UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA
TESE DE DOUTORADO

WANESSA MOREIRA GOES

SEQUENCIAMENTO E CARACTERIZAÇÃO DO GENOMA DA CEPA PH8 DE *LEISHMANIA AMAZONENSIS* COM ÊNFASE EM FATORES DE VIRULÊNCIA POSSIVELMENTE ENVOLVIDOS NO ESTABELECIMENTO E VISCERALIZAÇÃO DA INFECÇÃO

BELO HORIZONTE

2022

WANESSA MOREIRA GOES

SEQUENCIAMENTO E CARACTERIZAÇÃO DO GENOMA DA CEPA PH8 DE *LEISHMANIA AMAZONENSIS* COM ÊNFASE EM FATORES DE VIRULÊNCIA POSSIVELMENTE ENVOLVIDOS NO ESTABELECIMENTO E VISCERALIZAÇÃO DA INFECÇÃO

Tese de Doutorado apresentada ao Programa Interunidades de Pós-graduação em Bioinformática da Universidade Federal de Minas Gerais como pré-requisito para obtenção do título de Doutor em Bioinformática.

Orientadora: Prof[a]. Santuza Maria R Teixeira

Coorientadora: Prof[a]. Daniella Castanheira Bartholomeu

BELO HORIZONTE

2022

043 Goes, Wanessa Moreira.
Sequenciamento e caracterização do genoma da cepa PH8 de Leishmania amazonensis com ênfase em fatores de virulência possivelmente envolvidos no estabelecimento e visceralização da infecção [manuscrito] / Wanessa Moreira Goes. – 2022.
137 f. : il. ; 29,5 cm.

Orientadora: Profa. Dra. Santuza Maria R. Teixeira. Coorientadora: Profa. Daniella Castanheira Bartholomeu.
Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Leishmania. 3. Genoma. 4. Família Multigênica. 5. Fatores de Virulência. I. Teixeira, Santuza Maria Ribeiro. II. Bartholomeu, Daniella Castanheira. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 573:004

Ficha catalográfica elaborada pela bibliotecária Fabiane C M Reis – CRB 6 - 2680

UNIVERSIDADE FEDERAL DE MINAS GERAIS INSTITUTO DE CIÊNCIAS BIOLÓGICAS PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

FOLHA DE APROVAÇÃO

"Sequenciamento e caracterização do genoma da cepa PH8 de *Leishmania amazonensis* com ênfase em fatores de virulência possivelmente envolvidos no estabelecimento e visceralização da infecção"

Wanessa Moreira Goes

Tese aprovada pela banca examinadora cons tuída pelos Professores: Profª Santuza Maria Ribeiro Teixeira - Orientadora (Universidade Federal de Minas Gerais), Profª Daniella Castanheira Bartholomeu Coorientadora (Universidade Federal de Minas Gerais), Profª Angela Kaysel Cruz (Faculdade de Medicina de Ribeirão Preto/USP), Prof. Wanderson Duarte da Rocha (Universidade Federal do Paraná), Profª Camila Indiani de Oliveira (Ins tuto Gonçalo Moniz-Fiocruz/BA) e Prof. Gabriel da Rocha Fernandes (Ins tuto René Rachou-Fiocruz/MG)

Belo Horizonte, 20 de janeiro de 2023.

Documento assinado eletronicamente por **Aristoteles Goes Neto**, **Coordenador(a) de curso de pósgraduação**, em 09/02/2023, às 08:42, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A auten cidade deste documento pode ser conferida no site h ps://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2073114** e o código CRC **CCD17F83**.

AGRADECIMENTOS

À Profª Drª Santuza Maria Ribeiro Teixeira por ter me aceitado como aluna, pela orientação em todas as etapas deste trabalho, pelo incentivo na busca do conhecimento, pela contribuição em minha formação profissional e pessoal e pelos bons momentos vividos dentro do laboratório e fora deste.

À Profª. Drª Daniella Castanheira Bartholomeu pela coorientação durante o desenvolvimento deste trabalho, por toda a experiência em bioinformática transferida, pelo apoio, preocupação e incentivo durante todo o tempo.

À Profª Drª Ana Paula Fernandes pela colaboração na anotação de famílias multigênicas, principalmente proteínas A2, e também na contextualização deste trabalho. Além de toda a troca de conhecimento que foi essencial para o desenvolvimento deste trabalho.

Aos meus colegas, Ms. Anderson Coqueiro-dos-Santos e Dr. João Luís Reis-Cunha que deram uma enorme contribuição nas etapas de montagem e anotação do genoma da cepa PH8 de *Leishmania amazonensis*. Agradeço pela contribuição na escrita do artigo derivado deste trabalho, pelo incentivo e humildade em sempre estarem dispostos a ajudar.

Ao meu colega Carlos Rodolpho Ferreira Brasil, pela colaboração na anotação de famílias multigênicas, principalmente genes que codificam cinases em *Leishmania amazonensis*.

À Dra. Maria Fernanda Laranjeira da Silva pela colaboração na anotação de genes relacionados ao metabolismo de heme e ferro e por toda a troca de conhecimento acerca do tema trabalhado nesta tese.

Aos meus colegas de laboratório, Nailma Silva Aprígio dos Santos, Carlos Alberto de Almeida Júnior, Marina Batista, Renata Barbosa Peixoto, Jennifer Ottino, Juan Macedo e Jéssica pela companhia durante todo o período em que permaneci no laboratório.

Aos meus pais, Marivone Moreira dos Santos e Marcos Linhares Goes, pelo apoio incondicional, pelo incentivo e pela contribuição financeira e enriquecimento científico. Ao meu irmão Allyson Moreira Goes pela contribuição na construção de algumas figuras deste trabalho.

À Francielle Resende de Oliveira pela ajuda na formatação e correção gramatical do trabalho, apoio emocional e por ter estado comigo em todas as horas.

Ao programa de Pós-Graduação em Bioinformática da UFMG, em especial à Sheila e Tiago, pelo carinho, pela disponibilidade e pelo esforço para solucionar todos os problemas que apareceram durante o período do Doutorado.

À agência financiadora CAPES pela bolsa, que me auxiliou durante a minha estadia em Belo Horizonte e na realização deste trabalho.

Ao INCTV pelo financiamento do sequenciamento Pacbio que deu origem a todo o trabalho desenvolvido, juntamente com o sequenciamento Illumina financiado pelo grupo da Profª Drª Daniella Bartholomeu.

EPÍGRAFE

“Descobrir consiste em olhar para o que todo mundo está vendo e pensar uma coisa diferente”

Roger Von Oech

RESUMO

Leishmania amazonensis é um dos agentes etiológicos da leishmaniose cutânea, uma doença que apresenta 21 mil casos/ano no Brasil. Diferentes moléculas do parasito já foram estudadas por desempenharem um papel crucial na invasão e estabelecimento da infecção no hospedeiro mamífero, incluindo a visceralização de formas amastigotas, o que contribui para o aumento da patogênese da leishmaniose. Para aprofundar os estudos sobre esses fatores de virulência, é necessária a obtenção de genomas completos e com anotação adequada de diferentes cepas e isolados de *L. amazonensis*. Apesar de ter sido descrito em 2013, o genoma desse parasito não foi ainda completamente sequenciado, montado e anotado. No presente trabalho, relatamos o sequenciamento e a montagem do genoma da cepa PH8 de *L. amazonensis* utilizando uma estratégia baseada na combinação de *reads* de cobertura longa obtidos na plataforma PacBio, *reads* de Illumina com cobertura curta e ainda dados de sintenia com o genoma de *Leishmania mexicana*. Os *contigs* iniciais foram gerados usando apenas as *reads* PacBio e o montador Canu, e o pipeline IPA foi usado para remover *contigs* redundantes. A etapa de *scaffolding* foi executada usando SSPACE e o preenchimento de gaps com GapFiller, com base em *reads* curtas *paired-end* Illumina. Finalmente, a montagem foi polida usando Pilon, e os *scaffolds* foram ordenados com base nos cromossomos de *L. mexicana* usando Abacas. A montagem final, composta por 34 pseudocromossomos e 42 *scaffolds* não incorporados representa um genoma de ~32 Mb, além da sequência do maxicírculo de 18,1 kb. Em seguida, foi analisada a ocorrência de aneuploidias para vários cromossomos da cepa PH8 e a presença de expansões gênicas relacionadas à genes que codificam fatores de virulência como as amastinas, GP63 e proteínas cinases. Sequências repetitivas foram descritas pela primeira vez revelando a predominância de retroelementos e transposons de DNA. A anotação do conteúdo gênico de *L. amazonensis* foi conduzida utilizando duas abordagens: *ab initio* e baseada na transferência da anotação de um total de 8.2317 genes presentes no genoma de *L. mexicana*. Destes, 7999 são genes codificadores de proteínas e 318 deles classificados como pseudogenes. Diversas famílias multigênicas que codificam fatores de virulência, como proteínas A2, amastinas, metaloproteases GP63, fosfatases e cisteíno proteases, foram identificadas e comparadas com sua anotação no genoma de outras espécies de tripanosomatídeos. O genoma da cepa PH8 possui 29 genes que codificam todas as quatro subclasses de amastinas (α, β, γ e δ), 5 genes que codificam antígenos A2, 9 genes que codificam a metaloprotease GP63 e 76 genes de cisteíno proteases. Como foram recentemente reconhecidos como fatores de virulência essenciais para o estabelecimento da doença e progressão da infecção, foram também identificados 14 genes que codificam proteínas envolvidas no metabolismo de ferro e de heme do parasita, os quais foram comparados com o repertório gênico relacionado à essas vias presente em outros tripanosomatídeos. Com base nesse estudo genômico e em estudos recentes do nosso grupo descrevendo novos métodos de edição de genomas, abre-se a perspectiva para, através de genômica funcional, aprofundar o conhecimento sobre esses e outros fatores de virulência e, como consequência, acelerar o desenvolvimento de novas estratégias de controle da infecção causada por *L. amazonensis*.

Palavras-chave: *Leishmania amazonensis*, montagem, genoma, famílias multigênicas, fatores de virulência.

ABSTRACT

Leishmania amazonensis is one of the etiologic agents of cutaneous leishmaniasis, a disease with 21,000 cases/year in Brazil. Different molecules of the parasite have already been studied because they play a crucial role in the invasion and establishment of infection in the mammalian host, including the visceralization of amastigotes, which contributes to the increase in the pathogenesis of leishmaniasis. To deepen the studies on these virulence factors, it is necessary to obtain complete genomes with adequate annotations of different strains and isolates of *L. amazonensis*. Despite having been described in 2013, the genome of this parasite has not yet been completely sequenced, assembled and annotated. In the present work, we report the sequencing and assembly of the genome of the PH8 strain of *L. amazonensis* using a strategy based on the combination of long coverage reads obtained on the PacBio platform, reads from Illumina with short coverage and also synteny data with the *Leishmania mexicana* genome. The initial contigs were generated using only the PacBio reads and the Canu assembler, and the IPA pipeline was used to remove redundant contigs. The scaffolding step was performed using SSPACE and gap filling with GapFiller, based on Illumina paired-end short reads. Finally, the assembly was polished using Pilon, and the scaffolds were sorted based on *L. mexicana* chromosomes using Abacas. The final assembly, composed of 34 pseudochromosomes and 42 unincorporated scaffolds, represents a genome of ~32 Mb, plus the maxicircle sequence of 18.1 kb. Then, the occurrence of aneuploidies for several chromosomes of the PH8 strain and the presence of gene expansions related to genes that encode virulence factors such as amastins, GP63 and protein kinases were analyzed. Repetitive sequences were described for the first time revealing the predominance of retroelements and DNA transposons. The annotation of the gene content of *L. amazonensis* was conducted using two approaches: *ab initio* and based on transferring the annotation of a total of 8,2317 genes present in the genome of *L. mexicana*. Of these, 7999 are protein coding genes and 318 of them are classified as pseudogenes. Several multigene families that encode virulence factors, such as A2 proteins, amastins, GP63 metalloproteins, phosphatases and cysteine proteases, were identified and compared with their annotation in the genome of other trypanosomatid species. The genome of the PH8 strain has 29 genes that encode all four subclasses of amastins (α, β, γ and δ), 5 genes that encode A2 antigens, 9 genes that encode the GP63 metalloprotease and 76 cysteine proteases genes. As they were recently recognized as essential virulence factors for the establishment of the disease and progression of the infection, 14 genes were also identified that encode proteins involved in the iron and heme metabolism of the parasite, which were compared with the gene repertoire related to these pathways present in other trypanosomatids. Based on this genomic study and on recent studies by our group describing new genome editing methods, deepening the knowledge about these and other virulence factors and, as a consequence, accelerating the development of new strategies control of infection caused by *L. amazonensis*.

Keywords: *Leishmania amazonensis*, assembly, genome, multigene families, virulence factors

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

1.1. Sequenciamento e montagem de genomas

A história do sequenciamento de DNA e algoritmos complexos desenvolvidos para "montar" as sequências geradas teve início ainda no ano de 1953 com a descoberta da estrutura da molécula de DNA por Watson e Crick (1953) (WATSON; CRICK, 1953). Em 1964, Richard Holley realizou o sequenciamento do tRNA, sendo essa a primeira tentativa de sequenciar ácidos nucleicos (HOLLEY *et al*., 1965). Em 1977 foram propostos dois métodos, o de Maxam e Gilbert para sequenciamento de DNA (MAXAM; GILBERT, 1977) e o método de Sanger, baseado no uso didesoxinucleotídeos, também conhecido como terminação de cadeia (SANGER; NICKLEN; COULSON, 1977). Pode-se dizer que o método de Sanger causou uma revolução na biologia. O primeiro grande marco dessa nova era foi o sequenciamento do genoma humano, cujo primeiro rascunho foi publicado em 2001 simultaneamente por dois grupos (CRAIG VENTER *et al*., 2001; LANDER *et al.*, 2001).

Durante as últimas duas décadas houve uma enorme evolução das técnicas de sequenciamento, sobretudo após a publicação do genoma humano, quando a demanda por esta tecnologia começou a aumentar e o elevado custo era um impedimento na condução de novos projetos (revisado por VAN DIJK *et al.*, 2018). Dessa forma, nos anos seguintes foram desenvolvidas tecnologias de sequenciamento de segunda geração (do inglês *Next Generation Sequencing*), caracterizadas por paralelização massiva, automação aprimorada, maior velocidade e custo reduzido, comparado aos primeiros métodos desenvolvidos. Neste cenário surgiram as plataformas 454, em 2004, a plataforma SoliD em 2007 e por fim a plataforma que ainda hoje domina o mercado de sequenciamento, Illumina/Solexa (revisado por KCHOUK; GIBRAT; ELLOUMI, 2017). A tecnologia Illumina utiliza uma abordagem de sequenciamento por síntese, onde fragmentos de DNA previamente ligados a adaptadores são amplificados em milhares de reações de "PCR em ponte" que criam várias cópias idênticas da mesma sequência e, posteriormente cada nucleotídeo dessas sequências são determinados por essa abordagem, que emprega terminadores reversíveis (BENTLEY *et al*., 2008).

Embora as tecnologias da segunda geração tenham representado enorme avanços, uma limitação de todas essas plataformas NGS é o tamanho reduzido das *reads*, ou seja, dos fragmentos sequenciados. A partir de 2010, foram lançadas novas plataformas, desta vez caracterizadas por sequenciamento de molécula única (do inglês *Single Molecule Sequencing*, SMS) e em tempo real, as quais se enquadram na terceira geração do sequenciamento (SCHADT; TURNER; KASARSKIS, 2010). A primeira tecnologia SMS, comercializada pela Helicos Biosciences, lembrava o

sequenciamento Illumina, mas sem nenhuma amplificação em ponte, contudo, seu custo elevado e a produção de *reads* curtas (~32 pb) a tornou pouco atrativa (THOMPSON; STEINMANN, 2010). Posteriormente, em 2011 foi lançada uma nova tecnologia pela Pacific Bioscienses (PacBio), que introduziu de fato o sequenciamento de molécula única em tempo real (do inglês Single-Molecule Real-Time, SMRT) (RHOADS; AU, 2015). Por fim, mais recentemente, a empresa Oxford Nanopore Technologies introduziu o sistema de detecção de bases por meio de nanoporos imobilizados em membrana (MIKHEYEV; TIN, 2014). A grande vantagem dessas plataformas de sequenciamento em tempo real é sobretudo a geração de *reads* longas, além da exclusão da etapa de amplificação.

A plataforma PacBio permite a obtenção de *reads* de até 20 kb, sendo útil para resolver grandes *gaps* deixados em genomas montados com *reads* curtas, produzidas pela plataforma Illumina, por exemplo (JANG-IL SOHN; JIN-WU NAM, 2018; RHOADS; AU, 2015). Essa característica é essencial, sobretudo, para resolver regiões altamente repetitivas, que além de gerarem conflitos para os algoritmos desenvolvidos para lidar com *reads* curtas, também são extremamente caros computacionalmente (AJAY UMMAT AND ALI BASHIR, 2014; SUZUKI, 2019).

Apesar de produzir *reads* longas, a precisão das bases geradas na plataforma Pacbio é de apenas ~85%, ou seja ~15% das *reads* representam erros na chamada de base (SUZUKI, 2019). Nesse sentido, surgiram várias abordagens de montagens híbridas, que consideram o uso de *reads* longas e *reads* curtas para obter um genoma com maior qualidade, com poucos ou nenhum *gap* entre os *scaffolds* produzidos. Dessa forma, existem dois tipos de metodologias que podem ser abordadas: a primeira consiste na realização de montagem utilizando apenas *reads* longas e posteriormente corrigindo-as com o suporte de *reads* curtas; a segunda, ao contrário, considera a montagem inicial utilizando apenas *reads* curtas e posteriormente a junção de *contigs* gerados utilizando *reads* longas (JANG-IL SOHN; JIN-WU NAM, 2018).

Esses avanços tecnológicos possibilitaram a montagem de vários genomas por meio de abordagens híbridas e vários outros têm sido resequenciados, possibilitando o fechamento de *gaps* e a correção de regiões truncadas (DE MAIO *et al.*, 2019; KADOBIANSKYI *et al.*, 2019; MARGOS *et al.*, 2020; MILLER *et al.*, 2017; POLLO *et al.*, 2020; SAGAR M. UTTURKAR *et al.*, 2014). Com isso, ferramentas que foram desenvolvidas inicialmente para trabalhar com conjuntos de *reads* geradas por uma única plataforma, também estão sendo atualizadas para que possam se adequar as novas metodologias de montagens híbridas (GHURYE; POP, 2019; PRJIBELSKI *et al.*, 2020; SUZUKI, 2019).

Como consequência do avanço das tecnologias de sequenciamento e montagem, o número de sequências produzidas aumentou de forma exponencial. Até o presente momento, mais de 1,39

trilhões de bases de nucleotídeos provenientes de 239 milhões de sequências já foram depositadas no banco de dados GenBank (Genbank, 2022). Entre os genomas sequenciados, cujos dados estão disponíveis em inúmeros bancos de dados, encontram-se genomas de vários patógenos humanos, incluindo os genomas de protozoários parasitos causadores de doenças como a malária, doença de Chagas, leishmaniose e toxoplasmose. Em 2006, foi estabelecido o Eukaryotic Pathgen Database (EuPathDB) sob um programa do National Institutes of Health (NCBI) para criar Centros de Recusos de Bioinformática. EuPathDB consiste em uma coleção de 13 bancos de dados de bioinformática e dados experimentais relacionados a mais de 170 parasitas eucarióticos, organismos não parasitários de vida livre relevantes e hospedeiros patógenos selecionados. Dentre os bancos de dados que fazem parte desta coleção encontra-se o TritrypDB, criado em 2009, correspondendo ao primeiro banco de dados de parasitos da ordem Kinetoplastida que fornece acesso a conjuntos de dados genômicos integrados a outros dados ômicos, como por exemplo, dados de perfil de expressão e resultados proteômicos. Dessa forma, o TritrypDB suporta uma variedade de consultas complexas e retorna à comunidade científica anotações e curadoria atualizadas, além de permitir acesso a outras ferramentas de análises de dados, em escala genômica, que foram aos poucos incorporadas ao banco de dados. As informações armazenadas estão relacionadas a várias espécies da família Trypanosomatidae e incluem os gêneros *Blechomonas, Bodo, Crithidia, Endotrypanum, Leishmania, Leptomonas, Paratryoanosoma* e *Trypanosoma* que incluem protozoários parasitos e de vida livre. Esse banco de dados é, portanto, uma importante fonte de informação para estudos relacionados a espécies dessa família (ASLETT *et al*., 2010). A existência desses bancos de dados permite ao pesquisador realizar a integração de dados de sequenciamento e outras informações relevantes, bem como a utilização e o desenvolvimento de diversas ferramentas de análises computacionais. Nesse sentido, este trabalho, que tem como modelo de estudo uma espécie do gênero *Leishmania* busca contribuir para a expansão do conhecimento acerca dos tripanosomatídeos, além de aumentar o volume de informação curada no banco de dados TritrypDB.

1.2. *Leishmania* (*Leishmania*) *amazonensis* e a leishmaniose no Brasil e no mundo

O gênero *Leishmania* é composto por parasitos dixênicos obrigatórios (digenéticos) e compreende aproximadamente 53 espécies diferentes (AKHOUNDI *et al*., 2016; MOMEN; CUPOLILLO, 2000). Pelo menos 20 dessas espécies são patogênicas para os seres humanos e ~70 para outros mamíferos, onde se manifestam de formas clinicamente distintas (LAINSON; SHAW, 1987; MURRAY *et al.*, 2005). A maioria das espécies clinicamente relevantes pertencem aos subgêneros *Leishmania* e *Viannia*. Espécies do subgênero *Viannia* apresentam uma fase de

desenvolvimento no intestino posterior do inseto vetor com subsequente migração para o intestino médio, enquanto que espécies do subgênero *Leishmania* se desenvolvem no intestino médio e no intestino anterior (LAINSON; SHAW, 1987).

A leishmaniose é considerada uma doença tropical negligenciada causada por espécies de *Leishmania* spp., que são transmitidas aos hospedeiros mamíferos por flebotomíneos fêmeas infectadas (WHO, 2022). Devido à grande heterogeneidade de parasitas desse gênero, observa-se diferentes formas clínicas e patológicas dessa doença, consistindo na Leishmaniose Tegumentar (LT) e Leishmaniose Visceral (LV). Além da espécie parasitária infectante e dos fatores de tropismo e virulência de seus tecidos, o desenvolvimento de cada manifestação clínica depende de outros fatores que incluem a biologia do vetor, a genética do hospedeiro e seu sistema imunológico. A combinação desses elementos resulta no amplo espectro de manifestações clínicas da LT, incluindo a leishmaniose cutânea difusa (LCD), leishmaniose cutânea localizada (LCL) e leishmaniose mucosa (LM) (NATARAJAN *et al*., 2013).

De forma geral, a leishmaniose é endêmica em 92 países e, em todo o mundo, ocorrem anualmente 700.000 a 1 milhão de novos casos, resultando em 20.000 a 30.000 mortes por ano, principalmente associadas à LV. Ainda assim, cerca de 95% dos casos de LC ocorrem na América Latina, bacia do Mediterrâneo e Ásia Ocidental, principalmente no Afeganistão, Argélia, Bolívia, Brasil, Colômbia, Irã, Iraque, Paquistão, República Árabe Síria e Tunísia (WHO, 2022). Nos últimos 20 anos, 1.067.759 casos de LC foram notificados à Organização Pan-Americana da Saúde (OPAS), com 16.000 novos casos notificados somente no Brasil em 2020 (PAHO, 2021).

Dentre as mais de 20 espécies de *Leishmania* patogênicas ao homem, destaca-se a *Leishmania amazonensis* por poder causar praticamente todas as formas clínicas de LT e também LV em humanos (AKHOUNDI *et al.*, 2016, BARRAL *et al*., 1991) e cães (VALDIVIA *et al.*, 2017; TOLEZANO *et al.*, 2007). O fato de *L. amazonensis* ser capaz de causar infecção por LV é intrigante pois a LV é causada principalmente por espécies do Velho Mundo, que compreende o complexo *Leishmania donovani* e *Leishmania infantum chagasi* enquanto que, *L. amazonensis* é classificada como uma espécie autóctone do Novo Mundo (AKHOUNDI *et al.*, 2016). Além disso, enquanto LCL ou LM causadas por *Leishmania braziliensis* são caracterizadas por respostas exacerbadas de células T ao parasita e produção precoce de quantidades excessivas de citocinas pró-inflamatórias (por exemplo, IFN-γ, TNF-α e IL-6), que estão associadas com baixa carga parasitária, mas dano tecidual exacerbado, a LCD causada por *L. amazonensis* correlaciona-se com anergia das respostas das células T e grande quantidade de parasitas nas lesões (CONVIT ;RONDON; PINARDI, 1971, AFONSO; SCOTT, 1993, RUSSO *et al*, 1993, CASTES; TAPIA, 1998, SOONG; HENARD; MELBY, 2012). Atividade lítica de células NK contra macrófagos parasitados também foi relatada para pacientes com LT infectados por *L.* amazonensis, assim como

visto para infecções provocadas por *L. donovani* (BARRAL-NETO *et al*., 1995). Todos esses diferentes padrões de interação com o sistema imunológico do hospedeiro sugerem que *L. amazonensis* possui fatores de virulência distintos ou pode expressá-los de forma diferente em relação a outras espécies.

Diferentes espécies de *Leishmania* apresentam também diferentes graus de virulência, como ilustrado pelo fato de várias linhagens de camundongos serem resistentes a *L. major* e suscetível a *L. amazonensis* (MCMAHON-PRATT; ALEXANDER, 2004; PEREIRA; ALVES, 2008). Da mesma forma, diferentes cepas da mesma espécie de *Leishmania* podem ter fenótipos distintos de infectividade e metastáticos *in vivo* (ALVES-FERREIRA *et al*., 2015; DA FONSECA PIRES *et al*., 2014; WALKER *et al*., 2006). Uma comparação realizada entre as cepas LV79 e PH8 de *L. amazonensis*, por exemplo, mostrou que a cepa PH8 é mais virulenta em camundongos, e que os parasitas derivados de lesões são mais viáveis e mais infectantes *in vitro*. Enquanto camundongos desafiados com a cepa LV79 desenvolveram lesões que aumentavam até seis semanas após a inoculação e diminuíram depois disso (VELASQUEZ *et al*., 2016), a cepa PH8 mostrou gerar lesões de tamanho crescente na mesma linhagem de camundongo (CORTEZ *et al*., 2011). Além disso, comparações entre o proteoma de amastigotas de ambas as cepas mostrou que proteínas GP63 são altamente expressas na cepa PH8, enquanto que superóxido dismutase, tryparedoxin peroxidase e proteína de choque-térmico 70 são mais abundantes na cepa LV79, resultados estes que correlacionam com seus fenótipos (DE REZENDE *et al*., 2017).

Além das características citadas anteriormente, *L. amazonensis* (IFLA/BR/67/PH8) tem sido utilizada em diversos estudos que visam o desenvolvimento de uma vacina contra LC. Esta cepa foi escolhida porque seus antígenos induzem altos índices de estimulação para linfócitos de voluntários vacinados (NASCIMENTO *et al*., 1990), além de apresentar crescimento facilitado em meio não celular, ser internacionalmente conhecida e está bem definida taxonomicamente (MAYRINK *et al*., 2002). Em 1979, Mayrink e colaboradores desenvolveram uma vacina de promastigotas mortas composta por cinco estoques de diferentes isolados de *Leishmania* (Leishvacin®), incluindo a cepa PH8. Mais tarde, em 2002, a vacina polivalente foi testada juntamente com formulações vacinais monovalentes, mostrando que apenas a Leishvacin® e a vacina monovalente PH8 foram capazes de induzir proteção contra a LT e os maiores níveis de interferon-g (MAYRINK *et al.* 2002). A proteína Larp33, expressa em promastigotas de *L. amazonensis* também já foi descrita como candidata a um componente vacinal contra LC, induzindo uma resposta protetora mediada por células Th1 (FERNANDES *et al*., 1997). Desde então, outras moléculas da cepa PH8 já foram testados e avaliadas como potenciais candidatos vacinais (GRENFELL *et al*., 2010; MARTÍNEZ-RODRIGO *et al*., 2019; MARZOCHI *et al*., 1998; MONTALVO-ÁLVAREZ *et al*., 2008).

A distribuição de casos de LC e LV no mundo é concentrada em alguns países, sobretudo onde há má distribuição de renda. Mais de 95% dos casos mundiais de LV ocorre em apenas 10 países (Brasil, China, Etiópia, Índia, Iraque, Quênia, Nepal, Somália, Sudão do sul e Sudão) e, apesar da ampla distribuição, cerca de 95% dos casos de LC ocorre nas Américas, na bacia do Mediterrâneo e oeste da Ásia, principalmente no Afeganistão, Argélia, Bolívia, Brasil, Colômbia, Irã, Iraque, Paquistão, República Árabe Síria e Tunísia (Figura 1) (WORLD HEALTH ORGANIZATION, 2022).

Nas últimas décadas tem sido observado o aumento da incidência da doença em áreas endêmicas, mas também o surgimento de epidemias em áreas não endêmicas, o que foi associado principalmente a eventos de migrações populacionais e a melhora no diagnóstico e relato dos casos (PISCOPO; MALLIA, 2006).

Dentre as espécies associadas a LC no Brasil, *L. (L.) amazonensis* (*L. amazonensis*) é responsável por mais de 8% de todos os casos nas regiões norte e nordeste, onde também têm sido reportados casos de transmissão autóctone (INES *et al.*, 2011; PAULO *et al.*, 2007). *L. amazonensis* é, portanto, considerada o agente etiológico da leishmaniose cutânea difusa (LCD), caracterizada pelo aparecimento de múltiplas lesões bastante polimórficas (SILVEIRA *et al.*, 2009). Somente no Brasil ~40 mil pessoas são afetadas por LC e cerca de 17 mil novos casos são reportados por ano, baseado nos dados da OMS coletados entre 2015 e 2017 (WHO, 2022).

A transmissão da *L. amazonensis* ocorre entre humanos e aproximadamente outras 70 espécies de mamíferos por meio da picada de fêmeas de dípteros da família Psychodidae, sub-família Phebotominae, gênero *Lutzomya*, conhecidos genericamente por flebotomíneos. Por causa da proximidade com humanos, caninos (*Canis familiaris*) são os principais reservatórios do parasito em regiões urbanas e desenvolvem sobretudo, leishmaniose visceral canina (LVC) (AKHOUNDI *et al.*, 2016). Mas, outras espécies mamíferas também constituem a grande diversidade de reservatórios deste parasito, como por exemplo, roedores (K. ROSE *et al.*, 2004), marsupiais (DOUGALL *et al.*, 2009), edentados, carnívoros e primatas (PISCOPO; MALLIA, 2006).

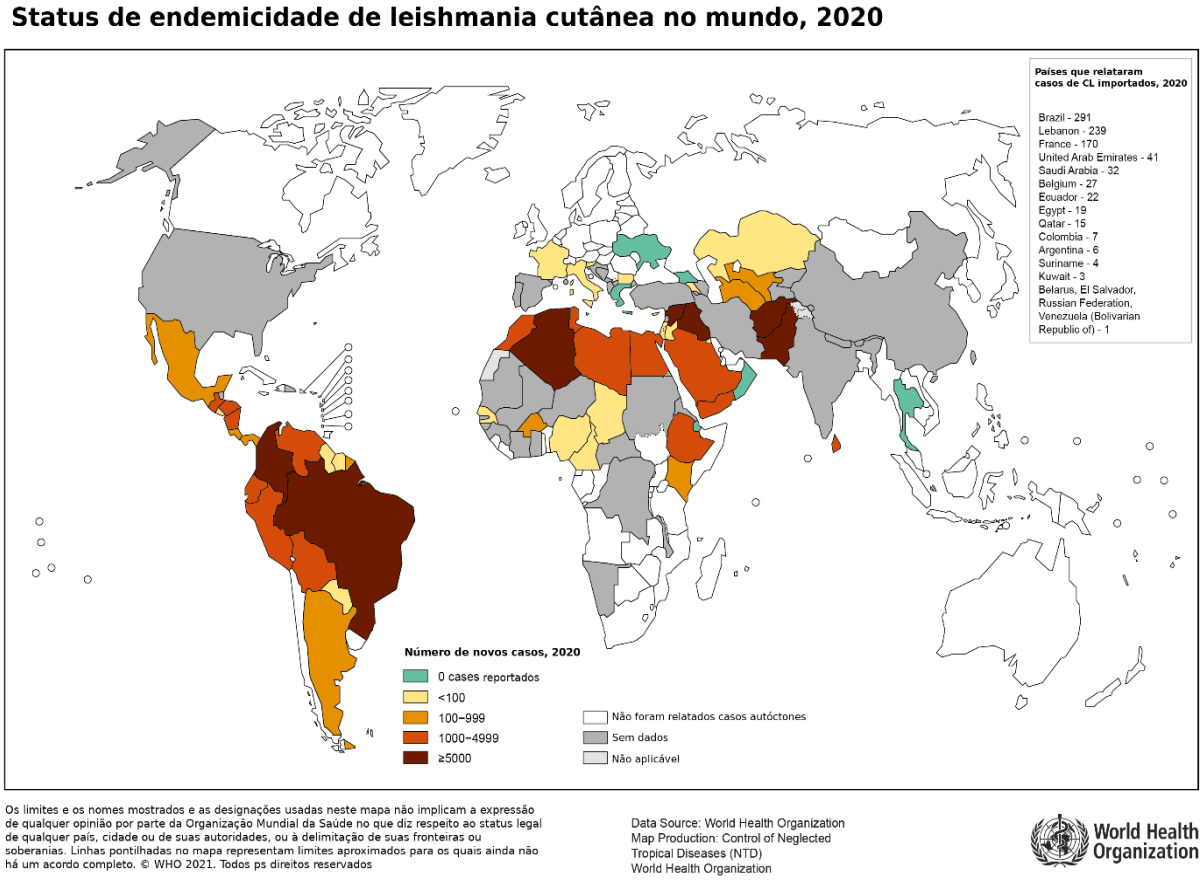

Figura 1. Distribuição de casos de leishmaniose cutânea relatados em 2020. Fonte: World Health Organization (2022).

Protozoários do gênero *Leishmania* alternam entre duas formas morfológicas principais durante seu ciclo de vida: promastigotas e amastigotas. Durante o repasto sanguíneo, fêmeas de insetos flebotomíneos adquirem *L. amazonensis* a partir de células infectadas do hospedeiro mamífero. Quando a mosca da areia se alimenta de cães, por exemplo, uma pequena ferida é formada no local da picada devido as suas peças bucais que são semelhantes à serra, então, há um acumulo de sangue nos capilares lesionados, levando a migração do parasito para o membro anterior do flebotomíneo. Neste local, promastigotas metacíclicas replicam-se e são transmitidas a outros cães e de forma análoga, a outros reservatórios vertebrados, incluindo humanos (Figura 2) (BATES, 2018; STEVERDING, 2017).

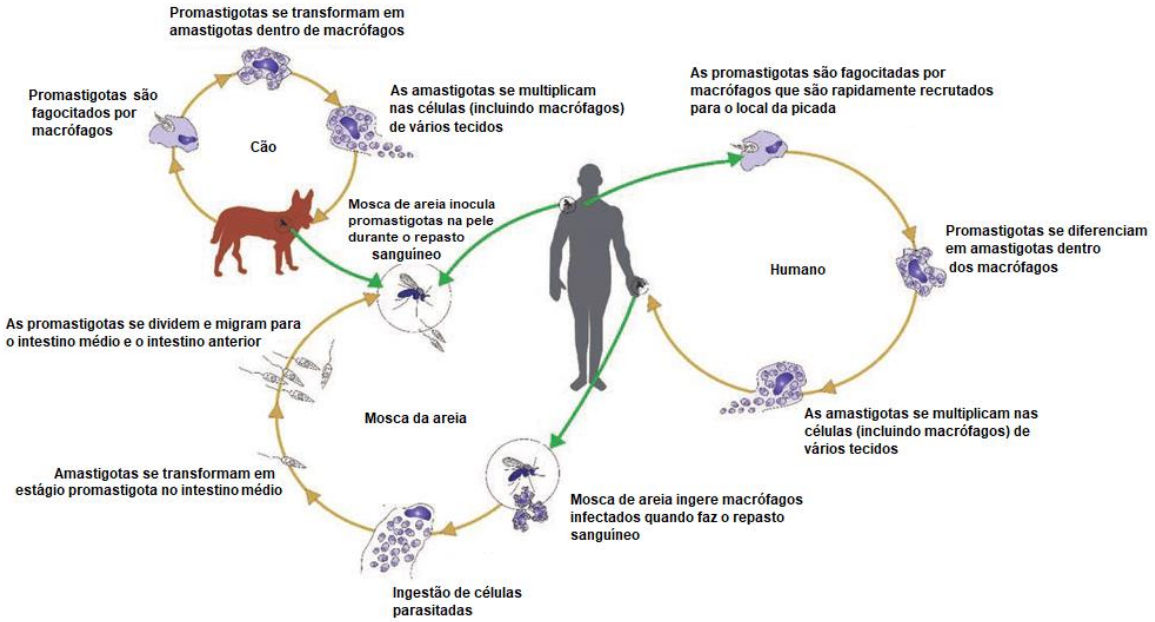

Figura 2. Ciclo de vida de *L. amazonensis* no hospedeiro humano e cães.
Fonte: ALMEIDA *et al.*, 2017, adaptada pelo autor.

Uma vez atingida a corrente sanguínea, promastigotas infectam células fagocíticas, induzindo a formação de fagossomos, onde ocorre a liberação de vários fatores de sobrevivência intracelular, de forma a modificar a biogênese do fagossomo e a inibição de vias pró-inflamatórias. Estes eventos promovem então, um ambiente propício para a diferenciação de promastigotas em amastigotas. Estas, por sua vez, se multiplicam por divisão binária simples e provocam a destruição do citoplasma celular, resultando no rompimento da célula e liberação das amastigotas na corrente sanguínea e no meio intercelular, o que facilita a infecção células vizinhas. Ao realizar um novo repasto sanguíneo, a mosca da areia acaba ingerindo estas células infectadas com o parasito, que ao atingir o trato digestivo se diferencia novamente em promastigotas, reiniciando o ciclo (STEVERDING, 2017).

A leishmaniose engloba um amplo espectro de manifestações clínicas, com diferentes sintomas que vão desde lesões cutâneas não ulcerativas localizadas ou difusas (leishmaniose cutânea localizada – LCL e leishmaniose cutânea difusa - LCD) até o acometimento de vários órgãos como linfonodos, baço, fígado, medula óssea e, menos comumente, rins (leishmaniose visceral - LV). Além disso, pode existir uma fase assintomática, principalmente em pacientes acometidos pela LV, causada pela *Leishmania (Leishmania) infantum* (*L. infantum*) no Brasil (BHATTACHARYA; ALI, 2013; HANDLER *et al.*, 2015; KEVRIC; CAPPEL; KEELING, 2015; SILVEIRA *et al.*, 2009).

LCL é a manifestação clínica mais frequente e é caracterizada pela presença de lesões cutâneas, principalmente úlceras, exclusivas no local da picada. Essas lesões geralmente ocorrem em áreas expostas da pele, como rosto, mãos e pernas, onde pápulas assintomáticas ou nódulos se

desenvolvem transformando-se em úlceras arredondadas. Contudo, essas lesões tendem a se curar espontaneamente após alguns anos, deixando uma cicatriz secundária (HANDLER *et al.*, 2015). Leishmaniose cutânea disseminada (LD) é uma variação da LC, na qual o parasito atinge outras áreas do corpo por meio do sistema linfático ou sanguíneo. Dessa forma, ocorre a formação de numerosas lesões ulceradas em várias áreas do corpo (SILVEIRA *et al.*, 2009). Já a LCD é uma forma rara e severa da LC, a qual acomete principalmente pessoas que não respondem adequadamente aos antígenos da *Leishmania*. É caracterizada pelo aparecimento de uma única lesão que se desenvolve de forma crônica, envolvendo múltiplas nodulações não ulceradas que podem cobrir grandes extensões cutâneas. Superfícies da face, orelhas e extensores, como joelhos e cotovelos são as regiões mais susceptíveis a esses tipos de manifestações (KEVRIC; CAPPEL; KEELING, 2015).

Por fim, a LV também conhecida como kala-azar é uma manifestação crônica e sistêmica da doença, na qual o parasito se multiplica em vários tipos de tecidos, causando o aumento do fígado, baço e anemia, além de ser caracterizada pelo aparecimento de manchas de Peyer no intestino, pulmões e pele. Contudo, o início dessa fase costuma ser assintomática devido ao período variável de incubação do parasito, sendo sucedido por sintomas que incluem febre irregulares, mal-estar, tremores e perda de peso (BHATTACHARYA; ALI, 2013; WHO, 2022).

1.3. Estudos do genoma de *L. amazonensis* e de outras espécies do gênero

Espécies do gênero *Leishmania* são considerados organismos diplóides, pois carregam em seu genoma duas cópias da maioria de seus cromossomos homólogos (BASTIEN; BLAINEAU; PAGES, 1992; IVENS *et al.*, 2005; MANNAERT *et al.*, 2012; PEACOCK *et al.*, 2007; ROGERS *et al.*, 2011), mas existem inúmeros trabalhos que relatam a ocorrência de eventos de aneuploidias para vários cromossomos, dependendo da espécie e da cepa considerada (DOWNING *et al.*, 2011; MANNAERT *et al.*, 2012; ROGERS *et al.*, 2011; STERKERS *et al.*, 2012; PATINO *et al.*, 2020). Outra característica bem estabelecida para o genoma de *Leishmania* é a organização dos genes codificantes de proteínas que acontecem em longas unidades transcricionais policistrônicas, levando a co-transcrição pela RNA Polimerase II de vários genes em uma única molécula de pré-RNA mensageiro, a qual é posteriormente processada em eventos de trans-splicing e poliadenilação (MARTÍNEZ-CALVILLO *et al.*, 2003, 2004).

A rede genoma de *Leishmania* (LGN) deu início em 1994 ao sequenciamento de várias espécies do mesmo gênero, os quais continuam em plena expansão. O consórcio EuLeish incluiu o Instituto Sanger, Instituto de Pesquisas Biomédicas de Seattle (SBRI) e vários laboratórios europeu, onde foram abordadas diferentes estratégias paralelas de sequenciamento, como o sequenciamento

completo de cosmídios, sequenciamento *shotgun* de cromossomos inteiros e sequenciamento de BACs (BASTIEN *et al.*, 1998; IVENS; BLACKWELL, 1996). O primeiro genoma completamente sequenciado foi o de *Leishmania major* MHOM/IL/81/Friedlin (LmjF). Este resultou na montagem de 36 cromossomos que variam entre 0.3 – 2.8 MB, do genoma haplóide de 32,8 MB. Esses resultados corroboraram as previsões feitas anteriormente por Wincker e colaboradores (1996), que mapearam 244 sondas de DNA específicas para cromossomos separados por *pulsed field gel electrophoresis* (PFGE). A partir dessa montagem foi possível realizar a predição de 8.272 genes codificadores de proteínas, dos quais 36% tiveram função putativa atribuída (IVENS *et al.*, 2005).

Em 2013 foi descrita a primeira tentativa de montagem do genoma de *L. amazonensis* (MHOM/BR/71973/M2269), a partir de parasitos obtidos de lesões cutâneas de pacientes infectados da cidade de Cafezal, Pará. A montagem do genoma foi conduzida por meio de uma estratégia mista utilizando *reads single-end* 454 com tamanho médio de 350 pb e *reads paired-end* Illumina com tamanho médio de 76 pb e 400 pb de tamanho médio do inserto, o que resultou na obtenção de 2.627 *scaffolds*, com N50 de 22.901 pb, totalizando 29,6 MB. Predição *ab initio* de genes foi capaz de modelar 8.100 estruturas gênicas com tamanho médio de 1.793 pb (REAL *et al.*, 2013). Os dados obtidos mostraram também a ocorrência de aneuploidias nessa espécie, uma vez que o mapeamento de *reads* Illumina contra o genoma de *L. mexicana* sugeriu a ocorrência de cromossomos supranumerários (cromossomos 7, 23 e 30) (REAL *et al.*, 2013).

Estudos mais recentes mostraram que aneuploidias é um fenômeno frequente em *L. amazonensis*. Dois isolados provenientes de cães com manifestações clínicas de doença visceral mostraram que a maioria dos cromossomos apresenta profundidade *reads* compatível com número de cópias dissômico, com exceção do cromossomo 30, ou seja, a mediana da densidade de *reads* foi maior apenas no cromossomo 30 quando comparada a mediana da profundidade de *reads* de todo o genoma. Este resultado confirma o que foi descrito em outros estudos, nos quais também foram relatadas aneuploidias entre espécies ou até mesmo entre isolados da mesma espécie (PEACOCK *et al.*, 2007; ROGERS *et al.*, 2011; VALDIVIA *et al.*, 2011, 2017). VALDIVIA *et al.* (2011) mostrou a ocorrência de aneuploidias em mosaico entre dois isolados de *L. (V.) peruviana*, onde o número normalizado de cópias dos cromossomos agrupou-se em torno de um padrão dissômico, tendo sido visto também desvios significativos dos valores não integrais evidentes para alguns cromossomos que apresentaram profundidade de *reads* intermediária entre os perfis dissômicos e trissômicos.

Em 2019 foram descritos dois trabalhos sobre sequenciamento e montagem de sequências das cepas CDC210-660 e UA301 de *L. amazonensis.* O primeiro foi obtido por sequenciamento na plataforma PacBio e Illumina MiSeq, resultando na montagem de 30.427.462 *reads paired-end* Illumina de 250 pb e 262.667 *reads* PacBio resultou em 92 *scaffolds*, obtendo N50 de 850.106 bp e totalizando 33.504.997 bp de tamanho total, correspondente ao genoma haplóide (BATRA *et al.*,

2019). Uma nova montagem, baseada somente em *reads* Illumina e guiada por referência (*L. mexicana* (MHOM/GT/2001/U1103)) atingiu o número de cromossomos preditos para espécies do complexo de *L. (L.) mexicana* igual a 34, totalizando 32.156.470 pb de tamanho total da montagem (PATINO *et al.*, 2020). Contudo, o trabalho de Patino e colaboradores possibilitou confirmar eventos de aneuploidia moderada, mostrando que 24 cromossomos eram dissômicos, 9 trissômicos e 1 tetrassômico. Além disso, foi relatada a variação no número de cópias de genes envolvidos na virulência, crescimento e sobrevivência do parasito, e nas distribuições dos genes multicópia em alguns cromossomos, além de um alto nível de heterozigosidade (PATINO *et al.*, 2020).

O número de cromossomos para *L. (V.) braziliensis*, *L. infantum*, *L. (L.) donovani* e *L. (L.) mexicana* também foi inicialmente predito por PFGE utilizando sondas específicas para cada espécie e posteriormente por sequenciamento e PCR de extremidades para confirmar a junção entre *contigs* (BRITTO *et al.*, 1998; DÉ *et al.*, 2012; PEACOCK *et al.*, 2007). O cariótipo molecular para espécies do Velho Mundo, *L. major*, *L. infantum* e *L. donovani* compreende 36 cromossomos cada, enquanto que espécies do Novo Mundo, *L. braziliensis* e o complexo *L. mexicana* (que inclui *L. amazonensis*) tem 35 e 34 cromossomos, respectivamente, evidenciados também em experimentos comparativos de PFGE. Estes possibilitaram detectar a ocorrência de rearranjos cromossômicos a partir da fusão dos cromossomos 8+29 e 20+36 para o complexo *L. mexicana* e entre os cromossomos 20+34 para *L. brazilienesis* (BRITTO *et al.*, 1998; WINCKER *et al.*, 1996). Com a disseminação das metodologias de sequenciamento, vários novos isolados foram sequenciados para cada uma dessas espécies sendo que alguns detalhes obtidos nessas montagens podem ser consultados na Tabela 1.

Os dados publicados dos genomas de *L. donovani* (HU3) e *L. infantum* (JPCM5) relatam a montagem de 36 cromossomos, como experimentalmente preditos. Para *L. donovani* foi alcançado pela primeira vez uma sequência genômica sem nenhum *gap*, considerando uma montagem *de novo* e híbrida, mais uma vez utilizando *reads paired-end* Illumina de 101 pb e *reads* PacBio com tamanho médio de 11.900 pb (CAMACHO *et al.*, 2019b). Da mesma forma, a montagem do genoma de *L. infantum* utilizou *reads paired-end* de 126 pb e *reads* PacBio com comprimento médio de 11.700 pb (GONZÁLEZ-DE LA FUENTE *et al.*, 2017). Além disso, ambos os genomas foram anotados a partir da transferência de anotação do genoma de *L. major* (cepa Friedlin), o que resultou em 8.595 e 8.976 genes codificantes de proteínas, respectivamente (CAMACHO *et al.*, 2019b; GONZÁLEZ-DE LA FUENTE *et al.*, 2017). Para o genoma de *L. infantum* esses resultados revelaram 495 novos genes anotados, além da correção de outros 100 e a descontinuação de 75, que foram excluídos devido a erros anteriores (GONZÁLEZ-DE LA FUENTE *et al.*, 2017).

Com relação aos outros genomas de espécies do Novo Mundo, montagens recentes têm chegado a nível cromossômico, alcançando o número de cromossomos previstos em experimentos

moleculares de cariótipo. A última publicação referente a montagem e anotação do genoma de *L. braziliensis* (MHOM/BR/75/M2904) por exemplo, relata a obtenção de 35 cromossomos contíguos, bem como a anotação de 8.395 genes codificadores de proteínas e 33 pseudogenes. Semelhantemente as últimas versões disponíveis para outras cepas, o genoma de *L. amazonensis* também considerou uma montagem *de novo* utilizando *reads* Pacbio (16.600 pb de tamanho médio) com posterior correção utilizando *reads* curtas (126 pb *paired-end*) (GONZÁLEZ-DE LA FUENTE *et al.*, 2019).

Tabela 1. Genomas de *Leishmania* depositados no repositório Genomas/NCBI e seus aspectos.

Espécie	Número de Acesso (Assembly/WGS)	Ano de publicação	*Software*	Tecnologia de sequenciamento	n° de *scaffolds*	N50 *(scaffolds/ contigs)*	Tamanho total
Leishmania amazonensis MHOM/BR/71973/M2269	LeiAma1.0/ APNT00000000.1	Jul/2013	Newbler v. 2.3; Velvet v. 0.7.56; Zorro	454; Illumina GAII	2.627	22,901 / 17,272	29,029,348 (29,6 lasted)
Leishmania amazonensis CDC210-660 (RZOD01)	ASM399250v1/ RZOD00000000.1	Jan/2019	Assembly v. 2.1	PacBio RSII	92	850,106	33,504,997
Leishmania amazonensis UA301	ASM531712v1	Maio/2019	SMALT v. 0.7.4	Illumina HiSeq	34	-	32,156,470
Leishmania braziliensis MHOM/BR/75/M2904	ASM284v2/CADA00000000.1	Abril/2007	PHRED/PHRAP	whole-genome shotgun - Sanger	138	992,961 / 63,680	32,068,771
Leishmania braziliensis MHOM/BR/75/M2903	GCA_000340355.2/AOSE00000000.2	Fev/2013	Newbler v. oct. 2015	Roche 454	744	1,030,512 / 62,201	35,210,150
Leishmania braziliensis IOC-L 3564	GCA_003304975.1/QFBG00000000.1	Jul/2018	SPAdes v. 3.1.10	IonTorrent	1.029	758,103 / 7,920	38,003,648

Leishmania braziliensis MHOM/BR/75/M2904	GCA_900537975.1	Dec/2018	HGAP3/ CLC Genomics Workbench v 5.0	PacBio RS II; Illumina HiSeq2000	35	1,063,631	32,301,632
Leishmania braziliensis MHOM/BR/75/M2904	GCA_902369275.1	Out/2019	IDBA_UD v. 1.1	Illumina HiSeq 2000	1	-	17,771 (mitocondrial)
Leishmania donovani BPK282A1	GCA_000227135.2	Fev/2011	Newbler	454; Illumina	36	1,024,085 / 45,436	32,444,968
Leishmania donovani Ld 2001	GCA_000283395.1/ALJU00000000.1	Ago/2012	Velvet v. 2.0	SOLiD	14,518	3,370	27,466,456
Leishmania donovani Ld 39	GCA_000316305.1	Dez/2012	Velvet v. 1.2.07	SOLiD	16.323	1,772	23,683,296
Leishmania donovani BHU 1220	GCA_000470725.1/AVPQ00000000.1	Set/2013	Bowtie v. 0.12.9	Illumina HiSeq	36	1,024,085 / 41,904	32,414,853
Leishmania donovani MHOM/IN/1983/AG83	GCA_001989975.1	Fev/2017	AllPaths v. Sep-2015	Illumina HiSeq	36	1,029,368 / 20,549	32,196,393
Leishmania donovani MHOM/IN/1983/AG83	GCA_001989955.1	Fev/2017	AllPaths v. 2015; STLab-assembler v. 2016	Illumina HiSeq	36	1,015,993 / 19,680	32,148,377

Leishmania donovani Pasteur	GCA_002243465.1	Ago/2017	HGAP v. 2	PacBio	37	1,079,609 / 770,314	33,545,875
Leishmania donovani LdCL	GCA_003719575.1	Nov/2018	HGAP v. 2; Celera Assembler v. 8.3; Canu v. 1.0	PacBio; Illumina MiSeq	36	1,067,468	32,959,864
Leishmania donovani FDAARGOS_361	GCA_003730215.1/RHLC00000000.1	Nov/2018	Canu v. 1.2	PacBio; Illumina	56	1,033,854	33,453,722
Leishmania donovani FDAARGOS_360	GCA_003730175.1/RHLD00000000.1	Nov/2018	Canu v. 1.2	PacBio; Illumina	71	828,097	34,011,430
Leishmania infantum JPCM5	GCA_000002875.2/CACT0100000.1	Abril/2007	phrap	Illumina Genome Analyzer II	76	1,043,848 / 302,093	32,122,061
Leishmania infantum TR01	GCA_003020905.1	Março/2018	Geneious v. 11.0.5.	Illumina HiSeq	36	1,046,244	32,009,138
Leishmania infantum JPCM5 (MCAN/ES/98/LLM-724)	GCA_900500625.1/UINB0100000.1	Ago/2018	CLC Bio; version 5.0; HGAP v2.3.0	Illumina HiSeq. 2000; PacBio RS II	36	1,055,293	32,803,248
Leishmania infantum	GCA_902369335.1	Out/2018	IDBA_UD (version 1.1.1)	Illumina HiSeq 2000	1	-	17,897 (mitocondrial)

Leishmania infantum HUUFS14	GCA_003671315.1/NSCO01000001	Out/2018	ABySS v. 1.56	Illumina HiSeq	2,507	29,848 / 25,367	32,578,914
Leishmania major Friedlin	GCA_000002725.2	Jan/1998	phrap	whole-genome shotgun - Sanger	36	1,091,540	32,855,089
Leishmania major SD 75.1	GCA_000250755.2/AFZI0100000.1	Fev/2012	Newbler v. MapAsmResearch-04/19/2010-patch-08/17/20106.0	Roche 454	36	1,022,795 / 89,399	31,242,750
Leishmania major LV39c5	GCA_000331345.1/AODR0100000.1	Jan/2013	Newbler v. March 2012	Roche 454	849	978,401 / 71,814	32,327,517
Leishmania major Friedlin	GCA_902369385.1	Out/2019	IDBA_UD (version 1.1.1)	Illumina HiSeq 2000	1	18,998	18,998 (mitocondrial)
Leishmania major Friedlin	GCA_902498725.1/CABVLB01000000.1	Out/2019	IDBA_UD (version 1.1.1)	Illumina HiSeq 2000	4	689	2,736 (mitocondrial)
Leishmania mexicana MHOM/GT/2001/U1103	GCA_000234665.4/CADB00000000.1	Fev/2011	Phrap	whole-genome shotgun - Sanger	588	1,044,075 / 164,930	32,108,741
Leishmania mexicana 215-49	GCA_000234665.4/CADB01000001	Jan/2019	Canu v. 1.6	PacBio RSII	55	825,953	32,057,209

L. amazonensis e *L. mexicana* são espécies que ocorrem em países do Novo Mundo e pertencem ao complexo *L. (L.) mexicana*, por este motivo e devido à similaridade já observada entre seus genomas (~92% de identidade e 99,87% de sintenia, mais ~80% do conjunto central de genes que codificam proteínas são compartilhados entre eles) (TSCHOEKE *et al.*, 2014), o genoma de *L. mexicana* é quase sempre usado como referência na montagem e anotação de *L. amazonensis*. Seu primeiro genoma montado e anotado foi publicado em 2011 e considerou uma abordagem *de novo* usando *reads* de sequenciamento por capilaridade gerado a partir de bibliotecas *whole-genome shotgun*. Os resultados indicaram uma montagem fragmentada em 929 *contigs*, dos quais 375 foram ordenados e juntados em 34 pseudocromossomos. Modelos de estruturas gênicas e anotação funcional foram transferidos do genoma *L. major* seguida de anotação manual. Dessa forma, 8.250 genes codificantes de proteínas foram preditos para *L. mexicana*, sendo apenas 2 exclusivos quando comparados aos genomas de *L. major*, *L. infantum* JPCM5 e *L. braziliensis* M2904, além de 132 genes multicópias. Análises de variação do número de cópias cromossômicas em *L. mexicana* mostrou que um e três cromossomos das cepas isoladas (U1103 e M379), respectivamente, possuem profundidade de *read* trissômica, além de não ter sido possível determinar a ploidia para outros 10 cromossomos em U1103 e para outros 2 em M379 (ROGERS *et al.*, 2011). Em 2019, outra cepa de *L. mexicana* foi sequenciada e seu genoma foi obtido a partir da montagem utilizando *reads* PacBio e Illumina *paired-end* de 250 pb, o que permitiu obter uma sequência mais contígua em 55 *scaffolds*, totalizando 32.057.209 pb de tamanho (BATRA *et al.*, 2019).

Para uma descrição completa do genoma de um eucarioto, é necessário também realizar o sequenciamento e montagem do genoma mitocondrial. Em tripanosomatídeos, o cinetoplasto contém uma enorme rede de DNA mitocondrial (kDNA), altamente condensada e formada por várias moléculas circulares conectadas. Entre 20 mil a 30 mil dessas moléculas possuem comprimento que varia de 0,5 kb a 2 kb (denominadas minicírculos) e outras dezenas, que variam de 20 kb a 40 kb de comprimento (denominadas maxicírculos) são as sequências que se assemelham ao genoma mitocondrial eucarioto (CHEN *et al.*, 1995; JENSEN; ENGLUND, 2012). No maxicírculo do kDNA são codificadas várias enzimas envolvidas na cadeia de transporte de elétrons, bem como rRNAs mitocondriais. (CHEN *et al.*, 1995; MARTYNKINA *et al.*, 1991; YATAWARA *et al.*, 2008). Além disso, análises comparativas mostraram uma alta conservação dessas regiões entre os cinetoplastídeos dos gêneros *Trypanosoma, Leishmania e Leptomonas*, tornando-os importantes marcadores moleculares tanto para estudos filogenéticos quanto para a detecção de novas espécies por meio de várias abordagens baseadas em qPCR (CAMACHO *et al*., 2019a; CECCARELLI *et al*., 2020; NOCUA *et al*., 2011; WESTENBERGER *et al*., 2006; YATAWARA *et al*., 2008).

Muitas ORFs (do inglês *open reading frame*) de maxicírculos encontram-se truncadas quando as sequências genômicas são analisadas, mas estão completas quando comparadas com o cDNA correspondente. Estudos realizados por vários grupos mostraram que os transcritos gerados precisam passar por um processo de edição antes de serem traduzidos. Esse processo pós-transcricional (conhecido como edição de RNA) ocorre por meio da atividade de RNAs guias (gRNAs), que são codificados principalmente nos minicírculos, apesar de alguns também serem codificados no maxicírculo (SIMPSON *et al*., 2015). Os gRNAs direcionam a inserção ou exclusão adequada de poucos ou milhares de resíduos de uridina (U), para criar ORFs funcionais nos transcritos derivados do maxicírculo. Esses mecanismos envolvem uma vasta gama de proteínas estruturais e enzimas catalisadoras. De forma bastante simples, uma riboendonuclease dependente de gDNA cliva o mRNA alvo após a formação do complexo gDNA-mRNA, posteriormente uma enzima conhecida por uridilil transferase adiciona resíduos de U's na extremidade 3' do fragmento de clivagem de mRNA, no caso de eventos de inserção ou uma exonuclease remove U's na extremidade 3' do fragmento. O processo é então finalizado pela ação de uma ligase que irá unir as duas fitas com extremidades livres (revisado por SIMPSON; SBICEGO; APHASIZHEV, 2003).

Dentre todas as espécies de *Leishmania*, a que possui o genoma mitocondrial mais bem caracterizado é o de *L. tarentolae*, pois é tido como modelo em estudos que visam entender os mecanismos de edição de RNA que ocorre em transcrições do maxicírculo. Existem aproximadamente entre 5000 e 10000 minicírculos e de 20 a 50 maxicírculos na mitocôndria deste parasito (SIMPSON, 1987). Para *L. amazonensis*, no entanto, não há nenhum estudo que descreva suas características, bem como os produtos gênicos e suas implicações nos processos celulares.

Análises de genômica comparativa possibilitaram obter vários outros *insights* a respeito do conteúdo gênico, variações no número de cópias cromossômicas e cópias de genes, além de permitir o estudo de famílias multigênicas relacionadas à virulência de *Leishmania* (BATRA *et al.*, 2019; BUTENKO *et al.*, 2019; EL-SAYED *et al.*, 2005; PEACOCK *et al.*, 2007; REAL *et al.*, 2013; ROGERS *et al.*, 2011). Esses estudos serão amplamente abordados nas seções seguintes.

1.4. Estudos de genômica comparativa de tripanosomatídeos

Estudos de genômica comparativa contribuem para o entendimento evolutivo entre diferentes espécies da mesma família ou entre diferentes famílias. Dessa forma, muitos aspectos que explicam diferenças no ciclo de vida, organização genômica, infectividade, patogênese parasitismo celular, tecidual ou mesmo diferenças no tipo de hospedeiro podem ser elucidados. Apesar da grande distância filogenética e de causarem doenças com manifestações completamente distintas, análises comparativas entre os genomas de *L. major*, *T. cruzi* e *T. brucei* revelaram um proteoma

compartilhado entre ambos, composto de ~6.200 genes, organizados em grandes *clusters* de genes direcionais sintênicos, além disso, *L. major* mostrou outros 482 genes compartilhados somente com *T. cruzi* e 74 com *T. brucei* e ~1.000 genes exclusivos (EL-SAYED *et al.*, 2005).

A conservação entre espécies de tripanosomatídeos se estende por longos blocos sintênicos nos genomas. Os genomas de *T. brucei* e *L. major*, por exemplo, apresentaram 110 blocos de sintenia com tamanhos de 19,9 e 30,7 Mb, respectivamente. Quebras na ocorrência da sintenia apareceram, sobretudo em regiões altamente repetitivas e próximas a regiões de troca de fita que separaram os *clusters* de genes direcionados (DGCs), onde grandes famílias multigênicas, retroelementos ou RNAs estruturais estão anotados (EL-SAYED *et al.*, 2005). Ao contrário desse padrão de blocos conservados interrompidos por regiões de troca de fita, não foram detectadas rupturas óbvias de sintenia quando comparados espécies de *Leishmania* (*L. major, L. infantum* e *L. braziliensis*). Além disso, alinhamentos mostraram que mais de 99% dos genes entre os três genomas são sintênicos e que a conservação nas sequências codificantes também é alta, o que revela organização genômica altamente conservada e a ocorrência de poucos rearranjos genômicos durante a especiação (PEACOCK *et al.*, 2007). Diferenças no conteúdo gênico entre *L. braziliensis, L. infantum* e *L. major* foram detectadas em cerca de 200 genes sendo que 47, 27 e 5 genes são exclusivos (espécies específicos ou únicos) para *L. braziliensis, L. infantum* e *L. major*, respectivamente. É razoável supor, portanto, que esses genes provavelmente contribuem para a patogênese diferencial causada por estas espécies (PEACOCK *et al.*, 2007; SMITH; PEACOCK; CRUZ, 2007). Análises do número de cópias gênicas também revelaram variações entre vários genes como: PSA2/GP46, amastinas, GP63, alfa e beta tubulina, proteína ribossomal 40S e chaperonas, apesar de compartilharem 56 grupos de genes ortólogos (ROGERS *et al.*, 2011).

Outra diferença encontrada entre esses genomas ocorre em relação a presença de genes envolvidos nas vias de RNA de interferência (RNAi). Embora tenham genomas muito semelhantes em relação ao conteúdo e organização dos genes, a presença de sequências específicas, como retrotransposons e uma maquinaria de RNAi ativa indicam uma diversidade maior entre essas espécies (LYE *et al*., 2010; SMITH; PEACOCK; CRUZ, 2007). Essa via mostrou ser funcional pela primeira vez em *T. brucei*, onde a transfecção de dsRNA longo, homólogo ao mRNA de α-tubulina causou regulação negativa do mRNA alvo e consequente inibição da síntese da proteína correspondente (NGÔ *et al*., 1998). Contudo, efeito semelhante não foi constatado em *T. cruzi* (DAROCHA *et al*., 2004) e espécies de *Leishmania* do Novo Mundo, como *L. major* e *L. donovani* (ROBINSON; BEVERLEY, 2003), indicando que não possuem uma via funcional de RNAi, o que foi posteriormente confirmado pela ausência de genes envolvidos com a maquinaria de RNAi (EL-SAYED *et al.* 2005; IVENS *et al.*, 2005). Ao contrário, Lye e colaboradores (2010) demonstraram que *L. braziliensis* e outras espécies do subgênero *Viannia* possuem um mecanismo ativo de RNAi,

que resultou na diminuição da expressão do gene repórter e dos genes endógenos alvos. Em *L. amazonensis* há indícios da presença de genes envolvidos com a maquinaria de RNAi, contudo esta parece não ser ativa nessa espécie. Uma tentativa de identificação de genes envolvidos nesse mecanismo revelaram a ausência do gene codificador da proteína Dicer com domínio Rnc (ribonuclease específica de RNA de fita dupla (dsRNA)) em *L. amazonensis*, mas a presença de outros genes codificadores de RNAs helicases com domínio DEAD/H box e Ribonuclease III, proteínas ERI, as quais estão envolvidas na formação do complexo ERI/DICER, e dois genes do complexo RISC (tudor e piwi, pertencentes a família das argonautas), demonstrando que ainda há resquícios de uma antiga via funcional de RNAi funcional (TSCHOEKE *et al.*, 2014).

Diferenças significativas também foram notadas quando famílias multigênicas presentes nos genomas de várias espécies de leishmanias foram analisadas. Esses genes são compostos por sequências repetidas *em tandem* ou não e que podem estar diretamente ligados a eventos de invasão celular, o que contribui para a infectividade do parasito; tropismo tecidual; visceralização e o tipo de doença causada. Genes que codificam proteínas GP63 por exemplo, são encontradas em todos os genomas já sequenciados de *Leishmania* e também nos genomas de *T. cruzi* e *T. brucei*, porém o número de cópias e a distribuição cromossômica, como já citado anteriormente, varia entre cada espécie. Estas proteínas por sua vez desempenham um papel fundamental na interação parasito-hospedeiro e representam a principal protease de superfície dessas espécies (OLIVIER *et al.*, 2012). Pesquisas realizadas no banco de dados Tritryp mostrou que para *L. major* há três conjuntos de genes GP63 distribuídos nos cromossomos 28 (um gene), 31 (um gene) e 10 (quatro genes) e que uma organização semelhante também é observada para *L. infantum* e *L. mexicana*. Em *L. braziliensis*, porém, uma quantidade maior e distribuição diferencial foi observada, onde nenhum gene de GP63 foi localizado no cromossomo 8, ao passo que 33 genes recentemente identificados foram identificados exclusivamente no cromossomo 10 e outros 6 genes ou fragmentos no cromossomo 31 (CASTRO NETO *et al.*, 2019).

A família multigênica das amastinas também pode ser encontrada tanto em *Leishmania* quanto em *T. cruzi* e *T. brucei*, variando no número de cópias e distribuição no genoma. Pelo fato de codificarem proteínas específicas de formas intracelulares (amastigotas), os genomas de *T. cruzi* e de *Leishmanias* possuem um número maior de cópias desses genes, quando comparados ao genoma de *T. brucei*, que não possui uma forma intracelular (JACKSON, 2010). Contudo, os genes de amastinas são mais abundantes em espécies do gênero *Leishmania*. Enquanto que em *T. cruzi* foram anotadas 12 sequências codificadoras de amastinas (KANGUSSU-MARCOLINO *et al.*, 2013) e apenas 4 em *T. brucei*, 57 genes foram inicialmente anotados em *L. major* e distribuídos em vários cromossomos (GONZÁLEZ-DE LA FUENTE *et al.*, 2017; LYPACZEWSKI *et al.*, 2018). Em *L. braziliensis* 55 genes codificadores de amastinas também foram identificados por meio de

alinhamento realizado contra a base de dados TritryDB, considerando as espécies *L. infantum* e *T. cruzi* como *queries*. Além disso, nessa espécie, as amastinas foram categorizadas dentro das quatro subfamílias descritas por Jackon (2010) e uma análise filogenética mostrou a proximidade evolutiva entre alpha e beta amastinas. Comparações entre as sequências de amastinas de *L.* braziliensis, *L. infantum* e *T. cruzi* revelou ainda uma significativa conservação tanto em sequência quanto nas estruturas previstas entre os diferentes membros da família (DE PAIVA *et al.*, 2015).

Estudos de filogenômica realizados entre *L. amazonensis*, *L. mexicana*, *L. infantum*, *L. braziliensis* e *L. major* identificou famílias de genes ortólogos de *Leishmania*, incluindo várias das famílias anteriormente citadas. Nesse sentido, 7.826 famílias de genes ortólogos foram identificadas, das quais 6.784 são compartilhadas por outras *Leishmanias ssp*, considerando as glicoproteínas GP63 e amastinas (REAL *et al.*, 2013). Em outro estudo filogenético, seis espécies de *Leishmania* e 28 outras espécies de protozoários, incluindo *L. amazonensis*, *T. cruzi* e *T. brucei* confirmaram a posição taxonômica de *L. amazonensis* no "complexo mexicana" além de mostrar que 7.076 grupos ortólogos são compartilhados entre os genomas *L. amazonensis*, *L. donovani*, *L. mexicana*, *L. infantum*, *L. braziliensis* e *L. major*. Dentre os ortólogos anotados com função conhecida, encontram-se novamente membros da família das amastinas, cisteínas peptidase, proteína ribossômica 40S, proteína quinase, dentre outras (TSCHOEKE *et al.*, 2014).

Por meio de filogenia baseada em genes ortólogos, Patino e colaboradores (PATINO *et al.*, 2020) também mostraram que, *L. amazonensis* (La_UA301) está intimamente relacionada com *L. mexicana*, e que estas se agrupam em um *cluster* próximo às espécies de *Leishmania* do Velho Mundo (*L. infantum*, *L. donovani* e *L major)*. Em outra análise, intra-espécie, foi revelado que existe uma estreita relação entre as sequências de La_UA301 e RZOD01.1 (isolado proveniente da Guiana Francesa), formando um nó independente. Contudo, essas relações precisam ser confirmadas, o que só será possível introduzindo um número maior de sequências provenientes de diferentes hospedeiros e regiões do genoma.

1.5. Genes codificadores de fatores de virulência de *L. amazonensis*

Espécies do gênero *Leishmania* possuem em seus genomas uma vasta gama de genes que codificam fatores de virulência, importantes para a invasão celular e manutenção de sua sobrevivência nos diferentes hospedeiros, incluindo moléculas que modulam ou contribuem para evasão da resposta imune (BIFELD; CLOS, 2015).

Diversos trabalhos têm caracterizado fatores de virulência em *Leishmania*. Estes são moléculas que permitem o contato do parasito com as células hospedeiras, o estabelecimento e manutenção da infecção, contribuindo dessa forma para a patogênese da doença

(MATLASHEWSKI, 2001). Nas formas promastigotas, os principais fatores de virulência estão no glicocálice, que reveste toda a superfície do parasito, e são codificados por famílias multigênicas. As camadas superficiais dos tripanosomatídeos em geral, apresentam uma diversidade significativa na composição, mas todas as moléculas encontram-se ancoradas a superfície por meio de um motivo de glicosilfosfatidilinositol (GPI) altamente conservado (MCCONVILLE; FERGUSON, 1993).

Glicanos fosfoglicilados ancorados em GPI é a estrutura celular predominante na superfície de promastigotas, sendo o lipofosfoglinaco (LPG) um dos glicoconjugados mais abundantes nesta forma de vida, com aproximadamente $5x10^6$ cópias por célula, uniformemente distribuídos (TURCO; DESCOTEAUX, 1992). Esta molécula é fundamental para a fixação do parasito no intestino dos flebotomíneos, além de conferirem resistência contra o sistema complemento humano, tornando-as essenciais para o estabelecimento da infecção por estes vetores (GRIMM; JENNI, 1993; SACKS *et al.*, 2000). LPGs de *L. amazonensis* (cepas PH8 e Josefa) mostraram ser eficazes na ativação e modulação da resposta imunológica inata, ativando TLR4, MAPKs e o inibidor de NF-κB, p-IκBα (NOGUEIRA *et al.*, 2016). Além disso, vários trabalhos têm mostrado o papel deste fator de virulência na modulação da resposta gerada pelos macrófagos infectados, impedindo por exemplo, a maturação do fagossomo e a produção de óxido nítrico, de forma a permitir que as promastigotas fiquem tempo suficiente dentro desse compartimento até se diferenciarem em amastigotas (ALBERT DESCOTEAUX; SALVATORE J. TURCO, 1999; DERMINE *et al.*, 2000, 2005; DESJARDINS; DESCOTEAUX, 1997; MOSMANN, 1983; STEPHEN M. BEVERLEY; SALVATORE J. TURCO, 1998). Contudo, LPG não parece ser um fator de virulência essencial na infectividade de *L. mexicana* (ILG; DEMAR; HARBECKE, 2001).

Alguns loci gênicos têm sido descritos por possuírem papel fundamental na invasão dos macrófagos, são os casos das proteínas GP63 e A2. Ambas são codificadas por famílias multigênicas e expressas predominantemente nas formas promastigota e amastigota, respectivamente, em *Leishmania*, mas também são expressas em tripanosomas (KULKARNI *et al.*, 2009). GP63 ou leishmanolisina pertence à família de metaloproteases dependentes de zinco são consideradas as principais proteases de superfície. Seu papel durante a infecção em *Leishmania* está associado principalmente a eventos que facilitam a interação do parasito com receptores dos macrófagos, permitindo sua posterior fagocitose (OLIVIER *et al.*, 2012).

Ainda no meio extracelular, GP63 atua inativando a cascata do sistema complemento, clivando C3b em iC3b, aumentando a resistência do parasita à lise mediada por complemento, mas permitindo sua opsonização. Segue facilitando a comunicação com macrófagos por meio de receptores de fibronectinas e reduzindo TNF, produção de óxido nítrico e IL-12 já dentro da célula. Foi visto também que GP63 pode ser secretada no meio extracelular por meio de exossomos ou

intracelular, por promastigotas e amastigotas, respectivamente, facilitando a absorção pelos macrófagos das promastigotas ou reduzindo a ativação celular e a atividade microbicida contra amastigotas e promastigotas (GUPTA; OGHUMU; SATOSKAR, 2013; MCGWIRE; CHANG, 1994; OLIVIER *et al.*, 2012). Macrófagos derivados da medula óssea e células B-1 peritoneais estimulados com vesículas extracelulares de *L. amazonensis* contendo LPG e GP63 contribuíram para o aumento de IL-6 e IL-10 e TNFα, favorecendo a sobrevivência do parasito e a progressão da doença (BARBOSA *et al.*, 2018).

O gene de A2 está diretamente relacionado ao tropismo diferencial de *Leishmania*. Esses genes codificam uma família de proteínas de 42 a 100 kDa composta por 40 a 90 cópias de uma sequência repetida de aminoácidos, que varia em seu tamanho, assim como o número de cópias desses genes varia por espécie (WEN-WEI ZHANG *et al.*, 1996). A expressão dessas proteínas ocorre principalmente em amastigotas de *L. donovani* e *L. infantum*, onde estão envolvidas com a visceralização do parasito (MATLASHEWSKI, 2001; ZHANG; MATLASHEWSKI, 2000). Em *L. major*, agente etiológico da leishmaniose cutânea, a sequência codificadora de A2 aparece como um pseudogene truncado (PEACOCK *et al.*, 2007). Mas há estudos que mostram que a expressão exógena desse gene aumenta a migração do parasito para fora da derme e sua consequente multiplicação em órgãos viscerais (ZHANG *et al.*, 2003). O mesmo aconteceu quando *L. tarentolae*, um parasito de lagarto, foi transfectada com A2 episssomal. Infecções *in vivo* aumentaram significativamente a capacidade de *L. tarentolea* sobreviver no fígado de camundongos BALB/c (MIZBANI *et al.*, 2011). Recentemente *L. amazonensis* tem sido associada a casos de leishmaniose visceral em cães e por isso o papel de moléculas envolvidas na visceralização deve ser investigado nesta espécie. Contudo, o gene que codifica a proteína A2 encontra-se colapsado, devido ao seu conteúdo repetitivo (VALDIVIA *et al.*, 2017).

Uma outra família multigênica envolvida na expressão de fatores de virulência em Leishmania são as amastinas. Esta família codifica proteínas de superfície com características hidrofóbicas, altamente glicosiladas de ~200 aminoácidos. São expressas tanto em amastigotas de *Leishmania* quanto de *T. cruzi*, porém sofreram uma expansão em espécies de *Leishmania* (ROCHETTE *et al.*, 2005). Em *T. brucei*, que não possui uma fase intracelular foram encontrados 4 genes altamente divergentes, porém sua função nesta espécie não está elucidada (ASLETT *et al.*, 2010). Jackson (2010) mostrou pela primeira vez um estudo evolutivo das amastinas de tripanosomatídeos e concluiu, baseado na diversidade das sequências, que essa família está dividida em quatro subfamílias que foram denominadas *alpha (α)*, *beta (β)*, *gama (*γ) e *delta (*δ)-amastinas. Além disso, verificou que elas possuem organização genômica e padrões de expressão distintos em *Leishmania* e *T. cruzi* (KANGUSSU-MARCOLINO *et al.*, 2013; TEIXEIRA; KIRCHHOFF; DONELSON, 1995) e que seu papel está associado ao aumento da infectividade em leishmanias.

Knockdown de δ-amastinas obtido por meio de RNA de interferência em *L. braziliensis* alterou a interação parasito-macrófago e prejudicou a viabilidade de amastigotas intracelulares, uma vez que o crescimento em macrófagos infectados *in vitro* foi menor quando comparado aos parasitas wt e falhou completamente em produzir infecção em camundongos BALB/c (DE PAIVA *et al.*, 2015). Além disso, a expressão do gene amastina também foi detectada predominantemente em amastigotas de várias cepas de *L. donovani* isoladas de pacientes com leishmaniose dérmica visceral pós cala-azar (SALOTRA *et al.*, 2006). Existe ainda evidências do papel dessas proteínas como epítopos de superfície, onde são reconhecidas por anticorpos IgGs do hospedeiro durante o processo de opsonização (NADERER; McCONVILLE, 2008).

Cisteína proteases ou peptidases também são importantes fatores de virulência expressos por grandes famílias multigênicas. Classificadas como enzimas proteolíticas, sua função biológica tem sido associada à degradação de nutrientes, modulação de mecanismos do sistema imunológico da célula hospedeira, ativação da inflamação em hospedeiros mamíferos, diferenciação celular do parasita e processamento de outras proteínas-chave (SILVA-ALMEIDA *et al*., 2012). Em espécies do complexo *L. mexicana*, as cisteína proteases representam a maioria das proteases expressas e têm sido caracterizadas a partir de infecções em modelos murinos. Em *L. amazonensis*, alta atividade proteolítica foi detectada em extratos de amastigotas, enquanto em promastigotas de fase exponencial ou estacionária essa atividade foi muito baixa (LASAKOSVITSCH *et al*., 2003). Além disso, verificou-se que essa propriedade está ligada à capacidade de degradar moléculas do MHC classe II, impedindo a consequente apresentação do antígeno ao sistema imunológico da célula hospedeira (DE SOUZA LEAO *et al*., 1995).

Vários estudos que empregam o uso de inibidores ou nocautes de cisteína proteases têm ajudado a compreender seu impacto no estabelecimento da infecção por *Leishmania* e na patogênese da doença. A supressão dos genes da cisteína protease diminuiu a virulência e a sobrevivência de macrófagos em parasitas nocaute de *L. infantum* em hamsters e de *L. chagasi* em culturas de células humanas (MUNDODI; KUCKNOOR; GEDAMU, 2005; POOT J, DENISE H, HERRMANN DC, MOTTRAM JC, COOMBS GH, 2006). Em *L. tropica*, parasitas tratados com N-Pip-F-hF-VS Fenil reduziram a viabilidade, crescimento e patogenicidade desta espécie (MAHMOUDZADEH-NIKNAM; MCKERROW, 2004). Outro inibidor de protease de cisteína demonstrou ter efeito no equilíbrio da resposta Th que ocorre através da clivagem de proteínas do MHC em *L. amazonensis*. A clivagem de moléculas de MHC classe II dentro do vacúolo parasitóforo pode inibir parcialmente a resposta imune ou ser substituída e mediada por produtos gênicos característicos de MHC classe I (DAS *et al*., 2001).

Várias proteínas envolvidas no metabolismo de ferro e de heme de *Leishmania* foram identificadas como fatores de virulência devido à sua importância no estabelecimento e progressão

da doença (LARANJEIRA-SILVA; HAMZA; PÉREZ-VICTORIA, 2020). Estudos com *L. amazonensis* mostraram que uma redutase férrica 1 (LFR1) reduz o ferro, que é translocado para o citosol do parasita pelo transportador de ferro ferroso LIT1, e os níveis de ferro citosólico são mantidos pelo exportador de ferro LIR1 (FLANNERY; RENBERG; ANDREWS, 2013; HUYNH; SACKS; ANDREWS, 2006; LARANJEIRA-SILVA *et al*., 2018; SARKAR; ANDREWS; LARANJEIRA-SILVA, 2019). O ferro também pode ser adquirido pela captação de heme via LHR1 e LFLVCRb, ou diretamente pela hemoglobina via endocitose do receptor de hemoglobina (HbR), como demonstrado em *L. amazonensis*, *L. major e L. donovani*, respectivamente (AGARWAL *et al*., 2013; CABELLO-DONAYRE *et al*., 2019; MIGUEL *et al*., 2013). Outros estudos com *L. amazonensis* também identificaram *Leishmania* mitoferrina 1 (LMIT1) como transportador de ferro para as mitocôndrias do parasita (MITTRA *et al*., 2016), onde ferro e heme são cofatores de *Leishmania* superóxido dismutase A (SODA) e ascorbato peroxidase (APX), respectivamente (MITTRA *et al*., 2017; XIANG *et al*., 2019). O ferro também é cofator das isoformas glicossômicas do SODB, que se mostra essencial para a *Leishmania* (DAVENPORT *et al*., 2018; PLEWES; BARR; GEDAMU, 2003). Por fim, LABCB3 foi identificado como um transportador ABC mitocondrial incomum necessário para a maturação de aglomerados citosólicos de ferro/enxofre em *L. major* (MARTÍNEZ-GARCÍA *et al*., 2016).

1.6. Genômica funcional de tripanosomatídeos

Na sua definição mais simples, entende-se por genômica funcional, o campo da biologia molecular que estuda a função do DNA (incluindo genes e elementos não gênicos), proteínas codificadas pelo mesmo ou produtos de ácidos nucleicos. Mas, de modo mais profundo, essa é a área que tenta descrever a complexa relação entre fenótipo e genótipo no nível do genoma, focando na coleção completa do DNA, RNA ou proteínas de um organismo. Portanto, é uma abordagem que visa o estudo da estrutura, função e regulação de todos os genes de maneira integrada e em um contexto definido, ou seja, no curso de uma doença ou em algum estágio do desenvolvimento (PEVSNER, 2015). Sendo assim, podemos dizer ainda que, a genômica funcional lida com aspectos dinâmicos, como a transcrição gênica, tradução, regulação da expressão gênica e interações proteína-proteína (BUNNIK; LE ROCH, 2013).

Levando em consideração o fato de que atribuir significado biológico aos genes em nível de genoma é um dos objetivos basais da genômica funcional (STEINMETZ; DAVIS, 2004), faz-se importante neste cenário a geração de grandes quantidades de dados capazes de fornecerem informações acerca de todos os elementos nele contido. Técnicas baseadas em *knockout*, RNA de interferência (RNAi), sistema CRISPR, sequenciamento e anotação de genomas e/ou tanscriptomas,

têm sido amplamente utilizadas para entender o papel de diversos genes de patógenos de grande importância clínica, como por exemplo, as leishmanioses (BARTHOLOMEU et al., 2021). Paralelo a geração massiva de dados e ao desejo de encontrar padrões biologicamente significativos, a bioinformática se estabelece como objeto de estudo essencial, que deve ser também amplamente empregada. A combinação de todo esse conhecimento possibilita, por fim, integrar as respostas obtidas por meio de análises genômicas, transcriptômicas, proteômicas e demais ômicas (LAPATAS et al., 2015).

Compreender os mecanismos de controle da expressão gênica têm ajudado a entender por exemplo, os diferentes fenótipos apresentados por espécies de *Leishmania* ou ainda a rápida adaptabilidade a ambientes hospedeiros distintos, observado também nas infecções por *T. cruzi* e *T. brucei* (BARTHOLOMEU; TEIXEIRA; CRUZ, 2021). Isso acontece porque existe uma relação direta entre os níveis de expressão de um gene em determinado contexto e sua função. Em busca de uma visão "macro" da função do genoma e/ou transcriptoma durante a infecção provocada por estes parasitos, vários estudos foram conduzidos valendo-se de análises de expressão diferencial de genes e ontologia gênica. Nesse sentido, as principais diferenças entre os transcriptomas dos estágios do flebotomíneo de *Leishmania* comparando organismos mantidos *in vivo* e *in vitro* foram relatadas, revelando uma assinatura única para cada estágio de diferenciação (INBAR et al., 2017). Além disso, macrófagos infectados por *L. major* e *L. amazonensis in vitro* revelaram uma resposta vigorosa e específica do parasita no início da infecção, que foi bastante atenuada logo após a sua entrada (FERNANDES et al., 2016). Ainda, fibroblastos humanos infectados com uma cepa virulenta e outra avirulenta de *T. cruzi* mostraram existir mudanças significativas na expressão de genes relacionados a proteínas de superfície, resposta imune e organização do citoesqueleto, principalmente após 96 horas de infecção, justificando os fenótipos de virulência contrastantes (BELEW et al., 2017; OLIVEIRA et al., 2020).

Neste cenário, ferramentas e protocolos de manipulação do genoma surgem como poderosos mecanismos na elucidação do papel de diversos elementos por ele codificados. Um trabalho pioneiro descreve o primeiro experimento envolvendo a deleção gênica em *L. major*, que resultou na substituição de um único alelo de DHFR-TS em promastigotas transfectadas com dsDNA linear, contendo sequências homólogas ao gene alvo (CRUZ; BEVERLEY, 1990). Posteriormente, Cooper *et al*. (1993) mostraram de forma análoga, a ruptura dos dois alelos do gene que codifica a molécula de adesão flagelar GP72 em *T. cruzi*, por recombinação homóloga. Tal feito foi alcançado após a transfecção de epimastigotas com um vetor plasmidial contendo os genes da neomicina fosfotransferase e da higromicina fosfotransferase flanqueados por sequências GP72. Por fim, em meados dos anos 2000, a tecnologia de interferência de RNA (RNAi) surgiu como uma importante ferramenta em análises de genética reversa, sendo inicialmente empregada em

experimentos de *knockdown* em *T. brucei*, como já relatado na seção anterior (NGÔ et al., 1998). Uma análise sistemática de RNAi e fenotípica foi realizada em 210 genes do cromossomo 1 de formas sanguíneas de *T. brucei* e fenótipos foram documentados para mais de 30% do total de genes alvos (SUBRAMANIAM et al., 2006).

Uma melhoria recente das técnicas de manipulação de genomas veio com a tecnologia de CRISPR/Cas9. Diversos trabalhos, como os de Cong *et al*. (2013), e Mali *et al*. (2013) demonstram o uso desse sistema para a alteração de material genético em células de mamíferos e humanos, respectivamente. A possibilidade de produção de RNAs guias que substituam o crRNA bacteriano permite a escolha de sequências alvo, desde que se tenha conhecimento da sequência de interesse. Os avanços no sequenciamento genético de diversos organismos vão ao encontro dessa tecnologia emergente, possibilitando a manipulação genética em diversos organismos.

Vários estudos para a aplicação do sistema CRISPR/Cas9 em diferentes espécies de *Leishmania* foram publicados demonstrando a validade para aplicação do sistema nesse grupo de organismos. Em 2015, Sollelis *et al* (SOLLELIS *et al*., 2015), foram responsáveis pelo primeiro trabalho empregando essa tecnologia em espécies desse gênero, promovendo o *knockout* em *Leishmania major* de três genes da proteína paraflagellar rod-2, que estão organizados *em tandem* no cromossomo 16. Os resultados do grupo revelaram o sucesso na obtenção de parasitos *knockout* a partir de análises por PCR, FISH, *western blotting* e imunofluorescência. Adicionalmente, foi realizado sequenciamento completo de clones *knockout* e parasitos *wild-type* (wt) para avaliação de efeitos *off target* decorrentes do uso do sistema. Os resultados estavam de acordo com o previsto, demonstrando a especificidade dessa ferramenta.

No mesmo ano, Zhang e Matlashewski induziram o *knockout* do gene da proteína transportadora de miltefosina em *L. donovani* utilizando um plasmídeo para a expressão de Cas9 e plasmídeos para a expressão dos guias *in vivo* pelos parasitos. Esse trabalho ainda demonstrou a precisão da ferramenta em termos de favorecimento do reparo mediado por sequência homologa, permitindo ainda a inserção de sequências no genoma editado (ZHANG; MATLASHEWSKI, 2015). Em 2017, os autores apresentaram novos resultados, demonstrando refinamento da metodologia além de demonstrar funcionalidade da mesma para as espécies *L. mexicana* e *L. major*. No mesmo ano, Beneke *et al* apresentaram um novo protocolo para o fornecimento de RNAs guia, a partir da expressão da T7 RNA polimerase e transfecção das sequências de DNA usadas em transcrições *in vitro* para promover síntese de sgRNAs *in vivo* (BENEKE *et al*., 2017).

Peng *et al.* publicaram, em 2015, o primeiro trabalho de CRISPR/Cas9 envolvendo *Trypanosoma cruzi*. Os autores transfectaram parasitos da cepa CL, com plasmídeos para expressão de Cas9 e de proteína fluorescente verde (eGFP). Posteriormente, sgRNAs, previamente transcritos *in vitro*, para interrupção do gene de eGFP foram dados aos parasitos com posterior avaliação de

fluorescência. A partir do segundo dia pós-transfecção foi possível observar diminuição dos níveis de fluorescência com acentuação desse perfil até o quinto dia. Posteriormente, os autores demonstraram a viabilidade do protocolo para indução de alteração em genes endógenos. Ensaios revelaram alterações funcionais na edição de genes de alfa-tubulina, histidina amônia-liase e um transportador de ácidos graxos. Outros resultados importantes do trabalho sugerem que a via de reparo mediado por micro-homologia (MMEJ) parece ser a via favorecida em *T. cruzi* (PENG *et al.*, 2015). Burle-Caldas *et al* em 2018 demonstrou a possibilidade de utilização da proteína recombinante SaCas9 para induzir *knockout* do gene GP72 em *Trypanosoma cruzi*. O mesmo resultado foi obtido com a utilização de Cas9 expressa constitutivamente pelos parasitos. Esses trabalhos abriram possibilidade para estudos mais profundos, a partir da realização de *knockouts* e *knock-ins* direcionados a elucidar melhor o mecanismo fisiopatológico das leishmanioses e doença de Chagas, assim como abrindo fronteiras para a identificação de novos alvos farmacológicos para tratamento e profilaxia dessas enfermidades (BURLE-CALDAS *et al*., 2018).

2. JUSTIFICATIVA

As leishmanioses são antropozoonoses consideradas um grave problema de saúde pública, representando um complexo de doença com importante espectro clínico e diversidade epidemiológica. A leishmaniose cutânea é considerada a forma de infecção mais comum causada pela *Leishmania* e afeta de 0.7-1 milhão de pessoas por ano (WHO, 2022).

No Brasil, *L. amazonensis* é a espécie responsável por mais de 8% de todos os casos de LC humana em regiões endêmicas, sobretudo nas regiões norte e nordeste (INES *et al.*, 2011; PAULO *et al.*, 2007). Apesar de raros, alguns casos com manifestações viscerais foram relatados em cães de regiões endêmicas para a leishmaniose visceral (sudeste brasileiro), cujo agente etiológico predominante é da espécie *Leishmania (L.) infantum* (ALVES SOUZA *et al.*, 2019; HOFFMANN *et al.*, 2013; VALDIVIA *et al.*, 2017, PORTO *et al.*, 2022). Somado a isso, características exclusivas relacionadas à progressão da doença e a resposta imune desencadeada pela *L. amazonensis* tornam essa espécie ainda mais complexa e desafiadora no que diz respeito à compreensão de todos os aspectos imunopatológicos que a envolvem. Diversos estudos relatam a susceptibilidade de camundongos C57BL/6, BALB/c e C3H.HeN à infecção por *L. amazonensis*, enquanto outras espécies também causadoras da LT, como *L. major* (subgênero *Leishmania*) e *L. braziliensis* (subgênero *Vianna*) não são capazes de estabelecer uma infecção persistente. Baixa taxa de ativação de células T, além da regulação negativa de moléculas do MHC de classe II, quimiocinas inflamatórias e seus respectivos receptores estão associadas a resposta estimulada pela infecção por *L. amazonensis*, enquanto *L. major* apresenta um perfil de resposta contrário a este (MENDES WANDERLEY et al., 2012). Outros aspectos moleculares e fenotípicos também já foram relatados para diferentes cepas da mesma espécie, como é o caso da cepa PH8 e LV79, onde a PH8 apresentou níveis de virulência mais altos refletidos em lesões maiores e não cicatrizantes, além de maior expressão de moléculas de superfície importantes durante o processo de invasão de células hospedeiras (DE REZENDE *et al*., 2017).

Diante do exposto é nítida a necessidade de gerar novas sequências para diferentes cepas de *L. amazonensis* afim de expandir o conhecimento acerca da diversidade patológica e molecular dessa espécie, além dos mecanismos de infecção e sobrevivência, como eventos de visceralização. Para isso faz-se necessário obter sequências de referências bem montadas e anotadas para diversas cepas, disponíveis em bancos de dados. Nesse sentido, para *L. amazonensis* já foram gerados três conjuntos de genomas, os quais estão disponíveis no repositório de genomas do NCBI, um depositado em 2013 (APNT00000000.1) e dois em 2019 (RZOD00000000.1 e ASM531712v1), onde ambos foram isolados de infecções humanas. O APNT00000000.1 está montado em 2.627 *scaffolds*, com um tamanho de genoma de ~29 MB e é o único com anotação disponível. Já o

RZOD00000000.1 está montado em 92 *contigs* com um tamanho de genoma de ~32 MB, enquanto o ASM531712v1 alcançou nível cromossômico, tendo sido montado em 34 *scaffolds*, totalizando ~32 MB, porém, essa última montagem foi baseada em referência, acarretando na transferência de possíveis erros provenientes do genoma utilizado como referência, no caso, *L. mexicana* MHOM/GT2001/U1103. Além disso, a cepa PH8 de *L. amazonensis* é a mais estudada no Brasil, tendo sido empregada em diversos estudos que visam investigar aspectos básicos da infecção, bem como estudos para o desenvolvimento de vacinas contra a leishmaniose tegumentar (ARAÚJO *et al*., 2008; DUARTE *et al*., 2016; FERRAZ COELHO *et al*., 2003; MARZOCHI *et al*., 1998; MAYRINK *et al*., 1979). Frente a este cenário é notória a necessidade de obtenção de genomas representativos de outras cepas de *L. amazonensis*, que sejam contíguos e melhores anotados, o que devido ao avanço das técnicas de sequenciamento torna possível este trabalho. Além disso, genomas completos permitem a realização de anotações consistentes e o estudo de vários genes e outros elementos envolvidos na infeção e progressão da patogênese da doença.

Sendo assim, para que haja o desenvolvimento de protocolos eficazes para o controle e combate da doença, bem como o desenvolvimento de novas vacinas e medicamentos é fundamental o entendimento dos elementos genômicos que regem o curso de eventos de invasão celular, diferenciação e visceralização de amastigotas de *L. amazonensis*, dentre eles, proteínas codificadas por famílias gênicas são importantes fatores de virulência amplamente estudados e evidentemente importantes em todos esses processos.

3. OBJETIVOS

3.1. Objetivo Geral

- Obter e caracterizar a sequência completa da cepa PH8 de *Leishmania amazonensis* e investigar a variabilidade dos genes sabidamente associados aos processos de invasão, diferenciação e visceralização dessa espécie.

3.2. Objetivos específicos

- Sequenciar e montar o genoma completo da cepa PH8 de *L. amazonensis* utilizando *reads* geradas nas plataformas Illumina e Pacbio;
- Anotar o genoma *de novo* e com base no genoma de referência *L. mexicana*;
- Analisar o conteúdo gênico, número de cópias de genes cópia simples e genes ortólogos;
- Investigar a variação no número de cópias cromossômicas e seu impacto no aparecimento de expansões gênicas;
- Anotar e analisar o genoma do maxicírculo da PH8;
- Realizar a anotação automatizada das principais famílias multigênicas envolvidas nos processos de invasão, diferenciação e visceralização da PH8: amastinas, GP63, proteínas A2, cisteína proteases, fosfatases e cinases;
- Comparar genes codificadores de proteínas do metabolismo de heme e ferro com outras espécies de *Leishmania*;
- Realizar a análise evolutiva dos genes codificadores de amastinas considerando os principais membros dos tripanosomatídeos: *L. amazonensis*, *L. braziliensis*, *L. donovani*, *L. infantum*, *L. major*, *L. mexicana*, *T. cruzi* e *T. brucei*;
- Caracterizar os genes codificadores de GP63 e proteínas A2 de *L. amazonensis*;
- Realizar um levantamento do repertório de cisteína proteases de *L. amazonensis* e comparar as demais espécies do gênero.

4. METODOLOGIA

4.1. Extração de gDNA e sequenciamento Ilumina e PacBio

DNA genômico (gDNA) foi obtido a partir de culturas de promastigotas da cepa PH8 de *L. amazonensis,* isolado de *Lutzomyia flaviscutellata* na cidade de Belém, Brasil. Essa cultura foi mantida no laboratório da Profa. Ana Paula Fernandes e gentilmente cedida para extração e sequenciamento do DNA. A extração e purificação do gDNA foi realizada no Laboratório de Genômica de Parasitos - UFMG, coordenado pela Profa. Daniella Castanheira Bartholomeu utilizando kit de extração de DNA da Life Technologies. A amostra foi enviada para a empresa sul-coreana Macrogen, e sequenciada utilizando TruSeq™ RNA and DNA Library Preparation Kits v2 para produzir *reads paired-end* de 350 pb de tamanho, na plataforma Illumina Hiseq 2000. Uma segunda amostra de *L. amazonensis* foi obtida a partir da extração e purificação de gDNA de $6x10^8$ promastigotas utilizando Wizard Genomic DNA Purification kit (Promega). A amostra foi sequenciada no Instituto Karolinska, Suécia, no laboratório do Dr. Bjorn Andersson, utilizando a plataforma PacBio RS II e o kit DNA Sequencing Reagent kit 4.0 v2, que produz *reads* superiores a 1kb.

4.2. Análise da qualidade de *reads* e montagem genoma da cepa PH8

A qualidade das *reads* brutas da Illumina foi verificada usando FastQC v0.11.3 (ANDREWS, 2020). Sequências de adaptadores e regiões de baixa qualidade foram removidas usando Trimmomatic v.0.39 (BOLGER; LOHSE; USADEL, 2014) e os seguintes parâmetros de corte: ILLUMINACLIP:TruSeq3-PE-2.fa:2:20:10 LEADING: 3, TRAILING: 3, SLIDINGWINDOW: 3: 15, MILÃO: 75.

Para obter a sequência completa do genoma da cepa PH8 foi considerada uma abordagem de montagem *de novo* com posterior correção, onde ~86x de *reads* Pacbio foram combinadas com ~43x de *reads* Illumina. O montador Canu v.1.5 (KOREN *et al.,* 2017) foi usado para a montagem de *contigs* iniciais, considerando apenas as *reads* PacBio. Posteriormente, parte do *pipeline* IPA (OTTO, 2007) foi usado para remover *contigs* redundantes (sobreposição ≥ 99%) e excluir *contigs* menores que 5kb. A etapa de *scaffolding* foi executada usando SSPACE v.3.0 (BOETZER *et al.,* 2011) e o preenchimento de *gaps* com o GapFiller v.1.10 (NADALIN; VEZZI; POLICRITI, 2012), com base em *reads* curtas *paired-end* Illumina. Finalmente, a montagem foi polida usando Pilon v.1.22 (WALKER *et al.*, 2014), onde novamente *reads* Illumina, alinhadas ao genoma pré-montado

foram usadas. Os *scaffolds* obtidos foram então ordenados com base nos cromossomos de *L. mexicana* usando Abacas v.1.3.1 (ASSEFA *et al.*, 2009), onde uma cobertura mínima igual a 20 foi considerada.

Para avaliar a qualidade da montagem do genoma, usamos uma estrutura projetada para avaliar montagens *de novo*, dnAQET e o genoma de *L. mexicana* U1103 (Release 52) foi passado como referência de genoma para alinhamento realizado para minimap2. Para a completude do genoma, foi empregada a métrica *Benchmarking Universal Single-Copy Orthologs* (BUSCO, v5.1.2) (MANNI *et al*., 2021), que é complementar ao N50. Pesquisas contra um conjunto de dados de genes ortólogos de cópia única que se espera estarem presentes na linhagem protista (euglenozoa_odb10) ou outros eucariotos (eukaryota_odb10) foram realizadas com a opção de predição de genes configurada para o *software* Augustus (NACHTWEIDE; STANKE, 2019), treinado para *L. tarentolae*. E, por último, a cobertura do genoma foi calculada a partir do mapeamento das bibliotecas de *reads* Illumina e Pacbio para a montagem do PH8. O algoritmo BWA-MEM foi executado com parâmetros padrão e a contagem de *reads* mapeadas para cada posição no genoma (-d flag) foi realizada usando ferramentas BED genomacov.

4.3. Anotação do genoma nuclear e do maxicírculo

Nosso pipeline prosseguiu com a realização da anotação automática do genoma da cepa PH8. Para isso a anotação disponível do genoma nuclear de *L. mexicana* MHOM/GT/2001/U1103 foi transferida usando Rapid Annotation Transfer Tool (RATT) (OTTO *et al.*, 2011). O algoritmo da ferramenta verifica a sintenia entre a sequência do genoma de referência e a sequência de consulta e descarta blocos sintéticos com identidade inferior a 40%. Por esta razão, o arquivo de saída foi interrogado para identificar possíveis genes que não são sintênicos ou que não estão presentes em nossa montagem. Em paralelo, uma etapa de anotação *de novo* foi realizada usando o *software* Companion, que executa o *software* AUGUSTUS, treinado para *L. mexicana*. Gráficos de alinhamento entre os *scaffolds* da PH8 e os cromossomos da referência foram construídos utilizando a ferramenta Nucmer v.4.0 (DELCHER *et al.*, 2002).

RepeatModeler v2.0.2 (http://www.repeatmasker.org/RepeatModeler) (FLYNN *et al*., 2020) foi usado para construir uma biblioteca de repetições *de novo* usando o banco de dados de *Leishmania* spp. (TritrypDB Release 52) e, em seguida, RepeatMasker v 4.1.2-p1 (TARAILO-GRAOVAC; CHEN, 2009) (http://www.repeatmasker.org) foi executado com o parâmetro de mecanismo de pesquisa configurado para utilizar “ncbi” como banco de dados na anotação de sequências repetitivas.

O genoma do maxicírculo da cepa PH8 foi identificado pelo alinhamento de *scaffolds* não incorporados em pseudocromossomos previamente identificados. Blastn foi usado para interrogar o banco de dados do NCBI considerando 90% de identidade, cobertura de 80% e e-value igual a 10-5. Para anotar o genoma do maxicírculo, foi utilizada a ferramenta RATT, porém o genoma de referência utilizado foi *L. tarentolae* (cepa UC), conforme descrito por Camacho e colaboradores (2019) (CAMACHO *et al*., 2019a).

4.4. Análise do número de cópias cromossômicas e cópias gênicas

A estimativa de ploidia foi feita usando CADIn (Coque**i**ro-dos-Santos *et al.*, submetido), que se baseia na frequência alélica de posições heterozigóticas e variações de r*ead deep* (RD). Para as estimativas de ploidia de frequência alélica, apenas posições SNP heterozigotas com duas (e apenas duas) variantes são consideradas. Para as estimativas de ploidia RD, apenas genes que têm >= 50% de seu comprimento coberto por *reads* são usados. Os testes iterativos de Grubb são usados para remover genes que apresentam coberturas discrepantes quando comparadas à cobertura do cromossomo correspondente, de modo que duplicações e deleções locais sejam excluídas. A cobertura gênica mediana é normalizada pela cobertura do genoma, e os valores normalizados são usados para estimar a somia de seus respectivos cromossomos. O suporte estatístico é avaliado usando o teste de classificação de Mann-Whitney-Wilcoxon, para avaliar se a soma cromossômica é menor que 0,5 e 1, e maior que 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5 e 5 vezes a cobertura do genoma haploide. Por fim, o CADIn gera arquivos tabulares com SNPs, informações de cobertura e profundidade de *read*, antes e depois da análise estatística, bem como gráficos de frequência alélica e RD para cada cromossomo. Para diferenciar genes de cópia única e expandidos, estabelecemos um ponto de corte de 1,80.

Para encontrar funções enriquecidas em genes expandidos, analisamos códigos de ontologia de genes super-representados usando ferramentas TritrypDB (versão 52) (AQUINO *et al*., 2021). Genes ortólogos de *L. mexicana* foram usados como entrada e a análise foi realizada para encontrar processos biológicos enriquecidos sob valor de $p < 0,05$.

4.5. Anotação e caracterização *in silico* de fatores de virulência codificados por grandes famílias multigênicas

Membros de famílias multigênicas que codificam fatores de virulência como amastinas, metaloproteases GP63, proteínas A2, fosfatases, cisteína proteases e quinases foram identificados usando um pipeline automatizado modificado (Apêndice I) desenvolvido inicialmente por Wei-

Wang e colaboradores (2021) (WANG *et al.*, 2021), o qual combina *scripts* escritos em perl e python. Resumidamente, genes candidatos com pelo menos 150 pb são previstos pela busca de membros anotados de cada uma das famílias em outras espécies de *Leishmania*, *T. cruzi* e *T. brucei* (TriTrypDB release 52), no genoma montado da cepa PH8. Nesta etapa, o algoritmo Blastn (versão ncbi-blast-2.11.0+) é executado, com os argumentos num_alignments e max_hsps configurados em 100 e o argumento perc_identity em 85. Posteriormente, os genes previstos são alinhados às transcrições da mesma espécie (configurações de argumento do BLAST: num_alignments e max_hsps definidos como 50, perc_identity não definido). Por fim, os limites dos genes candidatos são definidos com base nas coordenadas de correspondências aos genes modelo de cada família de genes considerados e, se necessário, foram realizadas correções manuais dos limites. Uma etapa de validação adicional dos genes candidatos previstos foi realizada usando InterProScan (QUEVILLON *et al*., 2005).

O arquivo fasta contendo as sequências de todos os genes previstos para amastinas, metaloproteases GP63, proteínas A2, fosfatases, cisteína proteases, quinases e proteínas do metabolismo de ferro e heme foi refinado de acordo com o resultado fornecido pela análise de domínio. Se nenhum domínio foi identificado para uma determinada sequência ou se o domínio previsto não correspondeu aos membros da família multigênica endereçada, a sequência foi removida do arquivo final. Nas demais sequências, também foram realizados ajustes manuais dos limites de sequência, quando necessário.

As assinaturas alfa (α), beta (β), gama (γ) e delta (δ)-amastinas foram descobertas de acordo com o método descrito por Jackson (2010). Alinhamentos entre as assinaturas de amastinas e sequências presentes no genoma PH8 foram realizados usando Blastp com e-value definido para 10^{-10}.

As sequências de aminoácidos de amastinas, GP63, proteínas A2 e heme/ferro foram alinhadas usando o programa MAFFT v7 entre cada família (KATOH *et al.*, 2002) e uma busca por domínios conservados foi realizada manualmente para todas elas. Para analisar os agrupamentos formados entre as subfamílias de amastinas e as distâncias evolutivas entre elas, foi realizada uma reconstrução filogenética de máxima verossimilhança (ML) usando PhyML v3.0 (GUINDON *et al.*, 2009). O melhor modelo de substituição foi definido como LG+G+I. Sequências de aminoácidos de amastinas de *L. braziliensis* M2904, *L. donovani* BPK282A1, *L. infantum* JPCM5, *L. major* Friedlin, *L. mexicana* U1103, *T. brucei* TREU927 e *T. cruzi* CL Brener foram baixados do TriTrypDB release 52 e foram alinhados juntamente com os genes da mesma família de *L. amazonensis* PH8 usando três programas diferentes: muscle v3.8.31 (EDGAR, 2004), mafft v7 (KATOH *et al.*, 2002) e clustalw v2.1 (THOMPSON; GIBSON; HIGGINS, 2002). Os três

alinhamentos resultantes foram combinados em um alinhamento de consenso usando M-Coffee v13.45 (WALLACE *et al.*, 2006).

A árvore ML foi reconstruída usando PhyML v3.0 para o alinhamento de 311 sequências de amastinas (GUINDON *et al.*, 2009). O melhor modelo de substituição foi definido para JTT+G (ARENAS, 2015) determinado usando ProtTest 3.4.2 (ABASCAL; ZARDOYA; POSADA, 2005) de acordo com a concordância entre o critério de informação de Akaike (AIC) e critério de informação Bayesiano (BIC).

O consenso da estrutura secundária das proteínas GP63 foi previsto pelo *software* PSIPRED V4.0 (MCGUFFIN; BRYSON; JONES, 2000) e o alinhamento das 9 sequências anotadas foi usado como entrada. PSIPRED é um método de previsão baseado em redes neurais feed-foward executadas no alinhamento. A predição da estrutura terciária foi realizada na ferramenta SWISS-MODEL (SCHWEDE *et al*., 2003), um servidor automatizado de modelagem de homologia estruturada de proteínas. Por último, a qualidade estereoquímica de uma estrutura de proteína foi verificada por Procheck (LASKOWSKI *et al.*, 1993).

5. RESULTADOS

5.1. Geração de *reads* curtas *paired-end* Illumina e *reads* longas PacBio de alta qualidade

Com o objetivo de obter a sequência completa do genoma da cepa PH8 de *L. amazonensis* foram geradas duas bibliotecas a partir do DNA extraído de formas promastigotas. A primeira foi obtida por meio de sequenciamento de fragmentos *paired-end* Illumina de tamanho médio de 350 pb, o qual produziu um total de 15.869.052 *reads* com 100 pb. Após análise de qualidade e remoção de regiões de baixa qualidade (Phred < 30) e dos adaptadores, o total de *reads* diminui para 12.286.419 (~77,5% do total obtido). O perfil de distribuição de Phred *score* ao longo das *reads* filtradas foi obtido após a execução da ferramenta FastQC e é mostrado na figura abaixo (Figura 3).

A segunda biblioteca de gDNA construída foi sequenciada pela plataforma PacBio em três *SMRT cells* e resultou na geração de 560.797 *subreads* com tamanho médio de 8.056 pb e N50 igual a 10.196 (Figura 4). A combinação dessas *subreads*, por sua vez, gerou um total de 349.217 *reads* com tamanho médio de 12.980, N50 de 17.985 e qualidade média igual a 0.83.

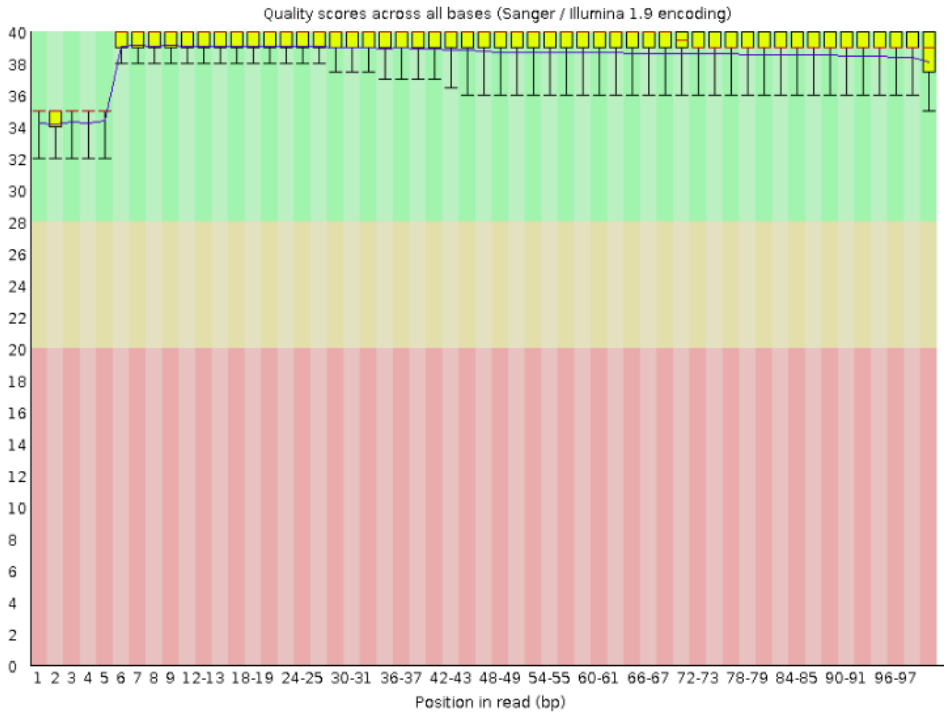

Figura 3. Gráfico de qualidade por base de *reads* filtradas provenientes de sequenciamento Illumina de gDNA de promastigotas de *L. amazonensis* cepa PH8. Qualidade de *reads* (*forward* e *reverse*) por base baseada na pontuação de phred *score* (eixo y). Ao final de todas as *reads* brutas (posições estão representadas no eixo x) ocorre a queda da qualidade para níveis abaixo de 30 e, portanto, foram removidas.

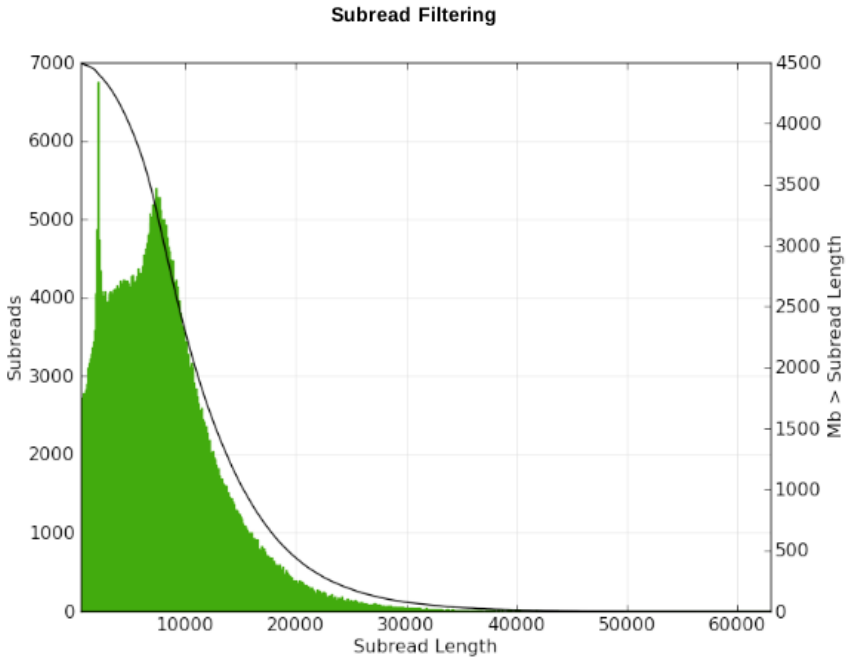

Figura 4. Gráfico de distribuição de tamanho de *subreads* longas em função da quantidade produzida pela plataforma PacBio RS II. A distribuição normal, representada pela cor verde mostra a distribuição do tamanho de *subreads* pela quantidade bruta de *subreads* produzidas. A curva em preto representa a distribuição do tamanho de *subreads* pela quantidade de *subreads* produzidas em megabases (Mb).

5.2. Montagem *de novo* do genoma nuclear e do maxicírculo mitocondrial de *L. amazonensis* cepa PH8

Reads longas de PacBio e *reads* curtas de Illumina quando combinadas são capazes de fornecer sequências mais contíguas (com menor número de *gaps*) e com número reduzido de erros, muito comuns em tecnologias que geram *reads* longas (MILLER *et al*., 2017). Por esse motivo, neste trabalho, foi empregada uma abordagem mista, onde os dois tipos de *reads* foram utilizados. Contudo, inicialmente apenas as *reads* PacBio foram usadas para a montagem. Este procedimento foi realizado utilizando o montador Canu, que foi capaz de montar 349.217 em 380 *contigs*. Posteriormente, o *pipeline* IPA foi usado para "limpar" a montagem inicial, removendo *contigs* redundantes, menores que 5 kb e aqueles com 90% de sobreposição. Essa etapa foi capaz de reduzir o número total de *contigs* para 329. A etapa de *scaffolding*, realizada com o objetivo de unir sequências não contíguas para formar *scaffolds* foi executada usando a ferramenta SSPACE e *reads paired-end* Illumina, resultando em 314 *scaffolds* com N50 igual a 188.152. Na tentativa de obter sequências ainda mais contíguas, GapFiller foi usado para o preenchimento de *gaps* existentes nas extremidades incorporadas de *reads* Illumina. No total, foram preenchidas apenas 4 de 15 locais onde haviam *gaps*, mas este resultado não influenciou no número de *scaffolds* já obtidos. Por fim, a ferramenta Pilon foi executada para "polir" o genoma montado, considerando inconsistências existentes entre *reads* Illumina mapeadas na montagem. Essa etapa, por sua vez, realizou correções

de *indels* (inserção e deleções; 29,76%), correções de base única (20,23%), eventos de substituição de blocos (20,23%), preenchimento de *gaps* e identificação de *misassemblies* locais que juntos representam os outros 29,77% das discrepâncias entre as *reads* Illumina e Pacbio encontrados em todos as sequências montadas (Figura 5).

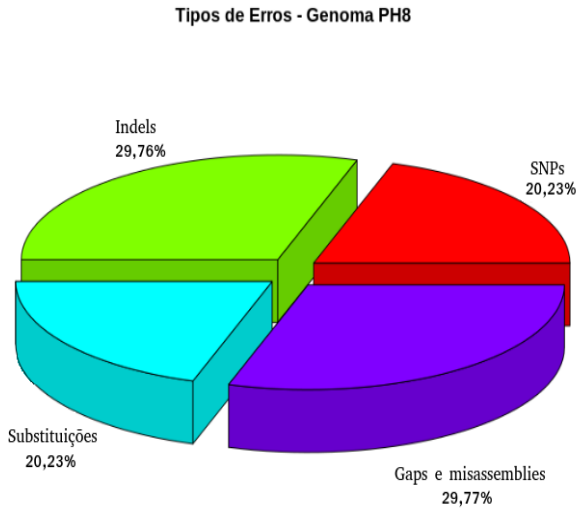

Figura 5. Principais tipos de erros de sequenciamento encontrados na montagem do genoma da PH8 e corrigidos posteriormente pela ferramenta Pilon.

Os 314 *scaffolds* obtidos foram alinhados, ordenados e orientados com base nos cromossomos de *L. mexicana* usando a ferramenta ABACAS, que além de encontrar regiões sintênicas entre as sequências alinhadas, projeta *primers* que são capazes de fechar *gaps* existentes. Dessa forma, o número de *scaffolds* obtidos foi reduzido para 77, onde 34 destes alinharam em regiões sintênicas à referência (Figura 6a-f), sendo, portanto, considerados como pesudochromossomos da PH8. Os outros 43 *scaffolds* não foram incorporados a estas moléculas, mas análises de identidade no banco de dados de nucleotídeos não redundantes (nr) evidenciou que um deles, com 41.489 pb, corresponde ao maxicírculo do genoma mitocondrial da PH8 e os outros 42 *scaffolds* representam sequências variando de 1 a 35 kb e estão localizados em regiões não sintênicas com o genoma de *L. mexicana*. Alguns desses outros *scaffolds* correspondem a sequências muito pequenas e outros, como mostrado adiante (Figura 13), possuem sequências truncadas que codificam famílias multigênicas. Além disso, o alinhamento da maioria dos *scaffolds* restantes mostrou similaridade com regiões cromossômicas da espécie de referência (Tabela 2) e outras espécies de *Leishmania*, principalmente *L. donovani*, como foi o caso da sequência correspondente ao maxicírculo.

Tabela 2. Porcentagem de identidade entre os *scaffolds* montados do genoma da PH8 e regiões cromossômicas de *L. mexicana* e outras leishmanias.

***Scaffold* PH8**	**Gene ID**	**Início-Final no cromossomo referência**	**Identidade**
scaffold180	Ldon kDNA	17331-17331	87.78%
scaffold197	Lmx.15 1036_mexFOS1_13k16.p1kpIBF_73 1036_mexFOS1_13k16.p1kpIBF_97 1036_mexFOS1_13k16.p1kpIBF_26	161694-164703 1-1868 1197-1792 1-1977	100% 99,93% 100% 100%
scaffold221	Lmx.30	655198-655839	99,92%
scaffold234	Lmx.33 1036_mexFOS1_13k16.p1kpIBF_9 1036_mexFOS1_13k16.p1kpIBF_17	685169-692483 944-1415 10963-11243	97,78% 95,23% 89,62%
scaffold235	Lmx.30 1036_mexFOS1_13k16.p1kpIBF_17	348298-348818 4824-9368	99,75% 99,25%
scaffold239	Lmx.11	497564-500236	99,78%
scaffold242	Lmx.27	480320-500350	98,81%
scaffold245	Lmx.12	384192-384870	99,70%
scaffold248	Lmx.33 1036_mexFOS1_13k16.p1kpIBF_9 1036_mexFOS1_13k16.p1kpIBF_17 1036_mexFOS1_13k16.p1kpIBF_105	685226-688314 897-1212 4118-5807 1-2042	96,95% 95,18% 92,95% 94,84%
scaffold253	Lmx.30 1036_mexFOS1_13k16.p1kpIBF_23 1036_mexFOS1_13k16.p1kpIBF_82	179418-184618 3058-6171 1263-2371	99,82% 99,90% 99,19%
scaffold256	Lmx.12	179418-184618	99,82%
scaffold258	Lmx.32	97239-98504	99,39%
scaffold259	Lmx.02	63172-71822	95,0%
scaffol260	Lmx.02	68033-71822	95,39%
scaffold262	1036_mexFOS1_13k16.p1kpIBF_1 1036_mexFOS1_13k16.p1kpIBF_21 1036_mexFOS1_13k16.p1kpIBF_104	8767-9874 1478-2696 125-2205	100% 81,15% 99,56%
scaffold264	Lmx.29	488816-490275	97,01%

scaffold266	1036_mexFOS1_13k16.p1kpIBF_17 1036_mexFOS1_13k16.p1kpIBF_9	9713-11704 2724-3419	90,66% 94,82%
scaffold267	exclusivo da cepa PH8 (baixa cobertura de alinhamento com outras espécies de *Leishmania*, apenas)	-	-
scaffold268	Lmx.19 1036_mexFOS1_13k16.p1kpIBF_21 1036_mexFOS1_13k16.p1kpIBF_104	651040 655046 872-1552 392-654	98,59% 100 100
scaffold270	Lmx.30 1036_mexFOS1_13k16.p1kpIBF_23	161110-163507 3058-6350	99,87% 99,78%
scaffold273	Lmx.14	438017-449922	85,69%
scaffold274	Lmx.22	529604-538607	73,71%
scaffold275	Lmx.30 1036_mexFOS1_13k16.p1kpIBF_82 1036_mexFOS1_13k16.p1kpIBF_23	161200-164027 1263-3155 3058-5671	99,78% 95,26% 99,61%
scaffold276	Lmx.28	1125878-1132084	90%
scaffold278	Lmx.22	258010-261219	99,53%
scaffold280	Lmx.20	1530016-1531853	98,65%
scaffold282	Lmx.22	267607-268815	99,75%
scaffold283	1036_mexFOS1_13k16.p1kpIBF_63 1036_mexFOS1_13k16.p1kpIBF_20	7661-8931 6615-7548	94,14% 99,68%
scaffold285	Lmx.30	627263-633658	99,31%
scaffold286	Lmx.30 1036_mexFOS1_13k16.p1kpIBF_342	652648-654381 1-2978	99,19% 91,71%
scaffold287	1036_mexFOS1_13k16.p1kpIBF_95 1036_mexFOS1_13k16.p1kpIBF_20	564-2098 1702-1961	98,14% 97,47%
scaffold291	Lmx.26	608324-615487	98,23%
scaffold292	Lmx.20	8174-10619	85,22%
scaffold293	Lmx.11	227125-231203	97,21%
scaffold294	Lmx.22	259027-261219	98,16%
scaffold295	Lmx.08	575755-582652	98,86%
scaffold296	Lbr.11	731-1104	90,52%
scaffold299	1036_mexFOS1_13k16.p1kpIBF_1	9221-10375	100%

scaffold301	Lmx.19	340897-43987	95,93%
scaffold302	Lmx.22	552570-554583	98,67%
scaffold309	1036_mexFOS1_13k16.p1kpIBF_1	1-345	93,98%
scaffold313	Lbr.11	1107-1480	88,77%
scaffold314	Lbr.11	58-436	89%

Os 34 pseudocromossomos da PH8 têm mais de 88% de suas sequências alinhadas ao genoma de referência de *L. mexicana* U1103 e a razão de base (N) ambígua muda de 0,001 para 0,057. Todos os outros parâmetros avaliados na montagem estão sumarizados na Tabela 3, onde é possível ver a distribuição dos *contigs* obtidos em relação ao tamanho, parâmetros como N50, N75 e porcentagem de conteúdo GC.

Tabela 3. Parâmetros avaliados no genoma montado de *L. amazonensis* cepa PH8.

***Contigs* Totais**	77
Contigs (>= 1000 bp)	77
Contigs (>= 5000 bp)	74
Contigs (>= 10000 bp)	61
Contigs (>= 25000 bp)	37
Contigs (>= 50000 bp)	34
Maior *contig*	3.400.190
N° de *scaffolds*	77
N° de cromossomos	34
N50	1.069.653
N75	661.168
GC (%)	59,62

A comparação com duas outras montagens de diferentes linhagens de *L. amazonensis* disponíveis em bancos de dados públicos (RZOD01 e UA301) mostrou valores semelhantes para todos os parâmetros avaliados, divergindo em maior grau apenas no número final de *scaffolds*, para

os quais o genoma da cepa UA301 está representado em exatamente 34 *scaffolds* (Tabela 4). Em contraste, a montagem do genoma da PH8 é significativamente melhorada em comparação com o genoma da cepa M2269, que é o único genoma atualmente disponível no TritrypDB.

Tabela 4. Comparação entre montagens de diferentes cepas de *L. amazonensis* disponíveis em bancos de dados públicos.

Cepa *L. amazonensis*	Plataforma e cobertura	Tamanho total (Mpb)	Nº. de *contigs* ou *scaffolds*	Conteúdo GC (%)	N50 (bp)	Tamanho Maior *contig* ou *scaffold*	Referências
PH8	Pacbio (86X)/ Illumina (43X)	31.970.850	77	59,62	1.069.653	3.400.190	Este trabalho
M2269*	454, Illumina (96X)	29.029.348	2.627	59,26	19.306	113.027	Real *et al.*, 2013
RZOD01	Pacbio (75X)	33.504.997	92	59,71	850.106	3.425.950	Batra *et al.*, 2019
UA301	Illumina (99X)	32.156.470	34	59,50	1.135.553	3.336.136	Patino *et al.*, 2019

*Esta é a montagem que está atualmente disponível no TritrypDB e considerado como genoma de referência.

Análises BUSCO foram empregadas para avaliar a completude da montagem da PH8 em comparação com um conjunto de dados de 130 marcadores de genes ortólogos de cópia única que se espera estarem presentes em 31 espécies de protozoários, incluindo a cepa *L. mexicana* U1103 e outras *Leishmania* spp (ROGERS *et al.*, 2011). Essa é uma métrica que auxilia a contiguidade ao avaliar a qualidade do genoma, e reflete uma estimativa da fração de genes que esperados na montagem em relação a um conjunto de genes ortólogos de determinado grupo taxonômico referência. Nesse sentido, os resultados mostraram que a sequência da PH8 contém 95,4% de conjuntos gênicos completos e apenas 4,6% de conjuntos ausentes em relação ao banco de dados euglenozoa, sendo assim, nenhum tendo sido classificado como duplicado ou fragmentado. Pesquisando contra 255 ortólogos de eucariotos, a montagem de PH8 indicou ter 50,6% de BUSCOs marcadores completos, dos quais 1,2% correspondem a conjuntos de genes duplicados (Tabela 5). Em relação à montagem de outras cepas de *L. amazonensis* a cepa PH8 alcançou

resultados similares, tendo como ponto negativo apenas os 4,6% de BUSCOs ausentes em sua montagem. Tal fato indica a não identificação de alguns ortólogos do filo Euglenozoa nesta montagem, devido a não existência de correspondências significativas ou a uma pontuação abaixo do intervalo de confiança para o perfil BUSCO. Na prática esse resultado demonstra que ainda há regiões não resolvidas na sequência do genoma, permanecendo *features* parciais ou ausentes.

Tabela 5. Comparação entre diferentes montagens de genomas de *L. amazonensis*.

	Número total de proteínas no banco de dados	BUSCOs Completos (%)	BUSCOs cópia única (%)	BUSCOs duplicados (%)	BUSCOs fragmentados (%)	BUSCOs ausentes (%)
Banco de dados: euglenozoa_odb10						
L. amazonensis PH8	130	95,4	95,4	0,0	0,0	4,6
L. amazonensis M2269		100,0	0,0	0,0	0,0	0,0
L. amazonensis APNT01.1		100,0	0,0	0,0	0,0	0,0
L. amazonensis RZOD01		100,0	98,5	1,5	0,0	0,0
L. amazonensis UA301		99,2	99,2	0,0	0,8	0,0
Banco de dados: eukaryota_odb10						
L. amazonensis PH8	255	50,6	49,4	1,2	0,0	49,4
L. amazonensis M2269		53,8	51,4	2,4	9,0	37,2
L. amazonensis APNT01.1		53,8	51,8	2,0	8,6	37,6
L. amazonensis RZOD01		52,5	49,8	2,7	9,0	38,5
L. amazonensis UA301		53,8	51,8	2,0	9,0	37,2

É importante ressaltar que cada conjunto BUSCO (completos, cópia única, duplicados, fragmentados e ausentes) representa uma lista de marcadores de genes ortólogos de uma linhagem diferente. Como mencionado anteriormente, no presente trabalho foram usadas duas linhagens distintas, sendo a primeira correspondente ao Filo Euglenozoa com 31 espécies incluindo *L. infantum*, *L. major*, *L. mexicana* e *L. tarentolae*, além de diversas espécies do gênero Trypanosoma, como *T. cruzi* e *T. brucei*. A segunda linhagem utilizada inclui outros 255 marcadores de genes ortólogos de cópia única que estão presentes em 70 espécies pertencentes ao domínio taxonômico Eucaryota. A combinação de ambos fornece uma visão mais ampla da completude da montagem da PH8, uma vez que as espécies classificadas dentro do filo euglenozoa pertencem ao domínio de organismos eucariontes e, portanto, existem milhares de genes conservados entre todas as espécies deste domínio (WANG *et al.*, 2021). Aqui é visto, por

exemplo, que esse compartilhamento de marcadores é de aproximadamente 50%, corroborando com os resultados obtidos para outras espécies pertencentes ao mesmo domínio taxonômico.

BUSCOs completos são subdivididos em cópia única e duplicados. Quando a busca realizada pelo algoritmo, baseado em perfis hmms (cada gene do banco de dados é representado por um perfil hmm), encontra um *match* com pontuação dentro do intervalo esperado de pontuações e dentro do intervalo esperado de alinhamentos de comprimento compatível com o perfil BUSCO, o gene buscado foi contabilizado como completo. Se o perfil foi encontrado uma única vez no genoma da PH8, o mesmo foi considerado como cópia única, caso contrário, duplicado. De forma análoga, quando o *match* marca dentro do intervalo de pontuações, mas não dentro do intervalo de alinhamentos de comprimento para o perfil BUSCO, o gene da PH8 foi classificado como um BUSCO fragmentado, indicando que o gene está apenas parcialmente presente. Por fim, perfis BUSCOs foram considerados ausentes quando nenhum dos parâmetros citados foram atendidos.

Figura 6a. Circle plots mostrando a conservação da sintenia entre os *scaffolds* de 01 a 06 da cepa PH8 de *L. amazonensis* e cromossomos de *L. mexicana*. No primeiro círculo (externo) estão os cromossomos (encima) de *L. mexicana* e *scaffolds* (embaixo) da PH8. Traços laranja e azuis representam genes localizados nas fitas *foward* e *reverse*, respectivamente (segundo círculo). Retângulos amarelos (terceiro círculo) indicam regiões de *gaps*. Retângulos verdes representam genes centrais ausentes e traços pretos são genes *singleton* (quarto círculo). LAMA_00: *scaffolds* não incorporado na montagem.

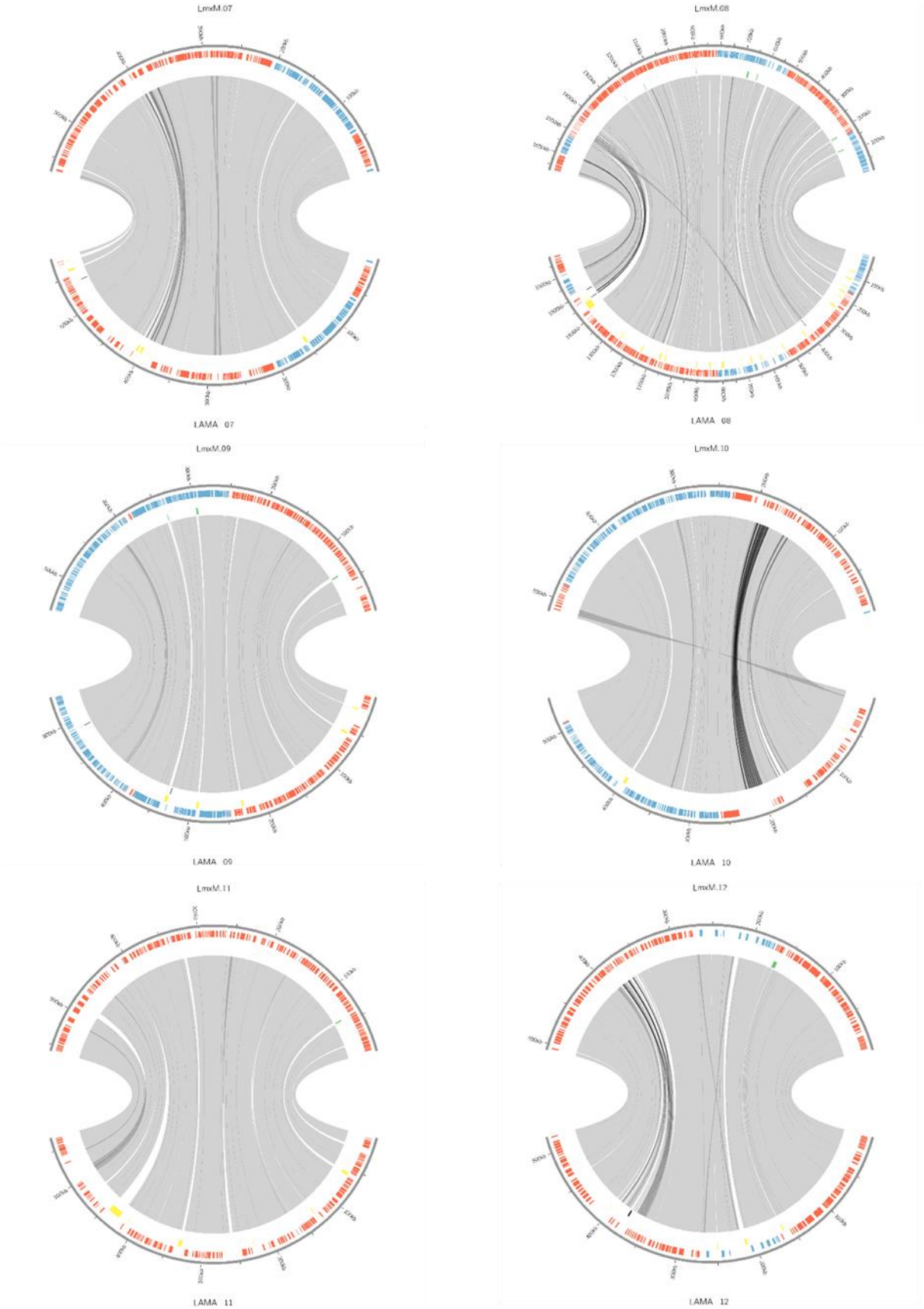

Figura 6b. Circle plots mostrando a conservação da sintenia entre os *scaffolds* de 07 a 12 da cepa PH8 de *L. amazonensis* e cromossomos de *L. mexicana*. No primeiro círculo (externo) estão os cromossomos (encima) de L. mexicana e *scaffolds* (embaixo) da PH8. Traços laranja e azuis representam genes localizados nas fitas *foward* e *reverse*, respectivamente (segundo círculo). Retângulos amarelos (terceiro círculo) indicam regiões de *gaps*. Retângulos verdes representam genes centrais ausentes e traços pretos são genes *singleton* (quarto círculo). LAMA_00: *scaffolds* não incorporado na montagem.

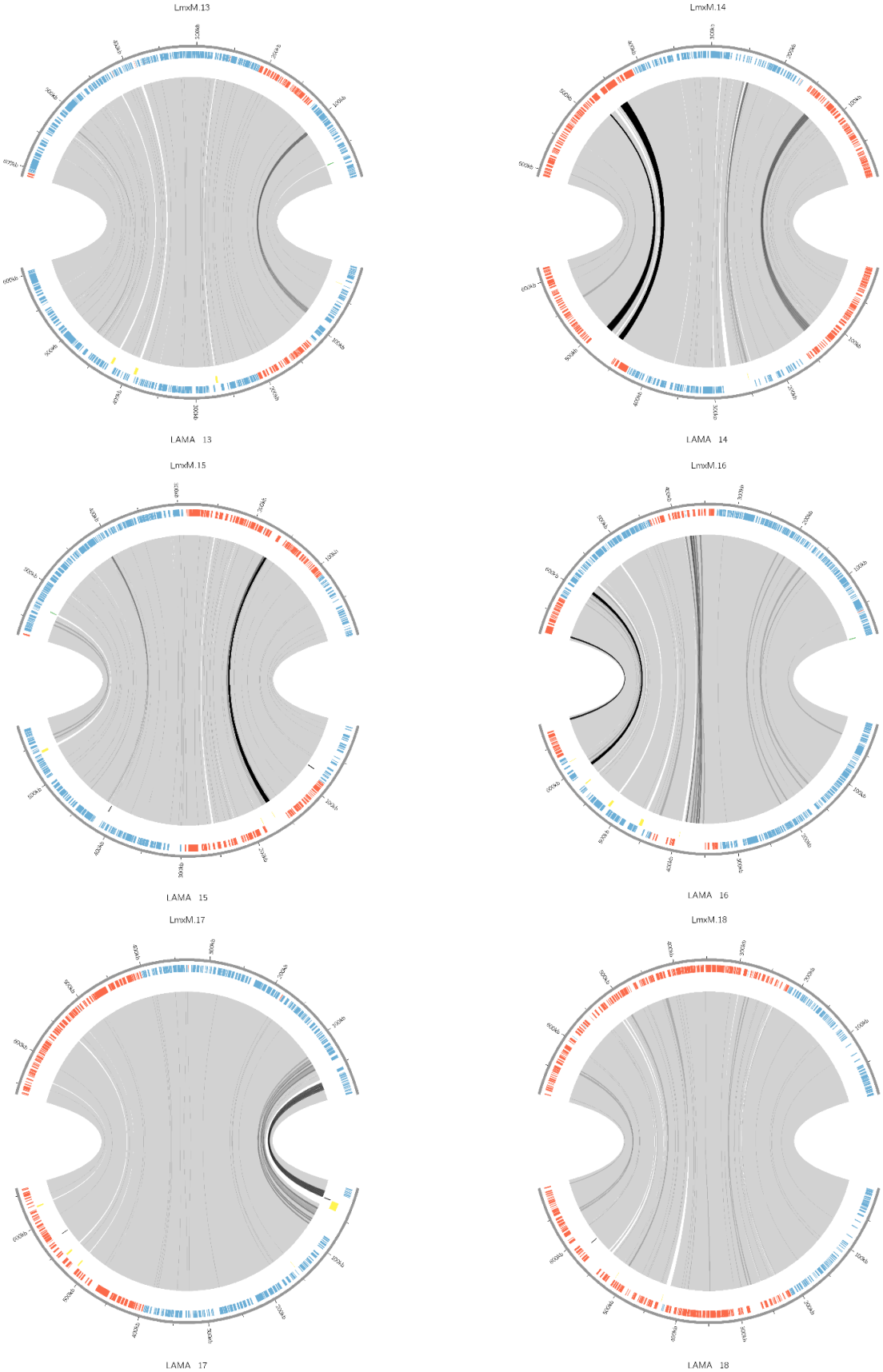

Figura 6c. Circle plots mostrando a conservação da sintenia entre os *scaffolds* de 13 a 18 da cepa PH8 de *L. amazonensis* e cromossomos de *L. mexicana*. No primeiro círculo (externo) estão os cromossomos (encima) de *L. mexicana* e *scaffolds* (embaixo) da PH8. Traços laranja e azuis representam genes localizados nas fitas *foward* e *reverse*, respectivamente (segundo círculo). Retângulos amarelos (terceiro círculo) indicam regiões de *gaps*. Retângulos verdes representam genes centrais ausentes e traços pretos são genes *singleton* (quarto círculo). LAMA_00*: scaffolds* não incorporado na montagem.

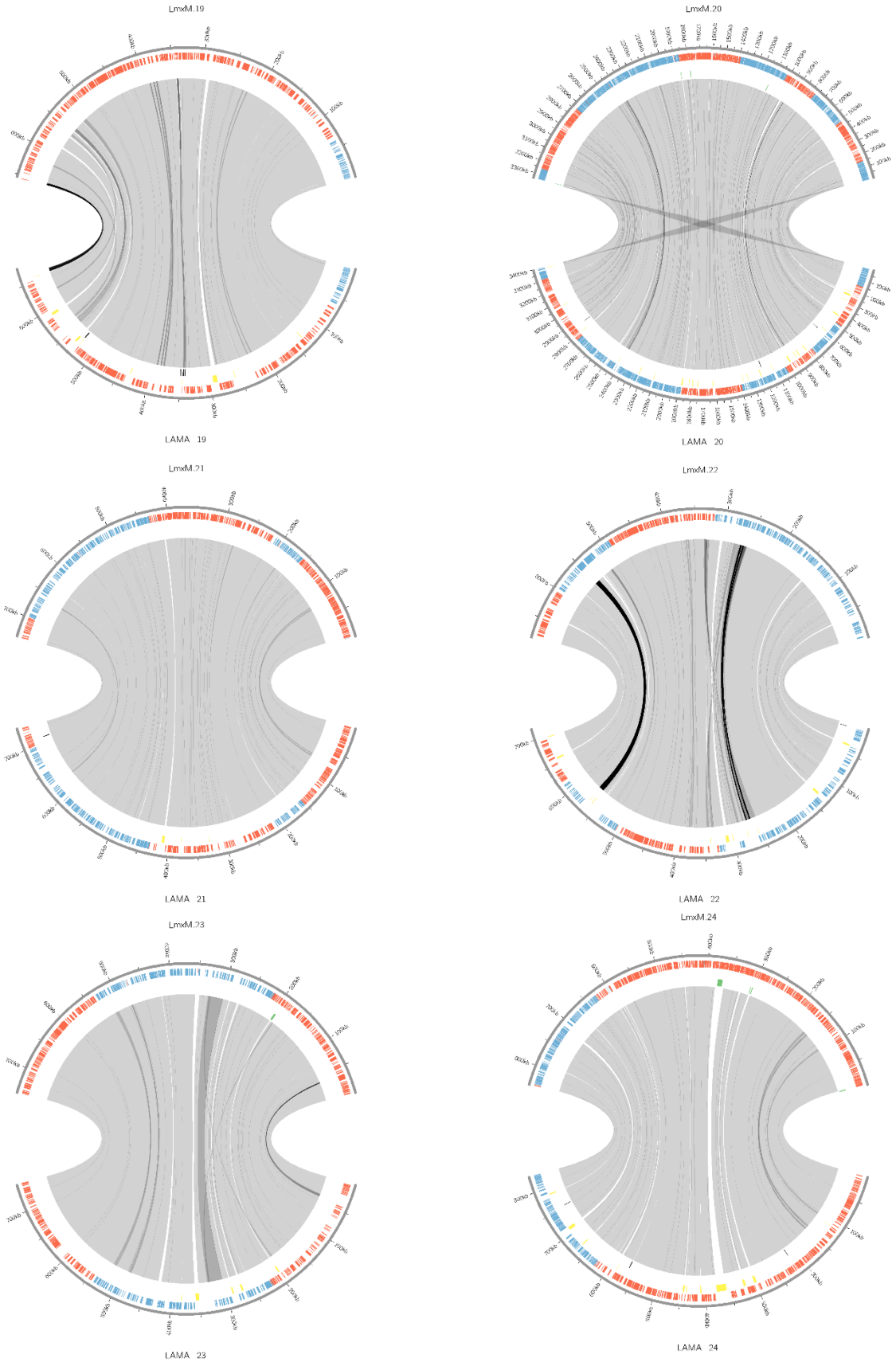

Figura 6d. Circle plots mostrando a conservação da sintenia entre os *scaffolds* de 19 a 24 da cepa PH8 de *L. amazonensis* e cromossomos de *L. mexicana*. No primeiro círculo (externo) estão os cromossomos (encima) de *L. mexicana* e *scaffolds* (embaixo) da PH8. Traços laranja e azuis representam genes localizados nas fitas *foward* e *reverse*, respectivamente (segundo círculo). Retângulos amarelos (terceiro círculo) indicam regiões de gaps. Retângulos verdes representam genes centrais ausentes e traços pretos são genes singleton (quarto círculo). LAMA_00: *scaffolds* não incorporados na montagem.

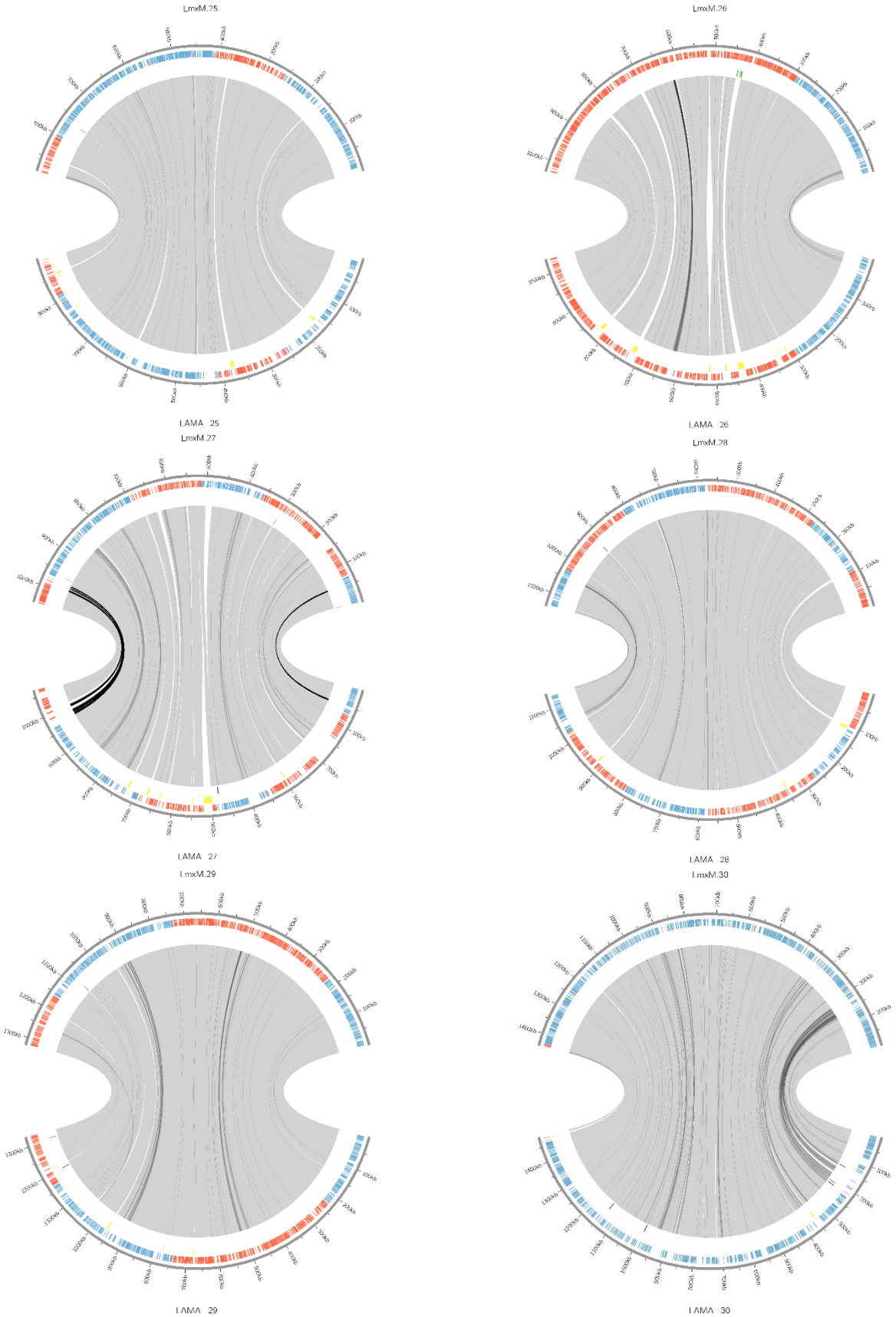

Figura 6e. Circle plots mostrando a conservação da sintenia entre os *scaffolds* de 25 a 30 da cepa PH8 de *L. amazonensis* e cromossomos de *L. mexicana*. No primeiro círculo (externo) estão os cromossomos (encima) de *L. mexicana* e *scaffolds* (embaixo) da PH8. Traços laranja e azuis representam genes localizados nas fitas *foward* e *reverse*, respectivamente (segundo círculo). Retângulos amarelos (terceiro círculo) indicam regiões de *gaps*. Retângulos verdes representam genes centrais ausentes e traços pretos são genes *singleton* (quarto círculo). LAMA_00: *scaffolds* não incorporados na montagem.

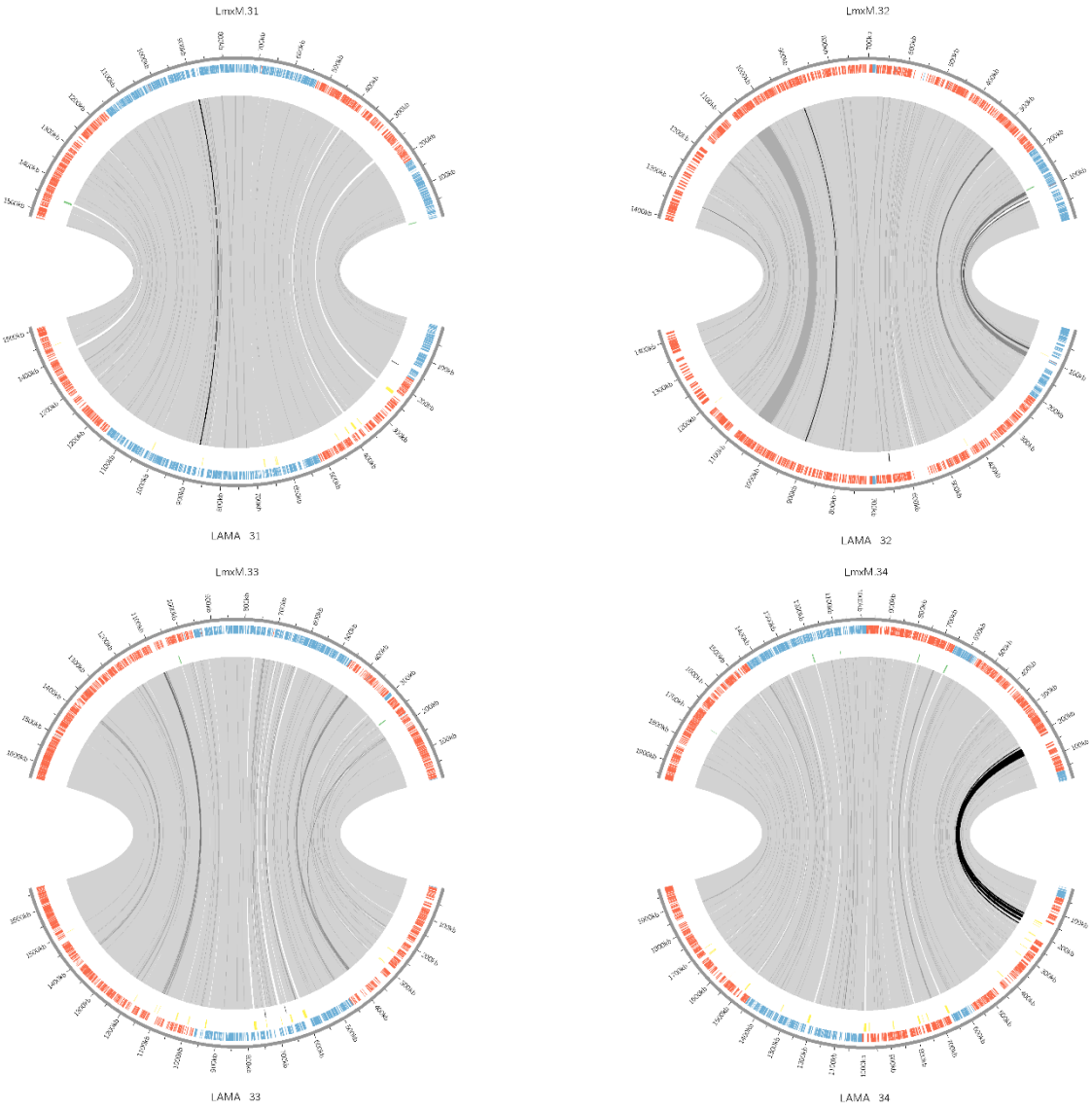

Figura 6f. Circle plots mostrando a conservação da sintenia entre os *scaffolds* de 31 a 34 da cepa PH8 de *L. amazonensis* e cromossomos de *L. mexicana*. No primeiro círculo (externo) estão os cromossomos (encima) de *L. mexicana* e *scaffolds* (embaixo) da PH8. Traços laranja e azuis representam genes localizados nas fitas *foward* e *reverse*, respectivamente (segundo círculo). Retângulos amarelos (terceiro círculo) indicam regiões de gaps. Retângulos verdes representam genes centrais ausentes e traços pretos são genes *singleton* (quarto círculo). LAMA_00: *scaffolds* não incorporado na montagem.

Levando em conta o tamanho estimado (média das montagens obtidas até o momento: ~31,8 Mb) para o genoma de *L. amazonensis* (BATRA *et al.*, 2019; PATINO *et al.*, 2020; REAL *et al.*, 2013), as *reads* aqui geradas representam uma cobertura média de sequenciamento de ~43x para *reads* Illumina e 86x para *reads* PacBio (Figura 7). As *reads* de Illumina e de *reads* Pacbio foram utilizadas na etapa de montagem *de novo* do genoma da PH8. *Reads* Illumina proporcionaram, sobretudo, a correção de erros produzidos pela plataforma Pacbio, que apresenta uma taxa média de erro de ~15% (ARDUI *et al.*, 2018).

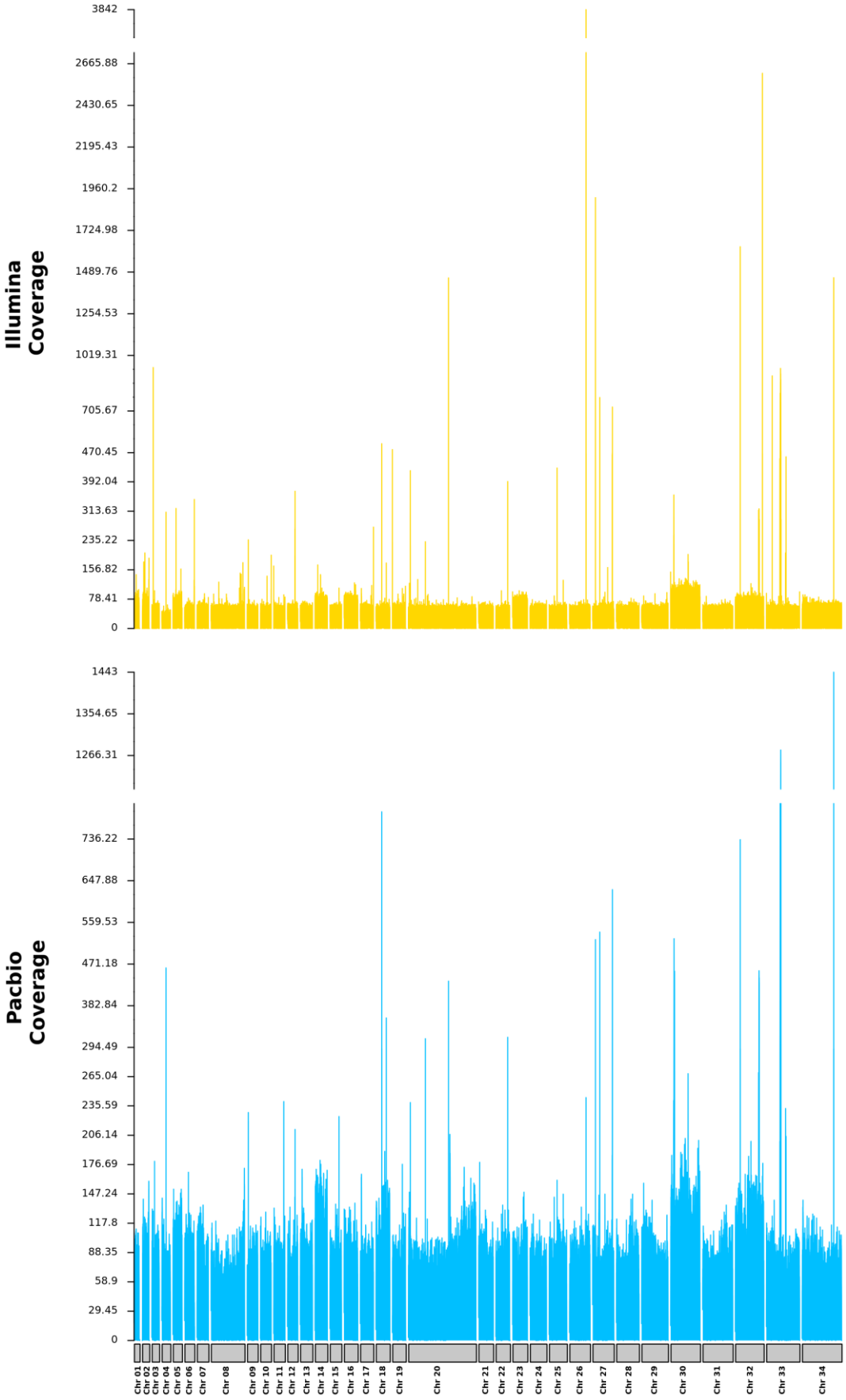

Figura 7. Gráfico de cobertura dos 34 pseudocromossomos preditos na montagem de *L. amazonensis*. As barras horizontais cinzentas representam os pseudocromossomos montados da cepa PH8. As faixas amarelas representam a cobertura usando *reads* Illumina e em azul as r*eads* Pacbio. A cobertura foi calculada como o número de *reads* por *bin*, onde os *bins* são janelas de contagem consecutivas curtas com o tamanho definido para 50 bp em ambas as análises.

Além disso, a montagem da PH8 teve uma qualidade média de 0,58 com a maioria dos *scaffolds* com qualidade igual a 0,65 (24 *scaffolds*) ou superior (21 *scaffolds*), como mostrado no histograma da Figura 8. Os *scores* de qualidade são calculados inicialmente a partir do alinhamento individual de cada *scaffold* aos cromossomos de referência de *L. mexicana*. Em seguida, é calculada uma pontuação de qualidade para a montagem usando a cobertura do genoma de referência fornecida pelos *scaffolds* em diferentes limiares de qualidade, a distribuição da pontuação de qualidade dos *scaffolds* individuais e a redundância dos mesmos na montagem. Os 34 pseudocromossomos tiveram mais de 88% de suas sequências alinhadas ao genoma de referência de *L. mexicana* U1103 e a razão de base (N) ambígua mudou de 0,001 para 0,057. Repetições teloméricas [(TTAGGG)n] foram identificadas nas extremidades 3' dos *scaffolds* 6, 17 e 32, e complemento reverso [(CCCTAA)n] nas extremidades 5' dos *scaffolds* 3, 7 e 19, previamente previsto por (CANO; SILVA, 2017; CHIURILLO *et al*., 2000; CONTE; CANO, 2005). Sequências conservadas presentes em regiões subteloméricas correspondentes a CSB1 (GTACAGT) e CSB2 (-GGAGAGGGTGT) também foram encontradas nos cromossomos 8 ao 34 para CSB1 e 8,10,19,20,24,26,27,30, 31, 32 e scaffold295 para CSB2.

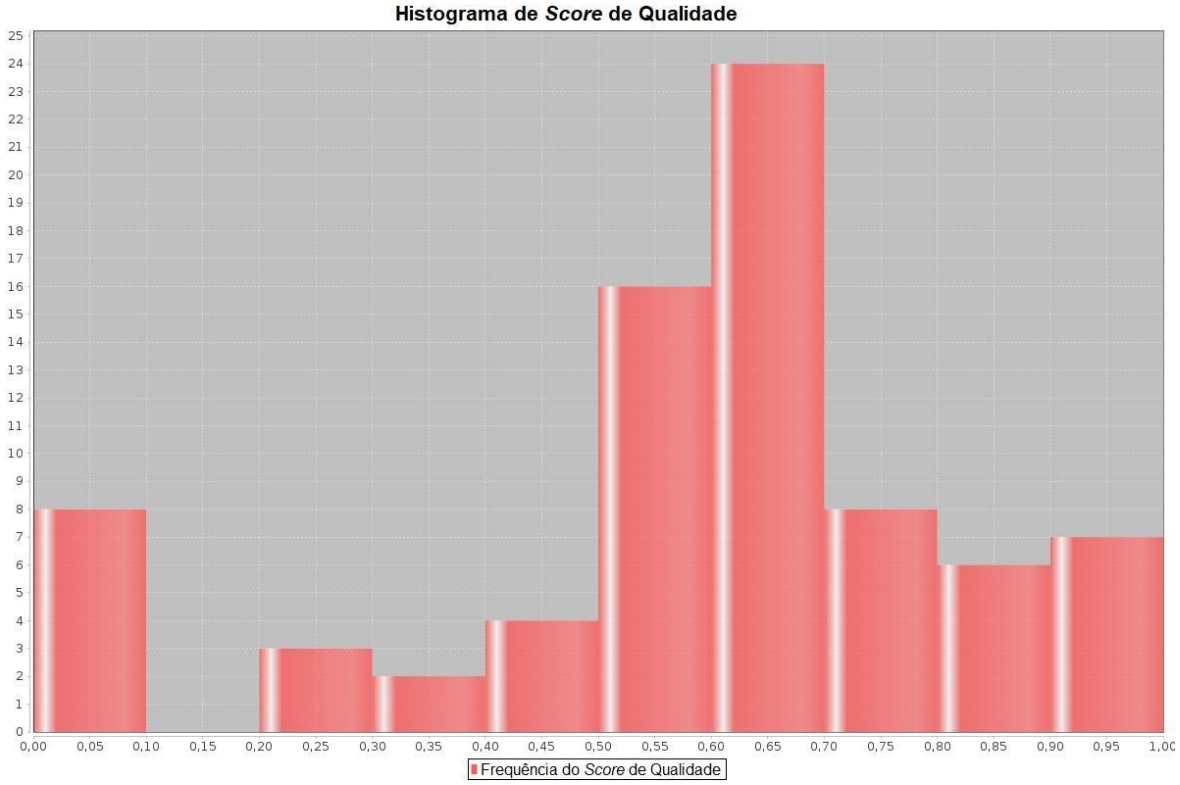

Figura 8. Histograma de *score* de qualidade dos 78 *scaffolds* obtidos na montagem da cepa PH8 de *L. amazonensis*. Os s*cores* de qualidade foram calculados pelo *software* dnAQET e são apresentados nas barras rosas dispostas no eixo x. A quantidade de *scaffolds* para cada s*core* de qualidade é representada no eixo y.

5.3. Anotação automática do genoma nuclear e do maxicírculo mitocondrial de *L. amazonensis* usando como referência os genomas de *L. mexicana* e *L. tarentolae*

Após alcançar o mais alto nível na montagem do genoma da PH8, ou seja, a obtenção de *scaffolds* com grande cobertura dos cromossomos preditos para *L. mexicana*, foi realizada uma transferência de anotação a partir da mesma. No total foram transferidas 8.225 *features* (genes), dentre os quais, codificantes de proteínas (putativas, hipotéticas e com função comprovada), codificantes de RNAs transportadores (tRNAs), RNAs ribossomais (rRNAs) e RNAs não codificantes (ncRNAs). No entanto, quando a predição gênica e anotação baseada em homologia foram realizadas pelo algoritmo AUGUSTUS, que realiza uma predição *ab initio* dos modelos gênicos, foram obtidos 8.317 genes, sendo 7.999 genes codificadores de proteínas (Arquivo GFF e fasta - Apêndice II). Além dos genes codificadores de proteínas, identificamos genes que codificam 15 rRNA e 92 tRNA, bem como 138 sequências que codificam RNAs não codificantes. Toda a informação foi armazenada em um arquivo GFF, que contém estrutura específica e informações de localização gênica e presença na fita positiva ou negativa. Esse último resultado significa que foram encontrados 190 novos genes anotados de *L. amazonensis*, que não foram descritos na anotação anterior de nenhum outro genoma de *L. amazonensis* (REAL *et al.*, 2013). Os resultados referentes a anotação estão sumarizados na Tabela 6. Dentre os novos genes identificados, 57 codificam fatores de virulência que serão detalhados na seção 5.5 deste trabalho, enquanto que os demais codificam em sua maioria proteínas hipotéticas ou com função desconhecida (118 genes). Apesar disso, fomos capazes de atualizar a anotação de ~280 genes anteriormente classificados nessa categoria. Com relação aos que codificam proteínas com função conhecida encontra-se 6 cópias do gene do fator de elongação 1-alpha (EF1-alpha) e o gene que codifica uma proteína anotada como associada ao cinetoplasto (KAP, do inglês *kinetoplast-asscociated protein*), ambas descritas como fatores de virulência no gênero *Leishmania.*

Tabela 6. Resumo da montagem e anotação do genoma da cepa PH8.

Geral		
	Transferência da Anotação	**Predição *ab initio***
Nº total de genes	8225	8317
Porcentagem codificante (%)	47.19	48.92
Genes codificadores de proteínas		
Genes	7.437	7.999
Pseudogenes*	788	318
Tamanho médio CDS (bp)	1.828,20	1.846,01
Conteúdo G+C (%)	60,15	62,05
RNAs		
ncRNA	270	138
tRNA	82	92
rRNA	13	15

*Pseudogenes foram inferidos com base na presença de *stop códons* prematuros e/ou quadros de leituras incompletos.

Uma análise comparativa (Tabela 7) também foi realizada a partir de dados coletados de genomas de *L. amazonensis* publicados anteriormente em 2013 e 2019. Melhora significativa foi obtida na montagem do genoma da PH8 se comparado aos genomas das cepas MHOM/BR/71973/M2269 e 210-660 (RZOD01), tendo sido alcançado o nível cromossômico do genoma na atual montagem. Com relação a cepa UA301, apesar de N50 um pouco menor e maior número de *contigs*, a cepa PH8 foi obtida por meio de estratégia *de novo*, enquanto que a montagem de UA301 foi guiada por referência, o que pode acarretar na transferência de erros, principalmente em regiões altamente variáveis, como por exemplo, regiões que codificam famílias de proteínas de superfície (revisado por LISCHER; SHIMIZU, 2017).

Tabela 7. Comparação entre genomas montados de *Leishmania*.

Cepa	N° *scaffolds*	N° *contigs*	N50 (pb)	Tamanho Maior *contig* (pb)	Conteúdo G+C (%)	Cobertura (x)	N° totais de genes
L. amazonensis PH8	34	77	1.069.653	3.400.190	59,6	60/48	8.317
L. amazonensis MHOM/BR/71973/M2269	2,627	3.199	19.306	113.027	59,26	96	8.127
L. amazonensis 210-660 (RZOD01)	92	92	850.106	3.425.950	59,71	75	NA
L. amazonensis UA301	34	34	1.135.553	3.336.136	59,49	99,1	NA
L. braziliensis MHOM/BR/75/ M2904	35	35	1.063.631	2.662.849	56,99	96/363	8.395
L. donovani LdCL	36	36	1.067.468	2.916.019	59,07	107	8.633
L. infantum JPCM5	36	36	1.055.293	2.743.073	59,56	100	8.796
L. major Friedlin	36	36	1.091.540	2.682.151	59,7	NA	8.412
L. mexicana U1103	588	337	825.953	3.343.498	59,78	NA	8.677

*NA's indicam dados não disponíveis nos bancos de dados ou ainda, inexistentes.

Comparações de anotações permitiram observar um aumento no número de genes preditos com relação à primeira versão do genoma de *L. amazonensis*, onde apenas 8.100 genes haviam sido anotados e também mostrou que a cepa PH8 possui uma média de genes codificadores de proteínas próxima aos demais genomas de *Leishmania*. Quando o conteúdo gênico de diferentes cepas de *L. amazonensis* foi comparado, sobretudo 12 genes ortólogos universais que são compartilhados entre todas as espécies de *Leishmania*, é notória a semelhança evolutiva da PH8 com o genoma da cepa UA301 e RZOD01 (Figura 9). Este resultado pode contribuir para futuros estudos comparativos, além de ajudar a resolver regiões de N's que ainda persistem no interior dos pseudocromossomos da PH8.

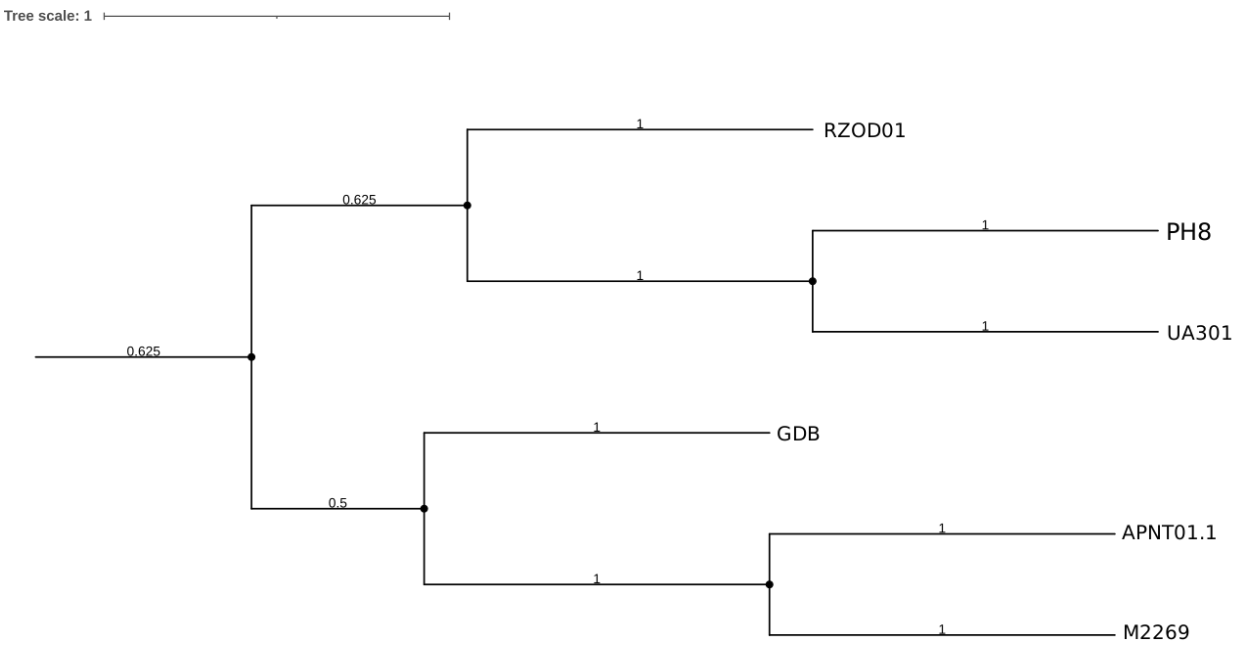

Figura 9. Descrição da relação filogenética entre cepas de *Leishmania amazonensis*. Reconstrução filogenética baseada em 12 genes ortólogos universais compartilhados entre todas as espécies de *Leishmania.* Os comprimentos dos ramos são desenhados proporcionalmente à distância genética, com valores mostrados no meio de cada ramo. UA301: Genoma colombiano de *L. amazonensis*, M2269: *L. amazonensis* MHOM/BR/71973/M2269 (TritrypDB); APNT01.1: *L. amazonensis* (acesso Genbank: APNT01.1); RZOD01.1: *L. amazonensis* cepa 210-660; GDB: *Leishmania (L.) amazonensis* Genoma DB.

Genes ortólogos universais (OU) são compartilhados entre todos os seres vivos, mas há ainda aqueles que estão presentes em todas as espécies de *Leishmania* e que foram anteriormente identificados por PATINO *et al.*, 2019. A lista de 12 genes OU inclui as seguintes proteínas: leucil-tRNA sintetase, valil-tRNA sintetase putativa, proteína de pré-translocação putativa, subunidade alfa, seril-tRNA sintetase putativa, proteína ribossomal semelhante a proteína L3, proteína de ligação ao GTP putativa, proteína ribossômica L11 putativa, proteína ribossômica 60S L10 putativa, proteína ribossômica L14 putativa, proteína ribossômica 40S S9 putativa, proteína ribossômica 40S S18 putativa, proteína ribossômica 40S S16 putativa.

A partir da anotação transferida foi possível ainda, a identificação de membros anotados para as principais famílias gênicas envolvidas em processos de invasão, diferenciação e visceralização de amastigotas de *L. amazonensis*, tais como: amastinas, GP63, proteína A2, cisteína proteases, fosfatases e cinases, os quais estão presentes nos 34 pseudocromossomos de *L. amazonensis* (Figura 10a-b), e em outras 20 sequências não incorporadas à montagem (Figura 11). Além dessas famílias multigênicas, foram também anotados genes que codificam proteínas envolvidas no metabolismo de heme e ferro devido ao fato de estarem diretamente envolvidas no estabelecimento da infecção.

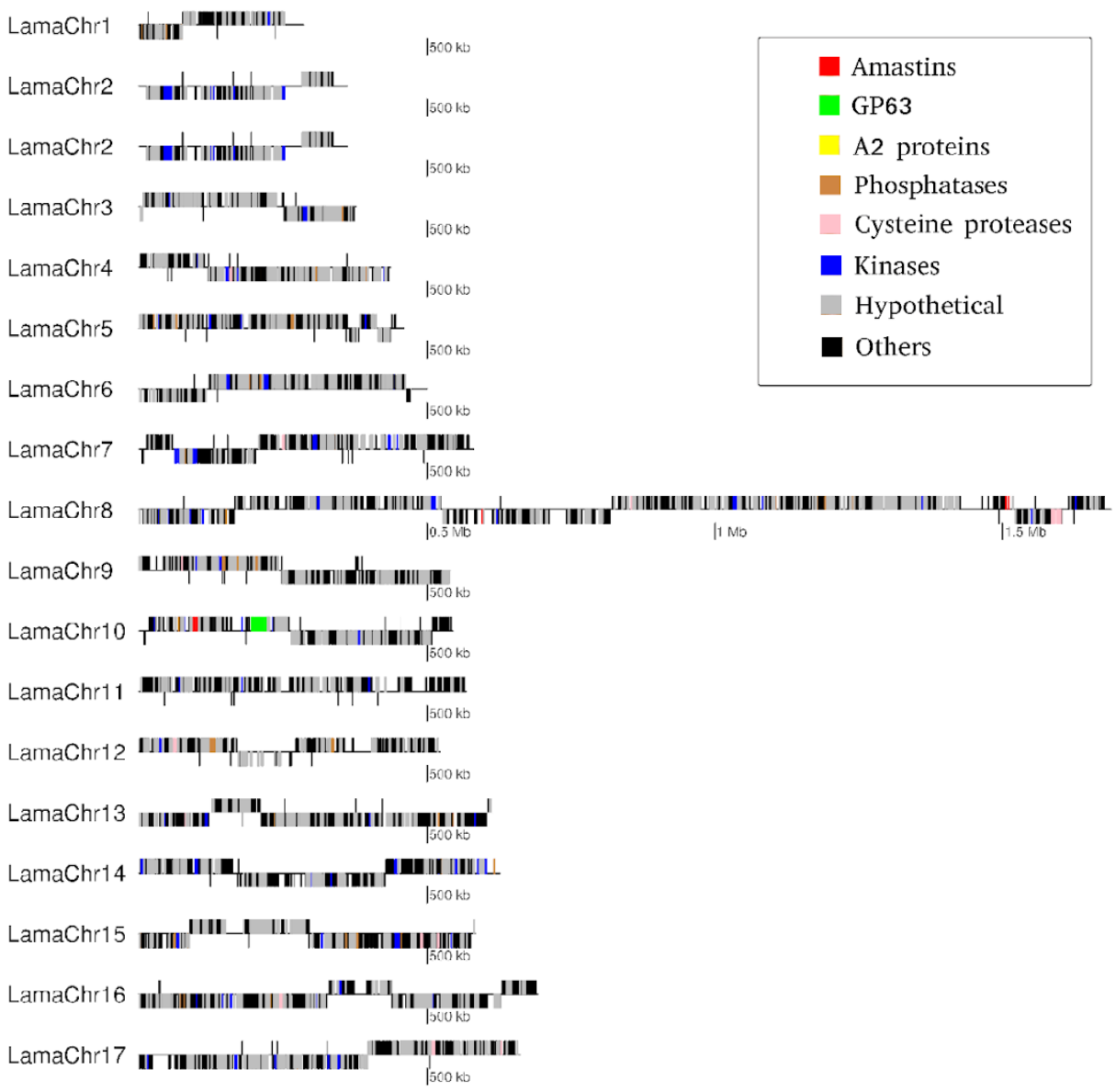

Figura 10a. Anotação do genoma da cepa PH8 de *Leishmania amazonensis* com destaque para as principais famílias multigênicas. No eixo y estão representados cada cromossomo, nomeados LamaChr1- LamaChr177. No eixo x encontra-se o tamanho de cada um em escala que varia de kb a Mb. Os retângulos na vertical ao longo da linha preta (fita) representam os genes. Genes voltados para cima indicam localização na fita *sense* e genes voltados para baixo, na fita *antisense*. Cada cor representa uma família de genes, como descrito na legenda e apenas as cores preta e cinza fazem referência a genes que codificam outros tipos de moléculas (proteínas hipotéticas, RNAs, proteínas codificadas por genes cópia única e outras famílias multigênicas, por exemplo).

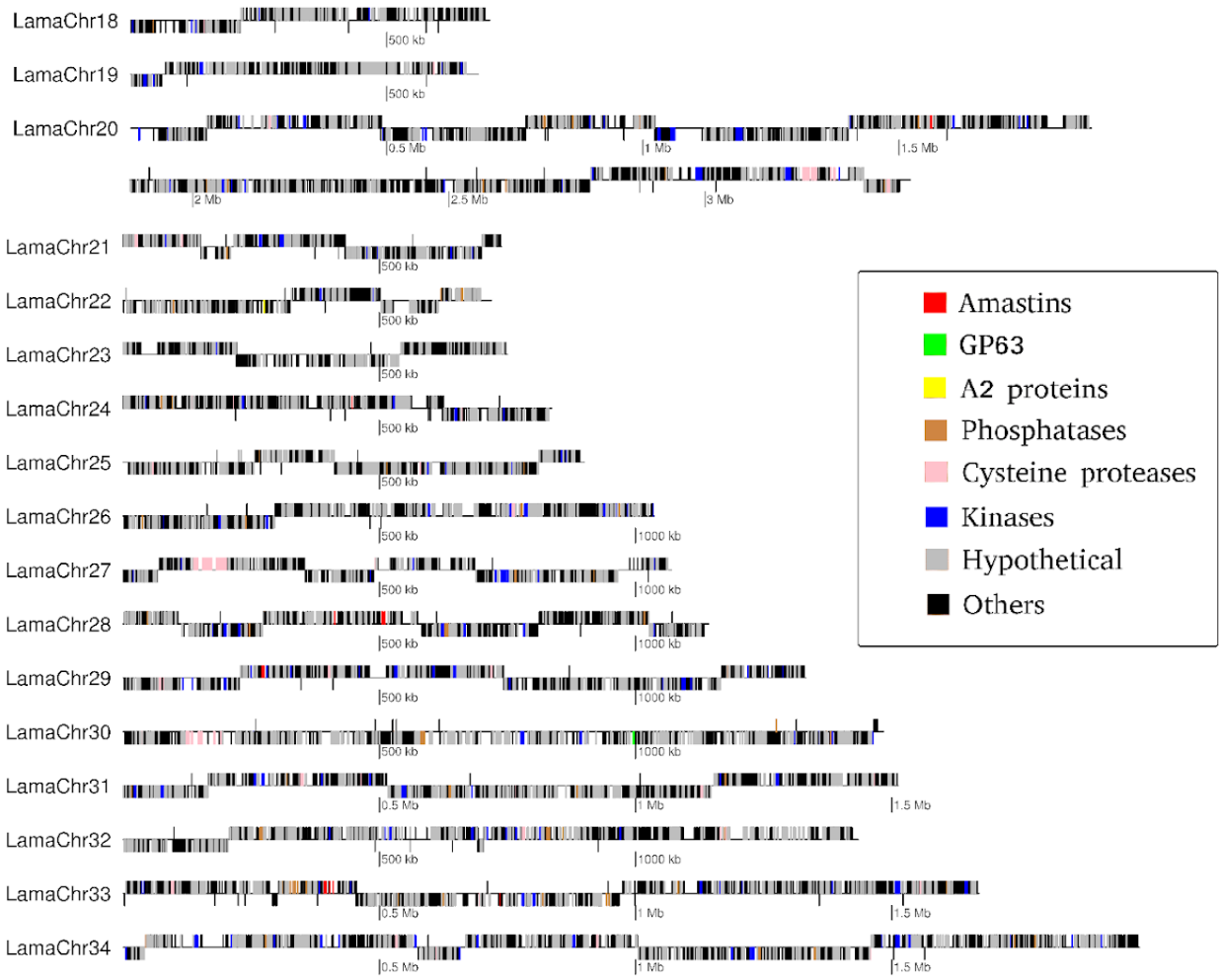

Figura 10b. Anotação do genoma da cepa PH8 de *Leishmania amazonensis* com destaque para as principais famílias multigênicas. No eixo y estão representados cada cromossomo, nomeados LamaChr18- LamaChr34. No eixo x encontra-se o tamanho de cada um em escala que varia de kb a Mb. Os retângulos na vertical ao longo da linha preta (fita) representam os genes. Genes voltados para cima indicam localização na fita *sense* e genes voltados para baixo, na fita *antisense*. Cada cor representa uma família de genes, como descrito na legenda e apenas as cores preta e cinza fazem referência a genes que codificam outros tipos de moléculas (proteínas hipotéticas, RNAs, proteínas codificadas por genes cópia única e outras famílias multigênicas, por exemplo).

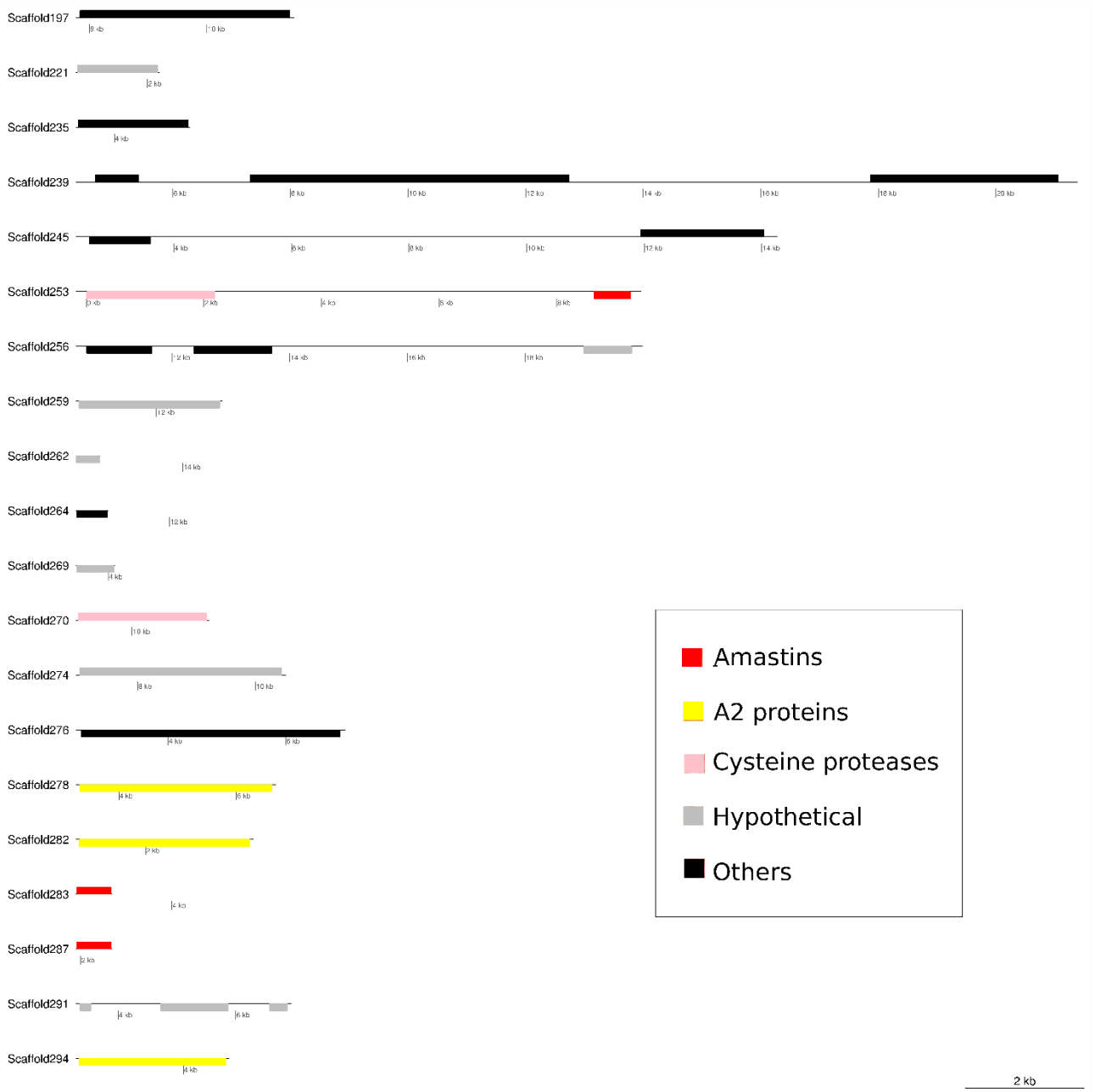

Figura 11. *Scaffolds* não incorporados à montagem. Cada *scaffold* é representado no eixo y. No eixo x estão as marcações de tamanhos que variam de kb a Mb. Os retângulos verticais ao longo da linha preta (fita) representam os genes (não estão em escala). Os genes voltados para cima indicam uma localização na fita *sense* e os genes voltados para baixo na fita *antisense*. Cada cor representa uma família de genes. Genes pretos indicam genes com anotação funcional transferidos do genoma de *L. mexicana*, mas que não pertencem a nenhuma das famílias aqui estudadas.

Alinhamento entre o genoma da PH8 e o genoma de referência de *L. mexicana* também permitiu a identificação de *clusters* de genes compartilhados e específicos da espécie. Nessa etapa foram considerados apenas os genes preditos no *workflow* da ferramenta Companion, que pode definir grupos ortólogos por meio do *software* ortomcl. Sendo assim, um conjunto central de 7.612 *clusters* de genes ortólogos foi encontrado sendo compartilhado entre os genomas da PH8 e de *L. mexicana*, enquanto 96 *clusters* de genes codificadores de proteínas são específicos de *L. amazonensis* e 141 de *L. mexicana*. Por outro lado, os resultados apresentados pela ferramenta

OrthoVenn2, aplicada aos conjuntos de proteínas *L. amazonensis*, *L. mexicana*, *L. donovani*, *L. infantum*, *L. major* e *L. braziliensis* indicam 8.025 *clusters* (7.746 *clusters* de genes ortólogos contendo ≥1 proteína de *L. amazonensis*), 1.299 *clusters* ortólogos com pelo menos duas espécies e 6.726 *clusters* de genes de cópia única. A maioria (7.120) dos *clusters* de genes ortólogos foram compartilhados por outras *Leishmania* spp. Vinte e nove famílias foram encontradas apenas em *L. amazonensis* e *L. mexicana* e 11 famílias foram exclusivas de *L. amazonensis* (Figura 12). Os cinco maiores aglomerados identificados em *Leishmania* spp. também foram identificados em *L. amazonensis* (Tabela S1; Apêndice III): proteínas semelhantes a amastinas (Grupo 1), dineínas (Grupo 2) glicoproteína GP63 (Grupo 3), proteína de cassete de ligação de ATP (Grupo 4) e histona H4 (Grupo 5). Sete cópias do gene tuzina e três de amastinas estão entre os genes únicos de *L. amazonensis*. Além disso, processos biológicos relacionados à modulação simbionte do processo hospedeiro e transporte de membrana são enriquecidos entre genes que codificam proteínas com funções desconhecidas. Esses achados indicam um número de cópias maior para o complexo tuzina-amastina em comparação com o sugerido por Real *et al.* (2013) e reforçam o importante papel que esta família, juntamente com a GP63, desempenha durante a infecção do hospedeiro.

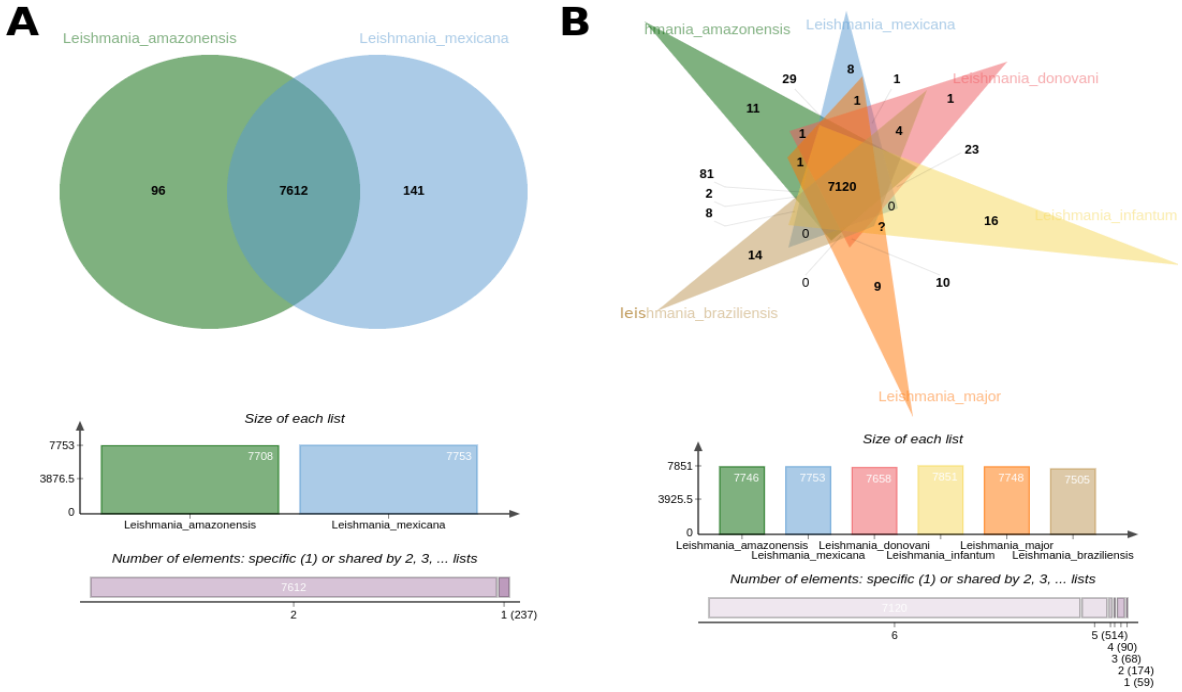

Figura 12. Diagramas de Venn de *clusters* ortólogos no gênero *Leishmania*. (A) Diagramas de Venn de genes codificadores de proteínas compartilhados e específicos entre as espécies *L. amazonensis* e *L. mexicana*. Os *clusters* de genes codificadores de proteínas compartilhadas e espécie-específicos no genoma alvo da cepa PH8 são representados à esquerda e em verde; e da cepa *L. mexicana* MHOM/GT/2001/U1103 – cromossomos de referência, à direita e em azul. O número do *cluster* em cada componente está listado abaixo do diagrama. **(B)** Diagramas de Venn de genes codificadores de proteínas compartilhados e espécie-específicos entre *Leishmania* ssp. O número do *cluster* em cada componente está listado abaixo do diagrama.

Sequências repetitivas ou de baixa complexidade (Tabela S2; Apêndice IV) ocupam 9,57% do genoma da cepa PH8 e são classificadas como retroelementos (1,58%), transposons de DNA (1,06%), círculos rolantes (0,21%), RNAs pequenos (0,41%), Satélites (0,16%), Simples (1,87%), baixa complexidade (0,27%) ou repetições não classificadas (4,01%). Outra característica observada é que aproximadamente 12% da sequência nos cromossomos são sequências repetitivas, em comparação com 86,5% em *scaffolds* menores (Figura 13). Como outros genomas de tripanosomatídeos, várias famílias multigênicas com membros gênicos organizados em repetições *em tandem* foram encontradas espalhadas em diferentes cromossomos do genoma PH8 (Figura 10a-b). Membros isolados de famílias multigênicas também são observados, como mostrado para as famílias gênicas de amastinas e GP63, que, juntamente com outras famílias gênicas que codificam fatores de virulência, serão melhor caracterizadas nas seções seguintes.

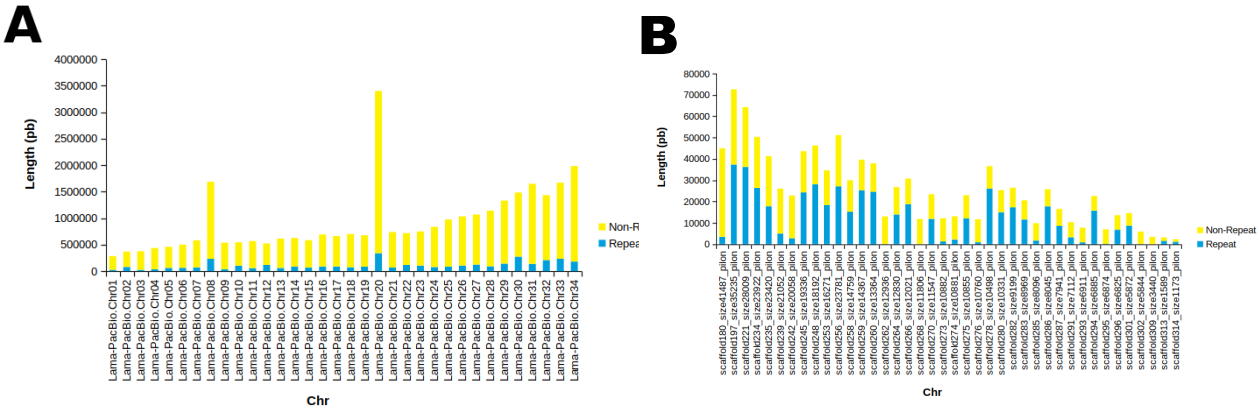

Figura 13. Composição repetitiva do genoma de *Leishmania amazonensis*. (A) Gráfico de barras empilhadas da relação entre conteúdo repetitivo e não repetitivo presente em pseudocromossomos. **(B)** Gráfico de barras empilhadas da relação entre conteúdo repetitivo e não repetitivo presente em pequenos *scaffolds*.

Por fim, um dos *scaffolds* não incorporados na montagem corresponde ao maxicírculo mitocondrial da PH8, o qual apresentou tamanho total de ~41 kb (41,489 pb) e alta similaridade com o maxicírculo de outras espécies de *Leishmanias*, como *L. tarentolae* (86,13%) e *L. braziliensis* (84.19%). Posteriormente foi realizada a remoção de regiões duplicadas desta sequência (identificadas por alinhamento contra si mesmo), resultando em um tamanho final de 18.137 pb.

A anotação do maxicírculo da PH8 cuja sequência foi identificada em um dos *scaffolds* não incorporados à montagem resultou na identificação genes que codificam rRNAs mitocondriais 12S e 9S e vários genes que codificam componentes da cadeia de transporte de elétrons, incluindo citocromo b (CYb), as subunidades I, II e III (COI, COII e COIII) do citocromo oxidase, e as subunidades 1, 4, 5 e 7 (ND1, ND4, ND5 e ND7) de NADH desidrogenase. Além destes, quatro ORFs de função desconhecida (MURF1, MURF2, MURF4, MURF5), também presentes no genoma de outros tripanosomatídeos, foram identificadas, como pode ser observado na Figura 14.

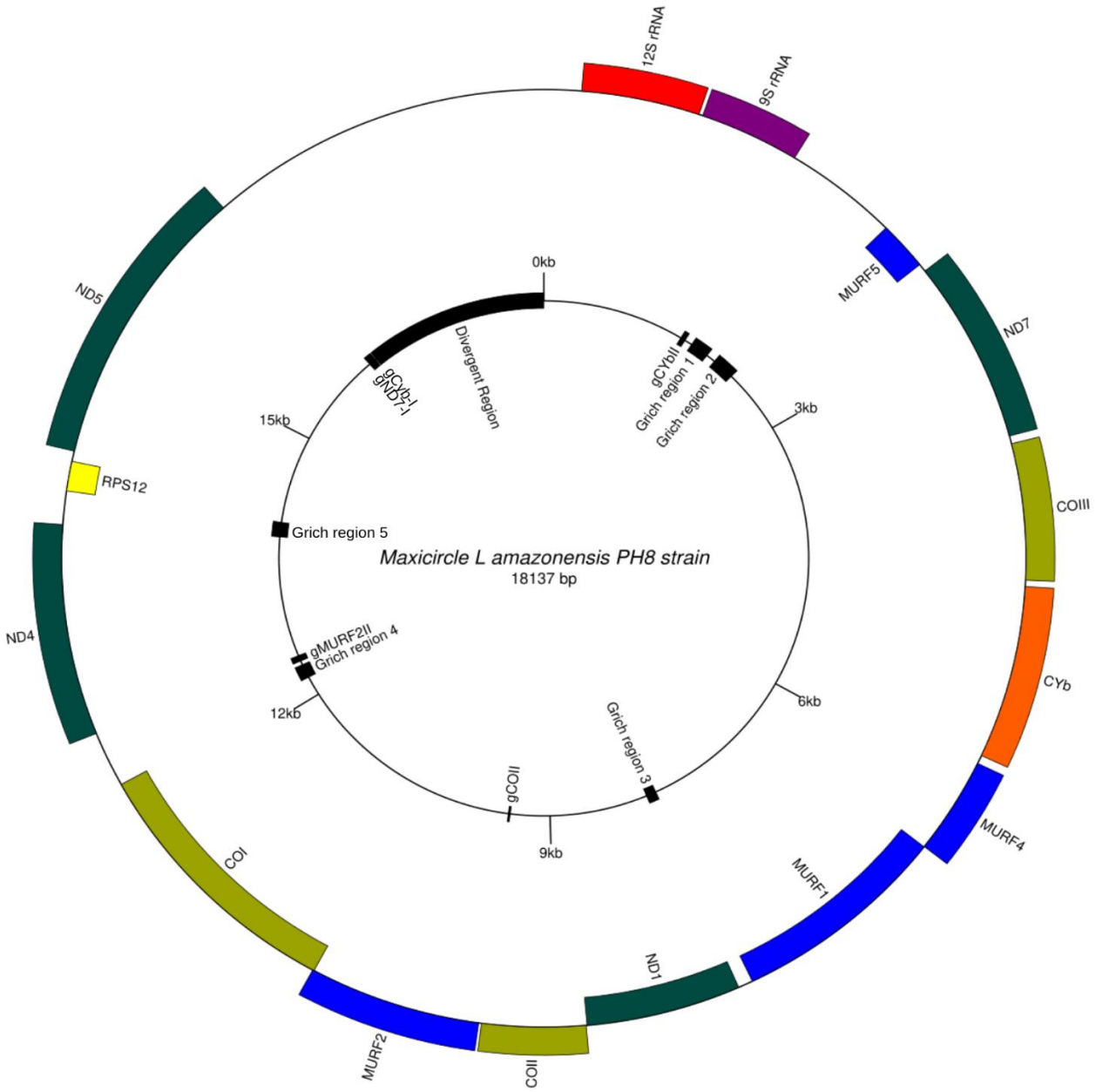

Figura 14. Anotação do maxicírculo da cepa PH8 *de L. amazonensis*. Os retângulos coloridos ao longo do círculo externo (fita) representam a localização de possíveis genes e criptogenes. Genes voltados para cima indicam localização na fita *sense* e genes voltados para baixo, na fita *antisense*. O retângulo no círculo interno (posições no genoma) representa uma região de repetição ~9 kb.

Como já descrito anteriormente, as janelas de leituras de vários genes do maxicírculo encontram-se truncadas e alguns transcritos gerados não possuem códons que iniciam a tradução, sendo, portanto, classificados como criptogenes. Para tornar estes genes funcionais é necessário que ocorra a edição dos transcritos, que resulta na adição ou remoção de resíduos de uridinas com a consequente restauração das janelas de leitura nos mRNAs. Esse evento de remodelação do mRNA pode afetar mais de 50% do comprimento final da molécula (SHAW *et al*., 1988; SIMPSON, 2003). Na figura 14 é possível notar longas regiões gênicas com potencial para codificar mais de 1 gene, como por exemplo, ND7, COIII, Cyb, MURF4, sendo inicialmente transcritos como pré mRNAs e,

portanto, todos considerados como criptogenes. Além disso, os resultados mostram que os gRNAs necessários para a edição destes criptogenes são codificados em diferentes localizações do maxicírculo e distantes dos genes que sofrerão edição. Apenas o gRNA que atua sobre o gene COII foi anotado próximo a este. Além disso, observou-se a presença de regiões ricas em guaninas (G), próximas a sequências codificadoras de gRNAs, as quais foram descritas por estarem envolvidas na edição de criptogenes (MASLOV *et al*., 1992).

A região codificadora do DNA do maxicírculo de *L. amazonensis* é muito semelhante às regiões codificadoras dos maxicírculos de *L. mexicana*, *L. infantum*, *L. donovani*, *L. braziliensis* e *L. major*, apresentando e 91,88%, 87,75%, 89,05%, 84,41%, 87,30% de identidade, respectivamente. A sintenia também é conservada entre diferentes espécies, porém, no genoma do maxicírculo de PH8, ORFs que codificam as proteínas ND3, ND8 e ND9 não foram detectados. Em vez disso, vários padrões de regiões ricas em G estavam presentes nessas posições, interrompendo a transcrição de mRNAs. Além disso, o maxicírculo de *L. amazonensis* possui um gene ortólogo da subunidade 6 da ATPase, denominado MURF4. Da mesma forma, o gene MURF1, presente em *L. amazonensis* é ortólogo do gene ND2 descrito nas demais espécies de *Leishmania*.

A região divergente (RD) do maxicírculo corresponde a porção não codificante desta molécula e recebe este nome por ser uma região altamente variável em nível de espécie. Na PH8, a DR tem aproximadamente 2.200 pb e pode ser subdividida em 2 porções distintas, sendo a primeira nomeada como P5 e P12, em referência a proximidade com os genes ND5 e rRNA 12S, respectivamente. Análise da cobertura em profundidade das *reads* nesta região evidenciou a baixa complexidade e a natureza repetitiva da região DR de modo geral. Mas, assim como visto em outras espécies, *L. amazonensis* apresenta um padrão de repetições distintas nestas sub-regiões, sendo que na sub-região P5 são encontrados um pequeno número de repetições (5) *em tandem* em contraste com um grande período (32), enquanto a sub-região P12 é composta por unidades com número de repetições moderados (5-10) com um pequeno período (10-23), que são organizadas em matrizes repetidas de comprimento variável. Além disso, tais repetições se caracterizam por serem ricas em adeninas (A) e timinas (T).

5.4. Amplificação do número de cópias de cromossomos e genes

Mudanças de somia são frequentemente observados em espécies de *Leishmania*. Tal fenômeno é descrito na literatura por afetar não somente diferentes espécies, mas também, diferentes cepas ou isolados. Em *L. amazonensis*, desvios do perfil dissômico esperado para cada cromossomo foram relatados, tendo alguns, apresentado número normalizado de cópias agrupando-se em torno de perfis dissômicos, trissômicos e tetrassômicos (ROGERS et al., 2011; PATINO *et al.*, 2020; VALDIVIA *et al.*, 2017).

Usando a mediana da densidade de *reads* de cada cromossomo normalizada pela mediana da profundidade de *reads* de todo o genoma, notou-se que a maioria dos cromossomos da cepa PH8 possui um padrão dissómico, enquanto os cromossomos 1, 2, 5, 14, 16, 23 e 32 têm um padrão de trissomia e o cromossomo 30, apresenta um padrão de tetrassomia (Figura 15A).

A análise de distribuição de frequência alélica para todos os cromossomos PH8 confirma as estimativas de somia cromossômica (Figura 15B). Cromossomos dissômico, que apresentam a razão entre a mediana da densidade de *read* do cromossomo e a mediana da profundidade de *read* do genoma igual a 1, exibiram picos para sítios de SNP heterozigotos apenas em 0,5, enquanto cromossomos trissômicos nessa razão próxima a 1,5 exibiram picos em 0,3 ou 0,6 (Figura 16), e o cromossomo 30 a razão foi próxima a dois, apresentando picos de SNP heterozigotos em 0,2, 0,5 e 0,8, consistentes com cromossomos tetrassômicos (Figura 15D). Apesar de possuir cópias extras, a profundidade de *reads* foi distribuída homogeneamente por todo o cromossomo 30, confirmando a hipótese de uma amplificação completa desse cromossomo (Figura 15C).

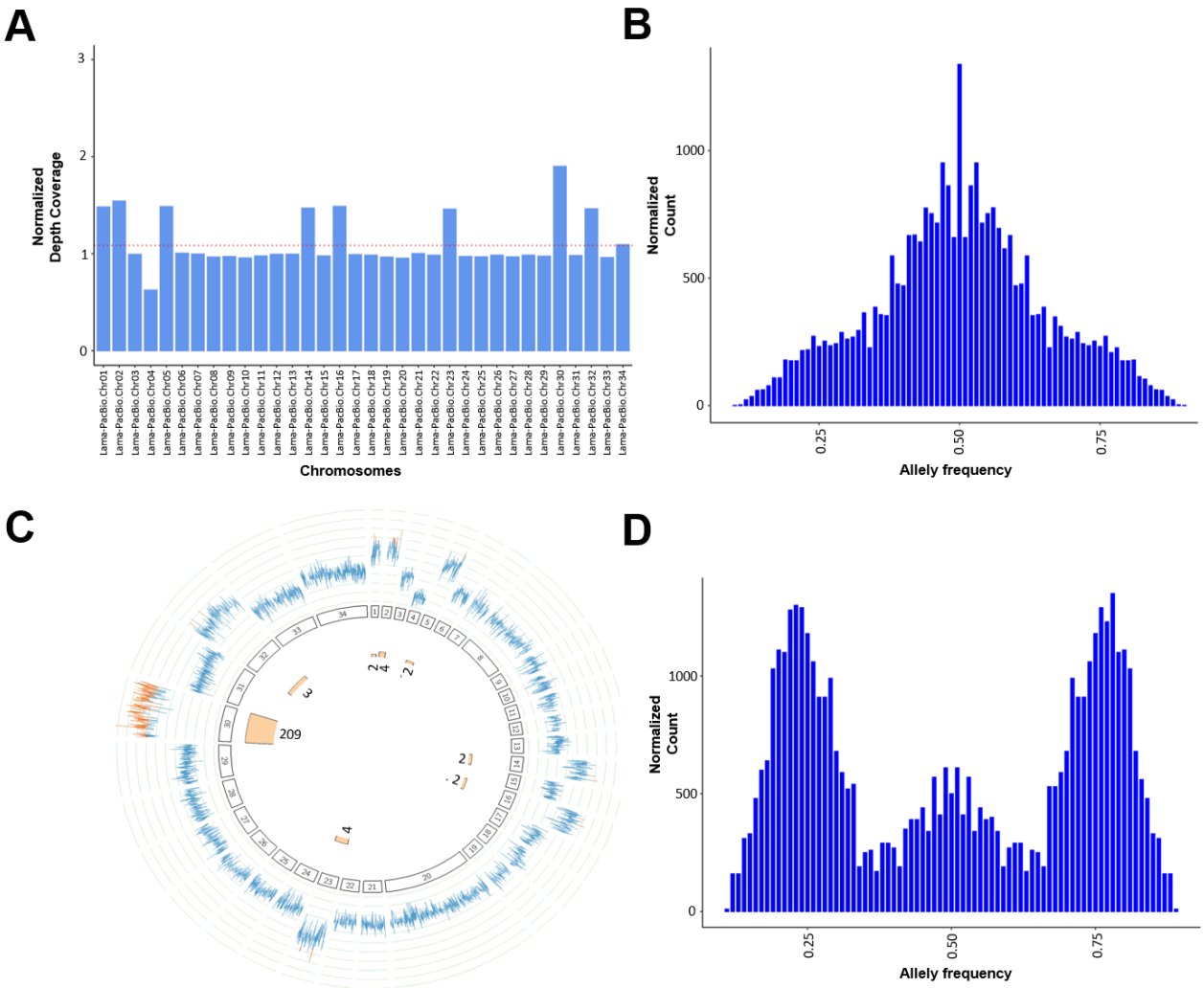

Figura 15. Variação cromossômica e do número de cópias gênicas e distribuição da frequência alélica normalizada na cepa PH8 de *L. amazonensis*. (A) As colunas azuis representam o número estimado de cópias haploides para cada cromossomo. A ploidia média do genoma é indicada pela linha vermelha pontilhada. **(B)** As linhas azuis representam contagens normalizadas das proporções em posições heterozigóticas para todos os cromossomos. **(C)** Profundidade de cobertura de *reads* em cada cromossomo (caixas internas) junto com o número de genes expandidos. A mediana da profundidade de *reads* é mostrada como um gráfico de linha para cada cromossomo; em azul são representadas regiões dissômicas e em vermelho regiões genômicas expandidas. O histograma interno exibe o número total de expansões gênicas identificadas em cada cromossomo. **(D)** Frequência alélica do cromossomo 30. As linhas azuis escuras representam contagens normalizadas das proporções em posições heterozigóticas para todos os cromossomos.

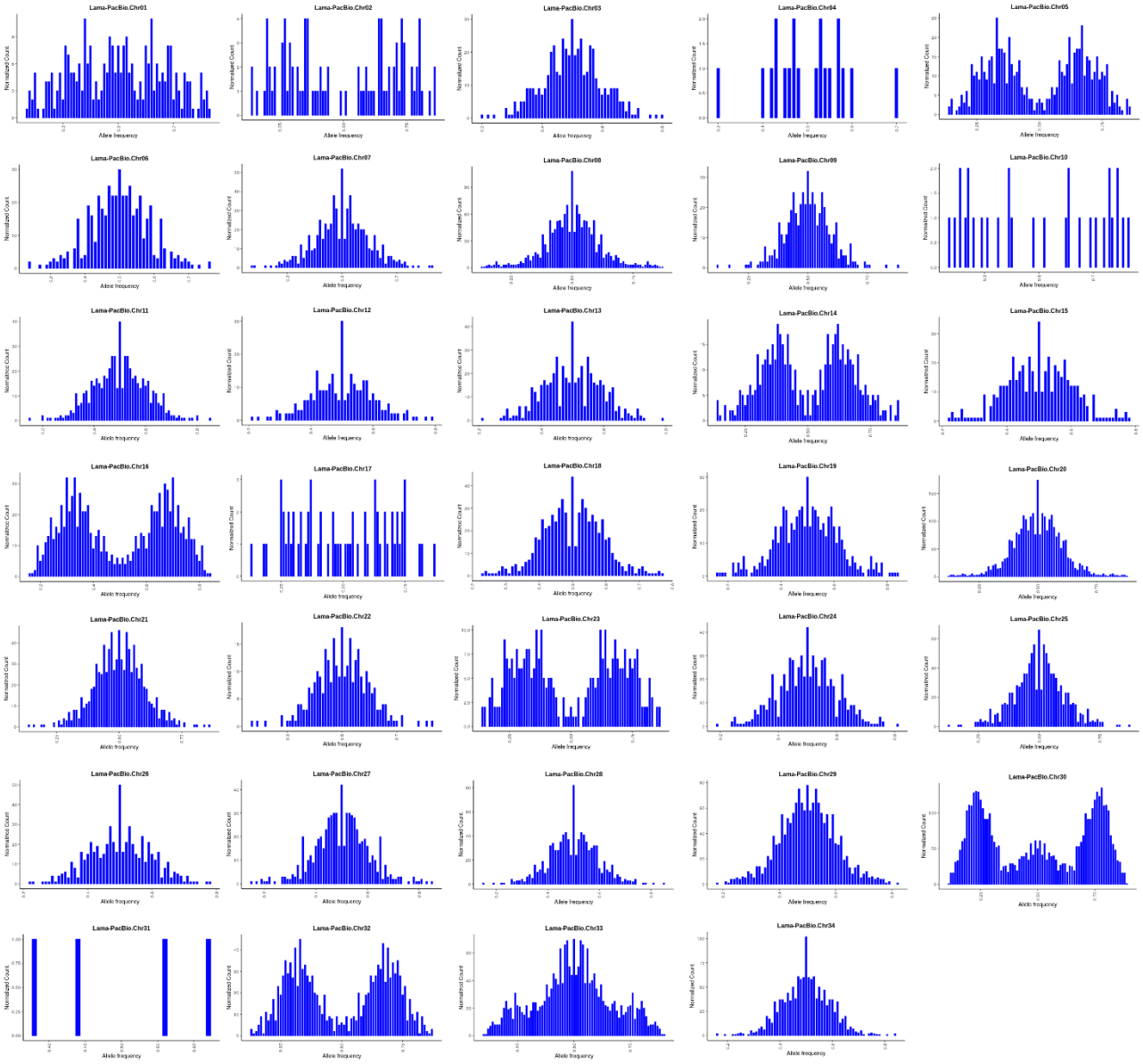

Figura 16. Distribuição das contagens da frequência alélica normalizada para a cepa PH8 de *L. amazonensis*. O eixo y representa as contagens normalizadas e o eixo x representa a frequência alélica.

Usando análise de profundidade de *reads* normalizada para avaliar a variação do número de cópias do gene na cepa PH8, identificamos 205 genes apresentando uma soma maior que 1,8. Dentre esses genes considerados supranumerários ou expandidos, 104 (~51%) codificam proteínas hipotéticas, com funções desconhecidas (Tabela S3; Apêndice V). Famílias multigênicas que codificam transportadores de aminoácidos, serina-treonina fosfatase, ferredoxinas, quinases, calpaínas e metaloproteases GP63 estão entre os genes supranumerários com função conhecida. Vários desses genes são caracterizados como fatores de virulência e estão discutidos com mais detalhes na próxima seção. A análise da ontologia gênica dos genes expandidos mostrou que este grupo é enriquecido por genes envolvidos em processos biológicos relacionados ao transporte transmembrana (8 genes) e metabolismo lipídico (13 genes), entre outros. No total esses genes estão representados em 447 cópias no genoma haplóide de PH8 (Figura 17).

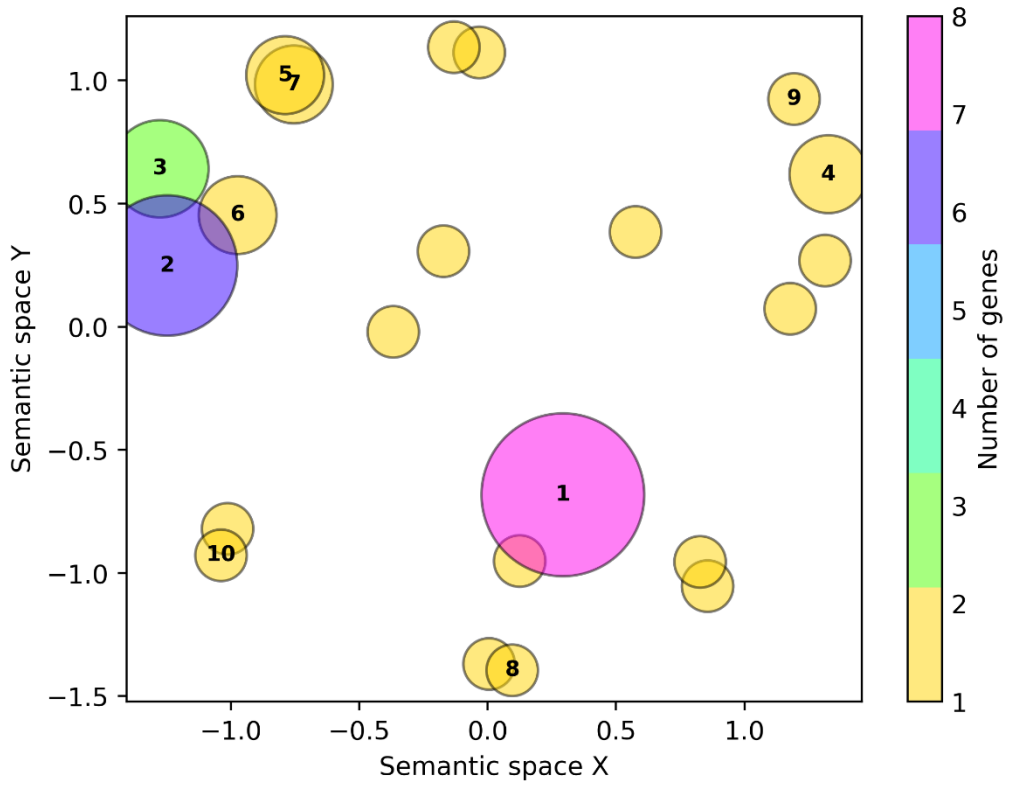

Figura 17. Vias biológicas enriquecidas com base em genes expandidos no genoma da cepa PH8. Os eixos x e y indicam a similaridade semântica entre os termos dos GOs enriquecidos. A escala de cores à direita representa o número de vezes que um termo GO foi anotado para um gene com múltiplas cópias.

5.5. Análises de famílias multigênicas descritas como fatores de virulência em *L. amazonensis* e outros representantes do gênero

L. amazonensis codifica diversas moléculas que atuam como fatores de virulência, principalmente em processos relacionados ao estabelecimento e manutenção da infecção. Dentre as moléculas mais estudadas estão as proteínas A2, amastinas e as leishmanolisinas GP63, cisteíno proteases, fosfatases e cinases, todas codificadas por genes multicópia. Não obstante, genes que codificam proteínas participantes de vias de metabolização do ferro e de heme também foram caracterizados por possuírem um grande papel durante o desenvolvimento da leishmaniose no hospedeiro (LARANJEIRA-SILVA *et al.*, 2020).

Após uma busca automatizada, baseada na similaridade e caracterização de domínios funcionais entre membros das famílias de proteínas A2, amastinas, GP63, cisteíno proteases, fosfatases e cinases, foi determinado o número de membros anotados para cada uma delas no

genoma da PH8 (Tabela 8) e pode ser comparado com outras espécies de *Leishmania*. *Scripts* combinados de Perl e Python realizaram o alinhamento de todos os genes anotados e disponíveis no banco de dados TritrypDB para cada uma das famílias citadas, contra as ORFs previstas em nosso genoma. Outras etapas de refinamento, que estabelecem os limites de cada gene e a identificação de domínios funcionais, permitiram a correção de genes anotados por anotação automática, transferidos de *L. mexicana*, e a identificação de novos genes.

Tabela 8. Número de membros das principais famílias multigênicas anotadas no genoma de *L. amazonensis* cepa PH8 e outra leishmanias.

Família / Espécie	Proteínas A2	Amastinas	GP63	Cisteino proteases	Cinases	Fosfatases
L. amazonensis PH8 strain	9 (4)	31 (2)	11(2)	76	254	124
L. amazonensis M2269 strain	0	20 (2)	2	66	245	97
L. braziliensis	0	55	21	63	261	127
L. donovani	4	33	3	58	258	122
L. infantum	1	68	13	69	266	128
L. major	0	63	5	64	265	130
L. mexicana	2	48	7	77	264	126

*Números entre parênteses indicam pseudogenes.

Para todas as famílias multigênicas analisadas, identificamos um número maior de membros no genoma da PH8, comparado ao genoma da cepa M2269, que é o único genoma de *L. amazonensis* anotado até o momento. Este resultado indica que vários genes não foram previamente anotados possivelmente porque este estudo inicial foi baseado em um genoma altamente fragmentado (REAL *et al.*, 2013). Para vários desses fatores de virulência, um repertório gênico mais completo foi obtido com a atual montagem do genoma PH8. Foi possível constatar, por exemplo, que, apesar de não ter sido identificada nenhuma cópia do gene que codifica o antígeno A2 na anotação da cepa M2269, o número de cópias de genes da família que codifica esse antígeno é, na realidade, significativamente maior em *L. amazonensis* quando comparado à outras espécies de Leishmania. Por outro lado, a família das amastinas sofreu uma maior expansão principalmente em *L. braziliensis*, mas também em *L. infantum* e *L. major*. A mesma analogia pode ser feita com

relação à família das GP63, tendo sofrido maior expansão também em *L. amazonensis*, *L. braziliensis* e *L. infantum*. Para as demais famílias não foram observadas diferenças significativas no número de membros quando consideradas as demais espécies de *Leishmania*.

Para as etapas seguintes de análises de famílias multigênicas que codificam fatores de virulência foram selecionadas quatro delas, onde o repertório de sequências completas no genoma da cepa PH8 de *L. amazonensis* foi estudado e comparado aos demais Tripanosomatídeos.

5.5.1. Família dos genes codificadores de proteínas A2

As proteínas A2 são descritas como importantes fatores de virulência devido a sua capacidade de resistir a temperaturas mais altas, encontradas nos órgãos viscerais do hospedeiro. Essa característica reflete no processo de visceralização da leishmania, no qual o parasito migra da derme para tecidos viscerais, como os do fígado e baço, e se multiplica (MIZBANI et al., 2011).

Durante o processo de curadoria automatizada e manual da anotação de famílias multigênicas foram consideradas 11 sequências de aminoácidos de proteínas A2, provenientes de *L. donovani*, *L. infantum*, *L. mexicana*, *L. major* e *L. chagasi*, as quais foram alinhadas contra o genoma da PH8 e apenas *hits* com pelo menos identidade > 85% e *e-value* < 10^{-5} foram considerados para efeito de anotação. Sendo assim, esta análise revela, pela primeira vez, o número de cópias de genes que codificam proteínas A2 no genoma de *L. amazonensis*. Enquanto que em *L. mexicana* foram constatadas 2 cópias, 4 cópias em *L. donovani* e 1 cópia em *L. infantum*, no genoma PH8, 9 sequências com homologia a genes que codificam proteínas A2 foram detectadas, contudo, apenas 5 possuem ORFs completas, sendo, portanto, 4 são pseudogenes. Dois dos cinco genes A2 foram encontrados no cromossomo 22 (Figura 12b), os outros três foram associados a pequenos *scaffolds* que não foram incorporados na montagem final (scaffold278, Scaffold282 e Scaffold294, Figura 13). As proteínas A2 codificadas por esses genes possuem entre 600 e 1.089 aminoácidos, dependendo do número de repetições, que constituem a maior porção da proteína. Esta estrutura da proteína A2 *de L. amazonensis* é semelhante à A2 de outras espécies de *Leishmania*, com uma sequência líder secretora seguida pelo módulo repetido de 10 aminoácidos VGP[Q/L]SVGPQS que ocorrem 40 a 90 vezes (CHAREST; ZHANG; MATLASHEWSKI, 1996; H; G, 1994; WEN-WEI ZHANG *et al*., 1996). Nas proteínas A2 da cepa *L. amazonensis* PH8, a sequência líder conservada e a repetição de 10 aminoácidos VGPQSVGPQS ou VGPLSVGPQS são intercaladas com agrupamentos de outra repetição, repetição SLLARSLLAR, que foram observadas apenas em proteínas A2 de *L. amazonensis* e *L. donovani* (Figura 18).

Um alinhamento realizado com sequências de aminoácidos de proteínas A2 de *L. amazonensis*, *L. mexicana*, *L. infantum* e *L. donovani* mostrou diferenças no número de repetições

que varia de 47 a 59 em *L. donovani* (LYPACZEWSKI *et al*., 2018), mas é limitado a 41 repetições em 4 dos 5 genes existentes na cepa PH8 de *L. amazonensis* (Figura 18 e Figura 19). Além disso, a mutação do aminoácido G para o aminoácido D (G → D) na 7ª posição de múltiplas repetições, descrita como sendo importante para o correto dobramento e função das proteínas A2, foi identificada apenas na proteína codificada pelo gene presente no scaffold 278 (Lama IV). Similar às proteínas A2 de *L. donovani* essa cópia de A2 possui mais de 90 repetições de 10 aminoácidos, muitas delas contendo a mutação G□□D (Figura 18). Em *L. mexicana* esta mutação ocorre na 2ª posição e gera um padrão de repetição único, com repetições VGPLSVDPQS (verde claro) e VDPQS (azul claro). Além do desvio no padrão de repetição, os genes A2 em *L. amazonensis*, assim como em *L. donovani*, também possuem sequências N-terminais conservadas.

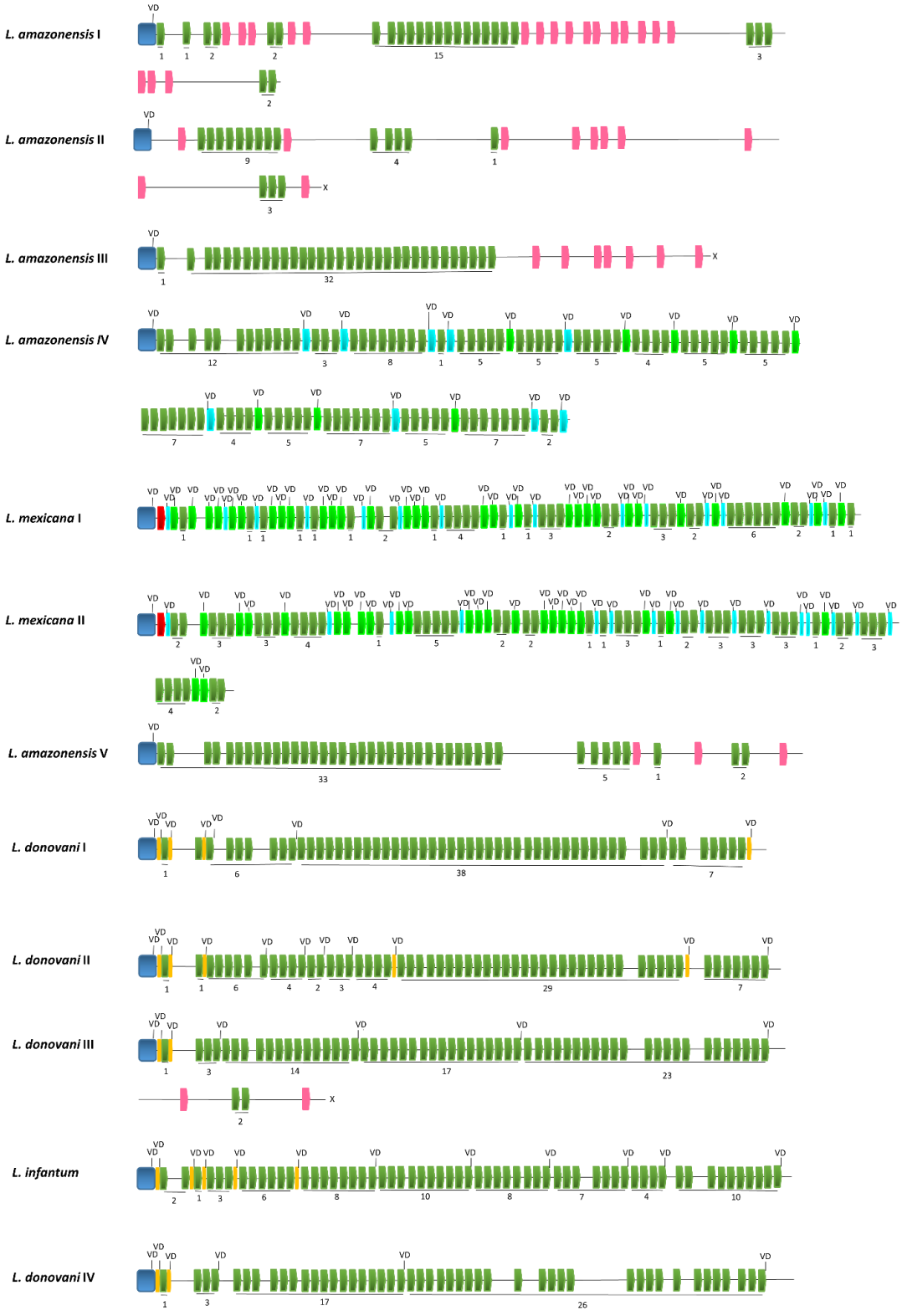

Figura 18. Alinhamento esquemático de proteínas A2 anotadas em diferentes espécies de *Leishmania*. Quadrados azuis e setas coloridas, respectivamente, representam regiões conservadas e repetições principais de cada proteína. Setas verdes escuras: repetições de VGP(L/Q)SVG(P/S)QS ou VGPQ(A/S)VGPLS ou VGPEAVGPLS ou VGPLSVGPQA; Setas amarelas: VGLPSVD, Setas cor-de-rosa: repetições SLLARSLLAR; Setas verdes claras: VGPLSVDPQS; Setas azuis claras: VDPQS. A linha preta representa lacunas no alinhamento e outras repetições não representadas nas sequências de *L. amazonensis*. Os números abaixo das repetições indicam o número de vezes que aparecem em cada intervalo. Mutações G☐☹D são indicadas por traços localizados no meio das setas verde claras e após as setas verdes escuras. A numeração que abrange grupos de setas indica o número de repetições neles contidos.

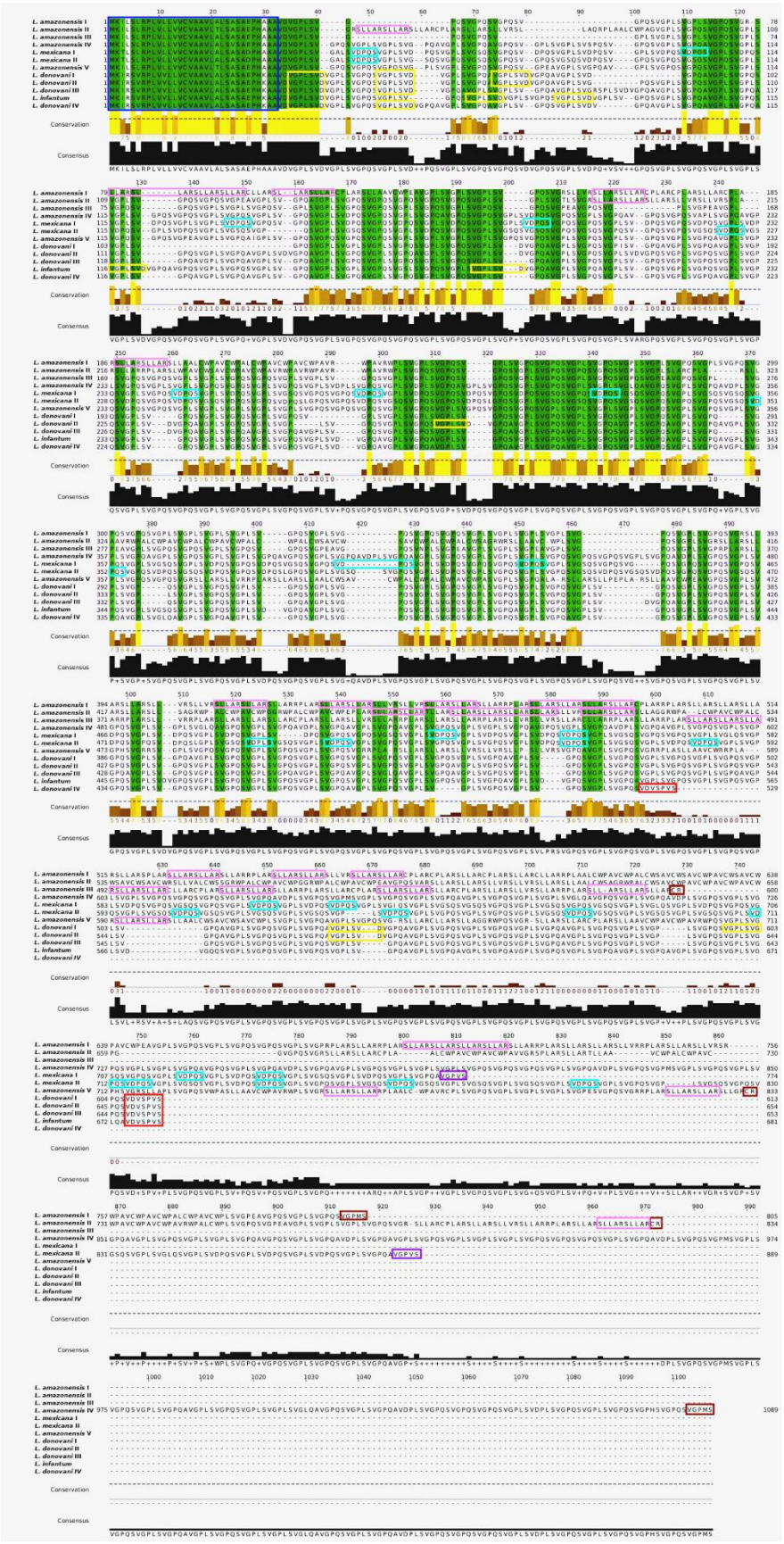

Figura 19. Alinhamento entre as sequências de aminoácidos das proteínas A2 anotadas em *L. amazonensis* (PH8), *L. mexicana, L. donovani* e *L. infantum*. O retângulo azul escuro representa os 32 aminoácidos sequência secretora; retângulos amarelos indicam repetições com inserções RV no final; retângulos azuis claros indicam repetições exclusivas de *L. mexicana* com inserções precoces de VD; retângulos rosa indicam repetições exclusivas de *L. amazonensis*; retângulos vermelhos indicam a sequência final de cada proteína.

5.5.2. Família dos genes codificadores de amastinas

As amastinas, codificadas por dezenas de cópias também são descritas como importantes fatores de virulência. Expressas em altos níveis em formas amastigotas, estudos mostram seu papel na interação parasito-macrófago. Queda na sobrevivência de amastigotas intracelulares é observada quando estes parasitos são depletados para os genes que codificam a classe mais abundante dessa família de proteínas (DE PAIVA et al., 2015).

Várias sequências codificadoras de membros da família das amastinas foram identificadas por meio da transferência da anotação do genoma de *L. mexicana.* Buscando identificar outras sequências dessa família no genoma da PH8, sequências homólogas de outras espécies da família da *Leishmania, T. cruzi* e *T. brucei* foram consideradas na busca. Dessa forma 269 sequências de aminoácidos de amastinas, provenientes de *L. brazilienzes, L donovani, L. infantum, L. major* e *L. mexicana*, 27 de *T. cruzi* e 4 de *T. brucei* foram alinhadas contra todas as ORFs de *L. amazonensis* e novamente apenas *hits* com pelo menos identidade > 85% e e-value < 10^{-5} foram considerados para efeito de anotação. Os resultados desse alinhamento mostraram que, em comparação com outras espécies de *Leishmania*, as amastinas apresentam menor número de cópias em *L. amazonensis*, com um total de 31 genes e 2 pseudogenes distribuídos em 10 dos 34 cromossomos; cromossomos 8, 10, 14, 20, 24, 27, 28, 29, 30 e 33 (Figura 12a-b), bem como em 3 pequenos *scaffolds* que não foram incorporados na montagem final (Figura 13). Os genes de amastinas estão organizados em pequenos *arrays em tandem* variando de 2 a 4 genes e o tamanho das proteínas codificadas varia entre 174 aa e 546 aa.

O alinhamento realizado entre as 29 sequências de amastinas (excluindo 2 pseudogenes) revelou que existe uma região de 11 aminoácidos (C- [IVLYF] - [TS] - [LFV] - [WF] -GX- [KRQ] - X - [DENT] - C) conservada entre as posições 461-472 do alinhamento em pelo menos 50% das sequências (Figura 20, a sequência consenso é mostrada na parte inferior do alinhamento), que corresponde à assinatura desta família, proposta por Rochette *et al.* (2005) (ROCHETTE *et al.*, 2005). Este resultado contrasta com o encontrado para amastinas de *L. braziliensis* e *L. infantum* e *L. major*, respectivamente, uma vez que nestas espécies a assinatura de amastinas foi conservada em todas as sequências. Além disso, a conservação do aminoácido na posição 10 da assinatura não foi observada, como esperado nestas espécies (DE PAIVA *et al*., 2015; ROCHETTE *et al.*, 2005). Finalmente, as regiões correspondentes aos quatro domínios transmembrana previstos mostraram alta conservação, com grande similaridade de sequência entre os membros da família das amastinas.

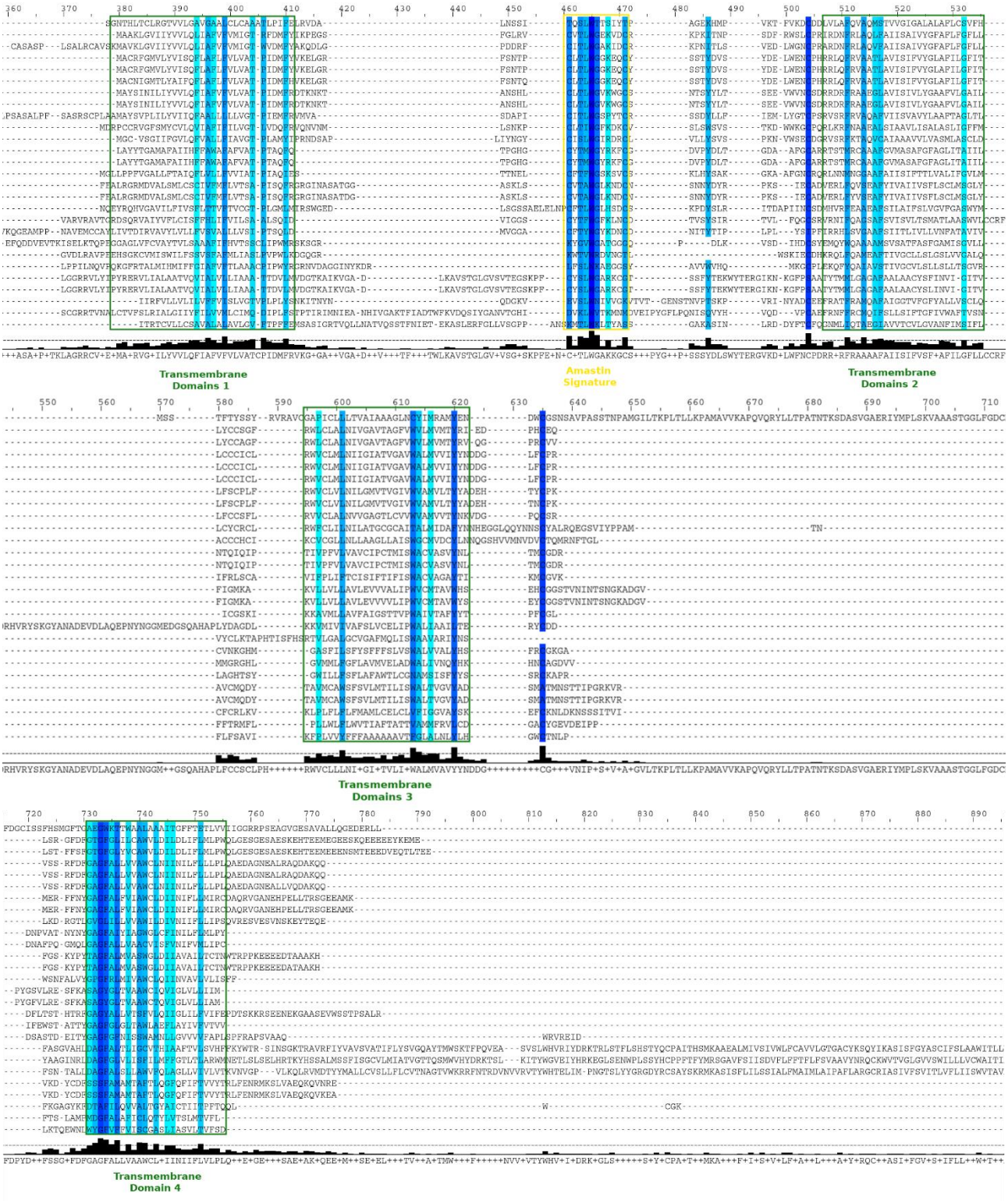

Figura 20. Caracterização *in silico* de sequências de amastinas anotadas no genoma de *L. amazonensis*. 29 sequências de aminoácidos foram alinhadas com a ferramenta MAFFT. A escala azul (claro a escuro) representa o grau de identidade entre os aminoácidos que são conservados em pelo menos 50% das sequências alinhadas. O histograma preto abaixo mostra o grau de identidade para todas as posições. A sequência de consenso é destacada abaixo dela. Os retângulos verdes representam os quatro domínios transmembrana e em amarelo a posição da sequência da assinatura das amastinas descrita por Rochette *et al.* (2005).

A filogenia mostrada na Figura 21 reforça a classificação de quatro subfamílias da amastinas de *L. amazonensis* e sua associação com a posição genômica. Dezesseis sequências de

amastinas pertencem à subfamília das δ-amastinas, a maior subfamília, e estão distribuídas nos cromossomos 8, 30 e 33. Quatro δ-amastinas estão agrupadas em uma posição de ramificação basal e foram classificadas como proto-δ-amastinas. Seis cópias pertencem à subfamília das β-amastinas; 4 a α-amastinas e 3 a γ-amastinas. A análise filogenética também indicou que, semelhante a outros membros da família Tripanosomatídeos, as β-amastinas se dividiram em duas linhagens intimamente relacionadas correspondendo a isoformas distintas localizadas nos cromossomos 10, 28 e 29. Da mesma forma, as α-amastinas também apresentaram maior divergência evolutiva, com genes surgindo a partir das δ-amastinas e outra isoforma proveniente de um nó independente.

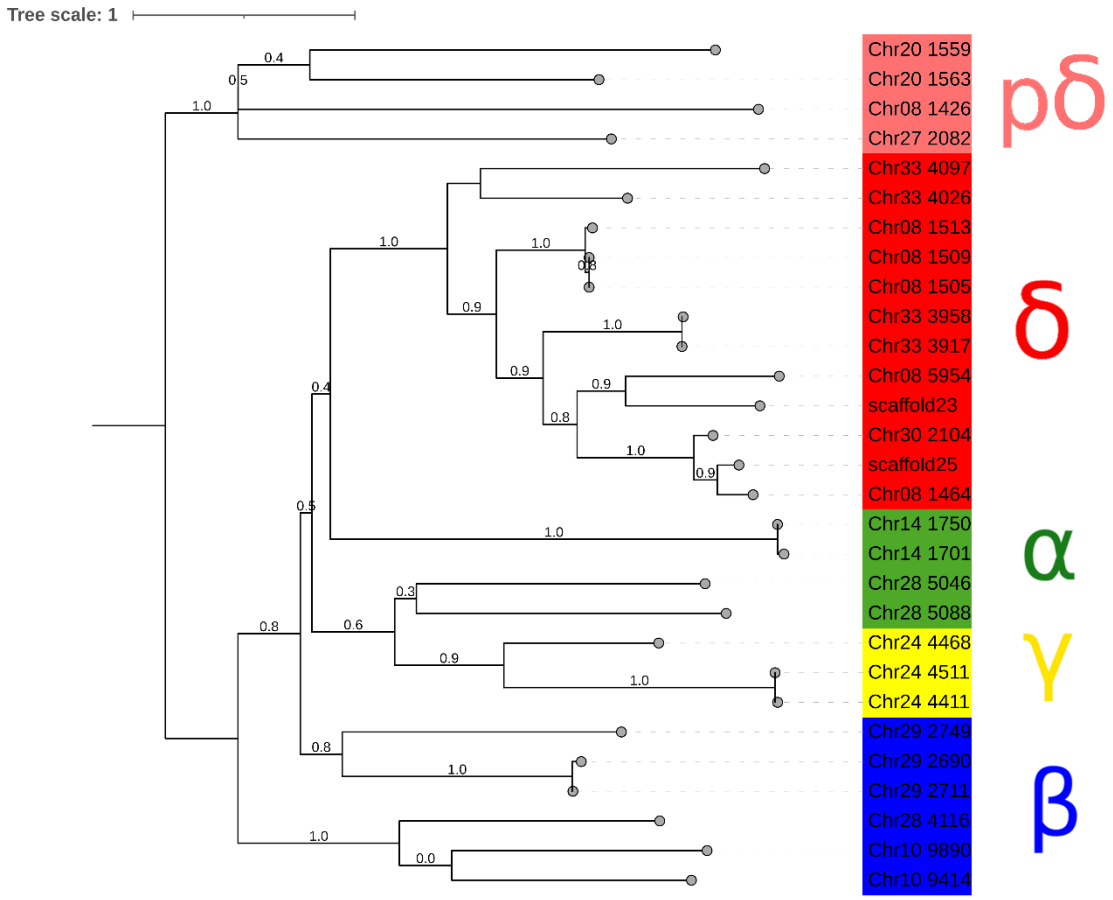

Figura 21. Árvore filogenética das sequências de aminoácidos de 29 amastinas de *L. amazonensis*. A estimativa filogenética foi realizada usando a máxima verossimilhança implementada pelo *software* Phyml. Os comprimentos dos ramos são desenhados proporcionalmente à mudança evolutiva, com valores de *bootstrap* mostrados no meio de cada ramo. Os círculos cinza representam os nós terminais da árvore, com cada uma das sequências alinhadas. O modelo de evolução de melhor ajuste foi definido como VT+G, determinado pelo ProtTest. A classificação em quatro subfamílias de amastinas mostradas à direita foi baseada em Jackson (2010). pδ = proto-delta; ô = delta; β = beta; a = alfa; γ = gama.

As proto-delta amastinas aparecem por ter evoluído de forma independente das demais subfamílias e por apresentarem maior semelhança de sequência a membros das delta-amastinas também foram classificadas dentro desta mesma subfamília. Além disso, a análise filogenética

realizada a partir de alinhamento revelou que existe uma relação evolutiva associada a essas subfamílias, uma vez que amastinas do mesmo tipo formam *clusters* que evidenciam a relação entre elas.

Árvore filogenética construída a partir do alinhamento entre 311 sequências de proteínas que codificam amastinas em *L. amazonensis, L. braziliensis, L. donovani, L. infantum, L. major, L. mexicana, T. cruzi* e *T. brucei* e 31 identificadas no genoma da PH8 reforçou as evidências de proximidade evolutiva entre *L. amazonensis* e *L. mexicana*, uma vez que todas as sequências de PH8 (LamaPacbio) e *L. mexicana* ou estão envolvidas em um dos 21 *clusters* formados, reunindo mais de duas sequências de amastinas ou encontram-se próximas a uma única sequência de *L. mexicana* (Figura 23). Também é possível identificar subfamílias de amastinas exclusivas de *L. donovani*, *L. infantum*, *L. major* e *T. cruzi*. Além disso, a análise realizada com o uso da ferramenta OrthoMCL previu 47 grupos ortólogos entre as amastinas anotadas para as espécies citadas e possibilitou a identificação de sub-famílias nas demais espécies de *Leishmania*, *T. cruzi* e *T. brucei*.

A topologia da árvore sugere ramificação precoce de alguns *clusters* ou clados espécie-específicos, formados sobretudo por alpha, beta e gamma-amastinas, indicando a presença de uma classe de amastinas conservada em *Leishmania* antes de sua radiação. A presença de clados espécie-específicos de δ-amastinas em ramos de árvores terminais (ramos marrons, rosas, verdes e roxos) sugere que vários genes de amastinas apareceram devido a pressões seletivas ambientais ou especiação de patógenos. Esse resultado reforça o que foi sugerido por Real e colaboradores (2013), além de adicionar novas evidências evolutivas, mostrando que as β-amastinas, por exemplo apresentam um grau de divergência maior ao que foi visto anteriormente. Essa constatação pode ainda sugerir o surgimento de uma nova subfamília de amastina, uma vez que as direções evolutivas são antagônicas, tendo uma classe sido originada a partir do mesmo nó que também originou o clado das δ-amastinas, enquanto a outra classe surgiu de um nó que irradiou também para o surgimento das α-amastinas.

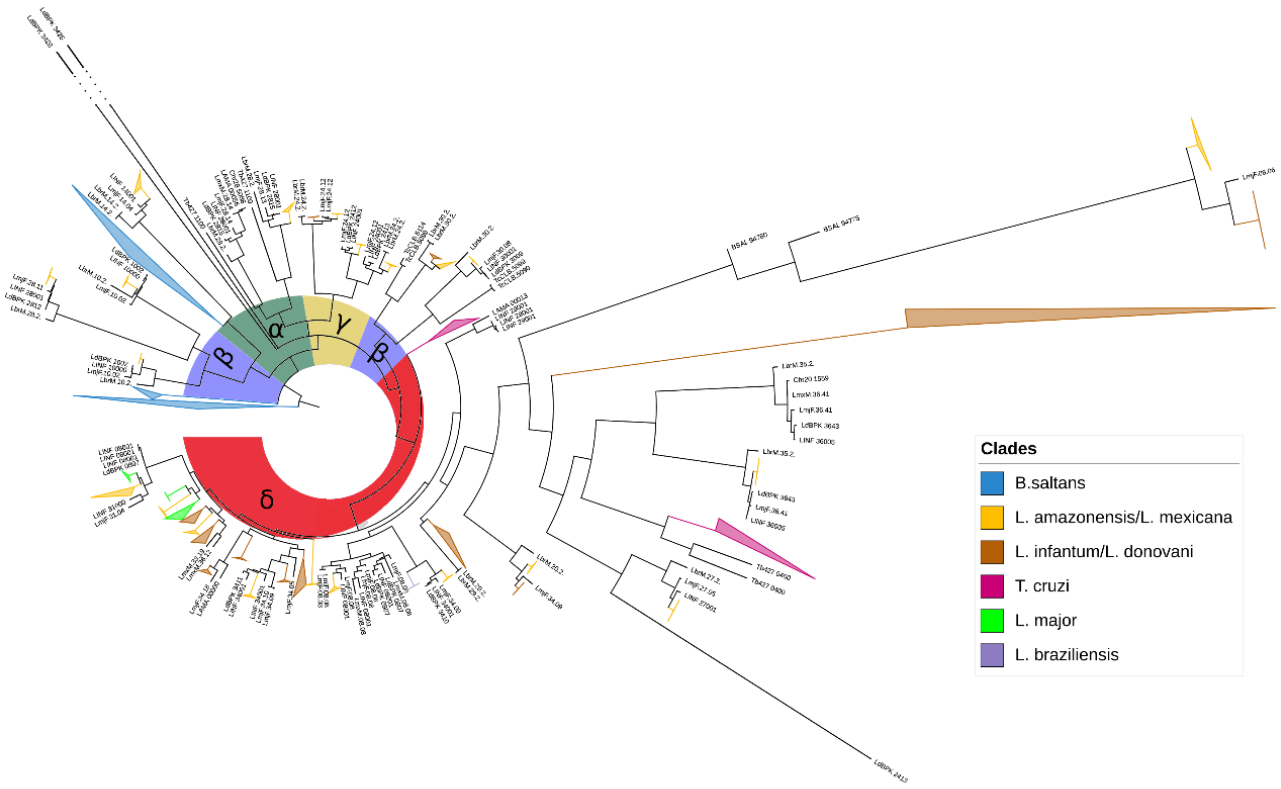

Figura 22. Filogenia de máxima verossimilhança de proteínas de superfície amastina em tripanosomatídeos. O filograma é representado por um consenso de 311 sequências de amastina. JTT+G foi usado como matriz de substituição para alinhamento de proteínas. Os clados colapsados destacados na cor laranja representam aglomerados de amastins da cepa PH8 de *L. amazonesnis* e *L. mexicana*. Clados colapsados de amastinas específicas de *L. infantum/L.donovani*, *T. cruzi*, *L. major* e *L. braziliensis* são destacados em marrom, rosa, verde e roxo, respectivamente. As subfamílias de amastinas são indicadas em vermelho (delta-amastinas), azul (beta-amastinas), verde (alfa-amastinas) e amarelo (gama-amastinas).

5.5.3. Família dos genes codificadores de GP63

Proteínas GP63 ou leishmanolisinas constituem as principais proteases de superfície expressa nas formas promastigotas de *Leishmania* (KULJARNI et al., 2009). Seu papel principalmente como facilitador da interação leishmania-macrófago é bastante estudado, além da sua capacidade de bloquear os efeitos microbicidas, permitindo consequentemente a progressão da infecção causada pelo parasito (OLIVIER et al., 2012).

As mesmas buscas também foram realizadas para a família das leishmanolisinas também denominadas GP63. A busca no genoma da PH8 utilizando um *pipeline* automatizado e posterior caracterização funcional, partindo de um conjunto de 593 sequências de GP63 de outras *Leishmanias*, *T. cruzi* e *T. brucei*, resultou em 11 genes que codificam proteínas GP63, dos quais 2 tinham códons de parada internos e, portanto, foram classificados como pseudogenes. A maioria desses genes está localizada no cromossomo 10, formando um único cluster composto por 7 genes, respectivamente. Outros 2 genes que codificam GP63 foram observados nos cromossomos 28 e 30.

Para caracterizar *in silico* as sequências de GP63 anotadas, foi realizado um alinhamento entre as 9 sequências de aminoácidos das proteínas desta família e posterior identificação de importantes domínios conservados (Figura 23A). Os resultados indicam a presença de dois núcleos funcionais previamente descritos (MCGWIRE; CHANG, 1994): o primeiro composto pelo domínio HEXXHA, que representa uma característica da família de enzimas metaloproteases, responsável pela ligação a um átomo de zinco, e também pelo motivo adesivo SRYD que é importante na ligação de receptores de superfície de macrófagos. No segundo núcleo funcional há um resíduo de histidina (posição 384 do alinhamento) que, coordenado aos resíduos de histidina do primeiro núcleo, participa da formação de uma estrutura que estabiliza o zinco. Sequências importantes para a internalização do parasita pelas células hospedeiras também são conservadas nesta família. Ensaios de inibição de internalização mostraram que essas sequências fazem parte de domínios que impactam o processo de interação com receptores específicos, facilitando a posterior internalização do parasita (PUENTES *et al*., 1999). O motivo de sete aminoácidos consecutivos, denominado KDELMAP também foi encontrado conservado no genoma da PH8 em todas as sequências alinhadas, no entanto, seu papel ainda não está claro (CASTRO NETO *et al*., 2019).

Predições da estrutura secundária e terciária de um dos membros do GP63 (localizado no cromossomo 10), realizadas nos *softwares* PSIPRED e SWISS-MODEL, respectivamente, mostram que em *L. amazonensis*, os genes desta família também são formados por folhas-beta, em sua maioria, mas também por estruturas secundárias alfa-hélice (Figuras 23B-C). Além disso, a estrutura 3D modelada a partir de outras estruturas tridimensionais de GP63s depositadas no banco de dados PDB possibilitou evidenciar a presença de domínios funcionais SRDY e HEXXHA em regiões alfa-hélice, enquanto domínios relacionados a processos de internacionalização são encontrados em regiões de folha-beta coordenado com o átomo de zinco. O domínio N-terminal não mostrou homologia suficiente com qualquer outra estrutura resolvida e, portanto, o peptídeo sinal característico e o pró-peptídeo não estão representados. Esta descoberta apoia a ideia de que esta região é altamente variável nesta família de proteínas de ligação ao zinco (SCHLAGENHAUF; ETGES; METCALF, 1998).

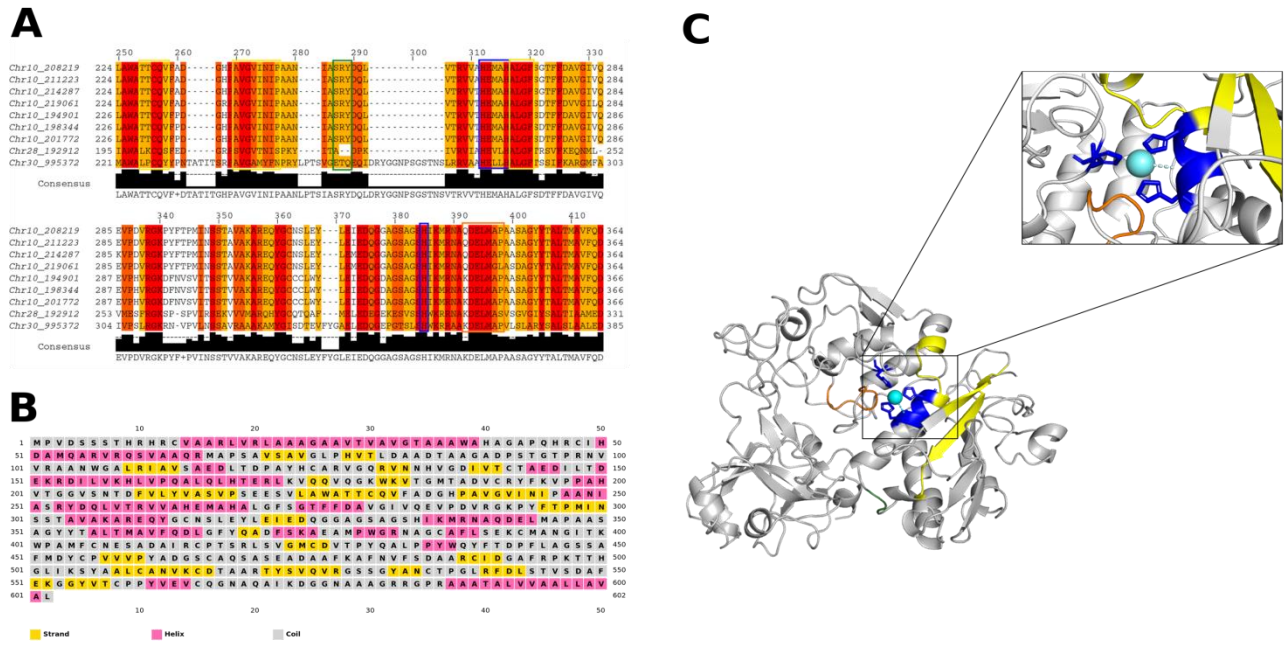

Figura 23. Caracterização in silico de sequências de aminoácidos de GP63 anotadas *em L. amazonensis*. (A) Alinhamento múltiplo de 9 sequências alinhadas com a ferramenta MAFTT. As cores vermelha e laranja representam o grau de identidade entre os aminoácidos que são conservados em pelo menos 80% das sequências alinhadas. Os retângulos azuis representam o domínio HEXXHA e o resíduo de histidina, que juntos formam uma "estrutura" coordenada que favorece a ligação do zinco. O retângulo verde marca o domínio SRYD e os retângulos amarelos indicam regiões importantes nos processos de internalização da célula hospedeira. O retângulo laranja destaca o motivo universalmente conservado de sete aminoácidos consecutivos denominado KDELMAP por Neto *et al.* (2019). **(B)** Predição de estrutura secundária 2D realizada no PSIPRED. Em amarelo são mostradas as sequências de fitas. A cor rosa representa sequências de estruturas helicoidais. Em cinza estão representadas as sequências de motivos de bobinas, formadas por alfa-hélices enroladas. **(C)** representação da estrutura terciária 3D (posição 100-574) de um GP63 localizado no cromossomo 10 (Chr10_208219) do genoma da cepa PH8. As cores na estrutura representam os mesmos domínios e motivos descritos no alinhamento. Na parte superior da figura destaca-se a ligação da molécula de zinco (azul claro).

A mesma análise também foi realizada para as demais proteínas codificadas por genes de GP63 no genoma da PH8 e foram vistas apenas variações estruturais, mas não quanto a localização de domínios importantes, os quais permanecem em regiões de folha-beta e alpha-hélice.

5.5.4. Família dos genes codificadores de cisteína proteases

Proteases são enzimas amplamente estudadas como fatores de virulência. Suas funções têm sido relacionadas à muitas atividades parasitárias, como invasão tecidual, sobrevivência em macrófagos e modulação da resposta imune do hospedeiro. As cisteíno proteases constituem uma das várias classes enzimáticas de endopeptidases e são recebem este nome devido ao aminoácido relacionado a catálise e a natureza de seu sítio catalítico, que neste caso é o resíduo de cisteína.

No genoma do PH8, foram anotados 76 genes codificadores de cisteíno proteases, o que representa 0,91% do total de genes dessa espécie. Esses genes estão distribuídos em vinte e três cromossomos, com formação de agrupamentos gênicos nos cromossomos 8 (cinco genes em *tandem*), 20 (nove e quatro genes em *tandem*) e 30 (três genes em *tandem*). A maioria dos genes é

atribuída ao Clã A; sendo trinta e nove semelhantes a calpaína, família C2, dez semelhantes à papaína, incluindo três genes semelhantes a catepsina L e um gene semelhante a catepsina B. Membros das famílias C12, C48 e C78 também estão presentes nesta espécie, sendo observado apenas um gene em cada uma, enquanto para as famílias C54 e C51 foram identificados quatro genes no total, dois em cada. Nas famílias designadas C19 e C65, correspondentes às hidrolases de ubiquitina e otubaína C-terminal, respectivamente, doze e dois genes estão presentes. Finalmente, para os clãs CF, CD e PC, foi anotado um gene dentro de cada família existente. Esta é a primeira vez que o repertório completo de cisteína proteases expressas em *L. amazonensis* é relatado e classificado nas diferentes famílias, além de refletir a identificação de 9 novos genes anotados para esta família em relação à anotação atual disponível para a cepa M2269.

5.6. Genes que codificam proteínas relacionadas ao metabolismo de heme e ferro

Ferro e heme são essenciais em muitas vias metabólicas conservadas na *Leishmania*. No entanto, além da falta de capacidade de síntese de heme, esses parasitas são incapazes de armazenar ferro ou heme (CHIN SHEN CHANG; KWANG-POO CHANG, 1985; FLANNERY *et al*., 2011; KOŘENÝ; OBORNÍK; LUKEŠ, 2013). Além disso, durante seu ciclo de vida digenético, a *Leishmania* passa por mudanças ambientais significativas desde o intestino médio e probóscide do flebotomíneo até os vacúolos fagolisossomais dos macrófagos de mamíferos, onde são expostos a grandes mudanças na disponibilidade de nutrientes (LARANJEIRA-SILVA; HAMZA; PÉREZ-VICTORIA, 2020; SCHAIBLE; KAUFMANN, 2004). Para sobreviver e replicar, esses parasitas devem adquirir ferro e heme de seus hospedeiros enquanto equilibram seus níveis intracelulares, pois o ferro livre e o heme podem ser citotóxicos quando em altas concentrações (DIXON; STOCKWELL, 2013). Nos últimos anos, várias proteínas envolvidas no metabolismo de ferro e de heme de *Leishmania* foram identificadas como fatores de virulência devido à sua importância no estabelecimento e progressão da doença (CHIN SHEN CHANG; KWANG-POO CHANG, 1985).

Uma vez que todas essas proteínas relacionadas ao metabolismo de ferro e de heme estão associadas à virulência do parasita, o genoma da PH8 foi interrogado para detectar a presença de genes que codificam cada uma das proteínas envolvidas. Dessa forma foram identificados 14 genes relacionados ao metabolismo de ferro e de heme. A Tabela 10 mostra o conjunto de genes de *L. amazonensis* em comparação com outras espécies de *Leishmania* e também com os genomas de *T. brucei* e *T. cruzi*.

Tabela 9. Similaridade de sequência de genes relacionados ao metabolismo de heme e ferro presentes em *Leishmania* spp, *T. cruzi* e *T. brucei*.

	L. amazonensis PH8	*L. mexicana*	*L. braziliensis*	*L. donovani*	*L. infantum*	*L. major*	*T. brucei*	*T. cruzi* (haplótipo Esmeraldo)
LIT1-1	100,00%	99,07%	73.84%	90,83%	87,96%	85,88%	_	_
LIT1-2	100,00%	98,38%	73,84%	90,83%	87,96%	85,88%	_	_
LFR1	100,00%	98,80%	72,72%	84,40%	84,40%	82,83%	30,46%	33,20%
LIR1	100,00%	98,79%	73,95%	88,10%	91,60%	87,76%	_	47,75%
LMIT1	100,00%	98,63%	86,94%	95,53%	95,53%	95,53%	53,01%	53,21%
SODA	100,00%	98,70%	72,61%	85,95%	85,95%	83,04%	63,45%	63,47%
SODB1	100,00%	92,23%	69,79%	86,98%	86,98%	83,33%	62,98%	63,37%
SODB2	100,00%	93,23%	72,82%	89,23%	89,23%	86,15%	68,72%	67,69%
APX	100,00%	9934%	84,11%	91,75%	91,75%	89,44%	_	61,36%
LHR1	100,00%	100,00%	68,02%	92,57%	93,14%	92,00%	25,90%	36,76%
ANCG5	100,00%	97,41%	70,47%	85,16%	85,24%	84,05%	45,95%	49,96%
ABCB3	100,00%	98,99%	80,36%	91,83%	91,94%	90,53%	59,50%	73,81%
Hbr	100,00%	99,15%	89,38%	97,03%	_	96,60%	_	_
LmFLVCRb	100,00%	98,06%	74,73%	90,06%	89,85%	86,18%	_	_

Os genes que codificam proteínas do metabolismo de heme e de ferro de *L. amazonensis* estão conservados entre *L. mexicana*, *L. major*, *L. donovani* e *L. infantum*, mas divergiram em maior grau dos genes de *L. braziliensis*, resultado que reforça a proximidade evolutiva de *L. amazonensis* com espécies relacionadas à LV. Inicialmente predito para estar localizado em um cromossomo diploide, as duas cópias em conjunto do gene LIT1 (LIT1-1 e LIT1-2) estão localizadas no cromossomo 30 tetraplóide PH8, sugerindo que as cópias extras favorecem a adaptação de *L. amazonensis*. Essa localização do genoma é semelhante à de *L. mexicana*, mas distinta dos genes de *L. major*, que estão localizados no cromossomo 31 (LmjF31.3060 e LmjF31.3070). Mesmo antes de sua correta localização cromossômica ser determinada, o gene LIT1 de *L. amazonensis* foi funcionalmente caracterizado pela geração de cepas *knockout* com LIT1-1 e LIT1-2 deletados. Os mutantes nulos de LIT1 cresceram normalmente em cultura axênica e não apresentaram defeitos de diferenciação em formas infecciosas. Consistente com um papel essencial para LIT1 no crescimento intracelular, a replicação intracelular de amastigota LIT1-mutantes foi completamente cessada (HUYNH; SACKS; ANDREWS, 2006).

6. DISCUSSÃO

A leishmaniose é um grave problema de saúde pública e está entre as dez principais doenças tropicais negligenciadas, com mais de 12 milhões de pessoas infectadas no mundo (WHO, 2022). Suas formas cutânea e mucocutânea registram uma média de 52.600 casos anualmente nas Américas, com uma taxa média de letalidade de 7% entre os casos de leishmaniose cutânea, mucocutânea e leishmaniose visceral (PAHO, 2021). Para que novos métodos de controle dessa doença possam ser desenvolvidos, é muito importante compreender as bases moleculares dos processos de interação parasito-hospedeiro, o que pode ser feito de forma aprofundada somente com os estudos de genomas dos seus agentes etiológicos. Como resultado desse trabalho de Tese, a sequência completa do genoma da cepa PH8 de *L. amazonensis*, que consiste em 34 pseudocromossomos, se torna disponível para os grupos interessados em estudos envolvendo o genoma desse parasito. Diferentemente da montagem do genoma da cepa M2269 de *L. amazonensis*, hoje a única disponível no banco de dados TriTrypDB, a montagem do genoma da PH8, foi obtida a partir de sequenciamento com duas plataformas complementares e o alinhamento com o genoma de referência de *L. mexicana*, totalizando ~32 Mb de tamanho. As estimativas de tamanho dos genomas das linhagens M2269, RZOD01 e UA301 previram tamanhos semelhantes na montagem final, mas diferentes níveis de contiguidade, chegando a 2.627, 92 e 34 *contigs*, respectivamente, mostrando que a qualidade das montagens obtidas foi bastante divergente. Apesar de sequências de genomas de outras cepas de *L. amazonensis* estarem disponíveis, apenas a anotação para a cepa M2269 pode ser consultada em bancos de dados. A obtenção de uma montagem mais contígua para a cepa PH8 só foi alcançada devido à abordagem híbrida adotada em nosso *pipeline*, que utilizou sequências longas geradas na plataforma PacBio e *reads* Illumina nas etapas de correções de *contigs scaffolding.*

Para entender melhor a dificuldade em alcançar montagens com níveis de contiguidade que correspondam ao número exato de cromossomos preditos para cada espécie, é necessário pontuar as principais características intrínsecas as plataformas usadas no sequenciamento, *softwares* utilizados em montagens e também características ligadas ao genoma da espécie. Em geral, o genoma de *Leishmania* possui sequências repetitivas em abundância espalhados por todos os cromossomos. De forma semelhante e ainda mais significativa, no genoma de *T. cruzi* essas sequências correspondem a ~50% do seu genoma (EL-SAYED *et al.*, 2005). Tais sequências estão localizadas, por exemplo, nos telômeros, elementos transponíveis e genes organizados em *clusters*, ocupando grandes extensões do genoma (REQUENA *et al.*, 2017; UBEDA *et al.*, 2014). Outra característica encontrada nos genomas é a presença de sequências altamente homólogas e variantes estruturais, como deleções, duplicações, inserções, inversões e translocações, as quais são dificilmente

reconstituídas utilizando apenas abordagens de sequenciamento de *reads* curtas (revisado por BARTHOLOMEU; TEIXIRA; CRUZ, 2021).

De fato, ao analisar a complexidade de diversos genomas de Trypanosomatídeos fica clara a necessidade da utilização de plataformas de sequenciamento que sejam capazes de abranger regiões cada vez maiores do genoma, além de detectar a organização espacial do DNA, capturando informações de pontos de contato na cromatina, o que é permitido pela técnica de sequenciamento Hi-C (BELTON et al., 2012). Nesse sentido, diversas plataformas foram desenvolvidas, como é o caso das plataformas Illumina, Pacbio e Nanopore, as quais, se usadas em conjunto possibilitam a resolução de *gaps* deixados por estratégias que geram um único tipo de *read*, seja ela curta ou longa (MILLER *et al.*, 2017). Outro aspecto a ser considerado na escolha da plataforma é a precisão no *base call*, que deve amenizar erros produzidos durante o sequenciamento. Da mesma forma, o número de *reads* geradas por corrida também é uma das características a serem analisadas, a fim de se obter uma cobertura e profundidade satisfatória do genoma em questão. Assim sendo, a possibilidade de produzir *reads* maiores que 8 kb de tamanho e combiná-las com *reads* curtas de 100 pb, que juntas cobririam o genoma em mais de 80X, com milhões de *reads* geradas, graças às plataformas Illumina e Pacbio, foi possível obter uma boa montagem do genoma da cepa PH8 de *L. amazonensis*. Ademais, a combinação de ambas as estratégias resolve a questão da precisão de bases sequenciadas, uma vez que a alta taxa de erro apresentada pela plataforma Pacbio pode ser reduzida alinhando *reads* curtas Illumina, que ao contrário apresentam maior precisão aliada a uma elevada profundidade de *reads* (MILLER *et al*., 2017). Como exemplo desta vantagem é possível ressaltar a montagem do genoma da cepa RZOD01, a qual foi obtida utilizando apenas *reads* Pacbio e chegou a um nível de contiguidade inferior ao da PH8, com 92 *contigs* (BATRA *et al*., 2019), mostrando que uma etapa de *scaffolding* para o fechamento de *gaps* e posterior correções de erros utilizando *reads* curtas melhora consideravelmente a montagem final. Outras espécies de *Leishmania*, como *L. braziliensis*, *L. donovani* e *L. infantum* tiveram a versão mais recente de seus genomas obtidas a partir da combinação das mesmas plataformas citadas anteriormente e uma melhora significativa também foi observada, tendo sido alcançado nível cromossômico em todas elas (GONZÁLEZ-DE LA FUENTE *et al.*, 2017; GONZÁLEZ-DE LA FUENTE *et al.*, 2019; LYPACZEWSKI *et al*., 2018).

Paralelamente à evolução das técnicas de sequenciamento, os algoritmos de montagem de genomas também sofreram grandes mudanças, devido a criação de novos tipos de arquivos gerados pelas plataformas de sequenciamento e as novas estratégias para a obtenção de genomas completos, como é o caso de abordagens híbridas, que consideram mais de 1 tipo de *read* (GIANI *et al.*, 2020). Para a montagem da PH8, inicialmente foram consideradas apenas *reads* longas Pacbio. Os algoritmos desenvolvidos para lidarem com estas sequências incluem o mapeamento de *reads*

brutas, correção de erro de *reads*, montagem de *reads* corrigidas e polimento de montagem. Os montadores de genoma de *read* longa normalmente são baseados em métodos de sobreposição, como algoritmos de consenso de layout de sobreposição (OLC) (DE LANNOY; DE RIDDER; RISSE, 2017). Ele primeiro gera os alinhamentos entre *reads* longas. Depois disso, ele calcula o melhor gráfico de sobreposição e, a partir do gráfico, é gerada a sequência de consenso dos *contigs*. O processo é então finalizado com a correção de erros utilizando *reads* longas ou curtas e por fim com o polimento da montagem, que tenta corrigir inconsistências de bases ainda existentes. No caso da montagem da PH8, foi ainda executada uma etapa de *scaffoling* usando *reads paired-end* Illumina, onde *contigs* gerados na etapa anterior são unidos em um mesmo *scaffold*, separados apenas por um *gap* de comprimento conhecido. Tal prática também foi adotada na montagem de outros genomas, incluindo *T. brucei*, que considerou também o uso de *reads* Hi-C combinado com *contigs* Pacbio (MÜLLER et al., 2018).

A combinação de todos os fatores descritos anteriormente leva ao sucesso ou fracasso na montagem de um genoma. O genoma de *T. cruzi*, por exemplo, foi obtido pela primeira vez em 2005 a partir de fragmentos gerados pelo método *whole genome shotgun* (WGS) e sequenciamento pelo método de Sanger, com um total de 1.192.680 *mate-pairs reads* de tamanhos médios de 2, 10, 35 e 90 kb. Sua montagem resultou em 5.517 *scaffolds* (8.780 *contigs*), totalizando 67 Mb referente ao seu tamanho haplóide (EL-SAYED *et al.*, 2005). Nessa época, a cepa CL Brener foi escolhida para o sequenciamento devido ao fato de ter sido isolado de *Triatoma infestans*, conhecimento acerca do processo de infecção em camundongos, tropismo para coração e células musculares, fase aguda sintomática em humanos infectados e suscetibilidade ao benzonidazol (revisado por REIS-CUNHA; BARTHOLOMEU, 2019). Análises do conteúdo do genoma desse parasito mostrou que além da grande quantidade de sequências repetitivas, a natureza híbrida desta cepa é uma característica que torna a montagem do genoma ainda mais complexa para os algoritmos montadores, que precisam incluir em seu código, uma etapa de reconstrução de haplótipos, ou seja, que consiga detectar variações de nucleotídeos únicos que distinguem sequências cromossômicas de seus pares homólogos. Essa dificuldade tem impacto na maioria das sequências de genomas depositadas em bancos de dados até o momento, possibilitando muitas vezes a obtenção de apenas uma representação haploide em mosaico. Isso é especialmente relevante para os genomas de *T. cruzi* e *Leishmania*, pois eles são propensos a extensas variações no número de cópias cromossômicas, como descrito neste trabalho e nas demais publicações para essas espécies (EL-SAYED *et al.*, 2005; PATINO *et* al., 2020; ROGERS *et al.*, 2011; VALDIVIA *et al.*, 2011, 2017).

Em 2009, Weatherly e colaboradores utilizaram *contigs* e *scaffolds* de *T. cruzi*, previamente montados, e com base em mapas de sintenia gerados com a espécie de *T. brucei* e em extremidades de sequências de *Bacterial Artificial Chromosome* (BAC), que possuíam sequências

presentes em *T. cruzi*, foi possível obter cromossomos completos e determinar a homologia entre os dois haplótipos para cada cromossomo de *T. cruzi*. Foram montados um total de 41 pares de pseudocromossomos, com tamanho variando de 78kb a 2,4Mb, os quais representavam 90% (2.1133 de 2.3216 genes) dos genes anotados (famílias multigênicas foram retiradas das análises. Apesar destes resultados, a montagem da cepa CL Brener, utilizada como referência, permaneceu fragmentada, pois nenhuma remontagem de *contigs* foi realizada, além de provavelmente conter regiões repetitivas colapsadas e ausência das extremidades dos cromossomos. A ausência de sintenia de determinados agrupamentos de genes com o genoma de *T. brucei* também dificultou a localização de alguns desses genes nos 41 pseudocromossomos de *T.cruzi*, deixando de fora 663 *contigs* anotados com um total de 2083 genes da montagem CL Brener de 2005. Tais dificuldades são atribuídas não somente a características intrínsecas ao genoma do parasito, mas também a estratégia escolhida no processo de montagem e anotação do genoma. Assim como para *T. cruzi*, a ordenação dos *scaffolds* obtidos para a cepa PH8 de *L. amazonensis* também foi realizado usando um genoma de referência, neste caso, o de *L. mexicana*. Essa estratégia foi adotada devido à ausência de um genoma bem montado e anotado da mesma espécie, sendo, portanto, usado o genoma mais próximo evolutivamente. Mas, como descrito anteriormente, essa abordagem possui vantagens e desvantagens, como a facilidade de alocação da maioria dos *scaffolds* dentro do genoma da PH8, determinando inclusive sua orientação, ao mesmo tempo que, devido à ausência de sintenia, não é possível atribuir algumas regiões a nenhum cromossomo, refletindo também na não transferência de anotação de diversos genes e outras *features*.

Com o sucesso na montagem da cepa PH8, diversas características intrínsecas ao genoma puderam ser analisadas a partir dos 34 pseudocromossomos obtidos, características essas que não puderam ser avaliadas nos estudos onde a montagem ainda resultou em um genoma extremamente fragmentado (cepa M2269), como também nos estudos nos quais a montagem foi bem sucedida, mas a anotação do genoma não foi realizada (cepa UA301). Entre as características analisadas nesse trabalho e que apresentam forte impacto nos estudos sobre a infecção por *L. amazonensis*, podemos citar a descrição do repertório completo de genes codificando o antígeno A2, o qual, como descrito a seguir, é uma proteína associada à capacidade de visceralização da infecção pelo parasito.

Virulência representa qualquer redução no *fitness* do hospedeiro após a infecção por um parasita (ABBATE; KADA; LION, 2015). Esses fatores são secretores, associados à membrana ou citosólicos por natureza. Essa definição caracteriza proteínas A2 como um dos fatores de virulência que desperta grande interesse nos estudos de infecções por LV devido a diversos aspectos: ser expressa somente nos estágios intracelulares do parasito; apresentar efeito na redução da virulência *in vivo* em parasitos *knockout* parciais e em parasitos de *L. donovani*, onde a expressão foi inibida por RNA *anti-sense*; aumento da carga parasitária no baço de camundongos infectados com *L.*

major expressando A2; imunidade protetora fornecida experimentalmente à camundongos imunizados com A2 recombinante ou vacina de DNA (Revisado por GARIN *et al.*, 2005). Inicialmente identificados em *L. infantum*, genes que codificam proteína A2 foram descritos como uma família de genes constituída por várias cópias organizadas *em tandem*, além de serem possuírem conteúdo altamente repetitivo (CHAREST; MATLASHEWSKI, 1994). Ao contrário de *L. major*, na qual apenas um pseudogene A2 está presente, com uma ORF de 159 nucleotídeos (ZHANG, W. W. *et al*., 2003), os genes A2 estão completamente ausentes em outras espécies do complexo *L. tropica* (*L. tropica* e *L. aethiopica*) e *L. braziliensis* (GHEDIN *et al*., 1997), ou seja, em espécies de *Leishmania* que não causam LV. No genoma PH8, os resultados apontam para a presença de 9 sequências com homologia a genes de A2, sendo 4 pseudogenes. Além disso, pela primeira vez essas sequências foram caracterizadas quanto a composição de suas sequências, destacando padrões de repetições distintas das observadas em *L. donovani*, *L. infantum* e *L. mexicana* (BATRA *et al*., 2019; GONZÁLEZ-DE LA FUENTE *et al*., 2017; LYPACZEWSKI *et al*., 2018). As diferenças observadas, no entanto, podem sugerir a ocorrência de pressões seletivas ambientais e consequente acúmulo de SNPs nesta região do genoma. Portanto, como sugerido anteriormente, o tropismo de *L. amazonensis* para órgãos viscerais (PORTO *et al*., 2022; TOLEZANO *et al*., 2007) pode estar ocorrendo devido a mutações pontuais (alterações na codificação de aminoácidos), uma vez que grandes eventos em escala cromossômica não foram observados.

A presença do gene A2 na cepa PH8 é consistente com relatos anteriores mostrando *que L. amazonensis* expressa proteínas A2 (CARVALHO *et al*., 2002) e que diferentes isolados foram identificados em cães infectados com manifestações clínicas de LV e LT (VALDIVIA *et al*., 2017). A não identificação destes genes na montagem da cepa 2269, única com a anotação disponível, provavelmente se deve a alta fragmentação de sua sequência, contudo, o pipeline adaptado ao presente trabalho pode ajudar na identificação desta família, ainda que seja encontrado apenas regiões parciais dos genes de A2 nesta cepa. Mas, diante da identificação dessa família em *L amazonensis* e em conjunto com o protocolo de edição de genes previamente desenvolvido pelo nosso grupo (GOES *et al*., submetido), o papel desses genes na virulência e capacidade de visceralização em infecções causadas por essa espécie poderá ser melhor investigado.

A anotação do genoma nuclear da PH8 resultou em 190 novos genes em comparação à anotação atualmente disponível para a cepa M2269 (REAL *et al*., 2013). Dentre os genes que codificam proteínas com função conhecida está o EF1-alpha e uma KAP, descrita por ser uma proteína associada ao cinetoplasto. O primeiro tem sido indicado como um importante fator de virulência, principalmente em espécies relacionadas a leishmaniose visceral como *L. donovani, L. infantum* e *L. chagasi* (NANDAN *et* al., 2002; TIMM *et al*., 2017; LEMOS-SILVA, TELLERIA,

TRAUB-CSEKÖ, 2021). Foi visto que essa proteína está relacionada à ligação e ativação da tirosina fosfatase 1 (SHP-1) em macrófagos, atuando como repressor das vias Toll e Jak/STAT, imunossuprimindo o hospedeiro (NANDAN *et* al., 2002). Além disso, análise da expressão gênica em infecções de *Lutzomya longiplapis* mostrou que esse gene está aumentado nas primeiras 6h pós-infecção, apontando um papel crucial para o estabelecimento da infecção no hospedeiro invertebrado (LEMOS-SILVA, TELLERIA, TRAUB-CSEKÖ, 2021). Já as KAPs tem sido associadas ao correto empacotamento do kDNA, sendo semelhantes a histonas H1 (XU *et al.*, 1996; XU, RAY, 1993). Inicialmente identificadas em *C. fasciculata* também têm demonstrado possuir um papel importante para o crescimento celular, respiração e transcrição mitocondrial, bem como replicação e/ou segregação do kDNA, demonstrado por meio de *knockouts* desses genes nessa espécie e em *T. brucei* (AVLIYAKULOV, LUKES, RAY, 2004; BECK *et al.*, 2013).

Como citado anteriormente, um dos genes também de grande importância para o estudo da infecção por *L. amazonensis* e que não havia sido identificado no estudo sobre o genoma de nenhuma das cepas incluindo a M2269 é o gene de A2. Além da família gênica de A2, genes que constituem duas outras importantes famílias multigênicas que atuam no processo de virulência da *Leishmania*, a saber, as amastinas e as metaloproteases GP63 podem ser agora melhor estudadas. Além disso, foram identificados os repertórios de genes codificadores de proteínas cisteíno proteases, cinases e fosafatases e ainda do conjunto de genes codificadores de proteínas envolvidas no metabolismo de ferro e heme.

A identificação de 11 novos genes de amastinas, quando comparada à anotação dessa família gênica no genoma da cepa M2269, representa um avanço na anotação dessa família que codifica esse fator de virulência e mostra a importância de se ter genomas mais completos. Essa melhoria permitiu uma reanálise evolutiva desta família em *L. amazonensis* e a ampliação do conhecimento acerca da organização das subfamílias nesta espécie. Semelhante à organização da família de genes de amastinas encontrada nos genomas de *T. cruzi*, *L. infantum* e *L. braziliensis* (ARAÚJO, P. R.; TEIXEIRA, 2011; DE PAIVA *et al.*, 2015; ROCHETTE *et al.*, 2005), apenas δ-amastinas foram encontradas associadas aos ortólogos do gene tuzina, como previsto por Jackson (2010). Além disso, foi vista uma divergência evolutiva dentro da subfamília β-amastinas, dando origem a 2 isoformas distintas. Diante dessas constatações, foi visto que, mais uma vez, em comparação com análises anteriores, a descoberta de novas cópias acrescentou evidências evolutivas, indicando que essa família continua sofrendo pressões seletivas e acumulando mutações ao longo dos anos. Ao comparar o repertório de genes de amastinas com outras espécies de *Leishmania*, *T. cruzi* e *T. brucei* e ainda, utilizando genes de *Bodo saltans* como grupo externo, maiores evidências evolutivas foram obtidas. Os resultados aqui apresentados corroboram estudos anteriores que α, γ e δ -amastinas representam grupos monofiléticos, enquanto β-amastinas se

dividem em duas linhagens intimamente relacionadas correspondendo às distintas isoformas já observadas. Outra observação é o fato de γ -amastinas não serem encontradas no gênero *Trypanosoma*, indicando que essa classe de amastinas existiu nesse gênero, mas foi perdida em algum ponto da evolução, fato que é reforçado pela posição entre as α e β-amastinas. Além das inferências filogenéticas realizadas com base nas sequências de amastinas, com o presente estudo descrevendo o repertório completo desses genes na cepa PH8 e com os novos protocolos de edição gênica por CRISPR/Cas9 que permitem a deleção de genes presentes em múltiplas cópias no genoma abre-se a também uma nova perspectiva para estudos sobre o papel das amastinas em *L. amazonensis*.

A terceira família analisada corresponde as metaloproteases GP63. Foram identificados 11 genes pertencentes a esta família no genoma da PH8, dos quais 2 são pseudogenes. As sequências foram caracterizadas quanto aos domínios estruturais e funcionais existentes: HEXXHA, SRYD, KDELMAP, corroborando com o que já está disponível na literatura (PUENTES *et al.*, 1999; CASTRO NETO *et al*., 2019). Além dos domínios destacados já foi relatado que estas proteínas possuem outras importantes regiões conservadas, as quais participam de processos de internalização por parte das células hospedeiras, como por exemplo, os trechos PAVGNIPA e KAREQYGC que aparecem conservados em todas as sequências da PH8 (MEDINA *et al*., 2016). Essa característica reforça o papel dessas proteínas em processos de interação parasito-hospedeiro, mas não exclui a necessidade de validação experimental. Além disso, a localização de domínios importantes para a internalização do parasito, em regiões de folha-beta e externamente na superfície da proteína, coincidindo com regiões imunogênicas, evidencia seu papel contra a resposta imune do hospedeiro (CASTRO-NETO *et al*., 2019). Estruturas folha-beta são mais propensas ao acúmulo de mutações que resultam em variações estruturais (ABRUSÁN; MARSH, 2016) e, portanto, atuam como abrigos para domínios funcionalmente importantes para a sobrevivência do parasito.

As análises estruturais de genes que codificam GP63 em *L. amazonensis* contribuem para aprofundar o conhecimento dos processos de evasão da resposta imune, uma vez que apresentam sequências N-terminais bastante variáveis (CASTRO-NETO *et al.*, 2019). Essas proteínas sofrem modificações pós-traducionais, nas quais perdem parte da região N-terminal, o que acaba contribuindo para sua variabilidade e dificulta sua representação em estruturas tridimensionais, como foi observado no modelo proposto para uma GP63 da cepa PH8. Foi visto também que de forma geral GP63 têm estruturas variáveis ao longo de toda a proteína. Ao se comparar dezenas de sequências de *Trypanosoma* e *Leishmania* descobriu-se que tais estruturas estão associadas à variabilidade do sítio de ligação ao zinco e supostamente a sua atividade (MA *et al*., 2011). Essa ligação entre diferentes estruturas e variabilidade da sequência também foi vista mais recentemente

em GP63 de *L. braziliensis*, confirmando o impacto funcional durante a ligação do substrato e consequentemente afetando a interação parasito-hospedeiro (SUTTER *et al*., 2017).

Neste trabalho também foi possível realizar o levantamento completo do repertório gênico de cisteíno proteases presente no genoma de *L. amazonensis*. Dessa forma destacou-se a existência de 76 genes que são classificados em 12 diferentes famílias. Uma nova contribuição dada aqui é o fato de que, ao contrário do que foi proposto por Silva Almeida *et al*. (2014), genes de cisteíno proteases também estão presentes no cromossomo 7, não sendo exclusivo de *L. mexicana*. Outra constatação é a distribuição não uniforme ao longo do genoma, ou seja, a abundância de cisteíno proteases varia bastante entre os diferentes cromossomos, estando concentrados principalmente nos cromossomos 8 (6 genes), 20 (20 genes) e 30 (10 genes). Esses resultados coincidem com o fato de suas localizações em cromossomos dissômicos e supranumerários, como é o caso de cromossomo 30. Tais achados, no entanto, são explicados pela organização dos genes em repetições *em tandem*, o que permite que os parasitas gerem rapidamente um grande número de transcritos que podem ser necessários em grandes quantidades (VICTOIR; DUJARDIN, 2002). Ainda, Rogers e colabores (2011) sugerem que *Leishmania* spp. pode ter uma estratégia para aumentar os níveis de mRNA duplicando genes em cromossomos dissômicos ou formando cromossomos supranumerários.

A diversidade de genes de cisteíno proteases observada na análise reforça a ideia de que esta classe de enzima é crucial para o ciclo de vida do parasita. Além disso, deixa claro sua importância como ferramentas de sobrevivência e adaptação e, consequentemente, como alvo importante em estratégias de vacinação e terapia. A maioria das cisteíno proteases de *L. amazonensis* pertencem ao Clan A, famílias C1 (CPA, CPB e CPC); C2; a qual inclui as proteínas semelhantes a calpaína, papaína e catepsina; e C19. Diversos estudos têm demonstrado o papel de enzimas CPA e CPB na infectividade do parasito, sobretudo em processos relacionados a interação parasito-hospedeiro. Promastigotas de *L. infantum* deficientes em CPA, por exemplo, provocaram uma diminuição significativa da virulência *in vitro* e *in vivo*, apesar de não apresentarem mudanças na replicação (DENISE *et al*., 2006). De forma análoga, parasitos duplo-*knockout* para genes CPB de *L. mexicana* mostraram redução de 80% na capacidade de infecção de macrófagos e retardo da lesão provocada em camundongos (MOTTRAM *et al*., 1996). Em outro estudo, onde o uso do inibidor de calpaína MDL28170 foi empregado também foi visto um importante papel nos estágios iniciais da infecção de macrófagos de mamíferos por *L. braziliensis* (ENNES-VIDAL *et al.*, 2019). Por fim, recentemente foi visto que cisteína peptidases deubiquitinantes, o que inclui as proteases específicas de ubiquitina (USPs, família C19) afetam a diferenciação de promastigotas metacíclicos em amastigotas, com significante perda do *fitness* durante a diferenciação, além de afetar também a infecção intracelular, sendo essenciais para a viabilidade de promastigotas (DAMIANOU *et al*., 2020).

Genes que codificam proteínas relacionadas ao metabolismo de ferro e de heme, apesar de não se caracterizarem como uma família multigênica, também constituem importantes fatores de virulência. Ferro e heme são essenciais em muitas vias metabólicas conservadas, incluindo transporte de elétrons, síntese e detecção de gás e transdução de sinal, mas apesar disso, quando presentes em altas concentrações causam efeitos citotóxicos para o parasito, devido a geração de radicais livres (LARANJEIRA-SILVA *et al*., 2020). Dada a importância dessas proteínas, a falta de genes que as codificam provoca efeitos danosos a leishmania. O transportador de membrana LFLVCRb por exemplo, possui um papel essencial na sobrevivência, replicação e virulência do parasito, evidenciado por experimentos simples e duplos-*knockouts* mediados por CRISPR/Cas9. Da mesma forma, leishmanias *knockouts* para o gene LIR1, apresentam um prejuízo significativo no efluxo de ferro e virulência, além de sensibilidade aumentada à toxicidade, demonstrando um papel crítico de LIR1 na manutenção da homeostase do ferro em *Leishmania* (LARANJEIRA-SILVA *et al.*, 2020).

Como o primeiro transportador de ferro ferroso putativo identificado em parasitas tripanosomatídeos, o LIT1 codifica uma proteína de 432 aminoácidos prevista para conter oito domínios transmembranares (HUYNH; SACKS; ANDREWS, 2006). Na PH8 notou-se que este gene está localizado no cromossomo 30 que é tetrassômico nesta cepa, sugerindo a existência de cópias extras que favorecem o *fitness* de *L. amazonensis*. Esse fato acaba sendo uma vantagem evolutiva adquirida pela leishmania, não sendo encontrados ortólogos do gene LIT em *T. brucei*, que não possui estágio intracelular ou em *T. cruzi*, que cresce como amastigotas intracelulares no citosol celular. Além disso, nem *T. cruzi* nem o genoma de *T. brucei* codificam homólogos dos transportadores de heme de Leishmania LFLVCRb ou do receptor de hemoglobina (HbR). Em contraste, ortólogos do LMIT1, que codifica um transportador de ferro para as mitocôndrias do parasita, e do LFR1, que codifica uma redutase férrica 1, estão presentes em todas as espécies de *Leishmania*, bem como em *T. cruzi* e *T. brucei*. Já os genes que codificam um homólogo do exportador de ferro LIR1, bem como a peroxidase ascorbato (APX), são encontrados em *Leishmania* spp e *T. cruzi*, mas não em *T. brucei*.

Também presentes no genoma de todos os parasitas tripanosomatídeos, superóxido dismutases ferro-dependentes (SODs) foram caracterizados como fatores de virulência em várias espécies de *Leishmania* e em *Tripanosomas*. A genética reversa tem sido usada para abordar os papéis de SODA de *L. amazonensis*, a isoforma mitocondrial SODB e o importador de ferro mitocondrial LMIT1. Tentativas de gerar mutantes nulos SODA, que não eram viáveis e o fenótipo quanto à capacidade de infecção de promastigotas metacíclicos sem um alelo SODA destacou o papel essencial em *L. amazonensis*. Mutantes parasitários com um alelo SODA deletado e, em consequência, com expressão reduzida de SODA, não se replicaram em macrófagos e foram

severamente atenuados em sua capacidade de gerar lesões cutâneas em camundongos (MITTRA *et al*., 2017). Da mesma forma, mutantes nulos de LMIT1 não são viáveis, consistente com a importância do ferro para a montagem de proteínas do *cluster* Fe-S (MITTRA *et al*., 2016). A caracterização do conjunto completo de genes envolvidos com ferro e metabolismo de heme em *L. amazonensis* aqui apresentado é um passo importante para um estudo funcional completo que preencherá uma das muitas lacunas em nossa compreensão da infecção por leishmania e poderá revelar novas estratégias para o controle da doença.

Dentre outras características intrínsecas ao genoma de *L. amazonensis* que puderam ser analisadas a partir dos 34 pseudocromossomos obtidos, estão as análises de sequências repetidas. Em geral, os genomas do gênero *Leishmania* são constituídos por sequências repetidas que variam entre 0,4 e 1 kb de tamanho. Essas sequências são encontradas amplamente distribuídas por todo o genoma, de forma cromossomo-específica (REQUENA *et al*., 2017; UBEDA *et al*., 2014). O conteúdo completo dessas sequências e sua organização só podem, entretanto, serem avaliados adequadamente quando se tem uma montagem completa. Por esse motivo, as *reads* longas Pacbio foram geradas também com o objetivo de resolver tais regiões no menor número possível de *contigs*, uma vez que porções altamente repetitivas do genoma podem ser sequenciadas em uma única *read*, como consequência da maior cobertura proporcionada por elas, podendo chegar a um tamanho maior que 25 kb (EID *et al*., 2009). Essa foi a primeira vez que o conteúdo repetitivo de *L. amazonensis* foi descrito, o qual representa 9,57% do genoma da PH8, corroborando com o observado para os genomas de *L. braziliensis* e *L. infantum* (PEACOCK *et al*., 2007). Desse total, retroelementos e transposons de DNA correspondem a aproximadamente 27% dos elementos repetitivos. Sendo assim e levando em consideração o papel que os mesmos tem na criação, eliminação ou modificação de genes, remodelando a estrutura e a função de seus genomas hospedeiros (WICKSTEAD; ERSFELD; GULL, 2003) é possível assumir que *L. amazonensis* esteja sofrendo constantes recombinações gênicas e assumindo sua própria trajetória na evolução dentro do gênero *Leishmania*. Contudo, uma análise minuciosa ainda é necessária para determinar o clado nos quais estes retroelementos estão inseridos, bem como se são elementos ativos ou apenas retroposons remanescentes.

Os resultados publicados até o momento indicaram a falta de retroelementos ativos em *Leishmania* (BHATTACHARYA, S.; BAKRE; BHATTACHARYA, 2002). Em *L. major* e *L. infantum* foram descritos apenas retroelementos degenerados como DIREs (degenerate ingi/L1Tc-related elements) e ingi/L1Tc (BRINGAUD *et al*., 2006; GHEDIN *et al*., 2004). Entretanto, as análises da sequência genômica de *L. braziliensis* permitiu a detecção de retroposons ativos (PEACOCK *et al*., 2007) integrados ao conjunto de genes que codificam o *spliced-leader* (AKSOY *et al*., 1987; VILLANUEVA *et al*., 1991). Na cepa PH8 de *L. amazonensis*, estas repetições podem

estar ainda associadas a regulação da expressão de diversos genes, uma vez que a abundância de retroposons extintos também pode ser atribuída à sua capacidade de participar de eventos recombinacionais que levam à amplificação genética (UBEDA *et al.*, 2014).

Repetições teloméricas presentes nas extremidades físicas dos cromossomos desempenham um importante papel na manutenção do genoma e progressão do ciclo celular. O DNA telomérico é formado por uma região de fita dupla (rica em citosina e guanina) e uma fita simples, rica em G que se projeta em direção às extremidades do cromossomo. (BLACKBURN, 1990). Tais características podem ter contribuído para uma montagem falha destas regiões e por isso a montagem da PH8 não é tida como completa, sendo apresentados pseudocromossomos de *L. amazonensis*. A sequência canônica 5' TTAGGG 3' presente nos cromossomos do gênero *Leishmania* foi detectada em apenas 6 extremidades dos *scaffolds* da PH8, além de outros 2 *scaffolds* (scaffold313 e scaffold313) não incorporados à montagem ricos em C, os quais podem ser provenientes dessas regiões. Sequências de N's também foram inseridas em regiões próximas aos telômeros e subtelômeros, reafirmando a dificuldade encontrada na montagem destas regiões. No entanto, essa dificuldade também já foi relatada para outros genomas, como por exemplo, na montagem da cepa HU3 de *L. donovani* (CAMACHO *et al*., 2019b) e para outras cepas de *L. amazonensis* (REAL *et al*., 2013; BATRA *et al*., 2019, TSCHOEKE *et al*., 2014), onde essas extremidades não podem ser encontradas.

Vários estudos mostraram que aneuploidias são altamente frequentes em diferentes espécies de *Leishmania*, assim como visto na cepa PH8 de *L. amazonensis*, e também nas demais cepas desta espécie (BATRA *et al*., 2019; PATINO *et al*., 2020). Nove cromossomos poliplóides foram descritos em *L. infantum*, quatro em *L. mexicana*, dois em *L. braziliensis* e um em *L. major* (ROGERS *et al*., 2011). Um estudo recente em que foram analisados os genomas de dois isolados de *L. amazonensis* obtidos de cães com manifestações clínicas de doença visceral, mostrou que a maioria dos cromossomos tem uma cobertura de *reads* com profundidade compatível com o número haploide de cada cromossomo, com exceção do cromossomo 30 (VALDIVIA *et al*., 2017), assim como observado em nosso estudo. Somada a isso, a variação do número de cópias (CNV) gênicas é outra característica associada ao genoma da *Leishmania* (IANTORNO *et al*., 2017; KRAMER, 2012). Esse fenômeno está associado à rápida mudança na expressão de certos fatores de virulência em resposta a mudanças no ambiente. A duplicação gênica tem sido proposta como mecanismo para aumentar a expressão gênica, uma vez que o controle da regulação transcricional está ausente em tripanosomatídeos (IVENS; BLACKWELL, 1996; PEACOCK *et al*., 2007; ROGERS *et al*., 2011). Nesse sentido, foi visto que diversos fatores de virulência, como as amastinas, GP63, fosfatases e cinases estão entre os genes que sofreram expansão, reforçando o seu papel na virulência do parasito e a sua contribuição para o aparecimento dos diversos fenótipos causados pela *L. amazonensis*.

O sequenciamento de *reads* longas possibilitou além da obtenção do genoma nuclear, a retenção de parte do genoma mitocondrial, correspondente ao maxicírculo da *L. amazonensis*. Esta é, portanto, a primeira descrição completa para esta molécula nessa espécie de *Leishmania* e pode agora contribuir para estudos evolutivos baseados no genoma mitocondrial. Até 2008, havia sido descrito apenas um fragmento de 8,4 kb referente à região codificadora do maxicírculo de *L. major* (YATAWARA *et al*., 2008) e outro de 17.028 pb de *L. donovani*, que representa quase toda a região codificadora do maxícírculo desta espécie. Curiosamente, neste último, foi observada a inserção de um minicírculo de 777 pb de comprimento total na região 3' do gene codificador de ND1. No entanto, essa característica não foi vista na anotação do maxicírculo da cepa PH8. Nos anos seguintes foram também publicadas as sequências completas dos maxícirculos de *L. braziliensis*, *L. major*, *L. infantum* e o conjunto completo de mRNAs mitocondriais editados em pan de *L. amazonensis* (cepa LV78), mas não a montagem desta última (CAMACHO; RASTROJO; *et al*., 2019; MASLOV, DMITRI A., 2010). No ano de 2022, uma coleção de regiões codificantes de maxircirculos de 26 linhagens prototípicas de espécies de tripanosomatídeos foram publicadas, incluindo *L. amazonensis*, mas não *L. mexicana*, a qual continua sem ter sido caracterizada. Apesar disso, a sequência do maxicirculo previamente disponibilizada, não foi anotada ou analisada quanto aos elementos nela contidos. Por fim, todas as espécies cuja o genoma mitocondrial já foi montado mostram que apesar da divergência evolutiva, no caso de *L. braziliensis* em relação a outras espécies patogênicas de *Leishmania*, uma notável conservação no ordenamento dos genes do maxicírculo (sintenia) foi mantida, corroborando com os resultados aqui descritos.

A ausência dos criptogenes ND8 (G1), ND9 (G2), G3, G4, ND3 (G5) editados observada aqui está de acordo com o que foi relatado no conjunto de mRNAs descritos por MASLOV *et al*. (2010) para *L. amazonensis*. A incapacidade de transcrever criptogenes não editados provocada pela presença das regiões ricas em G também foi vista em *L. tarentolae* e *L. donovani* (MASLOV, DMITRI A.; THIEMANN; SIMPSON, 1994; THIEMANN; MASLOV; SIMPSON, 1994), mas ausente nas demais espécies de *Leishmania*. Essa característica, no entanto, é comum em culturas que foram mantidas em laboratório por muito tempo, indicando que os minicírculos que codificam os gRNAs necessários para a edição desses criptogenes foram perdidos (SIMPSON, LARRY *et al*., 2000). Além disso, ao contrário de *L. infantum*, *L. donovani*, *L. braziliensis* e *L. major*, o maxicírculo de *L. amazonensis* possui um gene ortólogo da subunidade 6 da ATPase, denominado MURF4, que já foi descrito em *T. brucei* e *L. tarentolae*. Os transcritos deste gene são extensivamente editados na extremidade 5' através da adição e deleção de numerosas uridinas, criando potenciais códons de início e uma estrutura de leitura aberta contínua (BHAT; MYLER; STUART, 1991). Da mesma forma, o gene MURF1, presente em *L. amazonensis*, *L. tarentolae* e *L. donovani* é ortólogo do gene ND2 descrito nas demais espécies de *Leishmania*. Além da região

codificante (RC), o maxicírculo possui uma região não codificante e altamente variável entre as espécies de Leishmania, denominada região divergente (RD). Inicialmente seu papel foi atrelado exclusivamente ao controle da expressão gênica, abrigando promotores para o gene de rRNA12S e sequências essenciais para a replicação do minicírculo (CSB-I, -II, -III). Apesar de não estar ainda claro o papel das sequencias RD, a identificação de sítios de ligação da topoisomerase II e transcrito de diferentes tamanhos sugerem que esse papel possa ser ainda mais complexo (HORVATH *et* al, 1990; MYLER et al., 1993).

Com o conhecimento mais profundo acerca do conteúdo e organização gênica de famílias multicópias torna-se mais fácil sua manipulação utilizando ferramentas de edição de genomas, como é o caso das estratégias descritas previamente por colaboradores deste trabalho (GOES *et al.*, submetido), utilizando o sistema CRISPR-Cas9 do tipo II. Para estabelecer um protocolo eficiente de edição do genoma de *L. amazonensis* foram testadas duas estratégias para gerar mutantes nocaute do transportador de miltefosina (MT), o qual foi usado como prova de conceito. A primeira estratégia consiste na transfecção de promastigotas de *L. amazonensis* com plasmídeo pLDCN contendo a região codificadora da nuclease Cas9 de *S. pyogenes* (SpCas9) e com sgRNA transcrito *in vitro* contendo sequências do gene alvo e um fragmento de DNA que serviu como sequência doadora (stop códons + sítio de restrição) para reparo de HR. Sendo assim, os resultados mostraram que após a transfecção e digestão dos amplicons, os fragmentos de DNA previamente amplificados e submetidos à eletroforese em gel de agarose, confirmaram a ruptura do gene MT na população de parasitas transfectados em comparação a parasitas *wt*, que não tiveram seu DNA digerido. Na segunda estratégia, promastigotas de *L. amazonensis* foram transfectadas com um complexo de ribonucleoproteína (RNP) composto por Cas9 recombinante de *S. aureus* complexada com sgRNA transcrito *in vitro*, o qual também contém sequências do gene alvo e uma sequência de DNA doadora (stop códons + sítio de restrição). Sendo assim, o gene MT de *L. amazonensis* com o sítio de restrição incorporado, permitiu a digestão de produtos de PCR amplificados de DNA extraído de parasitas transfectados, gerando 3 fragmentos, enquanto a digestão de produtos de PCR de DNA extraído de parasitas wt gerou 2 fragmentos, uma vez que *L. amazonensis* já possui um sítio de restrição para a enzima utilizada no experimento. Então, juntos, esses resultados demonstraram que mutantes de *L. amazonensis* com um único gene interrompido podem ser gerados alguns dias após a transfecção, usando duas estratégias distintas baseadas na atividade de Cas9.

Dado que não há ainda na literatura relatos de *knockouts* gerados com essa ferramenta para *L. amazonensis*, e ainda que as espécies de *Leishmania* apresentam significativa variabilidade em seu genoma (LYPACZEWSKI *et al*., 2018; SAMARASINGHE *et al*., 2018) esse trabalho mostra-se importante como forma de estabelecer protocolos funcionais, possibilitando a realização de outras edições em outros alvos gênicos futuramente, principalmente genes codificadores de fatores

de virulência. Além disso, essas novas estratégias poderão auxiliar no entendimento dos mecanismos de resistência à diferentes fármacos apresentados por espécies de *Leishmania*, que já mostraram variação considerável na sensibilidade a estes compostos. Tais estudos tem impactos para a abordagem terapêutica e descoberta de novos fármacos, entre outros aspectos (CROFT; SUNDAR; FAIRLAMB, 2006).

Juntamente com esses protocolos de edição de genoma altamente eficientes, a disponibilidade de um genoma totalmente sequenciado e anotado abre infinitas possibilidades para estudos mais profundos sobre genes de *L. amazonensis*, principalmente quando famílias multigênicas são os genes de interesse. Além dos métodos de deleção gênica, o estudo da função gênica, utilizando outras estratégias de manipulação genética, como a geração de parasitas que super expressam uma proteína específica ou parasitas nos quais são criados genes repórteres ou marcados, certamente contribuirá para um melhor entendimento dos mecanismos moleculares envolvidos na interação parasita-hospedeiro e o estabelecimento da infecção. Sendo um parasita incomum e ainda pouco conhecido, que pode causar quase todos os tipos de manifestação da leishmaniose, o trabalho aqui apresentado abre novas portas para estudos tão necessários, que podem levar a métodos aprimorados de controle e eliminação da leishmaniose como um grave problema de saúde pública.

7. CONCLUSÕES

Neste trabalho, apresentamos a montagem dos genomas nuclear e mitocondrial da cepa PH8 de *Leishmania amazonensis* obtida a partir de uma abordagem híbrida que combina *reads* de sequenciamento longo e curto, bem como sintenia com o genoma de *Leishmania mexicana*. Além disso, fornecemos uma anotação que resultou em um número adicional significativo de genes anotados em comparação com o único *L. amazonensis* (cepa M2269) que está disponível no banco de dados TritrypDB. A análise do conteúdo gênico, bem como a variação no número de cópias cromossômicas, permitiu confirmar a natureza diplóide de *L. amazonensis* e também a existência de cópias extras, para alguns de seus cromossomos. Uma montagem mais completa de uma nova cepa de *L. amazonensis* permitiu uma melhor análise da organização gênica, principalmente em relação aos genes que codificam fatores de virulência que estão presentes como famílias gênicas multicópias. Essas análises, nas quais foram incluídos os dados de genomas de outras cepas de *L. amazonensis* bem como de outras espécies de *Leishmania* permitem aprofundar o conhecimento sobre a grande diversidade na população desse parasito e os impactos dessa diversidade sobre as manifestações clínicas na leishmaniose.

8. PERSPECTIVAS

Os resultados gerados neste trabalho contribuem para ampliar o conhecimento acerca do papel de diversos fatores de virulência na progressão da leishmaniose. Tendo isso em vista e somado as ferramentas de edição gênica previamente desenvolvidas, novos desafios são propostos e estão listados abaixo:

- Gerar e analisar sequências a partir de mRNA extraído de amastigotas e promastigotas por RNA-seq com o objetivo de avaliar o perfil de expressão gênica global, incluindo as famílias multigênicas A2, amastina, GP3 e cisteíno proteases ao longo do ciclo de vida do parasito;
- Obter linhagens *knockouts* usando o sistema CRISPR-Cas9 para estudar o papel de A2, amastinas e de outras famílias multigênicas;
- Expandir as análises funcionais para os demais fatores de virulência não caracterizadas até o momento, como as cinases, fosfatases e proteínas do metabolismo de ferro, que também já foram descritos por possuírem papel importante durante a infeção pela *Leishmania*.

9. REFERÊNCIAS BIBLIOGRÁFICAS

ABASCAL, F.; ZARDOYA, R.; POSADA, D. ProtTest: selection of best-fit models of protein evolution. **Bioinformatics (Oxford, England)**, v. 21, n. 9, p. 2104–2105, 1 maio 2005.

ABBATE, J. L.; KADA, S.; LION S. Beyond Mortality: Sterility As a Neglected Component of Parasite Virulence. **PLoS Pathog**, v. 11, n. 12, e1005229, 2015.

ABRUSÁN, G.; MARSH, J. A. Alpha Helices Are More Robust to Mutations than Beta Strands. **PLoS Computational Biology**, v. 12, n. 12, 1 dez. 2016.

AFONSO, L. C. C.; SCOTT P. Respostas imunes associadas à suscetibilidade de camundongos C57Bl/10 a *Leishmania amazonensis* **Infection and Immunity**, v. 61, p. 2952-2959, 1993.

AGARWAL, S. *et al*. Clathrin-mediated hemoglobin endocytosis is essential for survival of *Leishmania*. **Biochimica et Biophysica Acta (BBA) - Molecular Cell Research**, v. 1833, n. 5, p. 1065–1077, 1 maio 2013.

AJAY UMMAT AND ALI BASHIR. Resolving complex tandem repeats with long reads. **BIOINFORMATICS**, v. 30, n. 24, p. 3491–3498, 2014.

AKHOUNDI, M. *et al*. A Historical Overview of the Classification, Evolution, and Dispersion of *Leishmania* Parasites and Sandflies. **PLoS Neglected Tropical Diseases**, v. 10, n. 13, 2016.

AKSOY, S. *et al*. Multiple copies of a retroposon interrupt spliced leader RNA genes in the African

trypanosome, *Trypanosoma gambiense*. **The EMBO journal**, v. 6, n. 12, p. 3819–3826, 1987.

ALBERT DESCOTEAUX; SALVATORE J. TURCO. Glycoconjugates in *Leishmania* infectivity. **Biochimica et Biophysica**, p. 341–357, 1999.

ALVES-FERREIRA, E. V. C. *et al*. Differential Gene Expression and Infection Profiles of Cutaneous and Mucosal *Leishmania braziliensis* Isolates from the Same Patient. **PLOS Neglected Tropical Diseases**, v. 9, n. 9, p. e0004018, 14 set. 2015.

ALVES SOUZA, N. *et al*. Detection of mixed *Leishmania* infections in dogs from an endemic area in southeastern Brazil. **Acta Tropica**, v. 193, p. 12–17, 2019.

ANDREWS, S. **FastQC: a quality control tool for high throughput sequence data**. Disponível em: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>. Acesso em: 3 mar. 2020.

AQUINO, G. P. DE *et al*. Lipid and fatty acid metabolism in trypanosomatids. **Microbial Cell**, v. 8, n. 11, p. 262, 1 nov. 2021.

ARAÚJO, M. S. S. *et al*. Despite Leishvaccine and Leishmune trigger distinct immune profiles, their ability to activate phagocytes and CD8+ T-cells support their high-quality immunogenic potential against canine visceral leishmaniasis. **Vaccine**, v. 26, n. 18, p. 2211–2224, 24 abr. 2008.

ARAÚJO, P. R.; TEIXEIRA, S. M. Regulatory elements involved in the post-transcriptional control of stage-specific gene expression in *trypanosoma cruzi* - A review. **Memorias do Instituto Oswaldo Cruz**, v. 106, n. 3, p. 257–266, 2011.

ARDUI, S. *et al*. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. **Nucleic Acids Research**, v. 46, n. 5, p. 2159–2168, 2018.

ARENAS, M. Trends in substitution models of molecular evolution. **Frontiers in Genetics**, v. 6, n. OCT, p. 319, 2015.

ASLETT, M. *et al*. TriTrypDB: a functional genomic resource for the Trypanosomatidae. **Nucleic acids research**, v. 38, n. Database issue, p. D457-62, jan. 2010.

ASSEFA, S. *et al*. ABACAS: Algorithm-based automatic contiguation of assembled sequences. **Bioinformatics**, v. 25, n. 15, p. 1968–1969, 2009.

AVLIYAKULOV, N. K.; LUKES, J.; RAY, D. S. Mitochondrial histone-like DNA-binding proteins are essential for normal cell growth and mitochondrial function in *Crithidia fasciculata*. **Eukaryotic Cell**, v. 3, p. 518–526, 2004.

BARBOSA, F. M. C. *et al*. Extracellular Vesicles Released by *Leishmania (Leishmania) amazonensis* Promote Disease Progression and Induce the Production of Different Cytokines in Macrophages and B-1 Cells. **Frontiers in Microbiology**, v. 9, 21 dez. 2018.

BARRAL, A. *et al*. Leishmaniose na Bahia, Brasil: evidências de que a *Leishmania amazonensis* produz um amplo espectro de doenças clínicas. **The American Journal of Tropical Medicine and Hygiene,** v. 44, n. 5, p. 536-546, 1991.

BARRAL-NETO *et al*. Cytotoxicity in human mucosal and cutaneous leishmaniasis. **Parasite Imunnology**, v. 17, n; 1, p. 21-28, jan. 1995.

BARTHOLOMEU, D. C. et al. Genomics and functional genomics in *Leishmania* and *Trypanosoma cruzi*: statuses, challenges and perspectives. **Memórias do Instituto Oswaldo Cruz**, v. 116, n. 1, p. 1–21, 29 mar. 2021.

BASTIEN, P. *et al*. The complete chromosomal organization of the reference strain of the *Leishmania* genome project, *L. major* "Friedlin". **Parasitology Today**, v. 14, n. 8, p. 301–303, ago. 1998.

BASTIEN, P.; BLAINEAU, C.; PAGES, M. Leishmania: Sex, lies and karyotype. **Parasitology Today**, v. 8, n. 5, p. 174–177, 1 maio 1992.

BATES, P. A. Revising Leishmania's life cycle. **Nature Microbiology**, v. 3, n. 5, p. 529–530, 1 maio 2018.

BATRA, D. *et al*. Draft Genome Sequences of *Leishmania (Leishmania) amazonensis*, *Leishmania (Leishmania) mexicana*, and *Leishmania (Leishmania) aethiopica*, Potential Etiological Agents of Diffuse Cutaneous Leishmaniasis. **Microbiology Resource Announcements**, v. 8, n. 20, 16 maio 2019.

BECK, K. *Trypanosoma brucei* Tb927.2.6100 is an essential protein associated with kinetoplast DNA. **Eukaryot Cell**. v. 12, p. 970–978, 2013.

BELEW, T. et al. Comparative transcriptome profiling of virulent and non-virulent Trypanosoma cruzi underlines the role of surface proteins during infection. **PLoS Pathogens**, v. 13, n. 12, p. 1–23, 2017.

BELTON, J. M. et al. Hi-C: a comprehensive technique to capture the conformation of genomes. **Methods (San Diego, Calif.)**, v. 58, n. 3, p. 268–276, nov. 2012.

BENEKE, T. *et al*. A CRISPR Cas9 high-throughput genome editing toolkit for kinetoplastids. **Royal Society Open Science**, v. 4, n. 5, p. 1–16, 1 maio 2017.

BENTLEY, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. **Nature 2008 456:7218**, v. 456, n. 7218, p. 53–59, 6 nov. 2008.

BHAT, G. J.; MYLER, P. J.; STUART, K. The two ATPase 6 mRNAs of *Leishmania tarentolae* differ at their 3' ends. **Molecular and biochemical parasitology**, v. 48, n. 2, p. 139–149, 1991.

BHATTACHARYA, P.; ALI, N. Involvement and interactions of different immune cells and their cytokines in human visceral leishmaniasis. **Revista da Sociedade Brasileira de Medicina Tropical**, v. 46, n. 2, p. 128–134, 2013.

BHATTACHARYA, S.; BAKRE, A.; BHATTACHARYA, A. Mobile genetic elements in protozoan parasites. **Journal of genetics**, v. 81, n. 2, p. 73–86, 2002.

BIFELD, E.; CLOS, J. The genetics of *Leishmania* virulence. **Medical Microbiology and**

Immunology, v. 204, p. 619–634, 2015.

BLACKBURN, E. H. Telomeres: structure and synthesis. **THE JOURNAL OF BKH.OGKXL. CHEMISTRY**, v. 265, n. 11, p. 5919–5921, 1990.

BOETZER, M. *et al*. Scaffolding pre-assembled contigs using SSPACE. **BIOINFORMATICS APPLICATIONS NOTE**, v. 27, n. 4, p. 578–579, 2011.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: A flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114–2120, 2014.

BRINGAUD, F. *et al*. Evolution of non-LTR retrotransposons in the trypanosomatid genomes: *Leishmania major* has lost the active elements. **Molecular and biochemical parasitology**, v. 145, n. 2, p. 158–170, 2006.

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. 107–117, 5 nov. 1998.

BUNNIK, E. M.; LE ROCH, K. G. An Introduction to Functional Genomics and Systems Biology. **Advances in wound care**, v. 2, n. 9, p. 490–498, nov. 2013.

BURLE-CALDAS, G. A. *et al*. Assessment of two CRISPR-Cas9 genome editing protocols for rapid generation of *Trypanosoma cruzi* gene knockout mutants. **International Journal for Parasitology**, v. 48, n. 8, p. 591–596, 1 jul. 2018.

BUTENKO, A. *et al*. Comparative genomics of *Leishmania* (Mundinia). **BMC Genomics**, v. 20, n. 1, p. 1–12, 11 out. 2019.

CABELLO-DONAYRE, M. *et al*. *Leishmania* heme uptake involves LmFLVCRb, a novel porphyrin transporter essential for the parasite. **Cellular and Molecular Life Sciences 2019 77:9**, v. 77, n. 9, p. 1827–1845, 1 ago. 2019.

CAMACHO, E. *et al*. *Leishmania* Mitochondrial Genomes: Maxicircle Structure and Heterogeneity of Minicircles. **Genes**, v. 10, n. 10, p. 758, 26 set. 2019a.

CAMACHO, E. *et al*. Complete assembly of the *Leishmania donovani* (HU3 strain) genome and transcriptome annotation. **Scientific Reports**, v. 9, n. 1, p. 1–15, 1 dez. 2019b.

CANO, M. I.; SILVA, M. S. DA. **Frontiers in Parasitology - Molecular and Cellular Biology of Pathogenic Trypanosomatids**. 1. ed. Sharjah: Bentham Science Publishers, 2017. v. 1

CARVALHO, F. A. A. *et al*. Diagnosis of American visceral leishmaniasis in humans and dogs using the recombinant *Leishmania donovani* A2 antigen. **Diagnostic Microbiology and Infectious Disease**, v. 43, n. 4, p. 289–295, 1 ago. 2002.

CASTES, M.; TAPIA, F. J. Immunopathology of an tegumentary leishmaniasis. **Acta Científica Venezolana**, v. 49, p. 42-56, 1998.

CASTRO NETO, A. L. *et al*. In silico characterization of multiple genes encoding the GP63

virulence protein from *Leishmania braziliensis*: identification of sources of variation and putative roles in immune evasion. **BMC Genomics**, v. 20, n. 1, p. 20:108, 2019.

CECCARELLI, M. *et al.* Differentiation of leishmania (L.) infantum, leishmania (L.) amazonensis and leishmania (L.) Mexicana using sequential QPCR assays and high resolution melt analysis. **Microorganisms**, v. 8, n. 6, 1 jun. 2020.

CHAREST, H.; MATLASHEWSKI, G. Developmental Gene Expression in Leishmania donovani: Differential Cloning and Analysis of an Amastigote-Stage-Specific Gene. **MOLECULAR AND CELLULAR BIOLOGY**, v. 14, n. 5, p. 2975–2984, 1994.

CHAREST, H.; ZHANG, W. W.; MATLASHEWSKI, G. The developmental expression of *Leishmania donovani* A2 amastigote-specific genes is post-transcriptionally mediated and involves elements located in the 3'-untranslated region. **The Journal of biological chemistry**, v. 271, n. 29, p. 17081–17090, 1996.

CHEN, J. *et al.* The topology of the kinetoplast DNA network. **Cell**, v. 80, n. 1, p. 61–69, 13 jan. 1995.

CHIN SHEN CHANG; KWANG-POO CHANG. Heme requirement and acquisition by extracellular and intracellular stages of *Leishmania mexicana amazonensis*. **Molecular and Biochemical Parasitology**, v. 16, n. 3, p. 267–276, 1 set. 1985.

CHIURILLO, M. A. *et al*. Cloning and Characterization of *Leishmania donovani* Telomeres. **Experimental Parasitology**, v. 94, n. 4, p. 248–258, 1 abr. 2000.

CONG, L. *et al*. Multiplex genome engineering using CRISPR/Cas systems. **Science**, v. 339, n. 6121, p. 819–823, 15 fev. 2013.

CONTE, F. F.; CANO, M. I. N. Genomic organization of telomeric and subtelomeric sequences of *Leishmania (Leishmania) amazonensis*. **International journal for parasitology**, v. 35, n. 13, p. 1435–1443, nov. 2005.

CONVIT, J.; RONDON, A. J.; PINARDI, M. E. Diffuse cutaneous leishmaniasis – disease due to an immunological defect of host. **Transactions of the Royal Society of Tropical Medicine and Hygiene**, v. 66, p. 603-608, 1972.

COOPER, R.; DE JESUS, A. R.; CROSS, G. A. M. Deletion of an immunodominant Trypanosoma cruzi surface glycoprotein disrupts flagellum-cell adhesion. **The Journal of cell biology**, v. 122, n. 1, p. 149–156, 1993.

CORTEZ, M. *et al*. *Leishmania* promotes its own virulence by inducing expression of the host immune inhibitory ligand CD200. **Cell Host & Microbe**, v. 9, n. 6, p. 463–471, 1 jun. 2011.

CRAIG VENTER, J. *et al*. The sequence of the human genome. **Science**, v. 291, n. 5507, p. 1304–1351, 16 fev. 2001.

CROFT, S. L.; SUNDAR, S.; FAIRLAMB, A. H. Drug Resistance in Leishmaniasis. **Clinical**

Microbiology Reviews, 2006.

CRUZ, A. K.; TITUS, R.; BEVERLEY, S. M. Plasticity in chromosome number and testing of essential genes in Leishmania by targeting. **Proceedings of the National Academy of Sciences of the United States of America**, v. 90, n. 4, p. 1599–1603, 15 fev. 1993.

DA FONSECA PIRES, S. *et al*. Identification of virulence factors in *leishmania infantum* strains by a proteomic approach. **Journal of Proteome Research**, v. 13, n. 4, p. 1860–1872, 4 abr. 2014.

DAMIANOU, A. *et al*. Essential roles for deubiquitination in *Leishmania* life cycle progression. **PLoS Pathogens**, v. 16, n. 6, 1 jun. 2020.

DAROCHA, W. D. *et al*. Tests of cytoplasmic RNA interference (RNAi) and construction of a tetracycline-inducible T7 promoter system in *Trypanosoma cruzi*. **Molecular and Biochemical Parasitology**, v. 133, n. 2, p. 175–186, 2004.

DAS, L. *et al*. Successful therapy of lethal murine visceral leishmaniasis with cystatin involves up-regulation of nitric oxide and a favorable T cell response. **Journal of immunology (Baltimore, Md.: 1950)**, v. 166, n. 6, p. 4020–4028, 15 mar. 2001.

DAVENPORT, B. J. *et al*. SODB1 is essential for *Leishmania major* infection of macrophages and pathogenesis in mice. **PLOS Neglected Tropical Diseases**, v. 12, n. 10, p. e0006921, 1 out. 2018.

DÉ, F. *et al*. Genome sequencing of the lizard parasite *Leishmania tarentolae* reveals loss of genes associated to the intracellular stage of human pathogenic species. **Nucleic Acids Research**, v. 40, n. 3, p. 1131–1147, 2012.

DE LANNOY, C.; DE RIDDER, D.; RISSE, J. The long reads ahead:de novo genome assembly using the MinION. **F1000Research**, v. 6, 2017.

DE MAIO, N. *et al*. Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. **Microbial Genomics**, v. 5, n. 9, 2019.

DE PAIVA, R. M. C. *et al*. Amastin Knockdown in *Leishmania braziliensis* Affects Parasite-Macrophage Interaction and Results in Impaired Viability of Intracellular Amastigotes. **PLOS Pathogens**, v. 11, n. 12, p. e1005296, 7 dez. 2015.

DE REZENDE, E. *et al*. Quantitative proteomic analysis of amastigotes from *Leishmania (L.) amazonensis* LV79 and PH8 strains reveals molecular traits associated with the virulence phenotype. **PLoS Neglected Tropical Diseases**, v. 11, n. 11, p. 1–18, 2017.

DE SOUZA LEAO, S. *et al*. Intracellular *Leishmania amazonensis* amastigotes internalize and degrade MHC class II molecules of their host cells. **Journal of cell science**, v. 108 (Pt 10), n. 10, p. 3219–3231, 1995.

DELCHER, A. L. *et al*. Fast algorithms for large-scale genome alignment and comparison. **Nucleic Acids Research**, v. 30, n. 11, p. 2478–2483, 2002.

DENISE, H. *et al*. Studies on the CPA cysteine peptidase in the *Leishmania infantum* genome strain

JPCM5. **BMC molecular biology**, v. 7, 13 nov. 2006.

DERMINE, J. F. *et al*. *Leishmania* promastigotes require lipophosphoglycan to actively modulate the fusion properties of phagosomes at an early step of phagocytosis. **Cellular Microbiology**, v. 2, n. 2, p. 115–126, 2000.

DESJARDINS, M.; DESCOTEAUX, A. Inhibition of phagolysosomal biogenesis by the *Leishmania* lipophosphoglycan. **Journal of Experimental Medicine**, v. 185, n. 12, p. 2061–2068, 16 jun. 1997.

DIXON, S. J.; STOCKWELL, B. R. The role of iron and reactive oxygen species in cell death. **Nature Chemical Biology 2014 10:1**, v. 10, n. 1, p. 9–17, 17 dez. 2013.

DOUGALL, A. *et al*. New reports of Australian cutaneous leishmaniasis in Northern Australian macropods. **Epidemiology and Infection**, v. 137, p. 1516–1520, 2009.

DOWNING, T. *et al.* Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance. **Genome Research**, v. 21, n. 12, p. 2143–2156, 2011.

DUARTE, M. C. *et al*. Recent updates and perspectives on approaches for the development of vaccines against visceral leishmaniasis. **Revista da Sociedade Brasileira de Medicina Tropical**, v. 49, n. 4, p. 398–407, 1 jul. 2016.

EDGAR, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. **Nucleic acids research**, v. 32, n. 5, p. 1792–1797, 2004.

EID, J. *et al*. Real-time DNA sequencing from single polymerase molecules. **Science**, v. 323, n. 5910, p. 133–138, 2 jan. 2009.

EL-SAYED, N. M. *et al.* Comparative genomics of trypanosomatid parasitic protozoa. **Science**, v. 309, n. 5733, p. 404–409, 2005.

ENNES-VIDAL, V. *et al.* Calpains of *Leishmania braziliensis*: genome analysis, differential expression, and functional analysis. **Mem Inst Oswaldo Cruz**, v. 114, p. 1–12, 2019.

FERNANDES, A. P. *et al*. Immune responses induced by a *Leishmania (Leishmania) amazonensis* recombinant antigen in mice and lymphocytes from vaccinated subjects. **Revista do Instituto de Medicina Tropical de São Paulo**, v. 39, n. 2, p. 71–78, 1997.

FERNANDES, M. C. et al. Dual Transcriptome Profiling of Leishmania-Infected Human Macrophages Reveals Distinct Reprogramming Signatures. **mBio**, v. 7, n. 3, 2016.

FERRAZ COELHO, E. A. *et al.* Immune Responses Induced by the *Leishmania* (Leishmania) donovani A2 Antigen, but Not by the LACK Antigen, Are Protective against Experimental *Leishmania (Leishmania) amazonensis* Infection. **undefined**, v. 71, n. 7, p. 3988–3994, 1 jul. 2003.

FLANNERY, A. R. *et al*. LFR1 ferric iron reductase of *Leishmania amazonensis* is essential for the generation of infective parasite forms. **Journal of Biological Chemistry**, v. 286, n. 26, p. 23266–

23279, 1 jul. 2011.

FLANNERY, A. R.; RENBERG, R. L.; ANDREWS, N. W. Pathways of iron acquisition and utilization in *Leishmania*. **Current opinion in microbiology**, v. 16, n. 6, p. 716–21, 1 dez. 2013.

FLYNN, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. **PNAS**, v. 117, n. 17, p. 9451–9457, 2020.

GARIN, Y. J. F. et al. A2 gene of Old World cutaneous Leishmania is a single highly conserved functional gene. **BMC Infectious Diseases**, v. 5, n. 1, p. 1–7, 28 mar. 2005.

GHEDIN, E. *et al.* Antibody response against a *Leishmania donovani* amastigote-stage-specific protein in patients with visceral leishmaniasis. **Clinical and Diagnostic Laboratory Immunology**, v. 4, n. 5, p. 530, 1997.

GHEDIN, E. *et al*. Gene synteny and evolution of genome architecture in trypanosomatids. **Molecular and Biochemical Parasitology**, v. 134, n. 2, p. 183–191, abr. 2004.

GHURYE, J.; POP, M. Modern technologies and algorithms for scaffolding assembled genomes. **PLoS Computational Biology**, v. 15, n. 6, 1 jun. 2019.

GONZÁLEZ-DE LA FUENTE, S. *et al.* Resequencing of the *Leishmania infantum* (strain JPCM5) genome and de novo assembly into 36 contigs. **Scientific Reports**, v. 7, n. 1, p. 18050, 1 dez. 2017.

GONZÁLEZ-DE LA FUENTE, S. *et al*. Complete and de novo assembly of the *Leishmania braziliensis* (M2904) genome. **Memorias do Instituto Oswaldo Cruz**, v. 114, n. 1, p. 1–6, 1 jan. 2019.

GRENFELL, R. F. *et al*. Antigenic extracts of *Leishmania braziliensis* and *Leishmania amazonensis* associated with saponin partially protects BALB/c mice against *Leishmania* chagasi infection by suppressing IL-10 and IL-4 production. **Memórias do Instituto Oswaldo Cruz**, v. 105, n. 6, p. 818–822, 2010.

GRIMM, F.; JENNI, L. Human serum resistant promastigotes of *Leishmania infantum* in the midgut of Phlebotomus perniciosus. **Acta Tropica**, v. 52, n. 4, p. 267–273, 1 jan. 1993.

GUINDON, S. *et al*. Estimating maximum likelihood phylogenies with PhyML. **Methods in Molecular Biology**, v. 537, p. 113–137, 2009.

GUPTA, G.; OGHUMU, S.; SATOSKAR, A. R. Mechanisms of Immune Evasion in Leishmaniasis. In: **Advances in Applied Microbiology**. [s.l.] Academic Press Inc., 2013. v. 82p. 155–184.

H, C.; G, M. Developmental gene expression in *Leishmania donovani*: differential cloning and analysis of an amastigote-stage-specific gene. **Molecular and cellular biology**, v. 14, n. 5, p. 2975–2984, maio 1994.

HANDLER, M. Z. *et al.* Cutaneous and mucocutaneous leishmaniasis: Clinical perspectives. **Journal of the American Academy of Dermatology**, v. 73, n. 6, p. 897–908, 1 dez. 2015.

HOFFMANN, A. *et al*. *Leishmania amazonensis* in dog with clinical diagnosis of visceral leishmaniasis in Paraná State, Brazil – a case report. **Semina: Ciências Agrárias**, v. 33, n. 6Supl2, p. 3265–3270, 28 fev. 2013.

HOLLEY, R. W. *et al*. Structure of a ribonucleic acid. **Science**, v. 147, n. 3664, p. 1462–1465, 1965.

HORVATH, A. *et al*. Analysis of repeats in the divergent region of the maxicircle kinetoplast DNA of Crithidia oncopelti. **Molecular Biology**, v. 24, p. 1539–48, 1990.

HUYNH, C.; SACKS, D. L.; ANDREWS, N. W. A *Leishmania amazonensis* ZIP family iron transporter is essential for parasite replication within macrophage phagolysosomes. **Journal of Experimental Medicine**, v. 203, n. 10, p. 2363–2375, 2 out. 2006.

IANTORNO, S. A. *et al*. Gene Expression in *Leishmania* Is Regulated Predominantly by Gene Dosage. **mBio**, v. 8, n. 5, 1 set. 2017.

ILG, T.; DEMAR, M.; HARBECKE, D. Phosphoglycan Repeat-deficient *Leishmania mexicana* Parasites Remain Infectious to Macrophages and Mice. **Journal of Biological Chemistry**, v. 276, n. 7, p. 4988–4997, 16 fev. 2001.

INBAR, E. et al. The Transcriptome of Leishmania major Developmental Stages in Their Natural Sand Fly Vector. **mBio**, v. 8, n. 2, 1 mar. 2017.

INES, L. *et al*. Characterization of *Leishmania* spp. causing cutaneous leishmaniasis in Manaus, Amazonas, Brazil. **Parasitology research**, v. 108, n. 3, p. 671–677, 2011.

IVENS, A. C. *et al*. The genome of the kinetoplastid parasite, *Leishmania major*. **Science**, v. 309, n. 5733, p. 436–442, 15 jul. 2005.

IVENS, A. C.; BLACKWELL, J. M. Unravelling the *Leishmania* genome. **Current Opinion in Genetics and Development**, v. 6, n. 6, p. 704–710, 1996.

JACKSON, A. P. The evolution of amastin surface glycoproteins in trypanosomatid parasites. **Molecular Biology and Evolution**, v. 27, n. 1, p. 33–45, 2010.

JANG-IL SOHN; JIN-WU NAM. The Present and Future of De Novo Whole-Genome Assembly - PubMed. **Briefings in Bioinformatics**, v. 19, n. 1, p. 23–40, 2018.

JENSEN, R. E.; ENGLUND, P. T. Network News: The Replication of Kinetoplast DNA. **Annual Review of Microbiology**, v. 66, n. 1, p. 473–491, 13 out. 2012.

K. ROSE *et al*. Cutaneous leishmaniasis in red kangaroos: isolation and characterisation of the causative organisms. **International Journal for Parasitology**, v. 34, p. 655–664, 2004.

KADOBIANSKYI, M. *et al*. Hybrid genome assembly and annotation of *Danionella translucida*. **Scientific data**, v. 6, n. 1, p. 156, 26 ago. 2019.

KANGUSSU-MARCOLINO, M. M. *et al*. Distinct genomic organization, mRNA expression and cellular localization of members of two amastin sub-families present in *Trypanosoma cruzi*. **BMC**

Microbiology, v. 13, n. 1, p. 10, 2013.

KATOH, K. *et al*. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. **Nucleic acids research**, v. 30, n. 14, p. 3059–3066, 15 jul. 2002.

KCHOUK, M.; GIBRAT, J.-F.; ELLOUMI, M. Generations of Sequencing Technologies: From First to Next Generation. 2017.

KEVRIC, I.; CAPPEL, M. A.; KEELING, J. H. New World and Old World *Leishmania* Infections: A Practical Review. **Dermatologic Clinics**, v. 33, n. 3, p. 579–593, 2015.

KOREN, S. *et al*. Canu: Scalable and Accurate Long-Read Assembly via Adaptive k-mer Weighting and Repeat Separation. **Genome Research**, v. 27, n. 5, p. 722–736, 2017.

KOŘENÝ, L.; OBORNÍK, M.; LUKEŠ, J. Make It, Take It, or Leave It: Heme Metabolism of Parasites. **PLOS Pathogens**, v. 9, n. 1, p. e1003088, jan. 2013.

KRAMER, S. Developmental regulation of gene expression in the absence of transcriptional control: The case of kinetoplastids. **Molecular and Biochemical Parasitology**, v. 181, n. 2, p. 61–72, 1 fev. 2012.

KULKARNI, M. M. *et al*. *Trypanosoma cruzi* GP63 proteins undergo stage-specific differential posttranslational modification and are important for host cell infection. **Infection and Immunity**, v. 77, n. 5, p. 2193–2200, maio 2009.

LAINSON, R.; SHAW, J. J. **Evolution, classification and, geographical distribution**. 1. ed. Orlando: Academic Press Inc., 1987.

LANDER, E. S. et al. Initial sequencing and analysis of the human genome. **Nature**, v. 409, n. 6822, p. 860–921, 15 fev. 2001.

LAPATAS, V. et al. Data integration in biological research: an overview. **Journal of Biological Research**, v. 22, n. 1, p. 9, dez. 2015.

LARANJEIRA-SILVA, M. F. *et al*. A MFS-like plasma membrane transporter required for *Leishmania* virulence protects the parasites from iron toxicity. **PLOS Pathogens**, v. 14, n. 6, p. e1007140, 1 jun. 2018.

LARANJEIRA-SILVA, M. F.; HAMZA, I.; PÉREZ-VICTORIA, J. M. Iron and Heme Metabolism at the *Leishmania*–Host Interface. **Trends in Parasitology**, v. 36, n. 3, p. 279–289, 1 mar. 2020.

LASAKOSVITSCH, F. *et al*. Cloning and characterisation of a cysteine proteinase gene expressed in amastigotes of *Leishmania (L.) amazonensis*. **International Journal for Parasitology**, v. 33, n. 4, p. 445–454, 2003.

LASKOWSKI, R. A. *et al*. PROCHECK: a program to check the stereochemical quality of protein structures. **Journal of Applied Crystallography**, v. 26, n. 2, p. 283–291, 1 abr. 1993.

Leishmaniasis. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/leishmaniasis>. Acesso em: 1 jul. 2020.

LEMOS-SILVA, T.; TELLERIA, E. L.; TRAUB-CSEKÖ, Y.M. The gene expression of *Leishmania infantum chagasi* inside *Lutzomyia longipalpis*, the main vector of visceral leishmaniasis in Brazil. **Memórias do Instituto Oswaldo Cruz**, v. 126, 2021.

LISCHER, H. E. L.; SHIMIZU, K. K. Reference-guided de novo assembly approach improves genome reconstruction for related species. **BMC bioinformatics**, v. 18, n. 1, 10 nov. 2017.

LYE, L.-F. *et al*. Retention and Loss of RNA Interference Pathways in Trypanosomatid Protozoans. **PLoS Pathogens**, v. 6, n. 10, p. e1001161, 28 out. 2010.

LYPACZEWSKI, P. *et al*. A complete *Leishmania donovani* reference genome identifies novel genetic variations associated with virulence. **Scientific Reports**, v. 8, n. 1, p. 1–14, 1 dez. 2018.

MA, L. *et al*. An evolutionary analysis of trypanosomatid GP63 proteases. **Parasitology research**, v. 109, n. 4, p. 1075–1084, 2011.

MAHMOUDZADEH-NIKNAM, H.; MCKERROW, J. H. *Leishmania tropica*: cysteine proteases are essential for growth and pathogenicity. **Experimental parasitology**, v. 106, n. 3–4, p. 158–163, mar. 2004.

MALI, P. *et al*. RNA-guided human genome engineering via Cas9. **Science (New York, N.Y.)**, v. 339, n. 6121, p. 823–826, 15 fev. 2013.

MANNAERT, A. *et al*. **Adaptive mechanisms in pathogens: Universal aneuploidy in Leishmania**. **Trends in Parasitology**Elsevier Current Trends, 1 set. 2012.

MANNI, M. *et al*. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. **Molecular Biology and Evolution**, v. 38, n. 10, p. 4647–4654, 27 set. 2021.

MARGOS, G. *et al*. *Borrelia maritima* sp. Nov., a novel species of the borrelia burgdorferi sensu lato complex, occupying a basal position to north american species. **International Journal of Systematic and Evolutionary Microbiology**, v. 70, n. 2, p. 849–856, 2020.

MARTÍNEZ-CALVILLO, S. *et al*. Transcription of *Leishmania major* Friedlin chromosome 1 initiates in both directions within a single region. **Molecular Cell**, v. 11, n. 5, p. 1291–1299, 2003.

MARTÍNEZ-CALVILLO, S. *et al*. Transcription initiation and termination on *Leishmania major* chromosome 3. **Eukaryotic Cell**, v. 3, n. 2, p. 506–517, 2004.

MARTÍNEZ-GARCÍA, M. *et al*. LmABCB3, an atypical mitochondrial ABC transporter essential for *Leishmania major* virulence, acts in heme and cytosolic iron/sulfur clusters biogenesis. **Parasites and Vectors**, v. 9, n. 1, p. 1–17, 5 jan. 2016.

MARTÍNEZ-RODRIGO, A. *et al*. Immunization with the HisAK70 DNA Vaccine Induces Resistance against *Leishmania amazonensis* Infection in BALB/c Mice. 2019.

MARTYNKINA, L. P. *et al*. Structural organization of kinetoplast DNA and its compaction in the in vitro model system. **European journal of cell biology**, v. 56, n. 1, p. 123–31, out. 1991.

MARZOCHI, K. B. F. *et al*. Phase 1 Study of an Inactivated Vaccine against American Tegumentary Leishmaniasis in Normal Volunteers in Brazil. **Memórias do Instituto Oswaldo Cruz**, v. 93, n. 2, p. 205–212, 1998.

MASLOV, D. A. *et al*. An intergenic G-rich region in *Leishmania tarentolae* kinetoplast maxicircle DNA is a pan-edited cryptogene encoding ribosomal protein S12. **Molecular and cellular biology**, v. 12, n. 1, p. 56–67, jan. 1992.

MASLOV, D. A. Complete set of mitochondrial pan-edited mRNAs in *Leishmania mexicana amazonensis* LV78. **Molecular and biochemical parasitology**, v. 173, n. 2, p. 107, out. 2010.

MASLOV, D. A.; THIEMANN, O.; SIMPSON, L. Editing and misediting of transcripts of the kinetoplast maxicircle G5 (ND3) cryptogene in an old laboratory strain of *Leishmania tarentolae*. **Molecular and Biochemical Parasitology**, v. 68, n. 1, p. 155–159, 1 nov. 1994.

MATLASHEWSKI, G. *Leishmania* infection and virulence. **Medical Microbiology and Immunology**, v. 190, n. 1–2, p. 37–42, 2001.

MAXAM, A. M.; GILBERT, W. A new method for sequencing DNA. **Proceedings of the National Academy of Sciences of the United States of America**, v. 74, n. 2, p. 560–564, 1977.

MAYRINK, W. *et al*. A field trial of a vaccine against american dermal leishmaniasis. **Transactions of the Royal Society of Tropical Medicine and Hygiene**, v. 73, n. 4, p. 385–387, 1979.

MAYRINK, W. *et al*. Vaccination of C57BL/10 mice against cutaneous leishmaniasis using killed promastigotes of different strains and species of Leishmania. **Revista da Sociedade Brasileira de Medicina Tropical**, v. 35, n. 2, p. 125–132, 2002.

MCCONVILLE, M. J.; FERGUSON, M. A. J. The structure, biosynthesis and function of glycosylated phosphatidylinositols in the parasitic protozoa and higher eukaryotes. **Biochemical Journal**, v. 294, n. 2, p. 305–324, 1993.

MCGUFFIN, L. J.; BRYSON, K.; JONES, D. T. The PSIPRED protein structure prediction server. **Bioinformatics (Oxford, England)**, v. 16, n. 4, p. 404–405, 2000.

MCGWIRE, B.; CHANG, K. P. Genetic rescue of surface metalloproteinase (gp63)-deficiency in *Leishmania amazonensis* variants increases their infection of macrophages at the early phase. **Molecular and Biochemical Parasitology**, v. 66, n. 2, p. 345–347, 1994.

MCMAHON-PRATT, D.; ALEXANDER, J. Does the *Leishmania major* paradigm of pathogenesis and protection hold for New World cutaneous leishmaniases or the visceral disease? **Immunological reviews**, v. 201, p. 206–224, out. 2004.

MEDINA, L. S. *et al*. The gp63 Gene Cluster Is Highly Polymorphic in Natural *Leishmania (Viannia) braziliensis* Populations, but Functional Sites Are Conserved. **PLoS ONE**, v. 11, n. 9, 2016.

MENDES WANDERLEY, J. L. et al. Subversion of immunity by Leishmania amazonensis parasites: Possible role of phosphatidylserine as a main regulator. **Journal of Parasitology Research**, v. 2012, 2012.

MIGUEL, D. C. *et al.* Heme uptake mediated by lhr1 is essential for *Leishmania amazonensis* virulence. **Infection and Immunity**, v. 81, n. 10, p. 3620–3626, 2013.

MIKHEYEV, A. S.; TIN, M. M. Y. A first look at the Oxford Nanopore MinION sequencer. **Molecular Ecology Resources**, v. 14, n. 6, p. 1097–1102, 1 nov. 2014.

MILLER, J. R. *et al.* Hybrid assembly with long and short reads improves discovery of gene family expansions. **BMC Genomics**, v. 18, n. 1, p. 541, 19 jul. 2017.

MITTRA, B. *et al*. A Trypanosomatid Iron Transporter that Regulates Mitochondrial Function Is Required for *Leishmania amazonensis* Virulence. **PLOS Pathogens**, v. 12, n. 1, p. e1005340, 2016.

MITTRA, B. *et al*. The iron-dependent mitochondrial superoxide dismutase SODA promotes *Leishmania* virulence. **Journal of Biological Chemistry**, v. 292, n. 29, p. 12324–12338, 21 jul. 2017.

MIZBANI, A. *et al.* Effect of A2 gene on infectivity of the nonpathogenic parasite *Leishmania tarentolae*. **Parasitology Research**, v. 109, n. 3, p. 793–799, set. 2011.

MOMEN, H.; CUPOLILLO, E. Speculations on the Origin and Evolution of the Genus *Leishmania*. **Memórias do Instituto Oswaldo Cruz**, v. 95, n. 4, p. 583–588, 2000.

MONTALVO-ÁLVAREZ, A. M. *et al*. The *Leishmania* HSP20 is antigenic during natural infections, but, as DNA vaccine, it does not protect BALB/c mice against experimental *L. amazonensis* infection. **Journal of Biomedicine and Biotechnology**, v. 2008, n. 1, 2008.

MOSMANN, T. Rapid colorimetric assay for cellular growth and survival: Application to proliferation and cytotoxicity assays. **Journal of Immunological Methods**, v. 65, n. 1–2, p. 55–63, 16 dez. 1983.

MOTTRAM, J. C. *et al*. Evidence from disruption of the lmcpb gene array of *Leishmania mexicana* that cysteine proteinases are virulence factors. **Proceedings of the National Academy of Sciences**, v. 93, n. 12, p. 6008–6013, 11 jun. 1996.

MÜLLER, L. S. M. et al. Genome organization and DNA accessibility control antigenic variation in trypanosomes. **Nature 2018 563:7729**, v. 563, n. 7729, p. 121–125, 17 out. 2018.

MUNDODI, V.; KUCKNOOR, A. S.; GEDAMU, L. Role of *Leishmania (Leishmania) chagasi* amastigote cysteine protease in intracellular parasite survival: studies by gene disruption and antisense mRNA inhibition. **BMC molecular biology**, v. 6, 3 fev. 2005.

MURRAY, H. W. *et al.* Advances in leishmaniasis. **Lancet**, v. 366, n. 9496, p. 1561–1577, 29 out. 2005.

MYLER, P. J. et al. Structural organization of the maxicircle variable region of Trypanosoma

brucei: identification of potential replication origins and topoisomerase II binding sites. **Nucleic Acids Research**, v. 21, n. 3, p. 687, 2 fev. 1993.

NACHTWEIDE, S.; STANKE, M. Multi-Genome Annotation with AUGUSTUS. **Methods in molecular biology (Clifton, N.J.)**, v. 1962, p. 139–160, 2019.

NADALIN, F.; VEZZI, F.; POLICRITI, A. GapFiller: A de novo assembly approach to fill the gap within paired reads. **BMC Bioinformatics**, v. 13, n. SUPPL 1, 7 set. 2012.

NADERER, T.; MCCONVILLE, M. J. The Leishmania-macrophage interaction: A metabolic perspective. **Cellular Microbiology**, v. 10, n. 2, p. 301–308, fev. 2008.

NANDAN, D. *et al*. Leishmania EF-1alpha activates the Src homology 2 domain containing tyrosine phosphatase SHP-1 leading to macrophage deactivation. **The Journal of biological chemistry**, v. 277, n. 51, p. 50190–50197, oct. 2002.

NASCIMENTO, E. *et al*. Vaccination of humans against cutaneous leishmaniasis: cellular and humoral immune responses. **Infection and Immunity**, v. 58, n. 7, p. 2198, 1990.

NATARAJAN, G. *et al*. Mechanisms of immunopathology of leishmaniasis. **Pathogenesis of Leishmaniasis: New Developments in Research**, p. 1–13, 1 dez. 2013.

NGÔ, H. *et al*. Double-stranded RNA induces mRNA degradation in *trypanosoma brucei*. **Proceedings of the National Academy of Sciences of the United States of America**, v. 95, n. 25, p. 14687–14692, 8 dez. 1998.

NIH. **Genbak**. Disponível em: <https://www.ncbi.nlm.nih.gov/genbank/>. Acesso em: 20 jul. 2022.

NOCUA, P. *et al*. Redalyc.Secuencia parcial del genoma del maxicírculo de *Leishmania braziliensis*, comparación con otros tripanosomátidos. v. 16, p. 29–50, 2011.

NOGUEIRA, P. M. *et al*. Lipophosphoglycans from *Leishmania amazonensis* Strains Display Immunomodulatory Properties via TLR4 and Do Not Affect Sand Fly Infection. **PLOS Neglected Tropical Diseases**, v. 10, n. 8, p. e0004848, 10 ago. 2016.

OLIVIER, M. *et al*. *Leishmania* virulence factors: Focus on the metalloprotease GP63. **Microbes and Infection**, v. 14, n. 15, p. 1377–1389, dez. 2012.

OLIVEIRA, A. E. R. et al. Gene expression network analyses during infection with virulent and avirulent Trypanosoma cruzi strains unveil a role for fibroblasts in neutrophil recruitment and activation. **PLOS Pathogens**, v. 16, n. 8, p. e1008781, 1 ago. 2020.

OTTO, T. D. **IPA: Script to improve long read (pacbio) assemblies**. Disponível em: <https://github.com/ThomasDOtto/IPA>. Acesso em: 22 jun. 2020.

OTTO, T. D. *et al*. RATT: Rapid Annotation Transfer Tool. **Nucleic Acids Research**, v. 39, n. 9, 2011.

(PAHO), P. A. H. O. **LEISHMANIOSES: Informe epidemiológico das Américas**. [s.l: s.n.]. Disponível em: <https://iris.paho.org/handle/10665.2/55386>.

PATINO, L. H. *et al.* Genomic analyses reveal moderate levels of ploidy, high heterozygosity and structural variations in a Colombian isolate of *Leishmania (Leishmania) amazonensis*. **Acta Tropica**, v. 203, p. 105296, 1 mar. 2020.

PAULO, J. *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**, v. 6, p. 5, 2007.

PEACOCK, C. S. *et al.* Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. **Nature genetics**, v. 39, n. 7, p. 839–847, 2007.

PENG, D. *et al.* CRISPR-Cas9-mediated single-gene and gene family disruption in *Trypanosoma cruzi*. **mBio**, v. 6, n. 1, 30 dez. 2015.

PEREIRA, B. A. S.; ALVES, C. R. Immunological characteristics of experimental murine infection with *Leishmania (Leishmania) amazonensis*. **Veterinary Parasitology**, v. 158, n. 4, p. 239–255, 20 dez. 2008.

PEVSNER, J. **BIOINFORMATICS AND FUNCTIONAL GENOMICS**. third edit ed. Chichester: Wiley-Blackwell, 2015.

PISCOPO, T. V; MALLIA, A. C. Leishmaniasis. **Postgrad Med J**, v. 82, p. 649–657, 2006.

PLEWES, K. A.; BARR, S. D.; GEDAMU, L. Iron superoxide dismutases targeted to the glycosomes of *Leishmania* chagasi are important for survival. **Infection and Immunity**, v. 71, n. 10, p. 5910–5920, 1 out. 2003.

POLLO, S. M. J. *et al.* Benchmarking hybrid assemblies of Giardia and prediction of widespread intra-isolate structural variation. **Parasites and Vectors**, v. 13, n. 1, p. 1–13, 28 fev. 2020.

POOT J, DENISE H, HERRMANN DC, MOTTRAM JC, COOMBS GH, V. Virulence and protective potential of several Cysteine peptidase knockout strains of *Leishmania infantum* in hamsters. In: UTRECHT, P. J. (Ed.). **Experimental challenge models for canine leishmaniasis in hamsters and dogs, optimization and application in vaccine research**. [s.l.] Utrecht University press, 2006. p. 93–107.

PORTO, V. B. G. *et al*. Visceral leishmaniasis caused by *Leishmania (Leishmania) amazonensis* associated with Hodgkin's lymphoma. **Revista do Instituto de Medicina Tropical de Sao Paulo**, v. 64, 2022.

PRJIBELSKI, A. *et al*. Using SPAdes De Novo Assembler. **Current Protocols in Bioinformatics**, v. 70, n. 1, 1 jun. 2020.

PUENTES, F. *et al.* Leishmania: Fine mapping of the Leishmanolysin molecule's conserved core domains involved in binding and internalization. **Experimental Parasitology**, v. 93, n. 1, p. 7–22, set. 1999.

QUEVILLON, E. *et al*. InterProScan: protein domains identifier. **Nucleic acids research**, v. 33, n.

Web Server issue, jul. 2005.

REAL, F. *et al*. The genome sequence of *Leishmania (Leishmania) amazonensis*: Functional annotation and extended analysis of gene models. **DNA Research**, v. 20, n. 6, p. 567–581, 2013.

REIS-CUNHA, J. L.; BARTHOLOMEU, D. C. Trypanosoma cruzi Genome Assemblies: Challenges and Milestones of Assembling a Highly Repetitive and Complex Genome. **Methods in molecular biology (Clifton, N.J.)**, v. 1955, p. 1–22, 2019.

REQUENA, J. M. *et al*. Genomic cartography and proposal of nomenclature for the repeated, interspersed elements of the *Leishmania major* SIDER2 family and identification of SIDER2-containing transcripts. **Molecular and Biochemical Parasitology**, v. 212, p. 9–15, 1 mar. 2017.

RHOADS, A.; AU, K. F. PacBio Sequencing and Its Applications. **Genomics, Proteomics and Bioinformatics**, v. 13, n. 5, p. 278–289, 2015.

ROBINSON, K. A.; BEVERLEY, S. M. Improvements in transfection efficiency and tests of RNA interference (RNAi) approaches in the protozoan parasite Leishmania. **Molecular and Biochemical Parasitology**, v. 128, n. 2, p. 217–228, 2003.

ROCHETTE, A. *et al*. Characterization and developmental gene regulation of a large gene family encoding amastin surface proteins in *Leishmania* spp. **Molecular and Biochemical Parasitology**, v. 140, n. 2, p. 205–220, 2005.

ROGERS, M. B. *et al*. Chromosome and gene copy number variation allow major structural change between species and strains of Leishmania. **Genome Research**, v. 21, p. 2129–2142, 2011.

SACKS, D. L. *et al*. The role of phosphoglycans in Leishmania-sand fly interactions. **PNAS**, v. 97, n. 1, 2000.

RUSSO, D. M.; BARRAL-NETTO, M.; BARRAL, A.; REED, S. G. Human T-cell responses in Leishmania infections. **Progress in clinical parasitology**, v. 3, p. 119-144, 1993.

SAGAR M. UTTURKAR *et al*. Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences | Bioinformatics | Oxford Academic. **Bioinformatics**, v. 30, n. 19, p. 2709–2716, 2014.

SALOTRA, P. *et al*. Upregulation of surface proteins in *Leishmania donovani* isolated from patients of post kala-azar dermal leishmaniasis. **Microbes and Infection**, v. 8, n. 3, p. 637–644, mar. 2006.

SAMARASINGHE, S. R. *et al*. Genomic insights into virulence mechanisms of *Leishmania donovani*: Evidence from an atypical strain. **BMC Genomics**, v. 19, n. 1, p. 1–18, 28 nov. 2018.

SANGER, F.; NICKLEN, S.; COULSON, A. R. DNA sequencing with chain-terminating inhibitors. **Proceedings of the National Academy of Sciences of the United States of America**, v. 74, n. 12, p. 5463–5467, 1977.

SARKAR, A.; ANDREWS, N. W.; LARANJEIRA-SILVA, M. F. Intracellular iron availability

modulates the requirement for Leishmania Iron Regulator 1 (LIR1) during macrophage infections. **International Journal for Parasitology**, v. 49, n. 6, p. 423–427, 1 maio 2019.

SCHADT, E. E.; TURNER, S.; KASARSKIS, A. A window into third-generation sequencing. **Human Molecular Genetics**, v. 19, n. R2, p. R227–R240, 15 out. 2010.

SCHAIBLE, U. E.; KAUFMANN, S. H. E. Iron and microbial infection. **Nature Reviews Microbiology 2004 2:12**, v. 2, n. 12, p. 946–953, dez. 2004.

SCHLAGENHAUF, E.; ETGES, R.; METCALF, P. The crystal structure of the *Leishmania major* surface proteinase leishmanolysin (gp63). **Structure**, v. 6, n. 8, p. 1035–1046, 15 ago. 1998.

SCHWEDE, T. *et al.* SWISS-MODEL: an automated protein homology-modeling server. **Nucleic Acids Research**, v. 31, n. 13, p. 3381, 1 jul. 2003.

SHAW, J. M. *et al*. Editing of kinetoplastid mitochondrial mRNAs by uridine addition and deletion generates conserved amino acid sequences and AUG initiation codons. **Cell**, v. 53, n. 3, p. 401–411, 6 maio 1988.

SILVA-ALMEIDA, M. *et al*. Proteinases as virulence factors in *Leishmania* spp. infection in mammals. **Parasites and Vectors**, v. 5, n. 1, 2012.

SILVA-ALMEIDA, M. *et al.* Overview of the organization of protease genes in the genome of *Leishmania* spp. **Parasites and Vectors**, v. 7, n. 1, p. 1–7, 20 ago. 2014.

SILVEIRA, F. T. *et al*. Immunopathogenic competences of *Leishmania* (*V.*) *braziliensis* and *L.* (*L.*) *amazonensis* in American cutaneous leishmaniasis. **Parasite Immunology**, v. 31, n. 8, p. 423–431, ago. 2009.

SIMPSON, L. The Mitochondrial Genome of Kinetoplastid Protozoa: Genomic Organization, Transcription, Replication, and Evolution. **Annual Review of Microbiology**, v. 41, n. 1, p. 363–380, out. 1987.

SIMPSON, L. *et al*. Evolution of RNA editing in trypanosome mitochondria. **Proceedings of the National Academy of Sciences of the United States of America**, v. 97, n. 13, p. 6986, 6 jun. 2000.

SIMPSON, L. *et al*. Comparison of the Mitochondrial Genomes and Steady State Transcriptomes of Two Strains of the Trypanosomatid Parasite, *Leishmania tarentolae*. **PLOS Neglected Tropical Diseases**, v. 9, n. 7, p. e0003841, 2015.

SIMPSON, L.; SBICEGO, S.; APHASIZHEV, R. Uridine insertion/deletion RNA editing in trypanosome mitochondria: A complex business. **RNA**, v. 9, n. 3, p. 265–276, 1 mar. 2003.

SMITH, D. F.; PEACOCK, C. S.; CRUZ, A. K. Comparative genomics: From genotype to disease phenotype in the leishmaniases. **International Journal for Parasitology**, v. 37, n. 11, p. 1173–1186, set. 2007.

SOLLELIS, L. *et al*. First efficient CRISPR-Cas9-mediated genome editing in *Leishmania*

parasites. **Cellular Microbiology**, v. 17, n. 10, p. 1405–1412, 1 out. 2015.

SOONG, L.; HENARD, C. A.; MELBY, P. C. Immunopathogenesis of non-healing American cutaneous leishmaniasis and progressive visceral leishmaniasis. **Seminars in Immunopathology 2012 34:6**, v. 34, n. 6, p. 735–751, 11 out. 2012.

STEINMETZ, L. M.; DAVIS, R. W. Maximizing the potential of functional genomics. **Nature Reviews Genetics 2004 5:3**, v. 5, n. 3, p. 190–201, mar. 2004.

STEPHEN M. BEVERLEY; SALVATORE J. TURCO. Lipophosphoglycan (LPG) and the identification of virulence genes in the protozoan parasite Leishmania. **Trends Microbiology**, v. 6, n. 1, p. 35–40, 1998.

STERKERS, Y. *et al*. Novel insights into genome plasticity in Eukaryotes: mosaic aneuploidy in *Leishmania*. **Molecular Microbiology**, v. 86, n. 1, p. 15–23, out. 2012.

STEVERDING, D. The history of leishmaniasis. **Steverding Parasites & Vectors**, v. 10, n. 82, 2017.

SUBRAMANIAM, C. et al. Chromosome-Wide Analysis of Gene Function by RNA Interference in the African Trypanosome. **Eukaryotic Cell**, v. 5, n. 9, p. 1539, set. 2006.

SUTTER, A. *et al*. Structural insights into leishmanolysins encoded on chromosome 10 of *Leishmania (Viannia) braziliensis*. **Memórias do Instituto Oswaldo Cruz**, v. 112, n. 9, p. 617–625, 1 set. 2017.

SUZUKI, Y. Informatics for PacBio Long Reads. In: **Advances in Experimental Medicine and Biology**. [s.l.] Springer New York LLC, 2019. v. 1129p. 119–129.

TARAILO-GRAOVAC, M.; CHEN, N. Using RepeatMasker to identify repetitive elements in genomic sequences. **Current protocols in bioinformatics**, v. Chapter 4, n. SUPPL. 25, 2009.

TEIXEIRA, S. M. R.; KIRCHHOFF, L. V.; DONELSON, J. E. Post-transcriptional elements regulating expression of mRNAs from the amastin/tuzin gene cluster of *Trypanosoma cruzi*. **Journal of Biological Chemistry**, v. 270, n. 38, p. 22586–22594, 22 set. 1995.

TIMM, T. *et al*. The Eukaryotic Elongation Factor 1 Alpha (eEF1α) from the Parasite Leishmania infantum Is Modified with the Immunomodulatory Substituent Phosphorylcholine (PC). **Molecules**, v. 22, n. 12, nov. 2017.

THIEMANN, O. H.; MASLOV, D. A.; SIMPSON, L. Disruption of RNA editing in *Leishmania tarentolae* by the loss of minicircle-encoded guide RNA genes. **The EMBO Journal**, v. 13, n. 23, p. 5689, 12 dez. 1994.

THOMPSON, J. D.; GIBSON, T. J.; HIGGINS, D. G. Multiple sequence alignment using ClustalW and ClustalX. **Current protocols in bioinformatics**, v. Chapter 2, n. 1, jan. 2002.

THOMPSON, J. F.; STEINMANN, K. E. Single Molecule Sequencing with a HeliScope Genetic Analysis System. **Current Protocols in Molecular Biology**, v. 92, n. 1, p. 7.10.1-7.10.14, 1 out.

2010.

TOLEZANO, J. E. *et al*. The first records of *Leishmania (Leishmania) amazonensis* in dogs (Canis familiaris) diagnosed clinically as having canine visceral leishmaniasis from Araçatuba County, São Paulo State, Brazil. **Veterinary parasitology**, v. 149, n. 3–4, p. 280–284, 10 nov. 2007.

TSCHOEKE, D. A. *et al*. The comparative genomics and phylogenomics of *Leishmania amazonensis* parasite. **Evolutionary Bioinformatics**, v. 10, p. 131–153, 23 set. 2014.

TURCO, S. J.; DESCOTEAUX, A. The Lipophosphoglycan of *Leishmania* Parasites. **Annual Review of Microbiology**, v. 46, n. 1, p. 65–92, out. 1992.

UBEDA, J. M. *et al*. Genome-Wide Stochastic Adaptive DNA Amplification at Direct and Inverted DNA Repeats in the Parasite Leishmania. **PLOS Biology**, v. 12, n. 5, p. e1001868, 2014.

VALDIVIA, H. O. *et al.* Comparative genomic analysis of *Leishmania* (Viannia) peruviana and *Leishmania* (Viannia) *braziliensis*. v. 16, n. 715, 2011.

VALDIVIA, H. O. *et al.* Comparative genomics of canineisolated *Leishmania (Leishmania) amazonensis* from an endemic focus of visceral leishmaniasis in Governador Valadares, southeastern Brazil. **Scientific Reports**, v. 16, n. 7, 2017.

VAN DIJK, E. L. *et al*. The Third Revolution in Sequencing Technology. **Trends in Genetics**, v. 34, n. 9, p. 666–681, 1 set. 2018.

VELASQUEZ, L. G. *et al*. Distinct courses of infection with *Leishmania (L.) amazonensis* are observed in BALB/c, BALB/c nude and C57BL/6 mice. **Parasitology**, v. 143, n. 6, p. 692–703, 1 maio 2016.

VICTOIR, K.; DUJARDIN, J. C. How to succeed in parasitic life without sex? Asking *Leishmania*. **Trends in Parasitology**, v. 18, n. 2, p. 81–85, 2002.

VILLANUEVA, M. S. *et al*. A new member of a family of site-specific retrotransposons is present in the spliced leader RNA genes of *Trypanosoma cruzi*. **Molecular and Cellular Biology**, v. 11, n. 12, p. 6139, dez. 1991.

XU, C. W. *et al*. Nucleus-encoded histone H1-like proteins are associated with kinetoplast DNA in the trypanosomatid *Crithidia fasciculata*. **Molecular and Cellular Biology**, v. 16, p. 564–576, fev. 1996.

XU, C.; RAY, D. S. Isolation of proteins associated with kinetoplast DNA networks in vivo. **Proceeding of the National Academy of Sciences of USA**, v. 90, p. 1786–1789, mar. 1993.

WALKER, B. J. et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. **PLoS ONE**, v. 9, n. 11, p. e112963, 19 nov. 2014.

WALKER, J. *et al.* Comparative protein profiling identifies elongation factor-1β and tryparedoxin peroxidase as factors associated with metastasis in *Leishmania guyanensis*. **Molecular and Biochemical Parasitology**, v. 145, n. 2, p. 254–264, 1 fev. 2006.

WALLACE, I. M. *et al*. M-Coffee: combining multiple sequence alignment methods with T-Coffee. **Nucleic Acids Research**, v. 34, n. 6, p. 1692, 2006.

WANG, W. *et al*. Strain-specific genome evolution in *Trypanosoma cruzi*, the agent of Chagas disease. **PLOS Pathogens**, v. 17, n. 1, p. e1009254, 28 jan. 2021.

WATSON, J. D.; CRICK, F. H. C. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. **Nature**, v. 171, n. 4356, p. 737–738, 1953.

WEN-WEI ZHANG *et al*. Identification and overexpression of the A2 amastigote-specific protein in *Leishmania donovani*. **Molecular and Biochemical Parasitology**, v. 78, p. 79–90, 1996.

WESTENBERGER, S. J. *et al*. *Trypanosoma cruzi* mitochondrial maxicircles display species- and strain-specific variation and a conserved element in the non-coding region. **BMC Genomics**, v. 7, n. 1, p. 1–18, 22 mar. 2006.

WICKSTEAD, B.; ERSFELD, K.; GULL, K. Repetitive elements in genomes of parasitic protozoa. **Microbiology and molecular biology reviews: MMBR**, v. 67, n. 3, p. 360–375, set. 2003.

WINCKER, P. *et al*. The *Leishmania* genome comprises 36 chromosomes conserved across widely divergent human pathogenic species. **Nucleic Acids Research**, v. 24, n. 9, p. 1688–1694, 1996.

WORLD HEALTH ORGANIZATION. **Leishmaniasis**. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/leishmaniasis>. Acesso em: 20 mar. 2022.

XIANG, L. *et al*. Ascorbate-Dependent Peroxidase (APX) from *Leishmania amazonensis* is a reactive oxygen species-induced essential enzyme that regulates virulence. **Infection and Immunity**, v. 87, n. 12, 1 dez. 2019.

YATAWARA, L. *et al*. Maxicircle (mitochondrial) genome sequence (partial) of *Leishmania major*: Gene content, arrangement and composition compared with *Leishmania tarentolae*. **Gene**, v. 424, n. 1–2, p. 80–86, 15 nov. 2008.

ZHANG, H. X.; ZHANG, Y.; YIN, H. Genome Editing with mRNA Encoding ZFN, TALEN, and Cas9. **Molecular Therapy**, v. 27, n. 4, p. 735–746, 10 abr. 2019.

ZHANG, W. W. *et al*. Comparison of the A2 gene locus in *Leishmania donovani* and *Leishmania major* and its control over cutaneous infection. **Journal of Biological Chemistry**, v. 278, n. 37, p. 35508–35515, 12 set. 2003.

ZHANG, W. W.; MATLASHEWSKI, G. Analysis of antisense and double stranded RNA downregulation of A2 protein expression in *Leishmania donovani*. **Molecular and Biochemical Parasitology**, v. 107, n. 2, p. 315–319, 15 abr. 2000.

ZHANG, W. W.; MATLASHEWSKI, G. CRISPR-Cas9-mediated genome editing in *Leishmania donovani*. **mBio**, v. 6, n. 4, 21 jul. 2015.

APÊNDICE I – *Scripts* modificados utilizados para anotação de famílias multigênicas em *L. amazonensis.*

Para a busca completa:

```
##########################################
# files and inputs:

BLASTN_path="/home/wanessa/bin/x86_64/ncbi-blast-2.8.1+/bin/blastn" #path to blastn program

makeblastdb_path="/home/wanessa/bin/x86_64/ncbi-blast-2.8.1+/bin/makeblastdb" #path to blastn program



genomefasta="input_files/PH8_genome.fasta" # genome to perform search in (fasta format)

genefamilyfasta='input_files/family_nucleotide_sequence.fasta' # a collection of family member genes

all_transcripts="input_files/TriTrypDB-52_L.amazonensis_AnnotatedTranscripts.fasta" #All transcritps in the genome
genefamilygff="input_files/PH8_family_genes.gff" # coordinates of known family member genes, use empty.gff if you don't want to provide this
#

#Cut-offs:

perc_identity_cutoff=85

mapHspGap=0

minTcTSLength=150

##########################################



#make blast database

printf "\n\nmaking blast database from fasta file...\n"

${makeblastdb_path} -dbtype nucl -in ${genomefasta}



#blast input genes to the genome

printf "\n\nBLASTing input genes to the genome..."
```

```
${BLASTN_path} -db ${genomefasta} -query ${genefamilyfasta} -out percid_${perc_identity_cutoff}.blastout -num_threads 1 -num_alignments 100 -max_hsps 100 -perc_identity ${perc_identity_cutoff}


#parse blast, get gene candidates

printf "\n\nparsing BLAST results, obtaining gene candidates..."

perl mark_homology_pieces.pl -i percid_${perc_identity_cutoff}.blastout -g ${genomefasta} -maxhspgap ${mapHspGap} -minTcTSLength ${minTcTSLength} -gff ${genefamilygff} -familyfasta ${genefamilyfasta} > log_of_gene_num.txt
rm percid_${perc_identity_cutoff}.blastout



#blast back to all transcripts

printf "\n\nmaking blast database from fasta file...\n"

${makeblastdb_path} -dbtype nucl -in ${all_transcripts}

printf "\n\nBLASTing gene candidates to all transcripts in the genome..."

${BLASTN_path} -outfmt 5 -db ${all_transcripts} -query percid_${perc_identity_cutoff}.blastout.mintctslengthcutoff${minTcTSLength}.maxgaplenintcts${mapHspGap}.fasta -out ${perc_identity_cutoff}perc_${minTcTSLength}minlen_${mapHspGap}maxgap.blast2transcripts.blastoutxml -num_threads 1 -num_alignments 50 -max_hsps 50


#parse blast

printf "\n\nparsing BLAST results, filtering gene candidates..."

python3 parse_blastout.py --blastout ${perc_identity_cutoff}perc_${minTcTSLength}minlen_${mapHspGap}maxgap.blast2transcripts.blastoutxml --fasta percid_${perc_identity_cutoff}.blastout.mintctslengthcutoff${minTcTSLength}.maxgaplenintcts${mapHspGap}.fasta
rm ${perc_identity_cutoff}perc_${minTcTSLength}minlen_${mapHspGap}maxgap.blast2transcripts.blastoutxml



#adjust boundary

printf "\n\nadjusting gene boundaries..."

python3 adjust_start_stop_codon.py --fastafile 85perc_150minlen_0maxgap.blast2transcripts.blastoutxml.parsedNfiltered
```

```
printf '\n\ndone\n'
```

Script parse_blastout.py

```
#!/usr/bin/python3

from Bio import SeqIO
from Bio.Blast import NCBIXML
import argparse
import sys
import re

parser = argparse.ArgumentParser(description='please specify input file.')
parser.add_argument('--blastout', help='blastout xml file')
parser.add_argument('--fasta', help='blast input fasta file, for retrieving best match information')


args = parser.parse_args()
args_dict=vars(args) # convert namespace(args) to dict style
input_file=args_dict['blastout'] #extract the file option value from dict
input_fastafile=args_dict['fasta'] #extract the file option value from dict
#print (input_file)
genesPassFilter=dict()
gene_anno_pattern=re.compile('mucin')
max_aln_num=2 # Configured according to the family

if len(sys.argv)==1: # print help message if arguments are not valid
   parser.print_help()
   sys.exit(1)

def main():

        try:

                with open(input_file, "r") as filehandle, open("{}.parsedNfilterout".format(input_file),'w') as
writefilehandle:

                        writelog=open("log.parseblast",'w')
                        blast_records = NCBIXML.parse(filehandle) # we have a pair of input functions, read
and parse, where read is for when you have exactly one object, and parse is an iterator for when you can have lots of
objects, but instead of getting SeqRecord or MultipleSeqAlignment objects, we get BLAST record objects.


                        for blast_record in blast_records:
                                break_flag=0
                                for alignment in blast_record.alignments:
                                        for hsp in alignment.hsps:
                                                #coverage=hsp.align_length/blast_record.query_length
```

```
                                        perc_iden=hsp.identities/hsp.align_length
                                        aligned_gene_name=alignment.title
                                        #print (aligned_gene_name)
                                        if not (re.search(gene_anno_pattern,aligned_gene_name)
is None): # match gene_anno_pattern
                                                genesPassFilter[blast_record.query]=1
                                                break
                                        else:
                                                writefilehandle.write(blast_record.query)
                                                writefilehandle.write("\n")
                                                writefilehandle.write(hsp.query)
                                                writefilehandle.write("\n")
                                                writefilehandle.write(aligned_gene_name)
                                                writefilehandle.write("\n")
                                                break
                                break_flag+=1
                                if (break_flag>=max_aln_num):
                                        break

        except Exception  as e:
                print("Unexpected error:", str(sys.exc_info()))
                print("additional information:", e)

        try:
                with open(input_fastafile, "r") as handle, open("{}.parsedNfiltered".format(input_file),'w') as
writefilehandle2:
                        fasta_sequences = SeqIO.parse(handle,'fasta')
                        for entry in fasta_sequences:
                                name,seq,desc = entry.id, str(entry.seq), str(entry.description)
                                #print (name)
                                if name in genesPassFilter.keys(): # ortholog does not exists, print seq
                                        writefilehandle2.write(">")
                                        writefilehandle2.write(name)
                                        writefilehandle2.write("\n")
                                        writefilehandle2.write(seq)
                                        writefilehandle2.write("\n")

                                else:
                                        writelog.write("{} have orthologs, excluded from
output\n".format(name))

        except Exception  as e:
                print("Unexpected error:", str(sys.exc_info()))
                print("additional information:", e)

if __name__ == "__main__": main()
```

APÊNDICE II – Arquivos de anotação de genes e proteínas da cepa PH8 de *Leishmania amazoenensis*.

Link:
https://drive.google.com/file/d/127upKf6IdBFRugiCiVFcq2epj2mwaTTH/view?usp=share_link

APÊNDICE III – Clusters de genes ortólogos compartilhados entre *L. amazonensis, L mexicana, L. donovani, L. infantum, L. major* e *L. braziliensis*.

Link: https://docs.google.com/spreadsheets/d/1dl13Zy0-1rmu2DjNK778JvSoKwVCQji9/edit?usp=share_link&ouid=111669099015029993699&rtpof=true&sd=true

APÊNDICE IV – Sequências repetitivas caracterizadas na cepa PH8 de *Leishmania amazonensis*.

Link:
https://docs.google.com/spreadsheets/d/1l3MJX5SZ1r9dXIRRPoNlcvL9xohDK8lQ/edit?usp=share_link&ouid=111669099015029993699&rtpof=true&sd=true

APÊNDICE V – Genes expandidos na cepa PH8 de *L. (L.) amazonensis*.

Link: https://docs.google.com/spreadsheets/d/1bnRc4-54hwHbz_n1WfuGv0vXbMGDV-kj/edit?usp=share_link&ouid=111669099015029993699&rtpof=true&sd=true