UNIVERSIDADE FEDERAL DE PERNAMBUCO – UFPE
CENTRO DE BIOCIÊNCIAS – CB
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS BIOLÓGICAS – PPGCB

LUCAS CHRISTIAN DE SOUSA PAULA

ANÁLISE TRANSCRIPTÔMICA DE *Leishmania infantum*, AGENTE ETIOLÓGICO DA LEISHMANIOSE VISCERAL AMERICANA

Recife
2018

LUCAS CHRISTIAN DE SOUSA PAULA

ANÁLISE TRANSCRIPTÔMICA DE *Leishmania infantum*, AGENTE ETIOLÓGICO DA LEISHMANIOSE VISCERAL AMERICANA

Dissertação apresentada ao Programa de Pós-Graduação em Ciências Biológicas, Área de Concentração Biotecnologia, da Universidade Federal de Pernambuco como requisito parcial para obtenção do título de Mestre em Ciências Biológicas.

Orientador: Prof. Dr. Valdir de Queiroz Balbino

Recife
2018

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD

Paula, Lucas Christian de Sousa
Análise transcriptômica de *Leishmania infantum,* agente etiológico da leishmaniose visceral americana / Lucas Christian de Sousa Paula. – 2018.

74 f. : il.

Orientador: Valdir de Queiroz Balbino.
Dissertação (mestrado) – Universidade Federal de Pernambuco. Centro de Biociências. Programa de Pós-graduação em Ciências Biológicas, Recife, 2018.
Inclui referências.

1. Leishmaniose visceral I. Balbino, Valdir de Queiroz (orient.) II. Título.

616.9364 CDD (22.ed.) UFPE/CB – 2018 - 306

Elaborado por Bruno Márcio Gouveia - CRB-4/1788

LUCAS CHRISTIAN DE SOUSA PAULA

ANÁLISE TRANSCRIPTÔMICA DE *Leishmania infantum*, AGENTE ETIOLÓGICO DA LEISHMANIOSE VISCERAL AMERICANA

Dissertação apresentada ao Programa de Pós-Graduação em Ciências Biológicas, Área de Concentração Biotecnologia, da Universidade Federal de Pernambuco como requisito parcial para obtenção do título de Mestre em Ciências Biológicas.

Aprovado em: 08 / 02 / 2018

COMISSÃO EXAMINADORA:

_____

Dr. Valdir de Queiroz Balbino (orientador)
Universidade Federal de Pernambuco - UFPE

_____

Dra. Ana Maria Benko Iseppon
Universidade Federal de Pernambuco - UFPE

_____

Dr. Carlos Henrique Nery Costa
Universidade Federal do Piauí – UFPI

Aos meus pais, que nunca deixaram de acreditar na tão desacreditada palavra educação.

AGRADECIMENTOS

Ao Professor Dr. Valdir de Queiroz Balbino, pela oportunidade, confiança e principalmente pelos grandes ensinamentos. Minha sincera gratidão!

Ao Laboratório de Bioinformática e Biologia Evolutiva (LABBE) do Departamento de Genética da UFPE, pela viabilização deste trabalho.

Aos colaboradores do Laboratório de Leishmanioses do Instituto de Doenças Tropicais Natan Portella da Universidade Federal do Piauí, em especial ao Dr. Carlos Henrique Nery Costa, por cederem as amostras analisadas neste trabalho.

Aos companheiros da sala de Bioinformática do LABBE: Wilson Júnior, Raul Falcão e Dr. Sérgio Paiva, pela imensa ajuda e ensinamentos durante toda execução deste trabalho, além do nosso convívio descontraído, sadio e pela amizade.

Aos colegas do LABBE, pelo convívio e respeito.

Ao meu grande amigo Dr. Timbó Braga, que desde minha graduação tem ajudado, aconselhado e apoiado. Minha eterna gratidão!

À Capes e à Rede INTERSYS (23038.010050/2013-04) pelo apoio financeiro.

À Lidiane Gomes, felicidade dos meus dias contemporâneos, por toda cumplicidade, compreensão, apoio e respeito.

Aos meus pais, que sempre acreditaram na tão desacreditada palavra educação e sempre tentaram instruir-me da melhor maneira possível.

Aos meus irmãos, que de decerto, foram determinantes nos caminhos até então trilhados por mim.

À Deus, por todo conforto!

“A morte de um homem é uma tragédia; a de milhões, uma estatística.”

Josef Stalin.

RESUMO

Leishmania infantum é um protozoário parasita de grande relevância médica, por ser o agente etiológico da leishmaniose visceral, uma grave doença que pode ser fatal quando não tratada. Desde a última década, muita informação tem sido gerada acerca da genômica de tripanossomatídeos, o que tem possibilitado um melhor conhecimento dos mecanismos moleculares utilizados por estes protozoários para infectar seus hospedeiros. *L. infantum* teve seu genoma sequenciado em 2007, entretanto, análises de expressão gênica deste parasito têm sido limitadas ao uso de técnicas com baixa acurácia. O RNA-seq é uma ferramenta para análise de transcriptomas inovadora, com maior acurácia, sensibilidade e custo-benefício relativamente baixo, quando comparado às demais técnicas. Até a presente data, não há relatos na literatura de estudos que tenham utilizado RNA-seq para analisar o transcriptoma de *L. infantum*. Em vista disso, nós analisamos o transcriptoma de oito cepas de *L. infantum* utilizando RNA-seq e estratégia de montagem *de novo*. Inicialmente, nós realizamos comparações de seis softwares montadores de transcriptoma *de novo* (rnaSPAdes; IDBA-tran; Oases; Trinity; Trans-ABySS; SOAPdenovo-Trans) e cinco estratégias de tratamento de *short reads* (*pipelines* – P_0; P_1; P_2; P_3; P_4). Os montadores rnaSPAdes, Trinity e IDBA-tran obtiveram os melhores resultados para os critérios avaliados; enquanto que o P_0 obteve os melhores resultados dentre as estratégias de tratamento abordadas. Posteriormente, optando pelo montador rnaSPAdes, nós montamos *de novo* oito cepas de *L. infantum* cepa Teresina. Uma média de 26.410 transcritos foram montados, cerca de 30% destes foram classificados como não-anotados em relação ao genoma de referência; enquanto que 0,5 a 6,6% foram classificados como não-alinhados. Estes transcritos foram utilizados para predição de genes e um total de 88 novos possíveis genes foram preditos para *L. infantum*. Além disso, pode-se observar que mais de 90% do genoma foi constitutivamente expresso e que dentre os genes mais abundantes, foram detectados àqueles condificantes de proteínas de histonas, β-tubulina e α-tubulina. Nossas análises poderão servir como base para novos estudos acerca dos mecanismos de expressão gênica de *L. infantum*, bem como outros tripanossomatídeos.

Palavras-chave: RNA-seq. Montagem *de novo*. Trypanosomatidae. Sequenciamento de nova geração.

ABSTRACT

Leishmania infantum is a parasite protozoan of great medical relevance, due to be visceral leishmaniasis causative agent, a severe disease that may be lethal when untreated. Since last decade, many information has been generated about trypanosomatids genomics, that has made possible a better knowledge of molecular mechanisms used by these protozoans for to infect their host. *L. infantum* genome was sequenced in 2007, however, expression genic analyses of this parasite have been limited to the use of low accuracy techniques. RNA-seq is a revolutionary tool for transcriptome analyses, with greater accuracy, sensitivity and cost-benefit, when compared to the other ones. To date, there are no records on literature about studies that employed RNA-seq for *L. infantum* transcriptome analyses. Therefore, we analysed *L. infantum* transcriptome from eight strains using RNA-seq and *de novo* assembly strategy. Initially, we carried out comparisons among six *de novo* transcriptome assemblers software (rnaSPAdes; IDBA-tran; Oases; Trinity; Trans-ABySS; SOAPdenovo-Trans) and five short reads strategies of treatment (pipelines – P_0; P_1; P_2; P_3; P_4). rnaSPAdes, Trinity and IDB-tran assemblers, presented the best results for evaluated criteria, whereas the P_0 showed the best results among pipelines. Posteriorly, using rnaSPAdes, we *de novo* assembled eight *L. infantum* Teresina strain. An average of 26,410 transcripts were assembled, of these around 30% were classified as unannotated and between 0.5 to 6.6 were unaligned. These transcripts were utilized for novel genes prediction and a total of 88 novel putative genes were predicted for *L. infantum*. Therefore, was observed that more than 90% of *L. infantum* genes were expressed constitutively and among most abundant genes, were detected histones, β-tubulin and α-tubulin genes. Our analyses will serve as basis for future studies about *L. infantum* genic expression mechanism, as well as another trypanosomatids.

Key words: RNA-seq. *De novo* assembly. Trypanosomatidae. Next-generation sequencing.

LISTA DE ILUSTRAÇÕES

REVISÃO DA LITERATURA

ARTIGO 1

ARTIGO 2

LISTA DE TABELAS

ARTIGO 1

ARTIGO 2

LISTA DE ABREVIATURAS E SIGLAS

%	Porcentagem
~	Aproximadamente
<	Menor que
>	Maior que
≥	Igual ou maior que
α	Alfa
β	Beta
bp	Pares de base (*base pairs*)
CA	Califórnia
CEP	Comitê de ética na pesquisa
DE	Alemanha (*Germany*)
DNA	Ácido desoxirribonucleico (*desoxyribonucleic acid*)
e.g.	Por exemplo (*exempli gratia*)
et al.	Colaboradores
Fig.	Figura (*Figure*)
FPKM	*Fragments per kilobase of transcript per million mapped reads*
FTC	Fatal case
GC%	Percentual de guanina e citosina
HSP	Proteína de choque térmico (*heat shock protein*)
i.e.	Isto é (*id est*)
IDTNP	Instituto de Doenças Tropicais Natan Portela
kDNA	*Kinetoplast DNA*
k-mer	Comprimento de *k*
LaCTAD	Laboratório Central de Tecnologias de Alto Desempenho em Ciências da Vida
LPG	Lipofosfoglicano (*lipophosphoglycan*)

LT	Leishmaniose tegumentar
LV	Leishmaniose visceral
LVH	Leishmaniose visceral humana
Mb	Megabase (1 milhão de pares de base)
MP	Matriz peritrófica
NaCl	Cloreto de sódio
NCBI	*National Center for Biotechnology Information*
NFT	*Non-fatal cases*
NGS	Sequenciamento de nova geração (*Next-generation sequencing*)
nm	Nanômetro
NR	Banco de dados não-redundante (*Non-redundant database*)
P	*Pipeline*
pH	Potencial hidrogeniônico
PKC	Proteína quinase C (*protein kinase C*)
RNA	Ácido ribonucleico (*ribonucleic acid*)
RNA-seq	Sequenciamento de RNA (*RNA sequencing*)
rpm	Rotações por minuto
SINAN	Sistema de Informação de Agravos de Notificação
spp.	Espécies (*species*)
SRA	*Sequence Read Archive*
UTR	Região não-traduzida (*untranslated region*)

SUMÁRIO

1 INTRODUÇÃO

As leishmanioses são um complexo grupo de doenças tropicais negligenciadas que representam um importante problema de saúde pública que afeta a vida de milhões de pessoas em todo o mundo. A leishmaniose visceral (LV), que pode ser letal quando não tratada, registra entre 200 a 400 mil novos casos por ano, concentrando-se principalmente em países subdesenvolvidos e/ou em desenvolvimento. No Brasil, país com dimensões continentais, há uma estimativa de que até 6.300 novos casos ocorram a cada ano, concentrados principalmente na região Nordeste, com mais de 50% das notificações registradas. Considerando o período de 2001 a 2015, os estados do Ceará, Maranhão, Piauí e Bahia encerraram mais de 80% e quase 50% das notificações na região Nordeste e no Brasil, respectivamente. A LV é causada nas Américas pelo protozoário *Leishmania infantum*, que é transmitido por fêmeas infectadas de dípteros flebotomíneos hematófagos do gênero *Lutzomyia*.

A *L. infantum* é membro da família Trypanosomatidae, que obrigatoriamente agrupa organismos parasitas de uma grande diversidade de animais silvestres, sinantrópicos e domésticos, incluindo o homem e o cão. Este parasito possui um ciclo de vida digenético (*i.e.,* precisa de dois hospedeiros para completar seu ciclo biológico) bastante complexo, alternando entre hospedeiros intermediários (invertebrados) e definitivos (mamíferos). Além disso, no que diz respeito ao genoma de *L. infantum*, este protozoário apresenta genes organizados em longos *clusters* policistrônicos, quase completa ausência de íntrons e algumas características que para eucariotos são únicas ou mais comumente observadas na ordem Kinetoplastida.

Ao longo dos anos, diversos estudos têm possibilitado entender os mecanismos utilizados pela *L. infantum*, assim como por outras espécies da família Trypanosomatidae, para contornar as barreiras fisiológicas encontradas nos hospedeiros intermediários e definitivos. A descoberta e elucidação de estruturas e estratégias utilizadas por estes protozoários pavimentam o desenvolvimento de novas ferramentas para tratamento, profilaxia e controle das leishmanioses.

Na última década, com o advento da tecnologia das *ômicas*, tornou-se possível entender como parasitas interagem com seus hospedeiros a nível molecular, observando seus genomas e como estes são expressos em condições específicas. A partir de 2005, os primeiros genomas de *Leishmania* spp. e de outros

tripanossomatídeos foram sequenciados, tornando-se um marco histórico que tem impulsionado diversos estudos da biologia destes parasitos causadores de doenças.

Seguindo os trabalhos de genômica, a transcriptômica busca estudar quais genes estão sendo expressos em determinada condição e/ou momento metabólico. Para algumas espécies de *Leishmania*, estes trabalhos iniciaram-se mesmo antes do sequenciamento dos seus genomas. Entretanto, com a chegada da segunda geração de sequenciadores, os sequenciadores de nova geração (NGS), tornou-se possível sequenciar RNA (RNA-seq), o que tem permitido não somente identificar quais genes são expressos, mas também quantificá-los, descobrir sua precisa sequência nucleotídica, identificar regiões não-traduzidas (UTRs), sítios de *trans-splicing*, outros tipos de RNAs não-codificantes, novas isorformas e até mesmo utilizar dados de RNA-seq para corrigir montagens e anotações de genomas de referência.

Os avanços supracitados são possíveis graças a diversas áreas do conhecimento, dentre estas a bioinformática, uma ciência interdisciplinar que une, basicamente, as grandes áreas de Ciências Biológicas e Informática. Paralelamente ao avanço das *ômicas*, diversos *softwares* têm sido desenvolvidos e aprimorados para melhorar a acurácia de montagem e análise de sequenciamento de nucleotídeos. Não obstante, estabelecer um *pipeline* padrão para quaisquer trabalhos de bioinformática é uma missão difícil, tendo em vista que, a depender dos objetivos e organismo de interesse, a abordagem utilizada pode variar. Todavia, estabelecer comparações entre estratégias de montagens pode auxiliar novos usuários a definir qual melhor *pipeline* a seguir, ainda mais em estudos com tripanossomatídeos, que possuem um alto grau de sintenia entre si.

Após 10 anos ao sequenciamento do genoma da *L. infantum*, muita informação tem sido gerada, principalmente, acerca da genômica deste organismo; entretanto, ainda não fora utilizada a tecnologia de RNA-seq para analisar o transcriptoma total, o que poderá auxiliar em um melhor entendimento da biologia molecular deste parasito. Desta maneira, o uso do RNA-seq permitirá analisar de forma panorâmica os genes expressos da *L. infantum*, o que poderá subsidiar novos alvos para tratamento e controle da leishmaniose visceral.

2 REVISÃO DA LITERATURA

2.1 *Leishmania* spp. E AS LEISHMANIOSES

As leishmanioses são um complexo grupo de doenças negligenciadas que representam importante problema de saúde pública e possuem forte ligação com condições de pobreza (DESJEUX, 2004; ALVAR; YACTAYO; BERN, 2006). Este grupo compreende duas principais doenças, as leishmanioses tegumentar (LT) e visceral (LV) (ALVAR; YACTAYO; BERN, 2006; BERN; MAGUIRE; ALVAR, 2008), que manifestam distintas formas clínicas, desde úlceras cutâneas localizadas até manifestações sistêmicas letais (KAMHAWI, 2006).

Tais patologias são causadas por protozoários cinetoplastídeos do gênero *Leishmania* (Ross, 1903), pertencentes à família Trypanosomatidae, sendo todos os membros desta família obrigatoriamente organismos parasitas de insetos, mamíferos e outros (MASLOV; PODLIPAEV; LUKEŠ, 2001).

Possuindo pelo menos 23 espécies patogênicas de humanos e uma grande diversidade fenotípica, o gênero *Leishmania* é ainda subdivido em quatro subgêneros: *Leishmania*; *Viannia*; *Sauroleishmania*; e, o mais recentemente descrito, *Mundinia* (ESPINOSA et al., 2016), *Leishmania* e *Viannia* são os principais causadores de antropozoonoses, diferindo entre si, principalmente, pela região onde se multiplicam no trato digestivo do inseto vetor (LAINSON; SHAW, 1987). Os quatro subgêneros agrupam cerca de 52 diferentes espécies (MASLOV; PODLIPAEV; LUKEŠ, 2001; AKHOUNDI et al., 2016), dentre elas a *Leishmania (Leishmania) infantum*, agente etiológico da LV no continente americano.

A dispersão das espécies de *Leishmania* está intimamente associada à distribuição geográfica do seu vetor, os flebotomíneos. Ambos subgêneros, *Leishmania* e *Viannia*, são reportados e transmitidos no Novo Mundo por fêmeas de flebotomíneos do gênero *Lutzomyia* (DANTAS-TORRES et al., 2012).

2.2 EPIDEMIOLOGIA DA LEISHMANIOSE VISCERAL HUMANA

A leishmaniose visceral humana (LVH) possui manifestações clínicas e sintomatologias variáveis, acometendo principalmente órgãos e vísceras do sistema linfohematopoiético (*e.g.*, fígado, baço e medula óssea), podendo ser endêmica,

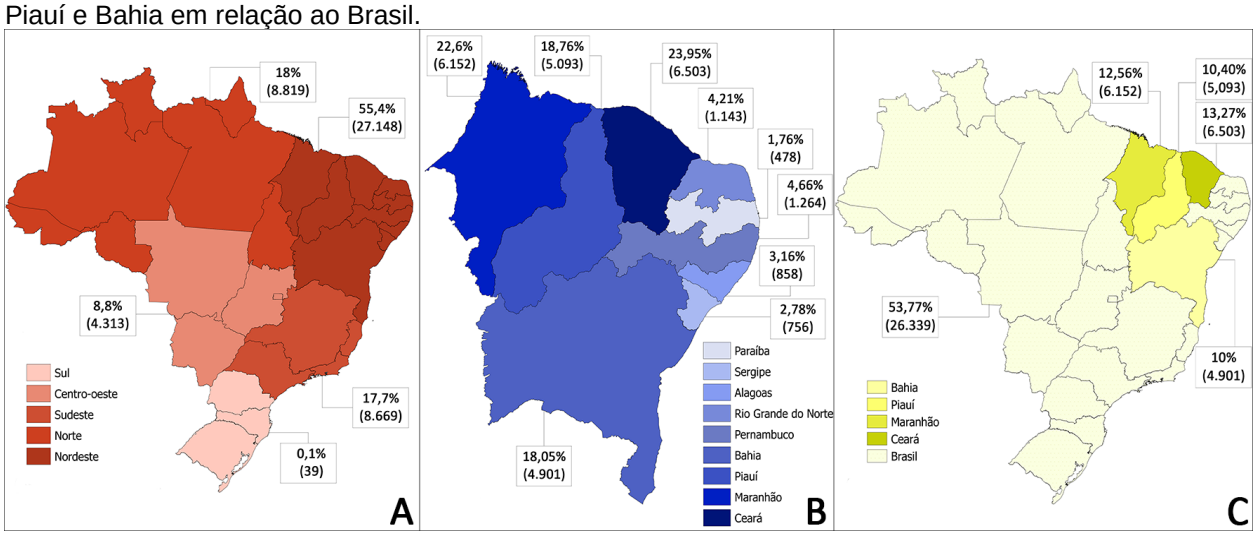
esporádica ou epidêmica (HERWALDT, 1999; MURRAY et al., 2005; CONCEIÇÃO-SILVA; ALVES, 2014).

No continente americano, a LVH é causada pela *L. infantum*, podendo ainda haver outras espécies incriminadas como causadoras da doença; o díptero flebotomíneo *Lutzomyia longipalpis* é unanimemente apontado como seu principal vetor (DANTAS-TORRES, 2009). A ocorrência da doença em determinada região necessita da presença do hospedeiro intermediário susceptível e do hospedeiro vertebrado igualmente susceptível (GONTIJO; MELO, 2004)

Por ser uma doença de populações negligenciadas, a LVH possui ligação direta com problemas socioambientais, tais como: pobreza; ocupação urbana não planejada; destruição ambiental; condições precárias de saneamento e habitação; e desnutrição (WERNECK, 2010). Estima-se que esta patologia acometa entre 200 mil a 400 mil pessoas por ano em todo o mundo, concentrando-se principalmente em países subdesenvolvidos e/ou em desenvolvimento, com 90% dos novos casos registrados em Bangladesh, Brasil, Etiópia, Índia, Sudão e Sudão do Sul (DESJEUX, 2004; MURRAY et al., 2005; ALVAR; YACTAYO; BERN, 2006; ALVAR et al., 2012).

No Brasil, estima-se que ocorram entre 4.200 e 6.300 novos casos por ano (ALVAR et al., 2012). Todas as regiões brasileiras possuem registro para LVH, destacando-se pela importância as regiões Norte, Nordeste e Sudeste que concentram mais de 90% das notificações (FIGURA 1-A). Segundo dados oficiais disponibilizados pelo Sistema de Informação de Agravos de Notificação – SINAN, contabilizando o período de 2001 a 2015, mais da metade (55,4%) dos casos brasileiros foram registrados somente na região Nordeste (FIGURA 1-A). Os Estados do Ceará, Maranhão, Piauí e Bahia representam juntos mais de 80% das incidências na região Nordeste e quase metade (46,23%) das notificações no país (FIGURA 1-B e 1-C) (BRASIL, 2017).

Figura 1 – Casos notificados de leishmaniose visceral humana no Brasil durante os anos de 2001 a 2015, segundo o Sistema de Informação de Agravos de Notificação – SINAN. Em **A**: Porcentagem de casos de LVH por Região em relação ao Brasil; **B**: Porcentagem de casos de casos de LVH por Estado em relação a Região Nordeste; **C**: Porcentagem de casos de LVH nos estados do Ceará, Maranhão, Piauí e Bahia em relação ao Brasil.



Fonte: O Autor (2018).

Devido a estes números alarmantes, desperta-se grande interesse e necessidade de estudos para desenvolvimento de abordagens alternativas que auxiliem a compreensão da epidemiologia da LVH, bem como um melhor entendimento do ciclo biológico da doença, assim como a elucidação dos mecanismos de interação parasito-hospedeiros, o que, consequentemente, poderá subsidiar o controle desta antropozoonose.

2.3 *Leishmania infantum*

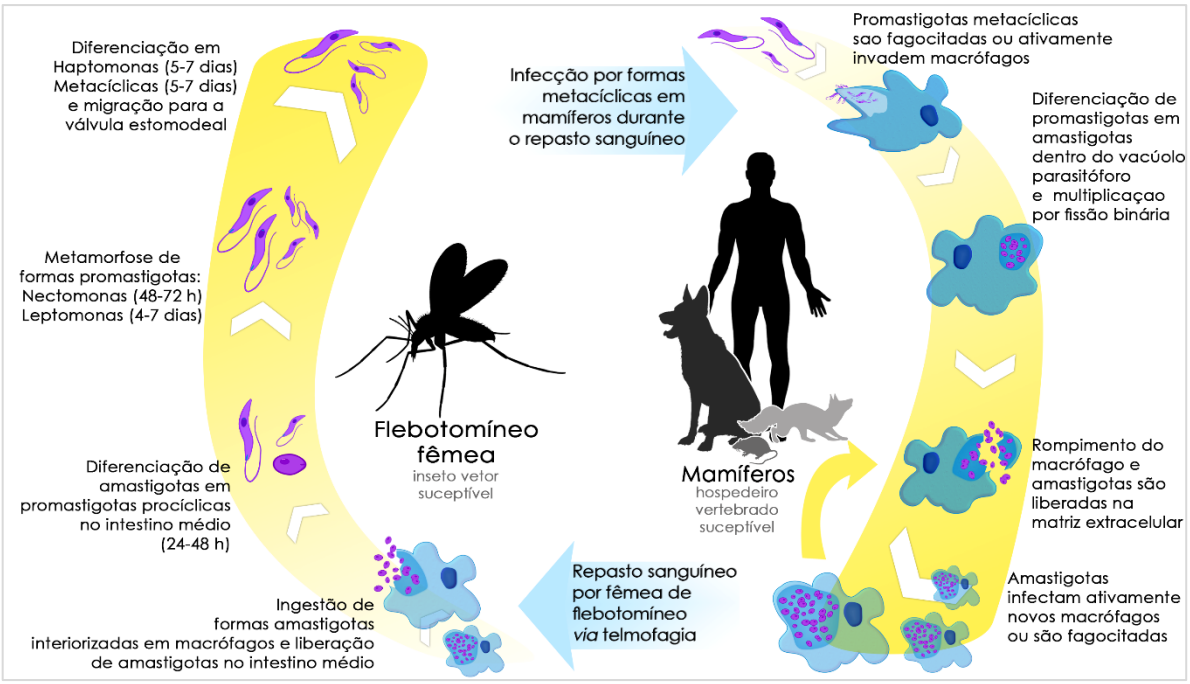
2.3.1 Biologia da Leishmania infantum

A *L. infantum* é um organismo diploide, sem ciclo sexual aparente e que apresenta algumas características, tais como: organização única do DNA mitocondrial [o cinetoplasto (*Kinetoplast DNA* - kDNA)]; glicossomos; transcrição policistrônica e mecanismos de *trans-splicing*, que para eucariotos são únicos ou mais comumente utilizadas pela ordem Kinetoplastida (CAMPBELL; THOMAS; STURM, 2003; ULIANA; RUIZ; CRUZ, 2008).

A *L. infantum*, bem como outras espécies do gênero, apresenta um ciclo de vida digenético bastante complexo que abrange hospedeiros definitivos (vertebrados) e hospedeiros intermediários (insetos vetores) em dois estágios básicos de

desenvolvimento: (1) promastigotas, a forma proliferativa encontrada dentro do trato digestivo de fêmeas de flebotomíneos infectadas; e (2) amastigotas, a forma proliferativa encontrada dentro de vários tipos de células de mamíferos (LEIFSON et al., 2007; TEIXEIRA et al., 2013), principalmente macrófagos, que são a primeira linha de defesa do hospedeiro (FIGURA 2) (PATINO; RAMÍREZ, 2017). Por ser membro do subgênero *Leishmania*, a *L. infantum* é tido como um parasita suprapilariano, estando o seu desenvolvimento restrito ao intestino médio dos flebotomíneos (LAINSON; SHAW, 1987; KAMHAWI, 2006).

Figura 2 - Ciclo de vida digenético da *Leishmania infantum*, alternando entre hospedeiros mamíferos e fêmeas de flebotomíneos.



Fonte: O Autor.

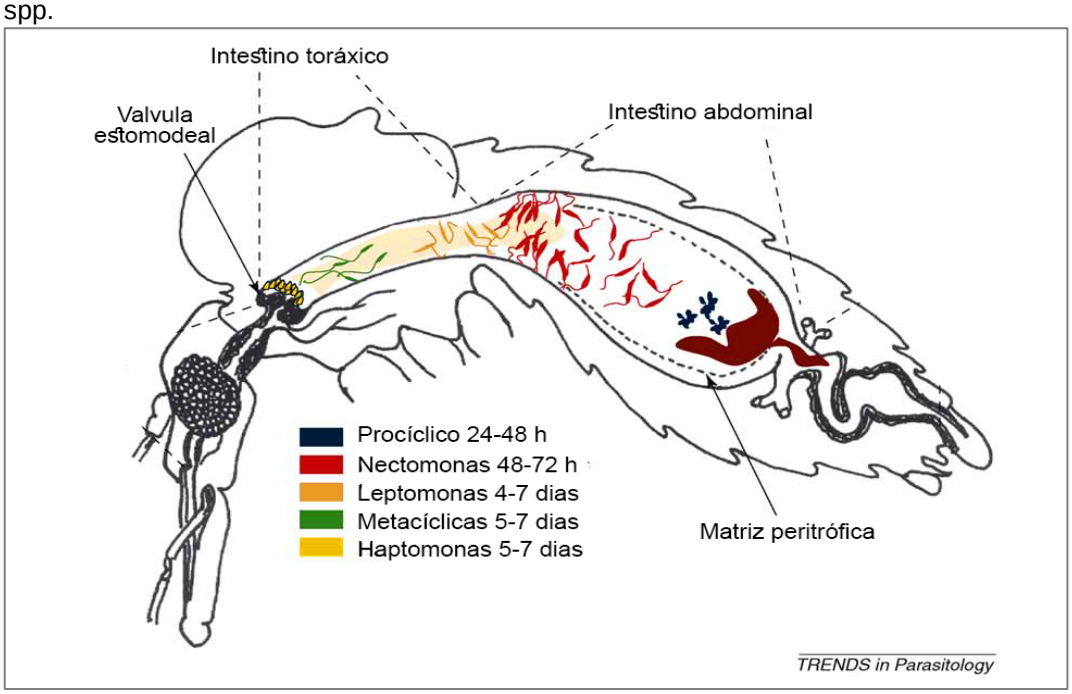
2.3.2 Interação parasito-hospedeiro intermediário

A transmissão da *L. infantum*, de um vetor infectado para um hospedeiro vertebrado susceptível, requer o desenvolvimento do parasito no intestino de um flebotomíneo competente (GONTIJO; MELO, 2004; CANTACESSI et al., 2015). Durante o repasto sanguíneo, *L. infantum* amastigotas, livres ou interiorizadas em macrófagos, são ingeridas e liberadas no intestino do flebotomíneo, possibilitando seu desenvolvimento através de vários estágios (Figura 3) até sua forma infectante (KIMA, 2014).

Após o repasto sanguíneo, as formas amastigotas presentes no sangue ingurgitado são envolvidas pela matriz peritrófica (MP) dentro do tubo digestivo do flebotomíneo (PIMENTA et al., 1997), onde diferenciam-se em formas com flagelos curtos e de rápida multiplicação, as promastigotas procíclicas (KAMHAWI, 2006) (FIGURA 2). Por sua vez, formas procíclicas se transformam em promastigotas nectomonas, que possuem flagelos mais longos e são mais hábeis, o que possibilita o parasito desvencilhar-se da MP e afixar-se no epitélio intestinal povoando o lúmen intestinal (KAMHAWI, 2006). A sobrevivência dos parasitos e, consequentemente, o progresso do ciclo de infecção/transmissão depende do sucesso deste processo logo após a degradação da MP e liberação do bolo fecal (TEMPONE; PITALUGA; TRAUB-CSEKÖ, 2014).

No lúmen intestinal, os parasitos assumem outra forma proliferativa, as promastigotas leptomanas, que constituirão a maior parte da população do parasito no intestino do flebotomíneo (WILSON et al., 2010). Após alguns dias,, formas promastigotas migram para a válvula estomodeal, onde é possível identificar duas formas distintas: promastigotas haptomonas; e metacíclicas (KAMHAWI, 2006). As haptomonas possuem flagelos mais curtos, pouca mobilidade e são responsáveis por formar um agregado de parasitos na válvula estomodeal; já as metacíclicas são o estágio infectivo e possuem um longo flagelo, bastante rapidez e são resistentes à fisiologia de defesa do hospedeiro mamífero (HANDMAN; ELSO; FOOTE, 2005; KAMHAWI, 2006; WILSON et al., 2010). Durante um novo repasto sanguíneo realizado por uma fêmea de flebotomíneo infectada, formas metacíclicas são regurgitadas no hospedeiro vertebrado, dando continuidade ao ciclo de infecção da *L. infantum*.

Figura 3 - Estágios de desenvolvimento da *Leishmania infantum* em fêmeas de *Lutzomyia* spp.



Fonte: Adaptado de Kamhawi (2006).

2.3.3 Interação parasito-hospedeiro vertebrado

Uma vez dentro da epiderme do hospedeiro vertebrado, algumas promastigotas metacíclicas interagem principalmente com macrófagos e são internalizadas em vacúolos parasitóforos, onde diferenciam-se intracelularmente em formas amastigotas (ALEXANDER; RUSSELL, 1992). Então, amastigotas se multiplicam por fissão binária até forçar o rompimento do macrófago e são liberadas para infectar outras células (ALCOLEA et al., 2010).

As amastigotas podem ser ingeridas em formas livres, ou interiorizadas em células do hospedeiro, por fêmeas de flebotomíneos durante o repasto sanguíneo (HANDMAN; ELSO; FOOTE, 2005).

2.3.4 Mecanismos moleculares de sobrevivência da Leishmania infantum

Devido às diferenças biológicas encontradas pela *L. infantum* e outras espécies do gênero em seus distintos hospedeiros, mecanismos de adaptação têm sido desenvolvidos a fim de sobrepujar tais barreiras fisiológicas (WRIGHT; SILVERSTEIN, 1983; ROSENTHAL et al., 1996; TEMPONE; PITALUGA; TRAUB-CSEÖ, 2014). Essas adaptações podem envolver expressão de genes que possibilitem ao parasito

escapar da resposta imune do hospedeiro ou que evitem a ação de enzimas digestivas do vetor (PETERS; SACKS, 2006).

Como o primeiro contato da *L. infantum* com seus hospedeiros ocorre com sua superfície celular, a expressão de genes relacionados ao seu glicocálice está diretamente ligada à sobrevivência e consequente sucesso da infecção (ALVES et al., 2014). Um desses principais componentes, o lipofosfoglicano (LPG - *lipophosphoglycan*), possui estrutura e expressão variável entre promastigotas procíclicas e metacíclicas de *L. infantum*, o que confere interações diferentes com seu vetor natural *Lu. longipalpis*, *i.e.,* a necessidade de se afixar ao epitélio intestinal por parte das formas procíclicas e a não-necessidade das formas metacíclicas (SOARES et al., 2002; DE ASSIS et al., 2012). Além disso, no hospedeiro vertebrado, o LPG possui função inibitória sobre a atividade da proteína quinase C (PKC) de macrófagos, uma vez que essas enzimas possuem a função de ativar células deste tipo (TURCO, 1999).

Além disso, promastigotas metacíclicas possuem uma proteína quinase serina/treonina, que inativa os componentes do sistema complemento imunológico do hospedeiro C3, C5 e C9 por fosforilação. Ao mesmo tempo que expressam uma proteinase denominada gp63 (também chamada leishmanolisina – *leishmanolysin*) que converte a opsonização do fator de complemento C3b , o que facilita a entrada do parasito nas células do hospedeiro vertebrado (KUMAR, 2013).

Um outro mecanismo de sobrevivência que tem sido observado em algumas espécies de *Leishmania*, é a secreção de um neuropeptídio inibitório de movimentos peristálticos, que pode interromper total ou parcialmente o peristaltismo do tubo digestivo dos flebotomíneos (VAIDYANATHAN, 2005), o que certamente reduz as chances de eliminação do parasito junto ao bolo alimentar.

2.4 GENÔMICA E TRANSCRIPTÔMICA APLICADAS AO ESTUDO DE *Leishmania infantum* E OUTRAS ESPÉCIES DE *Leishmania*

Avanços recentes na área da genômica e transcriptômica de tripanossomatídeos, bem como outras tecnologias das *ômicas*, juntamente com o barateamento do custo destas técnicas, têm revolucionado a pesquisa biológica na academia e na indústria, propiciando o desenvolvimento de estudos que possibilitam

revelar íntimos detalhes da relação molecular entre *Leishmania* spp. e seus hospedeiros (CALVETE; DOMONT, 2011; CANTACESSI et al., 2015).

2.4.1 Genômica

O primeiro sequenciamento de genoma de espécies de *Leishmania* foi divulgado no ano de 2005, com a publicação da descrição do genoma completo de *L. major* (IVENS et al., 2005), seguido pelos de *L. infantum* e *L. braziliensis* (PEACOCK et al., 2007), *L. mexicana* (ROGERS et al., 2011), *L. donovani* (DOWNING et al., 2011), *L. tarentolae* (RAYMOND et al., 2012) e *L. amazonensis* (REAL et al., 2013); sendo *L. donovani* a primeira espécie de *Leishmania* a ter o genoma sequenciado utilizando tecnologia NGS (*Next-generation sequencing*) (DOWNING et al., 2011; CANTACESSI et al., 2015).

Atualmente existe uma base de dados moleculares específica para tripanossomatídeos, o TriTrypDB (http://tritrypdb.org/tritrypdb) (ASLETT et al., 2009). No TriTrypDB, é possível encontrar, além dos genomas completos acima citados, *drafts* (*i.e.*, genomas incompletos) de outras espécies de *Leishmania* (LEPROHON et al., 2015). Além de informações de genômica, o TriTrypDB inclui também em sua base de dados informações disponíveis na literatura acerca de estudos transcriptômicos e proteômicos de tripanossomatídeos.

O genoma de *L. infantum* possui cerca de 33 Mb de extensão, distribuídas em 36 cromossomos (PEACOCK et al., 2007). É caracterizado pela quase completa ausência de íntrons e pela alta densidade de genes organizados em longos arranjos de *clusters* policistrônicos (ULIANA; RUIZ; CRUZ, 2008). Em 2007, através de sequenciamento utilizando abordagem de *shotgun*, *L. infantum* teve seu genoma completo publicado; foram montados 562 *contigs*, anotados inicialmente 8.154 genes codificantes e 41 pseudogenes (PEACOCK et al., 2007). Em 2011, o genoma de *L. infantum* foi ressequenciado utilizando tecnologia NGS, atualizando o número de genes codificantes de proteínas preditos para 8.241 (ROGERS et al., 2011).

Atualmente, no TriTrypDB há 8.381 genes anotados para *L. infantum* (versão 9.0). Apesar desse grande número de genes anotados, cerca de 60% possuem função conhecida e somente 19 são espécie-específicos para *L. infantum* (IVENS et al., 2005; PEACOCK et al., 2007; ROGERS et al., 2011). Deste modo, a grande maioria de suas sequências nucleotídicas codificantes são compartilhadas com outras espécies do

gênero (SMITH; PEACOCK; CRUZ, 2007) e até com outros membros da família Trypanosomatidae (EL-SAYED et al., 2005b; PEACOCK et al., 2007; SMITH; PEACOCK; CRUZ, 2007; ROGERS et al., 2011). Além do mais, estes estudos apontam, ainda, que a arquitetura do genoma (sintenia) ao longo da evolução entre este grupo de parasitos se mantém conservada.

Um ponto relevante e ao mesmo tempo desafiador, no que diz respeito à genômica destes parasitas, é identificar as várias cópias idênticas de genes organizados em tandem dentro do genoma (EL-SAYED et al., 2005a). Diversas proteínas altamente expressas possuem múltiplas cópias de genes, por exemplo, tubulinas, *heat shock proteins* (HSP), proteases e antígenos de membrana (ALONSO et al., 2016), estando a maioria destas relacionadas com a infectividade e virulência das *Leishmania* spp.

2.4.2 Transcriptômica

Nos últimos anos, diversas tecnologias têm sido desenvolvidas e aplicadas para análise de transcriptomas (PATINO; RAMÍREZ, 2017). Alguns dos principais objetivos destas técnicas são a identificação dos genes encontrados, em alguns casos a determinação da sequência nucleotídica precisa e sua abundância em diferentes condições (*e.g.,* estágio de desenvolvimento; presença/ausência de agentes inibitórios; alterações de pH; mimetização de *micro-habitat*) (MCGETTIGAN, 2013). Em *Leishmania* spp., estas análises ocorriam antes dos sequenciamentos dos genomas, utilizando-se principalmente da técnica de microarranjos (SAXENA et al., 2003; HOLZER; MCMASTER; FORNEY, 2006).

Abordagens transcriptômicas são particularmente importantes em organismos cuja regulação gênica é pós-transcricional (HAILE; PAPADOPOULOU, 2007), pois tem sido observado que diferenças em regiões UTR (*untranslated region*) do mesmo gene estão associadas com diferenças de expressão em diferentes estágios do ciclo de vida de tripanossomatídeos (MURRAY et al., 2007; FIEBIG; KELLY; GLUENZ, 2015).

Analisando transcriptoma e proteoma de *L. infantum*, foi possível observar que o genoma deste parasito é essencialmente expresso tanto em formas amastigotas, quanto em promastigotas (LEIFSO et al., 2007; MCCONVILLE et al., 2007). Por outro lado, os níveis de expressão de determinados genes podem ser modificados a

depender da etapa do ciclo de vida que se encontra o parasito (MCNICOLL et al., 2006).

Alcolea et al. (2009) analisaram o perfil de expressão diferencial de promastigotas procíclicas e metacíclicas de *L. infantum* de uma mesma população (separando-as através de técnicas de centrifugação) via técnica de microarranjos, e puderam observar que formas metacíclicas super-regulam vários genes envolvidos na biossíntese de lipofosfoglicanos, proteofosfoglicanos e glicoproteínas, todas estruturas relacionadas com a infectividade.

Entretanto, o uso das técnicas supracitadas pode dificultar, ou até mesmo não permitir, a detecção de novos transcritos, genes subexpressos, e outros produtos pós-transcricionais (WANG; GERSTEIN; SNYDER, 2009).

Outros trabalhos no âmbito da transcriptômica envolvendo *L. infantum* têm sido relatados na literatura, todavia voltados para a interação patógeno-hospedeiro, interessados na expressão de genes de fêmeas de flebotomíneos infectadas ou não-infectadas (JOCHIM et al., 2008).

Não obstante, com a chegada do advento da tecnologia NGS, o uso de sequenciamentos de RNA (RNA-seq – *RNA Sequencing*) tem-se mostrado promissor e sobressaindo-se em diversos aspectos às demais técnicas anteriormente aplicadas para análise de transcriptomas (WANG; GERSTEIN; SNYDER, 2009). Quando RNA-seq é utilizado, é possível identificar não somente a porcentagem dos genes diferencialmente expressos, mas também a identificação e análise de novos transcritos, RNAs não-codificantes, além de regiões UTR (WANG; GERSTEIN; SNYDER, 2009; PATINO; RAMÍREZ, 2017). Recentemente, Alonso et al. (2016) utilizaram dados de RNA-seq para reconstruir algumas regiões não-montadas do genoma de referência de *Le. major* (cepa Friedlin).

Abordagens utilizando RNA-seq têm sido empregadas para várias espécies de *Leishmania* (MITTRA; ANDREWS, 2013; RASTROJO et al., 2013; MARTIN et al., 2014; DILLON et al., 2015; FIEBIG; KELLY; GLUENZ, 2015; ALONSO et al., 2016). Por outro lado, até então não há publicações com o uso desta tecnologia para analisar o transcriptoma total de *L. infantum*.

2.5 BIOINFORMÁTICA

Com o aumento na demanda para análises de dados gerados a partir de NGS, diversos *softwares* têm sido desenvolvidos para avaliação da qualidade do sequenciamento (ANDREWS, 2010); tratamento das sequências geradas (JOSHI; FASS, 2011; GORDON; HANNON; GORDON, 2014); montagem de genomas e transcriptomas (BANKEVICH et al., 2012; HAAS et al., 2013); e avaliação qualitativa das montagens (LI et al., 2014; BUSHMANOVA et al., 2016).

Quando um genoma de referência é disponível (preferencialmente da mesma espécie ou de um organismo taxonomicamente relacionado), é possível mapear os *reads* gerados a partir do sequenciamento para reconstruir o transcriptoma utilizando programas de alinhamento específicos, como Bowtie [http://bowtie-bio.sourceforge.net/index.shtml (LANGMEAD et al., 2009)] e o TopHat [http://ccb.jhu.edu/software/tophat/index.shtml (KIM et al., 2013)]. Em contrapartida, na ausência ou quando se encontra genomas de referência incompletos, é possível utilizar uma abordagem similar àquela empregada em genomas e montar *de novo* o transcriptoma utilizando dados de RNA-seq (WANG; GERSTEIN; SNYDER, 2009).

Para montagem *de novo* de transcriptomas existem duas categorias de softwares: aqueles que utilizam algoritmos baseados em gráficos de sobreposição (*overlap-layout-consensus*); e aqueles baseados em gráficos *de Bruijn* (*de-Bruijn-graph*) (FLICEK; BIRNEY, 2010; SCHATZ; DELCHER; SALZBERG, 2010). A depender da escolha do tipo de sequenciamento (*long reads* ou *short reads*) uma ou outra abordagem pode ser mais bem empregada, para *short reads* tem-se obtido melhores resultados e então são mais utilizados montadores com algoritmos baseados em gráficos *de Bruijn* (LI et al., 2012).

3 OBJETIVOS

3.1 OBJETIVO GERAL

Analisar o transcriptoma total de cepas de *Leishmania infantum*, oriundas de área endêmica para transmissão de leishmaniose visceral no Nordeste brasileiro, por meio da tecnologia de sequenciamento de RNA (RNA-seq).

3.2 OBJETIVOS ESPECÍFICOS

a. Analisar sequências gênicas de *Leishmania infantum* através da tecnologia de NGS, sequenciadas a partir de cepas isoladas em Teresina, Piauí;
b. Construir um banco de dados local com as sequências obtidas a partir do sequenciamento do RNA total de *L. infantum*;
c. Avaliar diferentes *softwares* montadores de transcriptoma e estratégias de montagens *de novo* para estabelecer *pipelines* para analisar dados de RNA-seq de *Leishmania*;
d. Estabelecer *pipelines* para rotinas de análise de RNA-seq voltados para tripanossomatídeos, afim de facilitar estudos futuros, principalmente para usuários inexperientes;
e. Anotar a função predita dos genes identificados, através da utilização de *scripts* e programas de bioinformática específicos para esta finalidade.

4 COMPARISON OF TREATMENT STRATEGIES AND *DE NOVO* TRANSCRIPTOME ASSEMBLERS FOR *Leishmania* RNA-SEQ DATA ANALYSIS

Lucas Christian de Sousa-Paula[1]
E-mail: lcsousapaula@gmail.com

Wilson José da Silva Júnior[2]
E-mail: wilson_jsjunior@hotmail.com

Raul Maia Falcão[1]
E-mail: rmf4@cin.ufpe.br

Sérgio de Sá Leitão Paiva Júnior[1]
E-mail: sslpaiva@gmail.com

Valeria Claudiane Simeão Oliveira[3]
E-mail: valmpcsb@gmail.com

Vladmir Costa Silva[3]
E-mail: vladimir.costa@gmail.com

Valdir de Queiroz Balbino[1,*]
E-mail: vqbalbino@gmail.com

Carlos Henrique Nery Costa[3]
E-mail: chncosta@gmail.com

[1]Department of Genetics, Federal University of Pernambuco, Recife, Brazil
[2]Department of Agronomy, Federal Rural University of Pernambuco, Recife, Brazil
[3]Laboratory of Leishmaniasis, Tropical Diseases Institute “Natan Portella”, Federal University of Piauí, Teresina, Brazil

*Corresponding authors:
Valdir de Queiroz Balbino, Department of Genetics, Federal University of Pernambuco, Avenida da Engenharia, s/n, Cidade Universitária, Recife, PE – Brazil, Fone: +55 81 2126-8512, e-mail: valdir@ufpe.br

Submetido à revista Memórias do Instituto Oswaldo Cruz (memorias.ioc.fiocruz.br).

Abstract

Background

The RNA-seq has been chosen by many projects for transcriptome analysis. This technology has allowed the understanding of details at the molecular level of the relationship parasites-hosts. In the absence of a reference genome, RNA-seq reads can be assembled into a transcriptome using *de novo* assemblers. *Leishmania infantum* is major etiological agent of visceral leishmaniasis in the New World, but it has not yet had its transcriptome sequenced using NGS technologies. Then, we compared six transcriptome assemblers, five treatment strategies, sequenced and analysed *Leishmania infantum* RNA-seq.

Results

We sequenced the *Leishmania infantum* transcriptome by RNA-seq and a total 14,849,854 paired reads were yield. We compared six transcriptome assemblers and five treatments pipelines. The pipeline zero (P_0) obtained the better results regarding number of >500 and >1000 bp, database coverage and 95%-assembled isoforms. Among assemblers, rnaSPAdes had best result for 95%-assembled isoforms and database coverage, whereas IDBA-tran was better in the criteria length transcripts >500 and 1000 bp and Trinity had the highest percentage of remapped reads.

Conclusion

Our analysis will pave the way for further studies on the gene expression of *Leishmania*, especially for inexperienced users, serving as basis to determine workflows for RNA-seq analysis according to their own objectives.

Keywords: Next-generation sequencing, *de novo* assembly, *Leishmania infantum*, Trypanosomatidae

Background

During the last years the RNA-seq technology has been chosen by many projects for transcriptome analysis, in view of its greater accuracy, sensitivity and reliability (McGettigan 2013). When a reference genome (preferably of the same species) is available, it is possible to use it to anchor the reads generated by the RNA-seq and to assemble the transcriptome by reference. On the other hand, in the absence or incomplete reference genome, RNA-seq reads can be assembled into a transcriptome using *de novo* assemblers (*e.g.*, IDBA-Tran; rnaSPAdes; SOAPdenovo-Trans; Oases; Trans-ABySS; Trinity) (Birol et al. 2009, Bankevich et al. 2012, Schulz et al. 2012, Haas et al. 2013, Peng et al. 2013, Xie et al. 2014, Conesa et al. 2016).

Nevertheless, this technology has allowed the understanding of details at the molecular level of the relationship between parasites and their hosts, including trypanosomatids (Cantacessi et al. 2015). Trypanosomatids are a major public health problem, because they are causative agents of various diseases, such as

leishmaniasis, caused by genus *Leishmania*, which affects the lives of millions of people worldwide (Alvar et al. 2012).

Most of these parasites had their complete genomes published at the last decade (Peacock et al. 2007, Cantacessi et al. 2015), having as milestone the publication of the *Leishmania major*, *Trypanosoma cruzi* and *T. brucei* genomes, the so-called Tritryps (Berriman 2005, El-Sayed et al. 2005, Ivens et al. 2005). Currently, driven by the publication of genomes, many studies focus on the analysis of the *Leishmania* transcriptomes, have been aiming at the difference of gene expression in different conditions, stages of life cycle development and total transcriptome analysis (Alcolea et al. 2010, Rastrojo et al. 2013, Dillon et al. 2015).

L. infantum is the major etiological agent of visceral leishmaniasis in the New World (Alvar et al. 2012) and, as abovementioned, had its genome sequenced in 2007 (Peacock et al. 2007). However, it has not yet had its total transcriptome sequenced using next-generation sequencing technologies. The studies on gene expression of this parasite are limited principally by microarray (Rochette et al. 2009, Alcolea et al. 2010).

Currently, with the increase in demand to analyse the data generated from the next-generation sequencing, several software have been developed for *de novo* assembly (Birol et al. 2009, Bankevich et al. 2012, Schulz et al. 2012, Haas et al. 2013, Peng et al. 2013, Xie et al. 2014), reference assembly (Li et al. 2009, Kim et al. 2013) and evaluated quality of transcriptomes assemblies (Li et al. 2014, Bushmanova et al. 2016). In addition, it is possible to find in the literature some revisions about assembly strategies (Li et al. 2012, Chopra et al. 2014, Conesa et al. 2016), on the other hand, sometimes these papers do not embrace the main assemblers, besides what developers also update their programs by fixing some flaws and limitations of previous ones.

To allow inexperienced or not users to most easily choose the appropriate software and treatment strategies for *Leishmania* RNA-seq analysis according to their own objectives, we made comparisons of the main transcriptome assemblers reported in literature and treatment strategies (pipelines).

Material and methods

Leishmania's origin, culture and RNA extraction

The sample of *L. infantum* Teresina strain (MHOM/BR/2016/3075) obtained from bone marrow aspiration that was used in this study, come from routine collection by the Tropical Diseases Institute “Natan Portella” (IDTNP), Teresina, Piauí State, Brazil, for the diagnosis and treatment of patients, and is kept cryo-preserved in liquid nitrogen according to established protocols (Brasil 2006).

L. infantum Teresina strain (MHOM/BR/2016/3075) was reactivated and cultured in Schneider's medium (Sigma-Aldrich) at pH 7.2 supplemented with inactivated fetal bovine serum (10%), human urine (2%) and maintained at 26 °C. Expansion of the promastigotes for approximately 7 days was monitored and quantified by counting in Neubauer's chamber reaching a mean density of 4.2×10^7 cells/ml of medium in the final log phase of growth.

Cells were washed three times with 0.9% NaCl solution for removal from the medium and precipitated by centrifugation at 3,000 rpm for 10 minutes at 4 °C. The supernatant was discarded and the pellet was resuspended in RLT lysis buffer (RNeasy® Mini Kit – Qiagen, Hilden, DE). To ensure a better purity, the extracted RNA was treated with DNase (Qiagen, Hilden, DE) and purified with the RNeasy® Mini Kit (Qiagen, Hilden, DE) according to the manufacturer's instructions. Quantification of the sample was performed by absorbance at 260 nm using the Nanodrop ND 2000 spectrophotometer (Thermo Scientific, Wilmington, DE) and the RNA integrity assessment was checked on an Agilent 2100 bioanalyzer (Agilent Technologies, Palo Alto, CA) according to the manufacturer’s instructions.

Ethical approval

The protocol and informed consent, obtained from all the participants or their legal guardian, were approved by the Research Ethics Committee (CEP) of the Federal University of Piauí under the number 0116/2005. The study was conducted in accordance with the principles of the Declaration of Helsinki in resolution 196/96 of the National Health Council of the Ministry of Health that regulates research involving human beings in Brazil.

RNA Sequencing and *de novo* transcriptome assembly

Total RNA from high quality sample (RIN values > 7.5) was sequencing at the Laboratório Central de Tecnologias de Alto Desempenho em Ciências da Vida (LaCTAD) from the University of Campinas, Brazil. The cDNA libraries were generated

using the TruSeq RNA Sample Prep Kit v2 (Illumina, San Diego, CA) according to manufacturer's recommendations and sequencing was performed using the Illumina HiSeq 2500 platform on a HiSeq Flow Cell v4, with the HiSeq SBS Kit v4 and paired-reads 100 bp (2x).

Initially, quality raw reads were analysed by FastQC (Andrews 2010), adapters were removed (clipped reads) with FASTX-Toolkit (v. 0.0.13) (http://hannonlab.cshl.edu/fastx_toolkit/index.html) and then clipped reads were treated following five distinct pipelines (P) (Fig. 1): P_0, clipped reads were repaired using BBTools (www.jgi.doe.gov/data-and-tools/bbtools) and next assembled. P_1: repaired reads in P_0 were trimmed by length (>45) and quality (>30) and then assembled. P_2: repaired reads in P_0 were trimmed by length (16 initial bases were removed), repaired again and after assembled. P_3: repaired reads in P_2 were trimmed by length (>45) and quality (>30) and then assembled. P_4: repaired reads in P_2 were trimmed by length (>85) and quality (>20) and then assembled. We used the Sickle (v. 1.33) (Joshi & Fass 2011) for to trim reads in all abovementioned pipelines. All assemblies were evaluated by rnaQUAST (v 1.4.0) (Bushmanova et al. 2016) using *L. infantum* JPCM5 genome (release 9.0) as database reference available in TriTrypDB (http://tritrypdb.org/tritrypdb/).

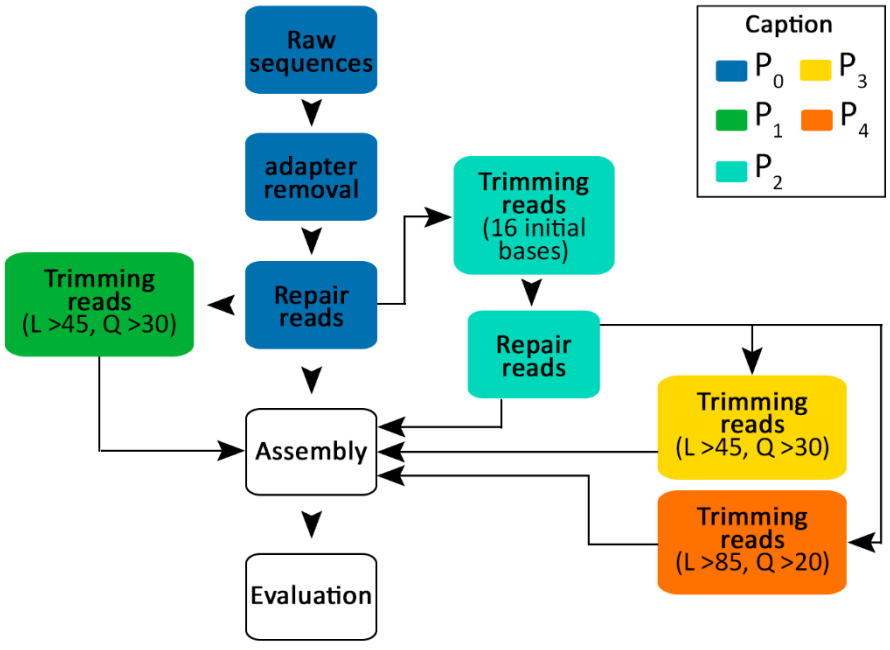


Fig. 1 Pipelines workflow used for assemblies. L: read length; Q: quality phred; P: pipelines.

Assemblers comparison and remapping reads

To perform the comparison, we used the following *de novo* transcriptome assemblers: IDBA-Tran (v. 1.1.1) (Peng et al. 2013); rnaSPAdes (v. 3.10.1) (Bankevich et al. 2012); SOAPdenovo-Trans (v. 1.03) (Xie et al. 2014); Oases (v. 0.2.08) (Schulz et al. 2012); Trans-ABySS (v. 1.5.5) (Birol et al. 2009); Trinity (v. 2.4.0) (Haas et al. 2013).

We ran rnaSPAdes, SOAPdenovo-Trans and Trinity with the default parameters (*k*-mer = 55, 23 and 25, respectively). The other assemblers we ran with various *k*-mer sizes: IDBA-Tran (*k* = 24, 26, 28, 30, 32, 34, 36), Oases (*k*= 37, 39, 42, 43, 45, 47, 49, 51) and Trans-ABySS (*k* = 26, 32, 40).

To evaluate the number of raw reads that contributed to each assembly, we remapped the raw reads using BWA aligner (Li & Durbin 2009) and used a custom python script for mapping statistics.

Results

Sequencing

Total RNA sequencing yielded 14,849,854 paired reads. The raw reads were deposited into the Short Read Archive (SRA) of the National Center for Biotechnology Information (NCBI) and can be accessed under the access number SRR6282525.

The raw reads were evaluated by FastQC (Andrews 2010) and Table 1 shows the statistics generated. Reads with a sequence length of 101 bp were obtained and with GC% of 59%.

Table 1 – Raw reads statistics using FastQC.

	R1	R2
File type	Conventional base calls	Conventional base calls
Encoding	Sanger / Illumina 1.9	Sanger / Illumina 1.9
Total Sequences	7,424,927	7,424,927
Sequences flagged as poor quality	0	0
Sequence length	101	101
%GC	59	59

%GC: GC content

***De novo* assembly**

We used paired-end reads to assemble the *Leishmania infantum* Teresina strain (MHOM/BR/2016/3075) transcriptome comparing six *de Bruijn* graph-based assemblers and five assemblies pipelines, resulting in 30 assemblies. Table 2 shows the basics characteristics and how much is cited in the scientific literature each transcriptome assembler evaluated in this study.

Table 2 – Basic characteristics of currently used and freely available short read transcriptome assemblers.

Assembler	Version (last release)	Method	Input seq format	Read pair format	Multiple libraries	Reference	Citations (total/2017[+])
IDBA-tran	1.1.1 (2012)	de Bruijn multiple *k*-mer	.fasta	interleaved only	yes	(Peng et al. 2013)	53/14
rnaSPAdes	1.03 (2012)	de Bruijn unique *k*-mer	.fastq	interleaved or separate	yes	(Bankevich et al. 2012)	2,320/755[*]
SOAPdenovo-Trans	1.03 (2017)	de Bruijn unique *k*-mer	.fastq, .fasta	interleaved or separate	yes	(Xie et al. 2014)	300/82
Oases	0.2.08 (2011)	de Bruijn multiple *k*-mer	.fasta	interleaved only	yes	(Schulz et al. 2012)	894/115
Trans-ABySS	1.5.5 (2016)	de Bruijn multiple *k*-mer	.fastq	interleaved or separate	yes	(Birol et al. 2009)	355/22
Trinity	2.4.0 (2017)	de Bruijn unique *k*-mer	.fastq	interleaved or separate	yes	(Haas et al. 2013)	1,415/428

*There is no a specific publication for rnaSPAdes, so assembler developers suggest citing SPAdes' paper.
[+]Number of citations until September 2017.

In Table 3, it is possible to observe the 30 assemblies' results. Assemblies generated a total ranging from 20,576– 68,192 assembled transcripts with coverage between 69–95% (Table 3). For assemblies evaluation, we used as criteria the number of 95%-assembled isoforms and database coverage, as suggested by Bushmanova et

al. (2016), in addition to these, we also take into account the number of assembled transcripts >500 and >1000 bp.

Assembly strategies and statistics

The pipelines were established to observe how assembly results may vary when raw reads are subjected to certain treatments. We used the statistics generated by rnaQUAST for evaluated the assemblies and, consequently, all five pipelines (Table 3). Using P_0, all assemblies had database coverage >90%, number of transcripts >500 bp between 39–85%, transcripts >1000 bp between 21–62% and 95%–assembled isoforms ranging between 39–54%. In P_1, assemblies presented database coverage from 88% to 93%, number of transcripts >500 bp between 36–80%, transcripts >1000 bp between 16–49% and 95%–assembled isoforms ranging between 29–45%. Assemblies with P_2 had database coverage >90%, number of transcripts >500 bp between 37–80%, transcripts >1000 bp between 16–49% and 95%-assembled isoforms ranging between 26–37%. In P_3, all assemblies had database coverage below 86%, number of transcripts >500 bp between 24–67%, transcripts >1000 bp between 7–28% and 95%-assembled isoforms ranging between 14–24%. Lastly, assemblies following P_4 presented database coverage between 69–90%, number of transcripts >500 bp between 22–63%, transcripts >1000 bp 5–22% and 95%-assembled isoforms ranging between 11–24%.

Table 3 – Assemblies results generated by rnaQUAST using *Leishmania infantum* JPCM5 genome as reference database.

	Assembler	T	T >500 bp	T >1000 bp	Aligned	Mis-assemblies	Database coverage	50%-assembled isoforms	95%-assembled isoforms	Un-annotated
P_0	IDBA-tran	20576	17566	12685	20464	163	89.7%	6530	3984	5230
	Oases	53210	38953	25843	52985	3259	94.2%	6464	3272	14881
	rnaSPAdes	24001	15158	10666	23769	154	91%	6583	4508	8706
	ABySS$^+$	35490	24812	17600	34975	626	93.8%	6930	4355	12372
	Trinity	33377	23699	15019	33200	151	93.6%	6699	3612	11382
	SOAP*	46406	18228	9659	41517	13	95.2%	6629	3467	22438
P_1	IDBA-tran	24907	19932	12182	24817	242	88.4%	6158	3234	6677
	Oases	55115	32831	17079	54922	2090	89.9%	5511	2394	17004
	rnaSPAdes	31297	17196	10487	31097	119	91.8%	6284	3796	11482
	ABySS$^+$	41540	24430	14274	41133	570	92.8%	6310	3386	14983
	Trinity	37363	23395	12396	37200	140	92%	6006	2831	12837
	SOAP*	51434	18676	8346	48170	15	93.3%	5806	2645	23549
P_2	IDBA-tran	26685	21443	13145	26586	230	89.5%	6158	3032	7077
	Oases	57317	32777	16385	57074	1482	92.2%	5472	2213	18939
	rnaSPAdes	41808	18294	9509	41613	56	91.7%	5770	2994	15405

	ABySS[+]	40842	24152	13900	40505	513	93.4%	6218	3121	15021
	Trinity	35826	22695	11913	35685	113	92.4%	5879	2593	12006
	SOAP[*]	52309	19103	8334	48862	12	94%	5780	2477	24152
P_3	IDBA-tran	30014	20063	8272	29921	248	75.6%	4672	1820	8613
	Oases	58172	20607	6278	57962	774	75.4%	3655	1212	20272
	rnaSPAdes	54770	17420	6526	54573	44	88.1%	4656	2040	19770
	ABySS[+]	54568	20203	7500	54257	384	85.1%	4570	1840	20086
	Trinity	42866	19023	6245	42703	98	82.4%	4311	1570	14331
	SOAP[*]	67339	16037	4440	65260	23	86.6%	4133	1450	28477
P_4	IDBA-tran	30192	19058	6734	30109	276	69%	3946	1374	8796
	Oases	57595	19530	5458	57387	913	71.8%	3219	941	20304
	rnaSPAdes	52127	18303	7175	51931	41	90.1%	4864	1996	18978
	ABySS[+]	56393	18758	5815	56080	380	79.2%	3794	1303	21146
	Trinity	43208	18143	5269	43037	105	77.9%	3795	1219	14616
	SOAP[*]	68192	14835	3459	66070	24	81.6%	3542	1104	28770

[*]SOAPdenovo-Trans.
[+]Trans-ABySS.
T: Transcripts; P_0 – P_4: Pipelines.

Discussion

Various next-generation sequencing technologies have been employed for trypanosomatids parasites studies, including some *Leishmania* species (Cantacessi et al. 2015). However, until now there are no published papers that have employed RNA-seq technology for analysis of total transcriptome of *Leishmania infantum*. Furthermore, we compared six *de novo* transcriptome assemblers and five pipelines for RNA-seq data treatment and assembly.

In the literature, it is possible to find publications that approach comparisons between assembly tools (Chopra et al. 2014), assembly algorithms (Li et al. 2012), as well evaluation tools for *de novo* transcriptome assemblies from RNA-seq data (Li et al. 2014, Bushmanova et al. 2016). These ones are very useful, in view of increasing use of -omics technologies by inexperienced users, serving as basis mainly in choosing the pipeline raw reads treatments and which assembler to use. Although it is a very difficult task to delineate a standard pipeline for any RNA-seq studies (Conesa et al. 2016), comparisons may be helpful and directing the new works, especially with organisms with remarkably syntenic genomes, as leishmanial parasites (Peacock et al. 2007, Cantacessi et al. 2015).

Comparison of pipelines

All five pipelines generated discrepant results, which was expected, as shown in the Fig. 2. As we used rnaQUAST to evaluate assemblies, it is important to note that the software calls isoforms all genes from the supplied database, in the present study we

used the reference genome *L. infantum* JPCM5 (release 9.0) available in TriTrypDB (http://tritrypdb.org/tritrypdb).

In P_0, raw reads were not trimmed by length and neither filtered by quality, only the adaptors were removed, whereas has been done trimming and filtering for most RNA-seq analysis in literature (Tao et al. 2017). This approach was chosen because of some *de novo* assembly tools, such as rnaSPAdes, correct reads before assembly. In view of this, in this approach, we seek to use the full-length and all sequenced reads. Moreover, Williams et al. (2016) suggest that the use of trimming in RNA-seq analysis workflows requires caution, since the decrease in reads length can generate spurious alignments when mapped in the reference.

Using longer reads, P_0 obtained the smallest number of transcripts assembled among all the pipelines and in all assemblies (Fig. 2). This is possibly unrelated if the assembler uses single or multiple *k*-mer or *k*-mer size used, *e.g.*, we ran Trinity with single *k*-mer 25 and 33,377 transcripts were assembled, whereas we ran IDBA-tran with multiple *k*-mers (see Methods) and 20,576 transcripts were assembled. On the other hand, rnaSPAdes was run with *k*-mer 55 and 24,001 transcripts were assembled. In addition, the P_0 showed the highest average coverage among pipelines (93%). Another important result of P_0 was the number of 95%-assembled isoforms (mean 46%, against 36%, 33%, 20% and 16% of other pipelines), this metric is calculated by a single transcript assembling a single database isoform, *i.e.*, an isoform is considered to be 95% -assembled if it has at least 95% covered by a single transcript (Bushmanova et al. 2016). This metric may suggest how much assembly is non-fragmented.

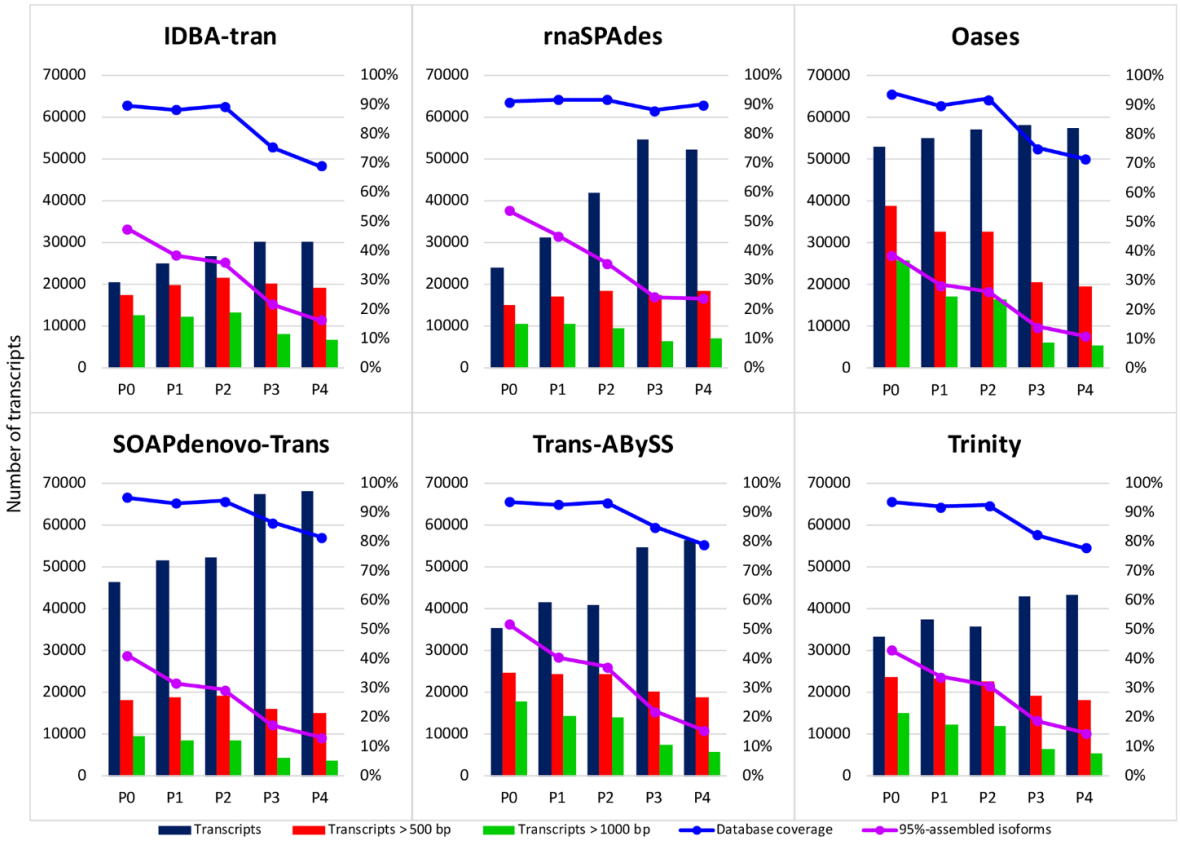


Fig. 2 Assemblies metrics for each transcriptome assembler generated by rnaQUAST. In primary x-axis are represented numbers of total assembled transcripts, transcripts >500 bp and >1000 bp (in bars). In secondary x-axis are represented coverage percentage and 95%-assembled isoforms (in lines). In y-axis, its plotted the results for all five pipelines used for each transcriptome assembler. The 95%-assembled isoforms percentage was calculated by the ratio of the number of 95%-assembled isoforms divided by the total number of isoforms in the database used (reference genome) which is 8381 genes (*Leishmania infantum* JPCM5 release 9.0).

In P_1 and P_2, we used most are commonly strategies in RNA-seq data analysis, as discard low-quality reads and very short reads (Tao et al. 2017), in addition, we eliminated bias in the first bases with poor-quality (see Fig. 1). Even in face of particulars different each assemblers, P_1 and P_2 had similar results in almost all six assemblers (except rnaSPAdes which presented discrepancy) with regard to the number of assembled transcripts, transcripts >500 bp and >1000 bp (Fig. 2 and Table 3). Compared to P_0, regarding the database coverage, presented closer values, 91% and 92% (P_1 and P_2, respectively) against 93% of P_0. Still compared to P_0, regarding the 95%-assembled isoforms, an average 36% and 33% were obtained by P_1 and P_2, respectively, against 46% of P_0.

Assemblies using P_3 and P_4 showed highest number of transcripts assembled, however, the mean database coverage was >90% (82% for P_3 and 78% for P_4). In addition, the lowest transcripts values >500 bp and >1000 bp were achieved by them

and worst results of 95%-assembled isoforms (Fig. 2 and Table 3). These data suggest that the assembly is rather fragmented, since the mean transcript >500 bp is approximately 40% and 38% for P_3 and P_4, respectively. They assembled many transcripts but assembled small ones.

Comparison of *de novo* transcriptome assemblers

We compared six *de novo* transcriptome assemblers described in the literature (see Methods and Table 2). All assembly tools used algorithms *de Bruijn* graph-based, *i.e.*, the assembler infers the precise sequence first chopping reads into much shorter *k*-mers and then using all the *k*-mers to form a *de Bruijn* graph (Li et al. 2012).

The assemblies by IDBA-tran showed database coverage range between 69–89.7%, the lowest among assemblers (Fig. 2 and Table 3), and remapped reads average 85% (Fig. 3). Furthermore, IDBA-tran assembled the lowest number of transcripts compared with other assemblers (mean 26,474.8 assembled transcripts). On the other hand, SOAPdenovo-trans assembled the highest mean of transcripts (57,136), whereas showed average database coverage ~90% (81.6–95.2%) and remapped reads average ~88% (85.6–90.1%). Probably, the number of assembled transcripts was not influence by the reads that contribute for the assembly, may be a matter of assembled transcripts with longer and shorter lengths.

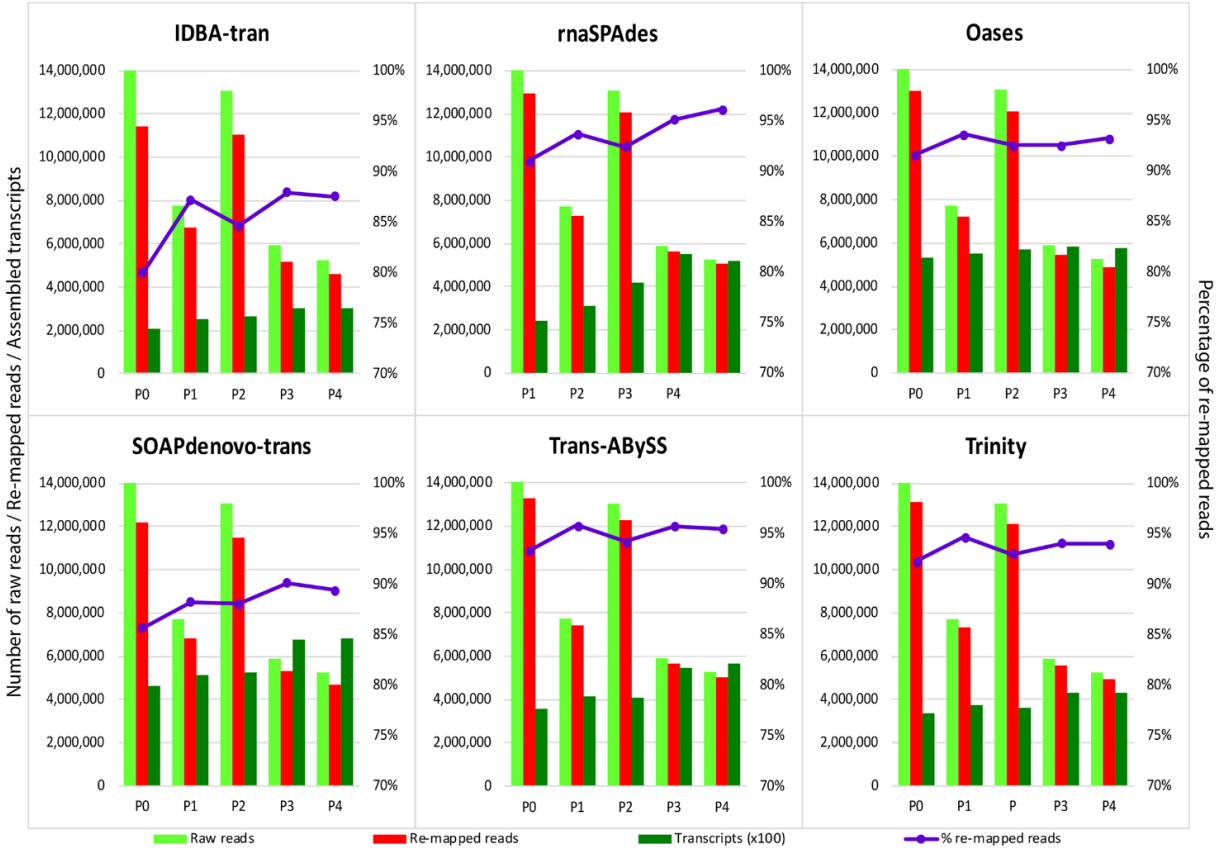


Fig. 3 Remapped reads in the assembled transcripts. In primary x-axis are represented numbers of raw reads, remapped reads and assembled transcripts (in bars). In secondary x-axis are represented percentage of remapped reads (in lines). In y-axis, its plotted the results for all five pipelines used for each transcriptome assembler.

Assemblies made by rnaSPAdes showed a database coverage average >90%, only assembly using P_3 had coverage <90% (88.1%). Another relevant result, combined with P_0, about 54% of database isoforms were covered at least 95% by a single transcript assembled, the best result of this criteria among all assemblies. Comparing also transcriptome assemblies, Bushmanova et al. (2016) used SPAdes (genome assembler, still did not have a specific algorithm for transcriptome assembler) against transcriptome assemblers *de Bruijn* graph-based and had better results for 95%-assembled isoforms. On the question remapped reads, >90% of the reads were remapped in all assemblies using rnaSPAdes (Fig. 3 and Table 4).

Trinity is tool most used for RNA-seq data analysis (Table 2). Assemblies combined with P_0–P_2 showed database coverage >90%, whereas assemblies with P_3 and P_4 had 82.4% and 77.9%, respectively (Fig. 2). With regards to the remapped reads, all assemblies using Trinity were >92% (Table 4), result higher than found by Chopra et al. (2014) comparing peanut transcriptome assemblies by Trinity. Furthermore,

between 14,5–43% of database isoforms were 95%-assembled by transcripts assembled by Trinity.

Assemblies with Trans-ABySS yielded between 35,490–56,393 transcripts, of these between 33.2–69.1% had at least 500 bp in length, whereas 10.3–49.6% had at least 1000 bp (Table 3). With regards to the database coverage, assemblies combined with P_0–P_2 showed >90%, like Trinity. An average 95% (93.3–95.8%) of raw read remapped in the assembled transcripts, acceptable results, in relation to those obtained by Chopra et al. (2014).

Table 4 Remapped reads in the assembled transcripts

	Assembler	Raw reads**	Remapped reads	%-remapped reads	Assembled transcripts
P_0	IDBA-tran	14,231,732	11,383,789	79.99%	20,576
	Oases	14,231,732	13,035,013	91.59%	53,210
	rnaSPAdes	14,231,732	12,951,266	91.00%	24,001
	ABySS+	14,231,732	13,282,894	93.33%	35,490
	Trinity	14,231,732	13,126,506	92.23%	33,377
	SOAP*	14,231,732	12,188,490	85.64%	46,406
P_1	IDBA-tran	7,730,816	6,738,747	87.17%	24,907
	Oases	7,730,816	7,237,012	93.61%	55,115
	rnaSPAdes	7,730,816	7,247,119	93.74%	31,297
	ABySS+	7,730,816	7,406,282	95.80%	41,540
	Trinity	7,730,816	7,316,777	94.64%	37,363
	SOAP*	7,730,816	6,823,842	88.27%	51,434
P_2	IDBA-tran	13,054,750	11,049,969	84.64%	26,685
	Oases	13,054,750	12,075,142	92.50%	57,317
	rnaSPAdes	13,054,750	12,069,404	92.45%	41,808
	ABySS+	13,054,750	12,298,592	94.21%	40,842
	Trinity	13,054,750	12,138,369	92.98%	35,826
	SOAP*	13,054,750	11,491,870	88.03%	52,309
P_3	IDBA-tran	5,904,936	5,190,998	87.91%	30,014
	Oases	5,904,936	5,462,088	92.50%	58,172
	rnaSPAdes	5,904,936	5,617,076	95.13%	54,770
	ABySS+	5,904,936	5,654,633	95.76%	54,568
	Trinity	5,904,936	5,555,514	94.08%	42,866
	SOAP*	5,904,936	5,321,383	90.12%	67,339
P_4	IDBA-tran	5,264,370	4,606,557	87.50%	30,192
	Oases	5,264,370	4,905,559	93.18%	57,595
	rnaSPAdes	5,264,370	5,062,502	96.17%	52,127
	ABySS+	5,264,370	5,023,977	95.43%	56,393
	Trinity	5,264,370	4,946,223	93.96%	43,208
	SOAP*	5,264,370	4,705,683	89.39%	68,192

*SOAPdenovo-Trans.
**Paired-end reads
+Trans-ABySS.
T: Transcripts; P_0 – P_4: Pipelines.

Conclusions

Given the significant importance in the study of human pathogenic parasites, such as *Leishmania* spp., analysis and *de novo* transcriptome assembly of these parasites may reveal new species-specific genes, which may be targets for the development of new treatment measures, besides provide robust information for better reference genome annotation and knowledge at the molecular level these parasites.

In summary, our results will pave the way for studies on the gene expression of *Leishmania* spp., especially for inexperienced users, serving as basis to determine workflows for RNA-seq analysis according to their own objectives.

Declarations

Acknowledgements

We are grateful to Tropical Diseases Institute “Natan Portella” and to LABBE/CB at the Federal University of Pernambuco for use its facilities.

Ethical approval

Sample collection was approved (process number 0116/2005) by the Research Ethics Committee (CEP) of Federal University of Piauí.

Consent for publication

Not applicable.

Availability of data and materials

All raw sequence data generated in this study were deposited in the NCBI Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) under accession numbers SRR6282525. All other supporting data are included as additional files.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was supported by the Coordination for Improvement of Personnel with Higher Education (CAPES), Bio-Computational Program.

Authors’ contributions

CHNC, VQB and VCS designed the project. CHNC, VCS and VCSO collected the samples. LCSP, RMF, WJSJ and VQB performed the assemblies, analyses and wrote the manuscript; VQB, CHNC, SSLPJ and VCS revised the manuscript; VQB, SLPJ, VCS and CHNC supervised the study and revised the definitive version of the manuscript. All authors read and approved the final manuscript.

References

Alcolea PJ, Alonso A, Gómez MJ, Moreno I, Domínguez M, Parro V, Larraga V 2010.

Transcriptomics throughout the life cycle of *Leishmania infantum*: High down-regulation rate in the amastigote stage. *Int. J. Parasitol.* 40: 1497–1516.

Alvar J, Vélez ID, Bern C, Herrero M, Desjeux P, Cano J, Jannin J, Boer M de 2012. Leishmaniasis worldwide and global estimates of its incidence. *PLoS One* 7.

Andrews S 2010. FastQC: A quality control tool for high throughput sequence data. FastQC: A Qual. Control tool high throughput Seq. data.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin A V., Sirotkin A V., Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* 19: 455–477.

Berriman M 2005. The Genome of the African Trypanosome *Trypanosoma brucei*. *Science (80).* 309: 416–422.

Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, Stazyk G, Morin RD, Zhao Y, Hirst M, Schein JE, Horsman DE, Connors JM, Gascoyne RD, Marra MA, Jones SJM 2009. De novo transcriptome assembly with ABySS. *Bioinformatics* 25: 2872–2877.

Brasil 2006. *Manual de Vigilância e Controle da Leishmaniose Visceral*. Editora do Ministério da Saúde, Brasília, DF.

Bushmanova E, Antipov D, Lapidus A, Suvorov V, Prjibelski AD 2016. RnaQUAST: A quality assessment tool for de novo transcriptome assemblies. *Bioinformatics* 32: 2210–2212.

Cantacessi C, Dantas-Torres F, Nolan MJ, Otranto D 2015. The past, present, and future of *Leishmania* genomics and transcriptomics. *Trends Parasitol.* 31: 100–108.

Chopra R, Burow G, Farmer A, Mudge J, Simpson CE, Burow MD 2014. Comparisons of de novo transcriptome assemblers in diploid and polyploid species using peanut (Arachis spp.) RNA-Seq data. *PLoS One* 9: 1–16.

Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczesniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A 2016. A survey of best practices for RNA-seq data analysis. *Genome Biol* 17: 13.

Dillon LAL, Okrah K, Hughitt K V., Suresh R, Li Y, Fernandes MC, Belew AT, Bravo HC, Mosser DM, El-Sayed NM 2015. Transcriptomic profiling of gene expression and RNA processing during *Leishmania major* differentiation.

Nucleic Acids Res. 43: 6799–6813.

El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran A-N, Ghedin E, Worthey EA, Delcher AL, Blandin G, Westenberger SJ, Caler E, Cerqueira GC, Branche C, Haas B, Anupama A, Arner E, Aslund L, Attipoe P, Bontempi E, Bringaud F, Burton P, Cadag E, Campbell DA, Carrington M, Crabtree J, Darban H, Silveira JF da, Jong P de, Edwards K, Englund PT, Fazelina G, Feldblyum T, Ferella M, Frasch AC, Gull K, Horn D, Hou L, Huang Y, Kindlund E, Klingbeil M, Kluge S, Koo H, Lacerda D, Levin MJ, Lorenzi H, Louie T, Machado CR, McCulloch R, McKenna A, Mizuno Y, Mottram JC, Nelson S, Ochaya S, Osoegawa K, Pai G, Parsons M, Pentony M, Pettersson U, Pop M, Ramirez JL, Rinta J, Robertson L, Salzberg SL, Sanchez DO, Seyler A, Sharma R, Shetty J, Simpson AJ, Sisk E, Tammi MT, Tarleton R, Teixeira S, Aken S Van, Vogt C, Ward PN, Wickstead B, Wortman J, White O, Fraser CM, Stuart KD, Andersson B 2005. The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. *Science* 309: 409–415.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N, Regev A 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* 8: 1494–1512.

Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, Sisk E, Rajandream MA, Adlem E, Aert R, Anupama A, Apostolou Z, Attipoe P, Bason N, Bauser C, Beck A, Beverley SM, Bianchettin G, Borzym K, Bothe G, Bruschi C V, Collins M, Cadag E, Ciarloni L, Clayton C, Coulson RM, Cronin A, Cruz AK, Davies RM, Gaudenzi J De, Dobson DE, Duesterhoeft A, Fazelina G, Fosker N, Frasch AC, Fraser A, Fuchs M, Gabel C, Goble A, Goffeau A, Harris D, Hertz-Fowler C, Hilbert H, Horn D, Huang Y, Klages S, Knights A, Kube M, Larke N, Litvin L, Lord A, Louie T, Marra M, Masuy D, Matthews K, Michaeli S, Mottram JC, Muller-Auer S, Munden H, Nelson S, Norbertczak H, Oliver K, O'Neil S, Pentony M, Pohl TM, Price C, Purnelle B, Quail MA, Rabbinowitsch E, Reinhardt R, Rieger M, Rinta J, Robben J, Robertson L, Ruiz JC, Rutter S, Saunders D, Schafer M, Schein J, Schwartz DC, Seeger K, Seyler A, Sharp S, Shin H, Sivam D, Squares R, Squares S, Tosato V, Vogt C, Volckaert G,

Wambutt R, Warren T, Wedler H, Woodward J, Zhou S, Zimmermann W, Smith DF, Blackwell JM, Stuart KD, Barrell B, Myler PJ 2005. The genome of the kinetoplastid parasite, *Leishmania major*. *Science (80-.).* 309: 436–442.

Joshi N, Fass J 2011. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14: R36.

Li Z, Chen Y, Mu D, Yuan J, Shi Y, Zhang H, Gan J, Li N, Hu X, Liu B, Yang B, Fan W 2012. Comparison of the two major classes of assembly algorithms: Overlap-layout-consensus and de-bruijn-graph. *Brief. Funct. Genomics* 11: 25–37.

Li H, Durbin R 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754–1760.

Li B, Fillmore N, Bai Y, Collins M, Thomson J a., Stewart R, Dewey C 2014. Evaluation of de novo transcriptome assemblies from RNA-Seq data. *bioRxiv*: 1–21.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078–2079.

McGettigan PA 2013. Transcriptomics in the RNA-seq era. *Curr. Opin. Chem. Biol.* 17: 4–11.

Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail MA, Peters N, Adlem E, Tivey A, Aslett M, Kerhornou A, Ivens A, Fraser A, Rajandream M-A, Carver T, Norbertczak H, Chillingworth T, Hance Z, Jagels K, Moule S, Ormond D, Rutter S, Squares R, Whitehead S, Rabbinowitsch E, Arrowsmith C, White B, Thurston S, Bringaud F, Baldauf SL, Faulconbridge A, Jeffares D, Depledge DP, Oyola SO, Hilley JD, Brito LO, Tosi LRO, Barrell B, Cruz AK, Mottram JC, Smith DF, Berriman M 2007. Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. *Nat. Genet.* 39: 839–847.

Peng Y, Leung HCM, Yiu SM, Lv MJ, Zhu XG, Chin FYL 2013. IDBA-tran: A more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. In: *Bioinformatics*,

Rastrojo A, Carrasco-Ramiro F, Martín D, Crespillo A, Reguera RM, Aguado B, Requena JM 2013. The transcriptome of *Leishmania major* in the axenic promastigote stage: transcript annotation and relative expression levels by

RNA-seq. *BMC Genomics* 14: 223.

Rochette A, Raymond F, Corbeil J, Ouellette M, Papadopoulou B 2009. Whole-genome comparative RNA expression profiling of axenic and intracellular amastigote forms of *Leishmania infantum*. *Mol. Biochem. Parasitol.* 165: 32–47.

Schulz MH, Zerbino DR, Vingron M, Birney E 2012. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 28: 1086–1092.

Tao S-Q, Cao B, Tian C-M, Liang Y-M 2017. Comparative transcriptome analysis and identification of candidate effectors in two related rust species (Gymnosporangium yamadae and Gymnosporangium asiaticum). *BMC Genomics* 18: 651.

Williams CR, Baccarella A, Parrish JZ, Kim CC 2016. Trimming of sequence reads alters RNA-Seq gene expression estimates. *BMC Bioinformatics* 17: 103.

Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, Huang W, He G, Gu S, Li S, Zhou X, Lam TW, Li Y, Xu X, Wong GKS, Wang J 2014. SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* 30: 1660–1666.

5 TRANSCRIPTOMICS ANALYSES OF *Leishmania infantum* TERESINA STRAINS, ETIOLOGICAL AGENT OF AMERICAN VISCERAL LEISHMANIASIS: ABUNDANCE AND NOVEL PUTATIVE GENES

Lucas Christian de Sousa-Paula[1]
E-mail: lcsousapaula@gmail.com

Wilson José da Silva Júnior[2]
E-mail: wilson_jsjunior@hotmail.com

Raul Maia Falcão[1]
E-mail: rmf4@cin.ufpe.br

Sérgio de Sá Leitão Paiva Júnior[1]
E-mail: sslpaiva@gmail.com

Valeria Claudiane Simeão Oliveira[3]
E-mail: valmpcsb@gmail.com

Vladmir Costa Silva[3]
E-mail: vladimir.costa@gmail.com

Valdir de Queiroz Balbino[1, *]
E-mail: vqbalbino@gmail.com

Carlos Henrique Nery Costa[3]
E-mail: chncosta@gmail.com

[1]Department of Genetics, Federal University of Pernambuco, Recife, Brazil
[2]Department of Agronomy, Federal Rural University of Pernambuco, Recife, Brazil
[3]Laboratory of Leishmaniasis, Tropical Diseases Institute “Natan Portella”, Federal University of Piauí, Teresina, Brazil

*Corresponding authors:
Valdir de Queiroz Balbino, Department of Genetics, Federal University of Pernambuco, Avenida da Engenharia, s/n, Cidade Universitária, Recife, PE – Brazil, Fone: +55 81 2126-8512, e-mail: valdir@ufpe.br.

Abstract

Background

Visceral leishmaniasis is a severe human disease widespread in the World and in American continent is caused by *Leishmania infantum.* Currently, next-generation sequencing technology has helped to understand at the molecular level the host-parasite interaction. *L. infantum* transcriptome was not analysed using NGS technology until then, as has been done for another trypanosomatids, previous analyses made to the use of microarrays, that have a lower accuracy than RNA-sequencing (RNA-seq).

Methodology/Principal findings

Through RNA-seq, we analysed eight *L. infantum* transcriptomes of Teresina strains. Total of ~123.9 million short sequence reads (101 bp) were produced, with database coverage > 90%. Between 19,840–36,676 transcripts were *de novo* assembled, of this about 30% were classified as unannotated and 0.5-6.6% did not align against the reference genome. 88 novel putative *L. infantum* genes were found. Regarding genic abundance, around 20 of 50 most expressed genes are annotated as ribosomal protein. Histones, β-tubulin and α-tubulin were some of the top expressed genes.

Conclusion/Significance

Herein, we present the firs report, to date, for *L. infantum* total transcriptome using RNA-seq data, analysing strains from Teresina, Piauí State. We observed that *L. infantum* genome was constitutively expressed and 88 novel putative genes were found. Our analyses will guide future studies to better understand the molecular mechanisms of *Leishmania*-host interaction.

Keywords: RNA-seq, Trypanosomatidae, *de novo* assembly, next-generation sequencing

Background

Visceral leishmaniasis (VL) is a severe human disease that affects organs and viscera of the lymphohematopoetic system, *e.g.,* liver and spleen [1,2] and can be potentially fatal when untreated [3]. VL is widespread in the World [4] and in American continent is caused by *Leishmania infantum,* a transmitted trypanosomatid to human and dogs (major disease domestic reservoir) by the bite of the infected females *Lutzomyia* [5].

Nowadays, next-generation sequencing (NGS) advent has enabled it possible to unravel deeply the functional mechanisms of the human-pathogens parasites genomes, that has helped to understand at the molecular level the host-parasite interaction [6]. *L. infantum* genome has around 33 Mb length organized in 36 chromosomes and was sequenced in 2007 using shotgun approach and resequenced using next-generation sequencing (NGS) technology in 2011 [2,7]. However, *L. infantum* transcriptome was not analysed using NGS technology until then, as has been done for another trypanosomatids that had their genomes also sequenced since the second decade of the 2000s [8–11]; limiting *L. infantum* transcriptomic analyses

made until then to the use of microarrays, that have a lower accuracy than RNA-sequencing (RNA-seq) [12].

Herein, through RNA-seq the *L. infantum* transcriptome from eight strains isolated and cultured from VL patients that either cured or had confirmed death after diagnosis were analysed. A mean of 26,410 *de novo* assembled transcripts were obtained, of these, around 30% and 0.6-6.6%, aligned against reference genome in regions unannotated and unaligned, respectively, then were used for novel genes prediction. In addition, about 90% of genome were expressed constitutively and gene abundance was calculated.

Material and methods

Ethical approval

The protocol and informed consent, obtained from all the participants or their legal guardian, were approved by the Research Ethics Committee (CEP) of the Federal University of Piauí under the number 0116/2005. The study was conducted in accordance with the principles of the Declaration of Helsinki in resolution 196/96 of the National Health Council of the Ministry of Health that regulates research involving human beings in Brazil.

Leishmania's origin and culture

L. infantum Teresina strain (MHOM/BR/2016/3075; MHOM/BR/2016/3967; MHOM/BR/2016/3991; MHOM/BR/2016/4018; MHOM/BR/2016/4084; MHOM/BR/2016/4090; MHOM/BR/2016/5564; MHOM/BR/2016/5570) were isolated from bone marrow aspirate come from group of eight patients with two clinical outcomes: five non-fatal cases (NFC), *i.e.*, diagnosed with VL and treated; and three fatal cases (FTC), *i.e.*, had a confirmed death after being diagnosed with VL. The collections were made by the Tropical Diseases Institute "Natan Portella" (IDTNP), Teresina, Piauí State, Brazil, and are kept cryo-preserved in liquid nitrogen according to established protocols [13].

L. infantum Teresina strains were reactivated and cultured in Schneider's medium (Sigma-Aldrich, Darmstadt, DE) at pH 7.2 supplemented with inactivated fetal bovine serum (10%), human urine (2%) and maintained at 26 °C. Expansion of the promastigotes for approximately 7 days was monitored and quantified by counting in

Neubauer's chamber reaching a mean density of 4.2 x 10^7 cells/ml in the log phase of growth.

RNA isolation and cDNA preparation

Cells were washed three times with 0.9% NaCl solution for removal from the medium and precipitated by centrifugation at 3,000 rpm for 10 minutes at 4 °C. The supernatant was discarded and the pellet was resuspended in RLT lysis buffer (RNeasy® Mini Kit – Qiagen, Hilden, DE). To ensure a better purity, the extracted RNA was treated with DNase (Qiagen, Hilden, DE) and purified with the RNeasy® Mini Kit (Qiagen, Hilden, DE) according to the manufacturer's instructions. Quantification of the samples was performed by absorbance at 260 nm using the Nanodrop ND 2000 spectrophotometer (Thermo Scientific, Wilmington, DE) and the RNA integrity assessment was checked on an Agilent 2100 bioanalyzer (Agilent Technologies, Palo Alto, CA) according to the manufacturer's instructions.

The cDNA libraries were generated using the TruSeq RNA Sample Prep Kit v2 (Illumina, San Diego, CA) according to manufacturer's recommendations.

RNA-seq data generation, pre-processing and *de novo* transcriptome assembly

Paired-end reads (101 bp) were obtained from the Illumina HiSeq 2500 platform at the Laboratório Central de Tecnologias de Alto Desempenho em Ciências da Vida (LaCTAD) from the University of Campinas, Brazil.

Raw reads quality metrics were assessed using FastQC [14], FASTX-Toolkit (v. 0.0.13) [15] was used to remove any remaining Illumina adapter sequences from reads, then clipped reads were repaired using BBTools (www.jgi.doe.gov/data-and-tools/bbtools). The *de novo* transcriptome assemblies were performed using rnaSPAdes (v. 3.10.1) with default parameters [16]. All assemblies were evaluated by rnaQUAST (v 1.4.0) [17] using as database reference the *L. infantum* JPCM5 genome (release 9.0) available in TriTrypDB (http://tritrypdb.org/tritrypdb/).

Novel genes detection and gene abundance

Unannotated transcripts and unaligned transcripts, from rnaQUAST result, were used for novel genes prediction. Due to the considerable number of unannotated transcripts, we filtered by average length of *L. infantum* genes (≥ 1,865 bp). After, Companion tool [18] was ran for annotation and prediction of these transcripts. Manually, sequences

that had no start and stop codon were discarded. Posteriorly, remaining predicted sequences were aligned against the non-redundant (NR) GenBank database using the BLAST tool (https://blast.ncbi.nlm.nih.gov/).

To evaluate the expressed genes abundance in each sample, using Cufflinks (v 2.2.1) [19] were calculate FPKM values (fragments per kilobase of transcript per million mapped reads) aligning reads (clipped and repaired) against *L. infantum* JPCM5 reference genome.

Results and discussion

RNA sequencing and *de novo* assembly

A total of five NFC *Leishmania* promastigote biological replicates and three FTC biological replicates were collected and sequenced using Illumina HiSeq 2500 platform. A total of ~123.9 million short sequence reads (101 bp) were produced across the eight samples. All the raw sequencing reads were deposited into the Short Read Archive (SRA) of the National Center for Biotechnology Information (NCBI) and can be accessed under the accession number SRR6282525 and SRP124821.

Currently, several studies have used NGS technologies to analyse parasites genomes and/or transcriptome [8,10,11,20]. However, *L. infantum* does not have its transcriptome totally sequenced using RNA-seq until then. Moreover, most *Leishmania* RNA-seq analysis has been carried out using reference assembly strategy, whereas we opted for *de novo* assembly strategy, which can allow a better reconstruction of regions unannotated in the reference genome.

For each sample, the reads were *de novo* assembled using rnaSPAdes and Table 1 displays the main evaluated assembly's metrics by rnaQUAST. The number of assembled transcripts varied between 19,840–36,676 (average 26,410 transcripts), number greater than the of annotated coding-genes for *Le. infantum* JPCM5 reference genome (8,331 genes available in TriTrypDB release 9.0). These high values may be due to more than one assembled transcript can reassemble an annotated gene, as may be seen in the Figure 1B. Moreover, database coverage has reached values higher than > 90% in all assemblies, similar values found by Rastrojo et al. (2013) analysing *Le. major* promastigotes RNA-seq, suggesting good sequencing coverage. The possibility of contamination can be ruled out because of the small number of assembled transcripts that are unaligned with the reference (Table 1) and, as observed

by Rastrojo and co-workers (2013), the lack of mapped reads throughout chromosome indicates that there is no DNA contamination.

Table 1 Main evaluated metrics assemblies from rnaQUAST for eight *Leishmania infantum* Teresina strains.

Metrics	FTC strain			NFC strain				
	3075 (%)	**3967 (%)**	**3991 (%)**	**5570 (%)**	**5564 (%)**	**4018 (%)**	**4090 (%)**	**4084 (%)**
Genes	8,381	8,381	8,381	8,381	8,381	8,381	8,381	8,381
Avg. number of exons per isoform	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01
Transcripts	24,001	36,676	28,739	25,398	29,985	20,049	19,840	24,184
Transcripts > 500 bp	15,158 (63.2%)	17,280 (47.1%)	16,123 (56.1%)	15,153 (59.7%)	15,913 (53.1%)	13,163 (65.7%)	13,377 (67.4%)	14,897 (61.6%)
Transcripts > 1000 bp	10,666 (44.4%)	9,279 (25.3%)	10,333 (36.0%)	10,236 (40.3%)	9,774 (32.6%)	10,319 (51.5%)	10,407 (52.5%)	10,526 (43.5%)
Aligned	23,769 (99.0%)	36,501 (99.5%)	28,119 (97.8%)	25,146 (99.0%)	29,766 (99.3%)	18,728 (93.4%)	19,429 (97.9%)	24,010 (99.3%)
Uniquely aligned	23,008 (95.9%)	35,879 (97.8%)	27,535 (95.8%)	24,303 (95.7%)	28,845 (96.2%)	17,845 (89.0%)	18,644 (94.0%)	23,188 (95.9%)
Multiply aligned	249 (1.0%)	241 (0.7%)	234 (0.8%)	249 (1.0%)	252 (0.8%)	251 (1.3%)	253 (1.3%)	242 (1.0%)
Unaligned	232 (1.0%)	175 (0.5%)	620 (2.2%)	252 (1.0%)	219 (0.7%)	1,321 (6.6%)	411 (2.1%)	174 (0.7%)
Avg. aligned fraction	98.6%	98.8%	98.8%	98.5%	98.6%	98.4%	98.4%	98.6%
Avg. alignment length	1,221.2	740.8	991.2	1,116.8	922.4	1,501.4	1,486.0	1,191.5
Avg. mismatches per transcript	0.261	0.188	0.215	0.226	0.207	0.29	0.247	0.262

Misassemblies	154	99	87	147	161	196	173	157
Database coverage	91.0%	91.5%	92.3%	91.0%	90.2%	90.6%	90.6%	90.8%
50%-assembled isoforms	6,583 (78.5%)	5,787 (69.0%)	6,460 (77.1%)	6,488 (77.4%)	6,223 (74.3%)	6,915 (82.5%)	6,815 (81.3%)	6,560 (78.3%)
95%-assembled isoforms	4,508 (53.8%)	3,160 (37.7%)	4,116 (49.1%)	4,528 (54.0%)	3,914 (46.7%)	5,501 (65.6%)	5,206 (62.1%)	4,549 (54.3%)
50%-covered isoforms	7,302 (87.1%)	7,742 (92.4%)	7587 (90.5%)	7,501 (89.5%)	7,553 (90.1%)	7,257 (86.6%)	7243 (86.4%)	7359 (87.8%)
95%-covered isoforms	6,736 (80.4%)	6,346 (75.7%)	6,736 (80.4%)	6,676 (79.7%)	6,312 (75.3%)	6,805 (81.2%)	6,798 (81.1%)	6,714 (80.1%)
Mean isoform coverage	95.7%	95.0%	95.9%	95.8%	95.0%	96.6%	96.5%	95.8%
Mean isoform assembly	83.2%	71.6%	79.4%	81.6%	77.5%	89.7%	87.8%	83.1%
Unannotated	8,706 (36.3%)	11,639 (31.7%)	9,876 (34.4%)	7,770 (30.6%)	9,542 (31.8%)	6,373 (31.8%)	6,631 (33.4%)	8,301 (34.3%)

Avg.: average.

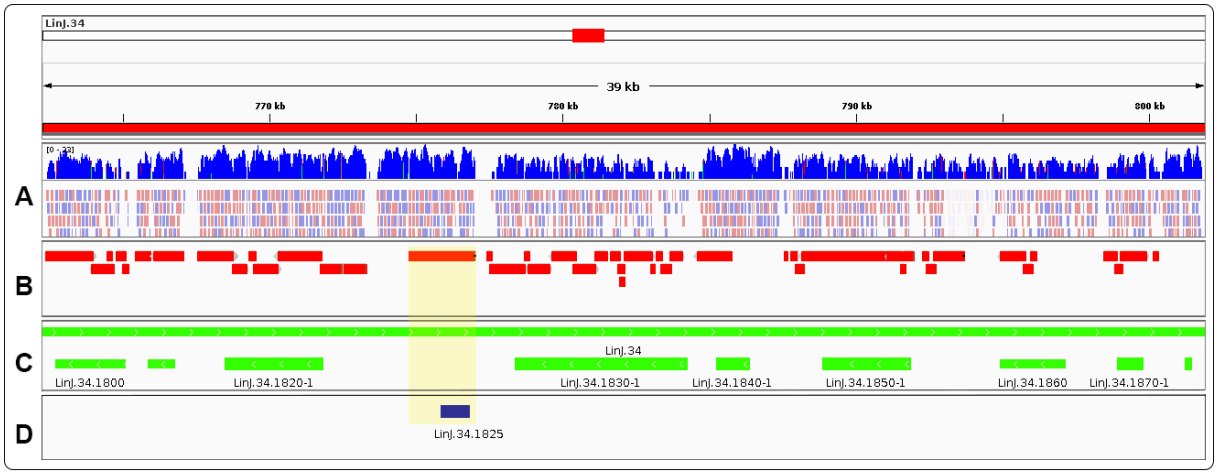


Figure 1: *Leishmania infantum* chromosome 34 partial representation. (A) *L. infantum* Teresina strain RNA-seq mapped reads against *L. infantum* JPCM5 reference genome; light blue bars indicate reverse reads; pink ones indicate forward reads; and the coverage is represented by blue waves. (B) Assembled transcripts by rnaSPAdes. (C) *L. infantum* JPCM5 chromosome 34 (continuous line) and annotated genes. (D) Putative novel gene unannotated in reference genome detected through RNA-seq data.

Gene abundance and novel putative genes

Expressed genes abundance was calculated by FPKM values for each assembly and Supplementary Table S1 showed the top 50 genes most abundant. It was possible to observe that 50 most abundant genes were similar among samples, even between FTC and NFT strains. That was expected because the sequenced samples were derived from the medium culture, *i.e.*, they were in the same condition during RNA isolation, furthermore differential expression analysis was not the initial objective of this work. Histones, β-tubulin and α-tubulin were most representative genes by FPKM values.

Histones are a part of the nucleosomal structure of eukaryotic chromatin, that are related to several activities, such as chromatin condensation, transcriptional activation, replication timing, and DNA repair [21]. Has been observed which *L. infantum* histones genes are up-regulated when in log phase, whereas are strongly down-regulated when parasites enter the stationary growth phase [22–24], confirming the involvement of these genes with the parasite proliferation phase, as can be noted in our results. In trypanosomatids, β- and α-tubulin are highly conserved proteins that interact with each other to form heterodimers, that in turn, polymerize to form microtubules [25]. Microtubules are responsible for several vital cell activities, such as the formation of mitotic spindle, as well cell shape and motility [26]. Therefore, microtubules have been indicated as targets for drugs action for the treatment of *L. infantum* infections [27].

Notwithstanding, as observed by Rastrojo et al. (2013) in *L. major* transcriptome, approximately, 20 out of 50 most expressed genes are annotated as ribosomal proteins. At the time, the authors suggested that a direct correlation between transcript levels and protein abundance could be a general rule in *Leishmania*. Our data corroborate such observations and reinforce this hypothesis.

Our data showed that a mean of 90% of *L. infantum* genes were constitutively expressed among eight strains. On the other hand, around 30% of *de novo* assembled transcripts were classified as unannotated by rnaQUAST, *i.e.*, the transcripts map against the reference genome, but not in a genic region (Table 1). In addition, between 0.5 to 6.6% of transcripts did not map against the reference, so were classified as unaligned (Table 1). Then, we treat as abovementioned in Material and Methods and 62 putative novel genes were found from unannotated transcripts and 26 ones from unaligned transcripts; In Table 2, *L. infantum* JPCM5 genes distributed per chromosomes and novel genes detected herein are represented. Interestingly, predicted genes from unannotated transcripts had hits with part of the *L. infantum* JPCM5 genome when used the BLASTn tool, however, when we used the BLASTx tool, there were no matches with sequences of *L. infantum*, but mainly with sequences from another *Leishmania* species and trypanosomatids (Supplementary Table S2). That is, are known genomic regions, but hitherto not annotated.

Table 2 *Leishmania infantum* JPCM5 distributed genes per chromosome and novel putative genes detected by *Leishmania infantum* Teresina strains RNA-seq data.

JPCM5 Chromosome	# Genes	# Novel genes
0	48	26*
1	745	0
2	549	0
3	433	0
4	412	1
5	386	0
6	347	2
7	337	1
8	324	2
9	299	0
10	292	1
11	266	2
12	257	1
13	248	2

14	230	3
15	211	1
16	180	1
17	175	0
18	175	3
19	173	0
20	168	0
21	167	0
22	167	1
23	163	4
24	162	1
25	160	0
26	147	0
27	141	2
28	135	2
29	134	2
30	133	1
31	126	10
32	119	4
33	113	7
34	97	3
35	85	2
36	74	3

*These putative genes need genes need to be confirmed

Figure 1B and C displays a single assembled transcript (≥ 1,865 bp) map in an unannotated region of the *L. infantum* JPCM5 chromosome 34; and Figure 1D reveals the novel predicted gene (950 bp) from RNA-seq data. Further details of the novel genes, as well as the genes that flank them, are available in the Supplementary Table S3.

In anothers *Leishmania* spp. studies using NGS technology for transcriptome analysis, novel genes that have not previously been detected have been reported [8,9]. A total of 58 out of 62 novel predicted genes since unannotated transcripts are hypothetical protein with unknow function; whereas 61.5% (16/26) of novel genes predicted from unaligned are hypothetical protein. Recently developed editing genome technologies, such as CRISPR-cas9 [28] can help to understand the unknown function these genes and, consequently, their participation in the parasite-host interaction.

We named the novel genes (from unannotated transcripts) following the standard nomenclature used by *L. infantum* JPCM5 genome, allocating the novel gene between previously annotated ones, *e.g.*, one novel gene was mapped between Linj.34.1820-1

and Linj.34.1830-1 genes, we named the new one as Linj.34.1825 (Figure 1); predicted genes since unaligned transcripts were allocated in chromosome zero named with the prefix LinfTER. However, further analyses are needed to confirm the presence of these 26 novel genes, given that were not mapped the genome.

Conclusions

We present herein the firs report, to date, for *L. infantum* total transcriptome using RNA-seq data, analysing strains from Teresina, Piauí State. We observed that *L. infantum* genome was constitutively expressed in promastigotes in the logarithmic phase of growth. In addition, were found and annotated 88 novel putative genes. In summary, our data available herein, can help and encourage future studies for a better understanding of the molecular mechanisms of parasite-host interaction of the main etiological agent of visceral leishmaniasis in the New World.

References

1. Conceição-Silva F, Alves CR. Leishmanioses do Continente Americano. 1ª edição. Conceição-Silva F, Alves CR, editors. Rio de Janeiro: Editora Fiocruz; 2014.
2. Rogers MB, Hilley JD, Dickens NJ, Wilkes J, Bates P a, Depledge DP, et al. Chromosome and gene copy number variation allow major structural change between species and strains of *Leishmania*. Genome Res. 2011;21: 2129–2142. doi:10.1101/gr.122945.111.Freely
3. de Araújo VEM, Morais MHF, Reis IA, Rabello A, Carneiro M. Early clinical manifestations associated with death from visceral leishmaniasis. PLoS Negl Trop Dis. 2012;6. doi:10.1371/journal.pntd.0001511
4. Alvar J, Vélez ID, Bern C, Herrero M, Desjeux P, Cano J, et al. Leishmaniasis worldwide and global estimates of its incidence. PLoS One. 2012;7. doi:10.1371/journal.pone.0035671
5. Dantas-Torres F. Canine leishmaniosis in South America. Parasit Vectors. 2009;2: 1–8. doi:10.1186/1756-3305-2-S1-S1
6. Cantacessi C, Dantas-Torres F, Nolan MJ, Otranto D. The past, present, and future of *Leishmania* genomics and transcriptomics. Trends Parasitol. Elsevier Ltd; 2015;31: 100–108. doi:10.1016/j.pt.2014.12.012
7. Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail MA, et al.

Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. Nat Genet. 2007;39: 839–847. doi:10.1038/ng2053

8. Rastrojo A, Carrasco-Ramiro F, Martín D, Crespillo A, Reguera RM, Aguado B, et al. The transcriptome of *Leishmania major* in the axenic promastigote stage: transcript annotation and relative expression levels by RNA-seq. BMC Genomics. 2013;14: 223. doi:10.1186/1471-2164-14-223
9. Aoki JI. RNA-seq transcriptional profiling of *Leishmania amazonensis* reveals an arginase-dependent gene expression regulation. 2017; 1–22. doi: 10.1371/journal.pntd.0006026
10. Inbar E, Hughitt VK, Dillon LAL, Ghosh K, El-Sayed NM, Sacks DL. The transcriptome of *Leishmania major* developmental stages in their natural sand fly vector. MBio. 2017;8: 1–18. doi:10.1128/mBio.00029-17
11. Greif G, Ponce de Leon M, Lamolle G, Rodriguez M, Piñeyro D, Tavares-Marques LM, et al. Transcriptome analysis of the bloodstream stage from the parasite *Trypanosoma vivax*. BMC Genomics. 2013;14: 149. doi:10.1186/1471-2164-14-149
12. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10: 57–63. doi:10.1038/nrg2484
13. Brasil. Manual de Vigilância e Controle da Leishmaniose Visceral. 1st ed. Ministério da Saúde. Brasília, DF: Editora do Ministério da Saúde; 2006.
14. Andrews S. FastQC: A quality control tool for high throughput sequence data. FastQC: A quality control tool for high throughput sequence data. 2010. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
15. Gordon A, Hannon GJ, Gordon. FASTX-Toolkit. 2014. Available: http://hannonlab.cshl.edu/fastx_toolkit/
16. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol. 2012;19: 455–477. doi:10.1089/cmb.2012.0021
17. Bushmanova E, Antipov D, Lapidus A, Suvorov V, Prjibelski AD. RnaQUAST: A quality assessment tool for de novo transcriptome assemblies. Bioinformatics. 2016;32: 2210–2212. doi:10.1093/bioinformatics/btw218
18. Steinbiss S, Silva-Franco F, Brunk B, Foth B, Hertz-Fowler C, Berriman M, et al. Companion: a web server for annotation and analysis of parasite genomes. Nucleic Acids Res. 2016;44: W29–W34. doi:10.1093/nar/gkw292

19. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28: 511–515. doi:10.1038/nbt.1621
20. Alonso G, Rastrojo A, López-Pérez S, Requena JM, Aguado B. Resequencing and assembly of seven complex loci to improve the *Leishmania major* (Friedlin strain) reference genome. Parasit Vectors. Parasites & Vectors; 2016;9: 74. doi:10.1186/s13071-016-1329-4
21. Kumar D, Rajanala K, Minocha N, Saha S. Histone H4 lysine 14 acetylation in *Leishmania donovani* is mediated by the MYST-family protein HAT4. Microbiology. 2012;158: 328–337. doi:10.1099/mic.0.050211-0
22. Soto M, Requena JM, Gomez LC, Navarrete I, Alonso C. Molecular characterization of a *Leishmania donovani infantum* antigen identified as histone H2A. Eur J Biochem. 1992;205: 211–216. doi:10.1111/j.1432-1033.1992.tb16770.x
23. Soto M, Requena JM, Quijada L, Alonso C. Organization, transcription and regulation of the *Leishmania infantum* histone H3 genes. Biochem J. 1996;318 (Pt 3: 813–819. Available: http://www.ncbi.nlm.nih.gov/pubmed/8836124
24. Soto M, Quijada L, Alonso C, Requena JM. Molecular cloning and analysis of expression of the *Leishmania infantum* histone H4 genes. Mol Biochem Parasitol. 1997;90: 439–447. doi:S0166-6851(97)00178-3 [pii]
25. Kohl L, Gull K. Molecular architecture of the trypanosome cytoskeleton. Mol Biochem Parasitol. 1998;93: 1–9. doi:10.1016/S0166-6851(98)00014-0
26. Hawkins T, Mirigian M, Selcuk Yasar M, Ross JL. Mechanics of microtubules. J Biomech. Elsevier; 2010;43: 23–30. doi:10.1016/j.jbiomech.2009.09.005
27. Escudero-Martínez JM, Pérez-Pertejo Y, Reguera RM, Castro MÁ, Rojo MV, Santiago C, et al. Antileishmanial activity and tubulin polymerization inhibition of podophyllotoxin derivatives on *Leishmania infantum*. Int J Parasitol Drugs Drug Resist. 2017;7: 272–285. doi:10.1016/j.ijpddr.2017.06.003
28. Doudna JA, Charpentier E. The new frontier of genome engineering with CRISPR-Cas9. Science (80-). 2014;346: 1258096–1258096. doi:10.1126/science.1258096

6 CONCLUSÕES GERAIS

Nós analisamos, pela primeira, o transcriptoma total da *Leishmania infantum* através de RNA-seq. Inicialmente, para fins de optimização e estabelecimento de protocolos destinados a tratamento e montagem de dados de RNA-seq de *Leishmania*, nós comparamos seis montadores *de novo* de transcriptomas descritos e bem aceitos na literatura, além de cinco estratégias de tratamento (*pipelines*) de sequencias brutas distintos. Nós observamos que ao depender do montador e da abordagem escolhida, os resultados podem variar, sendo os montadores rnaSPAdes, IDBA-tran e Trinity àqueles que melhores foram avaliados de acordo com os critérios aqui escolhidos, enquanto que *pipeline* isento de *trimming* e *filtering* apresentou bons resultados em relação aos outros comumente relatadas na literatura.

O genoma da *L. infantum* cepa Teresina foi constitutivamente expresso, com uma média superior a 90% entre as cepas. Foi possível observar que entre os genes mais expressos, histonas, β-tubulina e α-tubulina foram mais abundantes, confirmando a alta-regulação destes durante a fase de proliferação do parasito (fase logarítmica de crescimento). Além disso, observamos ainda, que, em torno de 20 dos 50 genes mais abundantes estão genes com função ribossomal, ratificando a hipótese de uma correlação direta entre o nível de transcrição e produção de proteínas em *Leishmania* seja uma regra geral. Além do mais, nós utilizamos os dados de RNA-seq para propor 88 novos genes até então não anotados no genoma de referência da *L. infantum*.

Por fim, o trabalho desenvolvido e apresentado aqui, além de analisar o transcriptoma total de *L. infantum* utilizando tecnologia NGS, poderá servir como base para futuros estudos que busquem fornecer um melhor entendimento dos mecanismos moleculares utilizados por este parasito para sobrepujar as barreiras fisiológicas de seus hospedeiros, podendo assim, subsidiar novos alvos para tratamento e controle das leishmanioses.

REFERÊNCIAS

AKHOUNDI, M.; KUHLS, K.; CANNET, A.; VOTÝPKA, J.; MARTY, P.; DELAUNAY, P.; SERENO, D. A Historical Overview of the Classification, Evolution, and Dispersion of *Leishmania* Parasites and Sandflies. **PLoS Neglected Tropical Diseases**, v. 10, n. 3, p. 1–40, 2016.

ALCOLEA, P. J.; ALONSO, A.; GOMEZ, M. J.; SANCHEZ-GOROSTIAGA, A.; MORENO-PAZ, M.; GONZALEZ-PASTOR, E.; TORANO, A.; PARRO, V.; LARRAGA, V. Temperature increase prevails over acidification in gene expression modulation of amastigote differentiation in *Leishmania infantum*. **BMC Genomics**, v. 11, n. 1, p. 31, 2010.

ALCOLEA, P. J.; ALONSO, A.; SÁNCHEZ-GOROSTIAGA, A.; MORENO-PAZ, M.; GÓMEZ, M. J.; RAMOS, I.; PARRO, V.; LARRAGA, V. Genome-wide analysis reveals increased levels of transcripts related with infectivity in peanut lectin non-agglutinated promastigotes of *Leishmania infantum*. **Genomics**, v. 93, n. 6, p. 551–564, 2009.

ALEXANDER, J.; RUSSELL, D. G. The interaction of *Leishmania* species with macrophages. **Advances in parasitology**, v. 31, p. 175–254, 1992.

ALONSO, G.; RASTROJO, A.; LÓPEZ-PÉREZ, S.; REQUENA, J. M.; AGUADO, B. Resequencing and assembly of seven complex loci to improve the *Leishmania major* (Friedlin strain) reference genome. **Parasites & Vectors**, v. 9, n. 1, p. 74, 2016.

ALVAR, J.; VÉLEZ, I. D.; BERN, C.; HERRERO, M.; DESJEUX, P.; CANO, J.; JANNIN, J.; DE BOER, M. Leishmaniasis worldwide and global estimates of its incidence. **PLoS ONE**, v. 7, n. 5, 2012.

ALVAR, J.; YACTAYO, S.; BERN, C. Leishmaniasis and poverty. **Trends in Parasitology**, v. 22, n. 12, p. 552–557, 2006.

ANDREWS, S. **FastQC: A quality control tool for high throughput sequence data**, 2010. Disponível em: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.

ASLETT, M.; AURRECOECHEA, C.; BERRIMAN, M.; BRESTELLI, J.; BRUNK, B. P.; CARRINGTON, M.; DEPLEDGE, D. P.; FISCHER, S.; GAJRIA, B.; GAO, X.; GARDNER, M. J.; GINGLE, A.; GRANT, G.; HARB, O. S.; HEIGES, M.; HERTZ-FOWLER, C.; HOUSTON, R.; INNAMORATO, F.; IODICE, J.; KISSINGER, J. C.; KRAEMER, E.; LI, W.; LOGAN, F. J.; MILLER, J. A.; MITRA, S.; MYLER, P. J.; NAYAK, V.; PENNINGTON, C.; PHAN, I.; PINNEY, D. F.; RAMASAMY, G.; ROGERS, M. B.; ROOS, D. S.; ROSS, C.; SIVAM, D.; SMITH, D. F.; SRINIVASAMOORTHY, G.; STOECKERT, C. J.; SUBRAMANIAN, S.; THIBODEAU, R.; TIVEY, A.; TREATMAN, C.; VELARDE, G.; WANG, H. TriTrypDB: A functional genomic resource for the Trypanosomatidae. **Nucleic Acids Research**, v. 38, n. SUPPL.1, p. 457–462, 2009.

BANKEVICH, A.; NURK, S.; ANTIPOV, D.; GUREVICH, A. A.; DVORKIN, M.; KULIKOV, A. S.; LESIN, V. M.; NIKOLENKO, S. I.; PHAM, S.; PRJIBELSKI, A. D.; PYSHKIN, A. V.; SIROTKIN, A. V.; VYAHHI, N.; TESLER, G.; ALEKSEYEV, M. A.; PEVZNER, P. A. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. **Journal of Computational Biology**, v. 19, n. 5, p. 455–477, 2012.

BATES, P. A. Transmission of *Leishmania* metacyclic promastigotes by phlebotomine sand flies. **International Journal for Parasitology**, v. 37, n. 10, p. 1097–1106, 2007.

BERN, C.; MAGUIRE, J. H.; ALVAR, J. Complexities of assessing the disease burden attributable to leishmaniasis. **PLoS Neglected Tropical Diseases**, v. 2, n. 10, 2008.

BRASIL, Secretaria de Virgilância em Saúde. **SINAN - Sistema de Informação de Agravos e Notificações**. Disponível em: <portalsinan.saude.gov.br>. Acesso em: 01 de dez. 2017.

BUSHMANOVA, E.; ANTIPOV, D.; LAPIDUS, A.; SUVOROV, V.; PRJIBELSKI, A. D. RnaQUAST: A quality assessment tool for de novo transcriptome assemblies. **Bioinformatics**, v. 32, n. 14, p. 2210–2212, 2016.

CALVETE, J. J.; DOMONT, G. B. Omic technologies to fight the neglect. **Journal of Proteomics**, v. 74, n. 9, p. 1483–1484, 2011.

CAMPBELL, D. A.; THOMAS, S.; STURM, N. R. Transcription in kinetoplastid protozoa: Why be normal? **Microbes and Infection**, v. 5, n. 13, p. 1231–1240, 2003.

CANTACESSI, C.; DANTAS-TORRES, F.; NOLAN, M. J.; OTRANTO, D. The past, present, and future of *Leishmania* genomics and transcriptomics. **Trends in Parasitology**, v. 31, n. 3, p. 100–108, 2015.

CLABORN, D. M. The biology and control of leishmaniasis vectors. **Journal of global infectious diseases**, v. 2, n. 2, p. 127–134, 2010.

CONCEIÇÃO-SILVA, F.; ALVES, C. R. **Leishmanioses do continente americano**. 1a ed. Rio de Janeiro: Editora Fiocruz, 2014.

DANTAS-TORRES, F. Canine leishmaniosis in South America. **Parasites & vectors**, v. 2, n. Suppl 1, p. 1–8, 2009.

DANTAS-TORRES, F.; SOLANO-GALLEGO, L.; BANETH, G.; RIBEIRO, V. M.; DE PAIVA-CAVALCANTI, M.; OTRANTO, D. Canine leishmaniosis in the Old and New Worlds: Unveiled similarities and differences. **Trends in Parasitology**, v. 28, n. 12, p. 531–538, 2012.

DE ASSIS, R. R.; IBRAIM, I. C.; NOGUEIRA, P. M.; SOARES, R. P.; TURCO, S. J. Glycoconjugates in New World species of *Leishmania*: Polymorphisms in lipophosphoglycan and glycoinositolphospholipids and interaction with hosts.

Biochimica et Biophysica Acta - General Subjects, v. 1820, n. 9, p. 1354–1365, 2012.

DESJEUX, P. Leishmaniasis: Current situation and new perspectives. **Comparative Immunology, Microbiology and Infectious Diseases**, v. 27, n. 5, p. 305–318, 2004.

DILLON, L. A. L.; OKRAH, K.; HUGHITT, K. V.; SURESH, R.; LI, Y.; FERNANDES, M. C.; BELEW, A. T.; BRAVO, H. C.; MOSSER, D. M.; EL-SAYED, N. M. Transcriptomic profiling of gene expression and RNA processing during *Leishmania major* differentiation. **Nucleic Acids Research**, v. 43, n. 14, p. 6799–6813, 2015.

DOWNING, T.; IMAMURA, H.; DECUYPERE, S.; CLARK, T. G.; COOMBS, G. H.; COTTON, J. A.; HILLEY, J. D.; DONCKER, S. De; MAES, I.; MOTTRAM, J. C.; QUAIL, M. A.; RIJAL, S.; SANDERS, M.; STARK, O.; SUNDAR, S.; VANAERSCHOT, M.; HERTZ-FOWLER, C.; DUJARDIN, J.; BERRIMAN, M. Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance. **Genome Res.**, v. 21, n. 12, p. 2143–2156, 2011.

EL-SAYED, N. M.; MYLER, P. J.; BARTHOLOMEU, D. C.; NILSSON, D.; AGGARWAL, G.; TRAN, A.-N.; GHEDIN, E.; WORTHEY, E. A.; DELCHER, A. L.; BLANDIN, G.; WESTENBERGER, S. J.; CALER, E.; CERQUEIRA, G. C.; BRANCHE, C.; HAAS, B.; ANUPAMA, A.; ARNER, E.; ASLUND, L.; ATTIPOE, P.; BONTEMPI, E.; BRINGAUD, F.; BURTON, P.; CADAG, E.; CAMPBELL, D. A.; CARRINGTON, M.; CRABTREE, J.; DARBAN, H.; DA SILVEIRA, J. F.; DE JONG, P.; EDWARDS, K.; ENGLUND, P. T.; FAZELINA, G.; FELDBLYUM, T.; FERELLA, M.; FRASCH, A. C.; GULL, K.; HORN, D.; HOU, L.; HUANG, Y.; KINDLUND, E.; KLINGBEIL, M.; KLUGE, S.; KOO, H.; LACERDA, D.; LEVIN, M. J.; LORENZI, H.; LOUIE, T.; MACHADO, C. R.; MCCULLOCH, R.; MCKENNA, A.; MIZUNO, Y.; MOTTRAM, J. C.; NELSON, S.; OCHAYA, S.; OSOEGAWA, K.; PAI, G.; PARSONS, M.; PENTONY, M.; PETTERSSON, U.; POP, M.; RAMIREZ, J. L.; RINTA, J.; ROBERTSON, L.; SALZBERG, S. L.; SANCHEZ, D. O.; SEYLER, A.; SHARMA, R.; SHETTY, J.; SIMPSON, A. J.; SISK, E.; TAMMI, M. T.; TARLETON, R.; TEIXEIRA, S.; VAN AKEN, S.; VOGT, C.; WARD, P. N.; WICKSTEAD, B.; WORTMAN, J.; WHITE, O.; FRASER, C. M.; STUART, K. D.; ANDERSSON, B. The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. **Science** (New York, N.Y.), v. 309, n. 5733, p. 409–15, 2005a.

EL-SAYED, N. M.; MYLER, P. J.; BLANDIN, G.; BERRIMAN, M.; CRABTREE, J.; AGGARWAL, G.; CALER, E.; RENAULD, H.; WORTHEY, E. a; HERTZ-FOWLER, C.; GHEDIN, E.; PEACOCK, C.; BARTHOLOMEU, D. C.; HAAS, B. J.; TRAN, A.-N.; WORTMAN, J. R.; ALSMARK, U. C. M.; ANGIUOLI, S.; ANUPAMA, A.; BADGER, J.; BRINGAUD, F.; CADAG, E.; CARLTON, J. M.; CERQUEIRA, G. C.; CREASY, T.; DELCHER, A. L.; DJIKENG, A.; EMBLEY, T. M.; HAUSER, C.; IVENS, A. C.; KUMMERFELD, S. K.; PEREIRA-LEAL, J. B.; NILSSON, D.; PETERSON, J.; SALZBERG, S. L.; SHALLOM, J.; SILVA, J. C.; SUNDARAM, J.; WESTENBERGER, S.; WHITE, O.; MELVILLE, S. E.; DONELSON, J. E.; ANDERSSON, B.; STUART, K. D.; HALL, N. Comparative genomics of trypanosomatid parasitic protozoa. **Science** (New York, N.Y.), v. 309, n. 5733, p. 404–409, 2005b.

FIEBIG, M.; KELLY, S.; GLUENZ, E. Comparative Life Cycle Transcriptomics Revises *Leishmania mexicana* Genome Annotation and Links a Chromosome Duplication with Parasitism of Vertebrates. **PLoS Pathogens**, v. 11, n. 10, 2015.

FLICEK, P.; BIRNEY, E. Sense from sequence reads: methods for alignment and assembly. **Nature Methods**, v. 7, n. 6, p. 479–479, 2010.

GONTIJO, C. M. F.; MELO, M. Leishmaniose visceral no Brasil: quadro atual, desafios e perspectivas. **Revista Brasileira de Epidemiologia**, v. 7, p. 338–349, 2004.

GORDON, A.; HANNON, G. J.; GORDON. **FASTX-Toolkit**. Disponível em: <http://hannonlab.cshl.edu/fastx_toolkit/>.

HAAS, B. J.; PAPANICOLAOU, A.; YASSOUR, M.; GRABHERR, M.; PHILIP, D.; BOWDEN, J.; COUGER, M. B.; ECCLES, D.; LI, B.; MACMANES, M. D.; OTT, M.; ORVIS, J.; POCHET, N.; STROZZI, F.; WEEKS, N.; WESTERMAN, R.; WILLIAM, T.; DEWEY, C. N.; HENSCHEL, R.; LEDUC, R. D.; FRIEDMAN, N.; REGEV, A. De novo transcript sequence recostruction from RNA-Seq: reference generation and analysis with Trinity. **Nature protocols**, v. 8, n. 8, p. 1–43, 2013.

HAILE, S.; PAPADOPOULOU, B. Developmental regulation of gene expression in trypanosomatid parasitic protozoa. **Current Opinion in Microbiology**, v. 10, n. 6, p. 569–577, 2007.

HANDMAN, E.; ELSO, C.; FOOTE, S. Genes and susceptibility to leishmaniasis. **Advances in Parasitology**, v. 59, n. 5, p. 1–75, 2005.

HERWALDT, B. L. **Leishmaniasis**. The Lancet, v. 354, n. 9185, p. 1191–1199, 1999.

HOLZER, T. R.; MCMASTER, W. R.; FORNEY, J. D. Expression profiling by whole-genome interspecies microarray hybridization reveals differential gene expression in procyclic promastigotes, lesion-derived amastigotes, and axenic amastigotes in *Leishmania mexicana*. **Molecular and Biochemical Parasitology**, v. 146, n. 2, p. 198–218, 2006.

IVENS, A. C.; PEACOCK, C. S.; WORTHEY, E. A.; MURPHY, L.; AGGARWAL, G.; BERRIMAN, M.; SISK, E.; RAJANDREAM, M. A.; ADLEM, E.; AERT, R.; ANUPAMA, A.; APOSTOLOU, Z.; ATTIPOE, P.; BASON, N.; BAUSER, C.; BECK, A.; BEVERLEY, S. M.; BIANCHETTIN, G.; BORZYM, K.; BOTHE, G.; BRUSCHI, C. V; COLLINS, M.; CADAG, E.; CIARLONI, L.; CLAYTON, C.; COULSON, R. M.; CRONIN, A.; CRUZ, A. K.; DAVIES, R. M.; DE GAUDENZI, J.; DOBSON, D. E.; DUESTERHOEFT, A.; FAZELINA, G.; FOSKER, N.; FRASCH, A. C.; FRASER, A.; FUCHS, M.; GABEL, C.; GOBLE, A.; GOFFEAU, A.; HARRIS, D.; HERTZ-FOWLER, C.; HILBERT, H.; HORN, D.; HUANG, Y.; KLAGES, S.; KNIGHTS, A.; KUBE, M.; LARKE, N.; LITVIN, L.; LORD, A.; LOUIE, T.; MARRA, M.; MASUY, D.; MATTHEWS, K.; MICHAELI, S.; MOTTRAM, J. C.; MULLER-AUER, S.; MUNDEN, H.; NELSON, S.; NORBERTCZAK, H.; OLIVER, K.; O'NEIL, S.; PENTONY, M.; POHL, T. M.;

PRICE, C.; PURNELLE, B.; QUAIL, M. A.; RABBINOWITSCH, E.; REINHARDT, R.; RIEGER, M.; RINTA, J.; ROBBEN, J.; ROBERTSON, L.; RUIZ, J. C.; RUTTER, S.; SAUNDERS, D.; SCHAFER, M.; SCHEIN, J.; SCHWARTZ, D. C.; SEEGER, K.; SEYLER, A.; SHARP, S.; SHIN, H.; SIVAM, D.; SQUARES, R.; SQUARES, S.; TOSATO, V.; VOGT, C.; VOLCKAERT, G.; WAMBUTT, R.; WARREN, T.; WEDLER, H.; WOODWARD, J.; ZHOU, S.; ZIMMERMANN, W.; SMITH, D. F.; BLACKWELL, J. M.; STUART, K. D.; BARRELL, B.; MYLER, P. J. The genome of the kinetoplastid parasite, *Leishmania major*. **Science**, v. 309, n. 5733, p. 436–442, 2005.

ESPINOSA, O. A.; SERRANO, M. G.; CAMARGO, E. P.; TEIXEIRA, M. M. G.; SHAW, J. J. An appraisal of the taxonomy and nomenclature of trypanosomatids presently classified as *Leishmania* and *Endotrypanum*. **Parasitology**, p. 1–13, 2016.

JOCHIM, R. C.; TEIXEIRA, C. R.; LAUGHINGHOUSE, A.; MU, J.; OLIVEIRA, F.; GOMES, R. B.; ELNAIEM, D.-E.; VALENZUELA, J. G. The midgut transcriptome of *Lutzomyia longipalpis*: comparative analysis of cDNA libraries from sugar-fed, blood-fed, post-digested and *Leishmania infantum chagasi*-infected sand flies. **BMC Genomics**, v. 9, n. 1, p. 15, 2008.

JOSHI, N.; FASS, J. **Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files**, 2011. Disponível em: <https://github.com/najoshi/sickle>.

KAMHAWI, S. Phlebotomine sand flies and *Leishmania* parasites: friends or foes? **Trends in Parasitology**, v. 22, n. 9, p. 439–445, 2006.

KIM, D.; PERTEA, G.; TRAPNELL, C.; PIMENTEL, H.; KELLEY, R.; SALZBERG, S. L. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. **Genome biology**, v. 14, n. 4, p. R36, 2013.

KIMA, P. E. *Leishmania* molecules that mediate intracellular pathogenesis**. Microbes and Infection**, v. 16, n. 9, p. 721–726, 2014.

KUMAR, A. ***Leishmania* and leishmaniasis**. New Delhi, India: Springer, 2013. v. 3

LAINSON, R.; SHAW, J. J. Evolution, classification and geographical distribution. In: PETERS, W. E.; KILLICK-KENDRICK, R. (Ed.). **The leishmaniases in biology and medicine**. Volume I. Biology and epidemiology. London: Academic Press, 1987. p. 120.

LANGMEAD, B.; TRAPNELL, C.; POP, M.; SALZBERG, S. L. Bowtie: An ultrafast memory-efficient short read aligner. **Genome biology**, n. 10, p. R25, 2009.

LEIFSO, K.; COHEN-FREUE, G.; DOGRA, N.; MURRAY, A.; MCMASTER, W. R. Genomic and proteomic expression analysis of *Leishmania* promastigote and amastigote life stages: The *Leishmania* genome is constitutively expressed. **Molecular and Biochemical Parasitology**, v. 152, n. 1, p. 35–46, 2007.

LEPROHON, P.; FERNANDEZ-PRADA, C.; GAZANION, E; MONTE-NETO, R.; OUELLETTE, M. Drug resistance analysis by next generation sequencing in *Leishmania*. **International Journal for Parasitology: Drugs and Drug Resistance**,

v. 5, n. 1, p. 26–35, 2015.

LI, B.; FILLMORE, N.; BAI, Y.; COLLINS, M.; THOMSON, J. a.; STEWART, R.; DEWEY, C. Evaluation of de novo transcriptome assemblies from RNA-Seq data. **Genome biology**, n. 0, p. 1–21, 2014.

LI, Z.; CHEN, Y.; MU, D.; YUAN, J.; SHI, Y.; ZHANG, H.; GAN, J.; LI, N.; HU, X.; LIU, B.; YANG, B.; FAN, W. Comparison of the two major classes of assembly algorithms: Overlap-layout-consensus and de-bruijn-graph. **Briefings in Functional Genomics**, v. 11, n. 1, p. 25–37, 2012.

MARTIN, J. L.; YATES, P. A.; SOYSA, R.; ALFARO, J. F.; YANG, F.; BURNUM-JOHNSON, K. E.; PETYUK, V. A.; WEITZ, K. K.; CAMP, D. G.; SMITH, R. D.; WILMARTH, P. A.; DAVID, L. L.; RAMASAMY, G.; MYLER, P. J.; CARTER, N. S. Metabolic Reprogramming during Purine Stress in the Protozoan Pathogen *Leishmania donovani*. **PLoS Pathogens**, v. 10, n. 2, 2014.

MASLOV, D. A.; PODLIPAEV, S. A.; LUKEŠ, J. Phylogeny of the Kinetoplastida: Taxonomic Problems and Insights into the Evolution of Parasitism. **Memorias do Instituto Oswaldo Cruz**, v. 96, n. 3, p. 397–402, 2001.

MCCONVILLE, M. J.; DE SOUZA, D.; SAUNDERS, E.; LIKIC, V. A.; NADERER, T. Living in a phagolysosome; metabolism of *Leishmania* amastigotes. **Trends in Parasitology**, v. 23, n. 8, p. 368–375, 2007.

MCGETTIGAN, P. A. Transcriptomics in the RNA-seq era. **Current Opinion in Chemical Biology**, v. 17, n. 1, p. 4–11, 2013.

MCNICOLL, F.; DRUMMELSMITH, J.; MÜLLER, M.; MADORE, É.; BOILARD, N.; OUELLETTE, M.; PAPADOPOULOU, B. A combined proteomic and transcriptomic approach to the study of stage differentiation in *Leishmania infantum*. **Proteomics**, v. 6, n. 12, p. 3567–3581, 2006.

MITTRA, B.; ANDREWS, N. W. IRONy OF FATE: Role of iron-mediated ROS in *Leishmania* differentiation. **Trends in Parasitology**, 2013.

MURRAY, A.; FU, C.; HABIBI, G.; MCMASTER, W. R. Regions in the 3’ untranslated region confer stage-specific expression to the *Leishmania mexicana* a600-4 gene. **Molecular and Biochemical Parasitology**, v. 153, n. 2, p. 125–132, 2007.

MURRAY, H. W.; BERMAN, J. D.; DAVIES, C. R.; SARAVIA, N. G. **Advances in Leishmaniasis**. Lancet (London, England), v. 366, n. 9496, p. 1561–77, 2005.

PATINO, L. H.; RAMÍREZ, J. D. RNA-seq in kinetoplastids: A powerful tool for the understanding of the biology and host-pathogen interactions. **Infection, Genetics and Evolution**, v. 49, p. 273–282, 2017.

PEACOCK, C. S.; SEEGER, K.; HARRIS, D.; MURPHY, L.; RUIZ, J. C.; QUAIL, M. A.; PETERS, N.; ADLEM, E.; TIVEY, A.; ASLETT, M.; KERHORNOU, A.; IVENS, A.; FRASER, A.; RAJANDREAM, M.-A.; CARVER, T.; NORBERTCZAK, H.;

CHILLINGWORTH, T.; HANCE, Z.; JAGELS, K.; MOULE, S.; ORMOND, D.; RUTTER, S.; SQUARES, R.; WHITEHEAD, S.; RABBINOWITSCH, E.; ARROWSMITH, C.; WHITE, B.; THURSTON, S.; BRINGAUD, F.; BALDAUF, S. L.; FAULCONBRIDGE, A.; JEFFARES, D.; DEPLEDGE, D. P.; OYOLA, S. O.; HILLEY, J. D.; BRITO, L. O.; TOSI, L. R. O.; BARRELL, B.; CRUZ, A. K.; MOTTRAM, J. C.; SMITH, D. F.; BERRIMAN, M. Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. **Nature Genetics**, v. 39, n. 7, p. 839–847, 2007.

PETERS, N.; SACKS, D. Immune privilege in sites of chronic infection: *Leishmania* and regulatory T cells. **Immunological Reviews**, 2006.

PIMENTA, P. F.; MODI, G. B.; PEREIRA, S. T.; SHAHABUDDIN, M.; SACKS, D. L. A novel role for the peritrophic matrix in protecting *Leishmania* from the hydrolytic activities of the sand fly midgut. **Parasitology**, v. 115, p. 359–369, 1997.

RASTROJO, A.; CARRASCO-RAMIRO, F.; MARTÍN, D.; CRESPILLO, A.; REGUERA, R. M.; AGUADO, B.; REQUENA, J. M. The transcriptome of *Leishmania major* in the axenic promastigote stage: transcript annotation and relative expression levels by RNA-seq. **BMC Genomics**, v. 14, n. 1, p. 223, 2013.

RAYMOND, F.; BOISVERT, S.; ROY, G.; RITT, J. F.; LÉGARÉ, D.; ISNARD, A.; STANKE, M.; OLIVIER, M.; TREMBLAY, M. J.; PAPADOPOULOU, B.; OUELLETTE, M.; CORBEIL, J. 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.

REAL, F.; VIDAL, R. O.; CARAZZOLLE, M. F.; MONDEGO, J. M. C.; COSTA, G. G. L.; HERAI, R. H.; WÜRTELE, M.; DE CARVALHO, L. M.; E FERREIRA, R. C.; MORTARA, R. A.; BARBIÉRI, C. L.; MIECZKOWSKI, P.; DA SILVEIRA, J. F.; BRIONES, M. R. D. S.; PEREIRA, G. A. G.; BAHIA, D. 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.

ROGERS, M. B.; HILLEY, J. D.; DICKENS, N. J.; WILKES, J.; BATES, P. A; DEPLEDGE, D. P.; HARRIS, D.; HER, Y.; HERZYK, P.; IMAMURA, H.; OTTO, T. D.; SANDERS, M.; SEEGER, K.; DUJARDIN, J.-C.; BERRIMAN, M.; SMITH, D. F.; HERTZ-FOWLER, C.; MOTTRAM, J. C. Chromosome and gene copy number variation allow major structural change between species and strains of *Leishmania*. **Genome Res**., v. 21, p. 2129–2142, 2011.

ROSENTHAL, L. A.; SUTTERWALA, F. S.; KEHRLI, M. E.; MOSSER, D. M. *Leishmania major*-human macrophage interactions: Cooperation between Mac-1 (CD11b/CD18) and complement receptor type 1 (CD35) in promastigote adhesion. **Infection and Immunity**, v. 64, n. 6, p. 2206–2215, 1996.

SAXENA, A.; WORTHEY, E. A.; YAN, S.; LELAND, A.; STUART, K. D.; MYLER, P. J. Evaluation of differential gene expression in *Leishmania major* Friedlin procyclics and metacyclics using DNA microarray analysis. **Molecular and Biochemical Parasitology**, v. 129, n. 1, p. 103–114, 2003.

SCHATZ, M. C.; DELCHER, A. L.; SALZBERG, S. L. Assembly of Large Genomes using Cloud Computing. **Genome research**, v. 20, n. 9, p. 1165–1173, 2010.

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, 2007.

SOARES, R. P. P.; MACEDO, M. E.; ROPERT, C.; GONTIJO, N. F.; ALMEIDA, I. C.; GAZZINELLI, R. T.; PIMENTA, P. F. P.; TURCO, S. J. *Leishmania chagasi*: Lipophosphoglycan characterization and binding to the midgut of the sand fly vector *Lutzomyia longipalpis*. **Molecular and Biochemical Parasitology**, v. 121, n. 2, p. 213–224, 2002.

TEIXEIRA, D. E.; BENCHIMOL, M.; RODRIGUES, J. C. F.; CREPALDI, P. H.; PIMENTA, P. F. P.; DE SOUZA, W. The Cell Biology of *Leishmania*: How to Teach Using Animations. **PLoS Pathogens**, v. 9, n. 10, p. 8–11, 2013.

TEMPONE, A. J.; PITALUGA, A. N.; TRAUB-CSEKÖ, Y. M. Aspectos Moleculares da Interação Flebotomíneo-*Leishmania*. In: CONCEIÇÃO-SILVA, F.; ALVES, C. R. (Ed.). **Leishmanioses do continente Americano**. Rio de Janeiro: Editora Fiocruz, 2014. p. 159–180.

TURCO, S. J. Adversarial relationship between the *Leishmania* lipophosphoglycan and protein kinase C of host macrophages. **Parasite Immunology**, v. 21, n. 12, p. 597–600, 1999.

ULIANA, S. R. B.; RUIZ, J. C.; CRUZ, A. K. *Leishmania* Genomics: Where Do We Stand? In: GRUBER, A.; DURHAM, A. M.; HUYNH, C.; PORTILLO, H. A. DEL (Ed.). Bioinformatics in Tropical Disease Research: A Practical and Case-Study Approach. Bethesda (MD): **National Center for Biotechnology Information**, 2008. p. 235–252.

VAIDYANATHAN, R. Isolation of a myoinhibitory peptide from *Leishmania major* (Kinetoplastida: Trypanosomatidae) and its function in the vector sand fly *Phlebotomus papatasi* (Diptera: Psychodidae). **Journal of medical entomology**, v. 42, n. 2, p. 142–152, 2005.

WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics. **Nature Reviews Genetics**, v. 10, n. 1, p. 57–63, 2009.

WERNECK, G. L. Expansão geográfica da leishmaniose visceral no Brasil. **Cadernos de Saúde Pública**, v. 26, n. 4, p. 644–645, 2010.

WILSON, R.; BATES, M. D.; DOSTALOVA, A.; JECNA, L.; DILLON, R. J.; VOLF, P.; BATES, P. A. Stage-specific adhesion of *Leishmania* promastigotes to sand fly midguts assessed using an improved comparative binding assay. **PLoS Neglected Tropical Diseases**, v. 4, n. 9, p. 1–9, 2010.

WRIGHT, S. D.; SILVERSTEIN, S. C. Receptors for C3b and C3bi promote phagocytosis but not the release of toxic oxygen from human phagocytes. **The Journal of experimental medicine**, v. 158, n. 6, p. 2016–23, 1983.

ANEXO A – NORMAS DA REVISTA MEMORIAS DO INSTITUTO OSWALDO CRUZ E COMPROVAÇÃO DE SUBMISSÃO À REVISTA

The Memórias' content is freely accessible to readers and no publication fees are charged to authors. The Memorias do Instituto Oswaldo Cruz has decided to simplify the requirements regarding the format of submitted manuscripts. From now on, all manuscripts may be submitted in any text format as long as the common subdivision of scientific articles are followed, e.g. introduction, materials and methods, results, discussion and references. For Reviews, Perspectives and similar articles, authors may use the sections that best suit the structure and content of the proposed manuscript. All manuscripts should contain, besides the title and abstract, full details of authors and institutions, acknowledgements of any technical or financial assistance as well as state any conflicts of interest. This flexible text format will be used for the initial analysis and peer review. If the manuscript is accepted, authors will be requested to edit the text in accordance with the publication style of the Memorias."

Upon acceptance, the manuscript should be arranged in the following format:

The manuscript should be prepared using standard word processing software and should be printed (font size 12) double-spaced throughout the text, figure captions, and references (must be up to 30 references), with margins of at least 3 cm. The figures should come in the extension tiff, with a minimum resolution of 300 dpi. Tables and legends to figures must be submitted all together in a single file. Figures, must be uploaded separately as supplementary file.

Running title: with up to 40 characters (letters and spaces)

Title: with up to 250 characters

Author's names: without titles or graduations

Institutional affiliations: full address of the corresponding author only

Abstracts: Provide an abstract of between 250- 300 words (100 words in case of short communications, technical notes, genome announcements or reviews). Abstracts of original articles should be structured into 5 sections as follows: BACKGROUND, OBJECTIVES, METHODS, FINDINGS and MAIN CONCLUSIONS, each section addressing respectively the problem, the aim of the study, the main methodological approach, the most important findings and the conclusions of the study.

Key words: 3-6 items must be provided. Terms from the Medical Subject Headings (Mesh) list of Index Medicus should be used.

Sponsorships: indicating the sources of financial support and change of address.

Introduction: should set the purpose of the study, give a brief summary (not a review) of previous relevant works, and state what new advance has been made in the investigation. It should not include data or conclusions from the work being reported.

Materials and Methods: should briefly give clear and sufficient information to permit the study to be repeated by others. Standard techniques need only be referenced.

Ethics: when reporting experiments on human subjects, indicate whether the procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional or regional) and with the Helsinki Declaration of 1975, as revised in 1983. When reporting experiments on animals, indicate whether the institution's or a national research council's guide for, or any national law on the care and use of laboratory animals was followed.

Results: should be a concise account of the new information discovered, with the least personal judgement. Do not repeat in text all the data in the tables and illustrations.

In case of describing New Species, should follow:

Name of the new species, authors (when it is the case), sp. nov., (Figs x-y)
[Ex: *An.* (*Nyssorhynchus*) *atacamensis* González and Sallum, sp. nov. (Figs 1-4)]

Previous reference to the new species (when it is the case)
[Ex: *An. pictipennis* of Rueda et al. (2008): 448.]

Diagnosis (or Description; all stages are described);

Type host (when it is the case);

Site of Infection (when it is the case);

Type-locality;

Type data and depository;

Other material examined (when it is the case);

Distribution;

Host-parasite data (such prevalence and other important data, when it is the same case);

Bionomics;

Etymology;

Taxonomic discussion (or simply DISCUSSION as internal title).

Discussion: should be limited to the significance of the new information and relate the new findings to existing knowledge. Only unavoidable citations should be included.

Acknowledgements: should be short and concise, and restricted to those absolutely necessary.

Author's contribution: state each author's contribution to the work.

References

Must be accurate. Only citations that appear in the text should be referenced. Unpublished papers, unless accepted for publication, should not be cited. Work accepted for publication should be referred to as "in press" and a letter of acceptance of the journal must be provided. Unpublished data should only be cited in the text as "unpublished observations", and a letter of permission from the author must be provided. The references at the end of the paper should be arranged in alphabetic order according to the surname of the first author. CLICK HERE [+]

Figures and tables must be understandable without reference to the text

Figures: presented in tiff format with a minimum of 300 dpi and photographs must be sharply focused, well contrasted, and if mounted onto a plate, the figures should be numbered consecutively with Arabic numbers. Magnification must be indicated by a line or bar in the figure, and referenced, if necessary in the caption (e.g., bar = 1 mm). Plates and line figures should either fit one column (8 cm) or the full width (16.5 cm) of the page and should be shorter than the page length to allow inclusion of the legend. Letters and numbers on figures should be of a legible size upon reduction or printing. A colour photograph illustrates the cover of each issue of the Journal and authors are invited to submit illustrations with legends from their manuscript for consideration for the cover.

Tables: should supplement, not duplicate, the text and should be numbered with Roman numerals. A short descriptive title should appear above each table, with any explanations or footnotes (identified with a, b, c, etc.) below.

Review: Papers in "review" format are accepted only by means of invitations made by the editor or associated editors.

Technical Notes: Technical Notes should communicate rapidly single novel techniques or original technical advances. The entire note should occupy no more than three printed pages including figures and/or tables (it means around 10 double-spaced typed Word file maximum). The text must not be not divided into sections. Therefore, the state of art must be very briefly presented; results must be rapidly presented and discussed at a time. Complementary tables and figures may be published as supplementary data. References must be limited to few essential ones and cited at the end of the note, using the same format as in full papers. A brief summary and three key words must be provided.

Short communications: should communicate rapidly single results or techniques. They should occupy no more than three printed pages including figures and/or tables. They should not contain excessive references. References should be cited at the end of the paper using the same format as in full papers. A brief summary and three key words must be provided.

Genome Announcement and Highlights: this section is dedicated to publish new genome information from eukaryote parasites, virus, bacteria and their respective vectors. Authors who wants a fast peer review and publication cycle for their research results covering new genome sequences, re-sequencing and comparative genome analysis as well as the expression pattern of genomes are invited to submitted papers under the short communication format.

Alternative format: manuscripts may be submitted following the "Uniform Requirements for Manuscripts Submitted to Biomedical Journals" produced by the International Committee of Medical Journal Editors also known as the Vancouver Style. In this case, authors should follow the guidelines in the fifth edition (Annals of Internal Medicine 1997; 126: 36-47, or at the website http://www.acponline.org/journals/resource/unifreqr/htm) and will be responsible for modifying the manuscript where it differs from the instructions given here, if the manuscript is accepted for publication.
Authors should also follow the Uniform Requirements for any guidelines that are omitted in these Instructions.

In case of clinical trials it's mandatory to inform the registration number of the REBEC platform.

A statement that the data/results of the manuscript are not plagiarism and have not been published elsewhere.

Once a paper is accepted for publication, the authors must provide:

Page charges: there will be no page charges.

Proofs: one set of page proofs will be supplied for the author to check for typesetting accuracy, to be returned by the stipulated date. No changes to the original manuscript will be allowed at this stage.

Gmail

Lucas Sousa-Paula <lcsousapaula@gmail.com>

NEW SUBMISSION - Memórias - Manuscript ID MIOC-2017-0503

1 message

Memorias MIOC <onbehalfof@manuscriptcentral.com> Thu, Nov 16, 2017 at 4:51 PM
Reply-To: memorias@fiocruz.br
To: lcsousapaula@gmail.com
Cc: memorias.artigo@fiocruz.br

16-Nov-2017

Dear Dr.(a) Sousa-Paula:

Your manuscript entitled "Comparison of treatment strategies and de novo transcriptome assemblers for Leishmania RNA-seq data analysis" has been successfully submitted online and is presently being given full consideration for publication in the Memórias do Instituto Oswaldo Cruz.

Your manuscript ID is MIOC-2017-0503.

Please mention the above manuscript ID in all future correspondence or when calling the office for questions. If there are any changes in your street address or e-mail address, please log in to ScholarOne Manuscripts at https://mc04.manuscriptcentral.com/mioc-scielo and edit your user information as appropriate.

You can also view the status of your manuscript at any time by checking your Author Center after logging in to https://mc04.manuscriptcentral.com/mioc-scielo.

Thank you for submitting your manuscript to the Memórias do Instituto Oswaldo Cruz.

Sincerely,
Memórias do Instituto Oswaldo Cruz
memorias@fiocruz.br