF 441 kb)

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
HOV carried out most bioinformatics analysis, participated in study conception, design and drafted the manuscript. JLR participated in gene and chromosome copy number calculations. GFR participated in gene and chromosome copy number calculations. RPB contributed in genome assembly and manuscript drafting. GCB participated in study design, coordination and participated. RG participated in study coordination and manuscript writing. DED participated in DNA quality control, sequencing and preliminary bioinformatic analysis. FP participated in study design and coordination. PB participated in study design and coordination. AGL participated in study design, coordination and manuscript writing. SB participated in study conception, design, coordination and manuscript writing. DCB participated in bioinformatic analysis, study design, coordination and manuscript writing. All authors read and approved the final manuscript.

Authors' information
Not applicable.

Availability of data and materials
Read and assembly files are available through the European Nucleotide Archive under the project number PRJEB7263.

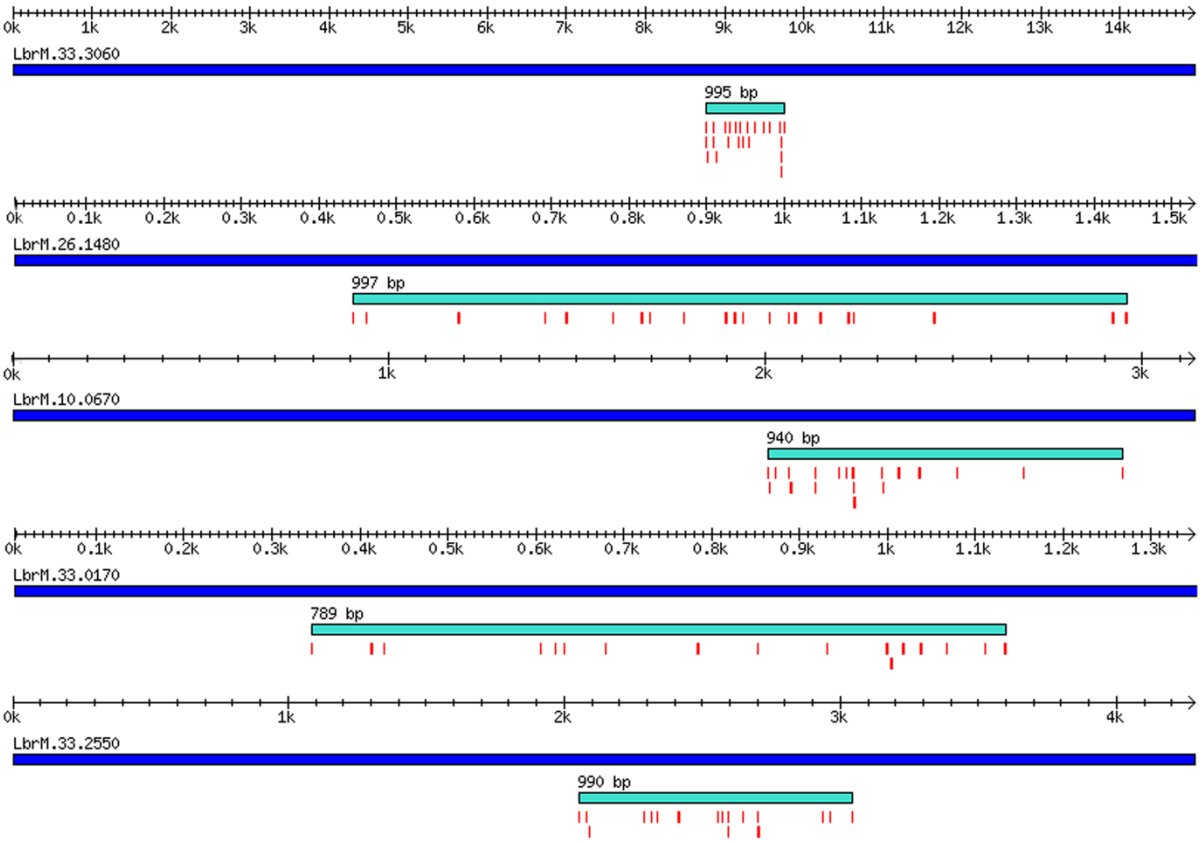
Acknowledgements
We thank Nick Dickens for his help with FPKM copy number calculations and the Genome Technology Access Center in the department of Genetics at Washington University School of Medicine for their support with next-generation sequencing. The Center is partially supported by NCI Cancer Center Support Grant #P30 CA91842 to the Siteman Cancer Center and by ICTS/CTSA Grant #UL1 TR000448 from the National Center for Research Resources (NCRR). Daniella C. Bartholomeu research was supported by Fundação de Amparo a Pesquisa do Estado de Minas Gerais (FAPEMIG), Instituto Nacional de Ciência e Tecnologia de Vacinas (INCTV)—Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). DCB is a CNPq research fellow. HOV, JLRC, GFRL received scholarships from CAPES and RPB received a scholarship from CNPq. Stephen Beverley and Deborah Dobson research was supported by NIH grants R01-AI29646 and R56-AI099364. Francine Pratlong and Patrick Bastien research was funded by the Institut de Veille Sanitaire, France.

Author details
[1]Laboratório de Imunologia e Genômica de Parasitos, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. [2]Department of Parasitology, U.S. Naval Medical Research Unit No. 6, Lima, Peru. [3]Universidad Peruana Cayetano Heredia, School of Public Health and Management, Lima, Peru. [4]Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, Missouri, USA. [5]Centre Hospitalier Universitaire de Montpellier, Departement de Parasitologie-Mycologie, Centre National de Reference des Leishmanioses, Montpellier, France.

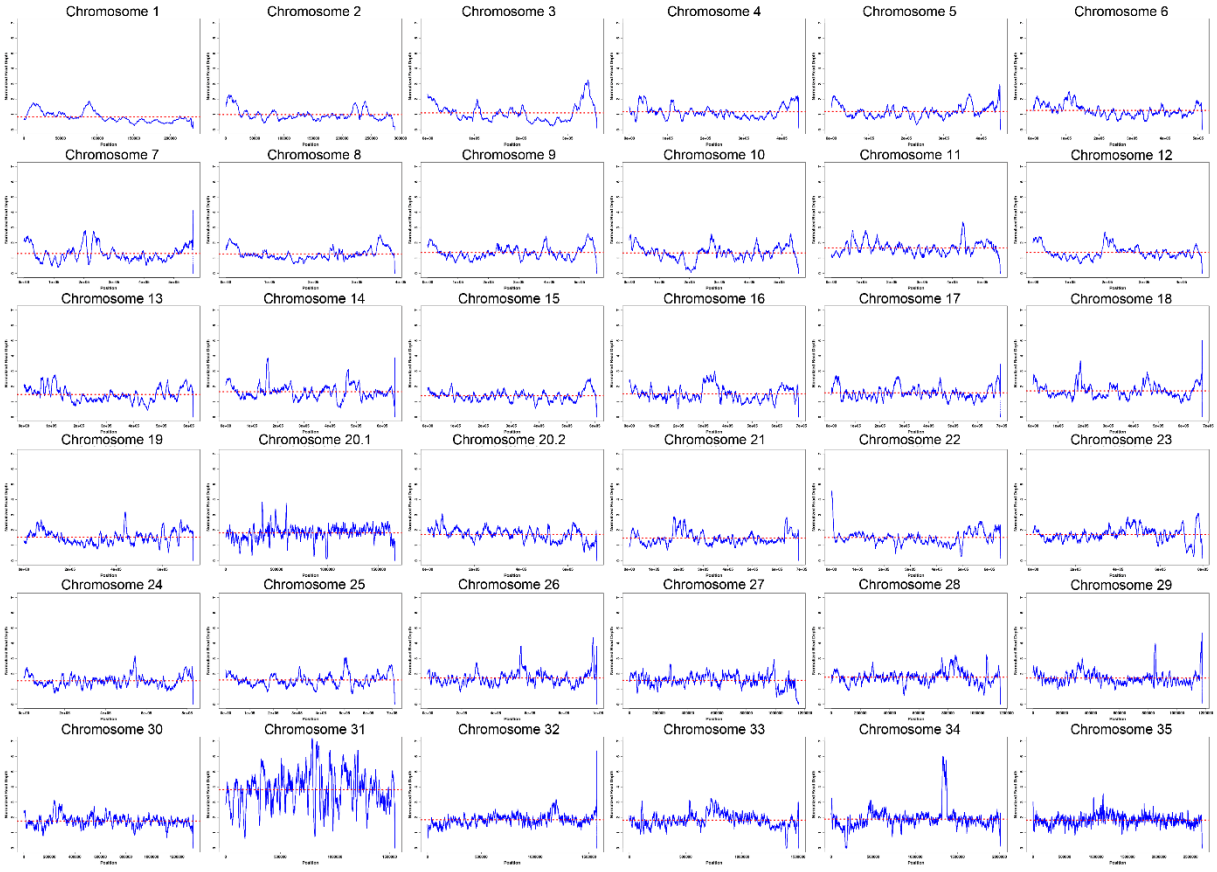
Received: 7 May 2015 Accepted: 9 September 2015
Published online: 18 September 2015

References
1. Kato H, Gomez EA, Caceres AG, Uezato H, Mimori T, Hashiguchi Y. Molecular epidemiology for vector research on leishmaniasis. Int J Environ Res Public Health. 2010;7(3):814–26.
2. Alvar J, Velez ID, Bern C, Herrero M, Desjeux P, Cano J, et al. Leishmaniasis worldwide and global estimates of its incidence. PLoS ONE. 2012;7(5):e35671.
3. Murray HW, Berman JD, Davies CR, Saravia NG. Advances in leishmaniasis. Lancet. 2005;366(9496):1561–77.
4. David CV, Craft N. Cutaneous and mucocutaneous leishmaniasis. Dermatol Ther. 2009;22(6):491–502.
5. Mimori T, Grimaldi Jr G, Kreutzer RD, Gomez EA, McMahon-Pratt D, Tesh RB, et al. Identification, using isoenzyme electrophoresis and monoclonal antibodies, of Leishmania isolated from humans and wild animals of Ecuador. Am J Trop Med Hyg. 1989;40(2):154–8.
6. Fraga J, Montalvo AM, Van der Auwera G, Maes I, Dujardin JC, Requena JM. Evolution and species discrimination according to the Leishmania heat-shock protein 20 gene. Infect Genet Evol. 2013;18:229–37.
7. Lucas CM, Franke ED, Cachay MI, Tejada A, Cruz ME, Kreutzer RD, et al. Geographic distribution and clinical description of leishmaniasis cases in Peru. Am J Trop Med Hyg. 1998;59(2):312–7.
8. Nolder D, Roncal N, Davies CR, Llanos-Cuentas A, Miles MA. Multiple hybrid genotypes of Leishmania (viannia) in a focus of mucocutaneous Leishmaniasis. Am J Trop Med Hyg. 2007;76(3):573–8.
9. Llanos-Cuentas EA, Roncal N, Villaseca P, Paz L, Ogusuku E, Perez JE, et al. Natural infections of Leishmania peruviana in animals in the Peruvian Andes. Trans R Soc Trop Med Hyg. 1999;93(1):15–20.
10. Oddone R, Schweynoch C, Schonian G, de Sousa CS, Cupolillo E, Espinosa D, et al. Development of a multilocus microsatellite typing approach for discriminating strains of Leishmania (Viannia) species. J Clin Microbiol. 2009;47(9):2818–25.
11. Odiwuor S, Veland N, Maes I, Arevalo J, Dujardin JC, Van der Auwera G. Evolution of the Leishmania braziliensis species complex from amplified fragment length polymorphisms, and clinical implications. Infect Genet Evol. 2012;12(8):1994–2002.
12. Rogers MB, Hilley JD, Dickens NJ, Wilkes J, Bates PA, Depledge DP, et al. Chromosome and gene copy number variation allow major structural change between species and strains of Leishmania. Genome Res. 2011;21(12):2129–42.
13. Victoir K, Dujardin JC. How to succeed in parasitic life without sex? Asking Leishmania. Trends Parasitol. 2002;18(2):81–5.
14. Sterkers Y, Lachaud L, Bourgeois N, Crobu L, Bastien P, Pages M. Novel insights into genome plasticity in Eukaryotes: mosaic aneuploidy in Leishmania. Mol Microbiol. 2012;86(1):15–23.
15. Delcher AL, Salzberg SL, Phillippy AM Using MUMmer to identify similar regions in large sequence sets. Current protocols in bioinformatics/editoral board, Andreas D Baxevanis [et al.] 2003, Chapter 10:Unit 10 13.
16. Downing T, Imamura H, Decuypere S, Clark TG, Coombs GH, Cotton JA, et al. Whole genome sequencing of multiple Leishmania donovani clinical isolates provides insights into population structure and mechanisms of drug resistance. Genome Res. 2011;21(12):2143–56.
17. Caroselli EE, Assis DM, Barbieri CL, Judice WA, Juliano MA, Gazarini ML, et al. Leishmania (L.) amazonensis peptidase activities inside the living cells and in their lysates. Mol Biochem Parasitol. 2012;184(2):82–9.
18. Cunningham ML, Titus RG, Turco SJ, Beverley SM. Regulation of differentiation to the infective stage of the protozoan parasite Leishmania major by tetrahydrobiopterin. Science (New York, NY). 2001;292(5515):285–7.
19. Waller JC, Alvarez S, Naponelli V, Lara-Nunez A, Blaby IK, Da Silva V, et al. A role for tetrahydrofolates in the metabolism of iron-sulfur clusters in all domains of life. Proc Natl Acad Sci U S A. 2010;107(23):10412–7.
20. Kaplan J, McVey Ward D, Crisp RJ, Philpott CC. Iron-dependent metabolic remodeling in S. cerevisiae. Biochim Biophys Acta. 2006;1763(7):646–51.
21. Huynh C, Sacks DL, Andrews NW. A Leishmania amazonensis ZIP family iron transporter is essential for parasite replication within macrophage phagolysosomes. J Exp Med. 2006;203(10):2363–75.

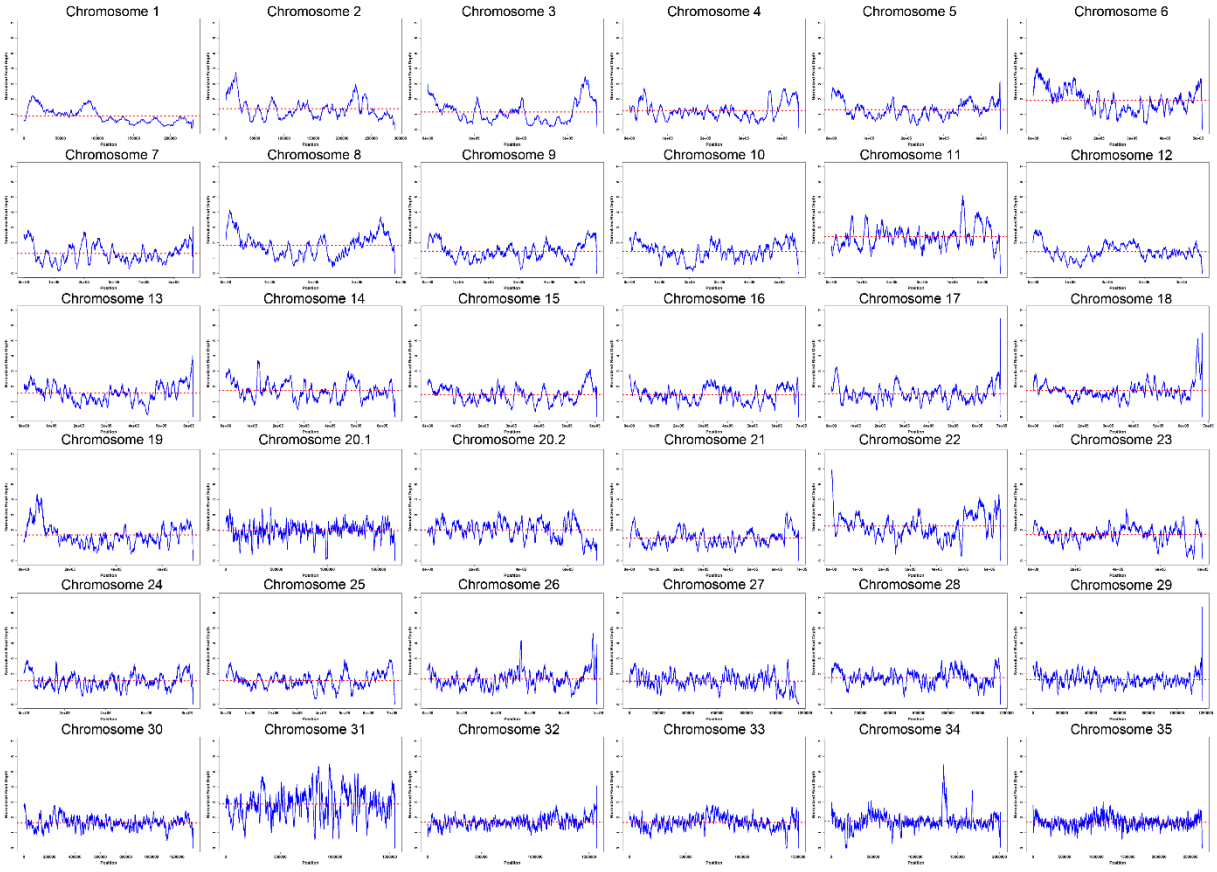
22. Huynh C, Andrews NW. Iron acquisition within host cells and the pathogenicity of Leishmania. Cell Microbiol. 2008;10(2):293–300.
23. Jackson AP. The evolution of amastin surface glycoproteins in trypanosomatid parasites. Mol Biol Evol. 2010;27(1):33–45.
24. Victoir K, Dujardin JC, de Doncker S, Barker DC, Arevalo J, Hamers R, et al. Plasticity of gp63 gene organization in Leishmania (Viannia) braziliensis and Leishmania (Viannia) peruviana. Parasitology. 1995;111(Pt 3):265–73.
25. Olivier M, Atayde VD, Isnard A, Hassani K, Shio MT. Leishmania virulence factors: focus on the metalloprotease GP63. Microbes and Infection/Institut Pasteur. 2012;14(15):1377–89.
26. Sterkers Y, Lachaud L, Crobu L, Bastien P, Pages M. FISH analysis reveals aneuploidy and continual generation of chromosomal mosaicism in Leishmania major. Cell Microbiol. 2011;13(2):274–83.
27. Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, Stitt M, et al. RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Res. 2012;40(Web Server issue):W622–7.
28. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18(5):821–9.
29. Tsai IJ, Otto TD, Berriman M. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. Genome Biol. 2010;11(4):R41.
30. Otto TD, Sanders M, Berriman M, Newbold C. Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. Bioinformatics. 2010;26(14):1704–7.
31. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.
32. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
33. Real F, Vidal RO, Carazzolle MF, Mondego JM, Costa GG, Herai RH, et al. The genome sequence of Leishmania (Leishmania) amazonensis: functional annotation and extended analysis of gene models. DNA Res. 2013;20(6):567–81.
34. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 2011;27(4):578–9.
35. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics. 2009;25(15):1968–9.
36. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303.
37. Cingolani P, Platts A, le Wang L, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly. 2012;6(2):80–92.
38. Motulsky HJ, Brown RE. Detecting outliers when fitting data with nonlinear regression - a new method based on robust nonlinear regression and the false discovery rate. BMC Bioinformatics. 2006;7:123.
39. Sievers F, Higgins DG. Clustal Omega, accurate alignment of very large numbers of sequences. Methods Mol Biol. 2014;1079:105–16.
40. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics. 2005;21(16):3448–9.
41. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7(3):562–78.
42. Li L, Stoeckert Jr CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13(9):2178–89.



Supplementary figure 1: Top five high SNP density genes in *L. (V.) peruviana*. Arrow ruler shows gene length in kilobases. High SNP density regions are shown in light blue and polymorhisms are shown as red vertical lines lines locted in their respective position.

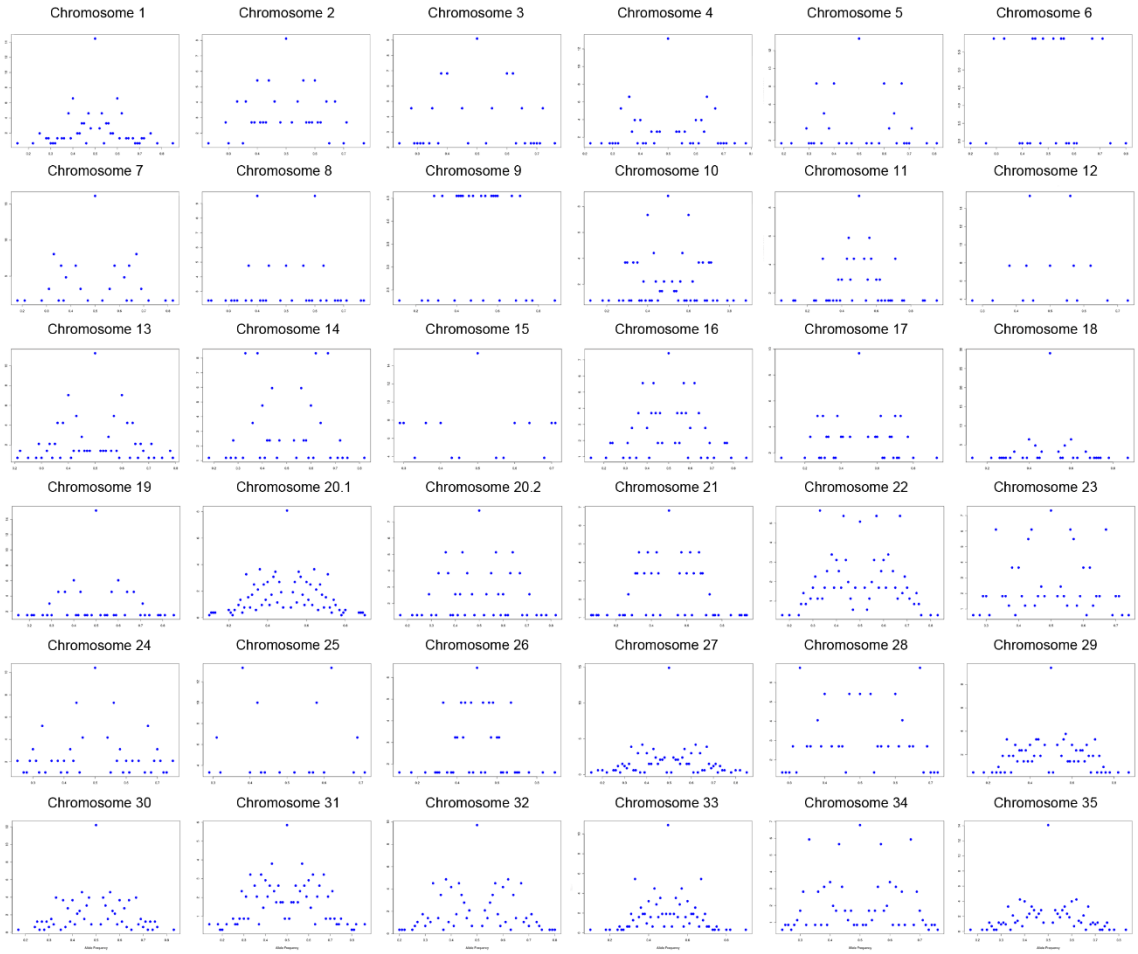


Supplementary figure 2: Normalized read depth for PAB-4377. The Y axis represents chromosome copy number and the X axis shows chromosome position. Estimated copy number for each chromosome is indicated by a dotted red line. Blue lines represent normalized read depth for each position using a slinding window of 10kb



Supplementary figure 3: Normalized read depth for LEM-1537. The Y axis represents chromosome copy number and the X axis shows chromosome position. Estimated copy number for each chromosome is indicated by a dotted red line. Blue lines represent normalized read depth for each position using a slinding window of 10kb

Supplementary figure 4: Distribution of normalized allele frequency counts for *L. (V.) peruviana* PAB-4377. Y axis represents normalized allele frequency counts percentage and X axis shows allele frequencies. Blue dots are located at the respective count for each heterozygous position.



Supplementary figure 5: Distribution of normalized allele frequency counts for *L. (V.) peruviana* LEM-1537. Y axis represents normalized allele frequency counts percentage and X axis shows allele frequencies. Blue dots are located at the respective count for each heterozygous position.

SNPs analysis in L. (V.) peruviana. Only shared SNPs between both isolates are reported. Number of SNPs and gene length are presented in nucleotides. Haploid copy number (HCN) estimated for the genome of PAB-4377 (PAB) and LEM-1537 (LEM). Of these, 14,244 SNPs (53.24%) were synonymous mutations and 12,462 (46.59%) were non-synonymous mutations. Additionally, eight SNPs mutating the annotated start codon (0.03%) and 36 mutating the annotated stop codon were found (0.13%). Only the 30 top genes out of the 6,114 are shown.

		HCN			
Gene Id	Product Description	PAB4377	LEM1537	Number of SNPs	Gene length
LbrM.33.3060	hypothetical protein, conserved	0.97	0.87	135	14943
LbrM.30.2340	hypothetical protein, conserved	0.98	0.75	83	11340
LbrM.34.5330	hypothetical protein, conserved	1.07	1.25	82	19875
LbrM.16.0180	hypothetical protein, conserved	0.82	0.79	69	13302
LbrM.35.1580	hypothetical protein, conserved	1.01	0.76	68	16767
LbrM.14.0770	hypothetical protein, conserved	0.68	0.39	63	12570
LbrM.35.3160	hypothetical protein, conserved	1.05	0.98	43	12582
LbrM.30.2160	endosomal trafficking protein RME-8, putative	1.2	1.19	40	7335
LbrM.02.0130	phosphatidylinositol kinase related protein, putative	0.51	0.47	39	14775
LbrM.30.1620	protein kinase, putative	1.3	1.23	38	5112
LbrM.33.3190	hypothetical protein, conserved	0.55	0.4	38	8253
LbrM.25.1000	hypothetical protein, conserved	0.56	0.33	37	19518
LbrM.16.1320	hypothetical protein, conserved	0.55	0.37	36	6750
LbrM.30.1440	hypothetical protein, conserved	0.58	0.51	36	6534
LbrM.08.0390	hypothetical protein, conserved	0.36	0.37	35	8442
LbrM.20.5140	protein kinase, putative	0.93	0.86	35	10629
LbrM.33.2620	hypothetical protein, unknown function	1.28	1.19	35	3492
LbrM.27.2140	calpain-like cysteine peptidase	0.9	0.8	34	16749
LbrM.31.1440	hypothetical protein, conserved	0.98	0.71	34	14886
LbrM.13.1230	hypothetical protein, conserved	0.36	0.21	33	16791
LbrM.19.1240	hypothetical protein, conserved	0.68	0.36	33	17184
LbrM.27.1700	diacylglycerol acyltransferase, putative	1.46	1.21	33	4887
LbrM.30.2290	zinc-finger protein, conserved	0.88	0.68	33	3534
LbrM.35.1630	hypothetical protein, conserved	1	0.7	33	3939
LbrM.33.3370	hypothetical protein, unknown function	0.7	0.69	32	6765
LbrM.20.3790	Cytoplasmic dynein 2 heavy chain (DYNC2H1), putative	1.05	0.82	31	12729
LbrM.27.0600	calpain-like cysteine peptidase	0.71	0.75	31	14919
LbrM.29.1940	hypothetical protein, conserved	1	0.8	31	5886
LbrM.31.1610	hypothetical protein, unknown function	1.44	1	31	3330
LbrM.31.2350	hypothetical protein, unknown function	1.42	1.36	31	3144

Indel analysis in L. (V.) peruviana. Only shared Indels between both isolates are reported. Gene length is presented in nucleotides. HCN estimated for the haploid genome of PAB-4377 (PAB) and LEM-1537 (LEM). A total of 1,014 (67.0%) variants were deletions, 146 (9.6%) insertions, 351 (23.19%) frameshifts and two stop codons (0.1%) were gained. Only the 30 top genes out of the 408 are shown.

		HCN			
Gene Id	Product Description	PAB4377	LEM1537	Number of nucleotides affected	Gene length
LbrM.17.0390	hypothetical protein, conserved	0.63	0.52	57	3480
LbrM.21.1080	hypothetical protein, conserved	0.76	0.56	42	2895
LbrM.34.2710	hypothetical protein, conserved	1.43	1.6	24	2133
LbrM.31.1470	hypothetical protein, conserved	0.82	0.66	21	4089
LbrM.32.3450	hypothetical protein, conserved	0.74	0.54	21	2469
LbrM.33.2950	hypothetical protein, conserved	0.91	0.77	21	3582
LbrM.07.1050	RNA binding protein-like protein	1.02	1.21	19	1377
LbrM.25.1000	hypothetical protein, conserved	0.56	0.33	18	19518
LbrM.34.4910	hypothetical protein, conserved	0.8	1.17	18	627
LbrM.29.0400	hypothetical protein, conserved	1.05	1.08	15	2127
LbrM.31.1230	hypothetical protein, conserved	0.71	0.43	15	2706
LbrM.32.2900	L-Lysine transport protein, putative (AAT16)	1.73	1.79	15	1416
LbrM.04.0920	hypothetical protein	0.57	0.21	12	1992
LbrM.05.0890	CYC2-like cyclin, putative	0.62	0.46	12	4248
LbrM.10.0340	hypothetical protein, conserved	0.76	0.71	12	6258
LbrM.13.0830	hypothetical protein, unknown function	0.76	0.69	12	3525
LbrM.17.0220	hypothetical protein, conserved	0.3	0.27	12	3750
LbrM.19.0680	protein kinase, putative	0.84	0.59	12	5604
LbrM.21.0610	dihydrolipoamide acetyltransferase precursorlike protein	0.78	1.13	12	789
LbrM.21.1230	hypothetical protein, unknown function	0.82	0.59	12	6261
LbrM.26.1200	hypothetical protein, conserved	0.54	0.42	12	2046
LbrM.26.1320	hypothetical protein, conserved	0.66	0.6	12	3570
LbrM.30.1620	protein kinase, putative	1.3	1.23	12	5112
LbrM.31.1060	hypothetical protein, conserved	1.06	0.86	12	5115
LbrM.32.2540	hypothetical protein, conserved	1.31	1.15	12	3684
LbrM.12.0370	hypothetical protein, unknown function	0.67	0.53	11	6204
LbrM.30.1560	p1/s1 nuclease	1.62	2.01	11	987
LbrM.06.0950	hypothetical protein, conserved	0.49	0.33	10	4965
LbrM.12.0200	hypothetical protein, conserved	0.57	0.43	10	2742
LbrM.22.0105	hypothetical protein	0.69	1.62	10	1470

Gene regions presenting high SNPs density. Conserved SNPs in PAB-4377 and LEM-1537 were selected using a sliding window of 1000 nucleotides in order to report the region with the highest SNPs density. Only the 30 top genes out of 270 are shown.

Gene Id	Gene length	SNPs Count	SNPs positions in gene
LbrM.33.3060	14943	23	8752 8755 8778 8849 8862 8884 9011 9047 9059 9139 9179 9182 9232 9275 9294 9379 9480 9557 9695 9704 9717 9718 9747
LbrM.26.1480	1524	21	438 455 574 685 713 773 810 821 864 919 930 941 975 1000 1008 1041 1077 1083 1187 1418 1435
LbrM.10.0670	3135	19	2002 2005 2022 2056 2063 2126 2127 2190 2209 2227 2230 2231 2303 2307 2349 2404 2503 2680 2942
LbrM.33.0170	1344	17	339 407 421 599 616 626 673 778 846 925 993 998 1012 1032 1061 1105 1128
LbrM.33.2550	4278	17	2046 2075 2084 2281 2310 2329 2409 2551 2568 2587 2589 2643 2694 2697 2928 2957 3036
LbrM.11.1070	1335	16	306 315 318 417 549 574 623 702 830 913 921 937 1164 1191 1279 1302
LbrM.33.1090	1743	16	371 380 654 672 725 726 768 804 832 912 1087 1215 1220 1264 1299 1356
LbrM.35.1610	3513	16	1265 1330 1332 1390 1573 1592 1667 1708 1714 1766 1854 1879 1947 2072 2074 2215
LbrM.10.1420	942	15	26 34 129 195 197 281 320 324 325 468 511 681 830 847 859
LbrM.16.0160	1047	15	97 292 333 434 448 472 560 573 604 625 867 871 950 987 1015
LbrM.31.1200	3453	15	614 772 783 785 898 919 1121 1261 1299 1373 1414 1452 1484 1526 1614
LbrM.34.2730	1947	15	128 265 419 605 682 799 836 839 847 868 890 892 949 1102 1125
LbrM.06.0030	1998	14	308 327 377 378 381 384 396 686 993 996 1158 1174 1244 1267
LbrM.27.1700	4887	14	2215 2265 2444 2487 2555 2558 2582 2599 2605 2883 2903 2936 3063 3165
LbrM.29.0470	8751	14	3850 3900 4069 4099 4214 4281 4300 4527 4531 4542 4591 4729 4776 4844
LbrM.30.1620	5112	14	3373 3385 3388 3491 3545 3590 3653 3660 4003 4036 4063 4124 4186 4353
LbrM.30.2290	3534	14	69 72 109 135 249 265 302 382 396 398 426 534 658 986
LbrM.33.3190	8253	14	6200 6297 6303 6330 6346 6351 6423 6456 6954 6985 6992 7005 7026 7200
LbrM.34.4830	3045	14	1459 1617 1620 1752 1791 1895 1940 1960 1977 2201 2290 2364 2382 2411
LbrM.34.4970	1485	14	207 254 345 346 374 399 475 520 613 615 997 1021 1121 1152
LbrM.06.0350	1140	13	174 273 291 330 348 353 363 367 477 495 706 864 1074
LbrM.14.0770	12570	13	7774 8035 8066 8122 8247 8272 8300 8328 8357 8653 8662 8710 8737
LbrM.16.1320	6750	13	2806 2852 2906 2976 3303 3305 3323 3371 3372 3445 3588 3594 3612
LbrM.20.1530	3504	13	251 254 459 481 621 750 834 872 903 957 959 1045 1138
LbrM.27.1560	1929	13	604 670 716 729 739 941 1025 1078 1149 1200 1252 1319 1566
LbrM.30.2340	11340	13	3670 3681 3687 3879 4063 4064 4144 4176 4298 4311 4325 4344 4402
LbrM.31.0540	1017	13	31 85 93 216 222 401 546 591 604 645 712 819 849
LbrM.31.1610	3330	13	818 892 1078 1095 1189 1208 1214 1450 1478 1515 1583 1613 1795
LbrM.31.2350	3144	13	711 779 886 1039 1048 1092 1158 1242 1267 1334 1403 1625 1709
LbrM.31.2750	1122	13	124 135 343 483 504 588 805 932 948 975 1069 1114 1119

Chapter 1: Supplementary Table 4

Gene Ontology analysis for chromosome 31. Statistical significant results were reported for 2 iron, 2 sulfur cluster binding and other related molecular functions. P-value after Benjamini and Hochberg false discovery rate correction.

GO-Id	p-value	GO Description	Gene Id
51537	1.08E-03	2 iron, 2 sulfur cluster binding	LbrM.31.2820;LbrM.31.2830;LbrM.31.2790; LbrM.31.2840;LbrM.31.2780;LbrM.31.2770
9055	1.85E-02	electron carrier activity	LbrM.31.2820;LbrM.31.2830;LbrM.31.2790; LbrM.31.0550;LbrM.31.2840;LbrM.31.2780; LbrM.31.2850;LbrM.31.2770
4198	1.85E-02	calcium-dependent cysteine-type endopeptidase activity	LbrM.31.0590;LbrM.31.0620;LbrM.31.0600; LbrM.31.0520;LbrM.31.0580;LbrM.31.0510
51536	1.85E-02	iron-sulfur cluster binding	LbrM.31.2820;LbrM.31.2830;LbrM.31.2790; LbrM.31.2840;LbrM.31.2780;LbrM.31.2850; LbrM.31.2770
51540	1.85E-02	metal cluster binding	LbrM.31.2820;LbrM.31.2830;LbrM.31.2790; LbrM.31.2840;LbrM.31.2780;LbrM.31.2850; LbrM.31.2770
4148	1.85E-02	dihydrolipoyl dehydrogenase activity	LbrM.31.2980;LbrM.31.2990;LbrM.31.0230
4197	3.81E-02	cysteine-type endopeptidase activity	LbrM.31.0590;LbrM.31.0620;LbrM.31.0600; LbrM.31.0520;LbrM.31.0580;LbrM.31.0510
8234	4.94E-02	cysteine-type peptidase activity	LbrM.31.0590;LbrM.31.0620;LbrM.31.0600; LbrM.31.0520;LbrM.31.0580;LbrM.31.0510

Chapter 1: Supplementary Table 5

Expanded tandem gene arrays in *L. (V.) peruviana* PAB-4377 and LEM-1537. HCN represents the mean HCN of each gene in the array.

***L. (V.) peruviana* PAB-4377**

OrthoMCL Id	Chromosome	Product Description	HCN	Genes in Array
OG5_126601	LbrM.01	long-chain-fatty-acid-CoA ligase	3	2
OG5_132061	LbrM.03	TATE DNA transposons	10	2
OG5_129349	LbrM.03	hypothetical protein, conserved	4	2
OG5_127165	LbrM.05	ATPase alpha subunit	3	2
OG5_126659	LbrM.06	60S ribosomal protein L23a, putative	2	2
OG5_130995	LbrM.07	3-hydroxyacyl-ACP dehydratase, putative	2	2
OG5_143904	LbrM.08	amastin-like surface protein, putative	4	3
OG5_127795	LbrM.09	hypothetical protein, conserved	2	3
OG5_129998	LbrM.10	zinc binding dehydrogenase-like protein	2	2
OG5_126605	LbrM.13	alpha tubulin	5	2
OG5_138263	LbrM.14	calpain-like cysteine peptidase,	2	2
OG5_142220	LbrM.15	ribonucleoprotein p18, mitochondrial precursor	2	2
OG5_128521	LbrM.21	DNA polymerase eta, putative	3	2
OG5_127374	LbrM.21	xanthine phosphoribosyltransferase (XRPT)	2	2
OG5_130211	LbrM.22	NADH-cytochrome b5 reductase, putative	2	2
OG5_140483	LbrM.24	hypothetical protein, conserved	2	2

OG5_126711	LbrM.26	glutathione peroxidase-like protein, putative	4	2
OG5_126623	LbrM.33	heat shock protein 83 (HSP83-2)	6	3
OG5_126611	LbrM.33	beta-tubulin	3	2
OG5_128620	LbrM.34	NADH-dependent fumarate reductase-like protein	8	3

***L. (V.) peruviana* LEM-1537**

OrthoMCL Id	Chromosome	Product Description	HCN	Genes in Array
OG5_126601	LbrM.01	long-chain-fatty-acid-CoA ligase, putative	3	2
OG5_132061	LbrM.03	TATE DNA transposons	23	2
OG5_129349	LbrM.03	hypothetical protein, conserved	5	2
OG5_127165	LbrM.05	ATPase alpha subunit	5	2
OG5_126659	LbrM.06	60S ribosomal protein L23a, putative	4	2
OG5_132061	LbrM.07	hypothetical protein	3	2
OG5_130995	LbrM.07	3-hydroxyacyl-ACP dehydratase, putative	5	2
OG5_140911	LbrM.09	hypothetical protein	2	2
OG5_127795	LbrM.09	hypothetical protein	3	3
OG5_129998	LbrM.10	hypothetical protein	2	2
OG5_145879	LbrM.10	zinc binding dehydrogenase-like protein	2	2
OG5_139854	LbrM.14	calpain-like cysteine peptidase	3	2
OG5_138263	LbrM.14	hypothetical protein	4	2
OG5_142220	LbrM.15	ribonucleoprotein p18	2	2
OG5_127365	LbrM.16	cytochrome c, putative	2	2
OG5_126617	LbrM.17	receptor-type adenylate cyclase b (RAC-B2)	2	5
OG5_127703	LbrM.17	zinc-finger protein ZPR1, putative	2	2
OG5_143079	LbrM.17	hypothetical protein	3	2
OG5_143904	LbrM.20.1	amastin-like surface protein	6	2
OG5_143904	LbrM.20.1	amastin-like surface protein	4	3
OG5_127374	LbrM.21	xanthine phosphoribosyltransferase (XRPT)	3	2
OG5_126922	LbrM.25	eukaryotic initiation factor 5a, putative (EIF5A2)	3	4
OG5_126611	LbrM.33	beta-tubulin	3	3
OG5_126623	LbrM.33	heat shock protein 83	6	3
OG5_127220	LbrM.34	mitochondrial phosphate transporter	2	2
OG5_128620	LbrM.34	NADH-dependent fumarate reductase-like protein	5	3

Expanded tandem gene arrays in *L. (V.) peruviana* and *L.(V.) braziliensis.* HCN represents the mean haploid copy number of each gene in the array.

L. (V.) peruviana

OrthoMCL Id	Chromosome	Product Description	HCN	Genes in Array
OG5_128620	LbrM.03	TATE DNA transposon	6.5	2
OG5_126623	LbrM.34	NADH-dependent fumarate reductase-like putative	6.2	3
OG5_132061	LbrM.33	heat shock protein 83	6	3
OG5_126611	LbrM.03	hypothetical protein	5.3	2
OG5_138263	LbrM.05	ATPase alpha subunit	4	2
OG5_129349	LbrM.07	3-hydroxyacyl-ACP dehydratase, putative	3.1	2
OG5_127165	LbrM.14	calpain-like cysteine peptidase, putative	3.1	2
OG5_126659	LbrM.06	60S ribosomal protein L23a, putative	3	2
OG5_130995	LbrM.01	long-chain-fatty-acid-CoA ligase, putative	2.9	3
OG5_127795	LbrM.33	beta-tubulin	2.9	3
OG5_127374	LbrM.09	hypothetical protein, conserved	2.7	3
OG5_142220	LbrM.21	xanthine phosphoribosyltransferase (XRPT)	2.3	2
OG5_129998	LbrM.15	ribonucleoprotein p18,	2.2	2
OG5_126601	LbrM.10	zinc binding dehydrogenase-like protein	2	2

L. (V.) braziliensis

OrthoMCL Id	Chromosome	Product Description	HCN	Genes in Array
OG5_126605	LbrM.13	alpha tubulin	8.6	2
OG5_143904	LbrM.08	amastin-like surface protein, putative	6.4	3
OG5_143904	LbrM.20.1	amastin-like surface protein, putative	5.6	3
OG5_126623	LbrM.33	heat shock protein 83-1 (HSP83-1)	4.4	3
OG5_143904	LbrM.20.1	amastin-like surface protein, putative	4.4	2
OG5_126588	LbrM.28	heat-shock protein hsp70, putative	3.8	2
OG5_128620	LbrM.34	NADH-dependent fumarate reductase, putative	3.7	3
OG5_126611	LbrM.33	beta-tubulin	3.6	2
NO_GROUP	LbrM.08	amastin-like protein	3.0	2
OG5_144952	LbrM.15	hypothetical protein	2.9	2
OG5_137181	LbrM.19	ATG8/AUT7/APG8/PAZ2, putative (ATG8B.1)	2.9	2
OG5_148241	LbrM.04	hypothetical protein, conserved in leishmania	2.9	4
NO_GROUP	LbrM.33	beta-tubulin	2.8	3
NO_GROUP	LbrM.04	hypothetical protein, conserved in leishmania	2.4	2
OG5_130385	LbrM.16	paraflagellar rod protein 2C	2.3	2
OG5_130729	LbrM.20.1	amastin-like surface protein, putative	2.2	3
OG5_157998	LbrM.32	3-hydroxyisobutyryl-coenzyme a	2.1	2
OG5_127518	LbrM.16	hypothetical protein	2.0	2

Expanded genes in *L. (V.) peruviana* PAB-4377. Only the top 30 out of 398 genes are shown.

Orthomcl ID	Chromosome	Gene Id	Product Description	HCN
OG5_128620	LbrM.34	LbrM.34.0820	NADH-dependent fumarate reductase-like protein , NADH-dependent fumarate reductase, putative	23.02
OG5_166669	LbrM.01	LbrM.01.0390	hypothetical protein, conserved	9.56
OG5_128334	LbrM.01	LbrM.01.0360	poly(A) export protein, putative	9.45
OG5_127067	LbrM.26	LbrM.26.1590	thimet oligopeptidase, putative,metallo-peptidase, Clan MA(E), Family M3	9.26
OG5_127295	LbrM.01	LbrM.01.0140	monothiol glutaredoxin, putative	9.22
OG5_126601	LbrM.01	LbrM.01.0520	long-chain-fatty-acid-CoA ligase, putative , fatty acyl CoA syntetase 1, putative	8.98
OG5_126613	LbrM.01	LbrM.01.0300	thioredoxin, putative	8.85
OG5_183236	LbrM.08	LbrM.08.0650	tuzin, putative	8.24
OG5_145016	LbrM.01	LbrM.01.0420	hypothetical protein, conserved	8.12
OG5_148121	LbrM.01	LbrM.01.0640	hypothetical protein, conserved	8.10
OG5_128371	LbrM.01	LbrM.01.0380	hypothetical protein, conserved	7.86
OG5_129543	LbrM.01	LbrM.01.0340	acidocalcisomal exopolyphosphatase, putative	7.11
OG5_126837	LbrM.01	LbrM.01.0210	CLC-type chloride channel, putative	7.10
OG5_148228	LbrM.01	LbrM.01.0320	hypothetical protein, conserved	6.69
OG5_126626	LbrM.01	LbrM.01.0080	carboxylase, putative	6.51
OG5_183100	LbrM.01	LbrM.01.0040	hypothetical protein, unknown function	5.98
OG5_127200	LbrM.01	LbrM.01.0060	MCAK-like kinesin, putative	5.78
OG5_183101	LbrM.01	LbrM.01.0070	hypothetical protein, unknown function	5.67
OG5_126967	LbrM.02	LbrM.02.0070	cytochrome b-domain protein, putative	5.48
OG5_183104	LbrM.01	LbrM.01.0350	hypothetical protein, conserved	5.16
OG5_129784	LbrM.01	LbrM.01.0310	pseudouridylate synthase-like protein	4.97
OG5_154526	LbrM.03	LbrM.03.0870	hypothetical protein, conserved	4.89
OG5_135379	LbrM.03	LbrM.03.0860	hypothetical protein, conserved	4.69
OG5_126657	LbrM.03	LbrM.03.0030	hypothetical protein	4.67
OG5_145811	LbrM.01	LbrM.01.0620	mitochondrial RNA editing ligase 1,RNA-editing complex protein,RNA editing ligase	4.53
OG5_154519	LbrM.03	LbrM.03.0660	hypothetical protein, conserved	4.37
OG5_148238	LbrM.02	LbrM.02.0540	protein kinase, putative	4.32
OG5_148223	LbrM.01	LbrM.01.0050	hypothetical protein, conserved	4.32
OG5_145841	LbrM.01	LbrM.01.0160	hypothetical protein, conserved	4.22
OG5_142669	LbrM.01	LbrM.01.0770	calcium/potassium channel (CAKC), putative	4.13

Expanded genes in *L. (V.) peruviana* LEM-1537. HCN shows the haploid copy number for the gene. Only the top 30 out of 942 genes are shown.

Orthomcl ID	Chromosome	Gene Id	Product Description	HCN
OG5_126613	LbrM.01	LbrM.01.0300	thioredoxin, putative	29.55
OG5_179440	LbrM.04	LbrM.04.1050	acyltransferase-like protein, copy 1	24.72
OG5_148121	LbrM.01	LbrM.01.0640	hypothetical protein, conserved	22.14
OG5_145016	LbrM.01	LbrM.01.0420	hypothetical protein, conserved	19.39
OG5_128334	LbrM.01	LbrM.01.0360	poly(A) export protein, putative	19.24
OG5_148228	LbrM.01	LbrM.01.0320	hypothetical protein, conserved	17.07
OG5_166669	LbrM.01	LbrM.01.0390	hypothetical protein, conserved	16.96
OG5_127295	LbrM.01	LbrM.01.0140	monothiol glutaredoxin, putative	14.97
OG5_128620	LbrM.34	LbrM.34.0820	NADH-dependent fumarate reductase-like protein , NADH-dependent fumarate reductase, putative	14.22
OG5_154519	LbrM.03	LbrM.03.0660	hypothetical protein, conserved	13.41
OG5_129543	LbrM.01	LbrM.01.0340	acidocalcisomal exopolyphosphatase, putative	12.29
OG5_151265	LbrM.03	LbrM.03.0880	hypothetical protein, conserved	12.27
OG5_183100	LbrM.01	LbrM.01.0040	hypothetical protein, unknown function	11.81
OG5_154526	LbrM.03	LbrM.03.0870	hypothetical protein, conserved	11.65
OG5_126837	LbrM.01	LbrM.01.0210	CLC-type chloride channel, putative	10.30
OG5_128371	LbrM.01	LbrM.01.0380	hypothetical protein, conserved	9.79
OG5_144559	LbrM.04	LbrM.04.0120	hypothetical protein, conserved	9.33
OG5_135379	LbrM.03	LbrM.03.0860	hypothetical protein, conserved	9.14
OG5_126967	LbrM.02	LbrM.02.0070	cytochrome b-domain protein, putative	8.80
OG5_166673	LbrM.03	LbrM.03.0040	hypothetical protein, conserved	8.75
OG5_183104	LbrM.01	LbrM.01.0350	hypothetical protein, conserved	8.43
OG5_151443	LbrM.05	LbrM.05.0240	hypothetical protein, conserved	8.30
OG5_127006	LbrM.03	LbrM.03.0050	D-3-phosphoglycerate dehydrogenase-like protein	8.13
OG5_128104	LbrM.01	LbrM.01.0030	DNA-damage inducible protein DDI1-like protein	7.89
OG5_154505	LbrM.02	LbrM.02.0300	hypothetical protein, conserved	7.87
OG5_128482	LbrM.03	LbrM.03.0830	hypothetical protein, conserved	7.86
OG5_143192	LbrM.06	LbrM.06.0570	deoxyuridine triphosphatase, putative,dUTP diphosphatase	7.50
OG5_145846	LbrM.04	LbrM.04.0650	hypothetical protein, conserved	7.38
OG5_126626	LbrM.01	LbrM.01.0080	carboxylase, putative	7.32
OG5_151438	LbrM.04	LbrM.04.1220	hypothetical protein, conserved	7.15

Single copy genes estimated by OrthoMCL. These genes were selected based in the presence of orthologs across all analyzed *Leishmania* species and the absence of paralogs in any of these species. Only 30 out of 6,899 ortholog groups are shown.

OrthoMCL Id	*L. (V.) braziliensis*	*L. (L.) mexicana*	*L. (L.) donovani*	*L. (L.) infantum*	*L. (L.) major*	*L. (S.) tarentolae*
LEISH1433:	LbrM.16.0860	LmxM.16.0850	LdBPK_160850.1	LinJ.16.0850	LmjF.16.0850	LtaP16.0830
LEISH1434:	LbrM.29.1760	LmxM.08_29.1650	LdBPK_291790.1	LinJ.29.1790	LmjF.29.1650	LtaP29.1810
LEISH1435:	LbrM.31.0130	LmxM.30.0130	LdBPK_310140.1	LinJ.31.0140	LmjF.31.0130	LtaP31.0140
LEISH1436:	LbrM.29.2130	LmxM.08_29.2150	LdBPK_292260.1	LinJ.29.2260	LmjF.29.2150	LtaP29.2270
LEISH1437:	LbrM.16.0260	LmxM.16.0250	LdBPK_160260.1	LinJ.16.0260	LmjF.16.0250	LtaP16.0250
LEISH1438:	LbrM.05.0900	LmxM.05.0920	LdBPK_050920.1	LinJ.05.0920	LmjF.05.0920	LtaP05.0990
LEISH1439:	LbrM.16.0720	LmxM.16.0730	LdBPK_160730.1	LinJ.16.0730	LmjF.16.0730	LtaP16.0700
LEISH1440:	LbrM.27.2490	LmxM.27.2305	LdBPK_272230.1	LinJ.27.2230	LmjF.27.2305	LtaP27.2380
LEISH1441:	LbrM.29.2240	LmxM.08_29.2260	LdBPK_292370.1	LinJ.29.2370	LmjF.29.2260	LtaP29.2420
LEISH1442:	LbrM.31.0420	LmxM.30.0310	LdBPK_310330.1	LinJ.31.0330	LmjF.31.0310	LtaP31.0330
LEISH1443:	LbrM.29.1290	LmxM.08_29.1210	LdBPK_291300.1	LinJ.29.1300	LmjF.29.1210	LtaP29.1370
LEISH1444:	LbrM.32.1990	LmxM.31.1810	LdBPK_321900.1	LinJ.32.1900	LmjF.32.1810	LtaP32.1930
LEISH1445:	LbrM.32.1100	LmxM.31.1000	LdBPK_321060.1	LinJ.32.1060	LmjF.32.1000	LtaP32.1090
LEISH1446:	LbrM.30.2600	LmxM.29.2640	LdBPK_302630.1	LinJ.30.2630	LmjF.30.2640	LtaP30.2620
LEISH1447:	LbrM.30.2650	LmxM.29.2690	LdBPK_302690.1	LinJ.30.2690	LmjF.30.2690	LtaP30.2680
LEISH1448:	LbrM.30.2190	LmxM.29.2240	LdBPK_302250.1	LinJ.30.2250	LmjF.30.2240	LtaP30.2270
LEISH1449:	LbrM.30.2560	LmxM.29.2610	LdBPK_302600.1	LinJ.30.2600	LmjF.30.2610	LtaP30.2580
LEISH1450:	LbrM.24.1760	LmxM.24.1700	LdBPK_241770.1	LinJ.24.1770	LmjF.24.1700	LtaP24.1850
LEISH1451:	LbrM.24.1810	LmxM.24.1750	LdBPK_241820.1	LinJ.24.1820	LmjF.24.1750	LtaP24.1900
LEISH1452:	LbrM.34.5200	LmxM.34.5260	LdBPK_355230.1	LinJ.35.5230	LmjF.35.5260	LtaP35.5220
LEISH1453:	LbrM.34.4780	LmxM.34.4820	LdBPK_354880.1	LinJ.35.4880	LmjF.35.4820	LtaP35.4810
LEISH1454:	LbrM.17.0520	LmxM.17.0530	LdBPK_170590.1	LinJ.17.0590	LmjF.17.0530	LtaP17.0560
LEISH1455:	LbrM.02.0440	LmxM.02.0460	LdBPK_020430.1	LinJ.02.0430	LmjF.02.0460	LtaP02.0390
LEISH1456:	LbrM.02.0590	LmxM.02.0610	LdBPK_020580.1	LinJ.02.0580	LmjF.02.0610	LtaP02.0530
LEISH1457:	LbrM.02.0600	LmxM.02.0620	LdBPK_020590.1	LinJ.02.0590	LmjF.02.0620	LtaP02.0540
LEISH1458:	LbrM.03.0590	LmxM.03.0690	LdBPK_030670.1	LinJ.03.0670	LmjF.03.0690	LtaP03.0640
LEISH1459:	LbrM.19.0560	LmxM.19.0250	LdBPK_190240.1	LinJ.19.0240	LmjF.19.0250	LtaP19.0210
LEISH1460:	LbrM.24.0800	LmxM.24.0780	LdBPK_240800.1	LinJ.24.0800	LmjF.24.0780	LtaP24.0860
LEISH1461:	LbrM.20.6010	LmxM.12.1320	LdBPK_120910.1	LinJ.12.0910	LmjF.12.1320	LtaP12.1130
LEISH1462:	LbrM.24.1130	LmxM.24.1120	LdBPK_241140.1	LinJ.24.1140	LmjF.24.1120	LtaP24.1230

CHAPTER II: The *Leishmania* metaphylome: a comprehensive survey of *Leishmania* protein phylogenetic relationships

Justification

As has been already mentioned, the leishmaniasis are characterized by their wide spectrum of clinical manifestations with different lines of evidence suggesting an association with the infecting *Leishmania* species and the host immune response (Murray *et al.*, 2005; David e Craft, 2009; Queiroz *et al.*, 2012). Genomic studies have shown a high conservation in the genomes of *Leishmania*, however there is a lack of studies regarding the organization of genes into families and their distribution across the different species.

In this chapter we employed a phylogenomics approach to analyze the distribution of genes across families in distinct *Leishmania* genomes in order to identify species specific expansions and other features that may be responsible for distinct clinical presentations, virulence and parasitism.

The results of this project were published on BMC genomics: doi:10.1186/s12864-015-2091-2.

Objectives

General Objective

The main objective described in this chapter was to characterize the methaphylome of *Leishmania* and identify features associated with parasite adaptations and disease phenotypes.

Specific Objectives

In order to achieve our proposed main goal, we followed these specific aims:

- Generate the phylomes *of L. (L.) infantum, L. (L.) donovani, L. (L.) major, L. (V.) braziliensis, L. (L.) mexicana* and *L. (S.) tarentolae.*
- Analyze presence and absence of gene families and assess species-specific expansions.
- Detect families important for parasitism in these species.

Valdivia et al. BMC Genomics (2015) 16:887
DOI 10.1186/s12864-015-2091-2

BMC Genomics

RESEARCH ARTICLE Open Access

The *Leishmania* metaphylome: a comprehensive survey of *Leishmania* protein phylogenetic relationships

Hugo O. Valdivia[1,2,3], Larissa L. S. Scholte[4], Guilherme Oliveira[4,5], Toni Gabaldón[6,7,8] and Daniella C. Bartholomeu[1,3*]

Abstract

Background: Leishmaniasis is a neglected parasitic disease with diverse clinical manifestations and a complex epidemiology. It has been shown that its parasite-related traits vary between species and that they modulate infectivity, pathogenicity, and virulence. However, understanding of the species-specific adaptations responsible for these features and their evolutionary background is limited. To improve our knowledge regarding the parasite biology and adaptation mechanisms of different *Leishmania* species, we conducted a proteome-wide phylogenomic analysis to gain insights into *Leishmania* evolution.

Results: The analysis of the reconstructed phylomes (totaling 45,918 phylogenies) allowed us to detect genes that are shared in pathogenic *Leishmania* species, such as calpain-like cysteine peptidases and 3'a2rel-related proteins, or genes that could be associated with visceral or cutaneous development. This analysis also established the phylogenetic relationship of several hypothetical proteins whose roles remain to be characterized. Our findings demonstrated that gene duplication constitutes an important evolutionary force in *Leishmania*, acting on protein families that mediate host-parasite interactions, such as amastins, GP63 metallopeptidases, cathepsin L-like proteases, and our methods permitted a deeper analysis of their phylogenetic relationships.

Conclusions: Our results highlight the importance of proteome wide phylogenetic analyses to detect adaptation and evolutionary processes in different organisms and underscore the need to characterize the role of expanded and species-specific proteins in the context of *Leishmania* evolution by providing a framework for the phylogenetic relationships of *Leishmania* proteins.
Phylogenomic data are publicly available for use through PhylomeDB (http://www.phylomedb.org).

Keywords: Phylogenomics, *Leishmania*, Homology prediction

Background

Leishmaniasis is a group of neglected tropical diseases caused by protozoan parasites belonging to the genus *Leishmania*. The disease is present in 98 countries causing more than 1.5 million cases per year [1, 2] and posing 350 million people at risk of infection [3].

Leishmania belongs to the Trypanosomatidae family that is composed of obligatory parasitic organisms. Members of this family can parasitize insects as their hosts, including monoxenic organisms such as *Crithidia, Leptomonas, Herpetomonas* and *Blastocrithidia*, whereas others can also parasitize vertebrates, such as in the digenetic genera *Trypanosoma* and *Leishmania*, or plants in the genera *Phytomonas* [4].

The *Leishmania* genus presents great phenotypic diversity represented by more than 30 different species, of which at least 20 are pathogenic to humans [5]. Phylogenetic analyses of the genus has further divided it into three subgenera named *Leishmania*, *Viannia* and *Sauroleishmania* [6–8].

The *Leishmania* subgenus is distributed throughout the Old and New Worlds, and it is transmitted by the bite of infected female sand flies of the genus *Phlebotomus* (Old

* Correspondence: daniella@icb.ufmg.br
[1]Laboratório de Imunologia e Genômica de Parasitos, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Av. Presidente Antonio Carlos, 6627 – Pampulha, Belo Horizonte, MG 31270-901, Brazil
[3]Centro de Investigaciones Tecnológicas, Biomédicas y Medioambientales, Lima, Peru
Full list of author information is available at the end of the article

World) and *Lutzomyia* (New World). The *Viannia* subgenus is exclusively found in the New World and is only transmitted by *Lutzomyia* sand flies [6]. In both subgenera, parasites are present as intracellular amastigotes inside phagolysosomes of phagocytes in the vertebrate host or as promastigote forms in the insect vector.

The *Sauroleishmania* subgenus that is present in the Old World is composed of non-human pathogenic *Leishmania* and it is assumed that it infects lizards through ingestion of infected *Sergentomya* sand flies [9]. Parasites of this subgenus are found as extracellular promastigotes or amastigote-like forms infecting monocyte-like cells or erythrocytes [6, 7, 10].

Leishmania parasites cause a wide spectrum of clinical manifestations that are classified into cutaneous (CL), mucosal (ML) and visceral leishmaniasis (VL). Previous studies have shown that clinical manifestation and treatment needs are associated with the infecting *Leishmania* species and the host immune response [11].

CL is primarily caused by *Leishmania (Leishmania) major, L. (Leishmania) mexicana, L. (Viannia) braziliensis* and other species of the Viannia subgenus. ML occurs in approximately 5 % of individuals with previous CL, most of who were infected with *L. (Viannia) braziliensis* [12]. VL is caused by *L. (Leishmania) infantum* and *L. (Leishmania) donovani*, which are included within the *L. donovani* complex [2].

Parasite-related factors modulate infectivity, pathogenicity, and virulence [2]. Promastigote virulence factors mediate invasion during the initial steps of an infection. For instance, lipophosphoglycan affects macrophage and dendritic cell functions and gp63 protects against complement mediated lysis and facilitates invasion [2, 13].

Candidate virulence factors in visceralizing parasites include the A2 gene family. This family has been detected in *L. (Leishmania) infantum, L. (Leishmania) donovani* and, as a non-expressed pseudogene, in the *L. (Leishmania) major* genome. All members of the A2 gene family are highly expressed during the amastigote stage, potentially allowing parasite survival at higher temperatures in visceral organs [14].

Over the last decade, *Leishmania* genome sequencing projects have resulted in the availability of a great amount of molecular data, including the genomes of *L. (Leishmania) major* Friedlin [15], *L. (Leishmania) infantum* JPCM5, *L. (Viannia) braziliensis* M2904 [16], *L. (Leishmania) amazonensis* M2269 [7] and several others draft assemblies that are available to the scientific community [17].

Comparative genomic studies have reported high synteny across *Leishmania* species despite a breach of 36–46 million years divergence between New World and Old World species [18]. Only 200 genes with differential distributions across *L. (Leishmania) major, L. (Leishmania) infantum,* and *L. (Viannia) braziliensis* have been described based on sequence similarity [16].

The identification of homologous genes is a critical step to understand the evolutionary history of an organism. Homologs can be divided into two types: orthologs, which originated through a speciation event from a common ancestor and paralogs, which resulted from a duplication event [19–21]. This classification is critical to understanding the diversification processes because duplication events are often related to a posterior functional divergence [22, 23].

Accurate predictions of homology relationships can be used to infer gene functionality [22], reconstruct species phylogenies, and characterize genomes based on their encoded genes [19]. For these purposes, different methods have been proposed. Most of them rely on sequence similarity between genes where function and homology are assessed from the most similar sequences [22]. These methods are fast; however, they have drawbacks because sequence similarity does not always have a direct relationship to functionality [22].

Phylogenomics, which analyzes genomic information in the context of its evolution, is a promising method for inferring homology relationships [24, 25]. This method establishes homology from an evolutionary perspective rather than relying only on sequence similarity [22]. It has also been previously used to reveal the origin and evolution of phenotypic characteristics and further our knowledge of metabolism, pathogenicity, and adaptation of an organism to its surroundings [24, 26–28].

In the current study, we employed a phylogenomics-based approach to analyze the phylomes of six *Leishmania* species to study their evolution and provide a comprehensive survey of the phylogenetic history of all proteins in *Leishmania*.

Methods

Sequence data

Predicted proteomes from six *Leishmania* species Predicted proteomes from six *Leishmania* species (*L. (Viannia) braziliensis, L. (Leishmania) mexicana, L. (Leishmania) major, L. (Leishmania) infantum, L. (Leishmania) donovani, Leishmania (Sauroleishmania) tarentolae*) and *Trypanosoma brucei* were downloaded from the TritrypDB V5 [17] (Table 1). Prior to the analysis, proteome data were filtered with a customized Perl script to select proteins starting with methionine, lacking internal stop codons, represented by the 20 IUPAC amino acid codes, and longer than 100 amino acids.

Phylome reconstruction

Phylome reconstruction for all species was done following an automated pipeline that was previously described [29] (Fig. 1). Briefly, a local database was created comprising all

Table 1 Proteomes selected for the construction of *Leishmania* phylomes

Species	NCBI ID	Total proteins	Valid proteins		Trees generated	Proteome coverage (%)
			#	%		
L. (Viannia) braziliensis	420245	8357	7942	95.0	7712	97.1
L. (Leishmania) donovani	981087	8033	7736	96.3	7550	97.6
L. (Leishmania) infantum	435258	8238	7974	96.8	7808	97.9
L. (Leishmania) major	347515	8400	8170	97.3	7849	96.1
L. (Leishmania) mexicana	929439	8250	7953	96.4	7796	98.0
L. (Sauroleishmania) tarentolae	5689	8452	7465	88.3	7203	96.5

proteomic data. For each protein sequence (seed), a Smith-Waterman search [30] was performed against the aforementioned database to retrieve highly similar proteins with a continuous alignment length of more than 50 % of the query sequence and e-value ≤ 1e-05.

Sets of similar protein sequences were aligned using MUSCLE v3.8 [31], MAFFT v6.712b [32] and KALIGN v2.04 [33]. Alignments were performed in the forward and reverse directions and combined using M-COFFEE [34]. Gaps were removed from the final alignment using trimAl v1.3 [35] with a consistency and gap score cutoffs of 0.1667 and 0.1, respectively.

Neighbor-joining trees were constructed for each trimmed alignment as implemented in BioNJ [36], and *T. brucei* protein sequences were used as the out-group. The resulting NJ tree was used as input for PhyML v3.0 [37] to create a maximum likelihood tree, allowing branch length optimization using different evolutionary models (JTT, LG, WAG, Blosum62, MtREV, VT and Dayhoff).

The two evolutionary models that better modeled the data were determined according to the Akaike Information Criterion (AIC) [38]. Maximum likelihood trees were derived using the two selected models. In all cases, we used a discrete gamma-distribution model with four rate categories plus invariant positions; the gamma parameter and fraction of invariant positions were estimated from the data. Tree support values were calculated with an approximate likelihood ratio test (aLTR) in PhyML [37].

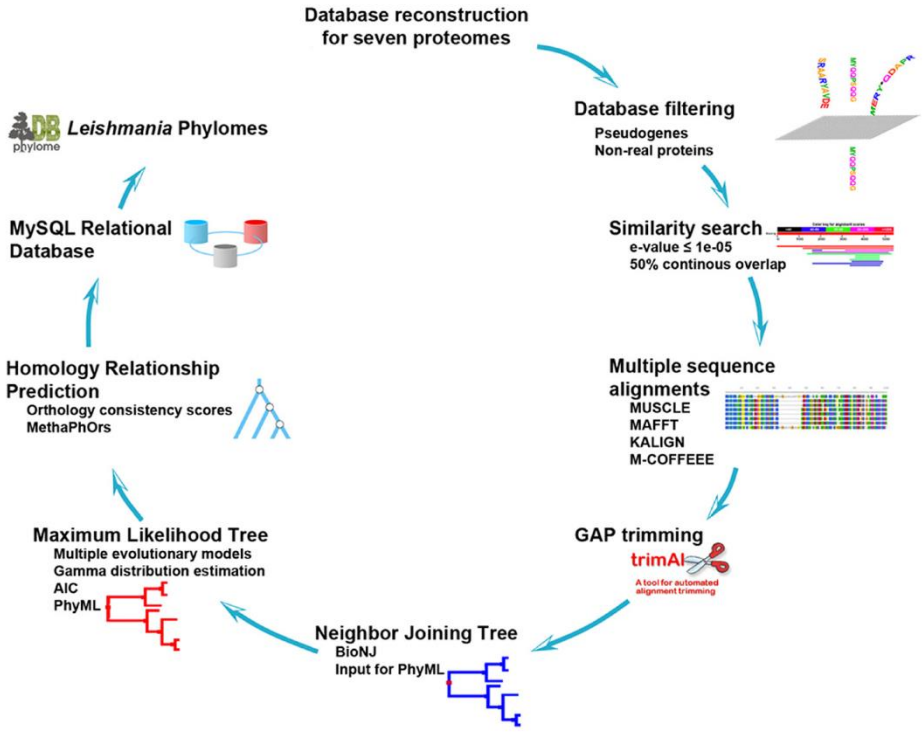


Fig. 1 Phylogenomics pipeline. Each protein was treated as a seed and compared against all proteins encoded in the database. Groups of similar proteins were aligned and trimmed to remove gap-enriched regions. The trimmed alignment was used to build a NJ tree, which was then employed to create a maximum likelihood tree using the two best evolutionary models selected by AIC. Lineage specific duplications and homology relationships were determined and a relational database was created to store and analyze phylomic data

All phylome-related data, including trees and alignments, can be downloaded and browsed through PhylomeDB [39] (www.phylomedb.org).

Detection of homology relationships

To identify orthologs and paralogs, we used a species-overlap algorithm [27] as implemented in the environment for tree exploration (ETE) v2 [40]. Shortly, this algorithm starts at each seed protein used for generating the tree and traverses it until reaching the root. Each internal node was labeled as a duplication or speciation event, depending on whether their daughter partitions showed genes from the same or different species.

Orthology and paralogy relationships derived from the analyses of each phylome were combined into a single prediction using the MetaPhOrs algorithm [41] with a cutoff consistency score of 0.5, meaning that orthology relationship between two genes is called if the majority of examined trees containing these two sequences are consistent with this prediction.

Detection of species-specific expansions

We analyzed the *Leishmania* methaphylome using ETE to identify families that were specifically expanded in each species since their diversification. For this purpose, we considered those duplications detected by the species overlap algorithm that only comprised paralogs as species-specific expansions. An in-house Perl script was subsequently used to filter out redundant paralogous and orthologous proteins and load them into a MySQL relational database.

Gene Ontology codes that were significantly overrepresented in expanded families were detected using the hypergeometric distribution analysis in BiNGO [42] with Benjamini and Hochberg false discovery rate correction (corrected p value <0.05).

Results and discussion

Phylome reconstruction

The *Leishmania* metaphylome was derived from comparative analyses of all proteins encoded by six *Leishmania* species and *Trypanosoma brucei*, which was included as the out-group. The selected set of species includes causal agents of CL (*L. (Viannia) braziliensis*, *L. (Leishmania) mexicana* and *L. (Leishmania) major*), ML (*L. (Viannia) braziliensis*), VL (*L. (Leishmania) infantum* and *L. (Leishmania) donovani*) and a non-human pathogenic *Leishmania* (*L. (Sauroleishmania) tarentolae*).

From an initial set of 49,730 *Leishmania* proteins, 47,240 (94.9 %) were analyzed after filtering for valid sequences resulting in 45,918 phylogenetic trees summarizing the evolutionary relationships of 46,667 proteins (98.8 % of all valid proteins). This coverage is greater than the ones obtained for other phylomes such as the *Schistosoma mansoni* (70 %) [24] or the pea aphid *Acyrthosiphon pisum* (67 %) [26], thereby underscoring the high quality and sequence conservation of the datasets.

The absence of trees for the remaining 573 proteins could be due to high divergence between these proteins and their homologs in the dataset. Alternatively, this set of remaining proteins may include species-specific proteins that did not present homologs due to their uniqueness (Additional file 1: Table S1). Finally, another possibility is the presence of errors in the gene models as has been previously suggested [24].

Species-specific expansions

It has been shown that gene duplication plays an important role in evolution that results in increased expression or novel functionalization and/or sub-functionalization [43, 44]. Duplicated or diversified paralogs may be kept in the genome if they provide a selective advantage [27]. Therefore, inspecting the functions of expanded families may provide evidence of these processes in the evolution of *Leishmania*.

The *Leishmania* metaphylome provides an overview of protein evolutionary relationships that can be explored to reveal events related to *Leishmania* diversification and adaptation. Using the species-overlap algorithm [40], we analyzed species-specific protein expansions in all *Leishmania* proteomes and reported the most expanded proteins for each species (Table 2, Additional file 1: Table S2).

Our results show that species-specific expansions vary greatly between species with *L. (Viannia) braziliensis* and *L. (Leishmania) donovani* accumulating the highest and lowest number of expansions, respectively (Fig. 2). Expanded proteins include well characterized families such as amastins, metalloproteinases, cysteine proteases and surface antigen proteins (Additional file 1: Table S2). These families are important virulence factors in *Leishmania* and reveal an evolutionary trend towards parasitism.

Over-represented Gene Ontology terms in expanded families also show species-specific adaptations (Fig. 2). However, common over-represented terms such as "glycosylation," "proteolysis," "cell adhesion" and "autophagy" are consistent with adaptation towards a parasitic lifestyle.

Glycosylation appears as an important mechanism of protein modification and may play a role in protein maturation and protein function in *Leishmania* [45]. Promastigote and amastigote stages express different types of proteophosphoglycans (PPGs) on their surfaces, and changes in the glycosylation of these proteins have resulted in striking reductions in promastigote and amastigote virulence in *L. (Leishmania) major* [46].

Proteolysis is a key component of pathogenesis in *Leishmania*, acting on several host intracellular proteins

Table 2 Top *Leishmania* species-specific protein expansions using the species-overlap algorithm

Species	Seed	Seed annotation	Expansions
L. (Viannia) braziliensis	LbrM.34.0020	TATE DNA Transposon	30
L. (Viannia) braziliensis	LbrM.08.1140	amastin-like protein	28
L. (Viannia) braziliensis	LbrM.10.0520	GP63, leishmanolysin,metallo-peptidase, Clan MA(M), Family M8	25
L. (Leishmania) major	LmjF.12.0755	surface antigen protein 2, putative	20
L. (Leishmania) mexicana	LmxM.08.0750	amastin-like protein, putative	19
L. (Sauroleishmania) tarentolae	LtaPcontig05711-1	Hypothetical protein, unknown function	14
L. (Sauroleishmania) tarentolae	LtaP10.0670	Major surface protease gp63, putative;GP63, leishmanolysin	12
L. (Leishmania) major	LmjF.34.1720	amastin-like surface protein, putative	12
L. (Leishmania) major	LmjF.12.0950	hypothetical protein, conserved	11
L. (Viannia) braziliensis	LbrM.30.0450	histone H4	10
L. (Leishmania) mexicana	LmxM.08.1080	cathepsin L-like protease, putative	8
L. (Viannia) braziliensis	LbrM.19.1530	glycerol uptake protein, putative	7
L. (Viannia) braziliensis	LbrM.02.0550	Retrotransposable element SLACS	7
L. (Leishmania) mexicana	LmxM.12.0870partial	surface antigen protein 2, putative	7
L. (Leishmania) major	LmjF.09.0156	ATG8/AUT7/APG8/PAZ2, putative (ATG8C.4)	7
L. (Leishmania) major	LmjF.08.1030	cathepsin L-like protease	7
L. (Leishmania) donovani	LdBPK_100380.1	folate/biopterin transporter, putative	5
L. (Leishmania) infantum	LinJ.10.0520	GP63, leishmanolysin,metallo-peptidase, Clan MA(M), Family M8 (GP63-3)	5
L. (Leishmania) infantum	LinJ.36.0010	phosphoglycan beta 1,3 galactosyltransferase 4 (SCG4)	5

such as cytoskeleton regulators, transcription factors or protein phosphatases [47, 48]. It has also been suggested that the direction of proteolytic activities towards degradative enzymes in phagolysosomes and major histocompatibility complex molecules may promote parasite survival by impairing host response and proper antigen presentation [49].

Autophagy has been shown to play an important function during *Leishmania* differentiation from procyclic to metacyclic promastigotes and into amastigotes with an increase in autophagosomes and protein degradation levels [50]. Additionally, degradation of glycosomes allows organelle renewal and enables the parasites to rapidly adapt to the new conditions within their various hosts [51].

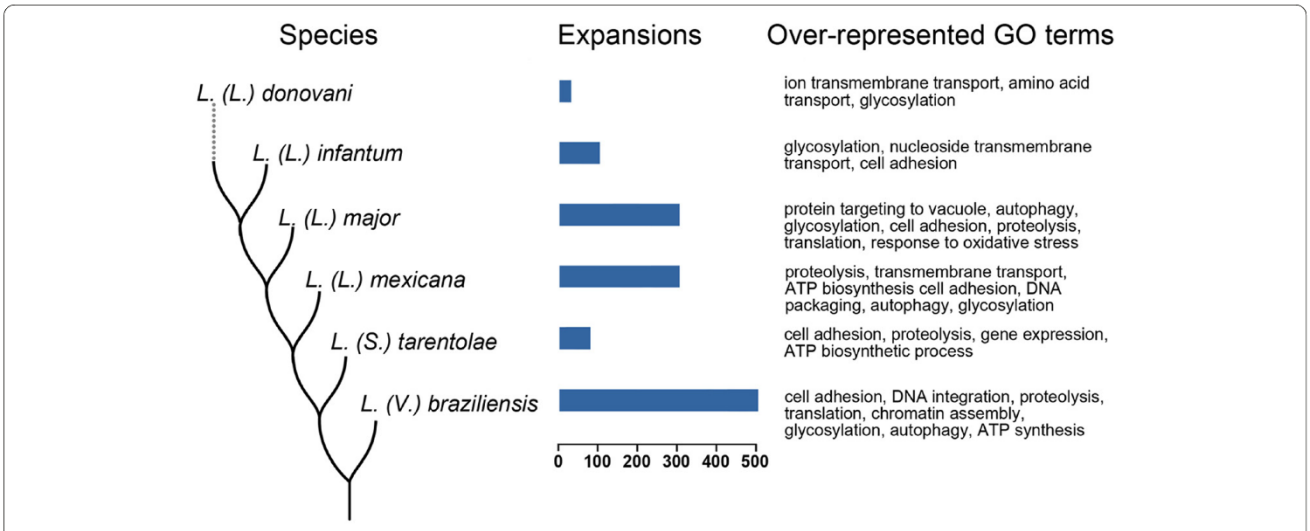


Fig. 2 Estimates of expansions in *Leishmania*. *Horizontal bars* indicate the number of genes in expanded families per species. On the *right* we show significantly over-represented GO terms (corrected *p* value <0.05) compared to the rest of the genome in the set of expanded gene families

Among the most expanded proteins in *L. (Viannia) braziliensis*, we detected the presence of TATE DNA transposons (Telomere-Associated Transposable Element) and SLACS (Spliced Leader Associated Conserved Sequence). SLACS are specific retroposons that are located between tandem arrays of spliced leader RNA genes while TATE transposons tend to be located at telomeres. These transposable elements are the source of most siRNA in *L. (Viannia) braziliensis* [52] that are generated by the RNAi machinery, which appears to be specific to the *Viannia* subgenus to downregulate the expression of mobile elements that can affect genome integrity [52].

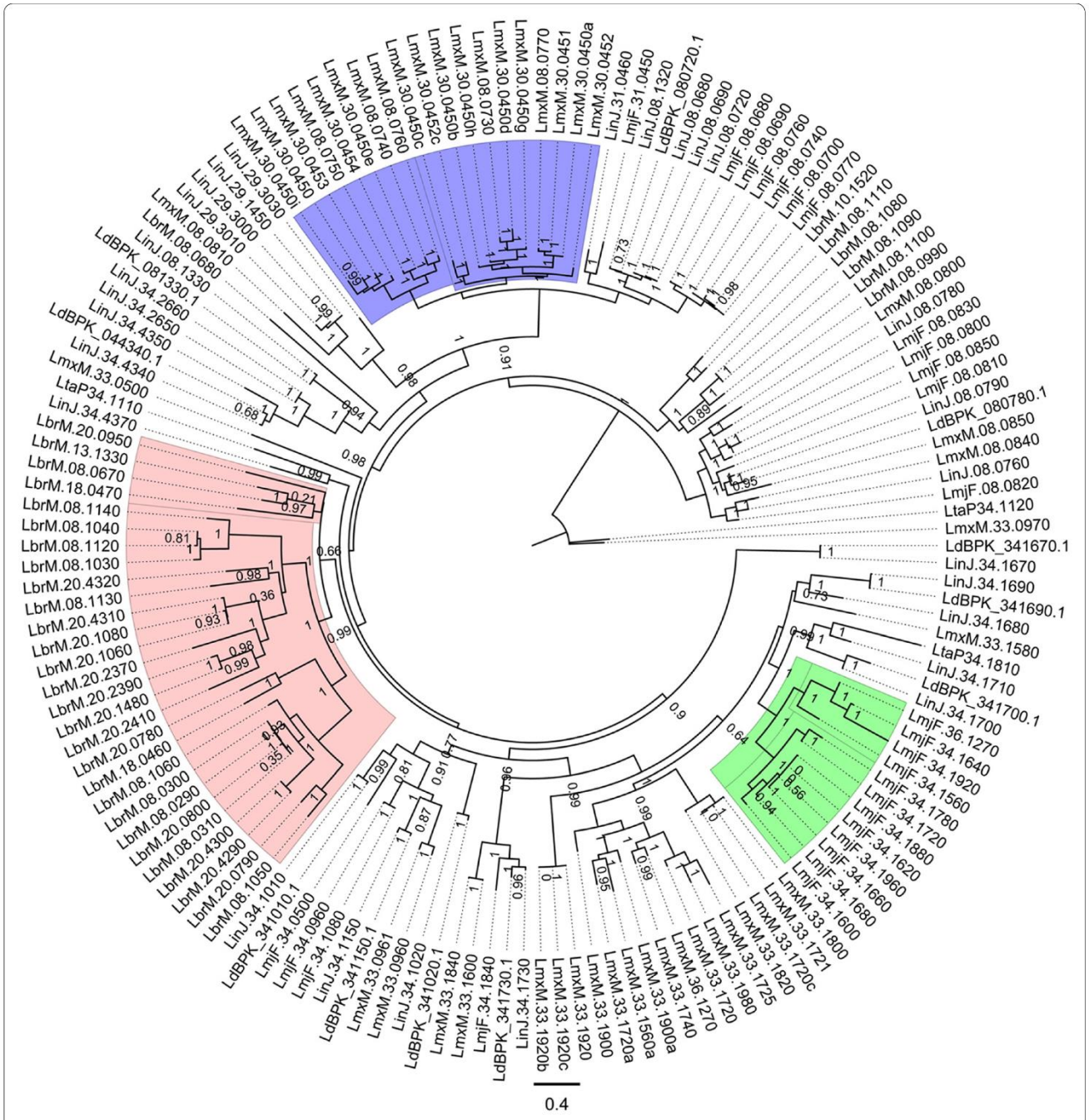


Fig. 3 Amastin phylogenetic tree. Phylogenetic relationships of 150 Amastin protein members using *L. (Viannia) braziliensis* LbrM.08.1140 as seed protein with JTT as the best-fit model. Numbers indicate support values computed by the approximate likelihood ratio test (aLTR). Colored regions show species-specific expansions as follows: *Rose*: *L. (Viannia) braziliensis*; *Green*: *L. (Leishmania) major*; *Blue*: *L. (Leishmania) mexicana*. Gene codes indicate the following species: LinJ: *L. (Leishmania) infantum*; LmxM: *L. (Leishmania) mexicana*; LmjF: *L. (Leishmania) major*; LdBPK: *L. (Leishmania) donovani;* LbrM: *L. (Viannia) braziliensis*; Lta: *L. (Sauroleishmania) tarentolae*

Another expanded protein family in *L. (Viannia) braziliensis* is amastin. This family of surface glycoproteins is highly expressed in amastigotes and, while their exact function is not known, they appear to mediate host-parasite interactions, allowing parasite infection and survival [53]. It has been previously shown that amastins are expanded in all *Leishmania* species compared to *Trypanosoma*, suggesting a functional adaptation [53]. The corresponding amastin phylogeny of our analysis comprises only proteins that originated after *Leishmania* diversification [53] and revealed that *L. (Viannia) braziliensis, L. (Leishmania) mexicana*, and *L. (Leishmania) major* have greatly expanded their delta-amastin repertoire compared to the visceral species included in the phylogenomic analyses (Fig. 3). (For an extensive evolutionary analysis of amastins in Trypanosomatids see Jackson [53]).

We have also detected expansions in the GP63 protein family in *L. (Viannia) braziliensis* and *L. (Sauroleishmania) tarentolae* (Table 2). These metallopeptidases participate in parasite interactions with both invertebrate and vertebrate hosts, ensuring parasite invasion and survival [54–57].

To perform a deeper analysis of the GP63 phylogenomic data, we conducted a case analysis using the tree that retrieved the largest number of homologs across all species (seed "LbrM.31.2240") (Fig. 4). This tree recovered 63 GP63 proteins from an initial set of 74 annotated GP63 plus 11 additional proteins that were annotated as hypothetical proteins in the input dataset. Proteins annotated as GP63 that were not present in this tree are shorter in

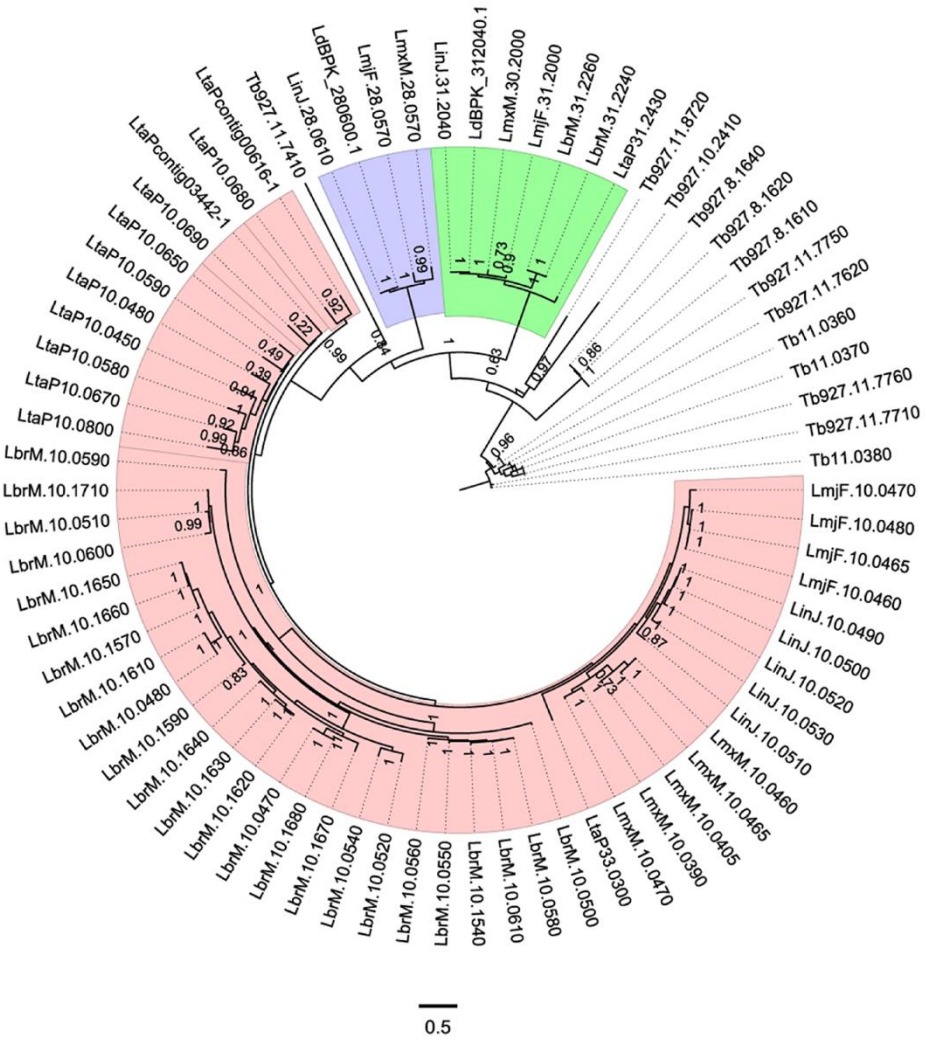


Fig. 4 GP63 phylogenetic tree. Phylogenetic tree for GP63 using *L. (Viannia) braziliensis* seed protein LbrM.31.2240 and WAG as the best-fit model. Numbers indicate support values computed by the approximate likelihood ratio test (aLTR). Colored regions denote GP63 distribution as follows: *Green*: Chromosome 31 GP63; *Blue*: Chromosome 28 GP63; *Rose*: Chromosome 10 GP63. Gene codes indicate the following species: LinJ: *L. (Leishmania) infantum*; LmxM: *L. (Leishmania) mexicana*; LmjF: *L. (Leishmania) major*; LdBPK: *L. (Leishmania) donovani;* LbrM: *L. (Viannia) braziliensis*; Lta: *L. (Sauroleishmania) tarentolae*; Tb: *T. brucei*

length, lack a peptidase domain, or have an incorrect annotation in the proteome dataset.

GP63 genes in the *Leishmania* subgenus range from two genes in *L. (Leishmania) donovani* to seven in *L. (Leishmania) infantum* and GP63. On the contrary, the GP63 repertoire has greatly expanded in *L. (Viannia) braziliensis* and *L. (Sauroleishmania) tarentolae* reaching up to 26 and 13 genes, respectively (Fig. 4).

Our analysis shows that the GP63 family appears to have suffered expansion events at different times during Trypanosomatids' evolution and can be divided in three distinct subfamilies located on chromosomes 31, 28, and 10 (Fig. 5). GP63 of chromosome 31 consists of a single GP63 gene present in all *Leishmania* species except *L. (Viannia) braziliensis*, where it is composed of two distinct isoforms that are located in an array (Figs. 4 and 5).

GP63 of chromosome 28 is present only in the *Leishmania* subgenus and is represented by one gene in *L. (Leishmania) major, L. (Leishmania) mexicana*, *L. (Leishmania) donovani* and *L. (Leishmania) infantum*, sharing more than 93 % similarity at the protein level.

GP63 of chromosome 10 constitutes a set of gene arrays in all *Leishmania* species except *L. (Leishmania) donovani,* where it is completely absent. The phylogeny shows that this subfamily branches with *T. brucei* GP63, supporting a common origin with subsequent gains and losses in *Leishmania* (Fig. 4). Among chromosome 10 GP63s, *L. (Sauroleishmania) tarentolae* and *L. (Viannia) braziliensis* stand out as the species with the highest number of expansions.

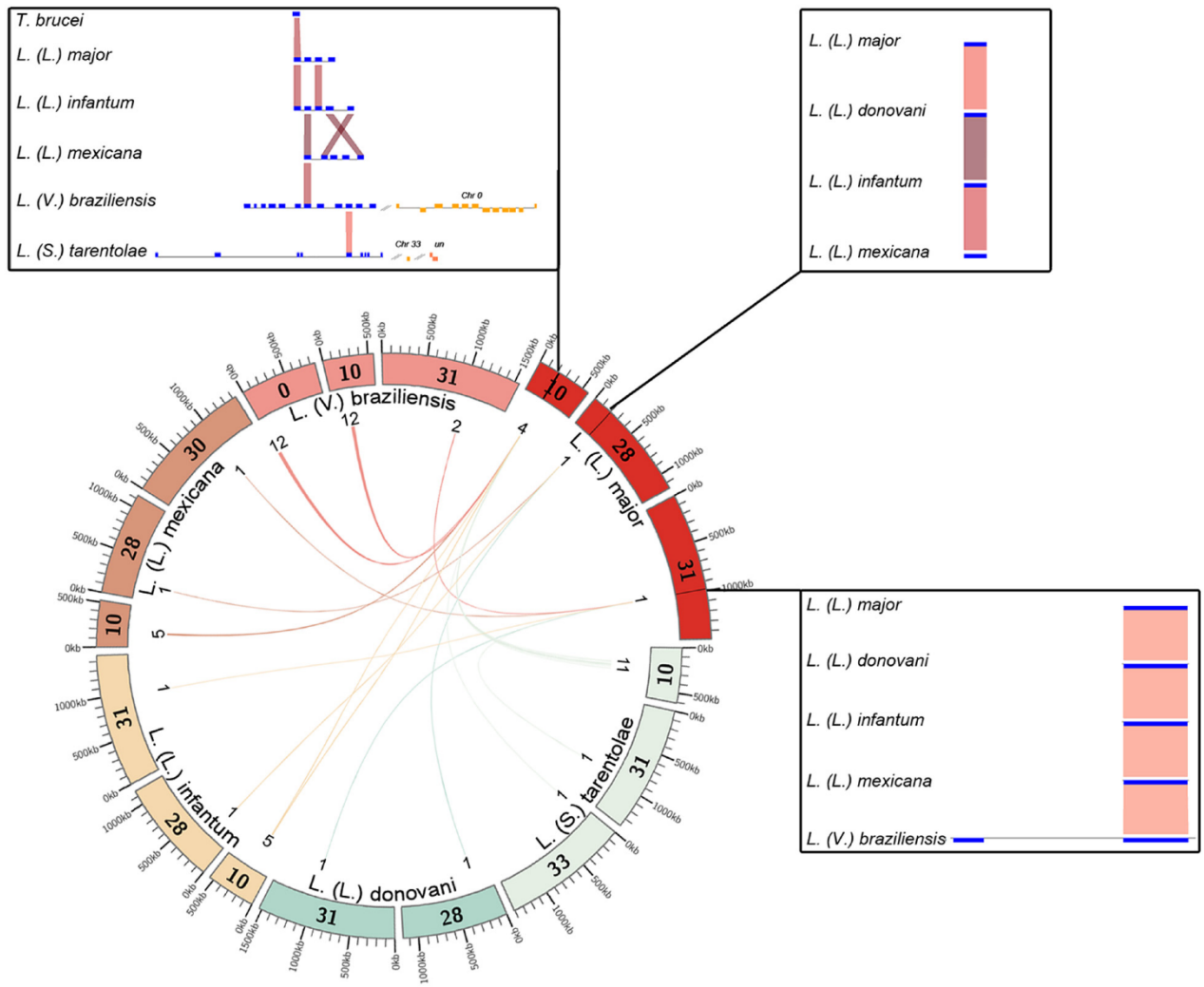


Fig. 5 GP63 chromosome distribution. Location of GP63 genes in *Leishmania* genomes. Boxes represent each GP63 gene per species for each chromosome, while inner links show best reciprocal blast-best hit, and numbers indicate GP63 genes on the respective chromosome; Chr 0: Chromosome 0; un: undetermined. *Inner lines* indicate blast results against the *L. (Leishmania) major* GP63 genes and show the position of each gene within its respective chromosome

Alignment data for the Chr 10 subfamily revealed that *L. (Sauroleishmania) tarentolae* Chr 10 GP63 proteins are shorter than those of *L. (Viannia) braziliensis* (291 versus 560 amino acids), lack predicted extracellular regions, and have a shorter peptidase domain. These characteristics may affect parasite host interaction and limit GP63 protease activity in *L. (Sauroleishmania) tarentolae*, as has been previously suggested [7]. Another possibility could be assembly completeness of the *L. (Sauroleishmania) tarentolae* genome, which may result in partial GP63 sequences [7].

Given that the long arrays in *L. (Viannia) braziliensis* are absent from the other *Leishmania* species, it is highly possible that this expansion occurred after the origin of the *Viannia* subgenus. Interestingly, it has been previously shown that GP63 is also present in high copy number in *L. (Viannia) peruviana* and *L. (Viannia) guyanensis* [58, 59].

This information suggests that large GP63 expansions in chromosome 10 are characteristic of the *Viannia* subgenus and could respond to an adaptation mechanism to the wider range of reservoirs and vectors that the species of this subgenus infect. In the case of *L. (Sauroleishmania) tarentolae*, GP63 expansions could be related to interactions with a different genus that serves as vector (*Sergentomya*) and the lizard host.

Histone 4 has also been shown to be differentially expanded in *L. (Viannia) braziliensis* with 10 genes. In the *Leishmania* subgenus, Histone 4 is reduced to three or less genes and is completely absent in *Sauroleishmania* (Fig. 6). However, the lack of Histone 4 in *Sauroleishmania* could likely result from the limitations in the current genome assembly of this species.

H4 expansions in *L. (Viannia) braziliensis* are not restricted to a single chromosome, suggesting derivation of novel loci through transposition. Sequence alignment of these expansions showed a conserved core with more than 80 % sequence similarity among all sequences and the presence of variable regions at the N and C terminal ends.

Post-translational modification analysis in histones of Trypanosomatids has revealed that H4 and H3 are heavily acetylated and methylated on the N-terminal tails in *Trypanosoma*, and these modifications change during parasite development [60]. Whether expansions and diversification in histone 4 of *L. (Viannia) braziliensis* have a role in transcriptional regulation in *Leishmania* remains to be investigated.

Our results also revealed species-specific expansions in cysteine peptidases (CPs) in *L. (Leishmania) mexicana, L. (Leishmania) major* and *L. (Viannia) braziliensis.* These expansions are located in tandem arrays in chromosome 8 (Fig. 7). Previous studies on Cathepsin-B have shown immunomodulatory roles suppressing the Th1 response, ensuring parasite survival in *L. (Leishmania) mexicana* and *L. (Leishmania) major* and that their activity could result

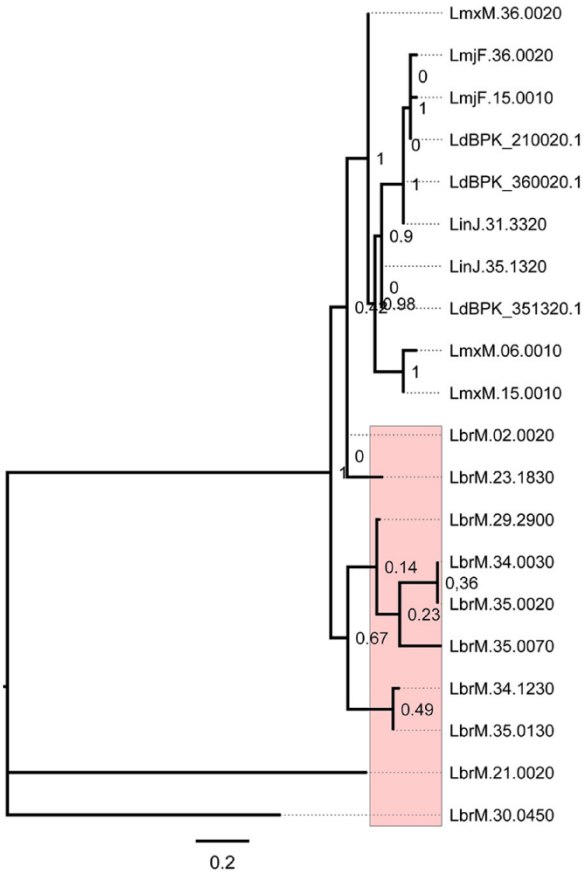


Fig. 6 Histone 4 phylogenetic relationship. Phylogenetic relationships for Histone 4 using *L. (Viannia) braziliensis* seed protein LbrM.30.0450 and JTT as the best-fit model. Numbers indicate support values computed by the approximate likelihood ratio test (aLTR). Due to their high sequence similarity, there is a large inconsistency in most nodes as reflected in the tree. Rose-colored regions indicate *L. (Viannia) braziliensis* expansions. Gene codes indicate the following species: LinJ: *L. (Leishmania) infantum*; LmxM: *L. (Leishmania) mexicana*; LmjF: *L. (Leishmania) major*; LdBPK: *L. (Leishmania) donovani*; LbrM: *L. (Viannia) braziliensis*

in different disease phenotypes in both species [61, 62]. The corresponding phylogeny of cysteine peptidases showed that cathepsin-L genes are exclusively located in chromosome 8, cysteine peptidases A in chromosome 19, and cathepsin-B in chromosome 29.

L. (Leishmania) mexicana, L. (Leishmania) major and *L. (Viannia) braziliensis* present eight, seven and three expansions of Cathepsin-L, respectively (Fig. 7). These expansions are organized into gene arrays and share more than 70 % similarity at the protein level.

RNA-expression data for *L. (Leishmania) major* retrieved from the Trytrip database [17] shows that these Cathepsin-L genes have, on average, a 1.7-fold increase in amastigotes versus procyclic promastigotes and up to a 1.8-fold increase between metacyclic versus procyclic promastigotes, which suggests that Cathepsin-L expression is modulated during parasite development with

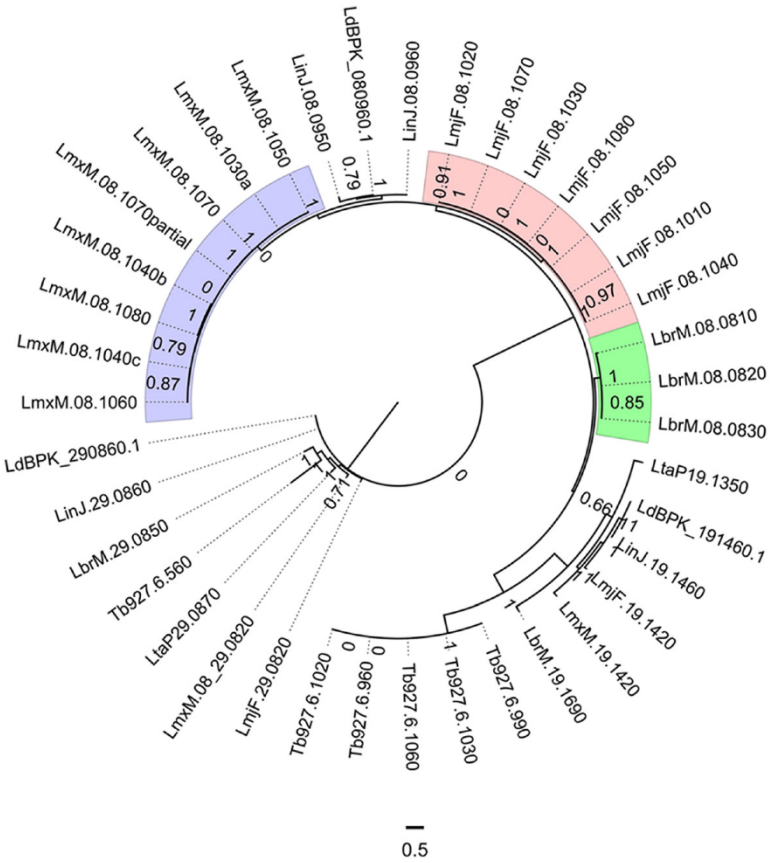


Fig. 7 Cysteine peptidases phylogenetic relationship. Phylogenetic relationships for cysteine peptidases using *L. (Leishmania) major* seed protein LmjF.08.1070 under the WAG model. Numbers indicate support values computed by the approximate likelihood ratio test (aLTR). *Blue, rose* and *green* colored regions represent *L. (Leishmania) mexicana, L. (Leishmania) major* and *L. (Viannia) braziliensis* specific cathpesin-L expansions, respectively. Gene codes indicate the following species: LinJ: *L. (Leishmania) infantum*; LmxM: *L. (Leishmania) mexicana*; LmjF: *L. (Leishmania) major*; LdBPK: *L. (Leishmania) donovani;* LbrM: *L. (Viannia) braziliensis*; Lta: *L. (Sauroleishmania) tarentolae*

expression increasing towards the infective and intracellular stages.

Orthology relationships in *Leishmania*

Using BioPerl:Trees, we extracted orthologs and paralogs for each seed protein to analyze the ones that are unique in each species and to look at their respective homologs.

A total of 28 trees summarizing the relationships of 72 genes were species-unique (Additional file 1: Table S3). From these, 25 trees belonged to *L. (Viannia) braziliensis* and comprised TATE DNA transposons, SLACS, a phosphatidic acid phosphatase, and hypothetical proteins. The remaining trees belonged to a folate biopterin transporter, an oligosaccharyl transferase in *L. (Leishmania) donovani*, and a hypothetical protein in *L (Leishmania) major*. The absence of a greater number of species-specific trees reflects the high conservation between *Leishmania* proteomes and underscores the importance of species-specific expansions. Another possibility is the variance in assembly completeness of *Leishmania* genomes that can limit an accurate assessment of orthology and paralogy relationships.

We found 299 trees comprising 1519 genes across five human pathogenic *Leishmania* species without orthologs in *L. (Sauroleishmania) tarentolae*. Protein families in these trees include histone 4, k39 kinesin, calpain-like cysteine peptidases, a2-rel and hypothetical proteins (Additional file 1: Table S4).

Calpain-like cysteine peptidases are predicted to encode large proteins with potential functions in signal transduction, cytoskeletal remodeling and membrane attachment during *Leishmania* differentiation [63, 64].

Previous studies have shown that disruption by gene targeting of a2-rel-related genes in *L. (Leishmania) donovani* generated mutants with reduced infectivity in mice and limited their proliferation in culture [65]; however, their specific function has not been elucidated yet.

We found a total of 11 trees that were shared by species of the *Leishmania donovani* complex without orthologs in *L. (Leishmania) major, L. (Viannia) braziliensis* nor *L. (Sauroleishmania) tarentolae* (Additional file 1: Table S5). Among these genes we found the presence of the A2 gene family that is the prototype of genes

involved in visceralization [66] and hypothetical proteins that remain to be characterized.

Leishmania species that are associated with CL include *L. (Viannia) braziliensis, L. (Leishmania) mexicana*, *L. (Leishmania) major* and occasionally *L. (Leishmania) infantum* [2]. We found a total of 15 trees specific for all these species comprising of 72 proteins, most of which are annotated as hypothetical (Additional file 1: Table S6).

Conclusions

Our results indicate that gene expansions are a common trait in *Leishmania* genomes and represent an important force in the evolution of these parasites. Major species-specific expansions in genes mediating host-parasite interactions reflect genome complexity and evolutionary processes that influence the wide spectrum of diseases that are caused by different *Leishmania* species.

An important limitation of the current study is the different assembly completeness across the *Leishmania* genomes analyzed. It is known that repetitions and head-to-tail duplicated genes are likely to suffer from assembly and annotation errors leading to partial sequences that could have been excluded during the filtering steps. In this sense, it might be possible that the exact number of expanded genes may vary with subsequent improvements of the current genome assemblies.

The *Leishmania* metaphylome appears as a promising resource to aid the scientific community in understanding the complexity of host-parasite relationships and highlighting areas of interest for additional experimentation. Further studies are needed to determine the function of relevant hypothetical proteins that were identified here, characterize species-specific expansions, and employ transcriptomic data to complement our results.

Additional file

Additional file 1: Supplementary tables. (XLSX 148 kb)

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

HOV carried out bioinformatics analysis, participated in study conception, design and drafted the manuscript. LLSS participated in study design, provided support with phylogenies and manuscript writing. GO participated in study design and manuscript writing. TG carried out bioinformatics analysis, contributed with data storage and publication on Phylome DB, manuscript writing. DCB participated in study design, coordination and manuscript writing. All authors read and approved the final manuscript.

Acknowledgements

We thank Leszek P. Pryszcz for his assistance with MetaPhOrs. DB group is funded by The National Institute of Science and Technology for Vaccines (Brazil) (MCT/CNPq, grant CNPq 573547/2008-4), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG, grant # APQ-04073-10, PPM-00219-13) and Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior (CAPES, grant # 051/2013). TG group research is funded in part by a grant from the Spanish ministry of Economy and Competitiveness (BIO2012-37161), a Grant from the Qatar National Research Fund grant (NPRP 5-298-3-086), and a grant from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC (Grant Agreement n. ERC-2012-StG-310325). GO group was funded by NIH-Fogarty (TW007012), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (REDE-56/11, RED-00014-14) and Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior (309312/2012-4).

Disclaimer

The views expressed in this article are those of the authors only and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, nor the U.S. Government.

Author details

[1]Laboratório de Imunologia e Genômica de Parasitos, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Av. Presidente Antonio Carlos, 6627 – Pampulha, Belo Horizonte, MG 31270-901, Brazil. [2]Department of Parasitology, U.S. Naval Medical Research Unit No. 6, Lima, Peru. [3]Centro de Investigaciones Tecnológicas, Biomédicas y Medioambientales, Lima, Peru. [4]Genomics and Computational Biology Group, Centro de Pesquisas René Rachou, Belo Horizonte, Brazil. [5]Instituto Tecnológico Vale – ITV, Belém, Brazil. [6]Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), Barcelona, Spain. [7]Universitat Pompeu Fabra (UPF), Barcelona, Spain. [8]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.

Received: 22 May 2015 Accepted: 15 October 2015
Published online: 30 October 2015

References

1. Alvar J, Velez ID, Bern C, Herrero M, Desjeux P, Cano J, et al. Leishmaniasis worldwide and global estimates of its incidence. PLoS One. 2012;7(5), e35671.
2. Murray HW, Berman JD, Davies CR, Saravia NG. Advances in leishmaniasis. Lancet. 2005;366(9496):1561–77.
3. Desjeux P. Leishmaniasis: current situation and new perspectives. Comp Immunol Microbiol Infect Dis. 2004;27(5):305–18.
4. Maslov DA, Podlipaev SA, Lukes J. Phylogeny of the kinetoplastida: taxonomic problems and insights into the evolution of parasitism. Mem Inst Oswaldo Cruz. 2001;96(3):397–402.
5. Banuls AL, Hide M, Prugnolle F. *Leishmania* and the leishmaniases: a parasite genetic update and advances in taxonomy, epidemiology and pathogenicity in humans. Adv Parasitol. 2007;64:1–109.
6. Bates PA. Transmission of *Leishmania* metacyclic promastigotes by phlebotomine sand flies. Int J Parasitol. 2007;37(10):1097–106.
7. Real F, Vidal RO, Carazzolle MF, Mondego JM, Costa GG, Herai RH, et al. The genome sequence of *Leishmania (Leishmania) amazonensis*: functional annotation and extended analysis of gene models. DNA Res. 2013;20(6):567–81.
8. Croan DG, Morrison DA, Ellis JT. Evolution of the genus *Leishmania* revealed by comparison of DNA and RNA polymerase gene sequences. Mol Biochem Parasitol. 1997;89(2):149–59.
9. LUMSDEN WHR, Evans D. Biology of the Kinetoplastida. Vol. 2. London: Academic Press Inc. (London) Ltd.; 1979.
10. Paperna I, Boulard Y, Hering-Hagenbeck SH, Landau I. Description and ultrastructure of *Leishmania zuckermani* n. sp. amastigotes detected within the erythrocytes of the South African gecko Pachydactylus turneri Gray, 1864. Parasite. 2001;8(4):349–53.
11. Kaye P, Scott P. Leishmaniasis: complexity at the host-pathogen interface. Nat Rev Microbiol. 2011;9(8):604–15.
12. Queiroz A, Sousa R, Heine C, Cardoso M, Guimaraes LH, Machado PR, et al. Association between an emerging disseminated form of leishmaniasis and

Leishmania (Viannia) braziliensis strain polymorphisms. J Clin Microbiol. 2012;50(12):4028–34.
13. McMahon-Pratt D, Alexander J. Does the *Leishmania major* paradigm of pathogenesis and protection hold for New World cutaneous leishmaniases or the visceral disease? Immunol Rev. 2004;201:206–24.
14. Zhang WW, Matlashewski G. Loss of virulence in *Leishmania donovani* deficient in an amastigote-specific protein, A2. Proc Natl Acad Sci U S A. 1997;94(16):8807–11.
15. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, et al. The genome of the kinetoplastid parasite, *Leishmania major*. Science. 2005;309(5733):436–42.
16. Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail MA, et al. Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. Nat Genet. 2007;39(7):839–47.
17. Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, Carrington M, et al. TriTrypDB: a functional genomic resource for the Trypanosomatidae. Nucleic Acids Res. 2010;38(Database issue):D457–462.
18. Lukes J, Mauricio IL, Schonian G, Dujardin JC, Soteriadou K, Dedet JP, et al. Evolutionary and geographical history of the *Leishmania donovani* complex with a revision of current taxonomy. Proc Natl Acad Sci U S A. 2007;104(22):9375–80.
19. Gabaldon T, Dessimoz C, Huxley-Jones J, Vilella AJ, Sonnhammer EL, Lewis S. Joining forces in the quest for orthologs. Genome Biol. 2009;10(9):403.
20. Fitch WM. Distinguishing homologous from analogous proteins. Syst Zool. 1970;19(2):99–113.
21. Descorps-Declere S, Lemoine F, Sculo Q, Lespinet O, Labedan B. The multiple facets of homology and their use in comparative genomics to study the evolution of genes, genomes, and species. Biochimie. 2008;90(4):595–608.
22. Eisen JA, Wu M. Phylogenetic analysis and gene functional predictions: phylogenomics in action. Theor Popul Biol. 2002;61(4):481–7.
23. Gabaldon T, Koonin EV. Functional and evolutionary implications of gene orthology. Nat Rev Genet. 2013;14(5):360–6.
24. Silva LL, Marcet-Houben M, Nahum LA, Zerlotini A, Gabaldon T, Oliveira G. The *Schistosoma mansoni* phylome: using evolutionary genomics to gain insight into a parasite's biology. BMC Genomics. 2012;13:617.
25. Gabaldon T. Large-scale assignment of orthology: back to phylogenetics? Genome Biol. 2008;9(10):235.
26. Huerta-Cepas J, Marcet-Houben M, Pignatelli M, Moya A, Gabaldon T. The pea aphid phylome: a complete catalogue of evolutionary histories and arthropod orthology and paralogy relationships for *Acyrthosiphon pisum* genes. Insect Mol Biol. 2010;19 Suppl 2:13–21.
27. Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldon T. The human phylome. Genome Biol. 2007;8(6):R109.
28. Jackson AP, Allison HC, Barry JD, Field MC, Hertz-Fowler C, Berriman M. A cell-surface phylome for African trypanosomes. PLoS Negl Trop Dis. 2013;7(3), e2121.
29. Huerta-Cepas J, Capella-Gutierrez S, Pryszcz LP, Denisov I, Kormes D, Marcet-Houben M, et al. PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. Nucleic Acids Res. 2011;39(Database issue):D556–560.
30. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981;147(1):195–7.
31. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004;5:113.
32. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30(14):3059–66.
33. Lassmann T, Sonnhammer EL. Kalign–an accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics. 2005;6:298.
34. Wallace IM, O'Sullivan O, Higgins DG, Notredame C. M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res. 2006;34(6):1692–9.
35. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972–3.
36. Gascuel O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol. 1997;14(7):685–95.
37. Guindon S, Delsuc F, Dufayard JF, Gascuel O. Estimating maximum likelihood phylogenies with PhyML. Methods Mol Biol. 2009;537:113–37.
38. Akaike H. A new look at the statistical model identification. Automatic Control IEEE Trans. 1974;19(6):716–23.
39. Huerta-Cepas J, Capella-Gutierrez S, Pryszcz LP, Marcet-Houben M, Gabaldon T. PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. Nucleic Acids Res. 2014;42(Database issue): D897–902.
40. Huerta-Cepas J, Dopazo J, Gabaldon T. ETE: a python Environment for Tree Exploration. BMC Bioinformatics. 2010;11:24.
41. Pryszcz LP, Huerta-Cepas J, Gabaldon T. MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. Nucleic Acids Res. 2011;39(5), e32.
42. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics. 2005;21(16):3448–9.
43. Vogel C, Chothia C. Protein family expansions and biological complexity. PLoS Comput Biol. 2006;2(5), e48.
44. Roth C, Rastogi S, Arvestad L, Dittmar K, Light S, Ekman D, et al. Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. J Exp Zool B Mol Dev Evol. 2007;308(1):58–73.
45. Rosenzweig D, Smith D, Myler PJ, Olafson RW, Zilberstein D. Post-translational modification of cellular proteins during *Leishmania donovani* differentiation. Proteomics. 2008;8(9):1843–50.
46. Spath GF, Lye LF, Segawa H, Sacks DL, Turco SJ, Beverley SM. Persistence without pathology in phosphoglycan-deficient *Leishmania major*. Science. 2003;301(5637):1241–3.
47. Halle M, Gomez MA, Stuible M, Shimizu H, McMaster WR, Olivier M, et al. The *Leishmania* surface protease GP63 cleaves multiple intracellular proteins and actively participates in p38 mitogen-activated protein kinase inactivation. J Biol Chem. 2009;284(11):6893–908.
48. Contreras I, Gomez MA, Nguyen O, Shio MT, McMaster RW, Olivier M. *Leishmania*-induced inactivation of the macrophage transcription factor AP-1 is mediated by the parasite metalloprotease GP63. PLoS Pathog. 2010;6(10), e1001148.
49. Silverman JM, Chan SK, Robinson DP, Dwyer DM, Nandan D, Foster LJ, et al. Proteomic analysis of the secretome of *Leishmania donovani*. Genome Biol. 2008;9(2):R35.
50. Besteiro S, Williams RA, Coombs GH, Mottram JC. Protein turnover and differentiation in *Leishmania*. Int J Parasitol. 2007;37(10):1063–75.
51. Michels PA, Bringaud F, Herman M, Hannaert V. Metabolic functions of glycosomes in trypanosomatids. Biochim Biophys Acta. 2006;1763(12):1463–77.
52. Atayde VD, Shi H, Franklin JB, Carriero N, Notton T, Lye LF, et al. The structure and repertoire of small interfering RNAs in *Leishmania (Viannia) braziliensis* reveal diversification in the trypanosomatid RNAi pathway. Mol Microbiol. 2013;87(3):580–93.
53. Jackson AP. The evolution of amastin surface glycoproteins in trypanosomatid parasites. Mol Biol Evol. 2010;27(1):33–45.
54. Jecna L, Dostalova A, Wilson R, Seblova V, Chang KP, Bates PA, et al. The role of surface glycoconjugates in *Leishmania* midgut attachment examined by competitive binding assays and experimental development in sand flies. Parasitology. 2013;140(8):1026–32.
55. Gomez MA, Contreras I, Halle M, Tremblay ML, McMaster RW, Olivier M. *Leishmania* GP63 alters host signaling through cleavage-activated protein tyrosine phosphatases. Sci Signal. 2009;2(90):ra58.
56. Corradin S, Ransijn A, Corradin G, Roggero MA, Schmitz AA, Schneider P, et al. MARCKS-related protein (MRP) is a substrate for the *Leishmania* major surface protease leishmanolysin (gp63). J Biol Chem. 1999;274(36):25411–8.
57. Brittingham A, Morrison CJ, McMaster WR, McGwire BS, Chang KP, Mosser DM. Role of the *Leishmania* surface protease gp63 in complement fixation, cell adhesion, and resistance to complement-mediated lysis. J Immunol. 1995;155(6):3102–11.
58. Victoir K, Dujardin JC, de Doncker S, Barker DC, Arevalo J, Hamers R, et al. Plasticity of gp63 gene organization in *Leishmania (Viannia) braziliensis* and *Leishmania (Viannia) peruviana*. Parasitology. 1995;111(Pt 3):265–73.
59. Steinkraus HB, Greer JM, Stephenson DC, Langer PJ. Sequence heterogeneity and polymorphic gene arrangements of the *Leishmania guyanensis* gp63 genes. Mol Biochem Parasitol. 1993;62(2):173–85.
60. Kumar D, Rajanala K, Minocha N, Saha S. Histone H4 lysine 14 acetylation in *Leishmania donovani* is mediated by the MYST-family protein HAT4. Microbiology. 2012;158(Pt 2):328–37.

61. Buxbaum LU, Denise H, Coombs GH, Alexander J, Mottram JC, Scott P. Cysteine protease B of *Leishmania mexicana* inhibits host Th1 responses and protective immunity. J Immunol. 2003;171(7):3711–7.
62. Judice WA, Manfredi MA, Souza GP, Sansevero TM, Almeida PC, Shida CS, et al. Heparin modulates the endopeptidase activity of leishmania mexicana cysteine protease cathepsin L-Like rCPB2.8. PLoS One. 2013;8(11):e80153.
63. Ono Y, Sorimachi H, Suzuki K. Structure and physiology of calpain, an enigmatic protease. Biochem Biophys Res Commun. 1998;245(2):289–94.
64. Mottram JC, Coombs GH, Alexander J. Cysteine peptidases as virulence factors of *Leishmania*. Curr Opin Microbiol. 2004;7(4):375–81.
65. Zhang WW, Matlashewski G. Characterization of the A2-A2rel gene cluster in *Leishmania donovani*: involvement of A2 in visceralization during infection. Mol Microbiol. 2001;39(4):935–48.
66. McCall LI, Zhang WW, Matlashewski G. Determinants for the development of visceral leishmaniasis disease. PLoS Pathog. 2013;9(1), e1003053.

This table contains proteins that were included in the analysis but that did not form part of any tree. It is highly possible that this set include species unique proteins, incomplete sequences and errors in the gene model. Only the first 30 lines of the table are shown (The complete table has 573 lines)

Gene Identifier	Product Description
LbrM.01.0260	long-chain-fatty-acid-CoA ligase, putative
LbrM.02.0710	repeat gene hypothetical protein
LbrM.02.0780	repeat gene hypothetical protein
LbrM.03.0010	hypothetical protein
LbrM.03.0960	hypothetical protein
LbrM.03.0970	hypothetical protein
LbrM.04.0020	hypothetical protein
LbrM.04.0830	hypothetical protein, conserved
LbrM.04.1180	31-O-demethyl-FK506 methyltransferase
LbrM.05.0240	hypothetical protein, conserved
LbrM.05.1110	DNA-directed RNA polymerase I largest subunit
LbrM.06.0420	60S ribosomal protein L19, putative
LbrM.06.1180	hypothetical protein
LbrM.07.0155	hypothetical protein, conserved
LbrM.07.0510	hypothetical protein
LbrM.07.0520	hypothetical protein, conserved
LbrM.08.0200	hypothetical protein, unknown function
LbrM.08.0370	hypothetical protein, conserved
LbrM.08.0400	hypothetical protein, conserved
LbrM.08.0560	hypothetical protein, conserved
LbrM.09.0470	hypothetical protein, conserved
LbrM.10.1030	hypothetical protein
LbrM.10.1720	GP63, leishmanolysin (GP63-2)
LbrM.11.0360	argonaute-like protein (AGO1)
LbrM.11.1100	hypothetical protein
LbrM.11.1130	hypothetical protein
LbrM.12.0320	hypothetical protein, conserved
LbrM.14.0480	hypothetical protein, conserved
LbrM.14.0940	Hypothetical repeat protein
LbrM.14.1690	proteophosphoglycan ppg1

Species-specific protein expansions for each species. These expansions were defined as duplications that only comprised paralogs detected by the species overlap algorithm. Only the first 30 lines of the table are shown (The complete table has 1316 lines).

Seed Identifier	Species	Protein Length	Product Description	Number of expansions
LbrM.34.0020	*L. (V.) braziliensis*	908	TATE DNA Transposon	30
LbrM.25.2610	*L. (V.) braziliensis*	1115	TATE DNA Transposon	30
LbrM.30.3220	*L. (V.) braziliensis*	953	TATE DNA transposons	30
LbrM.32.0040	*L. (V.) braziliensis*	767	TATE DNA Transposon	29
LbrM.25.2620	*L. (V.) braziliensis*	928	TATE DNA Transposon	29
LbrM.08.1140	*L. (V.) braziliensis*	209	amastin-like protein	28
LbrM.20.1060	*L. (V.) braziliensis*	192	amastin-like surface protein, putative	28
LbrM.08.1030	*L. (V.) braziliensis*	210	amastin-like protein	28
LbrM.08.1120	*L. (V.) braziliensis*	207	amastin-like surface protein, putative	28
LbrM.20.2390	*L. (V.) braziliensis*	192	amastin-like surface protein, putative	28
LbrM.13.1330	*L. (V.) braziliensis*	194	amastin	28
LbrM.20.0780	*L. (V.) braziliensis*	211	amastin-like surface protein, putative	28
LbrM.20.4310	*L. (V.) braziliensis*	192	amastin-like surface protein, putative	28
LbrM.20.4300	*L. (V.) braziliensis*	192	amastin-like surface protein, putative	28
LbrM.08.1130	*L. (V.) braziliensis*	212	amastin-like surface protein, putative	28
LbrM.08.0670	*L. (V.) braziliensis*	195	amastin-like protein	28
LbrM.20.2410	*L. (V.) braziliensis*	193	amastin-like surface protein, putative	28
LbrM.30.3200	*L. (V.) braziliensis*	1634	TATE DNA Transposon	27
LbrM.35.0030	*L. (V.) braziliensis*	1639	TATE DNA Transposon	27
LbrM.20.6070	*L. (V.) braziliensis*	1649	TATE DNA Transposon	27
LbrM.20.6140	*L. (V.) braziliensis*	1486	TATE DNA Transposon	27
LbrM.20.6120	*L. (V.) braziliensis*	1604	unspecified product	27
LbrM.05.1240	*L. (V.) braziliensis*	1531	hypothetical protein	27
LbrM.08.0310	*L. (V.) braziliensis*	231	amastin-like surface protein, putative	27
LbrM.31.1860	*L. (V.) braziliensis*	1639	TATE DNA Transposon	27
LbrM.30.3800	*L. (V.) braziliensis*	1336	TATE DNA transposons	27
LbrM.34.2200	*L. (V.) braziliensis*	1639	TATE DNA transposons	27
LbrM.20.6110	*L. (V.) braziliensis*	1722	TATE DNA Transposon	27
LbrM.24.2430	*L. (V.) braziliensis*	1639	TATE DNA Transposon	27

Species-specific trees for each *Leishmania* species. They are defined as trees with nodes belonging to the same species.

Seed ID	Species	Protein Length	Product Description	Proteins in tree
LbrM.25.2610	*L. (V.) braziliensis*	1115	TATE DNA Transposon	30
LbrM.25.2620	*L. (V.) braziliensis*	928	TATE DNA Transposon	29
LbrM.05.1240	*L. (V.) braziliensis*	1531	hypothetical protein	27
LbrM.20.6070	*L. (V.) braziliensis*	1649	TATE DNA Transposon	27
LbrM.20.6140	*L. (V.) braziliensis*	1486	TATE DNA Transposon	27
LbrM.07.0500	*L. (V.) braziliensis*	2052	hypothetical protein	26
LbrM.30.3210	*L. (V.) braziliensis*	1397	TATE DNA transposons	26
LbrM.11.1170	*L. (V.) braziliensis*	2025	TATE DNA transposons	25
LbrM.18.0010	*L. (V.) braziliensis*	629	TATE DNA Transposon	24
LbrM.20.6090	*L. (V.) braziliensis*	1470	TATE DNA Transposon	22
LbrM.29.0010	*L. (V.) braziliensis*	954	TATE DNA Transposon	21
LbrM.02.0550	*L. (V.) braziliensis*	1148	Retrotransposable element SLACS	7
LbrM.02.0750	*L. (V.) braziliensis*	1870	SLACS like gene retrotransposon element	6
LbrM.16.0780	*L. (V.) braziliensis*	977	hypothetical protein	6
LbrM.02.0690	*L. (V.) braziliensis*	526	SLACS like gene retrotransposon element	5
LbrM.08.0700	*L. (V.) braziliensis*	1783	SLACS	5
LbrM.16.1730	*L. (V.) braziliensis*	854	hypothetical protein	5
LbrM.02.0770	*L. (V.) braziliensis*	399	repeat gene hypothetical protein	4
LbrM.05.1230	*L. (V.) braziliensis*	491	hypothetical protein	4
LbrM.22.1370	*L. (V.) braziliensis*	152	hypothetical protein	4
LbrM.02.0680	*L. (V.) braziliensis*	514	hypothetical protein	3
LbrM.02.0700	*L. (V.) braziliensis*	318	hypothetical protein	3
LbrM.03.0020	*L. (V.) braziliensis*	181	hypothetical protein	3
LbrM.13.1570	*L. (V.) braziliensis*	378	hypothetical protein	3
LbrM.18.0500	*L. (V.) braziliensis*	269	phosphatidic acid phosphatase, putative	3
LdBPK_100380.1	*L. (L.) donovani*	168	folate/biopterin transporter, putative	5
LdBPK_351140.1	*L. (L.) donovani*	178	oligosaccharyl transferase-like protein	3
LmjF.12.0995	*L. (L.) major*	558	hypothetical protein	4

Trees shared by pathogenic leishmania species without orthologs in *L. (S.) tarentolae.*
The table show the tree and the number of homologs in:
1) *L. (V.) braziliensis,* 2) *L. (L.) major*, 3) *L. (L.) mexicana,* 4) *L. (L.) donovani,* 5) *L. (L.) infantum*
Only the first 30 lines of the table are shown (The complete table has 299 lines)

Seed Identifier	Length	Product Description	Lbr[1]	Lma[2]	Lme[3]	Ldo[4]	Lin[5]
LmjF.06.0010	100	histone H4	9	3	4	5	3
LmjF.15.0010	100	histone H4	9	3	4	5	3
LmjF.36.0020	100	histone H4	9	3	4	5	3
LmjF.14.1120	2976	kinesin K39, putative	4	4	7	1	4
LmjF.12.0950	482	hypothetical protein, cons	1	11	1	1	3
LmjF.12.1050	439	hypothetical protein, cons	1	11	1	1	3
LmjF.12.0840	314	hypothetical protein, cons	1	11	1	1	3
LmjF.12.0970	584	hypothetical protein, cons	1	11	1	1	3
LmjF.12.0770	482	hypothetical protein, cons	1	11	1	1	3
LmjF.12.0895	482	hypothetical protein, cons	1	11	1	1	3
LmjF.12.0820	482	hypothetical protein, cons	1	11	1	1	3
LmjF.12.1030	435	hypothetical protein, cons	1	11	1	1	3
LmjF.12.0715	482	hypothetical protein, cons	1	11	1	1	3
LmjF.12.0800	477	hypothetical protein, cons	1	11	1	1	3
LmjF.12.0930	482	hypothetical protein, cons	1	11	1	1	3
LmjF.36.4840	1207	hypothetical protein, cons	2	1	2	2	2
LmjF.29.0250	319	thymine-7-hydroxylase, putative	1	1	2	1	1
LmjF.33.0410	939	DNA mismatch repair protein, putative	3	2	2	2	2
LmjF.31.2520	847	hypothetical protein, cons	2	3	2	2	3
LmjF.31.2530	825	hypothetical protein, cons	2	3	2	2	3
LmjF.31.2540	845	hypothetical protein, cons	2	3	2	2	3
LmjF.27.2050	388	ribonucleoside-diphosphate reductase small chain, putative	2	2	2	1	2
LmjF.19.0570	548	hypothetical protein, cons	1	2	2	1	2
LmjF.19.0540	530	hypothetical protein, cons	1	2	2	1	2
LmjF.32.2950	151	nucleoside diphosphate kinase b	3	2	2	2	2
LmjF.27.0490	4553	calpain-like cysteine peptidase, putative	3	3	2	1	3
LmjF.32.3490	684	DEAD/DEAH box helicase, putative	2	2	2	2	2
LmjF.27.0510	5358	calpain-like cysteine peptidase, putative	3	3	1	1	3
LmjF.28.0780	455	RNA polymerase I	1	2	2	2	2
LmjF.20.0375	1399	hypothetical protein, cons	2	2	2	2	2

Chapter 2: Supplementary Table 5

Trees with orthologs across visceral *Leishmania* species
Total homologous proteins in:
1) *L. (L.) mexicana*
2) *L. (L.) donovani*
3) *L. (L.) infantum*

Seed Identifier	Protein Length	Product Description	Lme[1]	Ldo[2]	Lin[3]
LinJ.08.0140	585	hypothetical protein, conserved	1	1	1
LinJ.15.0900	461	nucleotide sugar transporter, putative	1	1	1
LinJ.01.0720	521	hypothetical protein, unknown function	0	1	2
LinJ.02.0720	433	hypothetical protein	1	0	2
LinJ.13.1450	327	alpha tubulin	1	0	2
LinJ.14.0020	1047	phosphatidylinositol 3-kinase 2, putative	1	1	1
LinJ.22.0670	487	A2 protein	2	0	1
LinJ.24.1510	3372	multi drug resistance protein-like	1	1	1
LinJ.28.3170	514	hypothetical protein, unknown function	1	1	1
LinJ.31.2550	497	hypothetical protein, conserved	1	1	1
LinJ.34.2330	259	hypothetical protein	1	1	1

Chapter 2: Supplementary Table 6

Trees with orthologs across cutaneous *Leishmania* species
Total homologous proteins in:
1) *L. (V.) braziliensis*
2) *L. (L.) major*
3) *L. (L.) mexicana*
5) *L. (L.) infantum*

Seed Identifier	Protein Length	Product Description	Lbr[1]	Lma[2]	Lme[3]	Lin[4]
LbrM.24.0470	658	hypothetical predicted transmembrane protein	3	3	3	1
LbrM.20.5340	341	calpain-like cysteine peptidase	2	2	2	2
LbrM.02.0170	368	RNA helicase, putative	2	3	2	0
LbrM.31.1840	487	thiolase protein-like protein	2	2	2	1
LbrM.25.2450	874	hypothetical protein, conserved	2	1	1	1
LbrM.33.1340	290	hypothetical protein, conserved	2	1	1	1
LbrM.13.0660	793	hypothetical protein, conserved	1	1	1	1
LbrM.15.0490	718	hypothetical protein, conserved	1	1	1	1
LbrM.22.0970	814	hypothetical protein, conserved	1	1	1	1
LbrM.24.0430	1331	hypothetical protein, unknown function	1	1	1	1
LbrM.31.1000	114	hypothetical protein, conserved	1	1	1	1
LbrM.32.3440	2412	hypothetical protein, conserved	1	1	1	1
LbrM.33.1350	470	hypothetical protein, conserved	1	1	1	1
LbrM.16.1120	2341	hypothetical protein, unknown function	1	1	1	0
LbrM.28.3150	140	hypothetical protein, unknown function	1	1	1	0

CHAPTER III: Comparative genomic analysis of *Leishmania* from an endemic focus in Governador Valadares

Background

The leishmaniasis are an important health problem in Brazil that presents both forms of disease (TL and VL). Data from geoestatistical modeling suggest that more than 30,189 TL and 4.889 VL cases occurred in 2010 across 3,895 and 4,889 municipalities (Karagiannis-Voules *et al.*, 2013). The highest number of TL cases occurred in the state of Pará (4,332), while Minas Gerais was the state with the most number of VL cases (693)(Karagiannis-Voules *et al.*, 2013).

In Minas Gerais, the municipality of Governador Valadares in the southeastern Brazilian state of Minas Gerais is an endemic area for TL and a focus of intense transmission of VL with more than 127 reported cases from 2007 to 2013 and up to 30% of domestic dogs positive by serology (Barata *et al.*, 2013). In this focus, human VL cases affect mainly males and children from 0-9 years with a lethality rate of 16% (Barata *et al.*, 2013).

In this chapter, we focused on the study of canine leishmaniasis in Governador Valadares using genomic data to study the dynamics of disease transmission and reveal the parasite population structure on this city.

Objectives

General Objective

The main objective of this chapter was to characterize the *Leishmania* population structure and dynamics in Governador Valadares.

Specific Objectives

In order to accomplish our main objective, we worked on the following specific aims:

- Analyze variants across isolates collected on this focus.
- Assess the parasite population structure.
- Determine the effective population size and evolutionary history of the parasites in this foci.
- Determine chromosome copy number variations.

- Characterize tandem duplicated gene arrays and dispersed duplicated genes.

Due to the results obtained during the execution of the project described in this chapter, we will divide it into two sections. The first section in the form of a manuscript that will be probably submitted by the time of the thesis defense describing the first isolation and a comparative genomic analysis of *L. (L.) amazonensis* in Governador Valadares and the second section of a population genomics study on *L. (L.) infantum* isolates from this focus.

Comparative genomics of canine isolated *Leishmania (Leishmania) amazonensis* from an endemic focus of visceral leishmaniasis in Governador Valadares, southeastern Brazil

Hugo O. Valdivia[1]; Laila Almeida[1]; Bruno M. Roatt[2]; Ricardo Toshio Fujiwara[1]; Celia Gontijo[2]; Alexandre B. Reis[3]; James Cotton[4]; Mathew Berriman[4]; Daniella C. Bartholomeu[1§]

[1] Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. Laboratório de Imunologia e Genômica de Parasitos, Instituto de Ciências Biológicas

[2] Universidade Federal de Ouro Preto

[3] Wellcome Trust Sanger Institute

[§]Corresponding author

Email addresses:

HOV: hvalrod@hotmail.com

LA: lailavalmeida@gmail.com

BMR: bmroatt@gmail.com

RTF: fujiwara@icb.ufmg.br

CG: gontijo@cpqrr.fiocruz.br

ABR: alexreisufop@gmail.com

JC: jc17@sanger.ac.uk

MB: mb4@sanger.ac.uk

DCB: daniella@icb.ufmg.br

Abstract

The leishmaniasis is a highly diverse group of diseases caused by kinetoplastid parasites of the *Leishmania* genus. Human pathogenic species are separated into the *Leishmania* and the *Viannia* subgenera according to their development site inside the alimentary tract of the sand fly. The clinical features of these diseases are highly diverse and can result in tegumentary or visceral compromise. These characteristics are the result of a complex interaction between the infecting parasite species and the host immune response.

Among the causative species in the Brazil, *Leishmania (Leishmania) amazonensis* is an important etiological agent of human cutaneous leishmaniasis accounting for more than 8% of all cases in endemic regions.

Here, we report the first isolation of *Leishmania (L.) amazonensis* from dogs with clinical manifestation of visceral leishmaniasis in Governador Valadares, an endemic focus in the southeastern Brazilian State of Minas Gerais. These isolates were characterized in terms of SNPs, chromosome and gene copy number variations. The results presented in this article will contribute to increase our knowledge about *Leishmania (L.) amazonensis* specific adaptations to infection and parasite survival and the transmission of this Amazonian species in a new endemic area of Brazil.

Introduction

The leishmaniasis encompasses a group of diverse clinical diseases caused by protozoan parasites of the genus *Leishmania*. These diseases are endemic in 98 countries posing a risk to 350 million people and causing 1.5 new million cases per year[1,2].

Leishmania are digenetic organisms with a phase of their lifecycle in an invertebrate host from the genus *Lutomzyia* in the New World or Phlebotomus in the Old World and a stage inside a mammalian host. To cope with the interplay between the invertebrate and mammalian hosts, *Leishmania* parasites present two different developmental stages: a motile flagellated extracellular promastigote form that develops within the digestive tract of the insect vector and a non-motile intracellular amastigote form that infects macrophages in the vertebrate host[3].

The *Leishmania* genus is comprised of up to 35 different species, of which at least 20 are pathogenic to humans[4]. These different species have been classified into two distinct subgenera (*Leishmania* and *Viannia*) according to their development site inside the alimentary tract of the sand fly[5]. Species from the *Viannia* subgenus present a phase of development at the hindgut and posterior migration to the midgut, whereas species from the *Leishmania* subgenus undergo intraluminal development in the midgut and foregut[5].

Leishmaniasis is known by producing a broad spectrum of clinical manifestations that are primarily associated to the infecting *Leishmania* species[2]. These distinct disease features have been classified into tegumentary (TL) and visceral leishmaniasis (VL).

Among the species associated to TL in Brazil, *L. (L.) amazonensis* accounts for more than 8% of all cases in the north and northeastern regions[6,7] and is considered the main etiological agent of diffuse cutaneous leishmaniasis (DCL) that is a type of TL characterized by the appearance of multiple non-ulcerative lesions[2].

Leishmaniasis is considered as a re-emergent and emergent disease with geographical expansion due to urbanization, human migration, man driven environmental modifications

and co-infection with other diseases[8]. This expansion has led to the emergence of new focuses of transmission and reactivation of previously controlled settings[9-11].

The municipality of Governador Valadares in the southeastern Brazilian state of Minas Gerais is an endemic area for TL and a focus of intense transmission of VL with more than 127 reported cases from 2007 to 2013 and up to 30% of domestic dogs positive by serology[9]. In this focus, human VL cases affect mainly males and children from 0-9 years with a lethality rate of 16% [9].

In this article, we present the results of a comparative genomic analysis of two *L. (L.) amazonensis* strains isolated from dogs with clinical manifestation of visceral leishmaniasis in the city of Governador Valadares. Our study has found important differences in terms of SNPs, chromosome and gene copy number variation that are unique to this species. This information will contribute to understand some of the mechanisms of *L. (L.) amazonensis* infection and survival as well as to provide conclusive evidence of the presence of this species in endemic areas for VL.

Results

Sample collection, serology test and genomic sequencing

We isolated 36 samples from bone marrow or lymph node aspirates and sera from dogs of the study area. We found 32 dogs (89%) with clinical manifestations of VL **(Supplementary Table 1)** and all samples (36) were positive by the EIE-LVC kit thus confirming VL infection **(Supplementary Figure 1)**. Genomes were sequenced using Illumina HiSeq 2000 NGS technology at The Welcome Trust Sanger Institute sequencing facility.

Mapping and PCA

Single nucleotide polymorphisms against the *L. (L.) infantum* JPCM5 reference genome were used to generate alternate sequences for each isolate that were subsequently used for principal component analysis.

PCA of these isolates showed a clear grouping of 34 isolates around the *L. (L.) infantum* reference genome while samples S3 and S6 were markedly divergent **(Figure 1)**. This result suggested that isolates S3 and S6 were somehow unrelated to the rest of isolates and therefore required a more specific analysis.

Figure 1: PCA of SNPs of Governador Valadares isolates: GV: Governador Valadares *L. (L.) infantum* isolates; JPCM5: *L. (L.) infantum* JPCM5 reference strain. PCA shows that isolates S3 and S6 to do not group with *L. (L.) infantum*.

Based on the PCA evidence we genotyped isolates S3, S6 and a randomly selected isolate from the samples grouping with *L. (L.) infantum* (S1) using primers specific to the *hsp70* gene followed by digestion with *Hae*III restriction enzyme, as previously described (REF). The results for S3 and S6 isolates matched the expected restriction profile of *L. (L.)*

amazonensis/*L. (L.) mexicana* species, whereas S1 isolate was confirmed as *L. (L.) infantum* (**Supplementary Figure 2**). Competitive mapping against the *L. (L.) mexicana* U1103 and *L. (L.) amazonensis* M2269 reference genomes resulted in more than 78% of the reads from both S3 and S6 isolates mapping specifically to *L. (L.) amazonensis* M2269 with a median genome coverage of 40.5 and 29.7, respectively **(Supplementary Figure 3)**. This result suggested a closer relation between both samples to *L. (L.) amazonensis*.

Genome assembly

In order to confirm that samples S3 and S6 were indeed *L. (L.) amazonensis* we employed a hybrid assembly approach to generate a draft genome sequence for each isolate[12]. This method resulted in 3,584 and 3,236 contigs with an N50 of 29,346bp and 26,692bp for isolates S3 and S6, respectively **(Table 1)**. These contigs comprised more than 30.5Mpb that is greater than the current *L. (L.) amazonensis* M2269 reference genome (version from 2013-07-25). The resulting contigs were subsequently ordered into 34 pseudochromosomes assuming a similar chromosomal organization than that of *L. (L.) mexicana*[13].

	Scaffolds			Contigs		
Variable\Sample	**M2269**	**S3**	**S6**	**M2269**	**S3**	**S6**
Number	2,627	3,293	2,545	2,944	3,584	3,236
Size	29.0Mb	30.8Mb	30.5Mb	29.0Mb	30.8Mb	30.5Mb
Longest	171,320	196,967	314,951	113,027	196,967	174,893
N50	22,901	32,050	33,999	19,306	29,346	26,692
Mean size	11,050	9,364	12,002	9,854	8,601	9,425

Table 1: Genome assembly results for samples S3 and S6. M2269 is the version 2013-07-25 of the *L. (L.) amazonensis* genome .

Phylogenetic inferences

The coding sequences of seven different *Leishmania* species were retrieved from the Trytrip database V10 and used for phylogenetic comparison against isolates S3 and S6. This dataset consisted of 66 loci shared by these species (95,734 nucleotides). jModelTest analysis for the

concatenated nucleotide dataset selected the TVM model with invariable sites as the best model.

Maximum likelihood analysis provided statistical reliability to most nodes in the three with bootstrap values of 1,000 (Figure 2) except on the position of the S6 isolate inside the *L. (L.) amazonensis* node that could be explained by an insufficient amount of discriminative positions in both isolates. Nevertheless, the tree supports a common origin of isolates S3 and S6 together with *L. (L.) amazonensis* M2269, clearly indicating that they belong to this species.

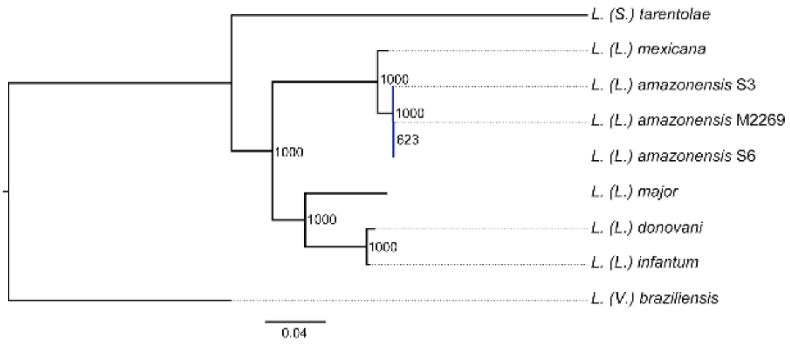


Figure 2: Maximum likelihood tree of 66 concatenated coding sequences (95,734 nucleotides). This result shows that isolates S3 and S6 correspond to *L. (L.) amazonensis*.

Chromosome and gene copy number variation

Chromosome copy numbers were estimated using the median read density of each chromosome normalized by the median read depth of the whole genome. Most chromosomes of both *L. (L.) amazonensis* isolates have a haploid copy number of one with the exception of Chr30 **(Figure 3, 4)**. This finding contrasts with previous results from studies conducted in other *Leishmania* species where a striking diversity in terms of aneuploidy was found that varies across different species and even isolates belonging to the same species[12,13].

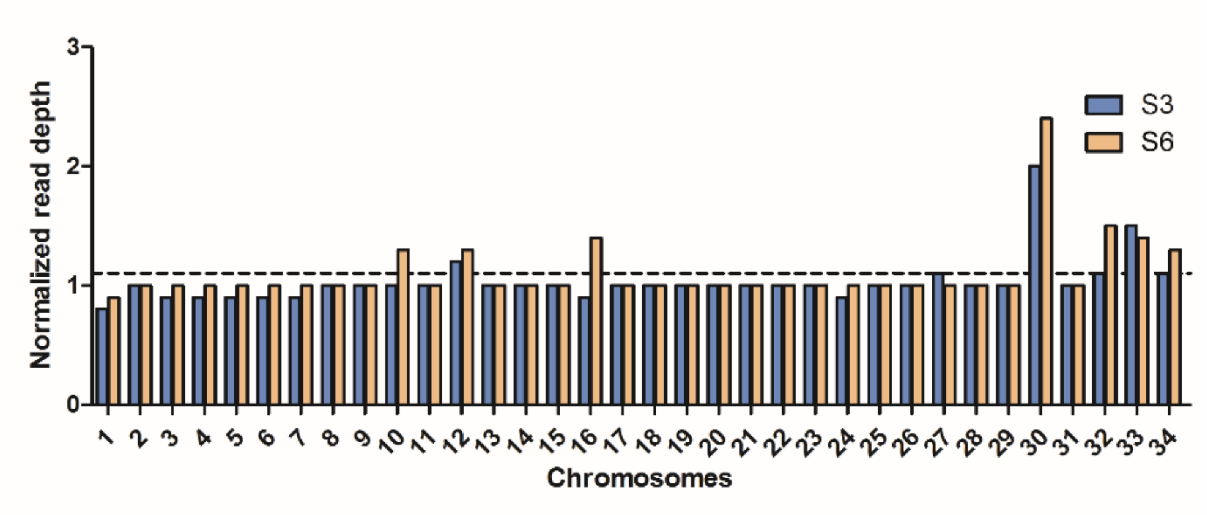


Figure 3: Chromosome copy number variation in *L. (L.) amazonensis* isolates S3 and S6. Columns represent the estimated haploid copy number for each chromosome. Mean genome ploidy is indicated by a dotted black line.

Chromosome 30 appears to be the only chromosome with significant increase in copy number. This chromosome is equivalent to chromosome 31 of Old World *Leishmania* and New World *Viannia* species if we assume a similar chromosomal organization than that of *L. (L.) mexicana*[13] due to two fusion events between chromosomes 8 and 29 and 20 and 36[14]. In both isolates, read depth of Chr30 is homogenously distributed along the entire chromosome, discarding segmental amplification **(Figure 4)**.

In order to assess chromosome ploidy and complement results from read depth based analyses we employed an allele frequency approach to identify disomic (clustering at 0.5 for heterozygous SNPs), trisomic (clustering at 0.33 and 0.67) and tetrasomic chromosomes (clustering at 0.25, 0.50 and 0.75).

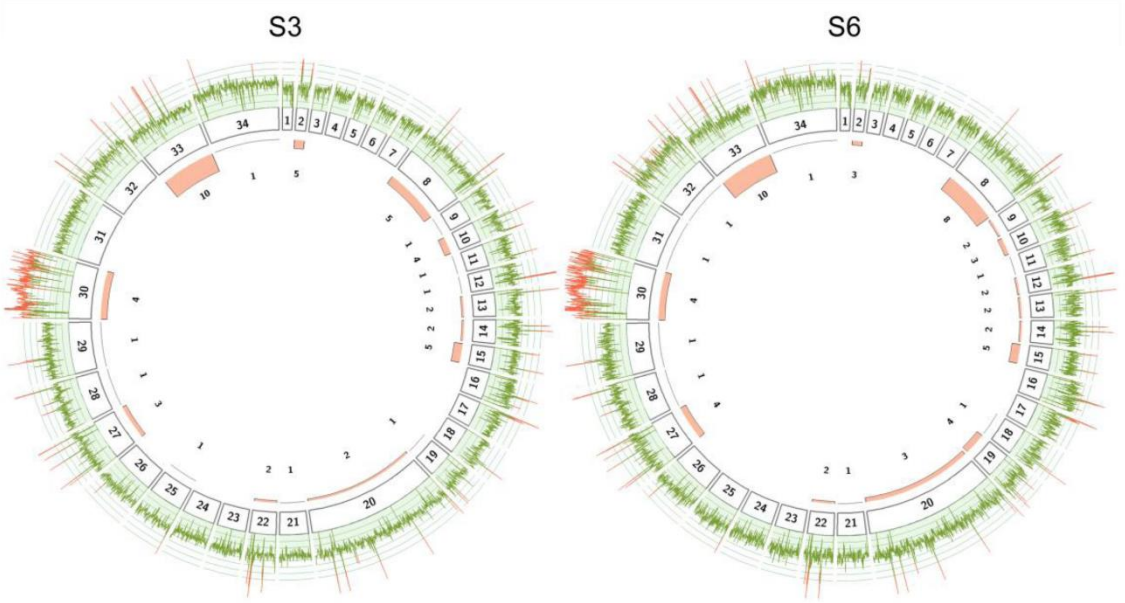


Figure 4: Chromosome and gene copy number variations in *L. (L.) amazonensis* isolates S3 and S6. Read depth from both isolates is shown as a green and red line plot for disomic and expanded regions of each chromosome, respectively. Internal histogram displays the total number of gene expansions in each chromosome.

Allele frequency profiles for heterozygous SNPs showed a marked disomic profile for all chromosomes in isolates S3 and S6 and a highly homogeneous structure within the cell population **(Figure 5, Supplementary figure 4 and 5)**. This result suggests that most chromosomes are disomic in both isolates. Contrastingly, chromosome 30 that is supernumerary by read depth has a disomic allele frequency pattern (**Supplementary figure 4 and 5**) that could indicate that it accumulates mutations in disomic alleles as has been shown in other *Leishmania* species [12,13].

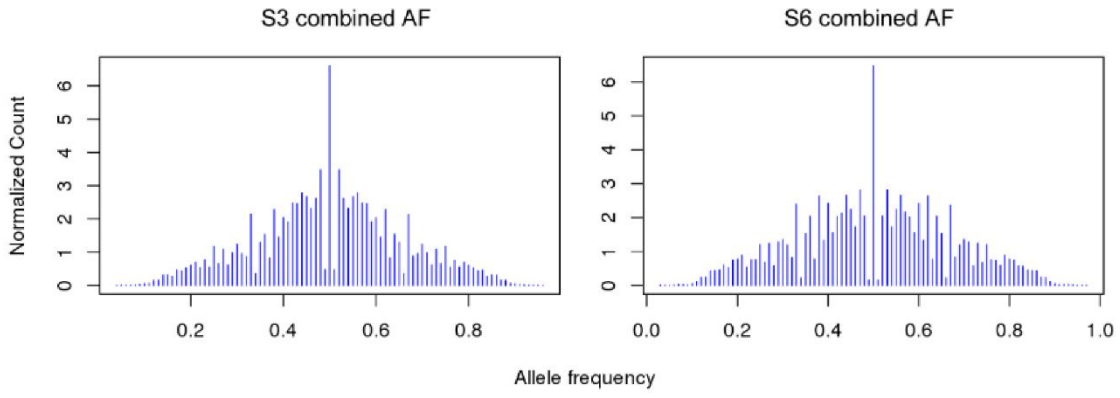


Figure 5: Normalized allele frequency distribution for *L. (L.) amazonensis* S3 and S6. Blue lines represent normalized counts of the proportions at heterozygous positions for all chromosomes. The results support a clear disomic tendency in the chromosomes of both isolates.

Gene copy number variations

It has been suggested that gene copy number variations in *Leishmania* can affect gene expression in response to the changing conditions within the host, contributing in part with the different disease tropism that is seen in *Leishmania*[13].

The gene copy number analysis resulted in 53 and 62 expanded genes in isolates S3 and S6 (**Supplementary Table 2**), respectively with 47 shared duplicated genes (**Figure 4 and Supplementary Table 3)**. Top expanded genes include a RNA helicase, a putative pyroglutamyl peptidase I (PPI) and several hypothetical proteins stressing the need of functional characterization. PPIs have been found in different organisms but a specific biological function has not been assigned yet. These proteins hydrolyze N-terminal-pyroglutamyl residues and confer stability to the modified peptides from aminopeptidase degradation and in some cases are crucial for functional activity[15]. A PPI in *Trypanosoma brucei* has been associated with protection against antimicrobial peptides suggesting that this enzyme could be an important virulence factor[16]. However, the corresponding ortholog in *L. (L.) major* appears to be a key factor during differentiation to metacyclic promastigotes[15]. Based on this evidence, the expanded ortholog in *L. (L.) amazonensis* is also likely to act during transition to infecting metacyclic promastigotes.

Gene ontology analysis on the expanded genes showed that enriched gene functions in this group are related to GTP catabolism **(Supplementary figure 6)**. GTPase proteins are crucial in vesicle formation, motility and union of vesicles to target compartments[17]. In *Leishmania,* GTPases play a major role during the regulation of vesicular transport in the exocytic and endocytic trafficking[18].

Another important characteristic of the *Leishmania* genomes is the presence of expanded tandem gene arrays that have been shown to vary greatly between species[13]. Our analysis found five tandem gene arrays in both *L. (L.) amazonensis* isolates **(Supplementary Table 4).** This tandem gene arrays include surface antigen protein 2 (PSA2), elongation factor 1 (EF-1α), ama1, HSP83 and beta tubulin.

The *Leishmania* surface antigen protein 2 (PSA-2), is a family of glycol-proteins that is expressed extracellularly in both parasite stages with overexpression in metacyclic promastigotes[19]. This family is involved in protecting the parasite from complement mediated lysis[20] and it may be also involved in host cell invasion due to the presence of leucine rich repeats that interact with the CR-3 receptor of macrophages[21]. Our results show the presence of an expanded tandem array of five PSA2 genes in *L. (L.) amazonensis* suggesting an important role of this virulence factor in this species **(Supplementary Table 4)**.

The *Leishmania* EF-1α is a tyrosine phosphatase-1 (SHP-1) binding protein that appears to be secreted to the phagosome. Experimental evidence shows that EF-1α targets host SHP-1 that is involved in macrophage inactivation by blocking the induction of nitric-oxide synthase in response to interferon-γ[22]. Consequently, this protein reverses the phenotype of infected macrophages towards a deactivated-like phenotype favoring parasite survival[22]. An expansion of a five EF-1α tandem array was found in our *L. (L.) amazonensis* isolates **(Supplementary Table 4)**. This expansion that is absent in *L. (L.) mexicana* may be particularly important for the characteristic disease phenotype that is associated to this species.

HSPs are well characterized proteins in *Leishmania* that maintain protein folding under stress conditions like the ones inside the phagosome and participate during differentiation in the lifecycle of *Leishmania*[23].

Discussion

L. (L.) amazonensis is an important cause of tegumentary leishmaniasis in Brazil been highly distributed in the north and northeastern regions of the country[6,7]. As part of a study aiming at characterizing the *Leishmania* isolates circulating in the city of Governador Valadares, Minas Gerais state, Brazil, we have sequenced the genomes of several *Leishmania* isolates form this focus. Genome sequencing of 36 isolates obtained from dogs revealed the presence of *L. (L.) amazonensis* in this endemic region of visceral and tegumentary leishmaniasis[25]. The genomic analysis herein performed of the two *L. (L.) amazonensis* isolates allowed us to explore some unique features in terms of chromosome and gene copy number that are unique to this species.

Our results show a clear and strong disomic trend in most *L. (L.) amazonensis* chromosomes that has not been reported in other *Leishmania* species so far analysed where genome plasticity and mosaic aneuploidy is a common trait[12,13]. Mosaic aneuploidy has been proposed as a rapid adaptive mechanism in *Leishmania* to deal with the different conditions inside its hosts[13,24]. In this sense, the strong disomic pattern seen in both *L. (L.) amazonensis* could be the result of low selection pressures.

Gene copy number variations in relation to *L. (L.) mexicana* show that species-specific expansions exist despite the high similarity, especially in expanded genes and tandem arrays in proteins involved in cell differentiation, cellular trafficking and parasite host interaction.

The different set of gene expansions in *Leishmania* that is known as intrachromosomal amplification appears to serve as a mechanism to modify gene dosage in the absence of transcriptional control[13]. In *L. (L.) amazonensis* this mechanism could be crucial for invasion and survival inside host macrophages with an important role for PSA2 and EF-1α

and also be partially responsible for the characteristic clinical phenotype that is associated to this species.

The Governador Valadares city is a re-emergent focus of visceral leishmaniasis with a high number of human cases due to *L. (L.) infantum* and a high prevalence of infected dogs[25]. To our knowledge, this article presents the first report of *L. (L.) amazonensis* in Governador Valadares and shows a potential risk to current control efforts in the area that have been designed considering the sole presence of *L. (L.) infantum*.

Importantly, both dogs infected with *L. (L.) amazonensis* presented similar clinical symptoms as the ones infected with *L. (L.) infantum* in the area. These symptoms included loss weight, lymphadenopathy, conjunctivitis, keratitis, anemia, ulcers, alopecia, dermatitis, onychogryphosis and vasculitis.

L. (L.) amazonensis infection in dogs has been reported in the Brazilian State of São Paulo[26] and other regions of Minas Gerais and together with our results demonstrate that *L. (L.) amazonensis* can affect peridomestic and domestic animals in urban settings increasing the risk of transmission to humans and possibly showing an urbanization process of this species. The finding of *L. (L.) amazonensis* in different ecological niches than the ones in north and northeastern regions of Brazil stresses the need to revise current serological and molecular *Leishmania* detection tests that are employed in endemic visceral leishmaniasis sites.

Our results indicate that the EIE-LVC kit and possibly other kits based on similar antigens can cross-react with *L. (L.) amazonensis*. This cross-reactivity may have resulted in an underestimation of the distribution and prevalence of *L. (L.) amazonensis* in Brazil and underscore the need to develop better diagnostic methods capable of discriminating at the species level.

Finally, surveillance efforts and control activities in this region should consider the presence of *L. (L.) amazonensis* and address putative vectors for this species and the risks of co-infections with *L. (L.) infantum* in human subjects.

Methods

Study site and sample collection

Samples were taken since 2008 from domestic dogs with clinical VL symptoms from the endemic focus of Governador Valadares in the southeastern Brazilian State of Minas Gerais (**Figure 6**). This city of 276,995 inhabitants is located at the bank of the Doce River at 455 meters above sea level. The region presents a tropical sub-humid climate[27] with a mean annual temperature of 29°C and a mean annual precipitation of 1,059 mm. The area is endemic for TL and a reemerging focus of VL with 127 human cases of VL reported from 2007 to 2013[28] and more than 30% of domestic dogs positive by serology[9]. We collected bone marrow or lymph node aspirates and serum for each dog that were subsequently used for *in vitro* culture of *Leishmania* and serology diagnosis.

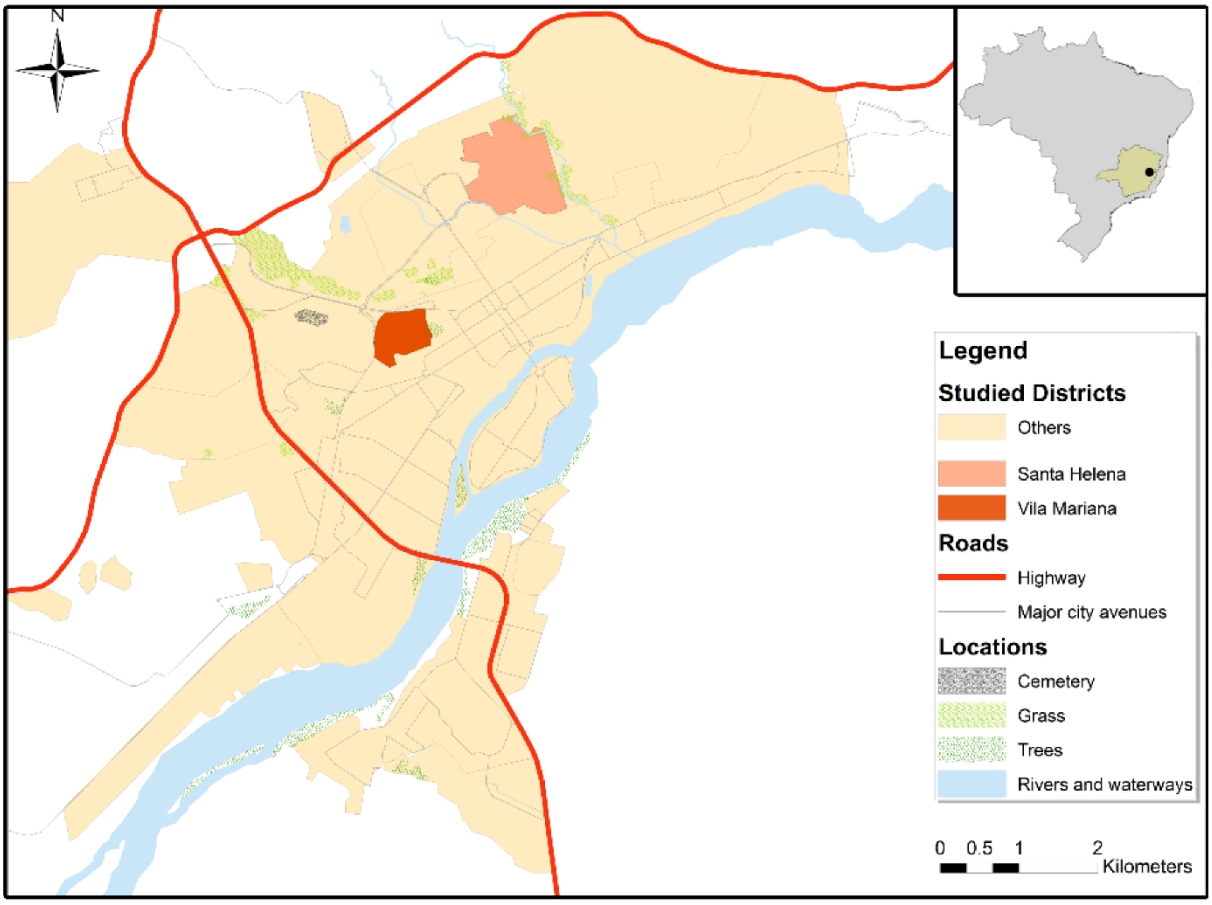


Figure 6: Map of the city of Governador Valadares. Inset shows the location of the Minas Gerais State in Brazil and the city of Governador Valadares City (upper right) and a city map showing the neighborhoods where isolates S3 and S6 were collected (Vila Mariana and Santa Helena, respectively)

Serology test

Sera from dogs were tested in triplicate by ELISA using the EIE-LVC kit supplied by Biomanguinhos following the manufacturer standard protocol. This kit consists of soluble antigens of *L. (L.) major* and has been widely used in public health laboratories in Brazil for the diagnosis and surveillance of canine visceral leishmaniasis.

Parasite isolates and sequencing

Libraries of 350bp were used to generate 100bp paired end reads for 36 samples at the Wellcome Trust Sanger Institute's sequencing facility by Illumina HiSeq.

The *L. (L.) infantum* JPCM5, *L. (L.) mexicana* U1103 and *L. (L.) amazonensis* M2269 reference genomes (version 10) were downloaded from the Tritryp database (http://tritrypdb.org/) for mapping and genome assembly steps.

Mapping and PCA

Initially, reads were mapped onto the *L. (L.) infantum* JPCM5 reference genome using Bowtie2[29] followed by SNP calling with SAMtools Mpileup[30], selecting sites with base quality scores ≥30, mapping quality scores ≥25, minimum coverage ≥ 10 reads and less than twice the median genome coverage.

Filtered SNPs were used to generate alternate sequences for each isolate in GATK[31] that were later used for PCA analysis using the Caret package in R[32].

Isolates S3 and S6 were subsequently mapped against the *L. (L.) mexicana* U1103 and *L. (L.) amazonensis* M2269 reference strains using Bowtie2[29].

Genome assembly and annotation

Reads from isolates S3 and S6 were filtered by quality using Trimmomatic[33] with a minimum base quality cutoff of 30, leading and trailing base qualities of 25, five bases sliding window, minimum per base average quality of 25 and a minimum read length of 65bp.

A combined *De novo* and reference based assembly approach[12] was employed for each sample. Briefly, for each sample we generated a *De Novo* assembly using Velvet 1.2.10[34] and a reference-based sequence with vcfutils[30] using the *L. (L.) amazonensis* M2269 genome as a template.

De novo and reference based sequences were combined in ZORRO[35] and the resulting hybrid assembly was extended and corrected with GapFiller[36] and iCORN2[37]. Contigs were scaffolded with SSPACE[38] and used to generate pseudochromosomes with ABACAS[39] assuming a similar chromosome organization as in *L. (L.) mexicana*[13].

Phylogenetic analysis

Nucleotide sequences from *L. (L.) amazonensis* M2269, *L. (L.) mexicana* U1103, *L. (L.) major* Friedlin, *L. (L.) infantum* JCPM5, *L. (L.) donovani* BPK282A1and *L. (Sauroleishmania) tarentolae* TarII were downloaded from the Tritryp database (http://tritrypdb.org/).

For each assembled genome of *L. (L.) amazonensis* (S3 and S6) a blastn search[40] was performed using a cutoff of $1e^{-5}$ against the *L. (L.) amazonensis* CDS and the best match was retrieved for each gene. Nucleotide sequences were filtered out using an in house Perl script in order to remove pseudogenes and partial or fragmented sequences.

CDS from all species were used as input for OrthoMCL[41] in order to select single copy genes with shared orthologs in all *Leishmania* species in the dataset. Each ortholog group was aligned using MUSCLEv3.8[42] and poorly aligned regions were removed with trimALv1.4[43]. Aligned ortholog groups were subsequently concatenated and jModelTest 2.1.5[44] was ran to carry out statistical selection of the best-fit models according to the Akaike Information

Criterion. Phylogenetic analysis of the concatenated dataset was performed using a Maximum Likelihood (ML) approach implemented in PhyML 3.0 with 1000 pseudoreplicates[45].

Allele frequency distribution

Allele frequencies for samples S3 and S6 were generated from filtered SAMtools results[30]. Briefly, for each heterozygous site we estimated the proportion of reads mapping to the alternate and reference base. Proportions were then grouped in bins from 0.01 to 1.0 and normalized by the sum of all allele frequencies for the respective chromosome. Plots of the distribution of allele frequencies were generated in R[32].

Chromosome and gene copy number analysis

To estimate haploid chromosome copy number, we normalized the median read depth for each chromosome by the median read depth of the whole genome using an in house perl script and figures were generated in Graph Pad Prism V5 and Circus[46].

Gene copy number variations were assessed by single copy gene normalization. Briefly, we used OrthoMCL[41] to select single copy genes with orthologs in *L. (V.) braziliensis*, *L. (L.) mexicana* and *L. (L.) major*, *L. (L.) infantum, L. (L.) donovani* and *L. (S.) tarentolae*. The mean read depth of these genes was then used to normalize each position along the genome. Gene copy numbers were furthered normalized by the average chromosome haploid copy number calculated from the allele frequency analysis. We employed a cutoff of 1.85 to discriminate between single copy and expanded genes.

In order to find enriched functions in expanded genes we analyzed overrepresented gene ontology codes using the hypergeometric distribution analysis with Benjamini and Hochberg false discovery rate correction implemented in BINGO[47].

References

1 Alvar, J. *et al.* Leishmaniasis worldwide and global estimates of its incidence. *PloS one* **7**, e35671, doi:10.1371/journal.pone.0035671 (2012).

2 Murray, H. W., Berman, J. D., Davies, C. R. & Saravia, N. G. Advances in leishmaniasis. *Lancet* **366**, 1561-1577, doi:10.1016/S0140-6736(05)67629-5 (2005).

3 Banuls, A. L., Hide, M. & Prugnolle, F. Leishmania and the leishmaniases: a parasite genetic update and advances in taxonomy, epidemiology and pathogenicity in humans. *Advances in parasitology* **64**, 1-109, doi:10.1016/S0065-308X(06)64001-3 (2007).

4 Peacock, C. S. *et al.* Comparative genomic analysis of three Leishmania species that cause diverse human disease. *Nature genetics* **39**, 839-847, doi:10.1038/ng2053 (2007).

5 Lainson, R., Shaw, J. J., Peters, W. & Killick-Kendrick, R. *Evolution, classification and geographical distribution*. (Academic Press, 1987).

6 Camara Coelho, L. I. *et al.* Characterization of Leishmania spp. causing cutaneous leishmaniasis in Manaus, Amazonas, Brazil. *Parasitology research* **108**, 671-677, doi:10.1007/s00436-010-2139-9 (2011).

7 de Oliveira, J. P. *et al.* Genetic diversity of Leishmania amazonensis strains isolated in northeastern Brazil as revealed by DNA sequencing, PCR-based analyses and molecular karyotyping. *Kinetoplastid biology and disease* **6**, 5, doi:10.1186/1475-9292-6-5 (2007).

8 Desjeux, P. Leishmaniasis: current situation and new perspectives. *Comparative immunology, microbiology and infectious diseases* **27**, 305-318, doi:10.1016/j.cimid.2004.03.004 (2004).

9 Barata, R. A. *et al.* Epidemiology of visceral leishmaniasis in a reemerging focus of intense transmission in Minas Gerais State, Brazil. *BioMed research international* **2013**, 405083, doi:10.1155/2013/405083 (2013).

10 Arce, A. *et al.* Re-emergence of leishmaniasis in Spain: community outbreak in Madrid, Spain, 2009 to 2012. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* **18**, 20546 (2013).

11 Arias, J. R., Monteiro, P. S. & Zicker, F. The reemergence of visceral leishmaniasis in Brazil. *Emerging infectious diseases* **2**, 145-146, doi:10.3201/eid0202.960213 (1996).

12 Valdivia, H. O. *et al.* Comparative genomic analysis of Leishmania (Viannia) peruviana and Leishmania (Viannia) braziliensis. *BMC genomics* **16**, 715, doi:10.1186/s12864-015-1928-z (2015).

13 Rogers, M. B. *et al.* Chromosome and gene copy number variation allow major structural change between species and strains of Leishmania. *Genome research* **21**, 2129-2142, doi:10.1101/gr.122945.111 (2011).

14 Britto, C. *et al.* Conserved linkage groups associated with large-scale chromosomal rearrangements between Old World and New World Leishmania genomes. *Gene* **222**, 107-117 (1998).

15 Schaeffer, M., de Miranda, A., Mottram, J. C. & Coombs, G. H. Differentiation of Leishmania major is impaired by over-expression of pyroglutamyl peptidase I. *Molecular and biochemical parasitology* **150**, 318-329, doi:10.1016/j.molbiopara.2006.09.004 (2006).

16 Morty, R. E., Bulau, P., Pelle, R., Wilk, S. & Abe, K. Pyroglutamyl peptidase type I from Trypanosoma brucei: a new virulence factor from African trypanosomes that de-blocks regulatory peptides in the plasma of infected hosts. *The Biochemical journal* **394**, 635-645, doi:10.1042/BJ20051593 (2006).

17 Stenmark, H. & Olkkonen, V. M. The Rab GTPase family. *Genome biology* **2**, REVIEWS3007 (2001).
18 Chenik, M. *et al.* Identification of a new developmentally regulated Leishmania major large RAB GTPase. *Biochemical and biophysical research communications* **341**, 541-548, doi:10.1016/j.bbrc.2006.01.005 (2006).
19 Devault, A. & Banuls, A. L. The promastigote surface antigen gene family of the Leishmania parasite: differential evolution by positive selection and recombination. *BMC evolutionary biology* **8**, 292, doi:10.1186/1471-2148-8-292 (2008).
20 Lincoln, L. M., Ozaki, M., Donelson, J. E. & Beetham, J. K. Genetic complementation of Leishmania deficient in PSA (GP46) restores their resistance to lysis by complement. *Molecular and biochemical parasitology* **137**, 185-189, doi:10.1016/j.molbiopara.2004.05.004 (2004).
21 Kedzierski, L. *et al.* A leucine-rich repeat motif of Leishmania parasite surface antigen 2 binds to macrophages through the complement receptor 3. *J Immunol* **172**, 4902-4906 (2004).
22 Nandan, D., Yi, T., Lopez, M., Lai, C. & Reiner, N. E. Leishmania EF-1alpha activates the Src homology 2 domain containing tyrosine phosphatase SHP-1 leading to macrophage deactivation. *The Journal of biological chemistry* **277**, 50190-50197, doi:10.1074/jbc.M209210200 (2002).
23 Lawrence, F. & Robert-Gero, M. Induction of heat shock and stress proteins in promastigotes of three Leishmania species. *Proceedings of the National Academy of Sciences of the United States of America* **82**, 4414-4417 (1985).
24 Sterkers, Y. *et al.* Novel insights into genome plasticity in Eukaryotes: mosaic aneuploidy in Leishmania. *Molecular microbiology* **86**, 15-23, doi:10.1111/j.1365-2958.2012.08185.x (2012).
25 Penaforte, K. M. *et al.* Leishmania infection in a population of dogs: an epidemiological investigation relating to visceral leishmaniasis control. *Revista brasileira de parasitologia veterinaria = Brazilian journal of veterinary parasitology : Orgao Oficial do Colegio Brasileiro de Parasitologia Veterinaria* **22**, 592-596, doi:10.1590/S1984-29612013000400022 (2013).
26 Tolezano, J. E. *et al.* The first records of Leishmania (Leishmania) amazonensis in dogs (Canis familiaris) diagnosed clinically as having canine visceral leishmaniasis from Aracatuba County, Sao Paulo State, Brazil. *Veterinary parasitology* **149**, 280-284, doi:10.1016/j.vetpar.2007.07.008 (2007).
27 Kottek, M., Grieser, J., Beck, C., Rudolf, B. & Rubel, F. World map of the Köppen-Geiger climate classification updated. *Meteorologische Zeitschrift* **15**, 259-263 (2006).
28 Saúde, M. d. & Saúde, d. d. V. e. (Ministério da Saúde Brasília, 2007).
29 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).
30 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
31 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).
32 Team, R. C. (ISBN 3-900051-07-0, 2014).
33 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).

34 Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* **18**, 821-829, doi:10.1101/gr.074492.107 (2008).
35 Real, F. *et al.* The genome sequence of Leishmania (Leishmania) amazonensis: functional annotation and extended analysis of gene models. *DNA research : an international journal for rapid publication of reports on genes and genomes* **20**, 567-581, doi:10.1093/dnares/dst031 (2013).
36 Boetzer, M. & Pirovano, W. Toward almost closed genomes with GapFiller. *Genome biology* **13**, R56, doi:10.1186/gb-2012-13-6-r56 (2012).
37 Otto, T. D., Sanders, M., Berriman, M. & Newbold, C. Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics* **26**, 1704-1707, doi:10.1093/bioinformatics/btq269 (2010).
38 Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578-579, doi:10.1093/bioinformatics/btq683 (2011).
39 Assefa, S., Keane, T. M., Otto, T. D., Newbold, C. & Berriman, M. ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* **25**, 1968-1969, doi:10.1093/bioinformatics/btp347 (2009).
40 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403-410, doi:10.1016/S0022-2836(05)80360-2 (1990).
41 Fischer, S. *et al.* Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **Chapter 6**, Unit 6 12 11-19, doi:10.1002/0471250953.bi0612s35 (2011).
42 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792-1797, doi:10.1093/nar/gkh340 (2004).
43 Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973, doi:10.1093/bioinformatics/btp348 (2009).
44 Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. jModelTest 2: more models, new heuristics and parallel computing. *Nature methods* **9**, 772, doi:10.1038/nmeth.2109 (2012).
45 Criscuolo, A. morePhyML: improving the phylogenetic tree space exploration with PhyML 3. *Molecular phylogenetics and evolution* **61**, 944-948, doi:10.1016/j.ympev.2011.08.029 (2011).
46 Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome research* **19**, 1639-1645, doi:10.1101/gr.092759.109 (2009).
47 Maere, S., Heymans, K. & Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21**, 3448-3449, doi:10.1093/bioinformatics/bti551 (2005).

Authors' contributions

HOV: carried out most bioinformatics analysis, participated in study conception, design and drafted the manuscript.

LA: participated in genome assembly, phylogenetic analysis and drafted the manuscripts.

AR: participated in study conception, design and drafted the manuscript.

BR: contributed in study conception, design and drafted the manuscript.

RTF: contributed in serology analysis

CG: contributed in genotyping analysis

JC: participated in study design, coordination and genome sequencing.

MB: participated in study design, coordination and genome sequencing.

DCB: participated in bioinformatic analysis, study design, coordination and manuscript writing

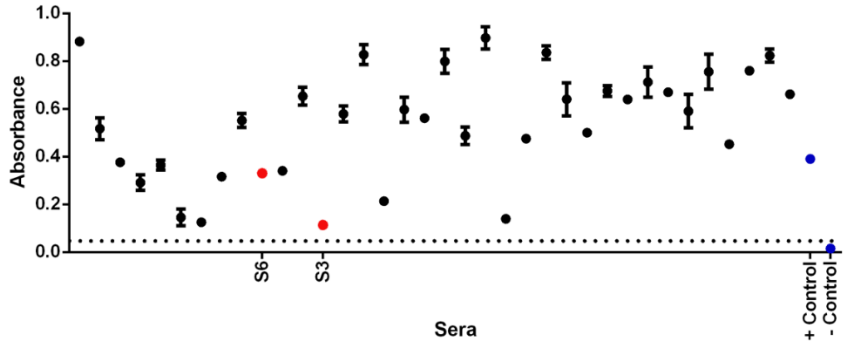
Competing Interests

None of the authors have any competing interests.

Acknowledgements

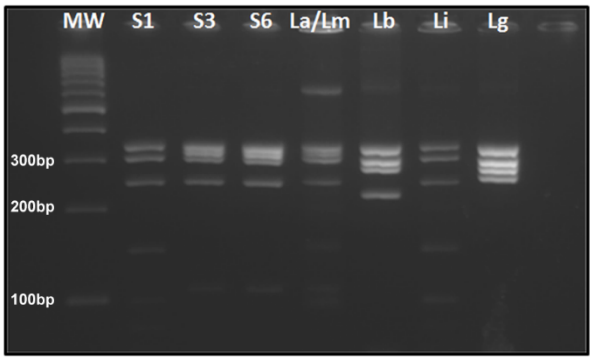
Daniella C. Bartholomeu research was supported by Fundação de Amparo a Pesquisa do Estado de Minas Gerais (FAPEMIG), Instituto Nacional de Ciência e Tecnologia de Vacinas (INCTV)—Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). DCB is a CNPq research fellow. HOV, received scholarship from CAPES.

Chapter 3: Supplementary Figure 1



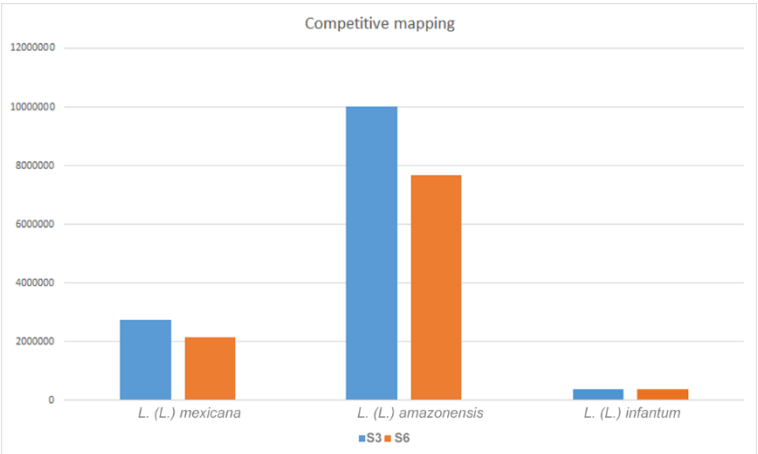
Supplementary figure 1: Elisa EIE-LVC kit results. Colors denote samples as follow: Blue, positive and negative controls; black, *L. (L.) infantum* isolates; red, *L. (L.) amazonensis* isolates.

Chapter 3: Supplementary Figure 2



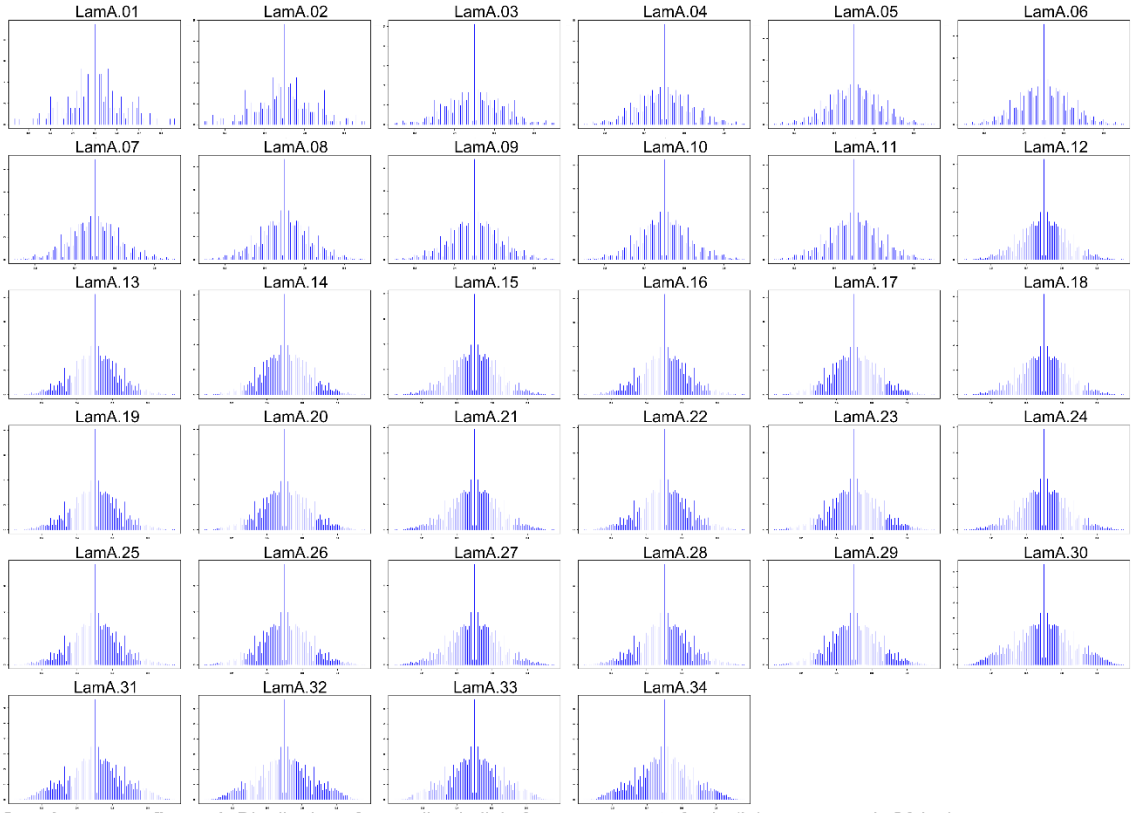
Supplementary figure 2: Genotyping of *Leishmania* isolates from Governador Valadares (GV). Genomic DNA was amplified using primers specific to the *hsp70* gene followed by digestion with HaeIII restriction enzyme and separation by electrophoresis in 4% agarose gel. MW – 100 bp molecular weight; S1 – GV sample S1 S3 – GV sample S3; S6 – GV sample S3; La/Lm – *Leishmania amazonensis* control (undistinguishable from *L. mexicana*); Lb – *L. braziliensis* control; Li – *L. infantum* control; Lg – *L. guyanensis* control

Chapter 3: Supplementary Figure 3



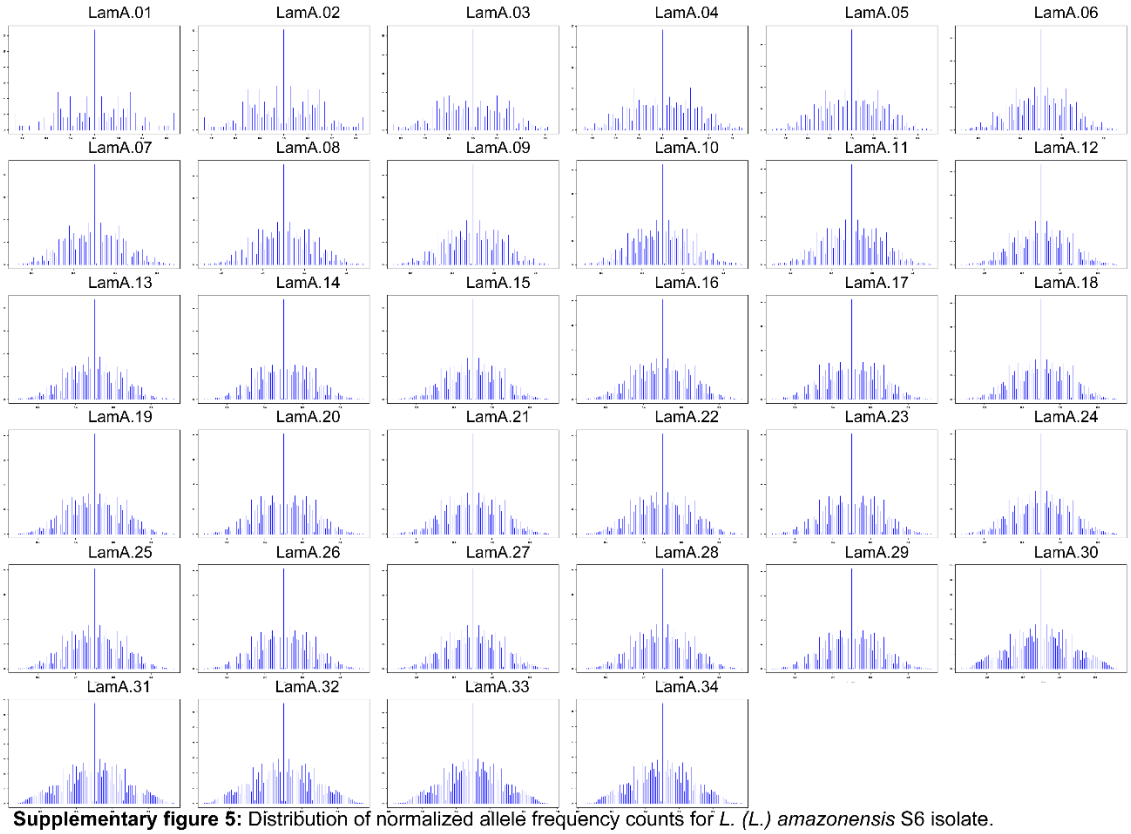
Supplementary figure 3: Competitive mapping of *L.(L.) amazonensis* S3 and S6 reads against the reference genomes of *L.(L.) mexicana*, *L.(L.) amazonensis* and *L.(L.) infantum*. Y axis shows the number of mapped reads, x axis denotes the species.

Chapter 3: Supplementary Figure 4

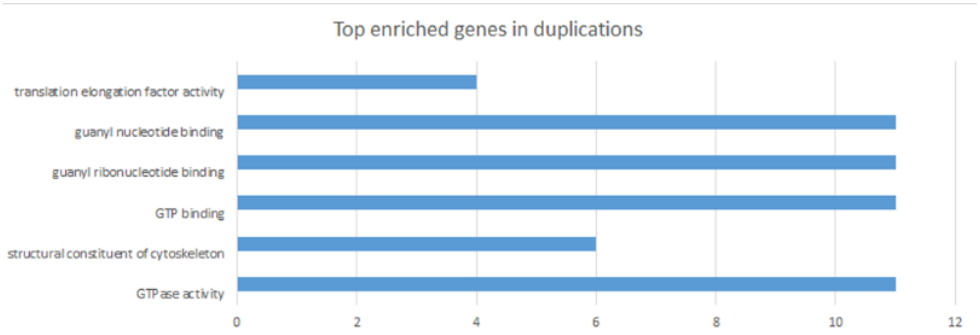


Supplementary figure 4: Distribution of normalized allele frequency counts for *L. (L.) amazonensis* S3 isolate.
Y axis represents the percentage of frequency counts and the X axis the allele.

Chapter 3: Supplementary Figure 5



Supplementary figure 5: Distribution of normalized allele frequency counts for *L. (L.) amazonensis* S6 isolate.
Y axis represents the percentage of frequency counts and the X axis the allele.



Supplementary figure 6: Enriched statistically significant gene ontology terms in the *L. (Leishmania) amazonensis* expansions. The y-axis show the enriched ontology term and the x-axis the number of genes associated to it.

Clinical symptomatology of dogs from Governador Valadares. Numbers indicate the following symptoms. 1: Loss of weight; 2: Linfadenopathy; 3: Conjunctivitis; 4: Keratitis; 5: Anaemia; 6: Ulcers/Nodules; 7: Alopecia; 8: Dermatitis; 9: Onychogryphosis; 10: Vasculitis

Sample code	Clinical symptoms	Neighborhood
C40	1, 2	Nossa Senhora das Graças
C41	1, 2, 6, 10	Santa Helena
C42	1, 2, 7	Grã-Duquesa
C43	1, 2, 3, 6, 7, 9, 10	Santa Helena
C46	1, 2	Santa Helena
C47	-	Santa Helena
C48	-	Grã-Duquesa
C49	-	Nossa Senhora das Graças
C50	-	Grã-Duquesa
J1	2, 5, 6, 8	Maravilha
J2	1, 2, 3, 4, 6, 8, 9, 10	Jardim Atalaia
J3	1, 2, 5, 8, 9	Jardim Atalaia
J4	1, 2, 5, 6, 7, 8	Nossa Senhora das Graças
J5	1, 2, 3, 4, 6, 7, 8, 9, 10	Jardim Atalaia/Azteca
J6	2, 6, 7, 8, 9, 10	Jardim do Trevo
J7	1, 2, 5, 8	Maravilha/Palmeiras
J8	1, 2, 5, 6, 8, 9, 10	Vila Mariquita
S-1	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	Altinópolis
S-10	1, 2, 5, 7, 8	Vila Isa/Vila Parque Ibituruna
S-11	1, 2, 6, 7, 8	Turmalina
S-12	1, 2, 3, 4, 8, 9, 10	Mãe de Deus
S-13	2, 3, 4, 6, 7, 10	Esperança
S-14	1, 2, 3, 4, 6, 8, 9	Jardim do Trevo
S-2	1, 2, 5, 6, 7, 9, 10	Chácara Braúna
S-3	2,3, 5, 6, 7, 8, 9, 10	Vila Mariana
S-4	1, 2, 5, 7, 8, 9	Vitória
S-5	2, 5, 6, 7, 8	Santa Rita
S-6	1, 2,3, 4, 5, 6, 7, 8, 9, 10	Santa Helena
S-7	1, 2, 3, 4, 5, 7, 8, 9, 10	Altinópolis
S-8	2, 3, 4, 5, 9	Altinópolis
S-9	1, 2, 3, 4, 5, 6, 7, 8	Altinópolis
V1	6, 8	Turmalina
V11	1, 2, 3, 5, 6, 7, 8, 9, 10	Turmalina
V13	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	Turmalina
V14	2,3, 5, 6, 7, 8, 9, 10	São Paulo
V4	2, 6, 8	Turmalina

Expanded genes in *L. (L.) amazonensis* isolates S3 and S6. Due to the absence of reference CDS for *L. (L.) amazonensis*, IDs correspond to the ortholog in *L. (L.) mexicana*. Only the top 30 genes out of 115 are shown.

Sample	Gene	Position	HCN
S6	LmxM.33.2070	LmxM.33_725558-728944	13.46096415
S3	LmxM.33.2070	LmxM.33_725558-728944	13.1720777
S3	LmxM.33.2040	LmxM.33_713213-714463	12.89297599
S6	LmxM.33.1990	LmxM.33_686832-689879	12.58013452
S3	LmxM.33.1990	LmxM.33_686832-689879	12.54335843
S3	LmxM.33.2060	LmxM.33_722991-723923	12.50612606
S6	LmxM.33.2040	LmxM.33_713213-714463	12.23178738
S6	LmxM.33.2060	LmxM.33_722991-723923	12.12420033
S3	LmxM.33.2050	LmxM.33_719840-722254	11.99601444
S6	LmxM.33.2050	LmxM.33_719840-722254	11.86561392
S3	LmxM.33.2000	LmxM.33_693151-693984	11.53244328
S6	LmxM.33.2000	LmxM.33_693151-693984	11.11774763
S3	LmxM.33.2020	LmxM.33_698252-700141	9.163110905
S3	LmxM.33.2010	LmxM.33_695934-696431	8.740015838
S3	LmxM.33.2030	LmxM.33_701680-705198	8.598561054
S6	LmxM.33.2020	LmxM.33_698252-700141	8.52789881
S6	LmxM.33.2010	LmxM.33_695934-696431	8.121249134
S6	LmxM.33.2030	LmxM.33_701680-705198	8.039877254
S6	LmxM.12.0867	LmxM.12_380281-381600	8.000643231
S3	LmxM.12.0867	LmxM.12_380281-381600	6.57789427
S6	LmxM.13.0280	LmxM.13_90325-91971	6.011368574
S6	LmxM.13.0390	LmxM.13_93414-95060	5.841662067
S6	LmxM.08.1171	LmxM.08_1668001-1669332	5.497551294
S3	LmxM.13.0280	LmxM.13_90325-91971	5.205761709
S3	LmxM.13.0390	LmxM.13_93414-95060	5.166413867
S6	LmxM.15.1050	LmxM.15_408603-408953	4.634631872
S6	LmxM.30.0480	LmxM.30_179649-180803	4.52306012
S6	LmxM.15.1160	LmxM.15_412848-413447	4.488707764
S6	LmxM.10.0390	LmxM.10_180317-182125	4.288285147
S3	LmxM.15.1050	LmxM.15_408603-408953	4.287645412

Expanded common genes in *L. (L.) amazonensis* isolates S3 and S6. HCN represents the mean haploid copy number for each gene. Due to the absence of reference CDS for *L. (L.) amazonensis*, IDs correspond to the ortholog in *L. (L.) mexicana.* Only the top 30 genes out of 47 are shown

Ortholog ID in *L. (L.) mexicana*	Annotation	HCN
LmxM.33.2070	hypothetical protein, conserved	13.3
LmxM.33.2040	hypothetical protein, unknown function	12.6
LmxM.33.1990	hypothetical protein, conserved	12.6
LmxM.33.2060	hypothetical protein, conserved	12.3
LmxM.33.2050	ATP-dependent RNA helicase, putative	11.9
LmxM.33.2000	Cysteine peptidase, Clan CF, family C15, pyroglutamyl-peptidase I, putative (PPI)	11.3
LmxM.12.0867	hypothetical protein, conserved	9.1
LmxM.33.2020	hypothetical protein, conserved	8.9
LmxM.33.2010	hypothetical protein, conserved	8.4
LmxM.33.2030	hypothetical protein, conserved	8.3
LmxM.13.0280	alpha tubulin	5.6
LmxM.13.0390	alpha tubulin	5.5
LmxM.15.1050	developmentally regulated protein, putative	4.5
LmxM.08.1171	unspecified product	4.4
LmxM.15.1160	tryparedoxin peroxidase	4.3
LmxM.30.0480	unspecified product	4.2
LmxM.10.0390	GP63, leishmanolysin	3.9
LmxM.29.1500	p1/s1 nuclease	3.5
LmxM.14.0400	hypothetical protein, conserved	3.5
LmxM.28.2770	heat-shock protein hsp70, putative	3.1
LmxM.22.1290	ribonucleoside-diphosphate reductase small chain, putative	3.0
LmxM.10.0470	GP63, leishmanolysin	3.0
LmxM.15.0440	unspecified product	2.8
LmxM.33.0960	amastin-like surface protein, putative	2.8
LmxM.15.1040	tryparedoxin peroxidase	2.7
LmxM.09.0891	polyubiquitin, putative	2.7
LmxM.30.0490	hypothetical protein, unknown function	2.7
LmxM.27.1570	hypothetical protein, conserved	2.5
LmxM.36.1280	tuzin-like protein	2.5
LmxM.27.0680	amino acid permease, putative	2.5

Tandem gene arrayss in *L. (L.) amazonensis* isolates S3 and S6. HCN represents the mean haploid copy number for the whole array in both isolates.

Chr	Genes in Array	Annotation	Ortholog IDs in *L. mexicana*	HCN
12	5	surface antigen protein 2, putative	LmxM.12.0870partial, LmxM.12.0890,LmxM.12.0891, LmxM.12.0980,LmxM.12.0990	5.6
17	5	elongation factor 1-alpha	LmxM.17.0080, LmxM.17.0081,LmxM.17.0082, LmxM.17.0084, LmxM.17.0085	3.0
29	2	ama1 protein, putative	LmxM.29.1410, LmxM.29.1420	4.0
32	3	heat shock protein 83-1	LmxM.32.0312, LmxM.32.0314, LmxM.32.0316	3.6
32	2	beta tubulin	LmxM.32.0792, LmxM.32.0794	3.8

Population genomics and evidence of clonal replacement in a canine-isolated population of *Leishmania (Leishmania) infantum* from Governador Valadares.

Abstract

Visceral leishmaniasis is a major health problem in Brazil, being highly endemic in the State of Minas Gerais, in particular affecting people with limited resources and little access to adequate health care. The region of Governador Valadares is a reemerging focus of intense transmission of VL and TL with more than 127 cases reported since 2007 to 2013 and 30% of positive domestic dogs.

In this subsection we will present the results of a study aimed at understanding the parasite population structure in the city using comparative genomic analysis of 36 samples from domestic dogs collected from 2008 to 2015.

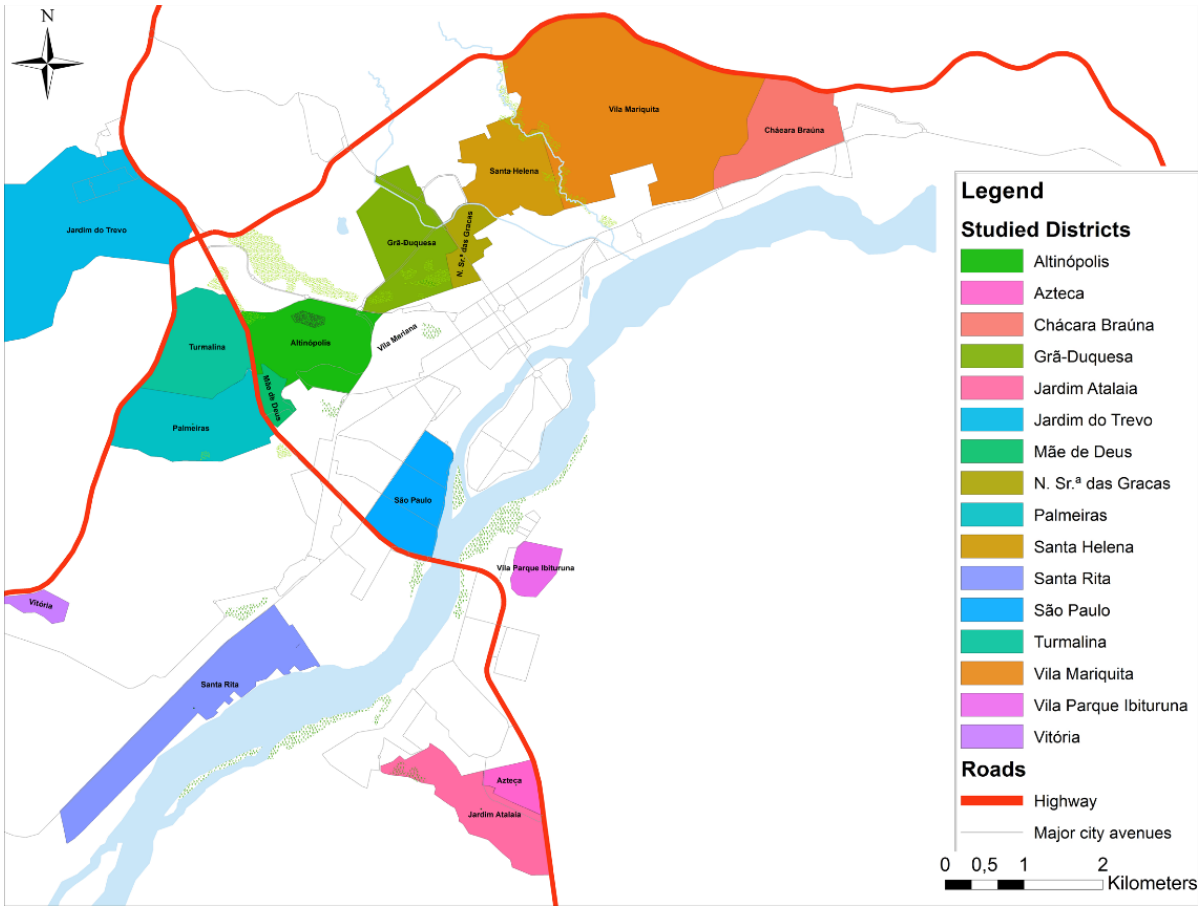
Analyses of this data were focused on the use of phylogenetic and population genetics methods to understand the evolution and the recent demographic history of parasites in this focus. Our results reveal the presence of two distinct sub-populations associated with two isolation periods and suggest a potential effect of climate on changes in the parasite population in this focus.

Methods

Study site and sample collection

Bone marrow and serum from each dog were collected in 2008, 2012 and 2015 from domestic dogs with clinical VL symptoms from Governador Valadares

Chapter 3: Figure 1**)** as already described in the previous section.



Chapter 3: Figure 1

Map of the city of Governador Valadares. Inset shows the location of the Minas Gerais State in Brazil and the city of Governador Valadares City (upper right) and a city map showing the neighborhoods where isolates were collected.

Parasite isolates and sequencing

We generated libraries of 350bp that were subsequently used to generate 100bp paired end reads at the Wellcome Trust Sanger Institute's sequencing facility by Illumina HiSeq.

The reference genomes of *L. (L.) infantum* JPCM5, *L. (L.) mexicana* U1103 and *L. (L.) amazonensis* M2269 (version 10) were downloaded from the Tritryp database (http://tritrypdb.org/) for all subsequent bioinformatics analyses.

Filtering and mapping

Reads filtered by quality using Trimmomatic (Bolger *et al.*, 2014) with minimum base quality cutoff of 30, leading and trailing base qualities of 28, minimum per base average quality of 20 and a minimum read length of 70bp.

Filtered reads were then mapped onto the *L. (L.) infantum* JPCM5, reference genome using Bowtie2 (Langmead e Salzberg, 2012). The fraction of reads mapped to each species was plotted using R and later used for CN analysis and variant calling.

Population structure analysis

SNPs among the isolates were called using the recommended parameters of GATK (Mckenna *et al.*, 2010). Briefly, mapped bam files were filtered for redundant reads and indels were re-evaluated via a local realignment. SNPs were called using the haplotype caller module and raw variants were filtered selecting sites with minimum raw coverage of 10, Root Mean Square mapping quality of 40, quality by depth greater than 2 and haplotype score greater than 13. Filtered SNPs were used to generate genomic sequence for each isolate in GATK (Mckenna *et al.*, 2010). These sequences were aligned among isolates with MAFFT (Katoh *et al.*, 2002), trimmed with TrimAL (Capella-Gutierrez *et al.*, 2009) and processed with a custom perl script to select non-biallelic SNPs that discriminate among isolates and create a numerical matrix for principal component analysis (PCA) and hierarchical clustering (HC) in R (Team, 2014).

The SNPs matrix of the *L. (L.) infantum* isolates was used for population analyses in Structure (Pritchard *et al.*, 2000) in a two-step process. First, we used 20 independent Markov Chain Monte Carlo (MCMC) runs with 10,000 burnin and a length of 100,000 for a range of population of one to eight. Then we re-executed Structure with the number of populations selected by the Delta K and Evanno methods with 20 MCMC runs of 500,000 burning and length of 750,000 in order to ensure convergence between runs.

Phylogenetic analysis

For all *L. (L.) infantum* isolates, alternate nuclear genome sequences were generated in GATK (Mckenna *et al.*, 2010) and aligned in MAFFT (Katoh *et al.*, 2002). The combined multiple sequence alignment was analyzed on jModelTest (Darriba *et al.*, 2012) for statistical selection of the best-fit models according to the Akaike Information Criterion (AIC), Decision Theory Method (DT) and Bayesian Information Criterion (BIC). Phylogenetic analysis of the concatenated dataset was based on Maximum Likelihood in PhyML 3.0 with 1,000 pseudoreplicates (Criscuolo, 2011) including the sequence of *L. (L.) infantum* JCPM5 as a reference.

Genome-wide assessment of within-host diversity

The within-host diversity (FWs) is a metric that describes the relation between the individual diversity to that of the population using estimations of heterozygosity (Auburn *et al.*, 2012; Manske *et al.*, 2012). This metric captures the overall diversity at the individual level and also the similarity between parasites and their relative proportions.

In order to estimate the FWs, we extracted biallelic SNPs with SAMtools to minimize confounding due to aneuploidy in *Leishmania*. We then employed Perl and R scripts to obtain the within-individual (*Hs*) and within-population heterozygosity (Hp).

Briefly, at each SNPs, we estimated heterozygosity as the proportion of reads mapping to reference and alternate alleles (p and q respectively) and the sample heterozygosity (Hs) was derived $Hs=1-(p^2+q^2)$. At the population level, the allele frequencies at each SNP were estimated as the total read counts for the two alleles across all samples in the population and the population heterozygosity was estimated (Hw).

Then, we extracted the minimum allele frequency (MAF) at each position and grouped them into ten bins ranging from 0 to 0.5. Then, we estimated within-individual and within-population heterozygosity for each MAF bins as the mean across the corresponding interval.

Within-host estimates were plotted against the corresponding within-population estimate for all intervals and the slope was used to calculate the FWs for each individual ($FWs=1-(Hw/Hs)$).

Recombination analysis

Heterozygous SNPs were phased in GATK (Mckenna *et al.*, 2010) using 5Kb windows and a custom Perl script to retrieve the two corresponding sequences for the phased region. Then, we performed a blast search against the *L. (L.) infantum* JPCM5 and *L. (L.) donovani* BPK282A1 reference genomes and extracted the sequence of the best match for the two reference strains for maximum likelihood phylogenetic analysis in PHYML using the GTR model with 100 pseudoreplicates. Median joining networks were constructed for phased haplotypes in PopArt.

Chromosome and gene copy number analysis

To estimate haploid chromosome copy number, gene copy number variations, expanded tandem gene arrays and gene enrichment we employed the same methodologies as described on the *L. (L.) amazonensis* section.

Allele frequency distribution

Allele frequencies for all isolates were generated from filtered SAMtools results (Li *et al.*, 2009). Briefly, for each heterozygous site we estimated the number of reads mapping to the alternate and reference bases. Counts were grouped in bins from 0.01 to 1.0 and we took the proportion respective to the sum of all allele frequencies by each chromosome and generated plots of the distribution of allele frequencies in R.

Divergence Time Estimation

Referenced-generated sequences of the *L. (L.) infantum* isolates were aligned against the *L. (L.) infantum* JCPM5 reference on MUSCLEv3.84 (Edgar, 2004) and poorly aligned regions were removed with trimALv1.4 (Capella-Gutierrez *et al.*, 2009).

The alignment was then run under a coalescent model on Beast2 (Bouckaert *et al.*, 2014) under the strict and relaxed log normal clocks in a MCMC run of 40 million. The calibration points were provided from divergence dates already estimated between Old World and New World *L (L.). infantum* (500 ya, SD 200 years) under a normal distribution model.

All information produced by BEAST was summarized onto a single “target” tree using the TreeAnnotator module of BEAST with Burnin of 30% of the samples and tree topology was represented using Figtree.

Bayesian skyline analysis were run for the Governador Valadares isolates to infer the effective population size on a 10,000,000 generation chain using the root calibration point estimated during the divergence analysis.

Climate data analysis

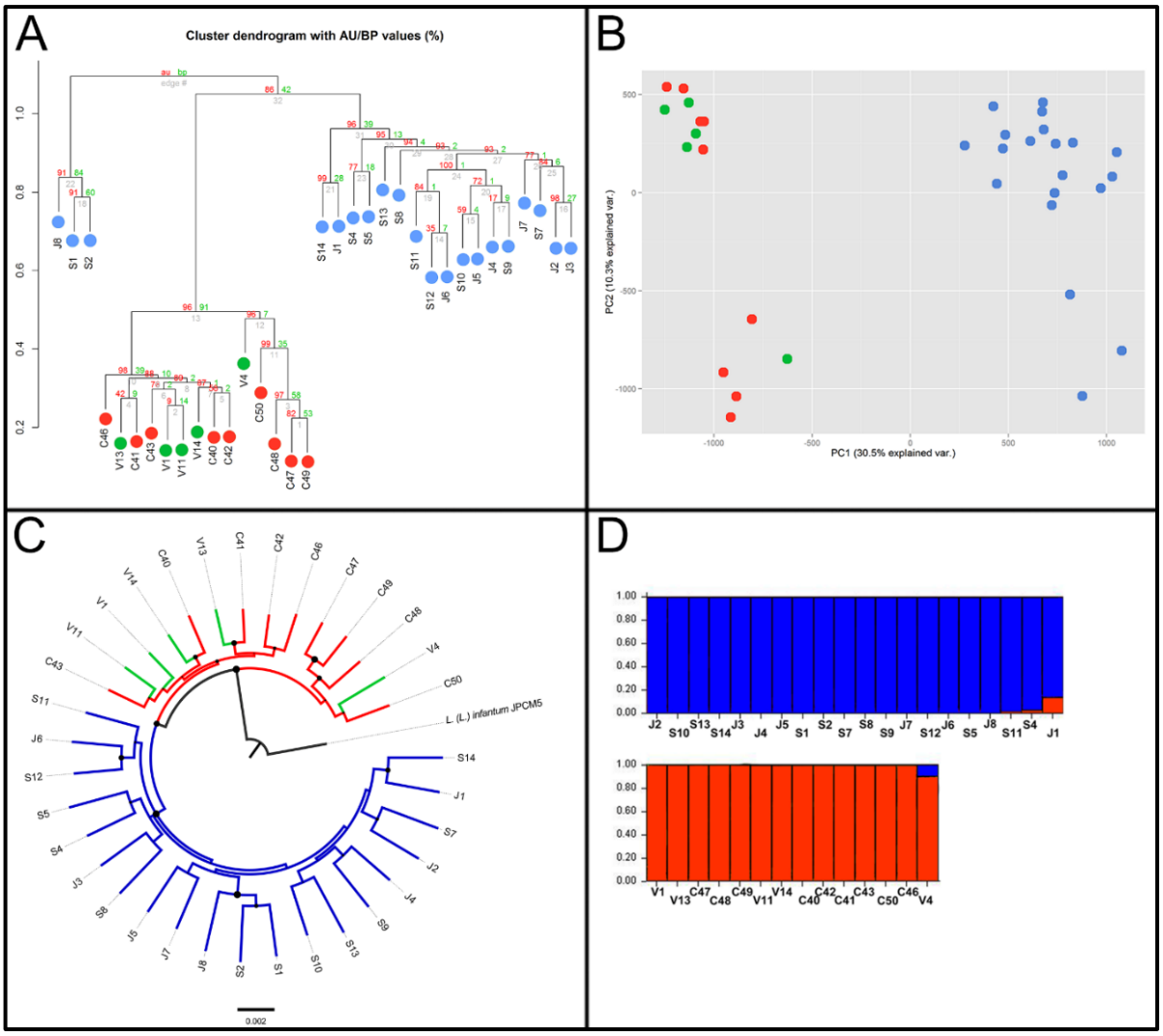
Daily precipitation and temperature data from 2008-2014 were obtained from the Instituto Nacional de Meteorologia (INMET). Mean daily values were obtained and used for cumulative and 3d surface plots on R.

In order to evaluate if there were substantial variation of precipitation through these years we employed TREND for statistical analysis of annual and seasonal rainfall time series using the Worsley likelihood ratio test.

Results

Phylogenetic and clustering analysis show the presence of two distinct sub-populations in Governador Valadares

We assessed the relatedness of the Governador Valadares isolates using 3,950 polymorphic sites using PCA, maximum likelihood, hierarchical clustering and Structure **(Chapter 3: Figure 2)**. The best-fit model for the data according to the Akaike and Bayesian Information Criteria (AIC and BIC) and Decision Theory method (DT) was the GTR (generalised time reversible) substitution model with four rate categories. This model was used on maximum likelihood analysis and our results provided statistical reliability to basal nodes of the tree with bootstrap values of 1000 and support a closer relationship among the Governador Valadares isolates than to the Old World *L. (L.) infantum* reference. Moreover, our findings show two well-defined sub-populations of *L. (L.) infantum* that correspond to two isolation periods (before 2015 and since 2015). The strength of the association between time of isolation and sub-population structure is statistically significant by Fisher exact test ($p<0.001$) whereas we did not see any association with geographical location of the isolates.



Chapter 3: Figure 2

(A) Hierarchical clustering (HC) with 1,000 bootstraps, the y-axis denotes the closeness of either individual point. (B) Principal component analysis (PCA), x and y axes show the first and second principal components, respectively. (C) Maximum likelihood phylogeny, size of black dots represent bootstrap support. (D) Structure results.
Figure shows that samples from 2008 and 2012 cluster together while samples isolated in 2015 cluster into different group.
Green and red colors in figures A, B and C represent samples from 2008 and 2012, respectively.
Red color bars in figure D represent samples from 2008 and 2012.
Blue denote samples collected in 2015.

There are fixed SNPs in coding sequences that differentiate among both populations

Given the presence of two sub-populations, we sought to assess the presence and location of fixed SNPs in each group. In this sense we found 86 and 45 SNPs that were exclusively present in the 2015 and the 2008-2012 sub-populations, respectively. In the 2015 sub-population, there were 30 SNPs located in 22 coding regions that comprise hypothetical proteins, phosphatases and transport

proteins. The gene that gathered the highest number of SNPs in the 2015 group corresponds to a ferric iron reductase whose sequence is altered in 6 amino acids. In the 2008-2012 population there were 15 SNPs affecting nine CDS that include surface antigen proteins, kinesins and hypothetical proteins **(Chapter 3: Table 1)**.

2015

Gene IDS	Annotation	SNPs
LinJ.34.0240, LinJ.08.0360, LinJ.09.0501, LinJ.12.0570, LinJ.14.0370, LinJ.17.1010, LinJ.24.0290, LinJ.24.1000, LinJ.26.1410, LinJ.29.2240, LinJ.29.2840, LinJ.35.3920	hypothetical protein	1 SNP each
LinJ.19.0800	ABC transport system ATP-binding protein (ABCF2)	
LinJ.34.1720	amastin-like surface protein	
LinJ.14.0020	phosphatidylinositol 3-kinase 2, putative	
LinJ.07.0330	protein kinase, putative	
LinJ.24.0260	protein phosphotase, putative	
LinJ.28.0690	protein transport protein SEC13, putative	
LinJ.22.0300	kinesin K39	2
LinJ.30.1630	ferric reductase transmembrane protein, putative (FR1)	8

2008-2012

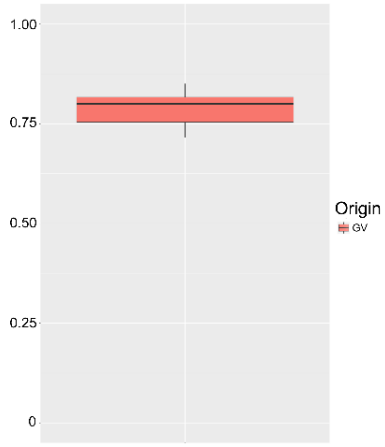
Gene IDS	Annotation	SNPs
LinJ.31.1660	3-ketoacyl-CoA thiolase-like	3
LinJ.12.0663	surface antigen protein 2	3
LinJ.14.1180	kinesin K39	1
LinJ.29.1600	Hypothetical protein	1
LinJ.20.0440	Metallo-dependent phosphatase-like	1
LinJ.22.0660	5'a2rel-related	1
LinJ.06.1360	Hypothetical	2
LinJ.34.1720	amastin-like surface protein	2
LinJ.34.2660	amastin-like surface protein	1

Chapter 3: Table 1

Fixed SNPs in CDS for the sub-populations of 2015 and 2008-2012. The table indicates the corresponding gene identifier, the respective annotation and the number of SNPs.

FWs shows lower within-host heterozygosity consistent

To characterize within-host diversity we used the FWs statistics using 9,119 biallelic sites. This metric that ranges from 0 to 1 evaluates the probability of parasites carrying different alleles at a given locus. Moreover, it captures the overall diversity within an individual and the similarity among the population. In the specific case of *Leishmania*, this metric will also be an indirect reflect of the levels of mosaic aneuploidy in the population. We observe that the Governador Valadares population presented FWs of 0.8 **(Chapter 3: Figure 3)** that serves as an indicator of low risk of outcrossing and low within-host diversity.



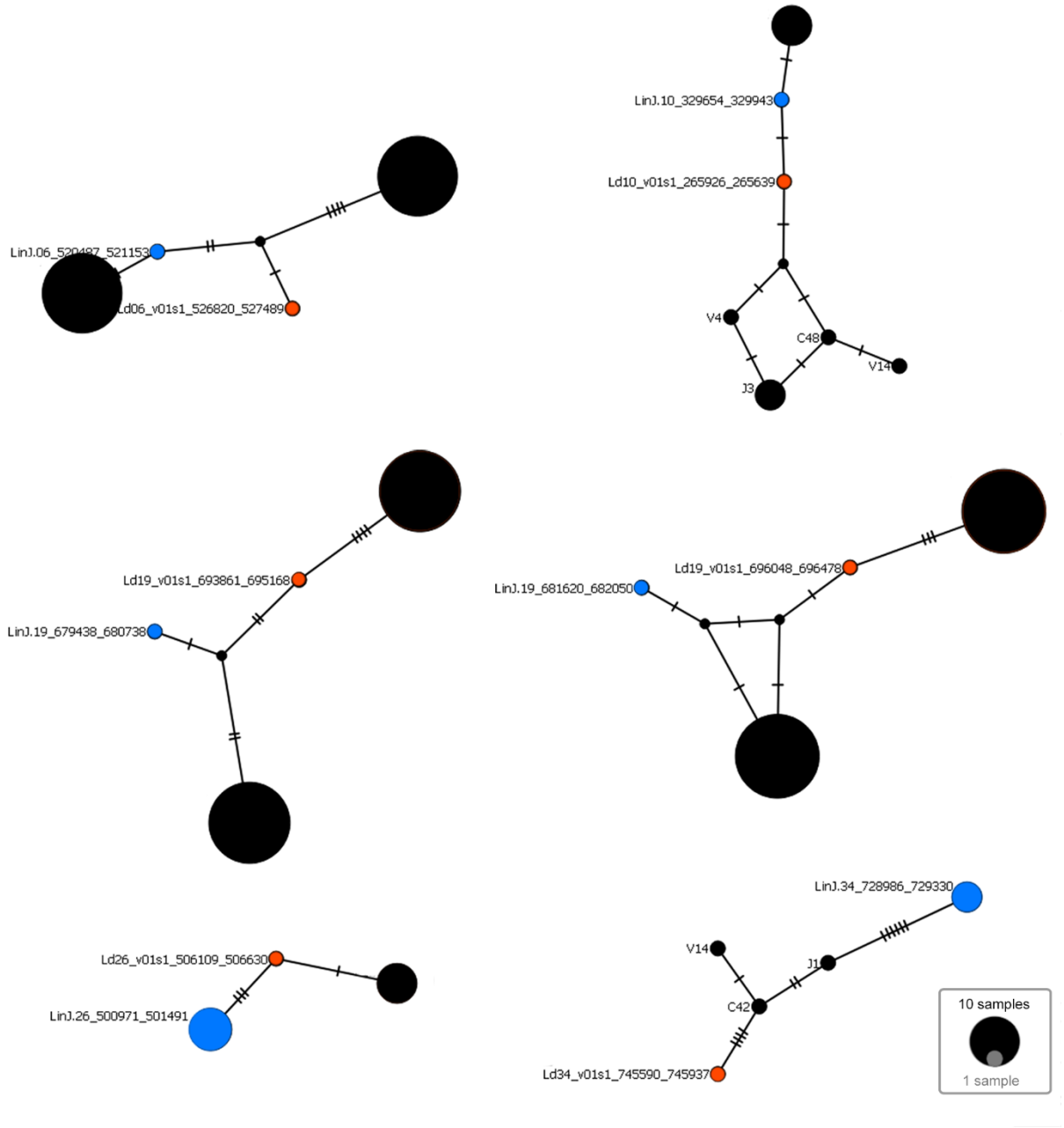
Chapter 3: Figure 3

Boxplot showing the distribution of FWs in the Governador Valadares isolates

Recombination analysis

To explore the ancestry of the Governador Valadares *L. (L.) infantum* isolates we reconstructed the haplotypes of phased heterozygous regions of the genome and compared them with one ancestor represented by the Old World JPCM5 line and another by the *L. (L.) donovani* BPK282 isolate based on a previous study (Auburn *et al.*, 2012). The inference of haplotypes was restricted to regions with equal or more than 4 phased heterozygous positions in at least one Governador Valadares isolate. Based on these criteria, six regions were selected and analyzed phylogenetically.

Our results show that two distinct clades among the haplotypes of the *L. (L.) infantum* isolates **(Chapter 3: Figure 4)**. One clade is similar to the JPCM5 line whereas the other is somewhat related to the *L. (L.) donovani* BPK282 reference.



Chapter 3: Figure 4

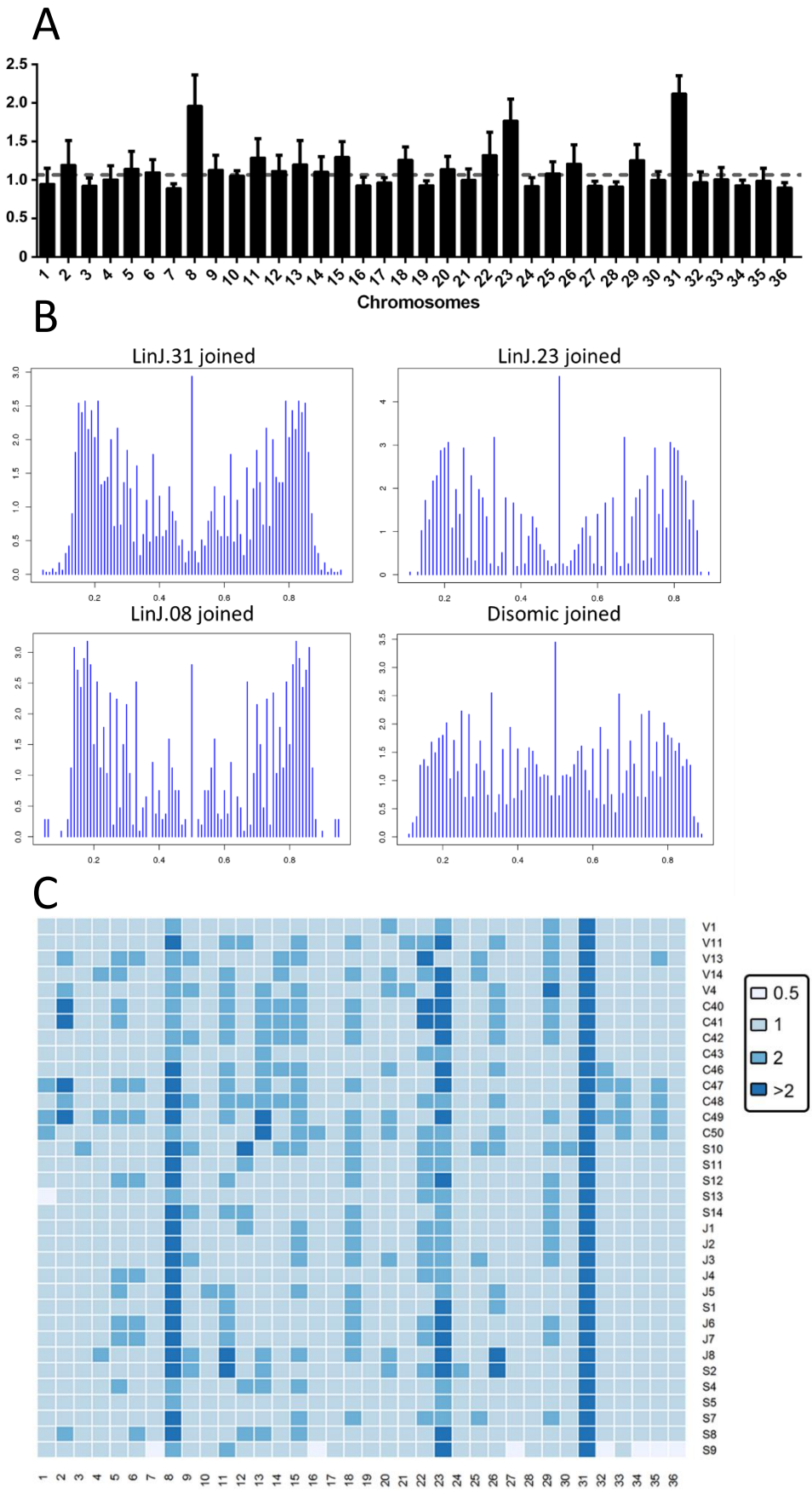
Median joining networks of phased haplotypes for six regions show two evolutionary lineages belonging to *L. (L.) infantum* and *L. (L.) donovani*. Size of the circles correspond to the number of similar samples in the respective node of the network. Blue and orange circles indicate the position of the *L. (L.) infantum* and *L. (L.) donovani* reference sequences, respectively.

The Governador Valadares population presents a heterogeneous pattern of chromosomal and gene copy number amplifications with an overall disomic tendency

Aneuploidy has been extensively described in *Leishmania* (Rogers *et al.*, 2011; Rogers *et al.*, 2014; Valdivia, Reis-Cunha*, et al.*, 2015) and we have also observed it in the Governador Valadares population **(Chapter 3: Figure 5 A,B,C).** Allele frequency and read depth based analyses allowed us to predict the ploidy of all chromosomes in the population.

In this analysis only three out of 36 chromosomes were diploid in all the GV isolates (17,19, 28) and this number increases to seven (7, 17, 19, 27, 28, 34, 36) if we take out isolate S9 **(Chapter 3: Figure 5 C)**. Our results show that chromosome 31 has the highest aneuploidy been pentasomic in all isolates. This chromosome has been previously shown to be expanded in all *Leishmania* species so far sequenced and has been suggested to confer some selective advantage to the parasite(Valdivia, Reis-Cunha*, et al.*, 2015). Other chromosomes that are expanded on most isolates are chromosome 8 and 23 whose ploidies range from tetrasomy to pentasomy **(Chapter 3: Figure 5 C)**.

Importantly, read depth on these chromosomes is uniformly distributed refuting region-specific amplification **(Chapter 3: Figure 5 D)**. This highly heterogeneous pattern has been also shown by other studies and indicates that some degree of aneuploidy is common among *L. (L.) infantum* and the rest of the genus (Rogers *et al.*, 2014).


A
Chromosomes
B
LinJ.31 joined
LinJ.23 joined
LinJ.08 joined
Disomic joined
C
0.5
1
2
>2

D

Chapter 3: Figure 5

Chromosome copy number analyses. Figures show read depth and allele frequency based estimations of chromosome ploidy in the Governador Valadares isolates.

A) Chromosome copy number (CCN) in the Governador Valadares *L. (L.) infantum* population. This graph shows the mean and standard deviation of the CCN estimated from the read depth analysis. A dotted line indicates the median haploid ploidy for all the genomes. The y-axis denotes the haploid CCN and the x-axis the respective chromosome.

B) Allele frequency of expanded and disomic chromosomes for all isolates. This figure shows tetrasomic and pentasomic profiles for chromosome 31; trisomic and tetrasomic profiles for chromosome 27 and chromosome 8 as well as a clear disomic trend for the rest of the chromosomes.

C) Heat map of individual ploidies for each sample. This figure shows the haploid CCN at the individual level, the y-axis indicates the isolate and the x-axis the chromosome.

D) Mean read depth at each chromosome, this figure shows the mean read depth for all isolates representing regions with increased read depth as red.

Gene copy number variations

Results from the CNV analyses were separated into expanded genes present in all the isolates tested and the ones expanded in each sub-populations **(Chapter 3: Table 2)**.

We only found two and one sub-population expanded genes in the 2008-2012 and 2015 groups, respectively. These genes are annotated as elongation factor 1-alpha, phosphoglycan beta 1,3 (SCG6) and paraflagellar rod protein. EF1-alpha has been described on the *L. (L.) amazonensis* chapter as a protein that favors parasite survival inside macrophages. The side chains galactose genes are a glycosylation gene family crucial for midgut attachment that modifies the phosphoglycan repeats of lipophosphoglycan (LPG)(Sacks *et al.*, 2000). These modified phosphoglycans promote the survival of the parasite within the vertebrate and invertebrate host and confer the ability of the natural sand fly

vector to transmit *Leishmania* (Dobson *et al.*, 2003). The paraflagellar rod proteins are a network of filaments that are located along the axoneme of trypanosomatids and are required for cell motility (Maga *et al.*, 1999).

CNV variations in *Leishmania* may contribute with the different tropism that is seen in this genus. In this sense, we aimed to explore expanded genes in all *L. (L.) isolates* to get a deeper understanding of species-specific expansions that may be only present in New World *L. (L.) infantum.* We found up to 40 expanded genes in all isolates distributed in 13 chromosomes.

Among the most expanded proteins we found a glucose transporter, an rrp6p homologue, many hypothetical proteins, amastins, GP63, among others. As it has been mentioned in the previous chapters, members from the amastin, GP63 and cysteine peptidase gene families have important roles as virulence factors permitting infection, immune evasion and parasite survival inside the vertebrate and invertebrate hosts. The other hypothetical proteins in this set remain to be characterized.

A

Subpopulation	Gene ID	Annotation	Normalized mean HCN
2008-2012	LinJ.25.2570	elongation factor 1-alpha	2.4
2008-2012	LinJ.17.0110	phosphoglycan beta 1,3 galactosyltransferase 6 (SCG6)	2.2
2015	LinJ.29.1880	paraflagellar rod protein 1D, putative	2.2

B

Gene ID	Annotation	Normalized mean HCN
LinJ.36.6550	glucose transporter 2 (GT2)	2.6
LinJ.34.4330	exosome subunit rrp6p homologue, putative	2.5
LinJ.34.4320	phosphatidylinositol-4-phosphate-5-kinase-likep ro tein	2.7
LinJ.34.4310	hypothetical protein, conserved	2.6
LinJ.34.4300	hypothetical protein, conserved	2.9
LinJ.34.4290	lipophosphoglycan biosynthetic protein (lpg2) (LPG2)	2.8
LinJ.34.2660	amastin-like surface protein, putative	4.4
LinJ.34.2650	amastin-like surface protein, putative	3.2
LinJ.34.2390	hypothetical protein, conserved	2.4
LinJ.34.1730	amastin-like surface protein, putative	9.8

LinJ.34.1680	amastin-like surface protein, putative	3.5
LinJ.33.0860	beta-tubulin	2.8
LinJ.33.0370	heat shock protein 83 (HSP83-3)	4.7
LinJ.33.0360	heat shock protein 83 (HSP83-2)	4.8
LinJ.31.1660	3-ketoacyl-CoA thiolase-like protein, putative	2.9
LinJ.29.2240	hypothetical protein, conserved	2.2
LinJ.29.1890	paraflagellar rod protein 1D, putative	2.3
LinJ.28.2060	Zinc transporter 3B (ZIP3B)	3.2
LinJ.28.2050	Zinc transporter 3A (ZIP3A)	3.4
LinJ.23.1060	beta-fructosidase-like protein, invertase-like protein, sucrose hydrolase-like protein	2.6
LinJ.22.0300	hypothetical protein	8.1
LinJ.21.2240	beta tubulin	3.2
LinJ.17.0200	elongation factor 1-alpha	2.8
LinJ.17.0190	elongation factor 1-alpha	2.6
LinJ.17.0170	elongation factor 1-alpha	2.8
LinJ.17.0100	elongation factor 1-alpha	2.8
LinJ.17.0090	elongation factor 1-alpha	2.8
LinJ.15.0730	hypothetical protein	3.4
LinJ.15.0490	hypothetical protein (pseudogene)	4.2
LinJ.13.1460	alpha tubulin	3.8
LinJ.13.0330	alpha tubulin	3.8
LinJ.10.0530	GP63, leishmanolysin, metallo-peptidase, Clan MA(M), Family M8 (GP63-4)	2.6
LinJ.10.0521	hypothetical protein, unknown function	2.8
LinJ.10.0520	GP63, leishmanolysin, metallo-peptidase, Clan MA(M), Family M8 (GP63-3)	3.8
LinJ.10.0510	GP63, leishmanolysin, metallo-peptidase, Clan MA(M), Family M8 (GP63-3)	2.2
LinJ.10.0500	GP63, leishmanolysin, metallo-peptidase, Clan MA(M), Family M8 (GP63-2)	7.7
LinJ.10.0490	GP63, leishmanolysin (GP63-1)	7.1
LinJ.08.1280	beta tubulin	1.9
LinJ.08.0960	cathepsin L-like protease	3.9
LinJ.08.0950	cathepsin L-like protease	2.2

Chapter 3: Table 2

CNV in Governador Valadares isolates. This table shows expanded genes in in the distinct sub-populations (A) and gene expansions in all isolates (B).

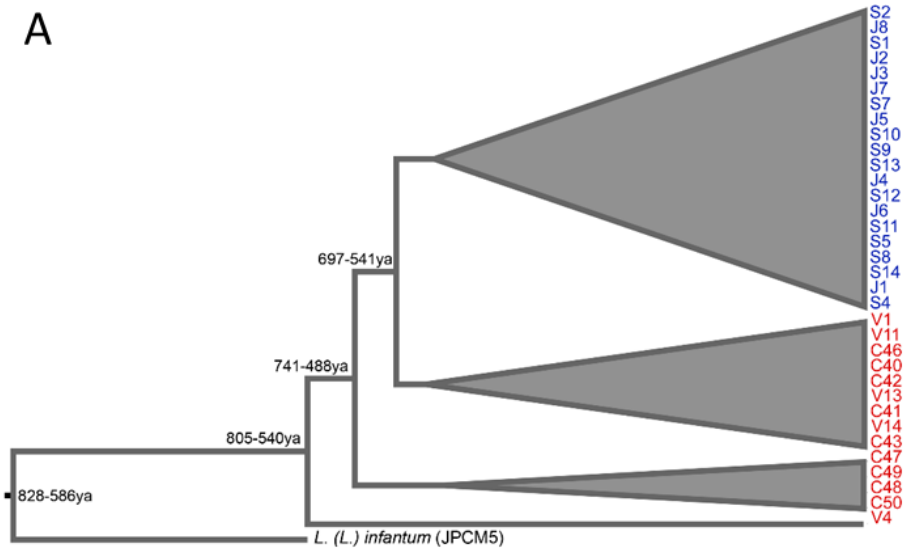
Divergence and skyline analysis suggest clonal replacement rather than in situ divergence and stabilization of the effective population size

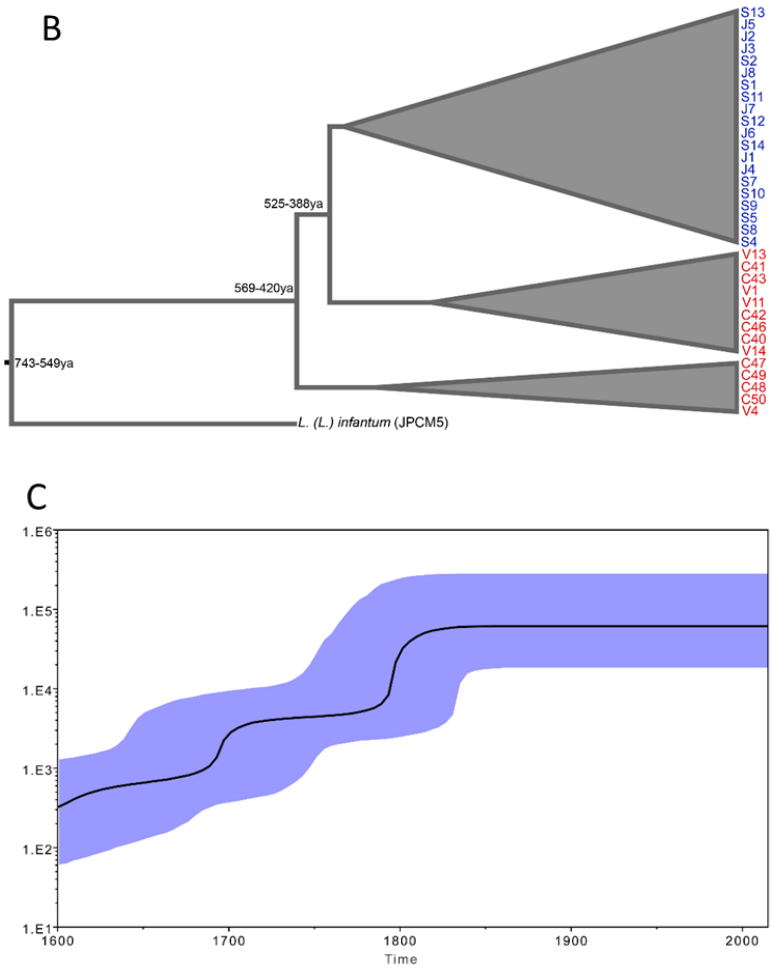
Bayesian based divergence time analysis using the strict and relaxed clock models resulted in fairly similar dates. The ages of divergences were inferred from 3,950 non-biallelic genotyped sites along

the genomes and calibrated against the timing of the split between Old World and New World *L. (L.) infantum*.

Dates estimated with this Bayesian approach points towards an earlier divergence of both populations after the introduction of *L. (L.) infantum* into the New World **(Chapter 3: Figure 6 A, B)**. In order to determine the parasite population history, we have constructed a Bayesian skyline plot **(Chapter 3: Figure 6 C)**. This plot shows that the Governador Valadares *L. (L.) infantum* population might have expanded since the 1600 up to 1800. Furthermore, the effective population size was estimated in 6.1 x 10^4 (95% HPD; 1.8x10^4-28x10^4) and the time of the most recent common ancestor between both populations was inferred between 697-541 and 525-388 years ago according to the relaxed and the strict clock models, respectively **(Chapter 3: Figure 6 A, B)**. This value is a reflect of the single geographic location under investigation and is similar to values obtained in Old World populations (Rogers *et al.*, 2014).

In summary, the time of divergence confirms that the changes we are seen in the *L. (L.) infantum* genotypes circulating in Governador Valadares are due to clonal replacement rather than recent in-situ divergence.



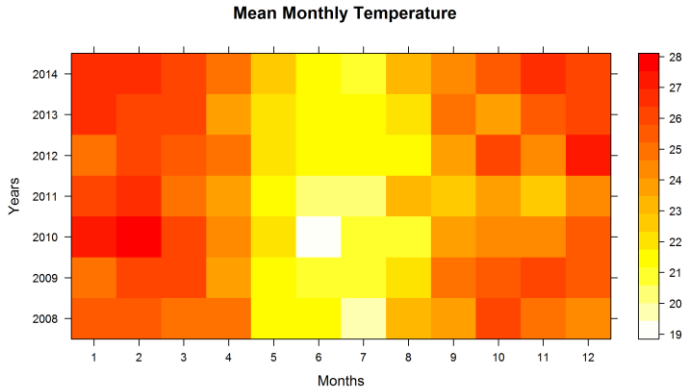


Chapter 3: Figure 6

Bayesian dating of *L. (L.) infantum* isolates. Figures A and B show the trees corresponding to the relaxed and strict clock models. Nodes are located at the mean divergence and numbers represent 95% confidence intervals. The result shows an earlier divergence of both *L. (L.) infantum* sub-populations in Governador Valadares. Figure C presents the Bayesian skyline plot showing changes in the effective population size through time. The Y-axis describes the Log scale population size whereas the X-axis show the time in years.

Climate data shows a significant change in precipitation during 2014

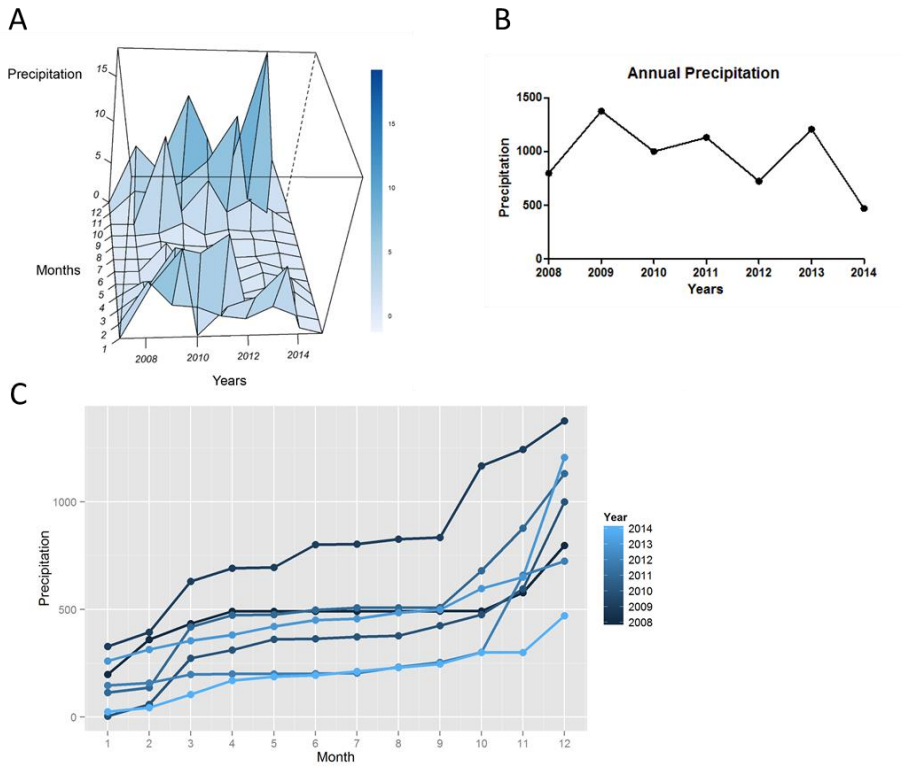
Analysis of daily temperature collected since 2008 up to 2014 by the INMET shows that temperature in Governador Valadares presents two distinct phases along the year **(Chapter 3: Figure 7)**. There is one colder season that starts in May and ends in August and a warmer season from September to April. However, we were not able to find any statistical significant difference in temperature cycles during these years.



Chapter 3: Figure 7

Heatmap of mean monthly temperatures since 2008 to 2014. Figure shows the presence of two cycles of colder and warmer temperatures in this region.

In terms of precipitation analysis, we found that precipitation also appears to be cyclic with a rainy season that starts in September and move up to March **(Chapter 3: Figure 8 A)**. Cumulative and annual precipitation plots suggest a decrease in the total rainfall in 2014 **(Chapter 3: Figure 8** Chapter 3: Figure 8**B,C)**. In order to assess if this difference in total rainfall are significant we compared annual precipitations since 2008 up to 2014 using he Worsley likelihood ratio test implemented in TREND. The result of this analysis shows a statistical significant result ($p<0.05$) for changes in the mean precipitation of 2014 compared to the other years evaluated.



Chapter 3: Figure 8

Precipitation data in Governador Valadares. Figure show the monthly (A), annual (B) and cumulative precipitation (C) in Governador Valadares since 2008 to 2014. The data shows a decay in the total rainfall during 2014.

Discussion

Among the leishmaniasis, VL is considered as the most aggressive form of these diseases and can be deadly if untreated (Murray, 2004; Murray *et al.*, 2005). VL is a serious problem in Brazil that is among the six countries that hold more than 90% of all VL cases in the world (Bangladesh, Brazil, Ethiopia, India, Nepal and Sudan) (Alvar *et al.*, 2012).

In Brazil VL is caused by *L. (L.) infantum* that is transmitted by the bite of infected *Lutzomyia longipalpis* sand flies been dogs the main reservoirs of the disease. Currently there is increased concern about the urbanization of VL in Brazil and the effects of climate change and human migration (Harhay *et al.*, 2011). These factors may have led to the emerging of foci of transmission in distinct cities in the country and have posed a challenge for control programs in the country.

The analysis of the Governador Valadares isolates represents to our knowledge the first genome-wide analysis of New World *Leishmania* and has increased our understanding of the *Leishmania* population structure that deserve to be explored in other endemic sites in the New World.

The city of Governador Valadares is an emerging focus of transmission of VL with both, *L (L.) infantum* and *L. (L.) amazonensis* species circulating in the area causing similar clinical symptomatology in infected dogs.

Our population structure analyses show the presence of two distinct sub-populations of *L. (L.) infantum* that are associated with the time of isolation. Importantly divergence dating on these isolates showed that both subpopulations diverged hundreds of years ago with no apparent crossing among them. This result supports recent clonal replacement in this city and discard the effects of in-situ divergence.

Within-host diversity in the parasite population is relatively low that is consistent with the predominant clonal mode of reproduction in *Leishmania*. However, FWs the value obtained of 0.8 indicates some level of recombination. It is worth noting that the FWs metric may be biased in *Leishmania* due to the effects of mosaic aneuploidy.

A previous study conducted with Old World *L. (L.) infantum* isolates found the presence of two distinct ancestor lineages, one similar to the JPCM5 *L. (L) infantum* line and the other closely related to *L. (L.) donovani* (Rogers *et al.*, 2014). Haplotype reconstruction at highly heterozygous regions conducted by us in Governador Valadares also corroborates this finding and show that New World *L. (L.) infantum* isolates share this common hybrid ancestry.

Chromosome and gene copy number variations are a major source of intra and interspecific variability in *Leishmania* and have been suggested to be an adaptation that favors parasitism and reflects a constitutive (unregulated) transcription in *Leishmania* (Rogers *et al.*, 2011; Sterkers *et al.*, 2012; Valdivia, Reis-Cunha*, et al.*, 2015). The GV population present a highly diverse pattern of ploidies with extensive variability among isolates regardless of their clustering in the two subpopulations described and the very recent divergence inside each subpopulation.

An important consequence of our findings is the rapid change in the predominant *L. (L.) infantum* population in GV. This rapid replacement may be due to significant lower rainfall that occurred in 2014 that may have affected the distribution of vectors in the city favoring one sub-population rather than the other. Another possibility that should be considered is the introduction of a different strain of *L. (L.) infantum* into this city. In this sense, further studies are needed to explore the effect that changes in rainfall or other variables have in the vector-reservoir-parasite triad.

GENERAL DISCUSSION

Leishmaniasis is a major health problem in endemic countries in tropical and sub-tropical areas and has been linked to poverty and little access to adequate health care. Control of these diseases has been difficult to achieve as evidenced by outbreaks in several countries (Silveira *et al.*, 2002; Velez *et al.*, 2012; Arce *et al.*, 2013; Uranw *et al.*, 2013) and an increase of endemic areas to 98 countries.

These diseases are characterized by a wide range of clinical symptoms associated to the more than 20 species of *Leishmania* pathogenic to human. This high species diversity is thought to be caused by gradual accumulation of mutations rather than sexual recombination. This premise is supported by the clonal theory of *Leishmania* that has been proposed a long time ago and affirms that asexual reproduction is the sole or main reproductive mechanism in this parasite (Tibayrenc e Ayala, 2002). This theory was founded on the high similarity among the offspring and their parents and other criteria like genetic markers and fixed heterozygosity (Tibayrenc *et al.*, 1993).

However, different lines of evidence have emerged that challenge this theory and suggest that sexual reproduction can occur in *Leishmania*. For instance, recent population studies have shown that *Leishmania* can experience sexual recombination events within and between species with distinct haplotypes coming from both parents (Rogers *et al.*, 2014; Seblova *et al.*, 2015; Kato *et al.*, 2016). This finding is further supported by evidence of *Leishmania* hybrids generated under controlled experiments as well as putative hybrids found in the nature that present different chromosomal contributions from each parent (Romano *et al.*, 2014; Seblova *et al.*, 2015; Kato *et al.*, 2016). Furthermore, evidence from a genomic population study shows a recent recombination event between *L. (L.) infantum* and *L. (L.) donovani* (Rogers *et al.*, 2014). Although further studies are needed to assess the relevance and frequency of sexual reproduction in *Leishmania*, these data challenge the traditional view of *Leishmania* as exclusively clonal parasite. Moreover, the finding of inter-species sexual recombination makes more complicated an accurate classification of species and adds an important cause of variability that has been largely overlooked.

During the last decade and thanks to the development of high throughput sequencing technologies, many genomic projects have been executed in *Leishmania*. These efforts have resulted in the availability of massive amounts of data for the scientific community. In this context, the growing number of genomic, transcriptomic and proteomic data has opened opportunities for broader and more extensive analysis that have increased our understanding of these complex parasites.

One of the most important characteristics revealed by genomics is the high degree of genomic conservation and synteny in *Leishmania* with a low number of species-specific genes despite a time of divergence of 36-46 million years between New World and Old World species (Peacock *et al.*, 2007).

This information has led to explore other mechanisms that could explain the distinct clinical manifestations, distribution, vectors and reservoirs that are seen in endemic regions. In the chapters presented in this thesis, we have used genomic and phylogenomic approaches aiming at a better understanding of the complexity of *Leishmania* and reveal some of the basis of variability.

The case of *L. (V.) peruviana* and *L. (V.) braziliensis* has been debated for a long time in the scientific community. These organisms are so closely related phylogenetically to the point that they were considered by some as a single species (Odiwuor *et al.*, 2012; Fraga *et al.*, 2013). Nevertheless, they are associated with different clinical symptomatology, vectors, reservoirs and distribution (Lainson *et al.*, 1979; Llanos-Cuentas *et al.*, 1999). Our results presented in chapter one strongly support maintaining the current classification as distinct species and show for the first time the high variability in chromosome and gene copy numbers in *L. (V.) peruviana*.

Further studies including a greater number of isolates of *L. (V.) peruviana* and *L. (V.) braziliensis* are needed to confirm the stability of CCN, CNV and SNPs seen in our study.

In this regard, previous evidence mainly from studies in Old World *Leishmania*, have shown that genome plasticity is an important source of heterogeneity at the intra and inter species level (Downing *et al.*, 2011; Rogers *et al.*, 2011; Rogers *et al.*, 2014). The analysis of genome structure in single cells using fluorescence in situ hybridization (FISH) and the use of next-generation sequencing

technologies have shown a wide range of variations in chromosomal content from cell to cell in a population generating an intra-strain genomic heterogeneity (Sterkers *et al.*, 2012) (Sterkers *et al.*, 2011). This high variation has been referred as mosaic aneuploidy and is proposed as a powerful adaptive mechanism in *Leishmania* to cope with the highly variable selective pressures in the vector and vertebrate hosts (Sterkers *et al.*, 2014).

Our results have also shown this high variability among closely related *L. (V.) peruviana* and *L. (V.) braziliensis* and even among more phylogenetically close isolates like the *L. (L.) infantum* from Governador Valadares. The theoretical wide range of ploidies seen on these isolates may confer a selective advantage for those individuals with the best-fit ploidy for distinct conditions contributing to parasite survival (Sterkers *et al.*, 2012). However additional studies are needed to confirm this premise given the potential deleterious effects of aneuploidy.

Additionally, further studies are needed to address the dynamics of mosaic aneuploidy during the interaction of the parasite with the vector and vertebrate hosts and assess the levels of ploidy in primary samples to rule out the effects of long term culturing.

Although there are not major differences in terms of gene content in the genomes of *Leishmania*, few studies have been conducted to explore the organization of gene families in these species and assess expansions and diversity among orthologous groups.

Identification of homolog genes is a critical step to understand the evolutionary context of an organism given the fact that duplication events can result in posterior functional divergence (Gabaldon *et al.*, 2009). For this reason, it is important to develop strong methods for accurate prediction of homology.

Traditional homology prediction approaches rely solely on sequence similarity and while they are fast and efficient, are also prone to errors in the homology assessment (Eisen e Wu, 2002). This occurs specially when there are multiple matches for the target gene as in the case of highly expanded families. For this reason, phylogenomic approaches constitute a robust platform that overcome

sequence similarity based methods and can provide additional information from a phylogenetic context.

Our results from the phylogenomics and comparative genomic studies showed extensive variability in the number of expanded genes among species and within strains. Importantly, several gene duplications and species-specific expansions affect gene families that mediate invasion, immune evasion and parasite survival (Valdivia, Scholte*, et al.*, 2015). Consequently, these duplications might be important contributors for the diversity and tropism seen in *Leishmania*. This extensive variability in multicopy genes have been reported in other studies across many *Leishmania* species (Rogers *et al.*, 2011) and may respond to the lack of transcriptional control in *Leishmania* acting as a mechanism to increase gene dosage of key genes.

Although there is some evidence retrieved from an RNA-sequencing study of increased RNA levels in expanded families (Rastrojo *et al.*, 2013), further studies are needed to assess the correlation between DNA and RNA levels and changes in expression among members of expanded families.

In spite of the increasing use of "OMICS" approaches, most genes associated to Leishmania virulence remain uncharacterized; limiting our understanding of the disease and hampering an efficient use of these results. Following this line, we have selected members of the Cathepsin-L family from our phylogenomics study for functional characterization.

We have chosen this family based on the finding of species-specific expansions in *L. (L.) mexicana*, *L. (L.) major* and *L. (V.) braziliensis*, increased RNA levels during the amastigote stage and the presence of related cysteine peptidases with characterized immunomodulatory roles in Th1 suppression.

We are currently generating knockout and superexpressors parasites whose pattern of infectivity will be compared with WT parasites in in vitro and in vivo experiments. Given our phylogenomics results for this family, we expect to find related functions in members of the *Viannia* subgenus.

Improved functional characterization of these and other genes with an increased focus on hypothetical, uncharacterized and species-specific expansions will provide useful insights for vaccine

design, novel therapies and increase our understanding of the mechanisms of *Leishmania* infection and immune evasion.

Currently, there is increasing concern regarding the effects that climate change, human activities and co-infection might have on the leishmaniasis. For instance, migration and urbanization with their implicit effects (deforestation, establishment of settings near the forest and breeding of animals) have led to an increase of TL and VL cases. Additionally, there is growing evidence of expansion of sand fly vectors, pathogenic *Leishmania* species and the re-emergence of infections in previously controlled areas (Desjeux, 2004; Arce *et al.*, 2013; Barata *et al.*, 2013).

It has been demonstrated that an improved knowledge of the population structure of pathogens can provide insights into transmission patterns, epidemiological features and insight into diagnostic development. In this sense the use of population genomics can be a valuable ally in the study of leishmaniasis in emergent and re-emergent foci.

The region of Governador Valadares is a re-emergent focus of leishmaniasis regardless the intense control efforts in this area. Our study in this city has revealed a dramatic shift in the *L. (L.) infantum* population in less than two years. Although we have not yet conclusive evidence, it is possible that changes in precipitation during 2014 have influenced on the sand fly population in the area and in turn selected strains from one parasite population over the other.

Regardless of the causes of this variation, the evidence of rapid clonal replacement is an important finding and underscores the need to consider other variables like climate change and migration in the study of the leishmaniasis as well as the potential effects of this rapid population change.

Previous evidence from a spatial-temporal analyses in French Guiana show a negative correlation between rainfall and the number of TL infections. This finding suggest that rainfall could be an indicator for risk in endemic sites with an increase in the number of cases of *Leishmania* after 2 months of relative decrease in rainfall (Roger *et al.*, 2013). Contrastingly, in the study described on the third chapter we did not find this correlation with the number cases. These contradictory results reveal the complexity of the leishmaniasis at each endemic site.

The finding of *L. (L.) amazonensis* is also another important consequence of our research, complementing previous studies that have isolated this species from previously unreported areas in Brazil in different ecological niches than the one originally described (Tolezano *et al.*, 2007; Oliveira *et al.*, 2015). This result also highlights the lack of studies in the New World regarding the distribution of *Leishmania*. For instance, previous evidence from Peru suggest a high diversity of species with areas of co-occurrence in the Andean and Amazon regions that has not been explored in other countries (Lucas *et al.*, 1998).

It is known that some vectors can transmit only certain species of *Leishmania* due to differences in proteolytic enzymes in the gut, parasite inability to escape from the peritrophic matrix, or inefficient midgut attachment (Pimenta *et al.*, 1994). In this regard, there is an urgent need to identify the putative vector of *L. (L.) amazonensis* in these new foci given the challenge to control activities due to differences in sand fly distribution and infectivity. Additionally, it is also crucial to determine if there are permissive vectors in these foci that could contribute to the spread of both *L. (L.) infantum* and *L. (L.) amazonensis*.

The presence of *L. (L.) amazonensis* in dogs from a domestic environment is also critical from an epidemiological point given the fact that dogs are the most important reservoirs of VL in domestic environments maintaining the parasite population and acting as a source of transmission.

Finally, the similar clinical signs from the ones shown in *L. (L.) infantum* may indicate the potential of visceralization of this species. This hypothesis is supported by the finding of *L. (L.) amazonensis* from dogs of a peridomestic environment in the Sao Paulo state with clinical signs undistinguishable from VL (Tolezano *et al.*, 2007). Importantly, that study has also shown issues with the specificity of serological tests used for VL. This cross reactivity may have led to misreporting of *L. (L.) amazonensis* and an underrepresentation of the distribution and cases associated to this species.

The importance and complexity of the leishmaniasis herein shown demand more integrated approaches to understand disease transmission and develop efficient intervention strategies, especially in emergent and re-emergent foci like Governador Valadares. In this sense, the use of

mathematical modelling, niche modelling and population genomics appear as promising areas of research.

We believe that these approaches can complement each other and in turn will allow a better understanding of the transmission of leishmaniasis in endemic sites. The information that can be provided by this endeavors is greatly needed in endemic settings to assess the impact of vector and reservoir control on disease prevalence and on the parasite population. Moreover, these efforts could offer the possibility of designing better control strategies and predict the potential effects of their implementation (Rock *et al.*, 2015).

CONCLUSIONS

In the context of *Leishmania* genomics, we have explored distinct *Leishmania* species through comparative genomics, phylogenomics and population genomics. In this process, we have:

- Detected important differences in SNPs, gene and chromosome copy numbers in *L. (V.) peruviana* and *L. (V.) braziliensis*.
- Provided support for the current classification of *L. (V.) peruviana* as a distinct species from *L. (V.) braziliensis*.
- Identified species-specific adaptations in pathogenic *Leishmania* species potentially acting on virulence factors and revealing the evolutionary history of these parasites.
- Show variation in species-specific expansions across various *Leishmania* species and the potential importance of these differences.
- Report the first detection of *L. (L.) amazonensis* in Governador Valadares.
- Identified *L. (L.) amazonensis* specific genomic features.
- Detected high levels of heterogeneity in chromosome copy numbers in a population of *L. (L.) infantum* in Governador Valadares.
- Evaluate the presence of two distinct sub-populations of *L. (L.) infantum* and a process of rapid clonal replacement.

BIBLIOGRAPHY

AKUFFO, H. et al. Leishmania aethiopica derived from diffuse leishmaniasis patients preferentially induce mRNA for interleukin-10 while those from localized leishmaniasis patients induce interferon-gamma. **J Infect Dis,** v. 175, n. 3, p. 737-41, Mar 1997. ISSN 0022-1899 (Print)

0022-1899 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/9041358 >.

ALVAR, J. et al. Leishmaniasis worldwide and global estimates of its incidence. **PLoS One,** v. 7, n. 5, p. e35671, 2012. ISSN 1932-6203 (Electronic)

1932-6203 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/22693548 >.

ARANA, M. et al. Biochemical characterization of Leishmania (Viannia) braziliensis and Leishmania (Viannia) peruviana by isoenzyme electrophoresis. **Trans R Soc Trop Med Hyg,** v. 84, n. 4, p. 526-9, Jul-Aug 1990. ISSN 0035-9203 (Print)

0035-9203 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/2091345 >.

ARCE, A. et al. Re-emergence of leishmaniasis in Spain: community outbreak in Madrid, Spain, 2009 to 2012. **Euro Surveill,** v. 18, n. 30, p. 20546, 2013. ISSN 1560-7917 (Electronic)

1025-496X (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/23929177 >.

ASLETT, M. et al. TriTrypDB: a functional genomic resource for the Trypanosomatidae. **Nucleic Acids Res,** v. 38, n. Database issue, p. D457-62, Jan 2010. ISSN 1362-4962 (Electronic)

0305-1048 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/19843604 >.

AUBURN, S. et al. Characterization of within-host Plasmodium falciparum diversity using next-generation sequence data. **PLoS One,** v. 7, n. 2, p. e32891, 2012. ISSN 1932-6203 (Electronic)

1932-6203 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/22393456 >.

BANULS, A. L. et al. Is Leishmania (Viannia) peruviana a distinct species? A MLEE/RAPD evolutionary genetics answer. **J Eukaryot Microbiol,** v. 47, n. 3, p. 197-207, May-Jun 2000. ISSN 1066-5234 (Print)

1066-5234 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/10847336 >.

BANULS, A. L.; HIDE, M.; PRUGNOLLE, F. Leishmania and the leishmaniases: a parasite genetic update and advances in taxonomy, epidemiology and pathogenicity in humans. **Adv Parasitol,** v. 64, p. 1-109, 2007. ISSN 0065-308X (Print)

0065-308X (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/17499100 >.

BARATA, R. A. et al. Epidemiology of visceral leishmaniasis in a reemerging focus of intense transmission in Minas Gerais State, Brazil. **Biomed Res Int,** v. 2013, p. 405083, 2013. ISSN 2314-6141 (Electronic). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/24000322 >.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics,** v. 30, n. 15, p. 2114-20, Aug 1 2014. ISSN 1367-4811 (Electronic)

1367-4803 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/24695404 >.

BOUCKAERT, R. et al. BEAST 2: a software platform for Bayesian evolutionary analysis. **PLoS Comput Biol,** v. 10, n. 4, p. e1003537, Apr 2014. ISSN 1553-7358 (Electronic)

1553-734X (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/24722319 >.

BRITTO, C. et al. Conserved linkage groups associated with large-scale chromosomal rearrangements between Old World and New World Leishmania genomes. **Gene,** v. 222, n. 1, p. 107-17, Nov 5 1998. ISSN 0378-1119 (Print)

0378-1119 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/9813266 >.

CAPELLA-GUTIERREZ, S.; SILLA-MARTINEZ, J. M.; GABALDON, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. **Bioinformatics,** v. 25, n. 15, p. 1972-3, Aug 1 2009. ISSN 1367-4811 (Electronic)

1367-4803 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/19505945 >.

CRISCUOLO, A. morePhyML: improving the phylogenetic tree space exploration with PhyML 3. **Mol Phylogenet Evol,** v. 61, n. 3, p. 944-8, Dec 2011. ISSN 1095-9513 (Electronic)

1055-7903 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/21925283 >.

DARRIBA, D. et al. jModelTest 2: more models, new heuristics and parallel computing. **Nat Methods,** v. 9, n. 8, p. 772, Aug 2012. ISSN 1548-7105 (Electronic)

1548-7091 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/22847109 >.

DAVID, C. V.; CRAFT, N. Cutaneous and mucocutaneous leishmaniasis. **Dermatol Ther,** v. 22, n. 6, p. 491-502, Nov-Dec 2009. ISSN 1529-8019 (Electronic)

1396-0296 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/19889134 >.

DESJEUX, P. Leishmaniasis: current situation and new perspectives. **Comp Immunol Microbiol Infect Dis,** v. 27, n. 5, p. 305-18, Sep 2004. ISSN 0147-9571 (Print)

0147-9571 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/15225981 >.

DOBSON, D. E. et al. Functional identification of galactosyltransferases (SCGs) required for species-specific modifications of the lipophosphoglycan adhesin controlling Leishmania major-sand fly interactions. **J Biol Chem,** v. 278, n. 18, p. 15523-31, May 2 2003. ISSN 0021-9258 (Print)

0021-9258 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/12604613 >.

DOWNING, T. et al. Whole genome sequencing of multiple Leishmania donovani clinical isolates provides insights into population structure and mechanisms of drug resistance. **Genome Res,** v. 21, n. 12, p. 2143-56, Dec 2011. ISSN 1549-5469 (Electronic)

1088-9051 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/22038251 >.

EDGAR, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. **BMC Bioinformatics,** v. 5, p. 113, Aug 19 2004. ISSN 1471-2105 (Electronic)

1471-2105 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/15318951 >.

EISEN, J. A.; WU, M. Phylogenetic analysis and gene functional predictions: phylogenomics in action. **Theor Popul Biol,** v. 61, n. 4, p. 481-7, Jun 2002. ISSN 0040-5809 (Print)

0040-5809 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/12167367 >.

FRAGA, J. et al. Evolution and species discrimination according to the Leishmania heat-shock protein 20 gene. **Infect Genet Evol,** v. 18, p. 229-37, Aug 2013. ISSN 1567-7257 (Electronic)

1567-1348 (Linking).

GABALDON, T. et al. Joining forces in the quest for orthologs. **Genome Biol,** v. 10, n. 9, p. 403, 2009. ISSN 1465-6914 (Electronic)

1465-6906 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/19785718 >.

GOTO, H.; LINDOSO, J. A. Current diagnosis and treatment of cutaneous and mucocutaneous leishmaniasis. **Expert Rev Anti Infect Ther,** v. 8, n. 4, p. 419-33, Apr 2010. ISSN 1744-8336 (Electronic)

1478-7210 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/20377337 >.

HARHAY, M. O. et al. Urban parasitology: visceral leishmaniasis in Brazil. **Trends Parasitol,** v. 27, n. 9, p. 403-9, Sep 2011. ISSN 1471-5007 (Electronic)

1471-4922 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/21596622 >.

IVENS, A. C. et al. The genome of the kinetoplastid parasite, Leishmania major. **Science,** v. 309, n. 5733, p. 436-42, Jul 15 2005. ISSN 1095-9203 (Electronic)

0036-8075 (Linking).

KARAGIANNIS-VOULES, D. A. et al. Bayesian geostatistical modeling of leishmaniasis incidence in Brazil. **PLoS Negl Trop Dis,** v. 7, n. 5, p. e2213, 2013. ISSN 1935-2735 (Electronic)

1935-2727 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/23675545 >.

KATO, H.; CACERES, A. G.; HASHIGUCHI, Y. First Evidence of a Hybrid of Leishmania (Viannia) braziliensis/L. (V.) peruviana DNA Detected from the Phlebotomine Sand Fly Lutzomyia tejadai in Peru. **PLoS Negl Trop Dis,** v. 10, n. 1, p. e0004336, Jan 2016. ISSN 1935-2735 (Electronic)

1935-2727 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/26735142 >.

KATO, H. et al. Molecular epidemiology for vector research on leishmaniasis. **Int J Environ Res Public Health,** v. 7, n. 3, p. 814-26, Mar 2010. ISSN 1660-4601 (Electronic)

1660-4601 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/20617005 >.

KATOH, K. et al. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. **Nucleic Acids Res,** v. 30, n. 14, p. 3059-66, Jul 15 2002. ISSN 1362-4962 (Electronic)

0305-1048 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/12136088 >.

KILLICK-KENDRICK, R. Phlebotomine vectors of the leishmaniases: a review. **Med Vet Entomol,** v. 4, n. 1, p. 1-24, Jan 1990. ISSN 0269-283X (Print)

0269-283X (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/2132963 >.

KILLICK-KENDRICK, R. et al. Leishmania in phlebotomid sandflies. V. The nature and significance of infections of the pylorus and ileum of the sandfly by leishmaniae of the braziliensis complex. **Proc R Soc Lond B Biol Sci,** v. 198, n. 1131, p. 191-9, Aug 22 1977. ISSN 0950-1193 (Print)

0950-1193 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/20640 >.

LAINSON, R.; READY, P. D.; SHAW, J. J. Leishmania in phlebotomid sandflies. VII. On the taxonomic status of Leishmania peruviana, causative agent of Peruvian 'uta', as indicated by its development in the sandfly, Lutzomyia longipalpis. **Proc R Soc Lond B Biol Sci,** v. 206, n. 1164, p. 307-18, Dec 31 1979. ISSN 0950-1193 (Print)

0950-1193 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/43976 >.

LAINSON, R.; SHAW, J. **The role of animals in the epidemiology of South American leishmaniasis**. 1979.

LAINSON, R. et al. **Evolution, classification and geographical distribution**. Academic Press, 1987. ISBN 0125521014.

LAINSON, R.; WARD, R. D.; SHAW, J. J. Leishmania in phlebotomid sandflies: VI. Importance of hindgut development in distinguishing between parasites of the Leishmania mexicana and L. braziliensis complexes. **Proc R Soc Lond B Biol Sci,** v. 199, n. 1135, p. 309-20, Nov 14 1977. ISSN 0950-1193 (Print)

0950-1193 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/22860 >.

LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. **Nat Methods,** v. 9, n. 4, p. 357-9, Apr 2012. ISSN 1548-7105 (Electronic)

1548-7091 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/22388286 >.

LI, H. et al. The Sequence Alignment/Map format and SAMtools. **Bioinformatics,** v. 25, n. 16, p. 2078-9, Aug 15 2009. ISSN 1367-4811 (Electronic)

1367-4803 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/19505943 >.

LLANOS-CUENTAS, E. A. et al. Natural infections of Leishmania peruviana in animals in the Peruvian Andes. **Trans R Soc Trop Med Hyg,** v. 93, n. 1, p. 15-20, Jan-Feb 1999. ISSN 0035-9203 (Print)

0035-9203 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/10492779 >.

LUCAS, C. M. et al. Geographic distribution and clinical description of leishmaniasis cases in Peru. **Am J Trop Med Hyg,** v. 59, n. 2, p. 312-7, Aug 1998. ISSN 0002-9637 (Print)

0002-9637 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/9715953 >.

LUKES, J. et al. Evolutionary and geographical history of the Leishmania donovani complex with a revision of current taxonomy. **Proc Natl Acad Sci U S A,** v. 104, n. 22, p. 9375-80, May 29 2007. ISSN 0027-8424 (Print)

0027-8424 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/17517634 >.

MAGA, J. A. et al. Genetic dissection of the Leishmania paraflagellar rod, a unique flagellar cytoskeleton structure. **J Cell Sci,** v. 112 (Pt 16), p. 2753-63, Aug 1999. ISSN 0021-9533 (Print)

0021-9533 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/10413682 >.

MANSKE, M. et al. Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing. **Nature,** v. 487, n. 7407, p. 375-9, Jul 19 2012. ISSN 1476-4687 (Electronic)

0028-0836 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/22722859 >.

MCKENNA, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. **Genome Res,** v. 20, n. 9, p. 1297-303, Sep 2010. ISSN 1549-5469 (Electronic)

1088-9051 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/20644199 >.

MIMORI, T. et al. Identification, using isoenzyme electrophoresis and monoclonal antibodies, of Leishmania isolated from humans and wild animals of Ecuador. **Am J Trop Med Hyg,** v. 40, n. 2, p. 154-8, Feb 1989. ISSN 0002-9637 (Print)

0002-9637 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/2919726 >.

MURRAY, H. W. Progress in the treatment of a neglected infectious disease: visceral leishmaniasis. **Expert Rev Anti Infect Ther,** v. 2, n. 2, p. 279-92, Apr 2004. ISSN 1478-7210 (Print)

1478-7210 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/15482193 >.

MURRAY, H. W. et al. Advances in leishmaniasis. **Lancet,** v. 366, n. 9496, p. 1561-77, Oct 29-Nov 4 2005. ISSN 1474-547X (Electronic)

0140-6736 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/16257344 >.

MYLER, P. J. et al. Leishmania major Friedlin chromosome 1 has an unusual distribution of protein-coding genes. **Proc Natl Acad Sci U S A,** v. 96, n. 6, p. 2902-6, Mar 16 1999. ISSN 0027-8424 (Print)

0027-8424 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/10077609 >.

ODIWUOR, S. et al. Evolution of the Leishmania braziliensis species complex from amplified fragment length polymorphisms, and clinical implications. **Infect Genet Evol,** v. 12, n. 8, p. 1994-2002, Dec 2012. ISSN 1567-7257 (Electronic)

1567-1348 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/22516226 >.

OLIVEIRA, E. F. et al. Leishmania amazonensis DNA in wild females ofLutzomyia cruzi (Diptera: Psychodidae) in the state of Mato Grosso do Sul, Brazil. **Mem Inst Oswaldo Cruz,** v. 110, n. 8, p. 1051-7, Dec 2015. ISSN 1678-8060 (Electronic)

0074-0276 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/26602870 >.

PEACOCK, C. S. et al. Comparative genomic analysis of three Leishmania species that cause diverse human disease. **Nat Genet,** v. 39, n. 7, p. 839-47, Jul 2007. ISSN 1061-4036 (Print)

1061-4036 (Linking).

PIMENTA, P. F. et al. Evidence that the vectorial competence of phlebotomine sand flies for different species of Leishmania is controlled by structural polymorphisms in the surface lipophosphoglycan. **Proc Natl Acad Sci U S A,** v. 91, n. 19, p. 9155-9, Sep 13 1994. ISSN 0027-8424 (Print)

0027-8424 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/8090785 >.

PRITCHARD, J. K.; STEPHENS, M.; DONNELLY, P. Inference of population structure using multilocus genotype data. **Genetics,** v. 155, n. 2, p. 945-59, Jun 2000. ISSN 0016-6731 (Print)

0016-6731 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/10835412 >.

QUEIROZ, A. et al. Association between an emerging disseminated form of leishmaniasis and Leishmania (Viannia) braziliensis strain polymorphisms. **J Clin Microbiol,** v. 50, n. 12, p. 4028-34, Dec 2012. ISSN 1098-660X (Electronic)

0095-1137 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/23035200 >.

RASTROJO, A. et al. The transcriptome of Leishmania major in the axenic promastigote stage: transcript annotation and relative expression levels by RNA-seq. **BMC Genomics,** v. 14, p. 223, 2013. ISSN 1471-2164 (Electronic)

1471-2164 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/23557257 >.

REAL, F. et al. The genome sequence of Leishmania (Leishmania) amazonensis: functional annotation and extended analysis of gene models. **DNA Res,** v. 20, n. 6, p. 567-81, Dec 2013. ISSN 1756-1663 (Electronic)

1340-2838 (Linking).

RIOUX, J. A. et al. Taxonomy of Leishmania. Use of isoenzymes. Suggestions for a new classification. **Ann Parasitol Hum Comp,** v. 65, n. 3, p. 111-25, 1990. ISSN 0003-4150 (Print)

0003-4150 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/2080829 >.

ROCK, K. S. et al. Uniting mathematics and biology for control of visceral leishmaniasis. **Trends Parasitol,** v. 31, n. 6, p. 251-9, Jun 2015. ISSN 1471-5007 (Electronic)

1471-4922 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/25913079 >.

ROGER, A. et al. Climate and leishmaniasis in French Guiana. **Am J Trop Med Hyg,** v. 89, n. 3, p. 564-9, Sep 2013. ISSN 1476-1645 (Electronic)

0002-9637 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/23939706 >.

ROGERS, M. B. et al. Genomic confirmation of hybridisation and recent inbreeding in a vector-isolated Leishmania population. **PLoS Genet,** v. 10, n. 1, p. e1004092, Jan 2014. ISSN 1553-7404 (Electronic)

1553-7390 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/24453988 >.

ROGERS, M. B. et al. Chromosome and gene copy number variation allow major structural change between species and strains of Leishmania. **Genome Res,** v. 21, n. 12, p. 2129-42, Dec 2011. ISSN 1549-5469 (Electronic)

1088-9051 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/22038252 >.

ROMANO, A. et al. Cross-species genetic exchange between visceral and cutaneous strains of Leishmania in the sand fly vector. **Proc Natl Acad Sci U S A,** v. 111, n. 47, p. 16808-13, Nov 25 2014. ISSN 1091-6490 (Electronic)

0027-8424 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/25385616 >.

SACKS, D. L. et al. The role of phosphoglycans in Leishmania-sand fly interactions. **Proc Natl Acad Sci U S A,** v. 97, n. 1, p. 406-11, Jan 4 2000. ISSN 0027-8424 (Print)

0027-8424 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/10618431 >.

SEBLOVA, V. et al. Natural hybrid of Leishmania infantum/L. donovani: development in Phlebotomus tobbi, P. perniciosus and Lutzomyia longipalpis and comparison with non-hybrid strains differing in tissue tropism. **Parasit Vectors,** v. 8, p. 605, 2015. ISSN 1756-3305 (Electronic)

1756-3305 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/26608249 >.

SILVEIRA, F. T. et al. An outbreak of cutaneous leishmaniasis among soldiers in Belem, Para State, Brazil, caused by Leishmania (Viannia) lindenbergi n. sp. A new leishmanial parasite of man in the Amazon region. **Parasite,** v. 9, n. 1, p. 43-50, Mar 2002. ISSN 1252-607X (Print)

1252-607X (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/11938695 >.

SINHA, S. et al. Diffuse cutaneous leishmaniasis associated with the immune reconstitution inflammatory syndrome. **Int J Dermatol,** v. 47, n. 12, p. 1263-70, Dec 2008. ISSN 1365-4632 (Electronic)

0011-9059 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/19126013 >.

STERKERS, Y. et al. Parasexuality and mosaic aneuploidy in Leishmania: alternative genetics. **Trends Parasitol,** v. 30, n. 9, p. 429-35, Sep 2014. ISSN 1471-5007 (Electronic)

1471-4922 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/25073852 >.

STERKERS, Y. et al. Novel insights into genome plasticity in Eukaryotes: mosaic aneuploidy in Leishmania. **Mol Microbiol,** v. 86, n. 1, p. 15-23, Oct 2012. ISSN 1365-2958 (Electronic)

0950-382X (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/22857263 >.

STERKERS, Y. et al. FISH analysis reveals aneuploidy and continual generation of chromosomal mosaicism in Leishmania major. **Cell Microbiol,** v. 13, n. 2, p. 274-83, Feb 2011. ISSN 1462-5822 (Electronic)

1462-5814 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/20964798 >.

TEAM, R. C. **R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2012**: ISBN 3-900051-07-0 2014.

TIBAYRENC, M.; AYALA, F. J. The clonal theory of parasitic protozoa: 12 years on. **Trends Parasitol,** v. 18, n. 9, p. 405-10, Sep 2002. ISSN 1471-4922 (Print)

1471-4922 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/12377258 >.

TIBAYRENC, M. et al. Leishmania and the clonal theory of parasitic protozoa. **Arch Inst Pasteur Tunis,** v. 70, n. 3-4, p. 375-82, Jul-Oct 1993. ISSN 0020-2509 (Print)

0020-2509 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/7802492 >.

TOLEZANO, J. E. et al. The first records of Leishmania (Leishmania) amazonensis in dogs (Canis familiaris) diagnosed clinically as having canine visceral leishmaniasis from Aracatuba County, Sao Paulo State, Brazil. **Vet Parasitol,** v. 149, n. 3-4, p. 280-4, Nov 10 2007. ISSN 0304-4017 (Print)

0304-4017 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/17720321 >.

TSUKAYAMA, P.; LUCAS, C.; BACON, D. J. Typing of four genetic loci discriminates among closely related species of New World Leishmania. **Int J Parasitol,** v. 39, n. 3, p. 355-62, Feb 2009. ISSN 1879-0135 (Electronic)

0020-7519 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/18817779 >.

URANW, S. et al. An outbreak investigation of visceral leishmaniasis among residents of Dharan town, eastern Nepal, evidence for urban transmission of Leishmania donovani. **BMC Infect Dis,** v. 13, p. 21, 2013. ISSN 1471-2334 (Electronic)

1471-2334 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/23327548 >.

VALDIVIA, H. O. et al. Comparative genomic analysis of Leishmania (Viannia) peruviana and Leishmania (Viannia) braziliensis. **BMC Genomics,** v. 16, p. 715, 2015. ISSN 1471-2164 (Electronic)

1471-2164 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/26384787 >.

VALDIVIA, H. O. et al. The Leishmania metaphylome: a comprehensive survey of Leishmania protein phylogenetic relationships. **BMC Genomics,** v. 16, p. 887, 2015. ISSN 1471-2164 (Electronic)

1471-2164 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/26518129 >.

VAN DER AUWERA, G. et al. Evaluation of four single-locus markers for Leishmania species discrimination by sequencing. **J Clin Microbiol**, Jan 22 2014. ISSN 1098-660X (Electronic)

0095-1137 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/24452158 >.

VELEZ, I. D. et al. An epidemic outbreak of canine cutaneous leishmaniasis in Colombia caused by Leishmania braziliensis and Leishmania panamensis. **Am J Trop Med Hyg,** v. 86, n. 5, p. 807-11, May 2012. ISSN 1476-1645 (Electronic)

0002-9637 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/22556078 >.

WINCKER, P. et al. The Leishmania genome comprises 36 chromosomes conserved across widely divergent human pathogenic species. **Nucleic Acids Res,** v. 24, n. 9, p. 1688-94, May 1 1996. ISSN 0305-1048 (Print)

0305-1048 (Linking). Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/8649987 >.