TALLES EDUARDO FERREIRA MACIEL

TRANSCRIPTOMA DE Leishmania (V.) braziliensis POR RNA-Seq: MONTAGEM DE TRANSCRIPTOMAS, ENRIQUECIMENTO DE ORFEOMA, ANÁLISE DE EXPRESSÃO E ANOTAÇÃO DOS GENES DIFERENCIALMENTE EXPRESSOS

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Bioquímica Agrícola, para obtenção do título de Doctor Scientiae.

VIÇOSA
MINAS GERAIS - BRASIL
2014

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
M152t
2014

Maciel, Talles Eduardo Ferreira, 1981-
Transcriptoma de *Leishmania (V.) braziliensis* por RNA-Seq : montagem de transcriptomas, enriquecimento de orfeoma, análise de expressão e anotação dos genes diferencialmente expressos / Talles Eduardo Ferreira Maciel. – Viçosa, MG, 2014.
x, 92f. : il. ; 29 cm.

Orientador: Juliana Lopes Rangel Fietto.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. *Leishmania braziliensis*. 2. Bioinformática. 3. Expressão diferenical. 4. Transcriptoma. I. Universidade Federal de Viçosa. Departamento de Bioquímica e Biologia Molecular. Programa de Pós-graduação em Bioquímica Agrícola. II. Título.

CDD 22. ed. 579.4

TALLES EDUARDO FERREIRA MACIEL

TRANSCRIPTOMA DE *Leishmania (V.) braziliensis* POR RNA-Seq: MONTAGEM DE TRANSCRIPTOMAS, ENRIQUECIMENTO DE ORFEOMA, ANÁLISE DE EXPRESSÃO E ANOTAÇÃO DOS GENES DIFERENCIALMENTE EXPRESSOS

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Bioquímica Agrícola, para obtenção do título de Doctor Scientiae.

APROVADA: 10 de fevereiro de 2014.

Fábio Ribeiro Cerqueira

Jerônimo Conceição Ruiz

Luciano Gomes Fietto

Moysés Nascimento

Juliana Lopes Rangel Fietto
(Orientadora)

À minha mãe Carmem Albino Ferreira Maciel *(in memoriam)*, pela luta, sacrifício, abdicação, amor incondicional, amizade, educação, incentivo e pelo exemplo de ser humano. Sem você, nada eu seria e nada teria conseguido.

Ao meu pai Tarcis Ferreira Maciel *(in memoriam)* por mostrar-me que "pequenas" situações podem ter um significado imenso, dependendo da maneira como às enxergamos e como enxergamos a vida;

À minha irmã, pela cumplicidade, apoio e torcida;

Ao meu irmão pela amizade, apoio e torcida;

Aos meus cunhados(as) pelos excelentes momentos compartilhados;

Aos meus sobrinhos Luiz Gustavo, Bruna, Júlia, Ana Júlia e Leonardo pelo incentivo transmitido pelo amor e carinho;

A minha avó, pelas orações diárias e pelos constantes ensinamentos;

Ao meu bom Deus, por colocar todos no meu caminho e me manter feliz e fortalecido neste mundo;

A todos os professores, heróis, que passaram por minha vida, desde a minha alfabetização.

Dedico

Ao meu único e verdadeiro amor, pelo carinho, companheirismo e compreensão e por me fazer uma pessoa melhor... Minha doce Kamila,

Aos meus filhos Gabriel e Iasmim, pelo amor incondicional demonstrado pelo carinho, palavras e gestos; que tem revelado o verdadeiro sentido da minha vida.

Ofereço

AGRADECIMENTOS

A Deus, por sua infinita bondade, por não me deixar cair, sempre segurando minha mão;

A Universidade Federal de Viçosa, em especial ao Departamento de Bioquímica e Biologia Molecular, pela oportunidade de realização do curso e sua contribuição para o meu crescimento profissional;

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela concessão da bolsa de estudo.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelo auxílio financeiro;

A professora Juliana Lopes Rangel Fietto pela orientação, paciência, compreensão, apoio, confiança, pelos ensinamentos de dedicação a ciência, pelos ensinamentos éticos e pela amizade.

Ao professor Jerônimo Conceição Ruiz pelo esforço em participar da defesa desta tese e pelas críticas construtivas.

A professor Luciano Gomes Fietto pela participação na defesa desta tese e pelas pertinentes observações.

Ao professor Fábio Ribeiro Cerqueira pela acessibilidade, paciência e sugestões.

Ao professor Moysés Nascimento pelos constantes ensinamentos, por disponibilizar-me o laboratório e pelos momentos de descontração proporcionados.

Ao secretário do programa de Pós-Graduação Stricto Sensu em Bioquímica Agrícola, Eduardo Pereira Monteiro por ajudar sempre que precisei.

Aos amigos do laboratório de Infectologia Molecular Animal, que sempre me ensinaram, ampararam e estimularam em todas as etapas deste trabalho, proporcionando assim momentos que jamais serão esquecidos.

Aos amigos do laboratório de Bioinformática pela agradável convivência.

Aos amigos bioinformatas Otávio, José Cleydison e Pedro pela amizade, ajuda e pelas incansáveis discussões sobre diversos temas de bioinformática.

A todos os colegas de Pós-Graduação da UFV pelos momentos compartilhados.

A todos que direta ou indiretamente ajudaram na realização deste trabalho.

Meus sinceros agradecimentos!!!

SUMÁRIO

INTRODUÇÃO GERAL 1
REVISÃO BIBLIOGRÁFICA 3
Leishmania e leishmaniose 3
Expressão Gênica em *Leishmania* 5
Sequenciamento paralelo maciço 7
Pirosequenciamento 8
Montagem 11
Anotação funcional 12
REFERÊNCIAS BIBLIOGRÁFICAS 14
TRANSCRIPTOMA DE *Leishmania* (*V.*) *braziliensis* POR RNA-Seq: MONTAGEM DE TRANSCRIPTOMAS, ENRIQUECIMENTO DE ORFEOMA, ANÁLISE DE EXPRESSÃO E ANOTAÇÃO DOS GENES DIFERENCIALMENTE EXPRESSOS 20
INTRODUÇÃO 20
MATERIAL E MÉTODOS 22
Estação de trabalho (computadores) 22
Delineamento dos tratamentos 22
Sequenciamento das bibliotecas transcriptômicas 23
Output e conversão de formatos 23
Análise da qualidade das *reads* sequenciadas 27
Tratamento inicial das *reads* 27
Referência utilizada para mapeamento 29
Panorama Geral 30
Mapeamento 30
Montagem das *reads* 30
Avaliação da qualidade da montagem 32
Anotação dos *contigs* 32
Análise estatística 33
RESULTADOS E DISCUSSÃO 36
Bibliotecas sequenciadas 36
Qualidade das *reads* 36
Tratamento das *reads* 39
Critérios para montagem 44

De novo assembly (montagem) 47
Teste de qualidade da montagem 50
Mapeamento para análise de enriquecimento do orfeoma 51
Mapeamento 1 51
Mapeamento 2 52
Mapeamento 3 52
Mapeamento 4 52
Merge dos mapeamentos 52
Erros de montagem e possibilidade de extensão de *gaps* no genoma anotado de *L. (V.) braziliensis* 52
Identificação de conflitos 55
Novas ORFs em *L. (V.) braziliensis* 56
Análise de pseudogenes 58
Extensão de ORFs incompletas 60
Panorama geral através do mapeamento das *reads* 62
Mapeamento das *reads* de cada biblioteca no genoma de *L. (V.) braziliensis* e *L.* (*L.*) *major* 62
Mapeamento das *reads* da biblioteca ET-MET 62
Mapeamento das *reads* da biblioteca ET-PRO 63
Mapeamento das *reads* da biblioteca NSL-MET 63
Mapeamento das *reads* da biblioteca NSL-PRO 64
Análise qualitativa da expressão gênica entre bibliotecas 66
Mapeamento dos *singlets* e *contigs* de cada biblioteca no conjunto de transcritos de *L. (V.) braziliensis* 66
ORFs exclusivas de cada bibliotecas 67
Expressão Gênica Diferencial em *L. (V.) braziliensis* 69
Anotação dos *contigs* diferencialmente expressos 80
Clusterização dos tratamentos 88
CONCLUSÕES 89
REFERÊNCIAS BIBLIOGRÁFICAS 90

RESUMO

MACIEL, Talles Eduardo Ferreira, D.Sc., Universidade Federal de Viçosa, fevereiro de 2014. **Transcriptoma de** Leishmania (V.) braziliensis **por RNA-Seq: montagem de transcriptomas, enriquecimento de orfeoma, análise de expressão e anotação dos genes diferencialmente expressos.** Orientadora: Juliana Lopes Rangel Fietto. Coorientadores: Abelardo Silva Júnior, Gustavo Costa Bressan, Márcia Rogéria de Almeida Lamego e Luís Carlos Crocco Afonso.

Os parasitos do gênero Leishmania, que causam um amplo espectro de desordens clínicas referidas comumente como leishmanioses, são um grande problema de saúde pública em vários países. A leishmaniose tegumentar americana está entre as endemias de maior importância em saúde pública no Brasil, devido a fatores como: ampla distribuição pelo território nacional, ocorrência de formas clínicas graves e limitações referentes tanto ao diagnóstico como ao tratamento, sendo a L. (V.) braziliensis uma das principais espécies de importância epidemiológica para a LTA no Brasil. Atualmente existem diversas tecnologias que permitem o sequenciamento do DNA em larga escala, sendo a plataforma 454/Roche utilizada neste trabalho. Assim, este trabalho utilizou ferramentas de bioinformática para montar e analisar o transcriptoma de L. (V.) braziliensis através do sequenciamento do transcriptoma de dois isolados (ET e NSL), que apresentam diferença significativa na virulência em modelo murino. Foram preparadas duas formas evolutivas para cada isolado: metacíclica (MET) e procíclica (PRO). Desta forma foram analisadas quatro bibliotecas. Após sequenciamento, os dados foram visualizados com o programa fastQC, tratados com FASTX-Tollkit e Prinseq-Lite e montados com programa Newbler. A montagem (Assembly) foi efetuada de duas maneiras distintas: primeiro efetuou-se a montagem com as reads de cada biblioteca e posteriormente, as reads das quatro bibliotecas foram alocados em arquivo único para realização de um novo assembly. As open reading frame (ORFs), que são regiões com potencial para codificar proteínas, foram preditas utilizando as sequências resultantes da montagem. A anotação foi efetuada através de duas abordagens: transferência de informações do genoma anotado automaticamente para as ORFs preditas e pela abordagem baseada em homologia de sequências através da ferramenta de anotação funcional Blast2GO. Após anotação, efetuou-se a análise da expressão gênica diferencial através de duas abordagens diferentes: a primeira, utilizou o método de Blind do pacote DESeq do R/Bioconductor e a segunda utilizou uma abordagem baseada em RPKM. Foram produzidas 3.095.724 reads, sendo 916.546, 589.554, 1.083.312 e 506.312 sequências para ET-MET (biblioteca 1), ET-PRO (biblioteca 2), NSL-MET (biblioteca 3) e NSL-PRO (biblioteca 4), respectivamente. Após o tratamento, utilizou-se para o restante das análises 2.899.230 sequências. Com o

intuito de validar algumas das análises, foi utilizado neste trabalho um segundo conjunto de reads (Illumina) baixado do banco de dados SRA (Sequence Read Archive) indexado ao NCBI, sendo este composto por 52.014.768 de reads paired end. Após o tratamento, utilizou-se para o restante das análises 47.377.233 de reads. Os resultados das análises com as reads sequenciadas neste trabalho e com os contigs montados, tal como o mapeamento destes no genoma anotado de L. (V.) braziliensis, produziu novas informações ao orfeoma anotado automaticamente de L. (V.) braziliensis. Após montagem, obteve-se 14.362, 13.145, 14.899 e 11.434 contigs maiores que 100 pb para as bibliotecas 1, 2, 3 e 4, respectivamente. Obteve-se como resultado da montagem, considerando as reads de todas as bibliotecas, 14.017 contigs. As ORFs preditas à partir dos contigs que não mapearam no genoma anotado foram utilizados para busca de novos genes de L. (V.) braziliensis. Como resultado, foi possível encontrar seis novos genes, 117 possíveis ORFs sem hits no banco de dados nr e 85 ORFs que, por algum motivo, deixaram de fazer parte do genoma anotado. Foram encontrados, ao se comparar as reads obtidas neste trabalho com o genoma anotado, 6.293 sítios com identidades diferentes, que pode ser devido a divergência alélica entre os isolados analisados ou devido ao polimorfismos de nucleotídeo único (SNPs).

ABSTRACT

MACIEL, Talles Eduardo Ferreira, D.Sc., Universidade Federal de Viçosa, February, 2014. **Transcriptome of** L. (V.) braziliensis **by RNA-Seq: assembly of transcriptomas, enrichment of orfeoma, expression analysis and annotation of differentially expressed genes.** Adviser: Juliana Lopes Rangel Fietto. Co-advisers: Abelardo Silva Júnior, Gustavo Costa Bressan, Márcia Rogéria de Almeida Lamego and Luís Carlos Crocco Afonso.

Parasites of the genus Leishmania, which cause a broad spectrum of clinical disorders referred to commonly as leishmaniasis, are a major public health problem in many countries. American cutaneous leishmaniasis is among the endemic most important in public health in Brazil, due to factors such as: wide distribution throughout the country, the occurrence of severe clinical forms and limitations relating to both diagnosis and treatment, with L. (V.) braziliensis being one of the main species of epidemiological significance to the LTA in Brazil. Currently there are several technologies that allow the DNA sequencing in large scale, being the 454/Roche platform used in this work. Thus, this study used bioinformatics tools for assembly and analyze the transcriptome of L. (V.) braziliensis through transcriptome sequencing of two isolates (ET and NSL), which present significant difference in virulence in murine model. Were prepared two evolutionary forms for each isolate: metacyclic (MET) and procyclical (PRO). Thus, four libraries were analyzed. After sequencing, the data were visualized with fastQC program, treated with FASTX-Tollkit and Prinseq-Lite and assembly with Newbler v.2.5.3 program. The assembly was conducted of two distinct ways: first performed the assembly whit the reads from each sample and then, the reads of the four samples were placed in single file to perform a new assembly. The open reading frame (ORF), which are regions with potential to encode a protein were predicted using the resulting assembly. The annotation was carried out using two approaches: transfer of information of automatically annotated genomic to predicted ORFs and by approach based on sequence homology by functional annotation tool Blast2GO. After annotation, performed the analysis of differential gene expression by two different approaches: first, was used the Blind method of DESeq package the R/Bioconductor and the second was used an approach based on RPKM. 3.095.724 reads were produced, with 916.546, 589.554, 1.083.312 and 506.312 sequences for ET-MET (sample 1), ET-PRO (sample 2), NSL-MET (sample 3) and NSL-PRO (sample 4), respectively. After treatment, was used for the remaining analysis 2.899.230 sequence. In order to validate some of the analysis, was used in this study, a second set of reads (Illumina) downloaded from the database SRA (Archive Sequence Read) indexed to NCBI, this being composed of 52.014.768 of reads paired end. After treatment, was used for

the remainder analysis 47.377.233 of reads. The results of the analysis with the reads sequenced this work and with the assembly contigs, such as mapping of these in annotated genome the L. (V.) braziliensis, produced new information to automatically annotated orfeoma of L. (V.) braziliensis. After assembly, we obtained 14.362, 13.145, 14.899 and 11.434 contigs larger than 100 bp for samples 1, 2, 3 and 4, respectively. It was obtained as a result of assembly, considering the reads from all samples, 14.017 contigs. The ORFs predicted from contigs not mapped the annotated genome were used to search for new genes of L. (V.) braziliensis. So, were found six new genes, 117 ORFs possible without hits in the nr database and 85 ORFs that, for some reason, no longer in the annotated genome. Were found, when comparing the reads obtained in this work with the annotated genome, 6.293 sites with different identities, which may be due to the allelic divergence between the isolates analyzed or due to single nucleotide polymorphisms (SNPs).

INTRODUÇÃO GERAL

As leishmanioses são um espectro de doenças de caráter zoonótico e sua transmissão ocorre de forma natural pela picada de fêmeas infectadas de mais de trinta espécies de flebotomíneos. Essa doença acomete o homem desde tempos remotos e vem proporcionando, nos últimos anos, um aumento do número de casos e ampliação de sua distribuição geográfica (Brandao-Filho et al., 2011). A leishmaniose é endêmica em 88 países distribuídos em quatro continentes, tendo altos índices de incidência em várias regiões do mundo, com cerca de 20 milhões de indivíduos infectados e 350 milhões de pessoas vivendo em área de risco (Do Monte-Neto et al., 2011). A cada ano são relatados mais de 1,5 milhões de novos casos de leishmaniose cutânea (Kovalenko et al., 2011) e 500.000 novos casos de leishmaniose visceral (Tiuman et al., 2011). A incidência da infecção é ainda superior quando casos subclínicos são incluídos. Como consequência, a leishmaniose é responsável por, aproximadamente, 70.000 mortes anuais (Desjeux, 2004).

Esta doença é diagnosticada em todos os estados brasileiros, manifestando-se sob diferentes perfis epidemiológicos. A leishmaniose tegumentar americana (LTA) está entre as endemias de maior importância em saúde pública no Brasil, devido a fatores como: ampla distribuição pelo território nacional, ocorrência de formas clínicas graves e limitações referentes tanto ao diagnóstico como ao tratamento, sendo L. (Viannia) braziliensis uma das principais espécies de importância epidemiológica para a LTA no Brasil (Amato et al., 2007).

Essa situação é agravada pela importância da leishmaniose canina na epidemiologia da doença, uma vez que os cães, que são os principais reservatórios destes parasitos, vivem em contato direto com humanos (Gramiccia e Gradoni, 2005, Martinez *et al.,* 2011).

Baseando-se no número expressivo de casos de leishmaniose no continente Americano, na morbidade causada pelas diferentes formas clínicas desta doença, na ineficiência dos tratamentos disponíveis, na ausência de métodos adequados de controle e profilaxia (Da Silva et al., 2010; Rapaport et al., 2013), torna-se de extrema importância o desenvolvimento de novas pesquisas que busquem novos alvos que possam ser utilizados para o desenvolvimento de novas abordagens de diagnóstico, tratamento (desenho racional de drogas) e prevenção (desenvolvimentos de vacinas para uso animal e humano).

Assim, é de fundamental importância a descoberta do universo de transcritos (proteínas e RNAs não codificadores) envolvidos nos processos de infectividade e virulência, os quais poderão ser utilizados como alvos em diversas aplicações biotecnológicas futuras. O

conhecimento aprofundado destes alvos é uma ferramenta interessante que poderá ser explorada com o objetivo de se desenvolver estratégias para intervir no processo infectivo e auxiliar no controle e erradicação desta enfermidade.

REVISÃO BIBLIOGRÁFICA

Leishmania e leishmaniose

Leishmania são parasitos digenéticos unicelulares intracelulares classificados de acordo com a Tabela 1, existindo ao menos 20 espécies deste gênero capazes de infectar mamíferos (Da Silva et al., 2010).

Tabela 1. Classificação taxonômica do parasito Leishmania**.**

Reino: Protista	Haeckel, 1866
Sub-Reino: Protozoa	Goldfuss, 1817
Filo: Sarcomastigophora	Honigberb e Balamuth, 1963
Sub-filo: Mastigophora	Deising, 1866
Classe: Zoomastigophorea	Calkins, 1909
Ordem: Kinetoplastida	Honigberg, 1963
Sub-ordem: Trypanosomatina	Kent, 1880
Família: Trypanosomatidae	Dofein, 1901

A comparação dos genomas das principais espécies de Leishmania tem revelado acentuada conservação no conteúdo genético e elevado grau de sintenia gênica. Estudos anteriores aos projetos de sequenciamento já haviam revelado que a maioria dos genes nestas espécies estão organizados em grandes clusters direcionais constitutivamente co-transcritos pela RNA Polimerase II (Johner et al., 2006). A transcrição policistrônica ocorre bidireccionalmente e tem início nas "regiões de troca de fitas" (SSR) (regiões conhecidas como switch) que são locais do cromossomo onde ocorre mudança na fita de DNA a ser transcrita (Martinez-Calvillo, Yan et al., 2003; Martinez-Calvillo, Nguyen et al., 2004).

As espécies de Leishmania do Velho Mundo possuem 36 pares de cromossomos, com tamanho variando de 0,28 a 2,8 Mb; enquanto que as espécies do Novo Mundo têm 34 ou 35 pares, sendo fusionados os cromossomos de número 8 com o 29 e o 20 com o 36 no complexo L. (L.) mexicana e os de número 20 com o 34 na espécie L. (V.) braziliensis (Britto, Ravel et al., 1998). Os projetos de sequenciamento tem revelado que o tamanho do genoma das diferentes espécies de Leihsmania é de aproximadamente 32,5 Mb, com exceção de L. (L.) donovani que tem em torno de 59,04 Mb.

Em 1972, introduziu-se o conceito de dois complexos ou grupos para as espécies de Leishmania do Novo Mundo: braziliensis e mexicana, baseando-se em estudos sobre o

comportamento de Leishmania em hospedeiros, meios de cultura e no tubo digestivo de flebotomíneos. Em 1987, a classificação foi modificada pela inclusão dos subgêneros Leishmania e Viannia (Lainson e Shaw, 1972; Shaw, 1994). Peacock et al. (2007), listam em seu trabalho as principais espécies de Leishmania que infectam humanos, assim como o tipo de manifestação clínica que causam.

O ciclo de vida das espécies deste gênero é constituído por dois estágios morfologicamente distintos: promastigotas extracelulares que residem no interior do intestino médio do inseto vetor, sendo essas as formas presentes no início da infecção e amastigotas intracelulares que residem no interior do fagolisossoma de macrófagos de hospedeiros mamíferos (Figura 1).

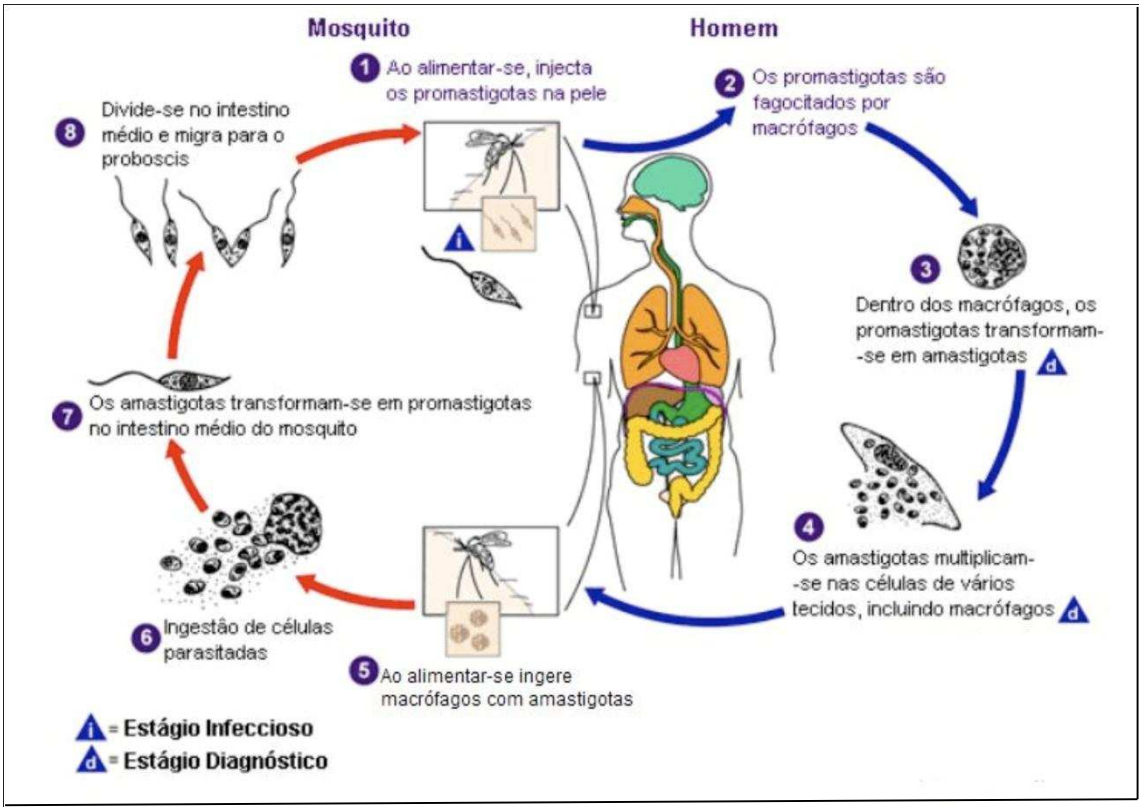

Figura 1. Esquema do ciclo de vida do parasito.
Fonte: adaptado de Leishmaniasis, 2011.

As leishmanioses podem manifestar-se sob diferentes formas clínicas, dependendo de fatores, tais como: espécie de Leishmania envolvida na infecção, vetor responsável pela transmissão e sistema imune do hospedeiro. A doença é categorizada em: (i) leishmaniose visceral (LV), forma mais grave, na qual os parasitos saem do sítio de inoculação e proliferam

no fígado, baço, linfonodos e medula óssea, podendo resultar em imunossupressão do hospedeiro e morte na ausência de tratamento (Cota, De Sousa e Rabello, 2011); (ii) leishmaniose tegumentar americana (LTA), que pode ser dividida em: (a) leishmaniose cutânea (LC), na qual os parasitos permanecem no local da infecção e causam ulceração de longa duração; (b) leishmaniose mucocutânea (LMC), com destruição crônica de mucosas e (c) leishmaniose cutânea difusa (LCD), na qual os parasitos causam lesões disseminadas não ulceradas (Tiuman et al., 2011). Ainda segundo este autor, mais de 90% dos casos de leishmaniose cutânea ocorrem no Brasil, Afeganistão, Arábia Saudita, Irã, Peru e Síria; e mais de 90% dos casos de leishmaniose visceral ocorrem no Brasil, Bangladesh, Índia e Sudão, sendo L. (Viannia) braziliensis uma das principais espécies de importância epidemiológica para a LTA no Brasil (Ashford, 2000). Além do Brasil, é importante ainda em toda a América Latina, ocorrendo desde a América Central até o norte da Argentina.

O quadro em que se encontra a leishmaniose é agravado pelo fato do cão ser considerado o principal reservatório deste parasito, sendo o cão acometido também pela doença que mostra-se como grave problema de saúde animal (Santos et al., 2010). Atualmente é considerada uma das seis principais doenças endêmicas de prioridade no mundo, sendo os cães a principal fonte da infecção para humanos. Estudos mostram redução da incidência da LV humana e canina após sacrifício dos cães soropositivos (Ashford et al., 1998; Nunes et al., 2010).

Expressão Gênica em Leishmania

Estes parasitos encontram extremas mudanças ambientais durante seu ciclo de vida e como consequência, respondem diferenciando-se em formas altamente adaptadas que os possibilitam invadir e proliferar-se dentro de seus hospedeiros.

Estudos realizados com espécies da família Trypanosomatidae têm revelado características bastante peculiares em relação ao processo de expressão gênica nesses microrganismos (Clayton, 2002; Teixeira e Darocha, 2003). Apesar de serem eucariotos, a transcrição dos genes em tripanossomatídeos, como em procariotos, é policistrônica, gerando pré-mRNAs que deverão ser processados no núcleo para formar mRNAs monocistrônicos maduros. Este fato faz com que a regulação da expressão gênica em Leishmania seja incomum (Myler et al., 2000; Saxena et al., 2007). Os mRNAs maduros são obtidos após poliadenilação-3´ coordenada e trans-splicing, nos quais 39 nucleotídeos (splicedleader) são

adicionados à extremidade 5'. A sequência do splicedleader (SL), encontrada somente em tripanossomatídeos e em nematódeos, é idêntica em todos os mRNAs de um mesmo organismo, mas é diferente entre as espécies (Donelson e Zeng, 1990).

Como consequência desta transcrição gênica peculiar, a expressão dos genes de Leishmania parece não ser regulada a nível transcricional. Reforçando este argumento, destaca-se o controle da expressão gênica estágio-específica em decorrência de diferenças na estabilidade dos mRNAs, que aparentemente é o principal mecanismo envolvido na regulação da expressão de vários genes de Leishmania, tais como: p63 (Brittingham et al., 2001), A2 (Charest, Zhang e Matlashewski, 1996) e hsp83 (Aly et al., 1994). Regiões não traduzidas dos mRNAs, principalmente as regiões 3' UTR parecem ser também responsáveis pela especificidade de expressão nos diferentes estágios do ciclo de vida destes parasitos (Brittingham et al. 2001). Outros mecanismos estão envolvidos neste processo, tais como: controle da degradação do mRNA, controle da tradução do mRNA e modificação e degradação de proteínas (Clayton & Shapira, 2007).

Vários trabalhos tem mostrado que a maioria dos genes de Leishmania são constitutivamente expresso e que a taxa daqueles diferencialmente expressos oscila entre 3% e 9% (Saxena et al., 2007; Guerfali et al., 2008; Rochette et al., 2008).

Holzer e colaboradores (2006), utilizaram microarranjos para analisar o perfil de expressão de 8.156 ORFs entre promastigotas e amastigotas derivados de lesão e entre estas mesmas amastigotas derivadas de lesão com amastigotas axênicos de L. (L.) mexicana. Neste estudo, apenas 3,5% dos genes apresentaram expressão diferencial entre promastigotas e amastigotas derivados de lesão. Já a comparação entre os amastigotas axênicos e os amastigotas derivados de lesão revelou que 175 genes são diferencialmente expressos (2,1%). Apenas 17 genes (0,2 do total) com expressão diferencial foram encontrados ao se comparar os amastigotas axênicos e os promastigotas. Estes dados sugerem que o reduzido número de genes diferencialmente expressos (gde) encontrados pode ser uma consequência do aumento na magnitude dos níveis de transcritos nas células sob condições axênicas (Holzer, Mcmaster et al., 2006).

De um total aproximado de mais de 8.200 genes encontrados nas diferentes espécies de Leishmania, aqueles diferentemente expressos entre espécies, apesar do baixo percentual, perfazem centenas de genes que podem ter papel importante nos processos celulares. As situações estudadas se referem à análise da expressão diferencial de formas morfológicas distintas do parasito (promastigota e amastigota) através do uso de microarranjos de DNA

(Cohen-Freue et al., 2007; Peacock et al., 2007). Apesar destes estudos mostrarem poucos genes diferencialmente expressos, quando comparado com estudos envolvendo organismos mais complexos, estes perfazem mais de 200 genes, aguçando a curiosidade para uma visão mais crítica sobre estes, que provavelmente devem estar envolvidos com mudanças morfológicas, bioquímicas e biológicas observadas em ambos estágios do ciclo de vida do parasito. Colaborando com essa visão, Saxena e colaboradores (2007) mostraram que para L. donovani a diferenciação de promastigota para amastigota envolve a mudança de expressão transiente e permanente de genes em escala temporal. Estes resultados mostraram que a transição de estágios frente à pressão ambiental pode ter importante influência na expressão gênica de espécies de Leishmania (Depledge et al., 2009). Estudos envolvendo a tradução de proteínas em estágios específicos do desenvolvimento de algumas espécies de Leishmania, como: L. (V.) panamensis (Walker et al., 2006), L. (L.) donovani (Bente et al., 2003), L. (L.) mexicana e L. (L.) infantum (Acestor et al., 2002), tem demonstrado que a expressão dos RNAm obtida por análise de microarranjos não tem correlação precisa com a tradução de proteínas (Mcnicoll et al., 2006; Leifso et al., 2007).

Análises de expressão gênica envolvendo Leishmania spp. tendem a se tornarem rotineiras com o desenvolvimentos das tecnologias de sequenciamento paralelo maciço, devido a facilidade em se obter os dados e pelo término dos projetos de sequenciamento dos genomas de L. (L.) major por Ivens et al. (2005), de L. (L.) infantum e L. (V.) braziliensis por Peacock et al. (2007), L. (L.) mexicana por Rogers et al. (2011) e L. (L.) amazonensis por Real et al. (2013).

Sequenciamento paralelo maciço

Até o começo da década de 70, era muito difícil obter a sequência de nucleotídeos de um fragmento de DNA, por menor que fosse. Este problema foi resolvido com o surgimento em 1977 de duas tecnologias: uma desenvolvida por Alan Maxam e Walter Gilbert (baseada em hidrólise química) e outra por Frederick Sanger e colaboradores (baseada em reações enzimáticas); sendo possível determinar a sequência de nucleotídeos de fragmentos maiores de DNA. Como consequência, surgiu o primeiro sequenciador automático, o qual utilizou a metodologia de Sanger (Sanger, Nicklen e Coulson, 1977) modificada por Edwards et al. (1990). Essa metodologia revolucionou as pesquisas científicas após se difundir rapidamente pelo mundo, sendo a base do início da era Genômica (Sanger, Nicklen e Coulson, 1977).

Após a publicação do draft do genoma humano (Venter et al., 2001), houve um avanço nas tecnologias de sequenciamento culminando no surgimento dos sequenciadores de segunda geração, os quais têm suprido as principais limitações do método convencional proposto por Sanger, tais como: tempo e capacidade de geração de dados; refletindo na diminuição do valor do sequenciamento e aumento significativo do output gerado (Mardis, 2008; Cullum, Alder e Hoodless, 2011).

Atualmente existem diversas tecnologias voltadas para o sequenciamento do DNA em larga escala, tendo sido a Roche a primeira empresa a desenvolver essa estratégia que se baseia na tecnologia de pirosequenciamento, utilizada neste trabalho (De Carvalho e Da Silva, 2010). A partir deste período, outros métodos foram desenvolvidos, como: o método Polony (Shendure et al., 2005) utilizado no sequenciador SOLiD (Applied Biosystems®) e o método de amplificação em ponte (BENNETT et al., 2005) utilizado no sequenciador Genome Analyser (Illumina). As principais diferenças entre os sequenciadores podem ser encontradas no trabalho de Metzker (2010).

Pirosequenciamento

O princípio desta tecnologia foi proposto por Hyman (1988); mas somente em 2005, com os aperfeiçoamentos propostos por Margulies et al. (2005), foi possível disponibilizar no mercado o primeiro sequenciador de segunda geração (sequenciador 454 GS20).

A eficiência e rapidez desta técnica foram comprovadas pelo resequenciamento do genoma da bactéria Mycoplasma genitalium (508.069 bases) com 96% de cobertura e 99.96% de precisão em um único processamento de quatro horas (Margulies et al., 2005).

Essa tecnologia dispensa clonagem in vivo e tem baixo custo comparado a outros métodos existentes, sendo baseado na detecção de fótons de luz produzidos, em quantidade proporcional ao número de nucleotídeos incorporados (Ronaghi, 2001; Mardis, 2008). Este método, pode ser dividido em três etapas: a) preparo da amostra, b) PCR em emulsão e c) sequenciamento (Figura 2). Em "a", o DNA é fragmentado aleatoriamente e ligado a adaptadores A e B em suas extremidades. Em "b", estes fragmentos são ligados à microesferas magnéticas por meio do pareamento do adaptador B com sequências curtas complementares presentes na superfície da microesfera, onde ocorre a amplificação deste fragmento. Desta forma, essas microesferas agem como reatores de amplificação individual produzindo milhares de cópias de um único molde. Em "c", essas microesferas são aplicadas a

uma placa contendo milhões de poços, de modo que cada orifício dessa placa receba uma única microesfera. Posteriormente, são adicionados os reagentes necessários para ocorrer a amplificação do DNA. Pirofosfatos inorgânicos (PPis) são liberados a cada incorporação de um nucleotídeo complementar a fita molde. Estes PPis são convertidos pela enzima ATP sulfurilase em ATP que proporciona a conversão mediada pela luciferase, de luciferina para oxiluciferina; gerando luz. Essa luz, característica do tipo de nucleotídeo incorporado, é registrada na forma de pico que em conjunto formam o pirograma. Por fim, as sequências complementares ao DNA molde são determinadas pela interpretação dos pirogramas gerados; sendo a intensidade do sinal de quimioluminescência proporcional ao total de moléculas de pirofosfato liberadas (Morozova e Marra, 2008) (Figura 2).

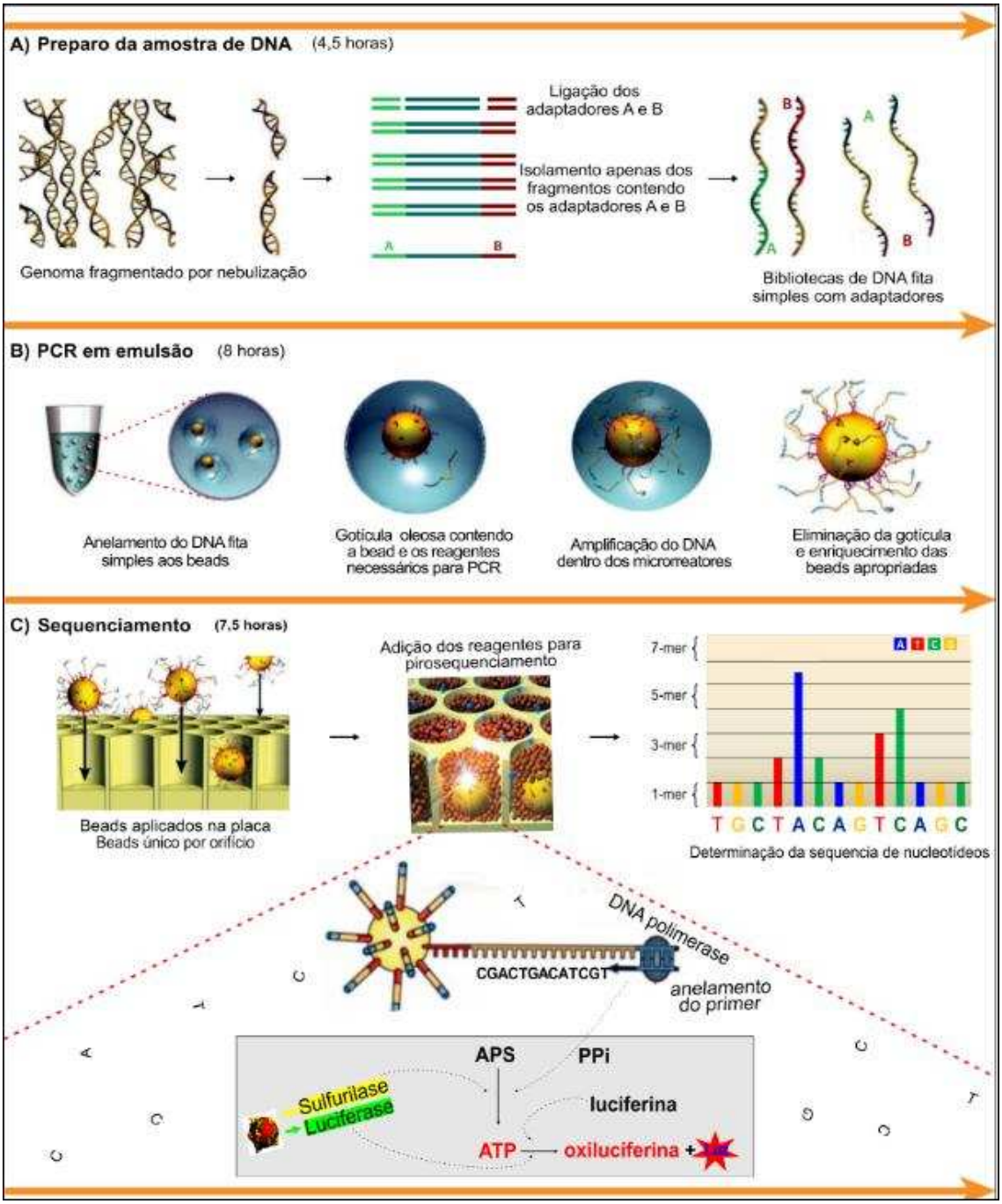

Figura 2. Etapas do sequenciamento na plataforma 454. O sequenciamento é dividido em três etapas: (a) preparo da amostra, (b) PCR em emulsão e (c) sequenciamento. (a) O DNA é fragmentado aleatoriamente e ligado aos adaptadores A e B em suas extremidades. (b) Os fragmentos são ligados às microesferas (beads) magnéticas por meio do pareamento do adaptador B com sequências curtas complementares presentes na superfície da microesfera. Um único tipo de fragmento se liga a uma determinada microesfera. As microesferas são capturadas individualmente em gotículas oleosas onde ocorre a PCR em emulsão. Milhares de cópias do fragmento alvo são produzidas nessa fase. (c) As microesferas ligadas às sequências alvo fita simples a serem sequenciadas são aplicadas na placa de sequenciamento de forma que apenas uma única bead é capturada por poço desta placa. Em seguida, são fornecidos os reagentes necessários para a reação de pirosequenciamento. Cada nucleotídeo incorporado, em cada um dos poços, liberará um pirofosfato que será convertido em luz e registrado na forma de pico. Assim, é gerado um pirograma para cada bead contida na placa; sendo estes posteriormente decodificados numa sequência de nucleotídeos única representativa do fragmento único ligado a bead. Fonte: adaptado de Margulies et al. (2005), p.377.

Uma limitação importante da plataforma 454 é a baixa eficiência na determinação de nucleotídeos que fazem parte de regiões homopoliméricas. Essa incerteza ocorre devido a não proporcionalidade entre a quantidade de nucleotídeos incorporados (quantidade de pirofosfato liberado) e a quantidade de luz emitida. Outra desvantagem é que o custo do sequenciamento com essa plataforma é superior comparado ao custo do sequenciamento utilizando as plataformas Solexa e SOLiD.

Montagem

O ideal seria sequenciar os transcritos no tamanho real em que ocorrem no interior das células. No entanto, a maioria das tecnologias de sequenciamento desenvolvidas até o momento utilizam a DNA polimerase para incorporação de nucleotídeos a cadeia de DNA nascente. Como consequência, o tamanho dos fragmentos gerados constitui numa limitação destas técnicas, uma vez que a processividade desta enzima é limitada. Para contornar este problema, é preciso fragmentar os transcritos (mRNAs e RNAs não codificantes) em tamanho menores para serem sequenciados. Após sequenciamento é preciso montar estes fragmentos. A montagem (agrupamento) das reads (sequências individuais geradas) consiste em eliminar a redundância do conjunto de fragmentos sequenciados pela sobreposição de regiões iguais a fim de produzir sequências únicas maiores denominadas contigs. Este procedimento facilita a anotação por homologia por aumentar o nível de confiabilidade dos resultados (Pertea et al., 2003). O ideal é que após a montagem, o número de contigs seja igual ou o mais próximo possível do número de transcritos da espécie.

Um passo anterior a montagem consiste na retirada de regiões que não fazem parte do genoma/transcriptoma do organismo sequenciado, tais como vetores, adaptadores e contaminações, pois estas sequências prejudicam o processo de montagem resultando em contigs não representativos da espécie e aumento do número de contigs gerados (Pertea et al., 2003).

Outro problema inerente aos sequenciadores de segunda geração é a incorporação errônea de nucleotídeos, pela DNA polimerase, a cadeia de DNA nascente. Como resultado, temos dentre os milhares de fragmentos amplificados alguns contendo determinados nucleotídeos incorporados de forma equivocada pela RNA polimerase. Para contornar este problema, torna-se necessário sequenciar a mesma posição do fragmento várias vezes (Bouck et al., 1998). É possível identificar, com a atribuição de valor de qualidade a cada base sequenciada, os prováveis nucleotídeos incorporados de forma equivocada.

Os pirogramas gerados pelo sequenciador são submetidos a um programa de base calling que efetua a leitura dos registros do sequenciador e identifica os nucleotídeos sequenciados. A atribuição de valor de qualidade aos nucleotídeo sequenciados permite, posteriormente eliminar e/ou mascarar do conjunto de dados, sequências (ou nucleotídeos) com baixo valor de qualidade (Ewing e Green, 1998; Ewing et al., 1998).

Atualmente, existem diferentes metodologias destinadas a montagem de transcriptomas e genomas sequenciados. Essas se baseiam, principalmente em duas categorias: de novo assembly (ab initio) e montagem por referência. A primeira categoria (montagem baseada no de novo assembly) contempla principalmente três abordagens: Overlap-layout-consensus, Greedy graph e grafo de de Bruijn ou caminho Euleriano (Miller et al., 2010).

Anotação funcional

O processo de anotação consiste em agregar o máximo de informação biológica à uma determinada sequência ou conjunto de sequências, que pode variar desde um transcrito até um genoma inteiro. Em outras palavras, podemos dizer que o processo de anotação consiste em buscar por padrões nas sequências biológicas e posteriormente, atribuir significado biológico a estes padrões. Isso é possível pela comparação das sequências a serem anotadas com inúmeros bancos de dados biológico. Esta abordagem permite identificar: códons de início e final da tradução, junção íntron-éxon, regiões codificadoras de proteínas, sítio de ligação ribossomal, regiões promotoras, regiões regulatórias; além da identificação de regiões conservadas, tais como: elementos repetitivos e os diferentes tipos de RNAs (Stein, 2001).

Atualmente existem diversas ferramentas destinadas a anotação funcional de sequências biológicas, sendo a ferramenta Blast2GO utilizada neste trabalho. Normalmente, estes métodos, utilizam termos de ontologias, como os presentes no banco de dados Gene Ontology (GO).

Blast2GO utiliza o resultado do algoritmo BLAST (Altschul et al., 1990). Posteriormente é efetuada diferentes etapas de mapeamento para vincular os hits do BLAST à informações funcionais armazenados no banco de dados Gene Ontology. Para isto, Blast2GO usa diferentes informações públicas disponibilizadas pelo NCBI, PIR e GO para conectar diferentes IDs de proteínas (nomes, símbolos, GIs, UniProts, etc.) à informações armazenadas no banco de dados Gene Ontology. Este banco de dados GO contém vários milhões de

produtos gênicos anotados funcionalmente para centenas de espécies. Todas as anotações são associadas a um código de evidência que fornece informações sobre a qualidade da atribuição desta função. Por fim tem-se o passo de anotação que consiste na seleção dos termos GOs a partir do conjunto de termos GOs obtidos pelo passo de mapeamento e suas associações com às sequências query (Conesa et al., 2005; https://www.blast2go.com/).

O resultado da anotação com Blast2GO depende dos parâmetros escolhidos, tais como banco de dados, valor de e-value, número de hits desejados, tipo de blast utilizado, cobertura do hit, busca de motivos conservados, dentre outros. Essa ferramenta possui funções gráficas e estatísticas que auxiliam na análise e interpretação dos resultados.

O objetivo deste trabalho foi, através de tecnologia de sequenciamento paralelo maciço e bioinformática, analisar o transcriptoma de L. (V.) braziliensis através do sequenciamento dos transcriptomas de dois isolados (ET e NSL), sendo um significativamente mais virulento e infectivo do que o outro. Foram preparadas duas formas evolutivas diferentes para cada isolado: uma preparação enriquecida em formas infectivas promastigotas metacíclicas (MET) e outra enriquecida de formas não infectivas promastigotas procíclicas (PRO). Este trabalho visou ainda analisar a expressão gênica diferencial entre as quatro bibliotecas de RNA-Seq sequenciadas, aqui definidas como: ET-MET, ET-PRO, NSL-MET e NSL-PRO.

REFERÊNCIAS BIBLIOGRÁFICAS

ACESTOR, N. et al. Establishing two-dimensional gels for the analysis of Leishmania proteomes. **Proteomics**, v. 2, n. 7, p. 877-879, 2002.

ALTSCHUL, S. F. et al. Basic local alignment search tool. **Journal Molecular Biology,** v. 215, n. 3, p. 403-10, 1990.

ALY, R. et al. A regulatory role for the 5' and 3' untranslated regions in differential expression of hsp83 in Leishmania. **Nucleic Acids Research,** v. 22, p. 2922-2929, 1994.

AMATO, V. S. et al. Treatment of mucosal leishmaniasis in Latin America: systematic review. **American Journal of tropical medicine and hygiene,** v. 77, n. 2, p. 266-74, 2007.

ASHFORD, D. A. et al. Studies on control of visceral leishmaniasis: impact of dog control on canine and human visceral leishmaniasis in Jacobina, Bahia, Brazil. **The American Journal of Tropical Medicine and Hygiene,** v. 59, n. 1, p. 53-7, 1998.

ASHFORD, R. W. The leishmaniases as emerging and reemerging zoonoses. **International Journal Parasitology,** v. 30, n. 12-13, p. 1269-81, 2000.

BENTE, M, et al. Developmentally induced changes of the proteome in the protozoan parasite Leishmania donovani. Proteomics. v. 3, n. 9, p. 1811-1829, 2003.

BOUCK, J. et al. Analysis of the quality and utility of random shotgun sequencing at low redundancies. **Genome Research,** v. 8, n. 10, p. 1074-84, 1998.

BRANDAO-FILHO, S. P. et al. Spatial and temporal patterns of occurrence of Lutzomyia sand fly species in an endemic area for cutaneous leishmaniasis in the Atlantic Forest region of northeast Brazil. **Journal Vector Ecology,** v. 36, p. S71-76, 2011.

BRITTINGHAM, A. et al. Regulation of GP63 mRNA stability in promastigotes of virulent and attenuated Leishmania chagasi. **Molecular and Biochemical Parasitology,** v. 112, p. 51-59, 2001.

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

CLAYTON, C. E. Life without transcriptional control? From fly to man and back again. **Embo Journal,** v. 21, n. 14, p. 3917-3917, 2002.

CLAYTON, C; SHAPIRA, M. Post-transcriptional regulation of gene expression in trypanosomes and leishmanias. **Molecular and Biochemical Parasitology**, v. 156, p. 93-101, 2007.

COHEN-FREUE, G. et al. Global gene expression in Leishmania. **International Journal Parasitology,** v. 37, p. 1077-1086, 2007.

CONESA, A. et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. **Bioinformatics,** v. 21, n. 18, p. 3674-6, 2005.

COTA, G. F.; DE SOUSA, M. R.; RABELLO, A. Predictors of Visceral Leishmaniasis Relapse in HIV-Infected Patients: A Systematic Review. **PLoS Neglected Tropical Diseases,** v. 5, n. 6, p. e1153, 2011.

CULLUM, R.; ALDER, O.; HOODLESS, P. A. The next generation: using new sequencing technologies to analyse gene regulation. **Respirology,** v. 16, n. 2, p. 210-22, 2011.

DA SILVA, M. S. et al. The Leishmania amazonensis TRF homologue binds and co-localizes with telomeres. **BMC Microbiology,** v. 10, p. 136, 2010.

DE CARVALHO, M. C. D. G.; DA SILVA, D. C. G. Next generation DNA sequencing and its applications in plant genomics. **Ciencia Rural,** v. 40, n. 3, p. 735-744, 2010.

DEPLEDGE, D. P. et al. Comparative Expression Profiling of Leishmania: Modulation in Gene Expression between Species and in Different Host Genetic Backgrounds. **PLoS Neglected Tropical Diseases,** v. 3, n. 7, p. e476, 2009.

DESJEUX, P. Leishmaniasis: current situation and new perspectives. **Computatiol Immunology Microbiology And Infectious Diseases,** v. 27, n. 5, p. 305-18, 2004.

DO MONTE-NETO, R. L. et al. Gene Expression Profiling and Molecular Characterization of Antimony Resistance in Leishmania amazonensis. **PLoS Neglected Tropical Diseases,** v. 5, n. 5, p. e1167, 2011.

DONELSON, J. E.; ZENG, W. A comparison of trans-RNA splicing in trypanosomes and nematodes. **Parasitology Today,** v. 6, n. 10, p. 327-34, 1990.

EDWARDS, A. et al. Automated DNA sequencing of the human HPRT locus. **Genomics,** v. 6, n. 4, p. 593-608, 1990.

EWING, B.; GREEN, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. **Genome Research,** v. 8, n. 3, p. 186-194, 1998.

EWING, B. et al. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. **Genome Research,** v. 8, n. 3, p. 175-185, 1998.

GRAMICCIA, M.; GRADONI, L. The current status of zoonotic leishmaniases and approaches to disease control. **International Jorunal Parasitology,** v. 35, n. 11-12, p. 1169-80, 2005.

GUERFALI, F. Z. et al. Simultaneous gene expression profiling in human macrophages infected with Leishmania major parasites using SAGE. **BMC Genomics,** v. 9, p. 238, 2008.

HYMAN, E. D. A new method of sequencing DNA. **Anal Biochemistry,** v. 174, n. 2, p. 423-36, 1988.

IVENS, A. C. et al. The genome of the kinetoplastid parasite, Leishmania major. **Science,** v. 309, n. 5733, p. 436-42, 2005.

JOHNER A.; KUNZ S.; LINDER M.; SAHAKUR Y.; SEEBECK T. Cyclic nucleotide specific phosphodiesterases of Leishmania major. **BMC Microbiology**, v**.**6, 2006.

KANEHISA, M.; GOTO, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. **Nucleic Acids Research,** v. 28, p. 27-30, 2000.

KOVALENKO, D. A. et al. Canine leishmaniosis and its relationship to human visceral leishmaniasis in Eastern Uzbekistan. **Parasitology Vectors,** v. 4, p. 58, 2011.

LEIFSO, K. et al. Genomic and proteomic expression analysis of Leishmania promastigote and amastigote life stages: the Leishmania genome is constitutively expressed. **Molecular Biochemistry Parasitololgy**. v.152, n. 1, p. 35-46, 2007.

LAINSON, R.; SHAW, J. J. LEISHMANIASIS OF THE NEW WORLD: TAXONOMIC PROBLEMS. **Br Med Bull,** v. 28, n. 1, p. 44-48, 1972.

LEISHMANIASIS. [Online]. 2014 July 01; Available from: URL:http://www.cdc.gov/dpdx/leishmaniasis/index.html.

MARDIS, E. R. Next-generation DNA sequencing methods. **Annual Review of Genomics and Human Genetics,** v. 9, p. 387-402, 2008.

MARGULIES, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. **Nature,** v. 437, n. 7057, p. 376-80, 2005.

MARTINEZ-CALVILLO, S., D. Nguyen, et al. Transcription initiation and termination on Leishmania major chromosome 3. **Eukaryot Cell**, v. 3, n. 2, p. 506-17, 2004.

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

MARTINEZ, V. et al. Canine leishmaniasis: the key points for qPCR result interpretation. Parasit Vectors, v. 4, p. 57, 2011.

MCNICOLL, F. et al. A combined proteomic and transcriptomic approach to the study of stage differentiation in Leishmania infantum. **Proteomics**. v. 6, n. 12, p. 3567-3581, 2006.

METZKER, M. L. Sequencing technologies - the next generation. **Nature Reviews Genetics,** v. 11, n. 1, p. 31-46, 2010.

MILLER, J. R.; KOREN, S.; SUTTON, G. Assembly algorithms for next-generation sequencing data. **Genomics**, v. 95, n. 6, p. 315-327, 2010.

MOROZOVA, O.; MARRA, M. A. Applications of next-generation sequencing technologies in functional genomics. **Genomics,** v. 92, n. 5, p. 255-64, 2008.

MYLER, P. J. et al. Genomic organization and gene function in Leishmania. **Biochem Soc Trans,** v. 28, n. 5, p. 527-31, 2000.

NUNES, C. M. et al. Relationship between dog culling and incidence of human visceral leishmaniasis in an endemic area. **Veterinary Parasitology,** v. 170, n. 1-2, p. 131-133, 2010.

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

PERTEA, G. et al. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. **Bioinformatics,** v. 19, n. 5, p. 651-652, 2003.

RAPAPORT, F. et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. **Genome Biology,** v. 14, n. 9, p. R95, 2013.

REAL, F. et al. The Genome Sequence of Leishmania (Leishmania) amazonensis: Functional Annotation and Extended Analysis of Gene Models. **DNA Research,** v. 20, n. 6, p. 567-81, 2013.

ROCHETTE, A. et al. Genome-wide gene expression profiling analysis of Leishmania major and Leishmania infantum developmental stages reveals substantial differences between the two species. **BMC Genomics,** v. 9, p. 255, 2008.

ROGERS, M. B. et al. Chromosome and gene copy number variation allow major structural change between species and strains of Leishmania. **Genome Research,** v. 21, n. 12, p. 2129-2142, 2011.

RONAGHI, M. Pyrosequencing sheds light on DNA sequencing. **Genome Research,** v. 11, n. 1, p. 3-11, 2001.

SANGER, F.; NICKLEN, S.; COULSON, A. R. DNA sequencing with chain-terminating inhibitors. **Proc Natl Acad Sci U S A,** v. 74, n. 12, p. 5463-5467, 1977.

SANTOS, J. M. et al. [Prevalence of anti-Leishmania spp antibodies in dogs from Garanhuns, in the middle scrub zone (Agreste) of Pernambuco]. **Rev Soc Bras Med Trop,** v. 43, n. 1, p. 41-45, 2010.

SAXENA, A. et al. Analysis of the Leishmania donovani transcriptome reveals an ordered progression of transient and permanent changes in gene expression during differentiation. **Molecular Biochemistry Parasitology,** v. 152, n. 1, p. 53-65, 2007.

SHAW, J. Taxonomy of the Genus Leishmania: Present and Future Trends and Their Implications. **Memorias Do Instituto Oswaldo Cruz,** v. 89, p. 471-478, 1994.

SHENDURE, J. et al. Accurate multiplex polony sequencing of an evolved bacterial genome. **Science,** v. 309, n. 5741, p. 1728-32, 2005.

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

STEIN, L. Genome annotation: from sequence to biology. **Nature Reviews Genetics,** v. 2, n. 7, p. 493-503, 2001.

TEIXEIRA, S. M.; DAROCHA, W. D. Control of gene expression and genetic manipulation in the Trypanosomatidae. **Genetics and Molecular Research,** v. 2, n. 1, p. 148-58, 2003.

TIUMAN, T. S. et al. Recent advances in leishmaniasis treatment. **Int J Infect Dis**, 2011.

VENTER, J. C. et al. The sequence of the human genome. **Science,** v. 291, n. 5507, p. 1304-51, 2001.

ZDOBNOV, E. M.; APWEILER, R. InterProScan--an integration platform for the signature-recognition methods in InterPro. **Bioinformatics,** v. 17, n. 9, p. 847-8, 2001.

WALKER, J. et al. Identification of developmentally-regulated proteins in Leishmania panamensis by proteome profiling of promastigotes and axenic amastigotes. **Mol Biochem Parasitol**. v. 147, n. 1, p. 64-73, 2006

TRANSCRIPTOMA DE Leishmania (V.) braziliensis POR RNA-Seq: MONTAGEM DE TRANSCRIPTOMAS, ENRIQUECIMENTO DE ORFEOMA, ANÁLISE DE EXPRESSÃO E ANOTAÇÃO DOS GENES DIFERENCIALMENTE EXPRESSOS

INTRODUÇÃO

L. (V.) braziliensis são protozoários da Ordem Kinetoplastida. Segundo Da Silva et al. (2010), o gênero Leishmania é composto por parasitos intracelulares obrigatórios e possui ao menos 20 espécies capazes de infectar mamíferos. Estes parasitos são um problema de saúde pública em muitos países causando um amplo espectro de desordens clínicas referidas comumente como leishmaniose que é uma doença de caráter zoonótico e sua transmissão ocorre de forma natural pela picada de fêmeas de flebotomíneos infectadas. Esta doença é endêmica, com altos índices de incidência em 88 países distribuídos em quatro continentes; existindo atualmente, aproximadamente 20 milhões de indivíduos infectados e 350 milhões de pessoas vivendo em área de risco (Do Monte-Neto et al., 2011). A cada ano são relatados mais de 1,5 milhões de novos casos de leishmaniose cutânea (Kovalenko et al., 2011) e 500.000 novos casos de leishmaniose visceral (Tiuman et al., 2011). Como consequência, a leishmaniose é responsável por cerca de 51.000 mortes anuais (Desjeux, 2004).

Baseando-se no expressivo número de casos de leishmaniose no continente Americano, na morbidade causada pelas diferentes formas clínicas desta doença, na ineficiência dos tratamentos disponíveis, na ausência de métodos adequados de controle e profilaxia (Da Silva et al., 2010; Rapaport et al., 2013); aliado ao desenvolvimento das novas tecnologias de sequenciamento; é de fundamental importância o sequenciamento do transcriptoma de L. (V.) braziliensis, pois este pode contribuir com inúmeras aplicações: enriquecimento de mapas genômicos (Wilcox et al., 1991), descoberta de genes (Mondego et al., 2011), caracterização de candidatos a polimorfismos de nucleotídeo único (SNPs) (Useche et al., 2001), identificação de marcadores moleculares específicos (Romanuik et al., 2009), identificação de erros de sequenciamento e montagem de genomas, complementação e/ou correções de anotações biológicas, dentre outras.

O princípio da tecnologia de sequenciamento utilizada neste trabalho foi proposto por Hyman em 1988, mas somente em 2005, com os aperfeiçoamentos propostos por Margulies et al. (2005), foi possível disponibilizar no mercado o primeiro sequenciador de segunda

geração que tem como princípio o pirosequenciamento. Esta tecnologia dispensa clonagem e tem baixo custo comparado a outros métodos existentes, sendo baseado na detecção de fótons de luz produzidos, em quantidade proporcional ao número de nucleotídeos incorporados (Ronaghi, 2001; Mardis, 2008).

As novas informações advindas deste trabalho serão importantes no desenvolvimento de novas pesquisas envolvendo diagnóstico, tratamento (desenho racional de drogas) e prevenção (desenvolvimento de vacinas alvo específicas); além de eliminar a propagação errônea de dados, uma vez que a abordagem de anotações de sequências biológicas são baseadas em identidade de sequências e por isso um único erro pode ser propagado para inúmeras outras sequências biológicas. Os novos conhecimentos poderão ainda serem explorados com o objetivo de se desenvolver estratégias para intervir no processo infectivo e auxiliar no controle e erradicação desta enfermidade.

O objetivo principal deste trabalho foi: sequenciar e montar, pela primeira vez por RNA-Seq, o transcriptoma de L. (V.) braziliensis através de dois isolados desta espécie que apresentam diferença significativa na virulência em modelo murino (ET e NSL), sendo duas formas evolutivas para cada isolado (metacíclica e procíclica); e comparar o conjunto gênico obtido com o orfeoma predito automaticamente a fim de identificar e corrigir erros de sequenciamento, montagem e anotação; além de identificar novos genes. Estes resultados permitirão ampliar os estudos acerca de genes, transcritos e proteínas em L. (V.) braziliensis. Por fim, este trabalho analisou a expressão gênica diferencial entre as quatro bibliotecas de RNA-Seq sequenciadas, aqui definidas como: ET-MET, ET-PRO, NSL-MET e NSL-PRO.

MATERIAL E MÉTODOS

Os procedimentos experimentais de isolamento dos parasitos até a obtenção das bibliotecas transcriptômicas de RNA-Seq estão descritos no trabalho de Santos (2012).

Estação de trabalho (computadores)

As análises de bioinformática foram realizadas em dois servidores localizados na Universidade Federal de Viçosa. Um no Instituto de Biotecnologia Aplicado à Agropecuária (BIOAGRO) e o outro localizado na Diretoria de Tecnologia da Informação - Central de Processamento de Dados (DTI-CPD). Ambos com sistema operacional Linux.

O primeiro possui dois processadores Intel® Xeon® E5410, 2.33 GHz, 32 GB de memória RAM e 4 terabyte (TB) de armazenamento. O segundo, situado no DTI-CPD, é um cluster composto por 24 processadores intel® Xeon® X5650, 2.67 GHz, sendo 1 nó com 48 GB de RAN, 15 nós com 24 GB de RAM, um nó teste com 16 GB de RAM e uma GPU com 24 GB de RAM. Cada máquina possui 3TB de armazenamento. No entanto, o cluster possuiu um storage (um único computador para armazenar dados) com 100 TB de armazenamento.

Delineamento dos tratamentos

Neste projeto foram sequenciados os transcriptomas de dois isolados de L. (V.) braziliensis (ET e NSL). Foram obtidas duas formas evolutivas diferentes para cada isolado: uma preparação enriquecida em formas infectivas promastigota metacíclica (MET) e outra enriquecida com formas não infectivas promastigota procíclica (PRO).

Desta forma foram construídas quatro bibliotecas transcriptômicas, aqui definidas como: ET-MET, ET-PRO, NSL-MET e NSL-PRO que, a partir de agora, poderão ser referidas como bibliotecas 1, 2, 3 e 4 respectivamente (Tabela 2). Não foram obtidas repetições biológicas e técnicas para os transcriptomas sequenciados.

Tabela 2. Definição dos tratamentos.

Isolado	Forma evolutiva	Biblioteca	Codificação
ET	MET	ET-MET	1
	PRO	ET-PRO	2
NSL	MET	NSL-MET	3
	PRO	NSL-PRO	4

ET= isolado de L. (V.) braziliensis de menor virulência; NSL= isolado de L. (V.) braziliensis de maior virulência; MET=formas promastigotas metacíclicas (infecciosas para mamíferos) e PRO=formas promastigotas procíclicas (não infecciosas para mamíferos).

Sequenciamento das bibliotecas transcriptômicas

As quatro bibliotecas transcriptômicas foram produzidas no Centro Avançado de Tecnologia em Genômica (CATG) do Instituto de Química da Universidade de São Paulo (IQ/USP), coordenado pelo professor Dr. Sérgio Verjovski-Almeida. Todos os procedimentos experimentais foram realizados de acordo com as instruções do fabricante, exceto por algumas pequenas modificações conforme descrito por Santos (2012).

Estas quatro bibliotecas foram sequenciadas utilizando o sequenciador de nova geração 454 GS FLX localizado na 454-Roche Life Science (Branford CT, USA), utilizando-se uma placa de sequenciamento para os tratamentos 1, outra placa para o tratamento 3 e uma placa para os tratamentos 2 e 4 (1/2 da placa para cada biblioteca). A utilização de uma placa inteira para formas infectivas (tratamentos 1 e 3) é explicada por estas serem responsáveis pelo início da infecção, sendo desta forma os alvos mais promissores e de maior interesse deste trabalho. A menor cobertura obtida para as bibliotecas 2 e 4 não afetou a análise de expressão gênica uma vez que foi efetuada uma normalização dos dados para contornar essa situação.

Output e conversão de formatos

Os sinais quimioluminescentes produzidos durante o pirosequenciamento, como resultado da incorporação de nucleotídeos à cadeia de DNA nascente, foram registrados na forma de picos. Em conjunto, esses picos formaram os pirogramas (flowgrams), tendo cada fragmento sequenciado seu próprio pirograma. Posteriormente, estes pirogramas foram interpretados por programa próprio do sequenciador para geração dos arquivos binários no formato Standard Flowgram File (**SFF),** que é a extensão padrão dos outputs dos

sequenciadores da Roche. Este arquivo binário foi convertido em dois outros arquivos: um com extensão fasta (formato baseado em texto que contêm a sequência de nucleotídeo do fragmento) e outro que contêm o valor de qualidade de cada nucleotídeo presente no arquivo de extensão fasta.

O número e tamanho dos arquivos, assim como a quantidade das reads sequenciadas em cada biblioteca, estão representados na Tabela 3.

Tabela 3. Número e tamanho dos arquivos gerados pelo sequenciador e quantidade de reads sequenciadas em cada biblioteca.

Isolado	Formas evolutiva	nº de arquivo	Tamanho do arquivo (Gb)	nº de reads
ET	MET	2	2,9	916.546
	PRO	1	1.9	589.554
NSL	MET	2	3,40	1083.312
	PRO	1	1.6	506.312
Total	**-----**	**6**	**9.8**	**3.095.724**

ET= isolado de menor virulência; NSL= isolado de maior virulência; MET=formas promastigotas metacíclicas (infecciosas para mamíferos) e PRO=formas promastigotas procíclicas (não infecciosas para mamíferos).

A ferramenta **sfffile** (Roche) disponibilizada pela Roche Life Science foi utilizada para junção dos dois arquivos da amostra 1. O mesmo foi efetuado para os dois arquivos da amostra 3. Os quatro arquivos resultantes, um para cada biblioteca, foram convertidos, com o software **sff2fastq** (https://github.com/indraniel/sff2fastq) para o formato ***.fastq** (formato baseado em texto que armazena, em um único arquivo, sequências biológicas e a qualidades Phred associada a estas sequências).

Com a finalidade de enriquecer e de aumentar a confiabilidade de algumas análises, baixou-se do banco de dados Sequence Read Archive (SRA - Arquivo de leituras de sequências) (Kodama et al., 2012), indexado ao National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov), o conjunto de dados armazenado sob o código ERR013296.sra, especificado abaixo (Tabela 4).

Tabela 4. Origem e características dos dados obtidos do banco de dados SRA.

Plataforma de sequenciamento	Illumina
Sequenciador	Genome Analyzer II
Id SRA	ERR013296
Molécula sequenciada	DNA
Isolado	MHOM/BR/75/M2904
Biblioteca	Paired end
Primers	randômicos
Tamanho do read	76 pb
Comprimento nominal	250 pb
Data de publicação	04/08/2010
Tipo do acesso	Público
Spots	26 milhões
Total de reads	52.000.000,00
Bases	4.0 Gbp
Tamanho do arquivo	4.1 G
Conteúdo GC	51.9%

Foi utilizado o plugin **Aspera Connect** para acelerar o download deste conjunto de dados. O SRA é um banco de dados destinado ao armazenamento de reads brutos gerados pelos sequenciadores de alto desempenho e/ou alinhamento destas reads (Coordinators, 2014).

O arquivo ERR013296.sra foi convertido nos arquivos ERR013296.sra_1.fastq e ERR013296.sra_2.fastq através do utilitário **fastq-dump** presente no pacote **SRA toolkit** (v.2.4.4) (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=software&m=software&s=software). A geração de dois arquivos se deve a estas reads pertencerem a uma biblioteca paired end, que produz duas reads para cada fragmento sequenciado (um para cada extremidade). Esta conversão foi necessária, pois as ferramentas utilizadas neste trabalho não são capazes de trabalhar com a extensão **sra.**

Foi utilizada a seguinte linha de comando para junção, com **sfffile**, dos dois arquivos da biblioteca 1:

sfffile -o 1_mais_2.sff 1.sff 2.sff

onde **1_mais_2.sff** representa o arquivo resultante da junção e **1.sff** e **2.sff** os arquivos brutos produzidos pela plataforma de sequenciamento. O mesmo foi efetuado para os dois arquivos da biblioteca 3.

Para conversão do formato **.sff** para o **.fastq** foi utilizada a seguinte linha de comando:

sff2fastq 1_mais_2.sff > -o 1_mais_2.fastq (exemplificado apenas para a biblioteca 1)

onde **1_mais_2.sff** representa o arquivo bruto e **1_mais_2.fastq** o arquivo após a conversão.

Este mesmo procedimento foi efetuado para as bibliotecas 2, 3 e 4. Assim, todos os arquivos foram convertidos para o formato ***.fastq**. Esta conversão foi necessária uma vez que alguns softwares trabalham exclusivamente com arquivos no formato ***.fastq** (este formato também pode ser representado por ***.fq**).

O aplicativo **sffinfo** fornecido pela Roche tem como função a extração de informações das reads a partir de arquivos ***.sff**. O algoritmo **sffinfo** foi empregado para conversão dos arquivos ***.sff** nos formatos ***.fna** (formato fasta) e ***.qual**. Para isto foi utilizada a seguinte linha de comando:

sffinfo -s 1_mais_2.sff > 1_mais_2.fna (exemplificado apenas para a biblioteca 1)
sffinfo -q 1_mais_2.sff > 1_mais_2.qual (exemplificado apenas para a biblioteca 1)

onde **1_mais_2.sff** representa o arquivo bruto, **1_mais_2.fna** corresponde ao arquivo contendo a sequência de nucleotídeos das reads e **1_mais_2.qual** corresponde ao arquivo contendo as informações de qualidade dos nucleotídeos do arquivo **1_mais_2.fna**.

O mesmo procedimento foi efetuado para as amostras 2, 3 e 4, sendo este passo necessário, uma vez que os arquivos resultantes da conversão foram utilizados como entrada por diversos dos softwares ou plataformas utilizadas neste trabalho.

Eventualmente, foi utilizado ainda o programa **sff_extract** para a conversão dos arquivos ***.sff** para ***.fasta** e ***.qual**, através da seguinte linha de comando:

sff_extract -o output_name 1_mais_2.sff

onde **1_mais_2.sff** representa o arquivo bruto e **output_name** o nome para os arquivos de saídas. Neste caso, foram produzidos três arquivos: extensão ***.fasta**, ***.qual** e ***.xml**. Este script pode ser utilizado ainda com outros parâmetros permitindo, por exemplo, trimar extremidades das reads e eliminar contaminações.

Análise da qualidade das reads sequenciadas

Várias características intrínsecas às reads sequenciadas são capazes de interferirem na montagem. Assim, as reads utilizadas neste trabalho foram analisadas com o software **fastQC** (http://www.bioinformatics.babraham.ac.uk/projects/download.html#fastqc) com a finalidade de projetar o tipo de tratamento necessário para deixar os dados o mais adequado possível para montagem. Este procedimento foi efetuado para cada uma das quatro bibliotecas obtidas neste trabalho e para o conjunto de dados obtido do banco de dados SRA.

Tratamento inicial das reads

Uma read típica é uma sequência relativamente curta (menor que 1000 pb), representada por uma sequência de nucleotídeos propensa a erros, principalmente em suas extremidades devida a limitada processividade da enzima DNA polimerase. Além da qualidade, inúmeros outros problemas no conjunto de dados precisam ser identificados e eliminados através de trimagens das reads. Sequências filtradas são aquelas eliminadas por completo do conjunto de dados e sequências trimadas são aquelas que tiveram, por algum motivo, uma parte removida; processos estes intitulado 'tratamento inicial dos dados'.

O tratamento das reads foi efetuado individualmente para cada biblioteca com a utilização dos programas **FASTX-Toolkit** (v.0.0.13; http://hannonlab.cshl.edu/fastx_toolkit/).

Dentre os tipos de tratamentos adotados, efetuou-se

- ✓ Trimagens de reads duplicadas.
- ✓ Trimagem de adaptadores.
- ✓ Exclusão de sequências super representadas.
- ✓ Trimagem de extremidades de reads com valor de Phred menor que 20.
- ✓ Trimagem de reads contendo mais que uma base ambígua, representada no conjunto de dados pela letra “N”.

- ✓ Trimagem de reads com valor de Phred médio menor que 20.
- ✓ Trimagem das reads menores que 50 pares de bases.

O valor Phred (Q) é atribuído a todos os nucleotídeos sequenciados e corresponde à probabilidade logarítmica negativa de uma base qualquer de uma read ter sido sequenciada erroneamente (Q = -10 log10 (Pe)), onde Q corresponde a qualidade e Pe à probabilidade de erro. Como exemplo, um valor de Phred de 30 significa 0,001% de chance de erro, ou seja 1 erro a cada 1000. Desse modo, quanto maior valor de Phred, melhor a qualidade do sequenciamento.

O tratamento das reads obtidas do banco de dados SRA foi efetuado com o software **Prinseq-Lite** (v. 0.20.4; Schmieder & Edwards 2011) devido a capacidade deste trabalhar com dados paired end. Para isto foi utilizada a seguinte linha de comando:

prinseq-lite.pl -verbose -fastq 1.fastq -fastq2 2.fastq -trim_qual_left 20 - trim_qual_right 20 -trim_ns_left 1 -trim_ns_right 1 -min_len 50 -min_qual_mean 20 -ns_max_n 1 -derep 1 -noniupac -out_bad null

onde:

prinseq-lite.pl corresponde ao script responsável por executar o programa,

-verbose, parâmetro responsável por mostrar na tela o andamento da análise,

-fastq parâmetro que especifica o 1º arquivo do par,

-fastq2 parâmetro que especifica o 2º arquivo do par,

1.fastq e **2.fastq** são os arquivo a serem analisados,

-trim_qual_left parâmetro que especifica a trimagem dos nucleotídeos com valor de Phred abaixo de 20 na extremidade 5' da read,

-trim_qual_right 20 parâmetro que especifica a eliminação dos nucleotídeos com valor de Phred abaixo de 20 na extremidade 3' da read,

-trim_ns_left 1 parâmetro que especifica a trimagem de base ambígua (N) na extremidade 5',

-trim_ns_right 1 parâmetro que especifica a trimagem de base ambígua (N) na extremidade 3',

-min_len 50 parâmetro que especifica a trimagem de reads menores que 50 pb,

-min_qual_mean 20 parâmetro que especifica a trimagem de reads que tenham valor de Phred médio abaixo de 20,

-ns_max_n 1 parâmetro que especifica a trimagem de reads contendo mais de um nucleotídeo ambíguo,

-derep 1 parâmetro que especifica a trimagem de reads idênticas,

-noniupac parâmetro que especifica a trimagem de reads contendo nucleotídeos diferentes de: A, C, T, G e N.

-out_bad null parâmetro que especifica a não geração de arquivos contendo sequências que não atendam aos critérios estabelecidos.

Após tratamento, estes dados foram novamente analisadas no programa **fastQC** visando verificar se as reads estavam aptas a serem utilizadas neste trabalho.

Referência utilizada para mapeamento

Os genomas de L. (V.) braziliensis e L. (L.) major utilizados como referência para o mapeamento das reads foram baixados no formato GenBank (***.gbk**) (formato utilizado para armazenar informações genômicas), do banco de dados de sequência de ácidos nucléicos GenBank (http://www.ncbi.nlm.nih.gov/genbank/). Este formato contém informações biológicas e bibliográficas e é considerado um dos mais padronizados formatos de armazenamento e descrição de dados biológicos. Posteriormente estes cromossomos foram unidos através de um script em perl visando otimizar o tempo das análises, uma vez que diversas análises foram realizadas uma única vez.

O genoma de L. (L.) major foi utilizado em algumas análises por estar mais bem fechado (36 cromossomos e 36 contigs) dentre as espécies de Leishmania com genoma sequenciado (Ivens et al., 2005). O genoma de L. (V.) braziliensis encontra-se bastante fragmentado (35 cromossomos e 955 contigs). A presença de 35 cromossomos em L. (V.) braziliensis se deve a fusão entre os cromossomos 20 e 34 (Britto et al., 1998).

Para expressão diferencial, devido a inúmeros erros de anotação encontrados nas ORFs anotadas automaticamente, utilizou-se como referência para mapeamento, o conjunto de consensos obtido após o mapeamento das reads de todas as bibliotecas no orfeoma anotado automaticamente de L. (V.) braziliensis.

Panorama Geral

Foi efetuado um mapeamento do conjunto de reads (obtidas da soma das reads das quatro bibliotecas), nas ORFs anotadas automaticamente de L. (V.) braziliensis com a finalidade de comparar a identidade entre as reads tratadas com o orfeoma de L. (V.) braziliensis. Esta abordagem permitiu ainda verificar a porcentagem de reads mapeadas, fornecendo uma ideia da representatividade do transcriptoma desta espécie proporcionada por nossos dados.

Por fim este conjunto de reads foi mapeado no genoma anotado desta espécie a fim de verificar o mapeamento de reads em regiões não correspondentes as ORFs, sendo este um indício de que nossos dados trará novas informações ao orfeoma desta espécie. As reads não mapeadas neste passo podem corresponder a regiões novas não representadas no genoma desta espécie, uma vez que este genoma encontra-se muito fragmentado.

Mapeamento

Os mapeamentos foram efetuados através do programa **CLC bio Genomics Workbench** (http://www.clcbio.com/products/clc-genomics-workbench/) e **bowtie 2** (Langmead & Salzberg, 2012).

Montagem das reads

Uma característica comum à maioria das tecnologias de sequenciamento atuais e que prejudicam a montagem é o limitado tamanho dos fragmentos de DNA sequenciados. Para contornar este problema, vários programas de montagem foram desenvolvidos com o intuito de comparar as reads ou parte destas em busca de regiões idênticas o suficiente para serem reunidas em uma sequência única denominada contig (Nagaraj, Gasser e Ranganathan, 2007).

Os quatro arquivos ***.sff**, referentes as quatro bibliotecas, foram utilizados de duas maneiras distintas para realização de um de novo assembly (montagem). Primeiro efetuou-se a montagem com as reads de cada biblioteca individualmente (montagens 1, 2, 3, e 4 para ET-MET, ET-PRO, NSL-MET e NLS-PRO, respectivamente). Posteriormente, as reads resultantes das quatro bibliotecas foram alocados em arquivo único para realização de um novo assembly (montagem 5). Foi utilizado, em todas as montagens, o default do programa **Newbler**; com exceção do percentual de cobertura que foi alterado de 40% para 90% após análises preliminares.

Foram efetuados, com o objetivo de verificar qual o melhor programa de montagem, testes com os montadores **CLC Genomics Workbench**, **MIRA**, **Newbler** e **Velvet**. O programa **Newbler v.2.5.3** (Roche) foi escolhido por produzir menor número de sequências (singlets mais contigs) (dados não mostrados).

Uma vez estabelecido o programa e os parâmetros identidade e cobertura, quatro montagens foram efetuadas para cada biblioteca. Uma para cada parâmetro: default, urt (use tip read), cDNA (admite splicing alternativo) e urt com cDNA simultaneamente. Não houve um parâmetro ideal para as montagens, reforçando que é preciso testar vários parâmetros para obtenção da melhor montagem, conforme pode ser visto no tópico "resultado e discussão".

As montagens 1 a 4 foram otimizadas através da utilização do programa **CAP3** (Huang & Madan, 1999), utilizando como parâmetros o default do programa (40% de cobertura e 90% de identidade).

Os programas e parâmetros utilizados para as montagens 1 a 4, estão representados na Figura 3.

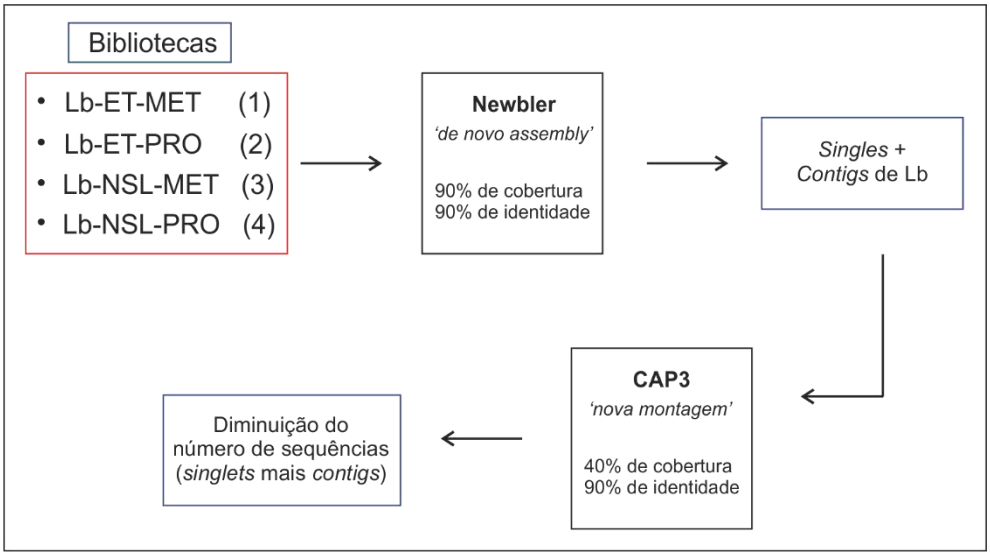

Figura 3. Organograma dos programas e parâmetros utilizados para execução das montagens 1 a 4.

A quinta montagem, com as reads das quatro bibliotecas, foi efetuada através da ferramenta **GS De Novo Assembler** do **Newbler** utilizando parâmetros default; exceção para cobertura que teve seu valor alterado de 40% para 90%.

Avaliação da qualidade da montagem

Após efetuar a montagem, conduziu-se um mapeamento das sequências obtidas (singlets mais contigs) nas ORFs anotadas automaticamente de L. (V.) braziliensis, sendo este procedimento efetuado através do programa **CLC bio Genomics Workbench** utilizando os valores default (50% de cobertura e 80% de identidade). Além de permitir inferir sobre a qualidade da montagem, essa análise forneceu conhecimento sobre a possível existência de novos genes.

Anotação dos contigs

O processo de anotação consiste na agregação do máximo de informação possível às sequências biológicas. Para algumas análises, a anotação dos transcritos foi efetuada através da transferência de informações do genoma anotado automaticamente de L. (V.) braziliensis para os contigs montados e mapeados. Para outras, utilizou-se a abordagem baseada em homologia de sequências através da plataforma de anotação funcional **Blast2GO** (Conesa et al., 2005; https://www.blast2go.com/). Basicamente, esta ferramenta executa três etapas: (1) busca por similaridade utilizando o algoritmo **BLAST** (Altschul et al., 1990), (2) mapeamento e (3) anotação. Na etapa de busca de similaridade (1), foi utilizado o **BLASTx** dentre os inúmeros tipos de BLASTs. O **BLASTx** traduziu a query para as seis frames possíveis e comparou cada uma destas com o conjunto de sequências proteicas contidas no banco de dados biológico nr (banco de dados do NCBI baixado em 06/10/2013). Considerou como possíveis homólogos os hits com valor de e-value menor que 1×10^{-10}. O banco de dados nr foi baixado e formatado localmente para busca pela ferramenta BLAST. Os demais parâmetros desta ferramenta não foram alterados.

Na etapa 2 (de mapeamento) foi executada em 3 passos: em “a” o algoritmo efetuou uma busca dos termos GO associado aos hits encontrados para cada query. Os arquivos 'gene info' e 'gene 2accession' fornecidos pelo NCBI na etapa 1 foram utilizados para obtenção do nome e/ou símbolos do gene. De posse destes nomes, foi realizada uma busca em “entradas espécie-específicas” da tabela de produtos gênicos do banco de dados do GO. Em “b”, os identificadores GI obtidos em “a” foram utilizados para busca de IDs no UniProt através de um arquivo de mapeamento disponibilizado pelo banco de dados PIR que inclui RefSeq, UniProt, Swissprot, PDB, PSD, GenPept e TrEMBL. Por último, os acessos resultantes do BLAST foram pesquisados diretamente no banco de dados do GO.

A etapa de anotação (3) consistiu em selecionar termos GOs obtidos na etapa 2 e associá-los à sequência query. A anotação é realizada através da aplicação de uma regra de anotação (RA) sobre os termos da ontologia encontrados. A regra visa encontrar as anotações mais específicas que tenha certo nível de confiabilidade. Neste passo, foi levado em consideração a identidade entre a query e o subject (hit), o tipo de experimento que resultou no termo GO e a estrutura do gráfico direto acíclico (GDA) do GO. Com exceção do valor de e-value, foram utilizados os parâmetros default do programa; configurados para otimizar a taxa entre termos GO encontrados (cobertura da anotação) e acurácia. Após estas três análises, os gráficos e as tabelas obtidas foram analisados.

Análise estatística

A análise da expressão gênica diferencial envolveu duas formas evolutivas diferentes para cada isolado: uma preparação enriquecida em formas infectivas promastigota metacíclica (MET) e uma preparação enriquecida de formas não infectivas promastigota procíclica (PRO). Esta análise auxiliou na elucidação do universo de transcritos codificadores de proteínas envolvidos nos processos de infectividade e virulência, os quais poderão ser utilizados como alvos em diversas aplicações biotecnológicas futuras.

O orfeoma representativo do nosso conjunto de dados, foi utilizado como template para a análise de expressão diferencial. Após o mapeamento com bowtie 2 e contagem com htseq-count, foi criada uma tabela contendo o número de reads de cada biblioteca que foram mapeadas em cada ORF.

Utilizou-se neste trabalho duas abordagens: uma para identificar os transcritos diferencialmente expressos e outra para avaliar a abundância dos transcritos; ambas para os quatro contrastes estabelecidos, os quais estão representados abaixo:

- ✓ ET-MET X ET-PRO (contraste 1)
- ✓ NSL-MET X NSL-PRO (contraste 2)
- ✓ ET-MET X NSL-MET (contraste 3)
- ✓ ET-PRO X NSL-PRO (contraste 4)

A primeira abordagem utilizou a metodologia de Blind através do pacote **DESeq** (Simon e Wolfgang, 2010) presente no Bioconductor (http://www.bioconductor.org/) utilizado o software livre R (R Development Core Team, 2012). O pacote **DESeq** foi escolhido por permitir utilizar dados sem replicata biológica, uma vez que as amostras utilizadas neste trabalho não possuem tais replicatas. Esta análise utiliza como input uma tabela como a exemplificada abaixo (Tabela 5).

Tabela 5. Dados hipotéticos para análise estatística pelo **DESeq**.

Identificação	reads da condição 1 mapeados	reads da condição 2 mapeados
ORF 1	2	2
ORF 2	0	0
ORF 3	35	3
ORF 4	0	3
ORF 5	5	0

Foram desconsideradas da análise as ORFs que não tiveram read mapeada ao menos por uma das bibliotecas comparadas, pois o valor zero influencia a estimativa da variância.

Foi realizada pelo DESeq uma normalização dos dados para tornar os valores de expressão gênica comparáveis, uma vez que o sequenciamento das bibliotecas gera quantidades e tamanhos diferentes de reads. Nesse método o fator de escala para uma determinada amostra é dado pela mediana dos valores de contagem de cada transcrito divido pela média geométrica das contagens de todas as amostras. Essa metodologia é baseada na suposição que a maioria dos transcritos não são diferencialmente expressos.

Após a normalização, calculou-se o valor de foldchange, que significa quantas vezes uma determinada ORF é mais expressa que outra numa dada condição. Posteriormente, estes valores foram transformados para escala logarítmica em base 2, visando estreitar a faixa dos valores comparados. Foram obtidos gráficos de dispersão considerando o valor de log2 do foldchange ajustado.

A segunda abordagem consiste numa análise que utiliza o valor de RPKM (**Reads P**er **K**ilobase of exon modelo per **M**illion mapped sequence reads) como normalização dos dados.

Esta normalização contorna os problemas gerados por quantidades distintas de reads por biblioteca e tamanho diferente dos transcritos utilizados como referência para mapeamento.

Os genes obtidos destas análises serão essenciais para aprofundar o conhecimento em diversas linhas de pesquisas relacionadas à infectividade/virulência de L. (V.) braziliensis e de outras espécies dentro do gênero Leishmania.

RESULTADOS E DISCUSSÃO

Bibliotecas sequenciadas

Após sequenciamento, efetuou-se o download de dois arquivo referente à amostra ET-MET através do site da 454 Life Science. Os arquivos ***.sff** das outras três amostras foram baixados do servidor do CATG através do protocolo de transferência de arquivo (FTP), sendo dois arquivos para a amostra NSL-MET, um arquivo para a amostra ET-PRO e um arquivo para a amostra NSL-PRO.

Como resultado, foram obtidas aproximadamente 3,1 milhões de reads referentes às quatro bibliotecas (Tabela 6).

Tabela 6. Panorama geral das bibliotecas de Leishmania (V.) braziliensis**.**

Isolado	Formas evolutiva	nº de reads	Variação de tamanho	Tamanho médio da read	% de reads ribossomais
ET	MET	916.546	46 - 1201	364	125.231=13,66
	PRO	589.554	49 – 1200	398	151.916=25,76
NSL	MET	1083.312	47 - 1201	370	312.160=28,81
	PRO	506.312	50 - 1201	385	108.095=21,35
Total	**-----**	**3.095.724**		**média = 379.25**	**média = 22,40**

Qualidade das reads

Várias características intrínsecas às reads sequenciadas interferem na montagem. Dentre estas, cita-se: conteúdo GC (percentual de guanina mais citosina), adaptadores, vetores, contaminação, sequências super-representadas, regiões de baixas qualidades, reads duplicadas, k-mers, bases ambíguas, quimeras, dentre outras. Estas características foram analisadas com o software **fastQC** utilizando como entrada os arquivos no formato ***.fastq**.

Como esperado, devido a limitada processividade da enzima DNA polimerase, as quatro bibliotecas apresentaram baixa qualidade nas últimas posições das reads (Figura 4). No eixo das abscissas (eixo X) tem-se a posição ou intervalo de posição das reads em pares de bases e no eixo das coordenadas (eixo Y) tem-se o valor Phred associado a cada posição ou intervalo de posições. A região verde (no topo da figura) corresponde a valores de Phred

considerados ideais para montagem. A região rosa (na parte inferior da figura) corresponde a valores de Phred considerados inapropriados para montagem e a região laranja (no meio da figura) corresponde a valores de Phred aceitáveis para montagem. A linha azul, representa o valor de qualidade médio ao longo das reads. As barras amarelas correspondem a um boxplot mostrando a variação do valor de qualidade Phred numa determinada posição ou intervalo de posições das reads, sendo a linha vermelha em seu interior a mediana dos valores de qualidade Phred. As projeções partindo das barras amarelas correspondem aos outliers (posições de algumas reads com valores de qualidades muito discrepante da média) (Figura 4).

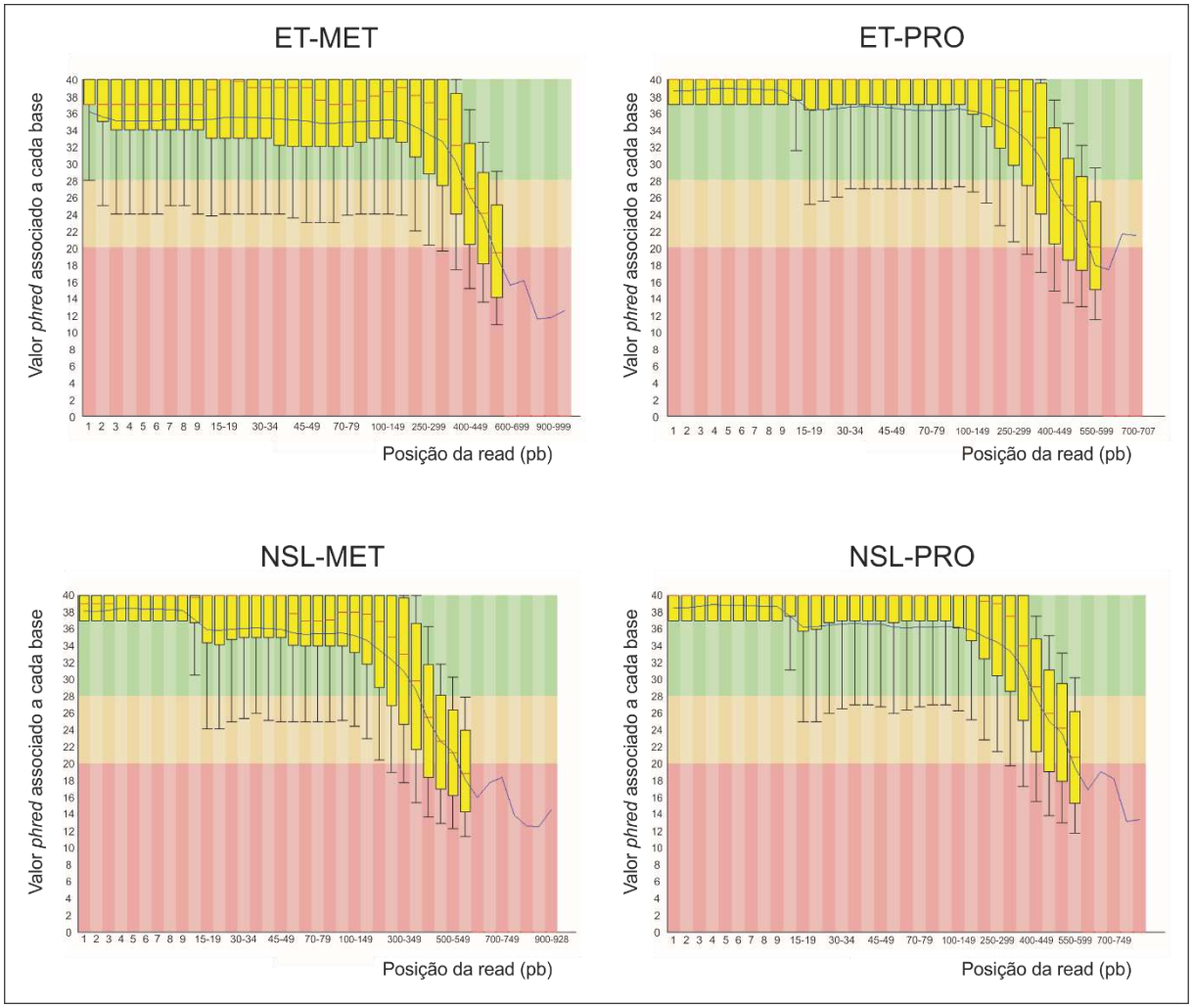

Figura 4. Análise de qualidade das reads das quatro bibliotecas sequenciadas. No eixo X, tem-se as posições das reads (pb) e no eixo Y tem-se o valor Phred associado a estas posições. A região verde no topo da figura corresponde a valores de Phred considerados ideais para montagem. A região rosa na parte inferior da figura corresponde a valores de Phred considerados inapropriados para montagem e a região laranja no meio da figura corresponde a valores de Phred aceitáveis para montagem. A linha azul, ao longo de todo o comprimento da read, representa o valor de qualidade média. As barras amarelas correspondem a um boxplot mostrando a variação do valor de qualidade Phred numa determinada posição ou intervalo de posições, sendo a linha vermelha no meio da barra amarela a mediana dos valores de qualidade Phred. As projeções partindo das barras amarelas correspondem ao outliers (posições de algumas reads com valores de qualidades muito discrepante da média).

Foram apresentados apenas os gráficos referentes à qualidade, embora o programa **fastQC** também tenha gerado gráficos referentes a:

- ✓ Estatísticas básicas,
- ✓ Score de qualidade por sequência,
- ✓ Tipo de base por posição,
- ✓ Conteúdo GC por base,

- ✓ Conteúdo GC por sequência,
- ✓ Conteúdo N por base,
- ✓ Distribuição do comprimento da sequência,
- ✓ Nível de duplicação das sequências,
- ✓ Sequências super representadas,
- ✓ Tipos de k-mers e sua representatividade.

Tratamento das reads

Os tipos de tratamentos efetuados são descritos, em seguida, para cada biblioteca.

A) ET-MET

1) Trimagem, com **fastx_trimmer**, dos 14 primeiros nucleotídeos das reads. Foi visto em outro gráfico fornecido pelo fastQC (dados não mostrados) que estes primeiros nucleotídeos correspondem à alguma sequência indesejada que foi incorporada, provavelmente, durante uma das etapas do sequenciamento. Este comportamento foi observado também para as outras três bibliotecas.
2) Trimagem, com **fastx_trimmer**, dos nucleotídeos após a posição 450, pois estes apresentaram baixo valor de qualidade. Este comportamento foi observado também para as outras três bibliotecas.
3) Trimagem, com **prinseq-lite.pl**, de reads com valor de Phred médio menor que 20. Com este tratamento, 3.235 sequências foram eliminadas (0,35 % do total).
4) Trimagem, com **prinseq-lite.pl**, de reads contendo mais de duas ambiguidades, representada pela letra 'N'. Um total de 1.375 sequências foram eliminadas (0,15% do total).
5) Trimagem, com **prinseq-lite.pl**, de reads menores 60 nucleotídeos. Um total de 67.216 sequências foram eliminadas (7,37% do total).

Com estes passos, foram descartadas 71.867 sequências (7,84 % do total).

B) ET-PRO

1) Trimagem, com **fastx_trimmer**, dos 24 primeiros nucleotídeos das reads.
2) Trimagem, com **fastx_trimmer**, dos nucleotídeos após a posição 425.
3) Trimagem, com **prinseq-lite.pl**, de reads com valor de Phred médio menor que 20. Com este tratamento, 530 sequências foram eliminadas (0,09% do total).
4) Trimagem, com **prinseq-lite.pl**, de reads contendo mais de duas ambiguidades. Foram eliminadas 2.139 sequências (0,36% do total).

5) Trimagem, com **prinseq-lite.pl**, de reads menores 60 nucleotídeos. Um total de 26.008 sequências foram eliminadas (4,42% do total).

Com estes passos, 28.677 sequências foram descartadas (4,86% do total).

C) NSL-MET

1) Trimagem, com **fastx_trimmer**, dos 24 primeiros nucleotídeos das reads.
2) Trimagem, com **fastx_trimmer**, dos nucleotídeos após a posição 450
3) Trimagem, com **prinseq-lite.pl**, de reads com valor de Phred médio menor que 20. Com este tratamento, 4.561 sequências foram eliminadas (0,42% do total).
4) Trimagem, com **prinseq-lite.pl**, de reads contendo mais de duas ambiguidades. Um total de 3.836 sequências foram eliminadas (0,36% do total).
5) Trimagem, com **prinseq-lite.pl**, de reads menores 60 nucleotídeos. Um total de 56.328 sequências foram eliminadas (5,2% do total).

Com estes passos, foram descartadas 64.725 sequências (5,97% do total).

D) NSL-PRO

1) Trimagem, com **fastx_trimmer**, dos 24 primeiros nucleotídeos das reads.
2) Trimagem, com **fastx_trimmer**, dos nucleotídeos após a posição 425.
3) Trimagem, com **prinseq-lite.pl**, de reads com valor de Phred médio menor que 20. Com este tratamento, 618 sequências foram eliminadas (0,13% do total).
4) Trimagem, com **prinseq-lite.pl**, de reads contendo mais de duas ambiguidades. Um total de 1.909 sequências foram eliminadas (0,38% do total).
5) Trimagem, com **prinseq-lite.pl**, de reads menores 60 nucleotídeos. Um total de 28.698 sequências foram eliminadas (5,69% do total).

Com estes passos, foram descartadas 31.225 sequências (6,17% do total).

Conforme esperado, a qualidade média das reads aumentou significativamente após os tratamentos realizados (Figura 5).

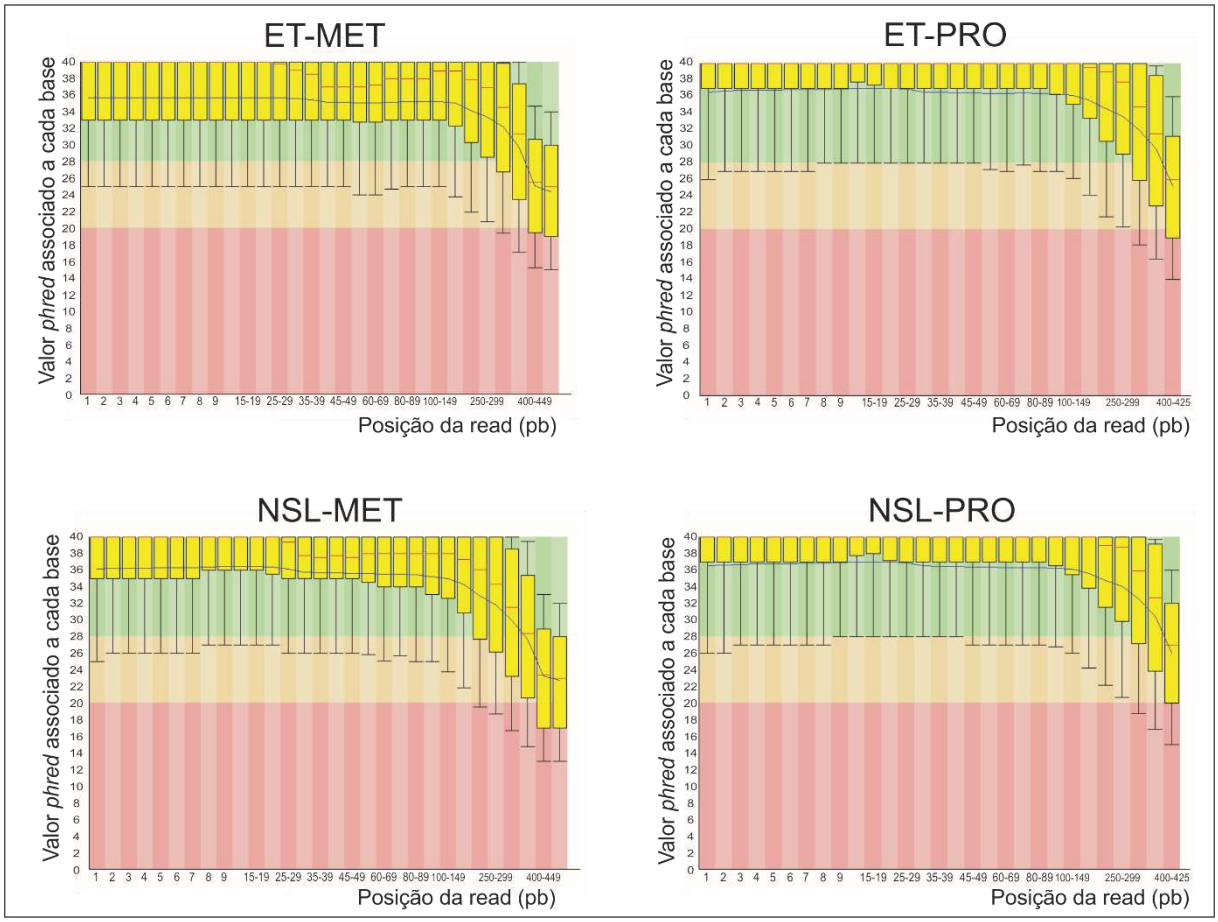

Figura 5. **Melhoria na qualidade média das reads após os tratamentos.** No eixo X tem-se as posições das reads (pb) e no eixo Y tem-se o valor Phred associado a estas posições. A região verde (no topo da figura) corresponde a valores de Phred considerados ideais para montagem. A região rosa (na parte inferior da figura) corresponde a valores de Phred considerados inapropriados para montagem e a região laranja (no meio da figura) corresponde a valores de Phred aceitáveis para montagem. A linha azul, ao longo de todo o comprimento da read, representa o valor de qualidade média. As barras amarelas correspondem a um boxplot mostrando a variação do valor de qualidade Phred numa determinada posição ou intervalo de posições, sendo a linha vermelha no meio da barra amarela a mediana dos valores de qualidade Phred. As projeções partindo das barras amarelas correspondem ao outliers (posições de algumas reads com valores de qualidades muito discrepante da média).

Os arquivos contendo as reads tratadas das quatro bibliotecas foram concatenados com a finalidade de efetuar a montagem 5. Em seguida, foram carregados no software **fastQC** visando comparar a qualidade das reads antes e após os tratamentos (Figura 6). Conforme esperado, o valor de qualidade das reads aumentou após tratamento.

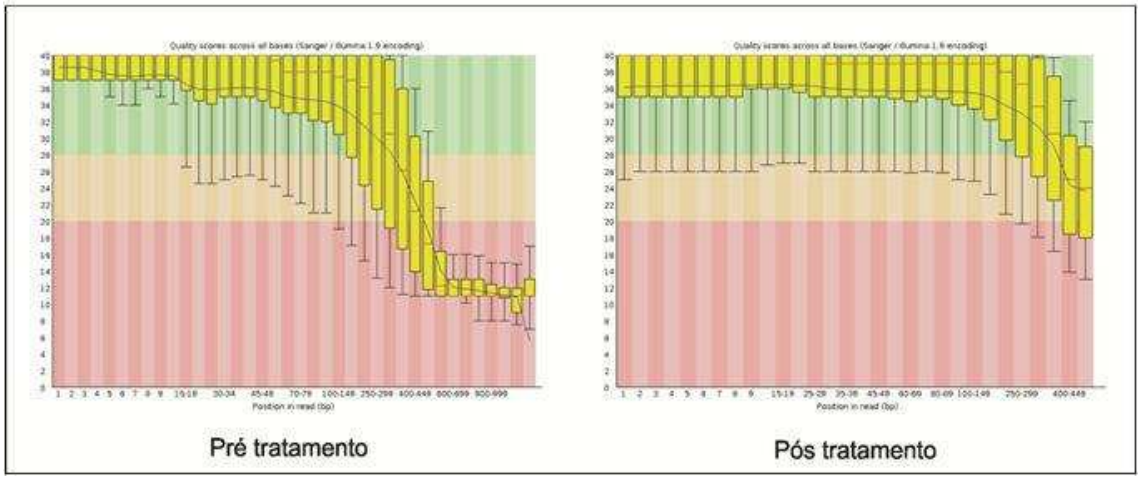

Figura 6. Qualidade das reads das quatro bibliotecas antes e após tratamento. No eixo X tem-se a posição da reads (pb) e no eixo Y tem-se o valor Phred associado a cada posição ou intervalo de posições. A região verde (no topo da figura) corresponde a valores de Phred considerados ideais para montagem. A região rosa (na parte inferior da figura) corresponde a valores de Phred considerados inapropriados para montagem e a região laranja (no meio da figura) corresponde a valores de Phred aceitáveis para montagem. A linha azul, ao longo de todo o comprimento da read, representa o valor de qualidade média. As barras amarelas correspondem a um boxplot mostrando a variação do valor de qualidade Phred numa determinada posição ou intervalo de posições, sendo a linha vermelha no meio da barra amarela a mediana dos valores de qualidade Phred. As projeções partindo das barras amarelas correspondem ao outliers (posições de algumas reads com valores de qualidades muito discrepante da média).

E) Reads paired end provenientes do SRA

O resultado após tratar as reads provenientes do SRA está detalhado na Tabela 7.

Tabela 7. **Número e tamanho de reads dos arquivos obtidos do SRA antes e após a trimagem.**

CARACTERÍSTICA	(nº/tamanho)
seqs. arquivo 1	26.007.384
bases do arquivo 1	1.976.561.184
tamanho médio do read do arquivo 1	76 pb
seqs. arquivo 2	26.007.384
bases do arquivo 2	1.976.561.184
tamanho médio do arquivo 2	76 pb
seqs. boas (em pares) *	23.059.397
bases boas (em pares)*	3.504.870.125
tamanho médio do par de reads*	151.99
singletons bons do arquivo 1*	1.085.194 (4.17%)
bases boas dos singletons do arquivo 1*	82.457.217
tamanho médio dos singletons do arquivo 1*	75.98
singletons bons do arquivo 2*	173.245 (0.67%)
bases boas dos singletons do arquivo 2*	13.163.937
tamanho médio dos singletons do arquivo 2*	75.98
seqs. ruins do arquivo 1*	1.862.793 (7.16%)
bases ruins do arquivo 1*	141.572.268
tamanho médio das seq. ruins do arquivo 1*	76.00
seq. ruins do arquivo 2*	1.085.194 (4.17%)
bases ruins do arquivo 2*	82,474,744
tamanho médio das seq. ruins do arquivo 2*	76.00
seqs. filtradas com **trim_qual_left**	217.828
seqs. filtradas com **min_len**	651.201
seqs. filtradas com **min_qual_mean**	1.342.573
seqs. filtradas com **ns_max_n**	40.924
seqs. filtradas com **derep**	1.407.455

*após tratamento, seq(s)= sequência(s), singletons corresponde a uma sequência do par, sendo a outra descartada devido a alguns dos parâmetros utilizados no tratamento das reads. Em negrito encontra-se parâmetros do **prinseq_lite** especificados na linha de comando.

A qualidade das reads provenientes do banco de dados SRA, antes e após o tratamento, está representada na Figura 7. Após tratamento, as sete primeiras posições das reads sequenciadas no sentido forward (5') passaram a apresentar valor de qualidade Phred médio igual a 38 (Figura 7c) e as nove primeiras posições das reads sequenciadas no sentido reverse (3') apresentaram valor de qualidade Phred médio próximo de 36 (Figura 7d).

Conclui-se pela análise desta figura que, provavelmente, as reads do SRA foram submetidas a tratamento prévio, antes de serem depositadas no NCBI, uma vez que possuem valor médio de qualidade acima de 20 em suas extremidades antes do tratamento (Figura 7a).

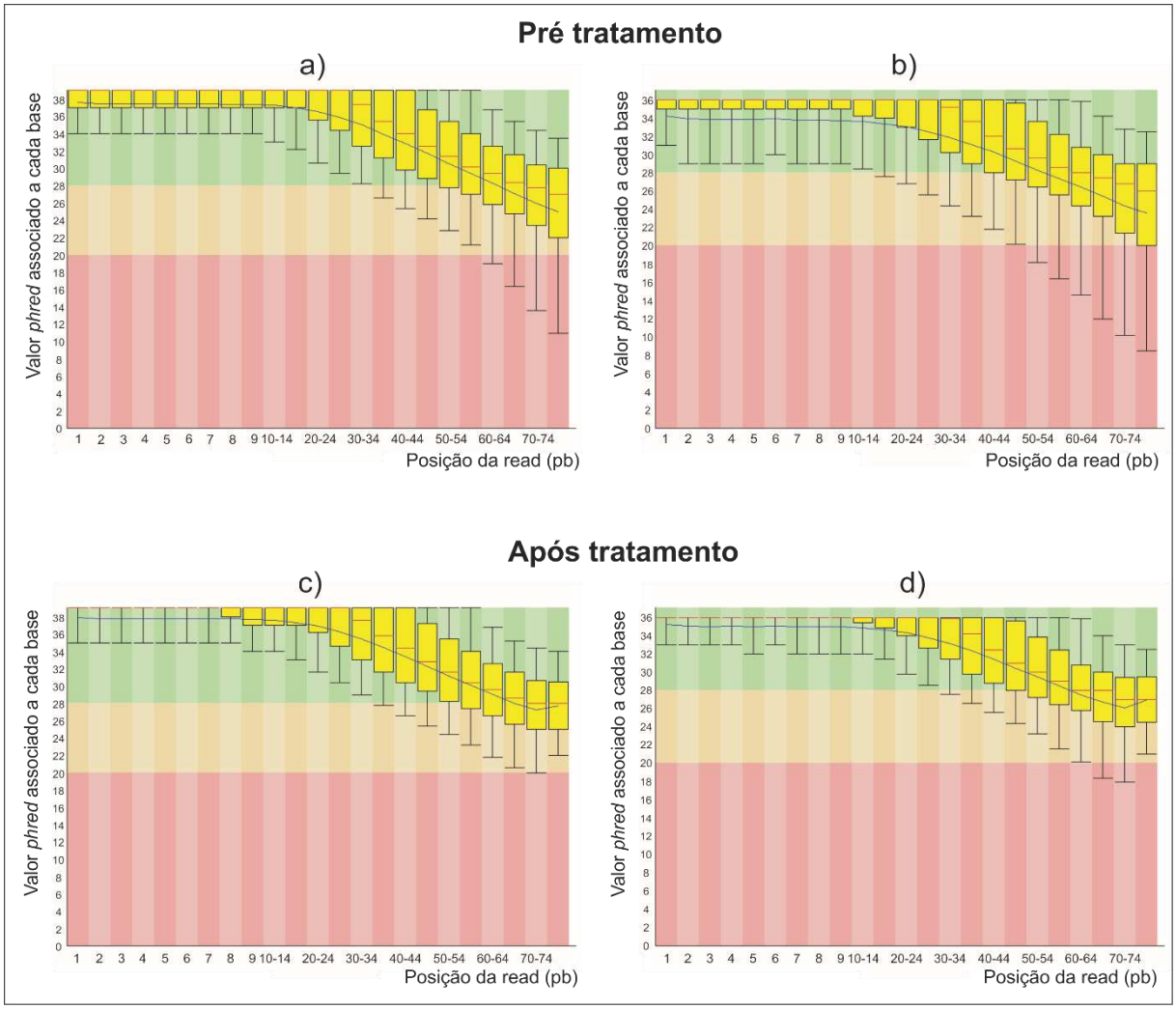

Figura 7. Qualidade das reads obtidas do SRA antes e após tratamento.

Com esta análise foram aproveitadas 47.377.233 reads (91,08%), sendo 4.639.717 descartadas (8.92%).

Critérios para montagem

Os genomas de espécies do gênero Leishmania possuem um conteúdo gênico e arquitetura genômica conservada, possuindo as espécies de L. major, L. infantum e L. (V.) braziliensis, respectivamente 8.298, 8.154 e 8.153 genes codificadores de proteínas (Ivens et al., 2005; Peacock et al., 2007; Smith, Peacock e Cruz, 2007). Por esta razão, é esperado obter como resultado da montagem um número total de sequências (singlets mais contigs) próximo deste valor.

Os programas e parâmetros influenciam a montagem de transcriptomas, devendo estes serem estabelecidos de acordo com os objetivos de cada trabalho. Uma vez escolhido o programa **Newbler**, o próximo passo foi estabelecer os valores dos parâmetros cobertura e identidade a serem utilizados. Uma montagem, com as reads tratadas, mostrou que não houve mudança significativa no somatório do número de singlets e contigs obtidos ao elevar o valor de cobertura de 40 para 90%. Desta forma, estabeleceu-se 90% de cobertura para realização da montagem (Tabela 8). Como resultado da montagem, tem-se singlets que são sequências únicas que não se alinharam com nenhuma outra e os contigs que são sequências únicas geradas pela sobreposição de regiões de duas ou mais sequências.

Tabela 8. Alteração do resultado da montagem ao elevar o valor da cobertura no Newbler.

Amostra	NEWBLER-40*	NEWBLER-90*
ET-MET	18.781	18.674
ET-PRO	22.089	22.155
NSL-MET	19.894	19.863
NSL-PRO	19.787	19.237

*Estes valores correspondem ao somatório do número de singlets e contigs.

Foi efetuada em seguida, com o intuito de diminuir ainda mais o número de sequências (singlets mais contigs), uma nova montagem com o programa **CAP3** utilizando o arquivo de saída do **Newbler**. Como resultado, houve uma diminuição do número de sequências utilizando os parâmetros default do CAP3 (40% de cobertura e 90% de identidade) (Tabela 9).

Tabela 9. Diminuição do número de sequências (singlets mais contigs) ao otimizar a montagem com o programa CAP3.

Biblioteca	Newbler	CAP3		
	(*Singlets*+*Contigs*)	*Singlets**	*Contigs**	(*Singlets*+*Contigs*)*
ET-MET	18.674	16.872	665	17.537
ET-PRO	22.155	20.335	703	21.038
NSL-MET	19.863	17.740	817	18.557
NSL-PRO	19.237	17.664	640	18.304

*Neste caso os singlets correspondem a singlets ou contigs do passo anterior que não colapsaram após utilização do **CAP3**; os contigs se referem a sequências resultantes da união de pelo menos duas sequências (singlets e/ou contigs) do **Newbler**.

Posteriormente o valor do parâmetro cobertura do **CAP3** foi alterado para 90%. Ao aumentar a estringência da montagem, o número de contigs aumentou. Por isto, utilizou-se 40% de cobertura a fim de obter o menor número de contigs (Tabela 10).

Tabela 10. Alteração do resultado da montagem ao elevar o valor do parâmetro cobertura no CAP3.

Amostra	Newbler*	% Cobertura	CAP3*
ET-MET	18.674	40	17.537
		90	18.065
ET-PRO	22.155	40	21.038
		90	21.480
NSL-MET	19.863	40	18.557
		90	18.822
NSL-PRO	19.237	40	18.304
		90	18.687

*Estes valores correspondem a soma do número de singlets e contigs.

Por fim, alterou-se o valor de identidade para o limite inferior aceito pelo **CAP3** que é 66% e não se observou diferenças significativas (dados não mostrados). Assim, definiu-se os parâmetros para a montagem final dos transcriptomas (Tabela 11).

Tabela 11. Parâmetros escolhidos para montagem do transcriptoma.

Passo	Programa	% de cobertura	% de identidade
Primeiro	**Newbler**	90	90
Segundo	**CAP3**	40	90

De novo assembly (montagem)

As montagens realizadas neste trabalho foram divididas em duas categorias. A primeira refere-se à montagem com as reads de cada biblioteca individualmente e a segunda refere-se à montagem utilizando as reads de todas as bibliotecas. Esta distinção é necessária porque a segunda categoria representa melhor o transcriptoma desta espécie, comparado ao resultado das montagens obtidas utilizando o conjunto de reads de cada biblioteca. No entanto a primeira foi necessária para analisar qualitativamente a expressão diferencial.

O resultado das quatro montagens, para cada biblioteca, está representado na Tabela 12. O número de sequências obtidas com a opção cDNA não foi o maior dentre os parâmetros testados, como esperado para a maioria dos organismos eucariotos. Inclusive foi com este parâmetro que se obteve a melhor montagem para ET-PRO e NSL-PRO. Isto pode ser explicado pelo fato das espécies de Leishmania praticamente não realizam cis-splicing. Não houve um parâmetro ideal para as quatro montagens, reforçando a ideia de que é preciso testar vários parâmetros para a obtenção da melhor montagem.

Tabela 12. Influência dos parâmetros urt e cDNA na montagem.

Parâmetros	ET-MET	ET-PRO	NSL-MET	NSL-PRO
default	14.639	14.185	**15.600***	12.278
urt	**14.373***	15.022	15.960	13.596
cDNA	15.865	**13.642***	19.933	**11.858***
cDNA/urt	21.090	24.437	25.210	24.476

*Os valores em negrito correspondem ao menor número de sequências (singlets mais contigs) obtidas para cada amostra.

Foram obtidos para a primeira categoria de montagem, 14.373, 13.642, 15.600 e 11.858 sequências (singlets e contigs) para as bibliotecas 1, 2, 3 e 4, respectivamente (Tabela 12). Posteriormente efetuou-se uma nova montagem com o programa **CAP3**, utilizando o arquivo

de saída do **Newbler**, visando diminuir os números de singlets e contigs obtidos anteriormente. Desta forma foi possível diminuir para 14.102, 13.600, 15.301 e 11.678 o número de sequências (*singlets* mais *contigs*) para as bibliotecas 1, 2, 3 e 4, respectivamente. A montagem com este programa resultou num número de sequências muito próximo do valor encontrado com **Newbler** e por isto não foi aproveitada. Outro fator que contribui para o não aproveitamento das montagens efetuadas pelo **CAP3** foi a eliminação do valor de qualidade das sequências geradas.

Visando aumentar a confiabilidade dos resultados, selecionou-se apenas as sequências maiores que 100 pb, uma vez que contigs pequenos tem maior probabilidade de alinhar com sequências não homologas. Desta forma, obteve-se como resultado final da montagem 14.362, 13.145, 14.899 e 11.434 sequência (singlets mais contigs) para as amostras 1, 2, 3 e 4, respectivamente (em negrito na Tabela 13).

Tabela 13. Número de contigs obtidos após a montagem do Newbler e do CAP3.

Amostra	Newbler*	Tamanho médio	N50	> 100pb*	>500 pb	>contig
ET-MET	14.373	1.755	2.148	**14.362**	11.580	24.357
ET-PRO	13.642	1.072	1.125	**13.145**	11.013	12.173
NSL-MET	15.600	1.623	1.850	**14.899**	11.231	21.369
NSL-PRO	11.858	1.033	1.069	**11.434**	9.037	13.434

*Estes valores correspondem a soma do número de singlets e contigs.

O número de sequências, após a montagem, foi maior para as bibliotecas 1 e 3 que possuem maior número de sequências. No entanto, o tamanho médio dos contigs nestas bibliotecas foi maior comparado ao tamanho dos contigs montados nas bibliotecas 2 e 4, que foram sequenciadas utilizando somente uma placa.

A montagem com o parâmetro default do **Newbler**, utilizando as reads de todas as amostras ($3{,}1 \times 10^6$ reads), resultou em 14.032 sequências (singlets mais contigs) (Tabela 14). Embora próximo ao número de sequências encontrado nas montagens individuais, o tamanho dos contigs obtidos foi maior. Este resultado ocorreu devido à maior cobertura proporcionada pelos dados. Após filtrar as sequências maiores que 100 pb, obteve-se 14.017 sequências como conjunto final da montagem (Tabela 14).

Tabela 14. Número de sequências (singlets e contigs) obtidos da montagem com as reads das quatro bibliotecas utilizando o Newbler.

Amostra	Sequências*	Tamanho médio	N50	> 100pb*	> 500 pb	> 1000	> contig
ET_e_NSL	14.032	2.426	3.237	**14.017**	10.738	8.275	27.782

*Estes valores correspondem a soma do número de singlets e contigs.

Como resultado, obteve-se ainda, 10.738 sequências maiores que 500 pb e 8.275 sequências maiores que 1.000 pb.

Alguns trabalhos mostraram que diferentes espécies de Leishmania possuem aproximadamente 8.200 genes comuns e poucos genes espécie-específicos (Ivens et al., 2005; Peacock et al., 2007; Smith, Peacock e Cruz, 2007). Considerando também que os genomas de espécies do gênero Leishmania são altamente conservados e que apresentam elevada sintenia gênica (Peacock et al., 2007), esperava-se, a princípio, neste trabalho obter um número total de sequências (singlets mais contigs) próximo aos 9.240 transcritos encontrado em L. major (Ivens et al., 2005; Smith, Peacock e Cruz, 2007), que seria o somatório do número: de genes, pseudogenes e RNA não codificadores (tRNAs, rRNAs, snoRNAs, snRNAs, dentre outros). No entanto, devido ao fato da transcrição gênica em L. (V.) braziliensis ser policistrônica, cada transcrito não corresponde necessariamente a um gene, e consequentemente, tem-se transcritos que não codificam proteínas e outros que codificam uma ou mais proteínas. Desta forma não foi possível, à princípio, estabelecer com precisão uma relação entre o número de transcritos obtidos com o número de genes desta espécie.

Estas 14.032 sequências totalizaram 26.974.518 de nucleotídeos e a distribuição de seus tamanhos está representada abaixo (Figura 8).

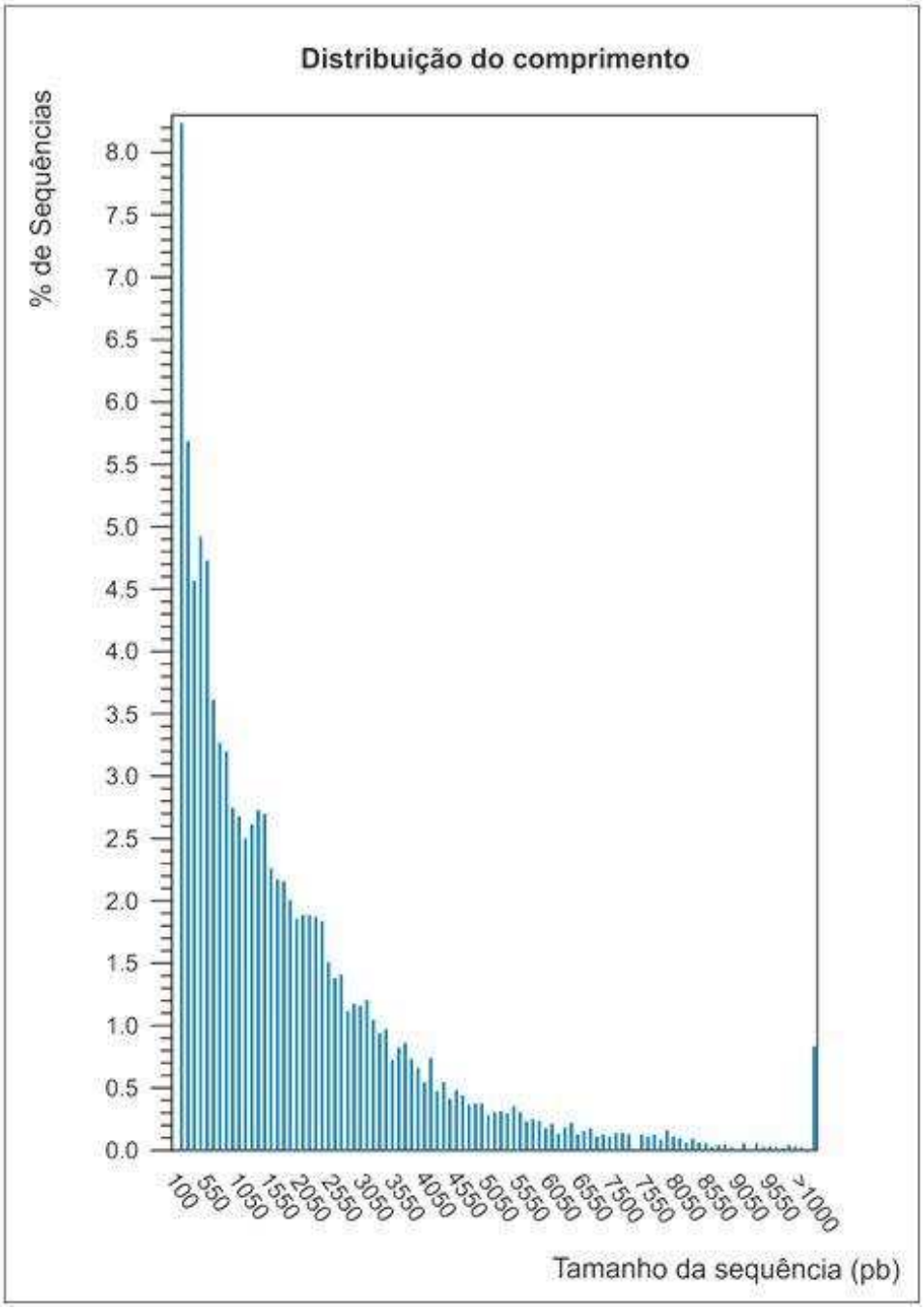

Figura 8. Distribuição do comprimento das sequências montadas (somatória de singlets e contigs) (eixo X) em relação ao percentual de sequências (eixo Y).

Teste de qualidade da montagem

Após efetuar a montagem, conduziu-se um mapeamento das sequências obtidas (singlets mais contigs) nas ORFs anotadas automaticamente do cromossomo 1 de L. (V.) braziliensis, sendo este procedimento efetuado através do programa **CLC bio Genomics Workbench** utilizando os valores default (50% de cobertura e 80% de identidade) (Figura 9).

Esta análise permitiu inferir que a montagem foi bem sucedida por representar bem as ORFs preditas automaticamente do cromossomo 1 desta espécie. Este cromossomo foi escolhido, para esta análise, por ser o menor dentre os 35 desta espécie, facilitando assim sua representação.

Figura 9. Mapeamento dos transcritos montados (singlets mais contigs) nas ORFs anotadas automaticamente do cromossomo 1 de L. (V.) braziliensis**.** As setas largas amarelas representam as ORFs que foram preditas nos sentidos senso (ponta da seta para direita) e anti-senso (ponta da seta para esquerda) do cromossomo. Tem-se abaixo destas setas uma linha preta contínua que representa o cromossomo e abaixo desta linha, encontram-se retas de diferentes tamanhos representando o mapeamento dos transcritos montados neste trabalho. Estes, geralmente, tem o mesmo tamanho das ORFs e encontram-se justamente abaixo delas.

Mapeamento para análise de enriquecimento do orfeoma

A estratégia adotada foi fazer quatro mapeamentos no genoma anotado de L. (V.) braziliensis: (1) com o somatório das reads das amostras 1 a 4 com tamanho entre 58 e 100 pb, (2) com o somatório das reads maiores que 100 pb deste mesmo conjunto de dados, (3) com as reads paired end provenientes do SRA e (4) com as singlets originadas após tratamento dos dados do SRA com prinseq-lite. Posteriormente efetuou-se um merge destes 4 mapeamentos, obtendo-se um mapeamento único.

Mapeamento 1

A filtragem das sequências com tamanho entre 58 e 100 pb foi necessária devido a necessidade de utilizar-se parâmetros mais rigorosos para o mapeamento, devido ao menor tamanho destas reads. Primeiramente, separou-se as 141.810 reads menores que 100 pb das 2.899.230 reads totais. Após teste de vários parâmetros de cobertura e identidade, optou-se por utilizar 95% de cobertura e 95% de identidade. Com estes parâmetros, foram mapeadas 84.239 reads (59.40% do total).

Mapeamento 2

Após testes, utilizou-se 90% para os parâmetros cobertura e identidade. Com estes parâmetros, foram mapeadas 1.891.269 de reads (68.66% do total de 2.754.344 reads).

Mapeamento 3

Após testes, utilizou-se 90% para os parâmetros cobertura e identidade para mapeamento das 46.118.794 reads paired end obtidas após o tratamento com prinseq-lite. As funções Map randomy e auto detectar paired distance foram habilitadas. Com estes parâmetros, foram mapeadas 36.205.892 reads (78.51% do total).

Mapeamento 4

Após testes, utilizou-se 95% para os parâmetros cobertura e identidade para mapeamento das 1.258.439 singlets que foram separadas do par durante o tratamento pelo prinseq-lite. Com estes parâmetros, foram mapeadas 944.005 reads (75.01% do total).

Merge dos mapeamentos

Os quatro mapeamentos foram combinados e o mapeamento resultante (merge 1) foi utilizado para comparação entre as reads mapeadas e o orfeoma anotado automaticamente de L. (V.) braziliensis.

Erros de montagem e possibilidade de extensão de gaps no genoma anotado de L. (V.) braziliensis

Embora depositados como 35 sequências representando os 35 cromossomos, o genoma de L. (V.) braziliensis encontra-se ainda bastante fragmentado, uma vez que estes 35 cromossomos são na verdade 955 contigs conectados pela letra “N”, que representa regiões que estão ausentes no genoma depositado. A análise deste genoma permitiu encontrar erros de montagem, que podem influenciar diversas análises ao utilizá-lo como referência. A Figura 10 representa um destes erros (representado pelo retângulo vermelho), que se inicia na posição 334.023 do cromossomo 6 de L. (V.) braziliensis.

Verifica-se ainda que as reads mapeadas são capazes de fechar um pedaço do gap, representado pela base ambígua “N” (lado esquerdo da linha azul mostrada por uma seta na parte inferior da figura).

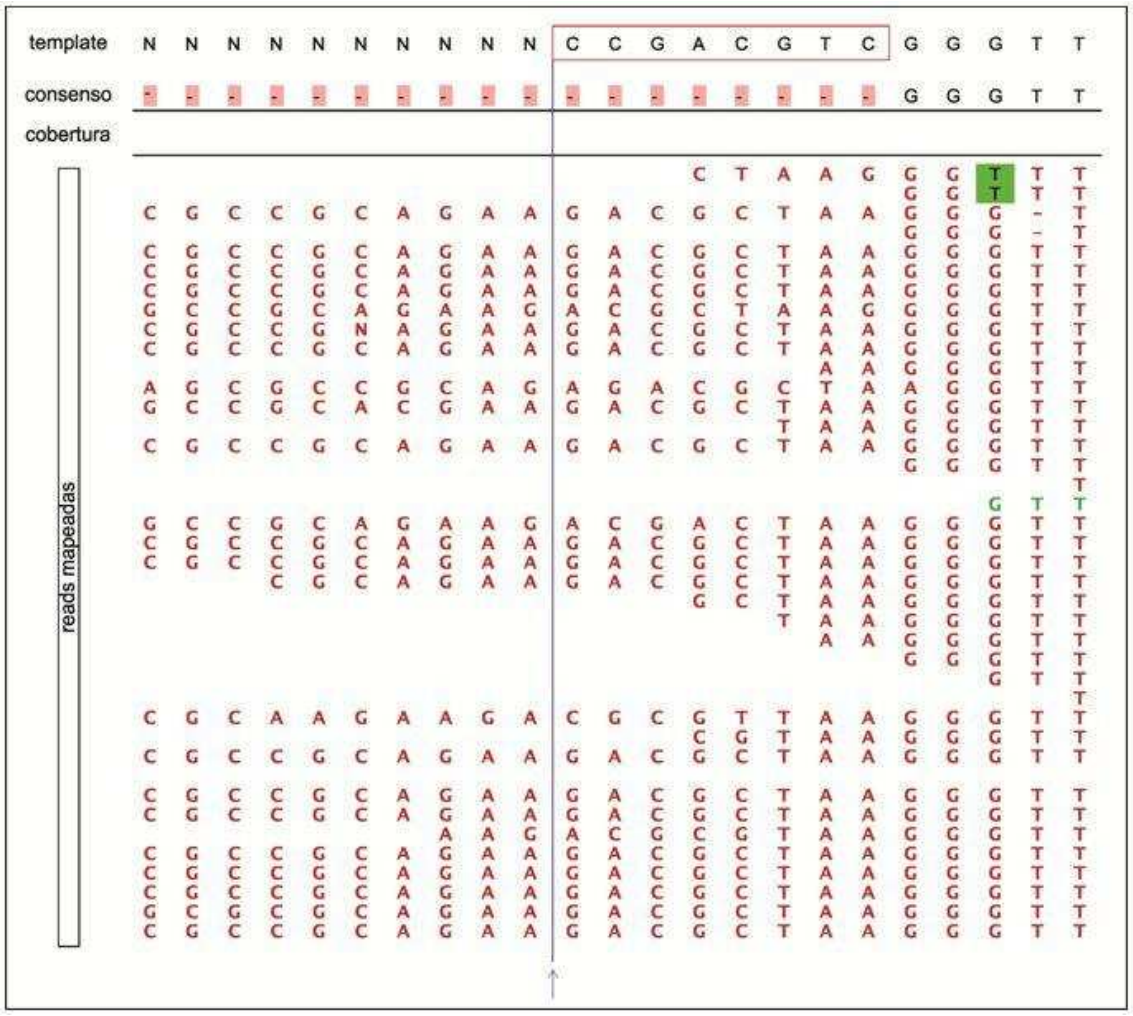

Figura 10. **Possibilidade de correção de erros e de extensão de gaps no genoma.** O mapeamento das reads evidencia um erro de montagem (retângulo vermelho), visto que a maioria das reads apresentam os mesmos nucleotídeos e estes diferem dos oito nucleotídeos do retângulo analisado. Verifica-se ainda que as reads mapeadas são capazes de fechar um pedaço do gap representado pela base ambígua “N” (lado esquerdo da linha azul sinalizada por uma seta na parte inferior da figura). Os nucleotídeos das reads utilizadas neste mapeamento apresentam valor Phred de qualidade acima de 20 (dados não mostrados).

Em seguida, encontram-se outros dois exemplos evidenciando a possibilidade de extensão de gaps no genoma anotado de L. (V.) braziliensis. O primeiro refere-se à posição 956.520 do cromossomo 2 de L. (V.) braziliensis (Figura 11).

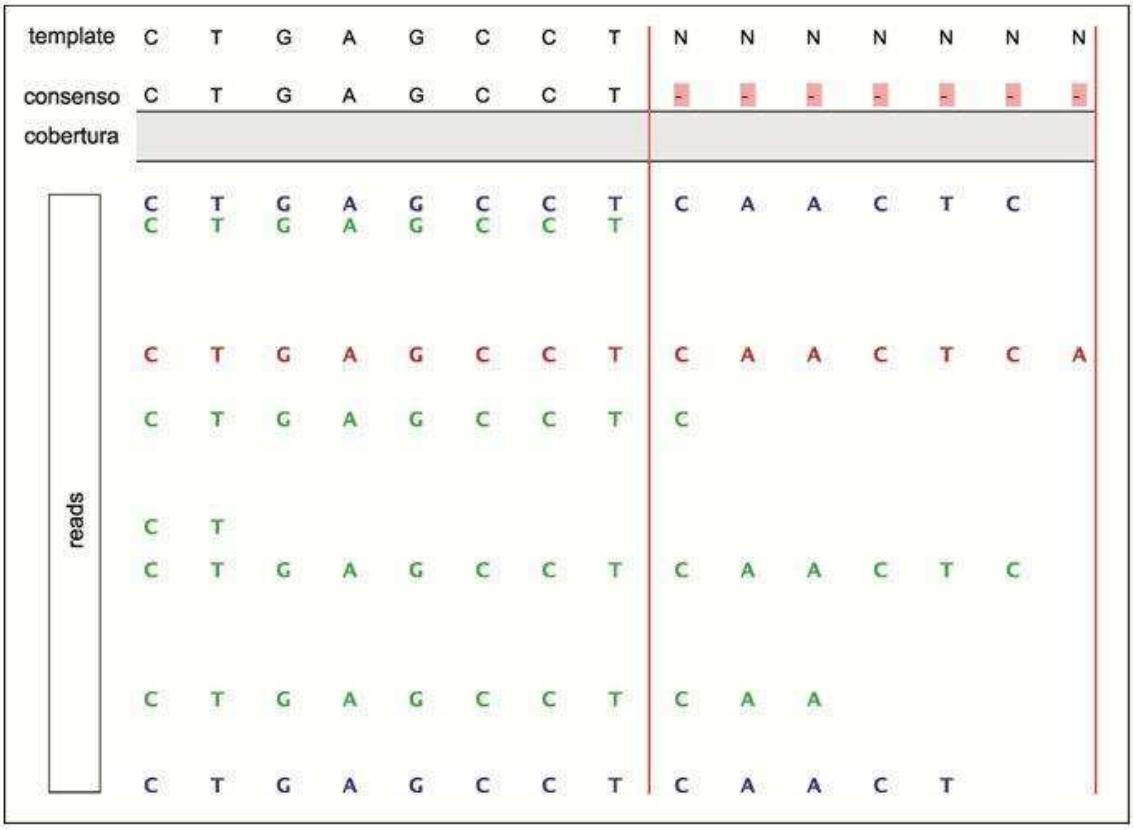

Figura 11. P**ossibilidade de extensão de gaps pelas reads mapeadas.** Os gaps são representados pela base ambígua "N" (lado direito da linha vermelha).

A figura acima mostra apenas seis reads contribuindo com novas informações. No entanto, devido à alta cobertura proporcionada pelos dados analisados, cada sítio deste genoma foi representado mais de 100 vezes, conforme representado na Figura 12.

Por fim, tem-se um exemplo evidenciando a possibilidade de extensão de gaps, baseando-se no mapeamento das reads (oriundas dos dois conjuntos de dados analisados) numa determinada posição do genoma. Neste exemplo, a extremidade 3` do pseudogene está incompleta e as reads mapeadas são capazes de estendê-la.

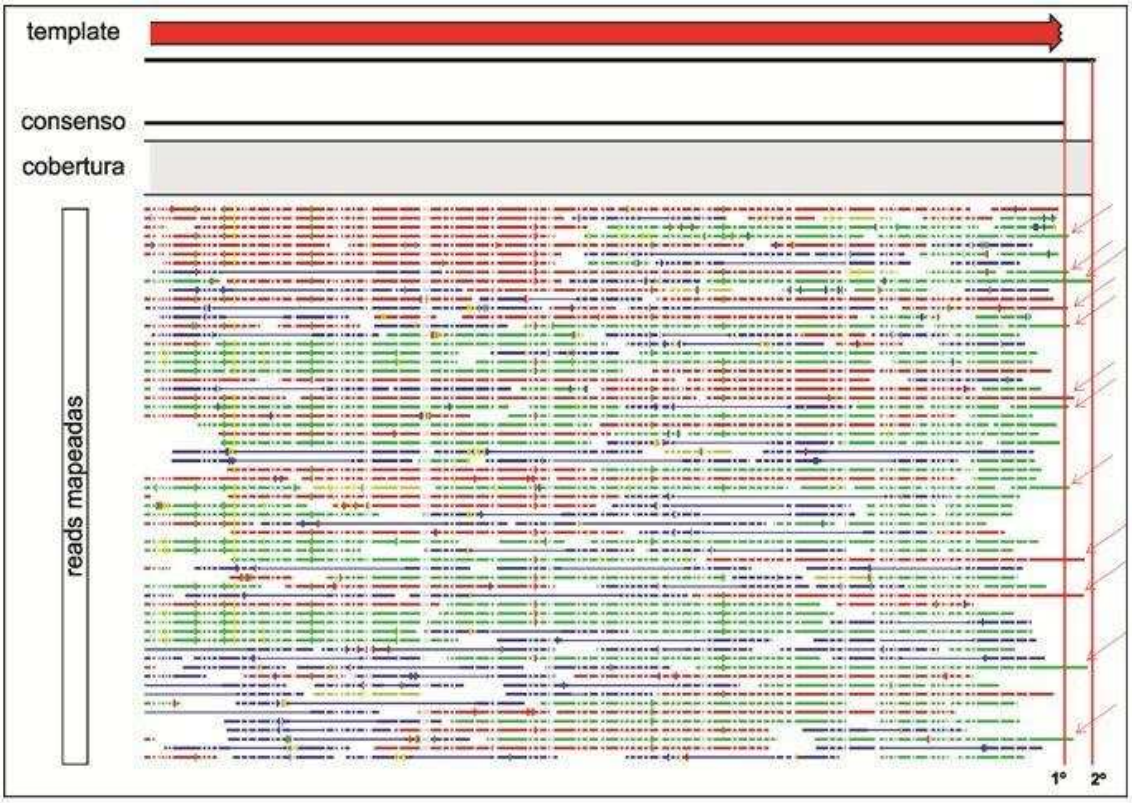

Figura 12. Extensão do pseudogene com base nas reads mapeadas. O mapeamento das reads dos dois conjuntos de dados evidencia a extensão do pseudogene (representado pela seta vermelha larga no topo da figura), visto que várias reads transcendem seu final, conforme pode ser visto pelas setas no espaço entre as linhas vermelhas. A cor azul representam as reads do SRA paired end, a cor verde representa as reads obtidas pelo 454 mapeadas na direção foward (direta) e a cor vermelha, as mapeadas da direção reverse (reversa).

Identificação de conflitos

Os quatro mapeamentos foram utilizados para a analisar posições do genoma com identidades diferentes das reads mapeadas (conflitos). Foram utilizadas três abordagens: primeiro efetuou-se uma análise dos conflitos comparando o genoma com as reads mapeadas provenientes da plataforma 454. Posteriormente, efetuou-se uma análise com as reads mapeadas proveniente do banco de dados SRA. Por fim, foram identificados os conflitos comuns aos dois conjuntos de dados.

Foram identificados 38.688 conflitos pela primeira abordagem e 17.629 pela segunda abordagem. O menor número de conflitos encontrados na abordagem 2 se deve, provavelmente, as reads mapeadas serem provenientes do mesmo isolado de L. (V.) braziliensis utilizado no sequenciamento genômico que é o MHOM/BR/75/M2904. Os conflitos comuns às duas abordagens perfazem um total de 6.293 sendo os 28 primeiros mostrados abaixo (Tabela 15).

Tabela 15. Conflitos comuns aos dois conjuntos de dados (454 e SRA).

Posição da Referência	454			SRA		
	Resíduo consenso	Comparação do conflito	IUPAC	Resíduo consenso	Comparação do conflito	IUPAC
3315	C	Referência: A. Reads: C	C	N	Referência: A. Reads: M	M
3602	N	Referência: G. Reads: R	R	N	Referência: G. Reads: R	R
18601	N	Referência: T. Reads: Y	Y	N	Referência: T. Reads: Y	Y
20183	N	Referência: C. Reads: Y	Y	N	Referência: C. Reads: Y	Y
20364	N	Referência: G. Reads: K	K	N	Referência: G. Reads: K	K
26328	N	Referência: C. Reads: Y	Y	N	Referência: C. Reads: Y	Y
26607	N	Referência: C. Reads: M	M	N	Referência: C. Reads: M	M
32621	G	Referência: A. Reads: G	G	N	Referência: A. Reads: R	R
33378	C	Referência: T. Reads: C	C	N	Referência: T. Reads: Y	Y
33412	T	Referência: G. Reads: T	T	N	Referência: G. Reads: K	K
36367	G	Referência: A. Reads: G	G	N	Referência: A. Reads: R	R
47870	N	Referência: G. Reads: R	R	N	Referência: G. Reads: R	R
47906	C	Referência: A. Reads: C	C	N	Referência: A. Reads: M	M
53045	N	Referência: A. Reads: R	R	N	Referência: A. Reads: R	R
53170	N	Referência: C. Reads: Y	Y	N	Referência: C. Reads: Y	Y
53288	N	Referência: G. Reads: R	R	N	Referência: G. Reads: R	R
63657	N	Referência: C. Reads: Y	Y	N	Referência: C. Reads: Y	Y
64046	N	Referência: C. Reads: M	M	N	Referência: C. Reads: M	M
77171	N	Referência: G. Reads: S	S	N	Referência: G. Reads: S	S
90400	N	Referência: A. Reads: R	R	N	Referência: A. Reads: R	R
90703	N	Referência: C. Reads: S	S	N	Referência: C. Reads: S	S
130742	C	Referência: T. Reads: C	C	N	Referência: T. Reads: Y	Y
151999	T	Referência: C. Reads: T	T	N	Referência: C. Reads: Y	Y
154582	G	Referência: A. Reads: G	G	N	Referência: A. Reads: R	R
155179	C	Referência: A. Reads: C	C	N	Referência: A. Reads: M	M
155256	C	Referência: T. Reads: C	C	N	Referência: T. Reads: Y	Y
158006	C	Referência: T. Reads: C	C	N	Referência: T. Reads: Y	Y
158107	T	Referência: C. Reads: T	T	N	Referência: C. Reads: Y	Y

***R** = G ou A (purina), **Y** = T ou C (pirimidina), **K** = G ou T (ceto), **M** = A ou C (amino).

A utilização destas informações para correção de erros no genoma de L. (V.) braziliensis é de fundamental importância por enriquecer e aumentar a confiabilidade de inúmeras análises futuras, como: descoberta de genes (Mondego et al., 2011), melhor caracterização de candidatos a polimorfismos de nucleotídeo único (SNPs) (Useche et al., 2001), identificação de marcadores moleculares específicos (Romanuik et al., 2009), complementação de anotações de genomas, dentre outras.

Novas ORFs em L. (V.) braziliensis

Os contigs montados que não mapearam no genoma de L. (V.) braziliensis foram utilizados para predição de ORFs, através de um script em perl. Foram obtidas 325 possíveis

novas ORFs, sendo estas utilizadas como query pelo blastx (configurado com: código genético padrão, valor de e-value 1 x 10^{-10}, seed igual a 7, opção de filtragem de baixa complexidade desabilitada e matriz BLOSUM 80) para busca no banco de dados nr (contendo sequências proteicas não redundantes) formatado localmente.

Após análise, foram encontradas seis proteínas que não estão representadas no orfeoma anotado desta espécie (Tabela 16).

Tabela 16. Proteínas novas para L. (V.) braziliensis

ESPÉCIE	FUNÇÃO	LOCUS	ESTATUS	START CONDO	STOP CÓDON	ID NCBI
L. infantum JPCM5	Hipotética conservada	LINJ_18_0850	CDS completa	ATG	TAG	XM_001464863.2
L. donovani	Like-metaloprotease	LDBPK_040820	CDS completa	ATG	TGA	XM_003858096.1
L. donovani	Hipotética conservada	LDBPK_270560	CDS completa	ATG	TAG	XM_003861864.1
Leishmania mexicana	Hipotética conservada	LMXM_02_0320	CDS completa	ATG	TGA	XM_003871582.1
Leishmania mexicana	Hipotética conservada	LMXM_09_0200	CDS completa	ATG	TAG	XM_003872651.1

Para comprovar este achado, as sequências de nucleotídeos destas proteínas foram utilizadas como referência para o mapeamento das reads tratadas dos dois conjuntos de dados (SRA e 454). Como resultado, toda a extensão da query foi coberta, reforçando que os genes que codificam estas proteínas são expressos em L. (V.) braziliensis.

Obteve-se ainda como resultado desta análise 80 outros genes, que embora contenham os locus referentes ao genoma anotado de L. (V.) braziliensis, não são encontrados neste genoma (Tabela 17).

Tabela 17. IDs do NCBI referentes a genes representados por nossos dados e ausente no genoma anotado de L. (V.) braziliensis.

ID NCBI				
XM_001561536.1	XM_001565654.1	XM_001565694.1	XM_003722978.1	XM_001566585.2
XM_001561635.1	XM_001565655.1	XM_001565695.1	XM_003722987.1	XM_001566584.2
XM_001561760.1	XM_001565657.1	XM_001565703.1	XM_003723004.1	XM_001565987.2
XM_001562107.1	XM_001565658.1	XM_001565704.1	XM_003723025.1	XM_001565698.2
XM_001562113.1	XM_001565661.1	XM_001565706.1	XM_003723056.1	XM_001565700.2
XM_001563094.1	XM_001565665.1	XM_001565707.1	XM_003723091.1	XM_001565674.2
XM_001564045.1	XM_001565666.1	XM_001565708.1	XM_003723092.1	XM_001565678.2
XM_001564048.1	XM_001565669.1	XM_001565709.1	XM_003723093.1	XM_001565679.2
XM_001564860.1	XM_001565671.1	XM_001565710.1	XM_001562106.2	XM_001565681.2
XM_001564863.1	XM_001565673.1	XM_001566300.1	XM_001562568.2	XM_001565663.2
XM_001564865.1	XM_001565683.1	XM_001566301.1	XM_001562569.2	XM_001565664.2
XM_001564868.1	XM_001565685.1	XM_001567309.1	XM_001562012.2	XM_001565029.2
XM_001564869.1	XM_001565688.1	XM_001562536.2	XM_001562011.2	XM_001564867.2
XM_001564870.1	XM_001565689.1	XM_001563705.2	XM_001561761.2	XM_001561467.2
XM_001565456.1	XM_001565692.1	XM_001568926.2	XM_001567306.2	XM_001564234.2
XM_001565457.1	XM_001565693.1	XM_003722935.1	XM_001566954.2	XM_001564047.2

Provavelmente ocorreu algum erro durante as atualizações deste genoma no banco de dados.

Após esta análise, ainda restaram 117 possíveis ORFs sem hits no nr, baseando-se nos critérios estabelecidos. Estas podem ser novas ORFs ainda não descritas por não possuírem identidade elevada com a ORFs depositadas neste banco de dados. Estes dados são reforçados pela eficiente montagem efetuada neste trabalho que permitiu encontrar possíveis bugs no genoma, conforme exposto na Tabela 17.

Análise de pseudogenes

Os 188 pseudogenes presentes no genoma de L. (V.) braziliensis foram analisados através do mapeamento combinado. Esta análise permitiu estender a maioria deles, seja na extremidade 5' ou 3'. Na maioria das vezes, o comprimento estendido não foi suficiente para alcançar o start e/ou o stop códon. Esta análise precisa ser melhor discutida em virtude do elevado polimorfismo encontrado, tanto entre os dois conjuntos de dados, como dentro do mesmo conjunto de dados.

Como exemplo do polimorfismo encontrado foi representado o mapeamento no pseudogene LBRM_07_0880 de L. (V.) braziliensis (Figura 13).

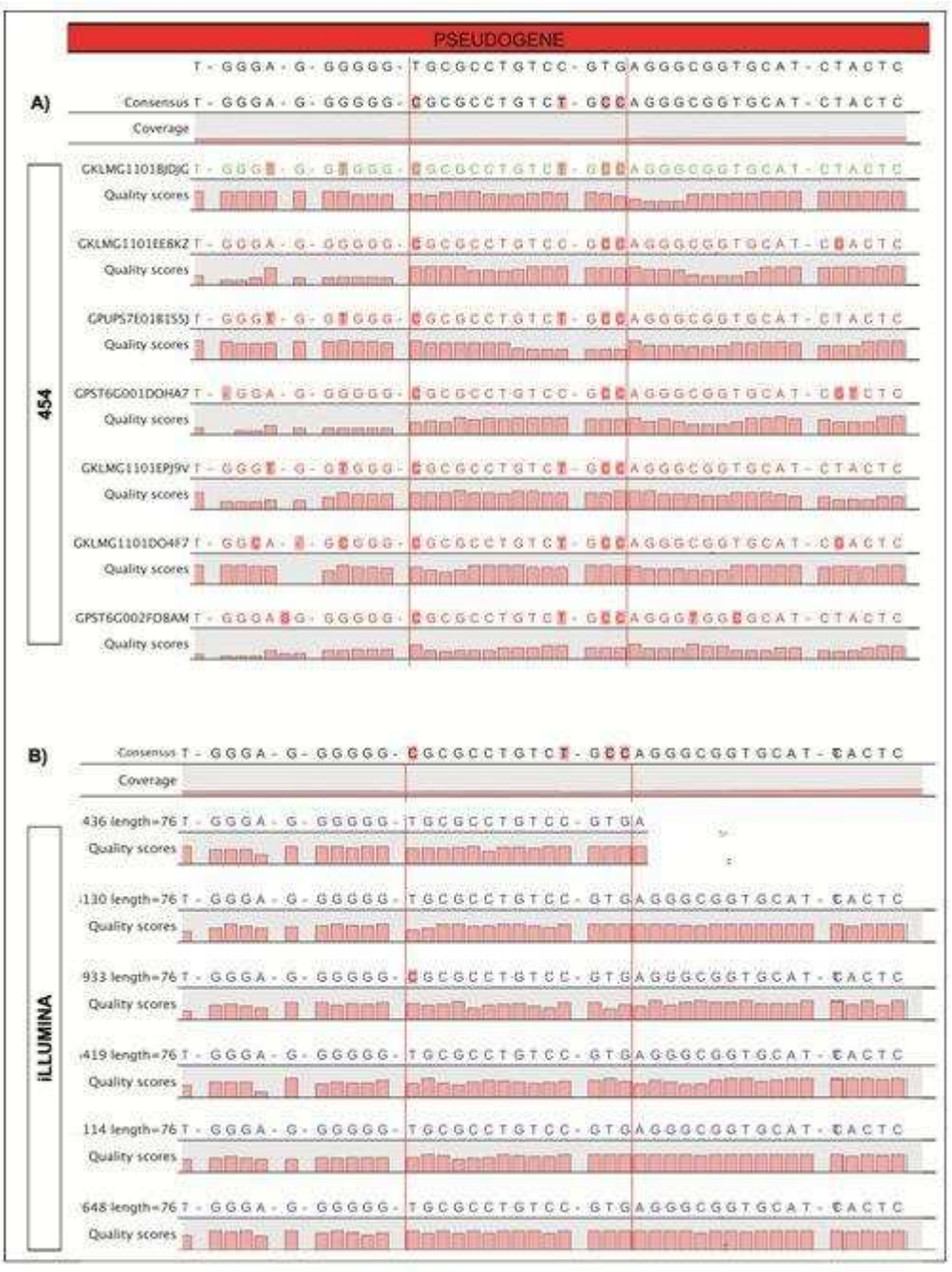

Figura 13. Mapeamento das reads tratadas, dos dois conjuntos de dados, no pseudogene LBRM_07_0880 de L. (V.) braziliensis (representado pelo retângulo vermelho). Em "a", tem-se o mapeamento das reads provenientes do 454/Roche; e em "b", tem-se o mapeamento das reads provenientes da plataforma Illumina.

Neste exemplo, quatro nucleotídeos da sequência consenso (marcados em vermelho) diferem da sequência do pseudogene. Observa-se, ao analisar a figura, que as reads mapeadas em "a" diferem da sequência do pseudogene enquanto as reads provenientes de "b" não interferem na sequência consenso, embora possuem nucleotídeos diferentes do pseudogene. As reads de "a" são provenientes de um isolado diferente do isolado utilizado como template. Desta forma, a variação genética entre os dois isolados pode ser a explicação para tais diferenças. Outra hipótese é que os pseudogenes encontrados no isolado MHOM/BR/75/M2904 podem não ser pseudogenes no isolado sequenciado neste trabalho e vice-versa. Análises futuras são necessárias para se chegar a uma conclusão.

Extensão de ORFs incompletas

Como esperado, devido à fragmentação do genoma de L. (V.) braziliensis, 140 das 7.809 ORFs anotadas estão incompletas (Tabela 18). O cromossomo 31 tem o maior número destas ORFs (12) enquanto nenhuma ORF incompleta foi encontrada para os cromossomos 6, 13 e 23.

O mapeamento combinado e um script em perl, foram utilizados para extensão destas ORFs. Como resultado, 125 delas tiveram suas extremidades aumentadas (89,89%). Em 21 destas foi possível encontrar o códon de início ou parada da tradução (16,08%).

Tabela 18. Locus das ORFs incompletas.

IDs das ORFs incompletas			
LBRM_01_0260	LBRM_14_1320_ENOL	LBRM_21_1980	LBRM_31_0900
LBRM_01_0550	LBRM_15_0460	LBRM_22_0440	LBRM_31_1130
LBRM_01_0640	LBRM_15_0500	LBRM_22_0580	LBRM_31_1270
LBRM_02_0040	LBRM_15_0730	LBRM_22_0620	LBRM_31_1850
LBRM_02_0290	LBRM_15_1280	LBRM_22_0830	LBRM_31_2070
LBRM_02_0500	LBRM_16_0310	LBRM_24_0450	LBRM_31_2210
LBRM_03_0500	LBRM_16_0460	LBRM_24_0470	LBRM_31_2230
LBRM_04_0210	LBRM_16_0780	LBRM_24_0560	LBRM_31_2500
LBRM_04_0260	LBRM_16_1520	LBRM_24_1160	LBRM_31_2630
LBRM_04_0680	LBRM_17_0110	LBRM_25_0340	LBRM_31_2860
LBRM_04_0790	LBRM_17_0120	LBRM_25_1010	LBRM_31_3400
LBRM_05_0390	LBRM_17_0230_YSB	LBRM_25_1290	LBRM_31_3410
LBRM_07_0520	LBRM_17_0470	LBRM_25_1380	LBRM_32_1230
LBRM_07_0920	LBRM_17_0490	LBRM_26_0370	LBRM_32_2860
LBRM_07_0920_B	LBRM_17_0820	LBRM_26_1250	LBRM_32_2930
LBRM_08_0230	LBRM_17_0990	LBRM_26_1380	LBRM_32_3410
LBRM_08_0380	LBRM_17_1560	LBRM_26_1700	LBRM_33_2770
LBRM_08_0550	LBRM_18_1000	LBRM_26_1860	LBRM_33_3050
LBRM_09_0200	LBRM_18_1110	LBRM_27_0260	LBRM_33_3350
LBRM_09_0410	LBRM_19_0680	LBRM_27_0620	LBRM_33_3410
LBRM_09_1070	LBRM_19_0880	LBRM_27_0890	LBRM_34_0070
LBRM_10_0390	LBRM_19_1150	LBRM_27_2150	LBRM_34_0520
LBRM_10_0530	LBRM_19_1160	LBRM_27_2810	LBRM_34_2060
LBRM_10_0550_P63-2	LBRM_19_1250	LBRM_28_1200	LBRM_34_2180
LBRM_10_0570_P63-4	LBRM_19_1530	LBRM_28_1580	LBRM_34_2561
LBRM_10_0980	LbrM20_V2.0430	LBRM_28_1870	LBRM_35_0880
LBRM_10_1050	LbrM20_V2.2080	LBRM_28_2970	LBRM_35_1410
LBRM_11_0170	LbrM20_V2.2090	LBRM_28_2980	LBRM_35_3570
LBRM_11_0260	LbrM20_V2.2100	LBRM_29_1830	LBRM_35_4630
LBRM_11_0620	LbrM20_V2.3530	LBRM_29_2260	LBRM_35_5500
LBRM_11_0930	LbrM20_V2.4280	LBRM_30_0990	LBRM_35_5740
LBRM_11_1170	LBRM_21_0180	LBRM_30_1491	LBRM_35_6480
LBRM_12_0420	LBRM_21_0270	LBRM_30_2500	LBRM_35_6491
LBRM_12_0730	LBRM_21_0940	LBRM_30_3640	LBRM_35_6940
LBRM_14_1310	LBRM_21_0970	LBRM_30_3730	LBRM_35_7400

As próximas análises estão relacionadas à expressão gênica.

Panorama geral através do mapeamento das reads

Mapeamento das reads de cada biblioteca no genoma de L. (V.) braziliensis e L. (L.) major

Primeiro mapeou-se as reads de cada biblioteca no genoma anotados de L. (V.) braziliensis. Em seguida, efetuou-se um mapeamento das reads não mapeadas em L. (V.) braziliensis no genoma de L. (L.) major. Desta forma foi possível estimar o universo de reads que não mapeou em nenhum dos dois genomas. O genoma de L. (L.) major foi escolhido para esta análise devido à elevada conservação das sequências de nucleotídeos entre estas duas espécies; e pelo fato do genoma de L. (V.) braziliensis encontrar-se muito fragmentado.

Esta análise foi realizada com o intuito de verificar o percentual de reads de cada biblioteca que não mapeou no genoma de L. (V.) braziliensis, pois estas reads provavelmente não mapearam por dois motivos: (1º) pelo fato do genoma desta espécie se encontrar muito fragmentado e (2º) por uma parte destas reads pertencerem a uma das seguintes categorias: novos genes, pseudogenes e quimeras (fragmentos de genes diferentes presentes no mesmo contig, que ocorrem devido a erros causados pela técnica ou devido a um processo biológico novo). Os resultados para cada amostra estão descritos em seguida.

Mapeamento das reads da biblioteca ET-MET

Das 844.679 reads de ET-MET, 657.422 (77,83%) mapearam no genoma de L. (V.) braziliensis, sendo 375.494 destas mapeadas no orfeoma de L. (V.) braziliensis (57,12% das mapeadas no genoma (Tabela 19).

Tabela 19. Mapeamento das reads da biblioteca 1 no genoma de L. (V.) braziliensis**.**

CATEGORIA	NO GENOMA* (%)	DO GENOMA QUE MAPEARAM NO ORFEOMA* (%)
L. (V.) braziliensis	657.422 (77,83)	375.494 (57,12)
L. (V.) major	118.066 (13,98)	------
Não mapearam	69.191 (8,19)	------
Total	**844.679 (100)**	------

*número de reads mapeadas

Das 187.257 reads restantes, 118.066 mapearam no genoma de L. (L.) major (13,98% do total geral). Assim, restaram 69.191 reads (8,19 % do total geral) que não mapearam no genoma de nenhuma das duas espécies (Tabela 19).

Mapeamento das reads da biblioteca ET-PRO

Das 560.877 reads de ET-PRO, 368.237 (65,65%) mapearam no genoma de L. (V.) braziliensis, sendo 146.885 destas mapeadas no orfeoma de L. (V.) braziliensis (39,89% das mapeadas no genoma) (Tabela 20).

Tabela 20. Mapeamento das reads da biblioteca 2 no genoma de L. (V.) braziliensis.

CATEGORIA	NO GENOMA* (%)	DO GENOMA QUE MAPEARAM NO ORFEOMA* (%)
L. (V.) braziliensis	368.237 (65,65)	146.885 (39,89)
L. (V.) major	146.419 (26,11)	------
Não mapearam	46.221 (8,24)	------
Total	560.877 (100)	------

*número de reads mapeadas

Das 192.640 reads restantes, 146.419 mapearam no genoma de L. (L.) major (26,11% do total geral). Assim, restaram 46.221 reads (8,24 % do total geral) que não mapearam no genoma de nenhuma das duas espécies (Tabela 20).

Mapeamento das reads da biblioteca NSL-MET

Das 1.018.587 reads de NSL-MET, 638.648 (62,70%) mapearam no genoma de L. (V.) braziliensis, sendo 267.994 destas mapearam no orfeoma de L. (V.) braziliensis (41,96% das mapeadas no genoma) (Tabela 21).

Tabela 21. Mapeamento das reads da biblioteca 3 no genoma de L. (V.) braziliensis**.**

CATEGORIA	NO GENOMA* (%)	DO GENOMA QUE MAPEARAM NO ORFEOMA* (%)
L. (V.) braziliensis	638.648 (62,70)	267.994 (41,96)
L. (V.) major	297,712 (29,23)	------
Não mapearam	82.227 (8,07)	------
Total	1.018.587 (100)	------

*número de reads mapeadas

Das 379.939 reads restantes, 297.712 mapearam no genoma de L. (L.) major (29,23% do total geral). Assim, restaram 82.227 reads (8,07% do total geral) que não mapearam no genoma de nenhuma das duas espécies (Tabela 21).

Mapeamento das reads da biblioteca NSL-PRO

Das 475.087 reads de PRO-NSL, 328.237 (69,09%) mapearam no genoma de L. (V.) braziliensis, sendo 132.469 destas mapeadas no orfeoma de L. (V.) braziliensis (40,36% das mapeadas no genoma) (Tabela 22).

Tabela 22. Mapeamento das reads do tratamento 4 no genoma de L. (V.) braziliensis**.**

CATEGORIA	NO GENOMA* / (%)	DO GENOMA QUE MAPEARAM NO ORFEOMA* / (%)
L. (V.) braziliensis	328.237 (69,09)	132.469 / (40,36)
L. (V.) major	103.417 (21,77)	------
Não mapearam	43.433 (9,14)	------
Total	475.087 (100)	------

*número de reads mapeadas

Das 146.850 reads restantes, 103.417 mapearam no genoma de L. (L.) major (21,77% do total geral). Assim, restaram 43.433 reads (9,14% do total geral) que não mapearam no genoma de nenhuma das duas espécies (Tabela 22).

Aquelas reads que mapearam no genoma de L. (V.) braziliensis e não mapearam no orfeoma podem ser regiões codificadoras deste genoma não preditas como ORFs. O

percentual de reads que enquadram nesta categoria é 42,88, 60,11, 58,04 e 59,64 para as bibliotecas 1, 2, 3 e 4, respectivamente. Este percentual variou de 42,88 para ET-MET à 60,11 para ET-PRO. Este alto percentual pode ainda ser devido ao peculiar processo de transcrição gênica destes organismos eucariotos, no qual uma região não codificadora do genoma é transcrita junto com a região codificadora (transcrição gênica policistrônica).

Com exceção de ET-MET, o percentual de reads mapeados em todas as categorias foi, relativamente, próximo. A comparação entre as categorias analisadas encontra-se abaixo (Tabela 23).

Tabela 23. Comparação do número de reads mapeadas entre as categorias analisadas, em porcentagem.

TRATAMENTO	GENOMA L. (V.) BRAZILIENSIS	ORFEOMA L. (V.) BRAZILIENSIS	GENOMA L. (L.) MAJOR	NÃO MAPEADAS
ET-MET	77,83	57,12	13,98	8,19
ET-PRO	65,65	39,89	26,11	8,24
NSL-MET	62,70	41,96	29,23	8,07
NSL-PRO	69,09	40,36	21,77	9,14

As reads que não mapearam em nenhuma das referências devem ser melhor analisadas, pois podem ser novos genes ou transcritos ainda não preditos devido ao genoma desta espécie se encontrar muito fragmentado. Uma parte destas reads podem ainda ser quimeras (fragmentos de genes diferentes fusionados compondo o mesmo contig, que podem ocorrer devido a erros causados pela técnica de sequenciamento ou devido a um processo biológico novo). A variação do percentual de reads, desta categoria, considerando todas as amostras foram 1,07%.

Após estas análises, observa-se que a maioria das reads, das quatro bibliotecas, foram mapeadas no genoma de L. (V.) braziliensis e L. (L.) major, sendo este um indício de que nossos dados representam bem a espécie estudada.

Análise qualitativa da expressão gênica entre bibliotecas

Mapeamento dos singlets e contigs de cada biblioteca no conjunto de transcritos de L. (V.) braziliensis

Foi efetuado, após a montagem do transcriptoma das quatro bibliotecas separadamente, um mapeamento das singlets e contigs maiores que 100 pb de cada biblioteca no transcriptoma de L. (V.) braziliensis (conjunto de ORFs mais RNA não codificadores).

Foram mapeadas 8.321 (57,94%), 7.401 (56,37%), 8.041 (53,97%) e 6.400 (55,97%) sequências (singlets mais contigs) das bibliotecas 1, 2, 3 e 4, respectivamente; sendo representados 6.714 (78,94%), 6.019 (70,77%), 6.381 (75,03%) e 5.719 (67,24%) dos transcritos anotados desta espécie (Tabela 24). Estes dados sugere uma representação diferencial do transcriptoma entre as bibliotecas. Além disto, sugere-se que muitas das sequências obtidas neste trabalho não estão representadas no transcriptoma predito; provavelmente devido ao genoma desta espécie estar fragmentado e a baixa cobertura quando se trabalha com as bibliotecas separadamente.

Tabela 24. Mapeamento das singlets e contigs de cada biblioteca nos transcritos anotados de L. (V.) braziliensis.

Amostra	Sequências obtidas após montagem*	Sequências mapeadas*	Percentual de sequências mapeadas	Transcritos anotado de L. (V.) braziliensis	Percentual de transcritos representados
ET-MET	14.362	8.321	57,94%	8.505	78,94%
ET-PRO	13.145	7.401	56,37%	8.505	70,77%
NSL-MET	14.899	8.041	53,97%	8.505	75,03%
NSL-PRO	11.434	6.400	55,97%	8.505	67,24%

*Estes valores correspondem a soma do número de singlets e contigs.

Como referência foi utilizado o conjunto de 8.505 transcritos de L. (V.) braziliensis depositados no banco de dados TriTrypDB (versão 6.0), o qual reuni informações genômicas de espécies patogênicas da família Trypanosomatidae (www.tritrypdb.org) (Laurentino et al., 2004).

Obteve-se, ao juntar as sequências das quatro bibliotecas, 90,71% de representação dos transcritos utilizados como referência para o mapeamento. Como hipótese, pode-se dizer que houve uma variação de 11,77 à 23,47% de representatividade se considerarmos a

possibilidade de presença de no máximo 90,71% dos transcritos (desconsiderando as novas ORFs presentes em nossos dados e ausentes no genoma); ou seja na amostra com maior representatividade (ET-MET) 11,77% dos transcritos preditos do genoma anotado não estão presentes no transcriptoma biológico desta amostra. Já na amostra de menor representatividade (NSL-PRO) este percentual foi de 23,47. A menor representatividade das amostras de formas procíclicas pode estar diretamente relacionada com a menor quantidade de dados gerada nestas bibliotecas, visto que foram sequenciadas em meia placa.

Assim como mencionado anteriormente, os transcritos utilizados como referência para mapeamento poderiam ter sido melhor representados pelas sequências provenientes das bibliotecas sequenciadas neste trabalho se o genoma de L. (V.) braziliensis anotado não estivesse tão fragmentado.

ORFs exclusivas de cada bibliotecas

Um banco de dados relacional foi criado com a finalidade de identificar quais das ORFs anotadas automaticamente foram representadas pelas sequencias resultantes da montagem de cada biblioteca. Cada uma das quatro tabelas deste banco contém a lista das ORFs anotadas automaticamente que foram representadas pelas sequências (singlets ou contigs) montadas neste trabalho. Foram representadas pelas bibliotecas ET-MET, ET-PRO, NSL-MET e NSL-PRO, 6.714, 6.019, 6.381 e 5.719 ORFs respectivamente. Esta análise foi efetuada comparando-se as tabelas em busca de ORFs exclusivas e comuns pela utilização da sintaxe do próprio banco de dados.

Primeiro, efetuou-se uma comparação das tabelas referentes a biblioteca 1 (ET-MET) e 3 (NSL-MET). Desta forma, verificou-se que 898 ORFs (10,56% do orfeoma) foram mapeados apenas por sequências (ao menos uma singlet ou contig) de ET-MET. Por outro lado, 566 ORFs (6,54% do orfeoma) foram mapeados apenas por sequências (ao menos uma singlet ou contig) de NSL-MET. Esta análise permitiu identificar o universo de ORFs exclusivas das formas infectivas do isolado virulento (NSL-MET) e exclusivos da forma infectiva do isolado não virulento (ET-MET). Uma análise mais aprofundada destas ORFs exclusivas pode auxiliar no entendimento do fenômeno diferencial de virulência nesta espécie. Um total de 5.812 ORFs (68,34% do orfeoma) foram mapeadas por pelo menos uma sequência (singlet ou contig) das formas metacíclicas; sendo estas ORFs representadas pela forma metacíclica de uma forma geral (Figura 14).

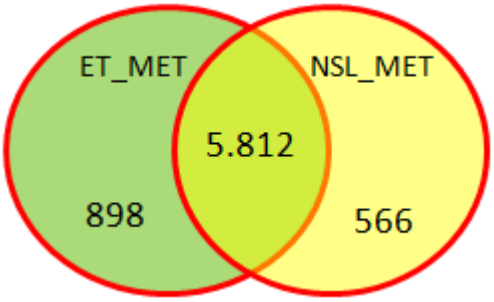

Figura 14. ORFs exclusivas e comuns às formas metacíclicas dos isolados ET e NSL.

Ainda é necessário verificar se estas 898 e 566 ORFs, exclusivas por esta análise, estão presentes nas bibliotecas de procíclicas. Esta informação permitirá categorizá-las como exclusivas ou não das formas metacíclicas.

Posteriormente, efetuou-se uma comparação das tabelas referentes a biblioteca 2 (ET-PRO) e 4 (NSL-PRO). Verificou-se que 1.238 ORFs anotadas (14,56% do orfeoma) foram mapeadas apenas por sequências (ao menos uma singlet ou contig) de ET-PRO. Por outro lado, 937 ORFs anotadas (11,02% do orfeoma) foram mapeados apenas por sequências (ao menos uma singlet ou contig) de NSL-PRO. Esta análise permitiu identificar o universo de ORFs exclusivas da forma não infectiva do isolado não virulento (ET-PRO) e exclusivas da forma não infectiva do isolado virulento (NSL-PRO). Um total de 4.781 ORFs (56,21% do orfeoma) tiveram mapeadas pelo menos uma sequência (singlet ou contig) das bibliotecas analisadas; sendo estas ORFs expressas na forma procíclica de forma geral (Figura 15).

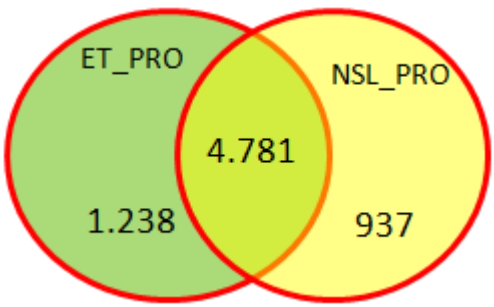

Figura 15. ORFs exclusivas e comuns às formas procíclicas dos isolados ET e NSL.

É necessário verificar se estas 1.238 e 937 ORFs, possivelmente exclusivas por esta análise, estão presentes nas bibliotecas de metacíclica. Esta informação permitirá categorizá-las como exclusivas ou não das formas procíclicas.

As ORFs anotadas foram classificados como exclusivas ou não de uma biblioteca de acordo com o critério estabelecido que foi o mapeamento de pelo menos uma sequência (singlet ou contig). Esta foi uma análise geral, sendo necessária a comprovação por outras analises, tal como expressão diferencial.

Expressão Gênica Diferencial em L. (V.) braziliensis

O transcriptoma é o conjunto de transcritos e suas quantidades em um estágio específico do desenvolvimento ou condição fisiológica. A descoberta de novos transcritos, assim como a quantificação destes é fundamental para entender os fenômenos biológicos que acontecem no interior das células. O sequenciamento do RNA tem permitido mensurações mais precisas do nível destes transcritos; existindo atualmente, diversos pacotes estatísticos destinados a análise de expressão gênica diferencial (Wang, Gerstein e Snyder, 2009).

É de fundamental importância o desenvolvimento de novas pesquisas que busquem novos alvos que possam ser utilizados em tratamento e prevenção da Leishmaniose sendo o conhecimento aprofundado destes alvos uma ferramenta interessante que poderá ser explorada com o objetivo de se desenvolver estratégias para intervir no processo infectivo e auxiliar no controle e erradicação desta infecção.

Como forma de identificar estes alvos, podemos destacar as análises quantitativas de expressão gênica envolvendo Leishmania spp., as quais tendem a se tornarem rotineiras com o desenvolvimentos das tecnologias de sequenciamento paralelo maciço de segunda geração; associado ao término dos projetos de sequenciamento dos genomas de L. (L.) major por Ivens et al. (2005), L. (V.) braziliensis por Laurentino et al. (2004), L. (L.) infantum, L. (L.) mexicana e L. (L.) amazonensis; devido a facilidade desta análise quando se tem o genoma sequenciado. Abaixo tem-se os gráficos de dispersão mostrando os contigs diferencialmente expressos dos quatro contrastes estabelecidos.

Isolado ET

Em vermelho, acima do valor zero no eixo da coordenada, estão os contigs mais expressos em ET-MET. Abaixo, também em vermelho estão os contigs mais expressos em ET-PRO. Foram encontrados nesta análise 866 contigs com expressão diferencial (5,96% do total) (Figura 16).

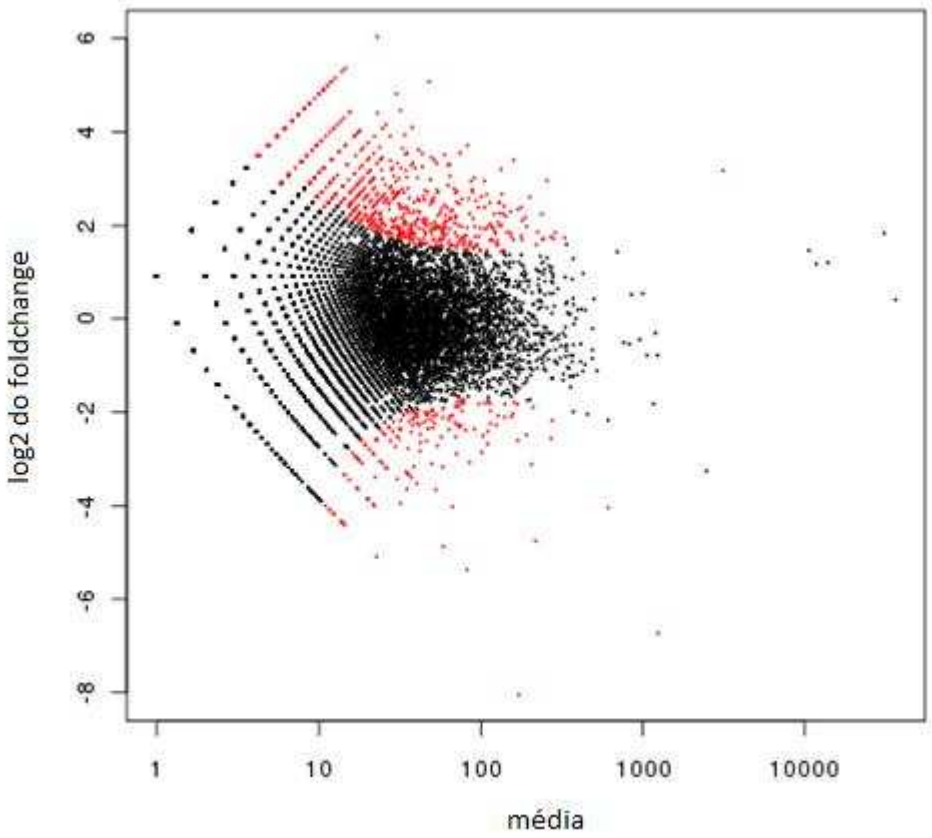

Figura 16. Diagrama de dispersão do log2 do foldchange versus a média evidenciando os contigs diferencialmente expressos. A cor vermelha marca os contigs detectados como diferencialmente expressos, tendo acima do 0 no eixo das coordenadas os mais expressos em ET-MET e abaixo os mais expressos em ET-PRO.

Isolado NSL

Em vermelho, acima do valor zero no eixo da coordenada, estão os contigs mais expressos em NSL-MET. E abaixo, também em vermelho estão os contigs mais expressos em NSL-PRO. Foram encontrados 169 contigs com expressão diferencial entre estes tratamentos (1,16% do total) (Figura 17).

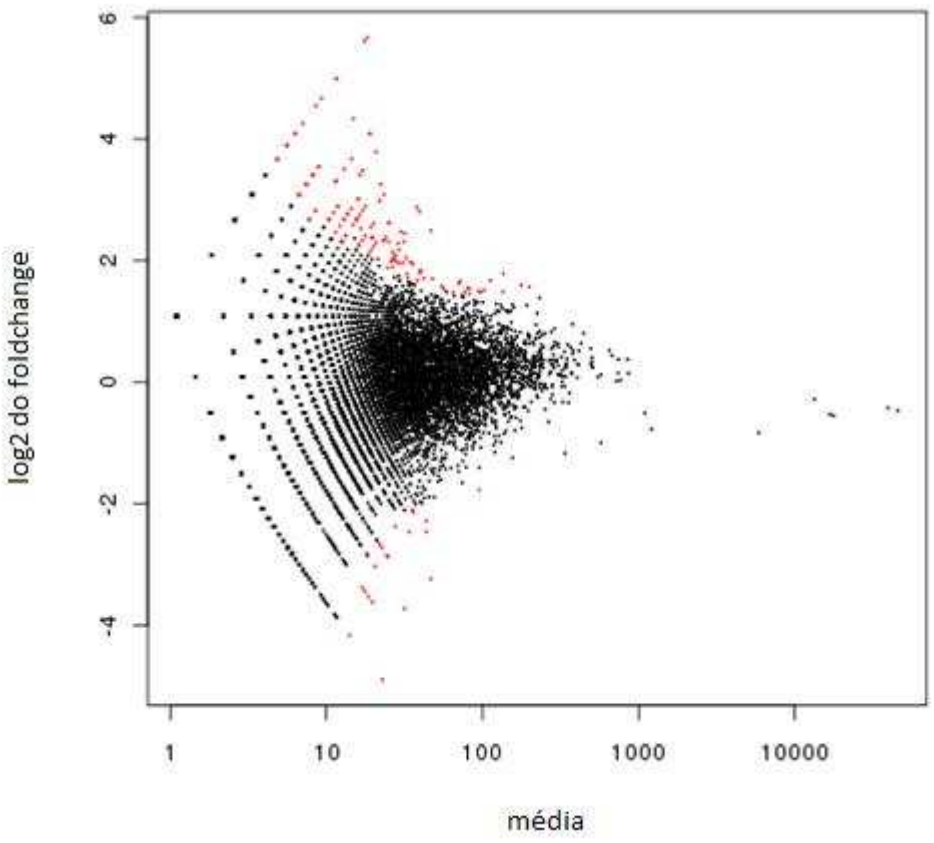

Figura 17. Diagrama de dispersão do log2 do foldchange versus a média evidenciando os contigs diferencialmente expressos. A cor vermelha marca os contigs detectados como diferencialmente expressos, tendo acima do 0 no eixo das coordenadas os mais expressos em NSL-MET e abaixo os mais expressos em NSL-PRO.

Comparação das metacíclicas

Em vermelho, acima do valor zero no eixo da coordenada, estão os contigs mais expressos em ET-MET. E abaixo, também em vermelho estão os contigs mais expressos em NSL-MET. Foram encontrados 587 contigs com expressão diferencial neste contraste (4,04% do total) (Figura 18).

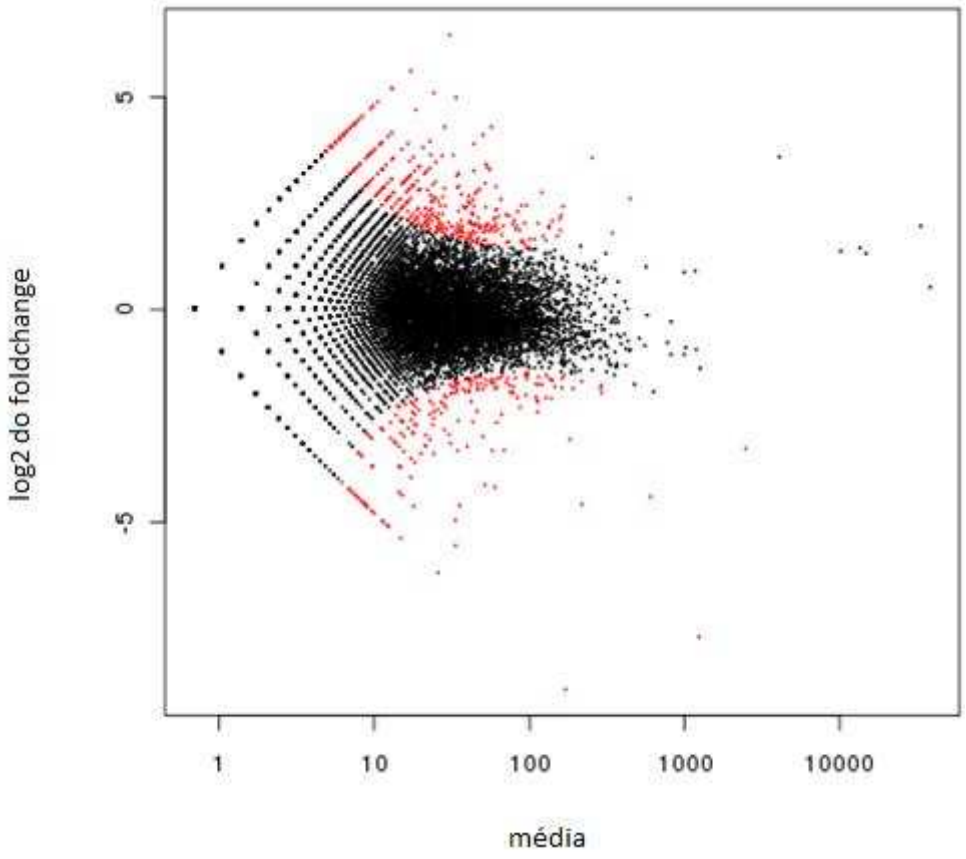

Figura 18. Diagrama de dispersão do log2 do foldchange versus a média evidenciando os contigs diferencialmente expressos. A cor vermelha marca os contigs detectados como diferencialmente expressos, tendo acima do 0 no eixo das coordenadas os mais expressos em ET-MET e abaixo os mais expressos em NSL-MET.

Comparação das procíclicas

Abaixo, tem-se a comparação do nível da expressão gênica entre ET-PRO com NSL-PRO. Em vermelho, acima do valor zero no eixo da coordenada, estão os contigs mais expressos em ET-PRO. E abaixo, também em vermelho estão os contigs mais expressos em NSL-PRO. Foram encontrados 195 contigs com expressão diferencial entre estes tratamentos (1,34% do total) (Figura 19).

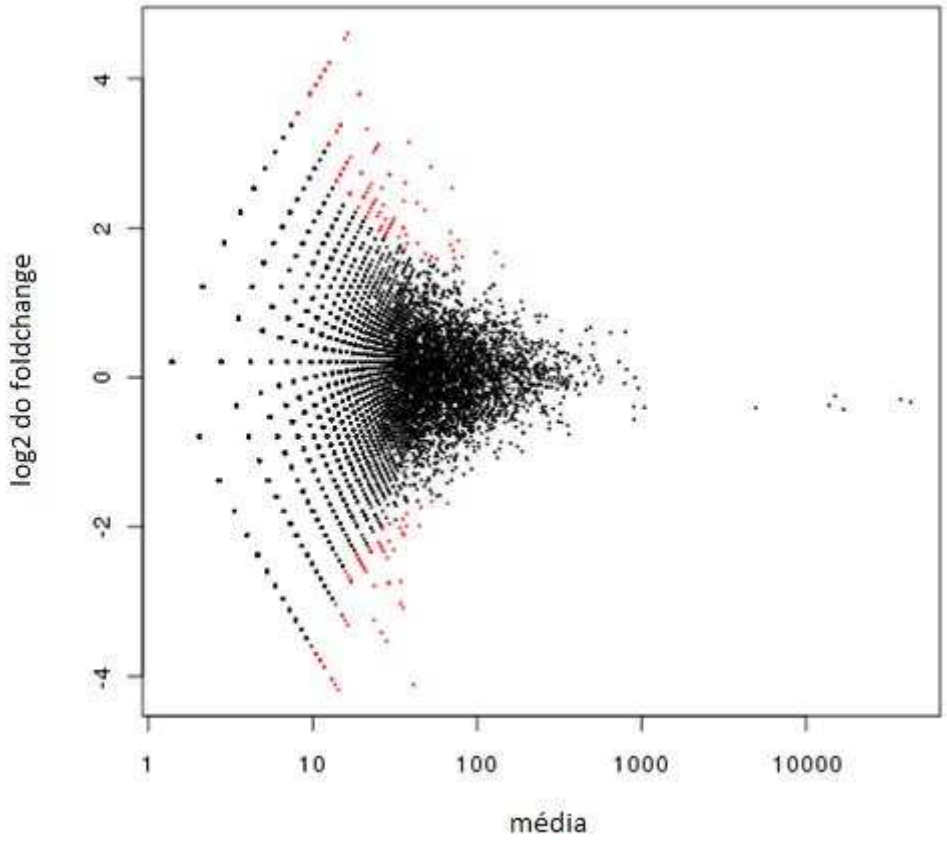

Figura 19. Diagrama de dispersão do log2 do foldchange versus a média evidenciando os contigs diferencialmente expressos. A cor vermelha marca os contigs detectados como diferencialmente expressos, tendo acima do 0 no eixo das coordenadas os mais expressos em ET-PRO e abaixo os mais expressos em NSL-PRO.

A segunda abordagem empregada utilizou o valor de RPKM (**Reads P**er **K**ilobase of exon modelo per **M**illion mapped sequence reads) como normalização dos dados. Este valor é obtido pela divisão do número de reads mapeados numa determinada referência pelo valor resultante da multiplicação do tamanho da referência em kilobase pelo número total de reads, de uma determinada biblioteca, que foram mapeados em todas as referências utilizadas (este último valor é divido por 10^6 para contornar o problema referente ao tamanho da biblioteca).

O valor de RPKM referente ao contig 1, considerando a Tabela 25 é:

Tabela 25. Exemplo de tabela necessária para análise de abundância.

Identificação do contig	Valor de Expressão	Valor de Expressão transformado	Comprimento do contig	reads mapeados	RPKM
Contig 1	65,28	6,03	18.276	848	65,28
Contig 2	64,54	6,01	17.067	783	64,54
Contig 3	32,23	5,01	12.878	295	32,23
Contig 4	22,22	4,47	10.508	166	22,22
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
Contig n	15,08	3,91	10.449	112	15,08
Total				**710833**	

$$RPKM = \frac{848}{(18276/1000)\ \text{x}\ (710833/1000000)} = 65{,}28$$

Foram selecionados para cada abordagem empregada, os 100 contigs com expressão diferencial mais acentuada em cada um dos contrastes estabelecidos.

Visando aumentar a confiabilidade dos resultados, foram selecionados apenas os contigs diferencialmente expressos comuns às duas metodologias empregadas. Desta forma, foram selecionados 40, 20, 34 e 38 contigs diferencialmente expressos comuns aos dois métodos empregados, para os contrastes 1, 2, 3 e 4, respectivamente (Tabelas 26, 27, 28 e 29). Foram eliminados deste conjunto os contigs que discordaram quanto a classificação (classificados como upregulation por um método e downregulation pelo outro), estando estes em negrito nas Tabelas 26, 27, 28 e 29.

Valores de expressão gênica positivos indicam contigs mais expressos em ET-MET e valores negativos, indicam contigs mais expressos em ET-PRO (Tabela 26).

Tabela 26. Valor de expressão gênica dos 40 contigs comuns, dentre os 100 com expressão gênica mais acentuada, identificados pelas duas abordagens utilizadas para o contraste 1 (ET-MET X ET-PRO).

Contig	Expressão gênica*	
	RPKM	**Método de blind**
00534	25,22	4,37
07491	26,26	5,08
07796	16,59	4,61
07783	15,39	4,23
07881	14,78	4,49
08166	12,72	4,91
08329	11,13	4,91
08243	10,38	4,23
08415	10,26	4,61
08609	9,62	5,00
08461	9,51	4,49
08269	9,37	3,91
08750	8,17	4,49
08676	7,62	4,08
08995	7,14	4,49
09224	6,99	4,82
09485	6,92	5,37
09131	6,84	4,61
09338	6,83	4,91
05795	6,54	4,16
09469	6,39	4,91
02828	6,32	5,08
09494	5,65	4,23
10256	5,56	5,30
09571	5,53	4,23
10040	5,49	4,91
06252	5,35	3,82
10208	5,34	5,00
09484	5,29	3,91
06789	5,29	4,43
06707	5,29	4,30
09644	5,23	4,08
09553	5,16	3,91
09883	5,05	4,23
05249	**-8,47**	**4,08**
05976	-9,24	-4,18
05958	-9,7	-4,34
05566	-10,85	-4,00
05646	**-11,49**	**4,08**
05273	-14,41	-4,41

* Valores de expressão gênica positivos indicam contigs mais expressos em ET-MET e valores negativos, indicam contigs mais expressos em ET-PRO.

Valores de expressão gênica positivos indicam contigs mais expressos em NSL-MET e valores negativos, indicam contigs mais expressos em NSL-PRO (Tabela 27).

Tabela 27. Valor de expressão gênica dos 20 contigs comuns, dentre os 100 com expressão gênica mais acentuada identificados pelas duas abordagens utilizadas para o contraste 2 (NSL-MET X NSL-PRO).

Contig	Expressão gênica*	
	Utilizando	**Método de blind**
08547	14,39	5,61
08810	9,08	4,09
09149	7,79	4,26
09221	6,65	3,68
09509	6,2	3,90
09599	5,68	3,68
10018	5,25	3,90
11456	4,68	5,00
11332	4,52	4,68
06974	4,38	3,09
12099	4,25	5,00
12858	4,13	5,68
08135	3,72	3,41
05347	3,59	3,79
08058	3,49	3,09
00729	3,47	3,68
02384	-3,6	-3,72
03392	-4,02	-3,43
03358	-4,2	-3,61
04875	-10,52	-4,89

* Valores de expressão gênica positivos indicam contigs mais expressos em NSL-MET e valores negativos, indicam contigs mais expressos em NSL-PRO.

Valores de expressão gênica positivos indicam contigs mais expressos em ET-MET e valores negativos, indicam contigs mais expressos em NSL-MET (Tabela 28).

Tabela 28. Valor de expressão gênica dos 34 contigs comuns, dentre os 100 com expressão gênica mais acentuada, identificados pelas duas abordagens utilizadas para o contraste 3 (ET-MET X NSL-MET).

Contig	Expressão gênica*	
	RPKM	**Método de blind**
02828	5,12	3,96
07252	29,97	3,84
07491	24,78	4,78
07549	20,09	4,28
07796	18,67	5,20
07783	14,37	3,94
08243	11,54	4,73
08156	10,85	4,12
08242	10,19	4,12
08415	9,88	4,42
08329	8,85	3,84
08461	8,43	3,94
08776	7,81	4,35
08780	7,75	4,35
08750	7,1	3,84
05795	6,17	3,89
09135	6,13	4,03
09224	5,98	4,03
09338	5,96	4,20
09571	5,67	4,35
09469	5,49	4,12
06472	5,19	3,94
09565	5,08	3,84
09727	4,96	4,03
10714	4,92	5,20
09767	4,89	3,94
09983	4,81	4,12
09314	-7,55	-4,36
05646	**-10,74**	**3,84**
05764	**-13,13**	**4,20**
06155	**-20,53**	**4,03**
06443	**-39,36**	**4,12**
06570	**-53,77**	**3,84**
06758	**-379,66**	**4,12**

* Valores de expressão gênica positivos indicam contigs mais expressos em ET-MET e valores negativos, indicam contigs mais expressos em NSL-MET.

Valores de expressão gênica positivos indicam contigs mais expressos em ET-PRO e valores negativos, indicam contigs mais expressos em NSL-PRO (Tabela 29).

Tabela 29. Valor de expressão gênica dos 38 contigs comuns, dentre os 100 com expressão gênica mais acentuada, identificados pelas duas abordagens utilizadas para o contraste 4 (ET-PRO X NSL-PRO).

Contig	Expressão gênica*	
	RPKM	**Método de blind**
04158	77,49	3,53
04795	17,69	3,53
05566	10,43	3,91
05846	10,32	4,53
05764	9,5	3,91
05958	8,41	3,8
06839	6,59	4,21
06563	6,57	3,8
07885	4,9	4,12
07763	4,73	3,8
08038	4,62	4,02
08146	4,33	3,8
08115	4,13	3,53
08135	4,12	3,53
08359	3,89	3,53
09476	3,87	4,6
08427	3,86	3,53
08874	3,83	3,91
01529	3,76	4,12
01334	3,7	3,07
01335	3,66	3,02
08887	3,55	3,53
05010	3,5	3,12
09430	3,41	3,8
09803	3,41	4,12
08216	-3,5	-3,7
08193	-3,58	-3,79
07969	-3,64	-3,7
02068	-3,65	-3,53
07394	-4,06	-3,7
07342	-4,1	-3,7
06462	-5,27	-4,04
05492	-6,52	-3,79
05234	-7,6	-4,04
04957	-8,03	-3,79
04875	-9,12	-4,18
04617	-9,68	-3,88
03421	**-25,41**	**3,8**

* Valores de expressão gênica positivos indicam contigs mais expressos em ET-PRO e valores negativos, indicam contigs mais expressos em NSL-PRO.

Assim, restaram 38, 20, 28, e 37 contigs diferencialmente expressos comuns aos dois métodos empregados, para os contrastes 1, 2, 3 e 4, respectivamente.

Anotação dos contigs diferencialmente expressos

Verificou-se ao analisar a anotação dos 38 contigs comuns, dentre os 100 com expressão gênica mais acentuada, identificados pelas duas abordagens utilizadas para o contraste 1 (ET-MET X ET-PRO) que a maioria dos contigs permaneceram sem anotação, pois não apresentaram similaridade significativa a nenhuma sequência dos bancos de dados analisados. Dentre os nove anotados, apenas 4 (10,53 %) não foram classificados como proteína hipotética. Encontra-se em negrito os contigs mais expressos em ET-PRO; os demais são mais expressos em ET-MET (Tabela 30).

Apenas dois destes quatro foram identificados ao mapear estas 38 sequências no conjunto de contigs anotados de L. (V.) braziliensis, utilizando 50% de cobertura e 80% de identidade, indicando que dois podem ser novos e não estão preditos no transcriptoma anotado de L. (V.) braziliensis.

Tabela 30. Anotação dos 38 contigs comuns, dentre os 100 com expressão gênica mais acentuada, identificados pelas duas abordagens utilizadas para o contraste 1 (ET-MET X ET-PRO). Encontra-se em negrito os contigs mais expressos em ET-PRO; os demais são mais expressos em ET-MET.

Contig	Descrição da proteína	Comprimento (pb)	E-value	Similaridade média (%)
00534	hipotética	6134	0.0	72.64
07491	-	1229	-	-
07796	-	1154	-	-
07783	-	1156	-	-
07881	-	1131	-	-
08166	-	1062	-	-
08329	-	1016	-	-
08243	-	1041	-	-
08415	-	1008	-	-
08609	-	953	6.00E-14	66.00
08461	Nome do produto desconhecido	988	-	-
08269	hipotética	1031	1.66E-57	69.33
08750	-	925	-	-
08676	-	933	-	-
08995	-	862	-	-
09224	infective insect stage-specific	820	4.65E-11	52.71
09485	-	764	-	-
09131	-	829	-	-
09338	-	799	-	-
05795	-	1705	-	-
09469	-	763	-	-
02828	-	2962		
09494	phosphoprotein phosphatase-like	766	1.33E-	67.2%
10256	-	643	-	-
09571	-	754	-	-
10040	-	675	-	-
06252	-	1569	-	-
10208	-	649	-	-
09484	-	767	-	-
06789	-	1409	-	-
06707	-	1439	-	-
09644	hipotética	740	2.60E-18	96.25
09553	-	753	-	-
09883	-	700	-	-
05976	**pre-mrna cleavage complex ii clp1-like**	**1648**	**0.0**	**61.16**
05958	**-**	**1654**	**-**	**-**
05566	**hipotética**	**1772**	**0.0**	**84.80**
05273	**hipotética**	**1874**	**0.0**	**57.08**

Verificou-se ao analisar a anotação dos 20 contigs comuns, dentre os 100 com expressão gênica mais acentuada, identificados pelas duas abordagens utilizadas para o contraste (NSL-MET X NSL-PRO) que a maioria dos contigs foram anotados (13). Dentre estes, apenas cinco (25,00 %) não foram classificados como proteína hipotética. Encontra-se em negrito os contigs mais expressos em NSL-PRO; os demais são mais expressos em NSL-MET (Tabela 31).

Um dado que chama atenção foi a maior expressão da proteína “infective insect stage-specific” na forma metacíclica do isolado ET. Devido a sua função, esta proteína provavelmente contribui para o estabelecimento da infecção por estes parasitos.

Tabela 31. Anotação dos 20 contigs comuns, dentre os 100 com expressão gênica mais acentuada, identificados pelas duas abordagens utilizadas para o contraste 2 (NSL-MET X NSL-PRO). Encontra-se em negrito os contigs mais expressos em NSL-PRO; os demais são mais expressos em NSL-MET.

Contig	Descrição da proteína	Comprimento (pb)	E-value	Similaridade média (%)
08547	-	965	-	-
08810	-	907	-	-
09149	hipotética	833	6.79E-34	80.75
09221	qc-snare	819	4.38E-14	96.50
09509	hipotética	764	8.44E-67	76.33
09599	-	745	-	-
10018	beta-fructosidase-like	679	1.30E-72	80.50
11456	-	499	-	-
11332	-	510	-	-
06974	hipotética	1358	0.0	80.50
12099	hipotética	441	3.60E-57	85.00
12858	-	364	-	-
08135	aptx_danre ame - aprataxin ame - forkhead-associated domain histidine triad-like protein short=fha-hit	1068	7.31E-158	69.95
05347	hipotética	1848	0.0	66.09
08058	hipotética	1077	2.61E-34	87.64
00729	mitochondrial structure specific endonuclease i (sse-1)	5452	0.0	84.55
02384	**vacuolar transporter chaperone**	**3243**	**0.0**	**70.25**
03392	**hypothetical transmembrane**	**2641**	**0.0**	**76.27**
03358	**-**	**2658**	**-**	**-**
04875	**hipotética**	**2018**	**5.48E-174**	**73.60**

Apenas um destes cinco foi identificado ao mapear estas 20 sequências no conjunto de contigs anotados de L. (V.) braziliensis, utilizando 50% de cobertura e 80% de identidade, indicando que quatro podem ser novos e não estão preditos no transcriptoma anotado de braziliensis.

Verificou-se ao analisar a anotação dos 28 contigs comuns, dentre os 100 com expressão gênica mais acentuada, identificados pelas duas abordagens utilizadas para o contraste 3 (ET-MET X NSL-MET) que a maioria dos contigs permanecerão também sem anotação, pois não apresentaram similaridade com nenhuma sequência dos bancos de dados analisados. Dentre os sete anotados, apenas quatro (14,29%) não foram classificados como

proteína hipotética. Encontra-se em negrito os contigs mais expressos em NSL-MET; os demais são mais expressos em ET-MET (Tabela 32).

Apenas um destes cinco foi identificado ao mapear estas 28 sequências no conjunto de contigs anotados de L. (V.) braziliensis, utilizando 50% de cobertura e 80% de identidade, o que significa que quatro são provavelmente novos contigs ainda não descritos no transcriptoma anotado desta espécie.

Tabela 32. Anotação dos 28 contigs comuns, dentre os 100 com expressão gênica mais acentuada, identificados pelas duas abordagens utilizadas para o contraste 3 (ET-MET X NSL-MET). Encontra-se em negrito os contigs mais expressos em NSL-MET; os demais são mais expressos em ET-MET.

Contig	Descrição da proteína	Tamanho (pb)	E-value	Similaridade média (%)
02828	-	2962	-	-
07252	-	1288	-	-
07491	-	1229	-	-
07549	-	1211	-	-
07796	-	1154	-	-
07783	-	1156	-	-
08243	-	1041	-	-
08156	-	1064	-	-
08242	hipotética	1043	1.75E-17	93.66
08415	-	1008	-	-
08329	-	1016	-	-
08461	Nome do produto desconhecido	988	6.01E-14	66.00
08776	hipotética	917	9.83E-35	75.90
08780	-	914	-	-
08750	-	925	-	-
05795	-	1705	-	-
09135	-	833	-	-
09224	infective insect stage-specific	820	4.665E-11	52.71
09338	-	799	-	-
09571	-	754	-	-
09469	-	763	-	-
06472	frataxin-like	1503	6.14E-57	88.00
09565	-	752	-	-
09727	-	713	-	-
10714	hipotética	576	1.998E-54	72.33
09767	-	716	-	-
09983	ubiquitin hydrolase	684	1.56E-73	66.30
09314	**-**	**803**	**-**	**-**

Verificou-se ao analisar a anotação dos 37 contigs comuns, dentre os 100 com expressão gênica mais acentuada, identificados pelas duas abordagens utilizadas para o contraste (ET-PRO X NSL-PRO) que a maioria dos contigs foram anotados (26). Destes, 12 (32,43%) não foram classificados como proteína hipotética. Encontra-se em negrito os contigs mais expressos em NSL-PRO; os demais são mais expressos em ET-PRO (Tabela 33).

Assim como aconteceu em ET-MET, a expressão da proteína "infective insect stage-specific" foi maior na forma metacíclica do isolado NSL. Esta proteína, provavelmente, contribui para o estabelecimento da infecção por estes parasitos.

Tabela 33. Anotação dos 37 contigs comuns, dentre os 100 com expressão gênica mais acentuada, identificados pelas duas abordagens utilizadas para o contraste 4 (ET-PRO X NSL-PRO). Encontra-se em negrito os contigs mais expressos em NSL-PRO; os demais são mais expressos em ET-PRO.

Contig	Descrição da proteína	Comprimento (pb)	E-value	Similaridade média (%)
04158	hipotética	2297	0.0	75.00
04795	hipotética	2044	0.0	69.25
05566	hipotética	1772	0.0	84.80
05846	hipotética	1686	1.52E-178	85.00
05764	dna repair	1716	0.0	65.30
05958	-	1654	-	-
06839	e3 sumo-protein ligase 2	1397	1.98E-17	37.00
06563	matrix	1468	0.0	60.77
07885	serine threonine phosphatase	1129	3.24E-143	91.30
07763	hipotética	1159	7.36E-10	64.50
08038	nudix hydrolase dihydroneopterin triphosphate pyrophosphohydrolase hydrolase	1087	7.76E-74	79.42
08146	-	1063	-	-
08115	hipotética	1070	0.0	68.00
08135	aptx_danre ame: full=aprataxin ame: full=forkhead-associated domain histidine triad-like protein short=fha-hit	1068	7.31E-158	69.95
08359	dynein-light chain-protein	1003	2.71E-47	73.65
09476	-	770	-	-
08427	hipotética	993	1.28E-149	75.18
08874	-	897	-	-
01529	hipotética	4009	0.0	68.20
01334	5 -3 exonuclease	4253	0.0	63.15
01335	hipotética	4252	0.0	65.55
08887	hipotética	893	4.28E-163	89.00
05010	hipotética	1968	1.49E-108	65.00
09430	-	783	-	-
09803	-	712	-	-
08216	**-**	**1048**	**-**	**-**
08193	**udp-glucoronosyl and udp-glucosyl**	**1054**	**0.0**	**60.00**
07969	**-**	**1105**	**-**	**-**
02068	**phosphoglycan beta arabinosyltransferase**	**3478**	**1.76E-136**	**44.18**
07394	**-**	**1258**	**-**	**-**
07342	**hipotética**	**1273**	**7.50E-81**	**71.60**
06462	**-**	**1504**	**-**	**-**
05492	**hipotética**	**1796**	**1.42E-170**	**60.40**
05234	**solanesyl diphosphate synthase**	**1890**	**0.0**	**73.0**

04957	**3-oxo-5-alpha-steroid 4-dehydrogenase-like**	**1987**	**0.0**	**61.16**
04875	**hipotética**	**2018**	**5.48E-174**	**73.60**
04617	**protein**	**2115**	**0.0**	**62.35**

Apenas quatro destes 12 foram identificados ao mapear estas 37 sequências no conjunto de contigs anotados de L. (V.) braziliensis, utilizando 50% de cobertura e 80% de identidade, indicando que oito podem ser novos e não estão preditos no transcriptoma anotado de L. (V.) braziliensis.

Clusterização dos tratamentos

Uma análise de agrupamento, utilizando distância euclidiana, foi realizada com os quatro tratamentos estabelecidos. Verifica-se, baseando no agrupamento, que as formas promastigotas são as mais próximas, formando um clado com ET-MET. Isoladamente ficou a forma metacíclica da amostra virulenta. Estes resultados sugerem que NSL-MET têm o universo de contigs mais diferenciado entre as amostras estudadas. Estes achados são reforçados pelos resultados encontrados na análise da expressão gênica, pois os contrastes contendo esta amostra foram os que tiveram maior número de contigs diferencialmente expressos. Estes dados também podem ajudar a explicar a virulência maior das formas metacíclicas de NSL, já que as formas metacíclicas de ET parecem ter uma expressão gênica similar as formas procíclicas não infectivas (Figura 20).

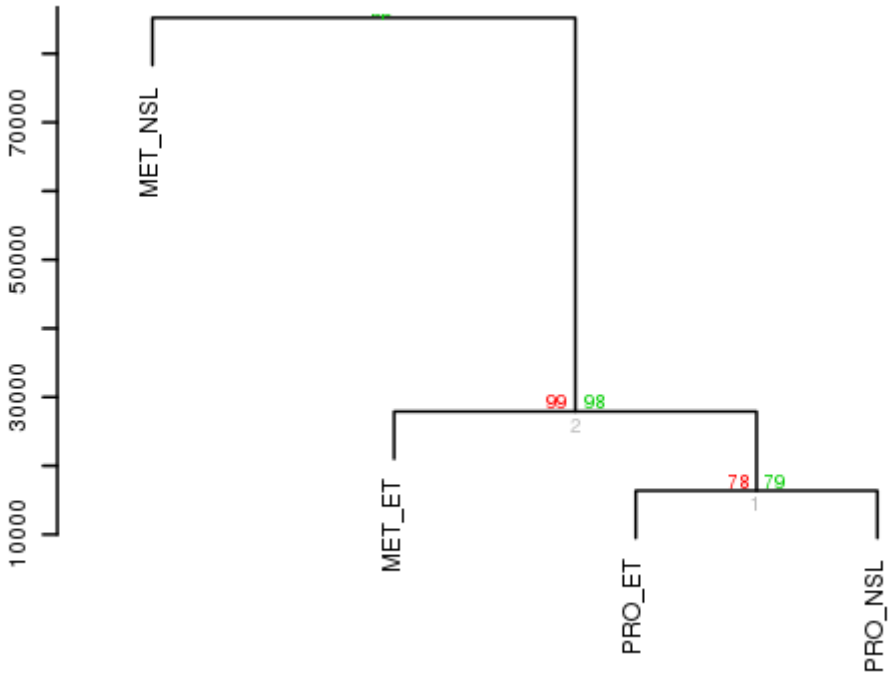

Figura 20. Agrupamento dos quatro tratamentos utilizando distância euclidiana.

CONCLUSÕES

As correções de erros no genoma e orfeoma anotado de L. (V.) braziliensis é de fundamental importância para enriquecer e aumentar a confiabilidade das análises futuras, como: descoberta de genes, correta caracterização de candidatos a polimorfismos de nucleotídeo único (SNPs), identificação de marcadores moleculares específicos, complementação de anotações de genomas, dentre outras. Esta análise é muito importante porque elimina a propagação errônea de conhecimentos, uma vez que as abordagens de anotações de sequências biológicas são baseadas, principalmente, em homologia de sequências.

Os resultados da análise de expressão diferencial serão importantes no desenvolvimento de novas pesquisas que busquem novos alvos que possam ser utilizados para o desenvolvimento de novas abordagens de diagnóstico, tratamento (desenho racional de drogas) e prevenção (desenvolvimentos de vacinas). O conhecimento aprofundado destes alvos é uma ferramenta interessante e será explorado com o objetivo de se desenvolver estratégias para intervir no processo infectivo e auxiliar no controle e erradicação desta infecção.

REFERÊNCIAS BIBLIOGRÁFICAS

ALTSCHUL, S. F. et al. Basic local alignment search tool. **Journal Molecular Biology,** v. 215, n. 3, p. 403-10, 1990.

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

CONESA, A. et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. **Bioinformatics,** v. 21, n. 18, p. 3674-6, 2005.

COORDINATORS, N. R. Database resources of the National Center for Biotechnology Information. **Nucleic Acids Research,** v. 42, n. 1, p. D7-D17, 2014.

DA SILVA, M. S. et al. The Leishmania amazonensis TRF (TTAGGG repeat-binding factor) homologue binds and co-localizes with telomeres. **BMC Microbiology** v. 10, p. 136, 2010.

DESJEUX, P. Leishmaniasis: current situation and new perspectives. **Computatiol Immunology Microbiology And Infectious Diseases,** v. 27, n. 5, p. 305-18, 2004.

DO MONTE-NETO, R. L. et al. Gene Expression Profiling and Molecular Characterization of Antimony Resistance in Leishmania amazonensis. **PLoS Neglected Tropical Diseases,** v. 5, n. 5, p. e1167, 2011.

HUANG, X.; MADAN, A. CAP3: A DNA Sequence Assembly Program. **Genome Research,** v. 9, p. 868-877, 1999.

HYMAN, E. D. A new method of sequencing DNA. **Anal Biochemistry,** v. 174, n. 2, p. 423-36, 1988.

IVENS, A. C. et al. The genome of the kinetoplastid parasite, Leishmania major. **Science,** v. 309, n. 5733, p. 436-42, 2005.

KODAMA, Y. et al. The Sequence Read Archive: explosive growth of sequencing data. **Nucleic Acids Research,** v. 40, n. Database issue, p. D54-6, 2012.

KOVALENKO, D. A. et al. Canine leishmaniosis and its relationship to human visceral leishmaniasis in Eastern Uzbekistan. **Parasites & Vectors,** v. 4, p. 58, 2011.

LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. **Nature Methods**, v. 9, p. 357-359, 2012.

LAURENTINO, E. C. et al. A survey of Leishmania braziliensis genome by shotgun sequencing. **Molecular Biochemistry Parasitology,** v. 137, n. 1, p. 81-6, 2004.

MARDIS, E. R. Next-generation DNA sequencing methods. **Annual Review of Genomics and Human Genetics,** v. 9, p. 387-402, 2008.

MARGULIES, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. **Nature,** v. 437, n. 7057, p. 376-80, 2005.

MONDEGO, J. M. et al. An EST-based analysis identifies new genes and reveals distinctive gene expression features of Coffea arabica and Coffea canephora. **BMC Plant Biology,** v. 11, p. 30, 2011.

NAGARAJ, S. H.; GASSER, R. B.; RANGANATHAN, S. A hitchhiker's guide to expressed sequence tag (EST) analysis. **Brief Bioinformatics,** v. 8, n. 1, p. 6-21, 2007.

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

PERTEA, G. et al. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. **Bioinformatics,** v. 19, n. 5, p. 651-2, 2003.

RAPAPORT, F. et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. **Genome Biology,** v. 14, n. 9, p. R95, 2013.

ROMANUIK, T. L. et al. Novel biomarkers for prostate cancer including noncoding transcripts. **Am Journal Pathology,** v. 175, n. 6, p. 2264-76, 2009.

RONAGHI, M. Pyrosequencing sheds light on DNA sequencing. **Genome Research,** v. 11, n. 1, p. 3-11, 2001.

RUDD, S. Expressed sequence tags: alternative or complement to whole genome sequences? **Trends in Plant Science,** v. 8, n. 7, p. 321-9, 2003.

SANTOS, R. D. F. Padrões de poliadenilação em moléculas de RNAs de Leishmania obtidas da análise de transcriptoma. 2012. 189 f. Tese (Doutorado em Bioquímica Agrícola) - Universidade Federal de Viçosa, Viçosa, MG, 2012.

SCHMIEDER, R.; EDWARDS R. Quality control and preprocessing of metagenomic datasets. **Bioinformatics**, v. 27, p. 863-864, 2011.

SIMON, A.; WOLFGANG, H. Differential expression analysis for sequence count data. **Nature Precedings,** v. Genome Biology, 2010.

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

TIUMAN, T. S. et al. Recent advances in leishmaniasis treatment. **International Journal of Infectious Diseases**, 2011.

USECHE, F. J. et al. High-throughput identification, database storage and analysis of SNPs in EST sequences. **Genome Informatics,** v. 12, p. 194-203, 2001.

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

WILCOX, A. S. et al. Use of 3' untranslated sequences of human cDNAs for rapid chromosome assignment and conversion to STSs: implications for an expression map of the genome. **Nucleic Acids Research,** v. 19, n. 8, p. 1837-43, 1991.