Universidade de São Paulo
Faculdade de Medicina de Ribeirão Preto
Programa Interunidades de Pós-Graduação em Bioinformática

Patrícia de Cássia Ruy

Identificação e análise *in silico* de ncRNAs empregando dados de RNA-seq em *Leishmania*.

Ribeirão Preto – SP
2017

Patrícia de Cássia Ruy

Identificação e análise *in silico* de ncRNAs empregando dados de RNA-seq em *Leishmania*.

Tese apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática, da Universidade de São Paulo, para obtenção do título de Doutor em Ciências.

Área de concentração: Bioinformática

Orientadora: Profa. Dra. Angela Kaysel Cruz

Ribeirão Preto – SP
2017

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Catalogação da Publicação
Serviço de Documentação
Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo

Ruy, Patrícia de Cássia.
Identificação e análise *in silico* de ncRNAs empregando dados de RNA-seq em *Leishmania*. Patrícia de Cássia Ruy ; orientadora Prof[a]. Dra. Angela Kaysel Cruz - Ribeirão Preto, 2017.
145 f.: il.

Tese (Doutorado) – Programa Interunidades de Pós-Graduação em Bioinformática - Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, 2017.

1. RNA não codificador 2. *Leishmania* 3. Transcriptoma 4. Bioinformática 5. ncRNA 6. RNA-seq

RUY, P.C. **Identificação e análise *in silico* de ncRNAs empregando dados de RNA-seq em *Leishmania*.** Tese apresentada à Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, para a obtenção do título de Doutor em Ciências.

Aprovado em:

Banca Examinadora:

Prof. Dr. _____ Instituição:_____
Julgamento: _____ Assinatura: _____

Prof. Dr. _____ Instituição:_____
Julgamento: _____ Assinatura: _____

Prof. Dr. _____ Instituição:_____
Julgamento: _____ Assinatura: _____

Prof. Dr. _____ Instituição:_____
Julgamento: _____ Assinatura: _____

Prof. Dr. _____ Instituição:_____
Julgamento: _____ Assinatura: _____

Dedico este trabalho a minha família.
Amo vocês!

AGRADECIMENTOS

À Profª. Dra. Angela Kaysel Cruz, por ser exemplo de amor a ciência, pela confiança, pelas discussões sempre empolgantes e por orientar meu trabalho.

Ao Dr. Peter Myler do *Seattle Biomedical Reasearch Institute – Seattle/USA* por me receber em seu laboratório e por discussões que ajudaram a definir os caminhos seguidos neste trabalho. Em especial gostaria de agradecer ao Dr. Elton Vasconcelos pelas discussões que contribuíram muito com este trabalho e pelo apoio de sempre.

Ao Sistema Nuvens Interconectadas (interNuvem USP), que disponibilizou a estrutura computacional utilizada no desenvolvimento deste trabalho e que se mostrou robusta e confiável em todos esses anos de utilização.

Ao Laboratório de Computação Científica Avançada - USP, por disponibilizar a estrutura de clusters que foram utilizadas em algumas análises deste trabalho.

Ao Centro de Medicina Genômica, pela parceria no sequenciamento das bibliotecas de RNA e por discussões sobre o projeto. Gostaria de agradecer em especial a Kamila Peroni pela dedicação e empenho em preparar as bibliotecas de RNA.

À Patricia Martorelli e Cristiane de Fátima Braulino, pelo trabalho junto a secretaria de pós-graduação.

Ao Raul, companheiro que escolhi para a vida, obrigada por existir! Agradeço também as inúmeras discussões científicas que auxiliaram o desenvolvimento deste trabalho, com você tudo fica melhor. JUNTOS! Amo você. ♡ ♡ ♡

Ao meu pai, Claudemir e minha mãe, Selma, meus eternos incentivadores e motivadores. Obrigada pelo amor de sempre. Vocês me ensinaram a ser uma pessoa melhor a cada dia. Amo vocês!

À Priscila e Pâmela, minhas muito amadas irmãs, que são prova de que amor não tem limites e sei que com vocês nossa rede de suporte sempre será forte. Amo vocês!

Ao meu avô Francisco e minha avó Maria, por serem exemplos de vida, o amor de vocês foi/é indispensável na minha vida. Obrigada por, de perto ou de longe, sempre olharem por mim. Saudades eternas vó. Amo vocês.

Ao meu avô João e minha avó Angelina, que foram exemplos importantes e que ajudaram a formar quem sou hoje.

À tia Lucéia, tia Cá, Rafael, tio Elias, tio Guta, Geni e Talita, vocês trazem ainda mais amor para minha vida. Exemplos a serem observados! Obrigada pelo apoio de sempre. Amo vocês!

À Suelen, que como prima-irmã sempre escutou meus desabafos, me apoiando no que fosse preciso.

À Carmen e Silvio, que me receberam de braços abertos e aumentaram ainda mais minha família. Agradeço também ao Érico e a Priscila, pelas boas conversas de sempre. Vocês estão em meu coração!

Aos amigos do laboratório: Felipe, Juliana, Natália, Lucas, Leandro e Angela, por todas as discussões científicas e não científicas e por tornarem o ambiente leve e bom de se trabalhar.

À Tânia e Viviane, por sempre cuidarem da gente, mesmo que demos muito trabalho.

À Thais Couto Laureano, minha primeira aluna de iniciação científica, por me ajudar a passar por essa experiência. Aprendemos muito juntas!

Este trabalho foi realizado com o suporte financeiro das seguintes entidades:

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES

Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq (305775/2013-8)

Fundação de Amparo a Pesquisa do Estado de São Paulo – FAPESP (Processo n° 2013/50219-9)

RESUMO

RUY, Patrícia de Cássia. **Identificação e análise *in silico* de ncRNAs empregando dados de RNA-seq em *Leishmania*.** 2017. 145 f. Tese de Doutorado - Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, Ribeirão Preto, 2017.

Análises de dados gerados em larga escala revelaram que os transcriptomas são mais extensos e complexos do que inferido previamente. Hoje é evidente que a maioria dos genomas de eucariotos são quase inteiramente transcritos e sob regulação atrelada a estágios de desenvolvimento. O controle da expressão gênica envolve RNAs não codificadores (ncRNAs) regulatórios, inclusive em processos pós-transcricionais. A regulação de expressão gênica em nível pós-transcricional é crucial em diferentes organismos, mas é particularmente central nos tripanossomatídeos. Os parasitos do gênero *Leishmania* (Ordem *Kinetoplastidae*, família *Trypanosomatidae*) provocam doenças infecto-parasitárias conhecidas como leishmanioses e em seu ciclo de vida apresenta-se sob três formas principais de desenvolvimento: promastigotas procíclicos, promastigotas metacíclicos e os amastigotas. Este trabalho teve como objetivo identificar e caracterizar computacionalmente ncRNAs putativos de *Leishmania* em diferentes estágios de desenvolvimento. A associação de abordagens em larga escala e ferramentas de bioinformática possibilitaram a identificação de 11.376 ncRNAs putativos em *L. braziliensis* e, em um estudo preliminar, de 37 ncRNAs putativos em *L. donovani*. Adicionalmente, o transcriptoma de *L. braziliensis* foi analisado comparativamente entre os três estágios de desenvolvimento do parasito. Em *L. donovani*, dos 37 ncRNAs putativos identificados, 34 estavam em UTRs (*Untranslated Regions*) e 3 em regiões intergênicas. Preditores de características específicas de ncRNAs foram utilizados e 32 ncRNAs putativos tiveram pelo menos uma predição positiva. Todos os candidatos estão conservados inteira ou parcialmente em pelo menos 3 espécies de *Leishmania*. Cinco ncRNAs de *L. donovani* foram confirmados por experimentos de *Northern blotting*. A análise do transcriptoma de *L. braziliensis* revelou uma diferença de expressão gênica entre os estágios de desenvolvimento que variou entre 52% e 71%, dependendo dos estágios comparados. Também foram definidos os limites das 5'UTRs e 3'UTRs de 81% e 38% das CDSs anotadas, respectivamente. Propusemos uma metodologia para identificação de ncRNAs

putativos utilizando dados de sequenciamento de RNA-total. Essa metodologia identificou 11.376 ncRNAs putativos em *L. braziliensis*, sendo que todos os candidatos foram analisados por programas preditores de características específicas de ncRNAs e apresentaram pelo menos uma predição positiva, além de não possuírem semelhança com domínios proteicos conhecidos. A análise de conservação demonstrou que de 27% a 41% dos ncRNAs putativos identificados são conservados em outras espécies de *Leishmania*. Foram encontrados de 27% a 38% de ncRNAs putativos com regulação atrelada ao estágio de desenvolvimento, dependendo dos estágios comparados. Assim, além da identificação e descrição de ncRNAs em *Leishmania* foram encontrados candidatos com regulação atrelada ao desenvolvimento e padrões foram descortinados com o processo de análise proposto e executado. Portanto, esse trabalho contribui significativamente para ampliar a compreensão dos processos de regulação de expressão gênica em *Leishmania* e oferecerá à comunidade um conjunto grande e importante de informações sobre a organização genética do parasito, diferenças genéticas e regulatórias ao longo do desenvolvimento, além de informações do transcriptoma de forma global.

Palavras-chave: Bioinformática, RNA não codificador, *Leishmania*, Transcriptoma, ncRNA, RNA-seq.

ABSTRACT

RUY, Patrícia de Cássia. ***In silico* identification and analysis of *Leishmania* ncRNAs using RNA-seq data.** 2017. 145 f. Tese (Doutorado) - Faculdade de Medicina de Ribeirão Preto / Universidade de São Paulo, Ribeirão Preto, 2017.

High-throughput data analyses indicated that transcriptomes are more extensive and complex than previously supposed. Currently, it is evident that most eukaryotic genomes are almost entirely transcribed and under regulation tied to developmental stages. Control of gene expression involves regulatory non-coding RNAs (ncRNAs), including post-transcriptional processes. Post-transcriptional regulation is particularly relevant for the regulation of gene expression in trypanosomatids, as compared to other organisms. Parasites of the genus *Leishmania* (Order *Kinetoplastidae*, family *Trypanosomatidae*) causes infectious-parasitic diseases known as leishmaniasis, and their life cycle comprises three development stages: procyclic promastigotes, metacyclic promastigotes and amastigotes. This study aimed to computationally identify and characterize *Leishmania* putative ncRNAs at different stages of development. Large-scale approaches combined with bioinformatics tools allowed the identification of 11,376 putative ncRNAs in *L. braziliensis* and, in a preliminary study, of 37 putative ncRNAs in *L. donovani*. In addition, the *L. braziliensis* the complete transcriptome was analyzed comparatively between the parasite development stages. In *L. donovani*, of the 37 putative ncRNAs identified, 34 were in UTRs (Untranslated Regions) and 3 in intergenic regions. Predictors of ncRNAs specific characteristics were used and 32 putative ncRNAs had at least one positive prediction. All candidates are conserved, partially or entirely, in at least three *Leishmania* species. Five *L. donovani* ncRNAs were confirmed by Northern blotting experiments. Analysis of the *L. braziliensis* transcriptome revealed differences in gene expression levels between developmental stages, ranging from 52% to 71%, depending on the compared stages. The boundaries of the 5'UTRs and 3'UTRs were also defined for 81% and 38% of the annotated CDSs, respectively. We developed a methodology for the identification of putative ncRNAs using total RNA sequencing data. This methodology allowed the identification of 11,376 putative ncRNAs in *L. braziliensis*, and all candidates were analyzed by predictive programs of ncRNAs specific characteristics and presented at least one positive prediction, in addition to bearing no similarity to known protein

domains. The analysis showed that from 27% to 41% of the putative ncRNAs identified are conserved in other *Leishmania* species. We found 27% to 38% of putative ncRNAs with regulation associated to the developmental stage, depending on the compared stages. Consequently, besides identification and characterization of ncRNAs in *Leishmania*, candidates with developmental-related regulation were found and patterns were uncovered with the proposed and implemented analysis. Thus, this work contributes significantly to improve understanding of gene expression regulation processes in *Leishmania* and will offer to the community important information about the parasite genetics and regulatory differences along the development, besides of *L. braziliensis* transcriptome information.

Keywords: Bioinformatics, non-coding RNA, *Leishmania*, Transcriptome, ncRNA, RNA-seq.

LISTA DE FIGURAS

Figura 1. Exemplos de ncRNAs já caracterizados em outros organismos..... 19
Figura 2. Imagens de pacientes diagnosticados com leishmaniose. 21
Figura 3. Representação esquemática do ciclo de vida digenético de *Leishmania* spp..... 22
Figura 4. Genes espécie-específicos em *Leishmania*..... 24
Figura 5. Distribuição de grupos ortólogos entre espécies de L. (*Leishmania*) e *L.* (*Viannia*). 24
Figura 6. Representação esquemática do processo de transcrição em *Leishmania*.26
Figura 7. Curva de crescimento de *Leishmania donovani* (linhagem Bob)..... 32
Figura 8. Representação esquemática da análise da fração de RNA não-polissômico de *L. donovani*..... 35
Figura 9. Região exemplo do alinhamento das *reads* no cromossomo 21 de *L. donovani*..... 40
Figura 10. Exemplos de ncRNAs putativos derivados de UTRs em *L. donovani*..... 43
Figura 11. ncRNA putativo intergênico..... 44
Figura 12. Porcentagem de *reads* com qualidade maior que 30..... 54
Figura 13. Imagem dos clusters gerados durante o sequenciamento..... 55
Figura 14. Distribuição de tamanho da biblioteca de cDNA na etapa de validação.. 56
Figura 15. Porcentagem de *reads* com qualidade maior que 30..... 57
Figura 16. Porcentagem de *reads* com qualidade maior que 30 por ciclo. 58
Figura 17. Qualidade das *reads* por base. 59
Figura 18. Alinhamento das bibliotecas por estágio de desenvolvimento de *L. braziliensis*..... 60
Figura 19. Contagem de reads por CDSs em cada biblioteca sequenciada..... 61
Figura 20. Distância entre as réplicas das bibliotecas. 62
Figura 21. *Heatmap* dos genes diferencialmente expressos para PROvsMETA..... 63
Figura 22. *Heatmap* dos genes diferencialmente expressos para METAvsAMA..... 64
Figura 23. *Heatmap* dos genes diferencialmente expressos para AMAvsPRO..... 65
Figura 24. Genes diferencialmente expressos na comparação entre os estágios de desenvolvimento de *L. braziliensis*..... 65
Figura 25. Distribuição do tamanho das 5'UTRs, CDSs e 3'UTRs de *Leishmania braziliensis*..... 82
Figura 26. Associação de abordagens usadas para identificação de ncRNAs putativos curtos e longos..... 84
Figura 27. ncRNA localizado interno a uma CDS mas na fita contrária..... 91
Figura 28. ncRNA putativos estágio específicos e consenso..... 92
Figura 29. Consenso de ncRNAs putativos após filtro com o Pfam..... 94
Figura 30. Distribuição dos tamanhos dos ncRNAs putativos..... 95
Figura 31. ncRNAs diferencialmente expressos na comparação entre os estágios de desenvolvimento de *Leishmania braziliensis*. 97
Figura 32. Gráfico *Elbow* gerado para o conjunto de dados de ncRNAs utilizando estatística gap. 102
Figura 33. Agrupamento do conjunto de dados de ncRNAs de *L. braziliensis*. 103
Figura 34. Avaliação da qualidade dos grupos gerados pela PAM..... 104
Figura 35. Clusterização hierárquica..... 105
Figura 36. Frequência RNAcon para o primeiro grupo do agrupamento hierárquico. 106
Figura 37. *Heatmap* complexo integrando dados de expressão diferencial (PROvsMETA) e características preditas dos ncRNAs de *L. braziliensis*..... 107

Figura 38. *Heatmap* complexo integrando dados de expressão diferencial (METAvsAMA) e características preditas dos ncRNAs de *L. braziliensis*. 108
Figura 39. *Heatmap* complexo integrando dados de expressão diferencial (AMAvsPRO) e características preditas dos ncRNAs de *L. braziliensis*. 109

LISTA DE TABELAS

Tabela 1. Informações do genoma de diferentes espécies de *Leishmania*. 23
Tabela 2. Regiões com alta densidade de *reads*. 41
Tabela 3. Caracterização dos ncRNAs identificados. 45
Tabela 4. Grau de conservação de sequência dos ncRNAs putativos em diferentes espécies de tripanossomatideos. 47
Tabela 5. Informações da primeira versão do sequenciamento teste feito no MiSeq. 56
Tabela 6. Informações da segunda versão do sequenciamento teste feito no MiSeq. 58
Tabela 7. Informações do sequenciamento no NextSeq dos três estágios de desenvolvimento de *L. braziliensis*..... 60
Tabela 8. Genes diferencialmente expressos entre promastigotas procíclico e metacíclico de *Leishmania braziliensis*: lista dos 25 genes apresentando DE mais marcante na análise comparativa. 66
Tabela 9. Genes diferencialmente expressos entre promastigota metacíclico e amastigota de *Leishmania braziliensis*: lista dos 25 genes apresentando DE mais marcante na análise comparativa. 67
Tabela 10. Genes diferencialmente expressos entre amastigota e promastigota procíclico de *Leishmania braziliensis*: lista dos 25 genes apresentando DE mais marcante na análise comparativa. 68
Tabela 11. Genes identificados como específicos de *L. braziliensis* em 2007 [41] e diferencialmente expressos. 72
Tabela 12. Classes GO enriquecidas para genes sobre-representados em PROvsMETA. 75
Tabela 13. Classes GO enriquecidas para genes sub-representados em PROvsMETA. 76
Tabela 14. Classes GO enriquecidas para genes sobre-representados em METAvsAMA. 77
Tabela 15. Classes GO enriquecidas para genes sub-representados em METAvsAMA. 78
Tabela 16. Classes GO enriquecidas para genes sobre-representados em AMAvsPRO. 79
Tabela 17. Classes GO enriquecidas para genes sub-representados em AMAvsPRO. 80
Tabela 18. Representação vetorial dos ncRNAs..... 89
Tabela 19. Número de ncRNAs identificados por cromossomo de *L. braziliensis*. ... 92
Tabela 20. ncRNAs diferencialmente expressos na comparação PROvsMETA..... 97
Tabela 21. ncRNAs diferencialmente expressos na comparação METAvsAMA..... 99
Tabela 22. ncRNAs diferencialmente expressos na comparação AMAvsPRO..... 100
Tabela 23. Características predominantes dos grupos formados pela PAM. 104

LISTA DE SIGLAS E ABREVIAÇÕES

BCV – *Biological Coefficient Variation*
BLAST – *Basic Local Alignment Search Tool*
CDS – *Coding Sequences*
CICS – *Conserved Intercoding Sequences*
DGC - *Directional Gene Clusters*
DNA – ácido desoxirribonucleico
EST – *Expressed Sequence Tags*
FPLC – *Fast Protein Liquid Chromatography*
ID – Identidade
MDS – *Multidimensional Scaling*
mL – mililitro
mM - milimolar
mRNA – ácido ribonucleico mensageiro
ncRNA – *non coding* RNA
nt – nucleotídeos
PTU – *polycistronic transcription unit*
pb – pares de base de nucleotídeos
pH – *potential of hydrogen*
RNA – ácido ribonucleico
RNAi – *RNA interference*
RNPs – ribonucleoproteínas
RNA-seq – sequenciamento de RNA
SIDER – *Short Interspersed DEgenerate Retroposons*
SL – *Spliced Leader* RNA
UTRs – *Untranslated Regions*
µg – micrograma

Sumário

1. INTRODUÇÃO 18

1.1 CONTEXTUALIZAÇÃO..... 18
1.1.1 RNAs não codificadores (ncRNAs) 18
1.1.2 Organismo modelo 20
1.1.2.1 Ciclo de vida do parasito 21
1.1.2.2 Genoma de *Leishmania*..... 22
1.1.2.3 Expressão gênica em *Leishmania*..... 25
1.1.2.4 Transcriptoma de *Leishmania* 27
1.1.2.5 ncRNAs em *Leishmania* 28
1.2 JUSTIFICATIVA E OBJETIVOS..... 29
1.2.1 Justificativa..... 29
1.2.2 Objetivo Geral 30
1.2.3 Objetivos Específicos..... 30
1.3 ORGANIZAÇÃO DO DOCUMENTO 30

2. *NON-CODING* RNAS EM *LEISHMANIA DONOVANI*.....31

2.1 VISÃO GERAL 31
2.2 MATERIAIS E MÉTODOS..... 31
2.2.1 Condições de cultura..... 31
2.2.2 Preparação das bibliotecas para identificação de ncRNAs – versão 1..... 32
2.2.3 Preparação das bibliotecas para identificação de ncRNAs – versão 2..... 33
2.2.4 Processamento das bibliotecas..... 33
2.2.5 Análise da correlação entre as réplicas 34
2.2.6 Identificação das regiões com alta densidade de reads 34
2.2.7 Teste de contaminação das bibliotecas 35
2.2.8 Bibliotecas de transcriptomas complementares..... 36
2.2.9 Informações de CICS e ESTs 36
2.2.10 Northern blotting..... 36
2.2.11 Caracterização computacional dos possíveis ncRNAs identificados..... 37
2.2.12 Conservação do ncRNAs identificados 38
2.3 RESULTADOS E DISCUSSÃO..... 38
2.3.1 Análise global dos dados de RNA-seq versão 1 e 2..... 38
2.3.1.1 RNA-seq versão 1 38
2.3.1.2 RNA-seq versão 2 48

3. ANÁLISE DO RNA-SEQ DE *LEISHMANIA BRAZILIENSIS* 49

3.1 VISÃO GERAL 49
3.2 MATERIAIS E MÉTODOS..... 49
3.2.1 Preparação das bibliotecas para identificação de ncRNAs..... 49
3.2.1.1 Fracionamento..... 49
3.2.1.2 RNA-total 50
3.2.2 Processamento das bibliotecas..... 50
3.2.3 Correlação entre as réplicas 51
3.2.4 Análise de expressão diferencial..... 51
3.2.4.1 Análise dos genes espécie-específicos de *L. braziliensis*..... 51
3.2.5 Enriquecimento Gene Ontology 51
3.2.6 Determinação dos limites das UTRs 52
3.2.6.1 5'UTR..... 52

3.2.6.2 3'UTR 53
3.3 RESULTADOS E DISCUSSÃO 53
3.3.1 Teste MiSeq versão 1 53
3.3.2 Teste MiSeq versão 2 57
3.3.3 Sequenciamento das bibliotecas de RNA total – NextSeq 58
3.3.3.1 Correlação entre as réplicas das bibliotecas de RNA total 61
3.3.3.2 Expressão diferencial 62
3.3.3.3 Enriquecimento de classes segundo *Gene Ontology* 75
3.3.4 Definição dos limites das UTRs 81

4. *NON-CODING* RNAS EM *LEISHMANIA BRAZILIENSIS* 83

4.1 VISÃO GERAL 83
4.2 MATERIAIS E MÉTODOS 83
4.2.1 Identificação dos ncRNAs putativos 83
4.2.2 Caracterização dos ncRNAs putativos 86
4.2.3 Conservação dos ncRNAs putativos 86
4.2.4 Expressão diferencial dos ncRNAs putativos 87
4.2.5 Busca por padrão 87
4.2.5.1 Representação vetorial dos ncRNAs 87
4.2.5.2 Agrupamento 89
4.3 RESULTADOS E DISCUSSÃO 91
4.3.1 Identificação dos ncRNAs putativos 91
4.3.2 Caracterização dos ncRNAs putativos 93
4.3.3 Conservação dos ncRNAs putativos 95
4.3.4 Expressão diferencial dos ncRNAs putativos 96
4.3.5 Busca por padrão 102
4.3.5.1 *Heatmaps* complexos – ncRNAs diferencialmente expressos 106

5. CONSIDERAÇÕES FINAIS 111

5.1 PRINCIPAIS CONTRIBUIÇÕES 114
5.2 PERSPECTIVAS 114

PRODUÇÃO CIENTÍFICA DURANTE O DOUTORADO 116

REFERÊNCIAS 117

ANEXOS 132

1. INTRODUÇÃO

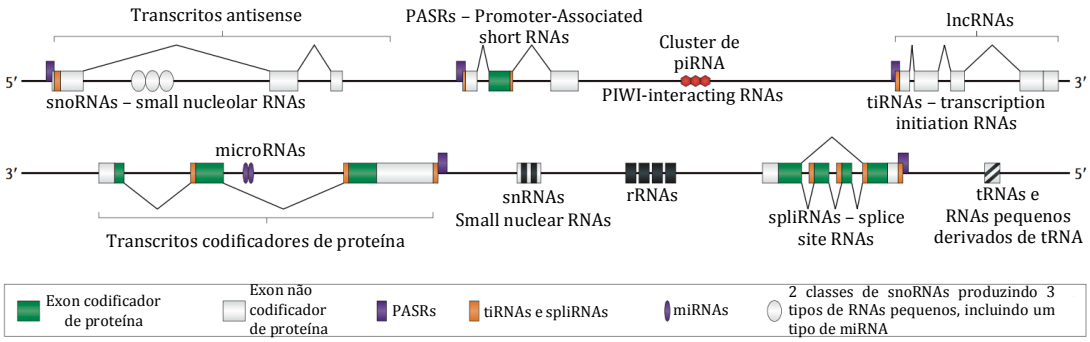
1.1 CONTEXTUALIZAÇÃO

1.1.1 RNAs não codificadores (ncRNAs)

As análises dos dados gerados em larga escala com as mais recentes ferramentas de sequenciamento de nova geração e as computacionais revelaram que os transcriptomas são mais extensos e complexos do que inferido previamente [1]. Hoje é evidente que a maioria dos genomas de eucariotos são quase inteiramente transcritos e sob regulação atrelada a estágios de desenvolvimento. Estratégias associando sequenciamento de bibliotecas de cDNA normalizadas (*deep sequencing*) com ancoragem dos transcritos aos cromossomos revelou dezenas de milhares de longos RNAs não codificadores de proteínas que podem ser intergênicos, *antisense*, sobrepostos e intrônicos e são dinamicamente expressos [2; 3; 4; 5].

Embora não possamos descartar a hipótese de que parte desses transcritos seja apenas "ruído" sem função biológica, os múltiplos papéis e a plasticidade de funções dos RNAs não codificadores que conhecemos, componentes críticos de várias maquinarias celulares, sugerem que existam outros tantos ncRNAs de fato funcionais. Para sustentar a relevância biológica desses transcritos há estudos que demonstram que vários deles são regulados durante o desenvolvimento, têm expressão tipo celular característico, localizam-se em compartimentos subcelulares específicos ou já foram associados a doenças humanas (revisado por [2]). ncRNAs regulatórios podem modular transcrição, tradução, estabilidade do mRNA e manutenção ou silenciamento de DNA por uma variedade de mecanismos, incluindo alterações de conformação de RNA, ligação a proteína, pareamento de bases com outros RNAs, e interações com DNA [6; 7; 8]. Apesar de várias classes/funções de ncRNAs serem conhecidas, estima-se que muitas outras funções ainda serão descobertas [9].

Figura 1. Exemplos de ncRNAs já caracterizados em outros organismos.



Representação gráfica de pequenos RNAs nucleolares (snoRNAs), RNAs ribossômicos (rRNAs), pequenos RNAs nucleares (snRNAs), RNAs não codificadores longos (lncRNAs), além de pequenos RNAs regulatórios como micro RNAs (miRNA), PIWI-*interacting* RNAs (piRNAs), *Promoter-Associated short* RNAs (PASRs), *splice site* RNAs (spliRNAs), transcription initiation RNAs (tiRNAs), RNAs pequenos derivados de snoRNA e RNAs pequenos derivados de tRNAs. Fonte: adaptado [10].

Existem diferentes tamanhos e tipos de ncRNAs conhecidos [10], que podem variar de ~22 nt (microRNA) até mais de 1 kb, como por exemplo o Xist RNA (*X-inactive-specific transcript RNA*) com 17kb [9]. Os ncRNAs <200 nt são conhecidos como curtos e são classificados em categorias de acordo com sua função, biogênese, características estruturais e seus ligantes (revisado em [11]) e os >200 nt são os ncRNAs longos (lncRNAs), que ainda não apresentam uma classificação satisfatória [12] (Figura 01).

O primeiro ncRNA caracterizado, o tRNA de alanina, descrito em 1965 [13], foi um dos grandes marcos da história de descoberta de ncRNAs, seguiu-se a revelação da maquinaria envolvida no processo de tradução. Mais de 30 anos depois, os RNAs de interferência, que atuam como reguladores da expressão gênica, foram descobertos, e o fenômeno de interferência por esses RNAs, RNAi (*RNA interference*), foi descoberto em plantas [14] e animais [15]. A partir de então um grande número de novos ncRNAs foi descoberto e sua importância vem aumentando (revisado em [10]), até que em 2012, o projeto ENCODE (*Encyclopedia of DNA Elements*) demonstrou que aproximadamente 80% do genoma humano e de camundongo transcreve ncRNAs [16]. Isso demonstra que a molécula de RNA é multi-tarefa, sendo que suas funções podem englobar tanto o domínio de codificação de proteína como de não-codificação e não somente ter mais de uma função incorporada mas também gerar vários produtos [10; 17].

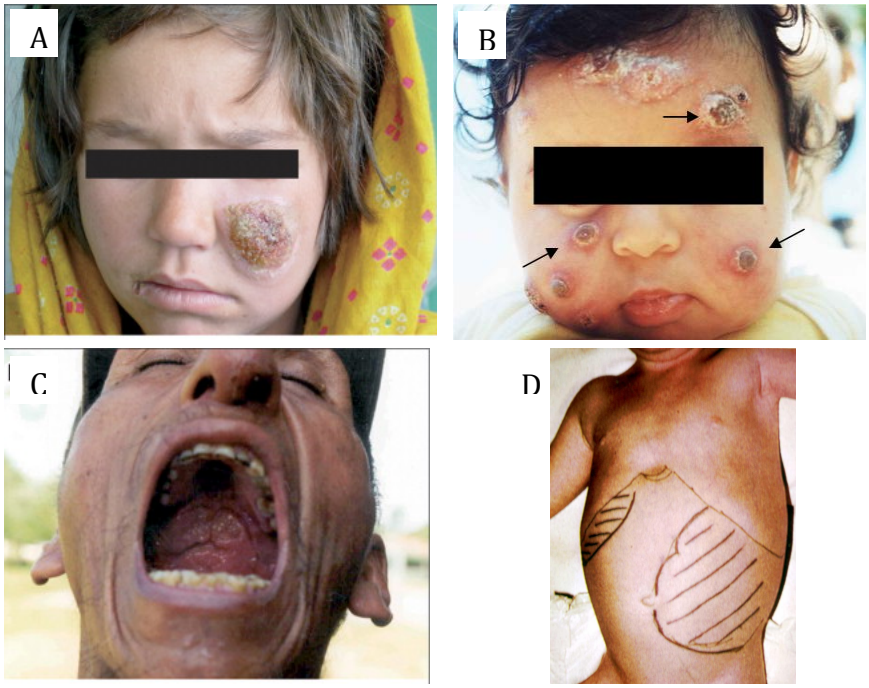
Os elementos funcionais de RNA estão se destacando cada vez mais devido a sua diversidade estrutural, regulatória e catalítica [18]. Vários RNAs regulatórios pequenos foram identificados em bactéria, regulando uma grande variedade de respostas adaptativas. Os pequenos RNAs regulatórios de bactéria estão envolvidos na regulação da tradução ou na estabilidades dos mRNAs [19; 20]. O controle da expressão gênica em nível pós-transcricional pode ocorrer em vários níveis, inclusive envolvendo elementos de RNA especificamente envolvidos em processos regulatórios pós-transcricionais [21]. Elementos *cis*-regulatórios presentes nas UTRs (*Untranslated Regions*) dos mRNAs são funcionais e envolvidos em regulação pós-transcricional da expressão gênica [22; 23]. A regulação pós-transcricional, é crucial em diferentes organismos e o mecanismo majoritário para a regulação da expressão gênica nos tripanossomatídeos [24].

1.1.2 Organismo modelo

Leishmania é um eucarioto ancestral do ponto de vista evolutivo que pertence à Ordem Kinetoplastidae, família *Trypanosomatidae*. Os tripanossomatídeos são responsáveis por patologias humanas de relevância para a saúde pública mundial por suas altas taxas de morbi-mortalidade e são ainda assim consideradas doenças negligenciadas pela Organização Mundial de Saúde (http://www.who.int/neglected_diseases/diseases/en/). O *Trypanosoma brucei* causa a doença do sono na África sub-sahariana, o *Trypanosoma cruzi*, a Doença de Chagas nas Américas do Sul e Central e 20 espécies de *Leishmania* causam a leishmaniose na África, América do Sul, Ásia e países do Mediterrâneo.

Os parasitos do gênero *Leishmania* provocam doenças infecto-parasitárias conhecidas como leishmanioses, as quais acometem o homem causando um amplo espectro de manifestações clínicas. Tais manifestações dependem da espécie infectante de *Leishmania* e da relação parasita-hospedeiro. Há um largo espectro de manifestações clínicas que variam de lesões de pele de cura espontânea (*L. major*) a manifestações mucosas desfigurantes (*L. braziliensis*) e a formas fatais como a doença visceral (*L. donovani/L. infantum/L. chagasi*) (Figura 02) [25].

Figura 2. Imagens de pacientes diagnosticados com leishmaniose.



A) Leishmaniose cutânea localizada; B) Leishmaniose cutânea difusa; C) Leishmaniose mucocutânea e D) Leishmaniose visceral. Fonte: [26] e do site https://www.ufpe.br/biolmol/Leishmanioses-Apostila_on_line/nas_americas.htm

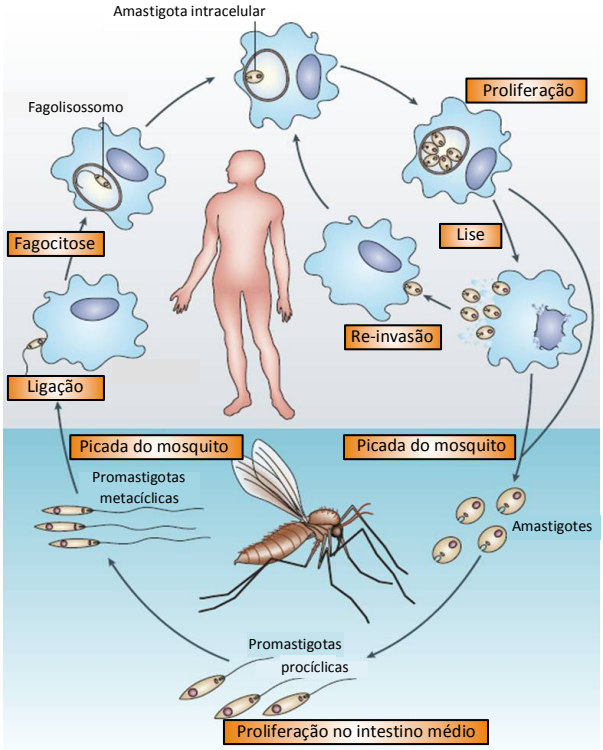
São conhecidas 20 espécies de *Leishmania* que causam a leishmaniose em aproximadamente 12 milhões de pessoas, com 2 milhões de novos casos registrados anualmente [27]. Não há vacinas disponíveis para humanos e os quimioterápicos em uso são insatisfatórios por sua alta toxicidade, regimes terapêuticos complexos e aparecimento de resistência em zonas endêmicas [28; 29; 30].

1.1.2.1 Ciclo de vida do parasito

A *Leishmania* completa seu ciclo de vida passando por três formas de desenvolvimento (Figura 03): o promastigota procíclico de flagelado livre, que é a forma proliferativa extracelular e habita o trato digestivo do inseto vetor (flebotomíneo); os promastigotas procíclicos sofrem diferenciação em metacíclicos, que é a forma infectiva; e o amastigota, que não tem flagelo livre, é a forma proliferativa intracelular e habita o fagolisossomo de macrófagos dos hospedeiros mamíferos [31; 32]. Os amastigotas dividem-se por fissão binária nesse ambiente hostil,

fugindo da resposta imune e protegendo-se do ambiente lítico e reativo pela produção de um arsenal próprio de fatores de resistência [33; 34].

Figura 3. Representação esquemática do ciclo de vida digenético de *Leishmania* spp.



Fonte: adaptado [35]

Nesse processo de adaptação aos diferentes ambientes hostis, as profundas alterações bioquímicas e morfológicas envolvem modificações importantes e rápidas no perfil de expressão gênica. Assim, conhecer os perfis globais de expressão gênica auxilia na compreensão dessas modificações e das especificidades de cada estágio de desenvolvimento.

1.1.2.2 Genoma de *Leishmania*

Os projetos genoma de *Leishmania* revelaram que os genes nesses organismos são organizados em clusters direcionais de genes, transcritos a partir da mesma fita de DNA. Experimentos de *run-on* nuclear no cromossomo 1 e 3 de *L. major* indicaram que a transcrição se inicia bidirecionalmente tanto entre dois clusters gênicos divergentes (cromossomo 1) [36] quanto convergentes (cromossomo 3) [37].

O projeto genoma de *L. major* (que integrou laboratórios da Inglaterra, França, Estados Unidos, Canadá e Brasil) foi finalizado em janeiro de 2005 e é sem dúvida um marco para o nosso conhecimento da genética desse parasito [38]. Esse projeto foi idealizado pela OMS/FIOCRUZ com o intuito de integrar estudos realizados por diferentes grupos de pesquisa, acelerar a compreensão da biologia do parasito e suas interações com os diferentes hospedeiros e visava ao desenvolvimento de vacinas, drogas e terapias mais eficazes. Atualmente, sabe-se que o número de cromossomos, sua ploidia, o tamanho do genoma e o número de genes identificados em *Leishmania* varia de acordo com as espécies (Tabela 01).

Tabela 1. Informações do genoma de diferentes espécies de *Leishmania*.

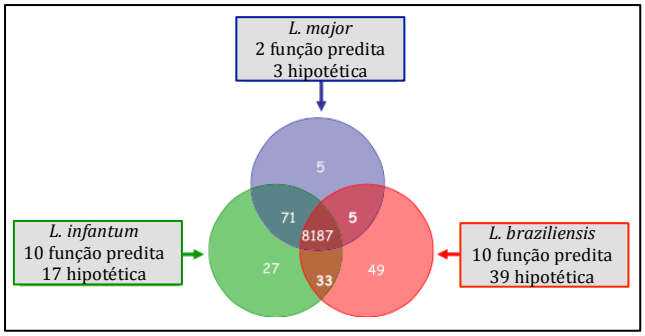
Característica	*L. major* *Friedlin*	*L. infantum* JPCM5	*L. mexicana* U1103	*L. braziliensis* M2904	*L. panamensis* PSC-1
Número de cromossomos	36	36	34	35	35
Variação do tamanho dos cromossomos (Mb)	0.27–2.68	0.28–2.67	0.27–3.34	0.23–2.69	0.27–2.61
Tamanho total (Mb)	32.85	31.92	30.94	31.24	30.69
Conteúdo GC (%)	59.72	59.58	59.72	57.72	57.56
Conteúdo de N (%)	<0.01	0.06	0.11	0.26	2.28
Número de *gaps*	9	440	410	876	553
Genes codificadores de proteínas	8,4	8,199	8,106	8,175	7,748*
Pseudogenes	89	61	99	195	185
Total de ncRNAs	893	136	96	109	163
Total de tRNAs	83	67	83	66	74
Total de rRNAs	63	11	13	6	17
Total de snRNA	6	8	-	7	5
Total de snoRNA	741	50	-	30	54

*Esse número menor de genes codificadores de proteínas anotados quando comparado às outras espécies de *Leishmania* é devido ao grande número de genes organizados em tandem, sendo que as cópias redundantes podem ter sido subestimadas por deficiência no processo de montagem utilizando *reads* de 454. Dados obtidos e adaptados de Llanes e colaboradores [39].

Em junho de 2007, um estudo comparativo de conteúdo gênico e arquitetura de *L. major*, *L. infantum* e *L. braziliensis* demonstrou que o genoma das três espécies apresentam elevado grau de sintenia e similaridade [40]. Também em 2007, Smith e colaboradores identificaram apenas ~200 genes diferencialmente distribuídos entre esses três genomas [41]. Sendo *L. braziliensis* o genoma mais

divergente, com 49 genes espécie-específicos, seguido por *L. infantum* com 27 genes espécie-específicos e *L. major* com apenas cinco (Figura 04).

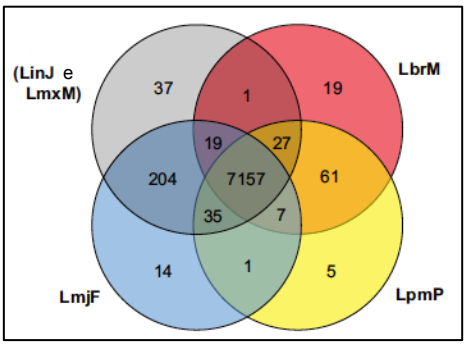
Figura 4. Genes espécie-específicos em *Leishmania*.



Fonte: adaptado [41].

Um estudo mais recente que avaliou a distribuição dos grupos de genes ortólogos entre espécies de *Leishmania* também revelou que *L. braziliensis* é a espécie mais divergente, com 19 grupos específicos desta espécie (Figura 05) [39].

Figura 5. Distribuição de grupos ortólogos entre espécies de L. (*Leishmania*) e *L.* (*Viannia*).



L. major (LmjF), *L. infantum* (LinJ), *L. mexicana* (LmxM), *L. braziliensis* (LbrM) and *L. panamensis* (LpmP). LinJ e LmxM foram agrupadas, tendo como número de grupos ortólogos específicos individuais 14 e 2, respectivamente. O valor de 37 apresentado é a soma dos valores individuais mais o número de grupos ortólogos comum as duas espécies (21). Fonte: [39]

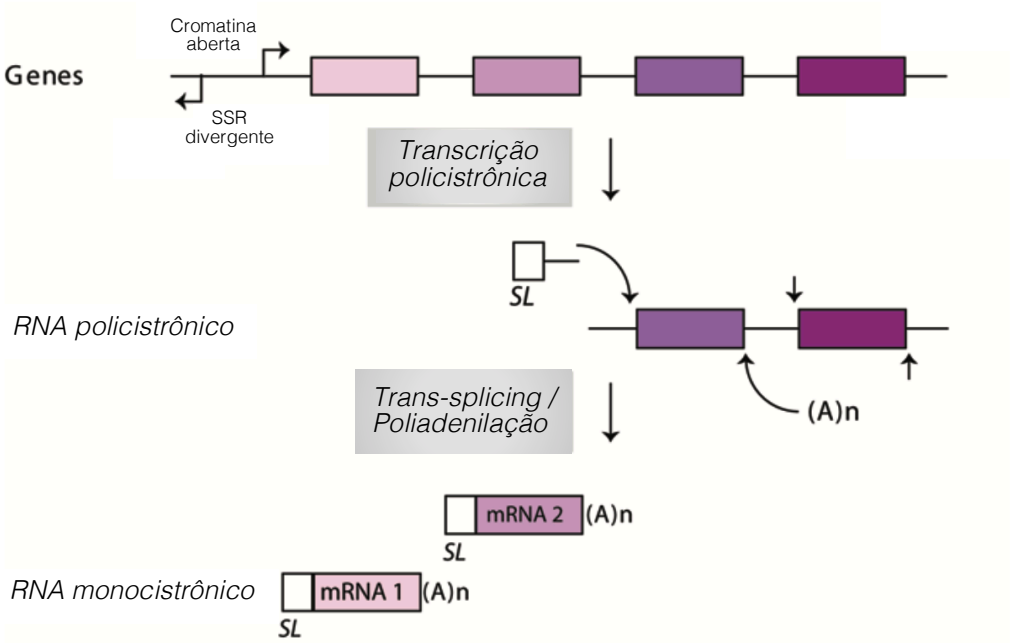
Esses resultados permitem especular que a diferença nas manifestações clínicas causadas pelas diferentes espécies de *Leishmania* pode ser mais complexa do que somente uma diferença de conteúdo gênico, podendo incluir diferenças na expressão de elementos regulatórios como os ncRNAs.

Alguns dos genes espécie-específicos de *L. braziliensis* apresentam grande divergência funcional, como os genes da maquinaria de RNAi (RNA de interferência). O termo RNAi foi definido para descrever eventos de regulação da expressão gênica que depende da atuação de pequenos RNAs regulatórios que por complementariedade com o RNA alvo direcionam os mRNAs alvo para destruição [15]. A atividade dessa via foi descrita em *Trypanosoma brucei* em 1998 [42]. Diversos estudos foram conduzidos para investigar a presença da maquinaria de RNAi em *L. major*, *L. infatum* e *T. cruzi*, sem sucesso [38; 40; 41; 43; 44]. Somente em *L. braziliensis* foi encontrada a maquinaria enzimática similar à de *T. brucei*, descoberta no processo de sequenciamento de seu genoma, disponibilizado em 2007 [40]. A funcionalidade da via "clássica" de RNAi foi confirmada posteriormente [45]. A aparente ausência da maquinaria clássica de RNAi em todas as espécies do subgênero *Leishmania* não exclui a possibilidade de que outros mecanismos e moléculas atuem em conjunto com ncRNAs como forma de regulação da expressão gênica nesses organismos. Esses pontos tornam ainda mais interessante o estudo de pequenos RNAs regulatórios nesses diferentes cenários, com a presença ou ausência dessa via "clássica" de RNAi, podendo trazer importantes esclarecimentos sobre o controle da expressão gênica tão peculiar nesses parasitos.

1.1.2.3 Expressão gênica em *Leishmania*

Os tripanossomatídeos exibem uma série notável de processos biológicos ímpares entre os eucariotos. As "novidades" moleculares descritas para os tripanossomatídeos incluem a ausência de condensação cromossômica máxima e da degeneração da membrana nuclear durante o ciclo celular; a transcrição policistrônica; o processamento do mRNA que ocorre por mecanismo de *trans-splicing*; a alta plasticidade genética e a edição de RNAs mitocondriais [46; 47; 48; 49; 50; 51]. Uma característica genética marcante dos tripanossomatídeos é a organização dos genes nos cromossomos. Os genes de *Leishmania* estão organizados como grandes unidades direcionais policistrônicas (DGC - *Directional Gene Clusters*) que podem compreender mais de uma centena de genes sem relação funcional e que carecem de sequências promotoras clássicas [38]. Como decorrência da transcrição policistrônica, a regulação da expressão gênica ocorre majoritariamente em nível pós-transcricional (Figura 06) [52; 53].

Figura 6. Representação esquemática do processo de transcrição em *Leishmania*.



Os genes estão organizados em unidades policistrônicas, que podem conter mais de uma centena de genes e, como a figura indica, a transcrição inicia-se em SSRs (*Strand Switch Region*) divergentes. Os clusters de genes são primeiramente transcritos em blocos policistrônicos. A seguir, ocorre o processamento por *trans-splicing* e poliadenilação de forma acoplada, dando origem aos mRNAs monicistrônicos maduros. Na figura, (A)n faz referência à cauda poliA e SL ao *Spliced Leader RNA* que traz a estrutura cap para cada mRNA maduro. Fonte: adaptado de [54].

O processamento dos grandes RNA precursores para formar os mRNAs monocistrônicos envolve a adição de uma sequência de 39 nucleotídeos, *spliced-leader* RNA (SL-RNA), na extremidade 5' de cada gene e da cauda poli-adenilada na extremidade 3', os dois processos ocorrem de forma acoplada (Figura 06) [51; 52]. Os sítios aceptores de *splicing* estão localizados a montante das CDSs e se caracterizam por um dinucleotídeo AG posicionado a jusante de uma trilha de polipirimidina. A maquinaria envolvida nesse processo de trans-splicing define espacialmente a posição de clivagem do transcrito a montante onde será adicionada a cauda poliA [51; 55]. Em *Leishmania major*, Dillon e colaboradores propuseram um modelo para representar a estrutura gênica (5'UTR, 3'UTR e CDS) empregando dados de transcriptoma (RNA-seq) [56]. Para as regiões 5'UTR e 3'UTR encontraram um tamanho mediano de 233 e 517 nt, respectivamente, considerando os sítios mais utilizados (sítio primário), indicando que essas regiões são maiores em *Leishmania* do que as encontradas em *Trypanosoma brucei* e *Trypanosoma cruzi* [57; 58],

condizendo com observações anteriores de que os genomas de *Leishmania* são mais compactos [59].

O controle da expressão gênica em nível pós-transcricional pode ocorrer em vários níveis inclusive de maneira integrada, seja no controle da abundância de RNA, ou no controle traducional ou pós-traducional [60; 61; 62]. Esses dois últimos níveis envolvem desde o controle na iniciação, elongação e terminação da tradução até a degradação nos níveis de proteína [62]. Com relação ao controle da expressão gênica referente a quantidade de mRNA, essa regulação pode ocorrer por meio de sequências regulatórias que interferem no gene próximo a ela (cis-elemento) ou que influencia a regulação de genes distantes (trans-elemento) .

Foi investigado o envolvimento desses elementos envolvidos na diferença de expressão gênica estágio-específica encontrada em *L. major* e *L. infantum* [63]. Entre esses elementos estão os SIDER (*Short Interspersed DEgenerate Retroposons*) que são resquícios de retroelementos que podem desempenhar atividades regulatórias e que foram identificados em *L. major*, *L. infantum* e *L. braziliensis* [63; 64; 65]. Rochette e colaboradores não encontraram associação dos SIDER com a diferença de expressão de transcritos, sugerindo que elementos regulatórios adicionais (não-SIDER) possam estar contribuindo para a diferença estágio-específica de acumulo dos mRNAs em *Leishmania* [63]. Regiões intergênicas conservadas, denominadas CICS (*Conserved Intercoding Sequences*), foram identificadas em diferentes espécies de *Leishmania*, sendo que essas sequências podem agir como cis ou trans-elementos regulatórios [66].

1.1.2.4 Transcriptoma de *Leishmania*

O sequenciamento genômico (*whole-genome sequencing*), tanto para organismos unicelulares como para organismos superiores transformou as ciências da vida mas o conhecimento do transcriptoma é necessário para o entendimento de como esse genoma dá origem a diferentes tipos celulares em um organimo e como esses genes são regulados [67]. Um marco na era pós-genômica foi a técnica de sequenciamento de RNA em larga escala (RNA-seq) que revolucionou a pesquisa em expressão gênica. Entender o transcriptoma é essencial para catalogar todas as espécies de transcritos (mRNAs, ncRNAs e pequenos RNAs); determinar a estrutura transcricional dos genes (sítio de iniciação, 5'UTR e 3'UTR, entre outros); e para

quantificar a mudança nos níveis de expressão de cada transcrito durante o desenvolvimento e sob diferentes condições [68; 69].

O emprego do sequenciamento em larga escala, permite ampliar o conhecimento acerca do transcriptoma, suas modulações frente a situações diversas e consequentemente acerca das vias de controle de expressão gênica do parasito, com maior rapidez e eficiência. O transcriptoma de *L. major* (forma promastigota) foi anotado pela primeira vez com os genes delimitados em regiões traduzidas e não traduzidas em 2013, utilizando dados de RNA-seq [70]. Neste trabalho foram encontrados anotações gênicas erradas além de novos genes não anotados, detectando um total de 10.285 transcritos.

1.1.2.5 ncRNAs em *Leishmania*

Em *Leishmania,* pequenos RNAs *sense* e *antisense* foram detectados há vários anos [71], sem que uma função fosse a eles atribuída. Mas um papel funcional para um ncRNA *antisense* só veio a ser descrito mais de uma década depois em *Trypanosoma brucei*, o primeiro tripanossomatídeo a ter a maquinaria de RNAi caracterizada [72; 73]. Mais recentemente, a anotação do genoma de *L. braziliensis* revelou a existência de maquinaria similar nessa espécie [40; 41]. No entanto, e a despeito de vários esforços, todas as análises computacionais ou experimentais falharam em identificar maquinaria de RNAi similar em *Trypanosoma cruzi*, *Leishmania major,* ou *Leishmania infantum* [38; 40; 43; 44]. Mas não se pode perder de vista que outros mecanismos de ação dos ncRNAs existem e foram descritos em organismos diferentes [3; 74; 75]. Foi demonstrado que ncRNAs, especialmente os curtos, são importantes moléculas na regulação pós-transcricional, principalmente os miRNAs (microRNAs) e siRNAs (small interfering RNAs) [18]. Em *Leishmania*, Dumas e colaboradores [76] demonstraram a presença de uma família de lncRNAs com regulação ligada ao estágio de desenvolvimento e em nosso laboratório, foi identificada uma classe distinta de ncRNAs [77].

As estratégias clássicas de sequenciamento são dirigidas para identificar/analisar mRNAs e/ou RNAs abundantes nas amostras, com isso, a descoberta de RNAs não canônicos, como ncRNAs pequenos ou RNAs pouco abundantes, pode ser dificultada. Sendo assim, estão sendo desenvolvidas estratégias alternativas para a construção de bibliotecas de RNAs não codificadores,

pequenos ou longos, que podem ser empregadas para a descoberta de ncRNAs em *Leishmania* [78; 79]. Como o trabalho de Lambertz e colaboradores que, utilizando RNA-seq, identificaram pequenos RNAs derivados de tRNAs e rRNAs enriquecidos em exossomos de *L. donovani* e *L. braziliensis* [80]. Outro exemplo foi o trabalho de Michaeli e colaboradores que, utilizando RNA-seq e associação a RNPs (Ribonucleoproteínas), identificaram novos ncRNAs putativos em *Trypanosoma brucei* [79]. Kolev e colaboradores analisando o transcriptoma de *T. brucei*, identificaram ncRNAs putativos derivados de UTR [57].

O poder informativo das estratégias de larga escala, como o RNA-seq, depende primordialmente dos processos *in silico* escolhidos para a exploração dos dados gerados. Dessa forma, o desenvolvimento e a integração de ferramentas computacionais que tratem da análise de expressão gênica são essenciais para que o grande volume de dados gerados seja aproveitado e de fato informativo.

1.2 JUSTIFICATIVA E OBJETIVOS

1.2.1 Justificativa

Assim, à luz das descobertas recentes sobre os papéis desempenhados por pequenos e longos ncRNAs, o modo de controle da expressão gênica em *Leishmania* e a disponibilidade de novas tecnologias tais como RNA-seq, torna-se central investigar possíveis modos de ação e funções destes ncRNAs no parasito, pois podem desempenhar papel fundamental na sobrevivência do parasito e patogênese e tornam-se alvos potenciais para terapêutica. Nessa direção, a emprego de técnicas de prospecção em larga escala de dados do transcriptoma para identificação de ncRNAs potencialmente funcionais possibilitará caracterizar possíveis padrões ainda não identificados para ncRNAs de *Leishmania*.

De forma que nesse trabalho pretendemos contribuir adicionando informação inédita, buscando uma classificação dos ncRNAs putativos encontrados computacionalmente no protozoário *Leishmania* e definindo padrões possíveis. Essa contribuição vai agregar dados aos trabalhos já conduzidos em tripanossomatídeos por outros e vai permitir que nosso grupo de pesquisa investigue o papel funcional de alguns dos RNAs identificados contribuindo assim para ampliar o entendimento dos processos de regulação da expressão gênica em *Leishmania*.

1.2.2 Objetivo Geral

Integração de ferramentas computacionais para exploração de dados do transcriptoma de *Leishmania* em busca da identificação de ncRNAs e da diferença de expressão entre três estágios de desenvolvimento do parasito.

1.2.3 Objetivos Específicos

1) Conduzir uma análise global e comparativa dos transcriptomas dos três principais estágios de desenvolvimento do parasito para identificar ncRNAs computacionalmente.
2) Anotar as características estruturais e funcionais dos RNAs pela utilização de preditores específicos e definir as regiões 5'UTR e 3'UTR dos genes anotados no genoma de *Leishmania braziliensis*.
3) Desenvolver *scripts* para seleção de informações essenciais para posterior exploração dos dados de cada um dos ncRNAs identificados.
   i) Criar uma pontuação para cada ncRNA identificado de acordo com as características encontradas nas predições.
   ii) Criar automaticamente representações vetoriais de cada um dos ncRNAs identificados.
4) Utilizar técnicas de prospecção de dados para explorar de maneira automática a informação latente contida nos dados, elucidando relações desconhecidas entre os ncRNAs e suas características e também as similaridades e especificidades de cada estágio.

1.3 ORGANIZAÇÃO DO DOCUMENTO

Esta tese está organizada em cinco capítulos principais. O Capítulo 1 contém a fundamentação teórica. Os Capítulos intermediários 2, 3 e 4 apresentam a visão geral dos dados discutidos no capítulo, a metodologia utilizada e os resultados e discussão. Especificamente, o Capitulo 2 apresenta a identificação e análise de ncRNAs putativos em *Leishmania donovani*; o Capítulo 3 contém a análise dos dados de sequenciamento de RNA total (RNA-seq) de *Leishmania braziliensis* e o Capítulo 4 apresenta a identificação e análise de ncRNAs putativos em *Leishmania braziliensis*. O Capítulo 5 apresenta a conclusão do trabalho, com as principais contribuições e perspectivas.

2. *NON-CODING* RNAS EM *Leishmania donovani*

2.1 VISÃO GERAL

Iniciamos a busca por ncRNAs putativos em *Leishmania donovani*. Esta espécie de *Leishmania* era a única para a qual estava descrito um método simples, robusto e reprodutível para obtenção de promastigotas metacíclicos a partir de culturas de promastigotas em fase estacionária de crescimento e outra para obtenção de culturas de amastigotas axênicos [81].

Inicialmente nossa proposta era empregar uma estratégia de fracionamento de ribonucleoproteínas (RNPs) não polissômicas, como fonte dos transcritos a serem sequenciados. Nosso intuito era evitar os mRNAs e os ncRNAs clássicos, todos associados à maquinaria de tradução. Dessa forma, pretendíamos obter uma fração enriquecida de RNAs não codificadores integrados a partículas de RNPs menos densas e evitaríamos também a inclusão de pequenos RNAs livres, uma fração mais contaminada de produtos de degradação, gerados pela manipulação. Essa estratégia foi descrita por Rederstorff e colaboradores [78] e abordagem similar também empregada em *Trypanosoma brucei* [79].

Os principais resultados deste capítulo resultaram no primeiro artigo publicado como fruto deste trabalho (em co-autoria principal compartilhada com Felipe F. Castro) [82].

Todos os arquivos de resultados citados estão disponíveis sob solicitação (akcruz@fmrp.usp.br).

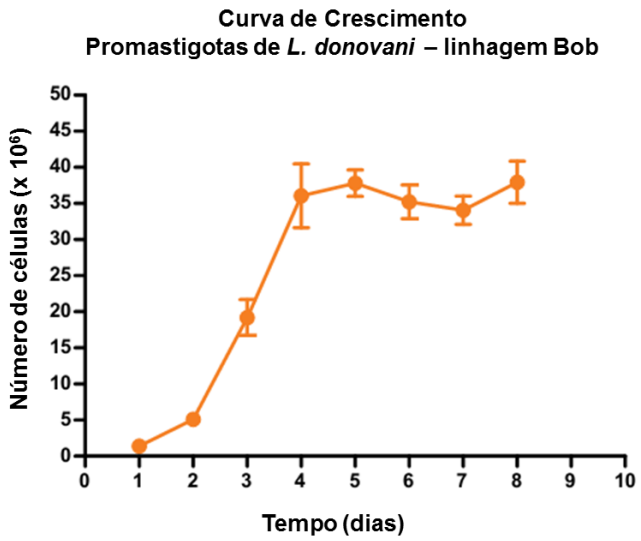
2.2 MATERIAIS E MÉTODOS

2.2.1 Condições de cultura

As três formas de vida de *L. donovani* foram obtidas de acordo com Goyard e colaboradores [81]. A curva de crescimento foi iniciada com o inóculo de 2 x 10^5 células/ml em meio de cultura M199, pH 7.4, suplementado com 10% de Soro Fetal Bovino (Figura 7). Os promastigotas procíclicos foram coletados no dia 3 da cultura, em fase exponencial de crescimento. Na fase estacionária de crescimento (dia 5 de cultura), a população de promastigotas que está enriquecida de formas metacíclicas (~10%) foi coletada. Os amastigotas axênicos foram obtidos a partir de promastigotas procíclicos de terceiro dia da cultura pela indução da diferenciação para amastigotas axênicos em Meio de Diferenciação, pH 5.5, suplementado com 10

mM de succinato de sódio e 25% de Soro Fetal Bovino. Experimento realizado por Ramon F. Santos.

Figura 7. Curva de crescimento de *Leishmania donovani* (linhagem Bob).



A curva de crescimento foi iniciada com o inóculo de 2 x 10^5 células/ ml em meio de cultura M199, pH 7.4, suplementado com 10% de Soro Fetal Bovino. Os promastigotas procíclicos foram coletados no dia 3 da cultura; a fase estacionária de crescimento no dia 5 de cultura. Os amastigotas axênicos foram obtidos a partir de promastigotas procíclicos de terceiro dia da cultura pela indução da diferenciação para amastigotas axênicos em Meio de Diferenciação. Fonte: Ramon F. Santos

2.2.2 Preparação das bibliotecas para identificação de ncRNAs – versão 1

Foram construídas bibliotecas de RNAs inseridos em partículas Ribonucleoproteicas (RNPs) não polissômicas, fracionadas por tamanho por FPLC (*Fast Protein Liquid Chromatography*). Foi utilizado o extrato de núcleo e citoplasma de *L. donovani* e o conteúdo de RNAs das frações selecionadas de cada amostra foi analisado por eletroforese digital utilizando o equipamento Bioanalyzer (*Agilent Technologies*).

Foram produzidas 18 amostras, triplicatas da fração nuclear e citosólica de cada um dos três estágios do ciclo de vida de *L. donovani*. Utilizando cultura axênica foram selecionados: promastigotas procíclicos, de fase logarítmica, promastigotas de fase estacionária (população que conta com ~10% de promastigotas metacíclicos) e amastigotas. Os RNAs foram processados para a construção das bibliotecas para o sequenciamento MiSeq-Illumina, com uma adaptação feita ao protocolo do kit *TruSeq® Stranded mRNA Sample Preparation*, pois a maioria dos protocolos

comerciais são desenhados para eucariotos superiores. Foi removido o primeiro passo do protocolo que tem a finalidade de purificar moléculas de mRNA contendo cauda poli(A). A fragmentação em pedaços menores também foi eliminada porque os ncRNAs já estavam num tamanho médio de 150 nucleotídeos. Os produtos foram purificados e enriquecidos por PCR (*Polymerase Chain Reaction*) para criar a biblioteca final de cDNAs. Experimentos idealizados e executados por Ramon F. Santos, pós-doutor do laboratório. As bibliotecas foram sequenciadas utilizando a plataforma MiSeq-Illumina (FMRP-USP).

2.2.3 Preparação das bibliotecas para identificação de ncRNAs – versão 2

Para a construção das bibliotecas da segunda versão do sequenciamento foi abandonado o fracionamento subcelular, de forma a recuperar todo o conteúdo de RNPs das células. Além disso, com o objetivo de evitar a alta contaminação com RNA ribossômico (rRNA) as condições de ultracentrifugação foram modificadas (150.000 × *g*). A partir desse ponto, foi utilizado o kit *TruSeq® Stranded mRNA Sample Preparation* com as mesmas adaptações da versão 1 do sequenciamento (Seção 3.1). Experimentos também idealizados e executados por Ramon F. Santos. As bibliotecas foram sequenciadas utilizando a plataforma MiSeq-Illumina (FMRP-USP).

2.2.4 Processamento das bibliotecas

Para a remoção de adaptadores foi utilizado o programa cutadapt (http://journal.embnet.org/index.php/embnetjournal/article/view/200), versão1.4.1. A verificação da qualidade, tamanho e número de *reads* foi feita com FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

Nas primeiras versões do sequenciamento, as *reads* foram alinhadas com o genoma de referência de *L. donovani* versão 6.0 disponível no TriTrypDB [83]. O mapeamento das reads foi feito com o programa Bowtie2 [84] com parâmetros específicos para bibliotecas *paired-end*. Os parâmetros utilizados no alinhamento foram *–N 1* que permite apenas 1 *mismatch* no alinhamento e *--local* que não exige que a *read* alinhe de ponta a ponta. O programa SAMtools [85] foi utilizado para converter arquivos sam para o formato binário (arquivo bam), além da ordenação,

indexação e contagem das *reads* mapeadas por cromossomo nos arquivos bam. Esses arquivos de alinhamento no formato bam foram visualizados e analisados com o programa Artemis [86].

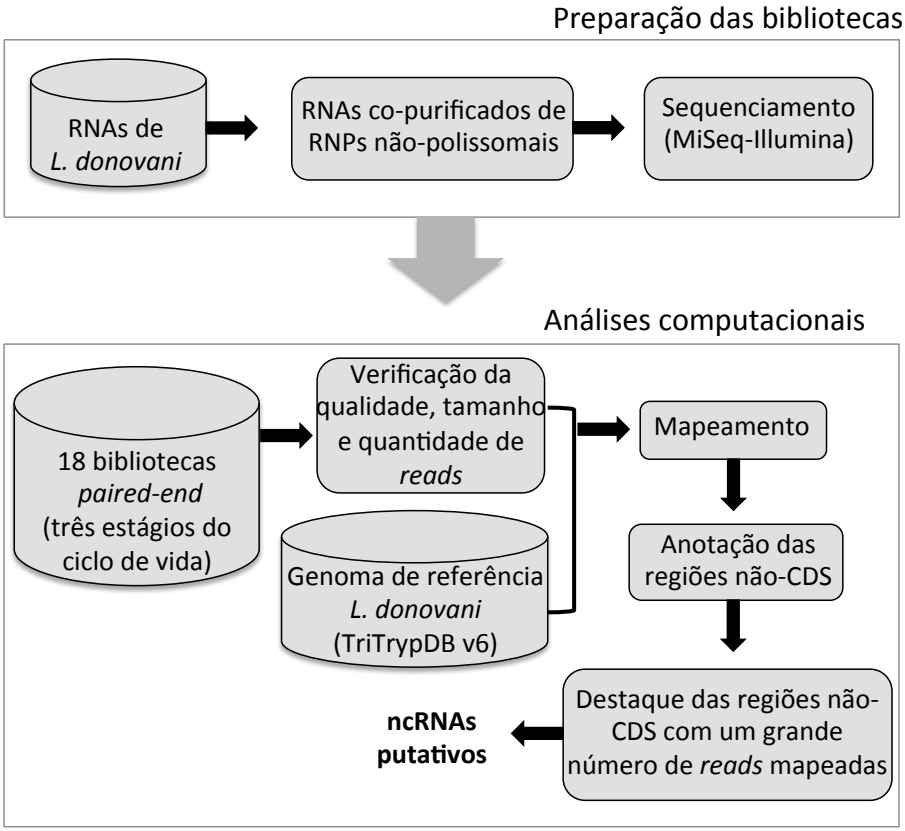
2.2.5 Análise da correlação entre as réplicas

Foi desenvolvido um *script* R [87] que faz a leitura de um arquivo com a contagem do número de *reads* por região (contagem realizada com HTSeq, seção 2.2.6) e realiza a etapa de normalização. O agrupamento das réplicas, ou seja, as similaridades/diferenças entre as réplicas analisadas foram observadas pelo gráfico MDS (*Multidimensional Scaling*) e a distância BCV (*Biological Coefficient Variation*). BCV é o coeficiente de variação com o qual a verdadeira abundância (desconhecida) de um transcrito varia entre as réplicas das amostras de RNA [88]. Assim, a análise de expressão diferencial só é realizada caso a diferença entre as réplicas seja menor do que a diferença entre as bibliotecas dos diferentes estágios de desenvolvimento analisados.

2.2.6 Identificação das regiões com alta densidade de *reads*

As regiões contendo alta densidade de *reads* mapeadas foram consideradas representativas das regiões com maior potencial de serem ncRNAs. Para a identificação automática desses potenciais ncRNAs na análise da primeira versão do sequenciamento utilizamos o programa HTSeq [89] (*script* htseq-count). Como um dos arquivos de entrada, o HTSeq utiliza um arquivo gff com a anotação das regiões de interesse na análise. Neste projeto, essas regiões são aquelas entre as CDSs anotadas no genoma, não considerando inicialmente os possíveis ncRNAs presentes em regiões de CDSs. Sendo assim, foi necessário o desenvolvimento de um *script* Perl que alterasse o arquivo gff com a anotação de CDSs (disponível no TriTrypDB) para conter a anotação de regiões não CDSs. Algumas regiões com erro de anotação foram definidas manualmente. A figura 8 apresenta um esquema resumindo o processamento das bibliotecas sequenciadas.

Figura 8. Representação esquemática da análise da fração de RNA não-polissômico de *L. donovani*.



Preparação e análises computacionais aplicadas as bibliotecas de RNA-seq de *L. donovani*. Fonte: autoral

2.2.7 Teste de contaminação das bibliotecas

Por causa do baixo número de *reads* no mapeamento com o genoma nuclear de *Leishmania* na segunda versão do sequenciamento (Seção 2.3.1.1) foi feita a verificação de possíveis contaminações com transcritos não-*Leishmania*. Para tanto, foi empregado o programa BLAST (*Basic Local Alignment Search Tool*) [90] com os parâmetros *--outfmt 7* para que o arquivos de saída estejam no formato tabular e *--num_threads 7* para que sejam utilizados 7 núcleos do *cluster* durante a execução. Como banco de dados, foi utilizado o não redundante (NR) do NCBI, considerando assim sequências de todos os organismos disponíveis em domínio público. O *cluster* do LCCA (Laboratório de Computação Científica Avançada) da USP foi utilizado. Os resultados obtidos com o BLAST foram processados por um *script* Perl que fazia o agrupamento e a contagem dos *hits* obtidos em cada biblioteca.

2.2.8 Bibliotecas de transcriptomas complementares

Em colaboração com Peter Myler (SBRI - *Seattle Biomedical Research Institute*) foram utilizadas, de maneira combinada, bibliotecas de RNA total (mRNAs), *ribosome profilling* de *L. donovani* e de bibliotecas direcionadas à detecção/enriquecimento de sítios aceptores de SL RNA de *L. donovani* (dados não publicados). A técnica de *ribosome profilling* fornece um cenário completo da tradução pois permite posicionar quantitativamente os ribossomos em seus moldes de mRNA [91; 92]. A biblioteca de SL RNA é enriquecida em regiões 5'UTR e são utilizados iniciadores complementares ao SL RNA de *Leishmania*. Detalhes das três metodologias utilizadas no sequenciamento dessas bibliotecas podem ser encontrados em Jensen, BC e colaboradores [93].

2.2.9 Informações de CICS e ESTs

Resultados de estudos anteriores em nosso laboratório foram combinados com nossos resultados de sequenciamento para conduzir a análise da presença de ncRNAs. Em 2002, Iribar [94] observou a existência de ESTs (*Expressed Sequence Tags*) de *L. major,* geradas a partir de uma biblioteca de mRNAs, que não seriam classificadas como CDSs por serem muito curtas ou terem conteúdo AT acima do padrão de sequências codificadoras em *Leishmania* [95]. Essas ESTs foram por nós mapeadas no genoma de *L. donovani* juntamente com os transcritos identificados por RNA-seq (Fig. 9). Adicionamos ainda sequências intercodificadoras conservadas (CICS, *Conserved Intercoding Sequences*) entre espécies de *Leishmania*, que haviam sido identificadas computacionalmente por Vasconcelos e colaboradores [66]. Os CICS são potenciais cis-elementos regulatórios presentes em transcritos de *Leishmania.* Todos os mapeamentos foram feitos utilizando BLAST com parâmetros *--outfmt 7* para que o arquivos de saída estivessem no formato tabular e *--word_size 7* partindo de sequência inicial de *match* menor que o padrão. O mapeamento de ESTs e CICS foi feito como um adjuvante de identificação de potenciais transcritos ou ncRNAs.

2.2.10 *Northern blotting*

Para os experimentos de *Northern blotting*, o RNA dos promastigotas de fase logarítmica e estacionária e dos amastigotas foi extraído a partir das células

cultivadas como descrito acima (Seção 2.2.1). As células foram lavadas em PBS, suspensas em 1 ml de reagente de Trizol (*Invitrogen*) e armazenadas a -70 °C. Posteriormente, as amostras foram processadas para extração do conteúdo de RNA total segundo instruções do fabricante (*Invitrogen*) e a quantidade total foi mensurada por fluorescência (*Quant-iT™ RiboGreen® RNA Reagent/Invitrogen*).

Uma alíquota de 10 µg de RNA total de cada amostra foi carregada em gel de poliacrilamida 6% contendo 8 M de ureia (gel TBE/ureia). Ao final, o RNA fracionado no gel foi transferido para membrana de nylon (*Amersham Biosciences*) para condução dos experimentos de hibridação. As marcações foram feitas por incorporação de α-P^{32}-dCTP, por *random priming* [96]. Estes experimentos foram executados por Ramon F. Santos, Viviane Ambrósio e Tânia Defina.

2.2.11 Caracterização computacional dos possíveis ncRNAs identificados

Para a caracterização dos possíveis ncRNAs preditos foram utilizados 4 programas preditores idealizados para a identificação de ncRNAs. O primeiro foi o RNAcon [97], uma ferramenta que utiliza SVM (*Support Vector Machine*), que tem como parâmetro a composição de nucleotídeos para classificar a sequência como codificadora ou não codificadora. Após essa classificação, o programa faz a predição da estrutura (programa IPknot [98]) em que são filtradas algumas propriedades para classificar o possível ncRNA em 18 categorias diferentes. O segundo programa utilizado foi o RNAspace [99], uma plataforma para anotação de ncRNAs que considera bancos de dados de domínios conhecidos de ncRNAs para fazer buscas nas sequências. O terceiro programa utilizado foi o PORTRAIT [100] que utiliza métodos *ab initio* que avaliam o potencial codificador de uma sequência por meio de SVM. O quarto programa utilizado foi o snoscan que utiliza métodos com modelo probabilístico na busca por snoRNAs guiados por metilação [101]. Além disso, foram feitas buscas em bancos de dados de ncRNAs conhecidos: fRNAdb [102], que inclui outros 8 bancos de dados, NONCODE (versão 4.0) [103], que inclui todos os tipos de ncRNAs (exceto tRNAs e rRNAs), e lncRNAdb [104], que apresenta ncRNAs longos de eucariotos. Foi utilizado o programa BLAST com parâmetros *--outfmt 7* para que o arquivos de saída estejam no formato tabular e *--word_size 7* para permitir a inclusão de sequências de *match* menor que o padrão.

2.2.12 Conservação do ncRNAs identificados

Sendo a conservação de sequência um parâmetro indicador de pressão seletiva e portanto possível relevância funcional, foi investigada a conservação dos ncRNAs identificados. Para isso, uma análise de similaridade de sequência foi realizada com os genomas de *L. major* Friedlin, *L. infantum* JPCM5, *L. amazonensis* MHOMBR71973M2269, *L. braziliensis* MHOMBR75M2903, *L. tarentolae* ParrotTarII, *L. enriettii* LEM3045, *T. brucei* TREU927 (TriTrypDB – versão 28). Um *hit* foi considerado positivo com um *e-value* $<10^{-5}$.

2.3 RESULTADOS E DISCUSSÃO

2.3.1 Análise global dos dados de RNA-seq versão 1 e 2

As duas primeiras versões de sequenciamento foram construídas usando abordagens alternativas para fracionamento de RNPs não polissômicas de *L. donovani*. As bibliotecas foram sequenciadas (MiSeq, Illumina) e a análise global revelou que as estratégias de purificação/preparação dos ncRNAs resultaram em coberturas baixas do transcriptoma.

2.3.1.1 RNA-seq versão 1

Na primeira versão do sequenciamento foram consideradas as frações nucleares e citosólicas separadamente, em triplicatas de três estágios de desenvolvimento de *L. donovani* em cultura axênica: promastigotas procíclicos, promastigotas metacíclicos e amastigotas. Considerando as 3 réplicas, no total foram obtidas para a fração citosólica 1.263.336 *reads* para procíclicos, 495.916 para metacíclicos, 875.192 para amastigotas e considerando a fração nuclear, 1.672.108 *reads* procíclicos, 1.064.026 metacíclicos e 1.396.664 amastigotas.

Foram consideradas as bases com qualidade acima de 30 (qualidade Phred). Após o mapeamento com o programa Bowtie2 com o genoma de referência de *L. donovani*, observamos que apesar de o número de *reads* mapeadas ter sido satisfatório, a quantidade de *reads* mapeadas no cromossomo 27 era de aproximadamente 50% e 80% nas frações citosólica e nuclear, respectivamente. Neste cromossomo estão os genes dos RNAs ribossômicos (rRNAs) e foi precisamente nessa região que houve o maior número de *reads* alinhadas. Dessa

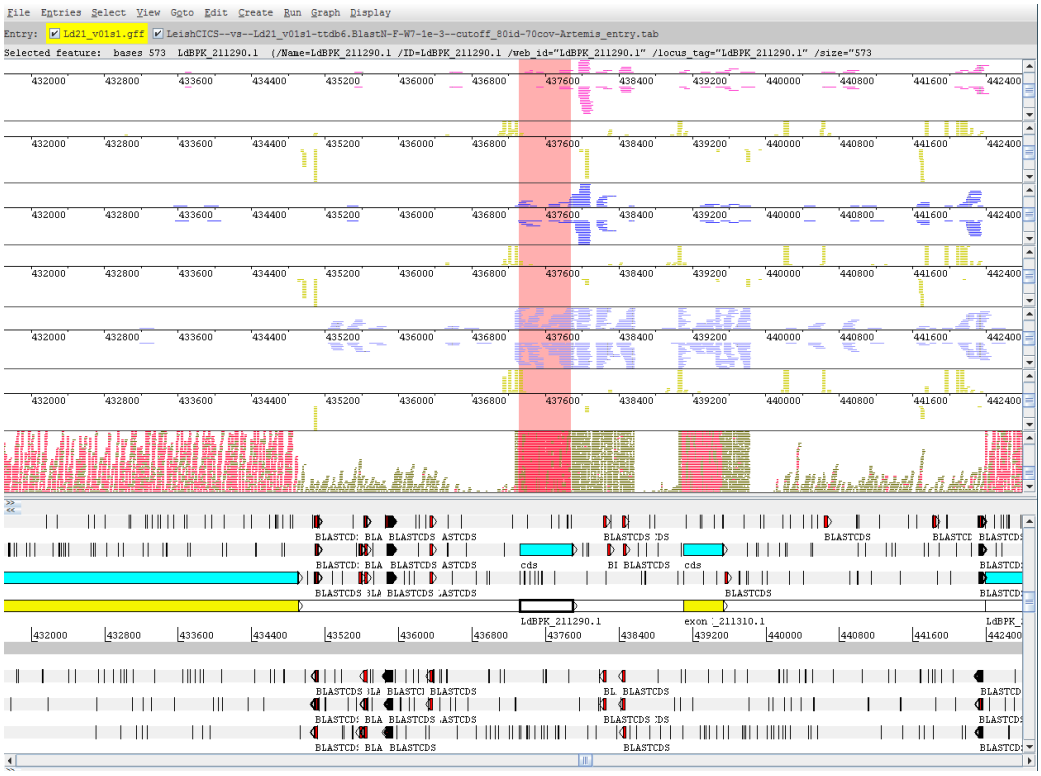
maneira, poucas *reads* restaram para cobrir outras regiões do genoma, o que levou a uma baixa cobertura do conjunto de transcritos do parasito.

O elevado número de *reads* mapeadas na região de rRNAs é frequentemente um problema em experimentos de RNA-seq devido a sua abundância [105]. Um trabalho de 2012, sobre RNA-seq comparativo de patógenos e hospedeiros, encontrou que aproximadamente 80% dos transcritos eram rRNAs, em bactérias e células eucarióticas. Ficou demonstrado que a depleção dos rRNAs é essencial para aumentar o conteúdo, de fato informativo, das amostras [106]. Um trabalho recente que utilizou RNPs e RNA-seq para identificar os ncRNAs em *Trypanosoma brucei* encontrou grande abundância de rRNAs em suas amostras (28% do total de *reads*) [79]. Assim, ainda que tenha sido idealizada uma estratégia para evitar as frações polirribossômicas, a contaminação com rRNAs foi alta, provavelmente em decorrência da contaminação com ribossomos livres, e um novo protocolo de preparação das bibliotecas foi estabelecido.

2.3.1.1.1 Identificação e caracterização dos possíveis ncRNAs

Apesar da baixa cobertura, a primeira versão do sequenciamento apresentou resultados suficientes para análise do padrão dos transcritos e os resultados preliminares foram promissores (Figura 9).

Figura 9. Região exemplo do alinhamento das *reads* no cromossomo 21 de *L. donovani*.



Representação da região do gene LdBPK_211290 visualizada no programa Artemis. Os quadrados brancos, azuis e amarelos representam as CDSs anotadas e a fita cinza clara com traços verticais pretos (*stop códons*) representam as 6 *frames* de leitura (3 acima – diretas e 3 abaixo - reversas). Os quadrados vermelhos representam *hits* BLAST contra o banco de dados de CICS e ESTs (detalhados na seção 3.6). As linhas horizontais empilhadas, verdes e vermelhas, representam as reads dos transcritos sequenciados para RNA total e *Ribosome profilling,* respectivamente. Na parte superior da figura estão representadas as *reads* (não sendo fita específico) dos transcritos dos três estágios de desenvolvimento: fase log (procíclico) (linhas horizontais empilhadas, azul claro), fase estacionária (linhas horizontais empilhadas, azul escuro) e amastigota (linhas horizontais empilhadas, rosa). Para cada estágio de desenvolvimento, estão representados os transcritos obtidos com o sequenciamento de sítios de SL (linhas horizontais empilhadas, amarelas). Fonte: autoral

A combinação dos dados de RNA-seq gerados no laboratório, como descrito acima, usando as regiões com densidade mais alta de *reads*, com os dados das bibliotecas construídas a partir de RNA total, de *ribosome profiling*, e das regiões 5'UTR (sítios de SL) (essas disponibilizadas por P. Myler), permitiram selecionar 37 regiões para caracterização/identificação de potenciais ncRNAs (Tabela 02). As informações foram obtidas por meio de verificação manual de cada região destacada pelo HTSeq, sendo que a definição de 5'UTR e 3'UTR foram definidas pela sobreposição entre a biblioteca de RNA total e *ribosome profiling*.

Tabela 2. Regiões com alta densidade de *reads*.

ncRNA ID	Tamanho predito (nt)	Coordenada inicial	Coordenada final	Anotação - *Ribosome profiling* versus RNA total	Gene	Presença de CICS	Presença de ESTs	Presença de SL
Ld_ncRNA1	190	519680	519869	3' UTR	LdBPK_061340.1_LdBPK_061350.1	-	ok	ok
Ld_ncRNA2	586	520139	520724	5' UTR	LdBPK_061350.1_LdBPK_061360.1	-	ok	ok
Ld_ncRNA3	402	527740	528141	3' UTR	LdBPK_061360.1	-	ok	ok
Ld_ncRNA4	280	272530	272809	5' UTR	LdBPK_080650.1_1	-	-	ok
Ld_ncRNA5	235	275572	275806	3' UTR	LdBPK_080650.1_2	ok	-	ok
Ld_ncRNA6	415	46465	46879	intergênico	LdBPK_100120.1_LdBPK_100130.1_1	ok	-	ok
Ld_ncRNA7	145	47446	47590	intergênico	LdBPK_100120.1_LdBPK_100130.1_2	-	-	-
Ld_ncRNA8	334	48274	48607	5' UTR	LdBPK_100130.1_1	-	-	ok
Ld_ncRNA9	235	50856	51090	3' UTR	LdBPK_100130.1_2	-	-	ok
Ld_ncRNA10	100	51531	51630	3' UTR	LdBPK_100130.1_3	ok	-	-
Ld_ncRNA11	505	52377	52881	3' UTR	LdBPK_100130.1_4	ok	-	-
Ld_ncRNA12	410	545959	546368	3' UTR	LdBPK_101440.1	ok	-	ok
Ld_ncRNA13	154	452039	452192	5' UTR	LdBPK_151130.1	ok	-	ok
Ld_ncRNA14	343	449718	450060	3' UTR	LdBPK_151150.1	ok	-	ok
Ld_ncRNA15	163	7750	7912	3' UTR	LdBPK_200030.1	-	-	-
Ld_ncRNA16	199	593912	594110	3' UTR	LdBPK_201320.1_1	-	-	ok
Ld_ncRNA17	298	594668	594965	3' UTR	LdBPK_201320.1_2	-	-	ok
Ld_ncRNA18	127	595307	595433	intergênico	LdBPK_201320.1_LdBPK_201330.1_1	-	-	-
Ld_ncRNA19	325	596090	596414	intergênico	LdBPK_201320.1_LdBPK_201330.1_2	-	-	-
Ld_ncRNA20	126	384833	384958	3' UTR	LdBPK_211160.1	ok	-	-
Ld_ncRNA21	64	385245	385308	5' UTR	LdBPK_211170.1	ok	-	ok
Ld_ncRNA22	667	437909	438575	3' UTR	LdBPK_211290.1	ok	-	ok
Ld_ncRNA23	136	746275	746410	3' UTR	LdBPK_212160.1	-	-	-
Ld_ncRNA24	250	277300	277550	3' UTR	LdBPK_300920.1	ok	-	ok
Ld_ncRNA25	334	821517	821850	5' UTR	LdBPK_302200.1	-	ok	ok
Ld_ncRNA26	280	814282	814561	3' UTR	LdBPK_302210.1_3	-	-	-
Ld_ncRNA27	253	814939	815191	3' UTR	LdBPK_302210.1_4	-	-	-
Ld_ncRNA28	388	815470	815857	3' UTR	LdBPK_302210.1_5	-	-	ok
Ld_ncRNA29	550	816136	816685	3' UTR	LdBPK_302210.1_6	ok	-	ok
Ld_ncRNA30	190	817603	817792	3' UTR	LdBPK_302220.1	-	-	-
Ld_ncRNA31	116	4758	4873	3' UTR	LdBPK_340010.1	ok	ok	
Ld_ncRNA32	100	1431069	1431168	3' UTR	LdBPK_343430.1_1	-	-	ok
Ld_ncRNA33	83	1433230	1433312	3' UTR	LdBPK_343430.1_2	-	-	-
Ld_ncRNA34	208	1435056	1435263	3' UTR	LdBPK_343440.1	-	-	-
Ld_ncRNA35	141	914907	915047	5' UTR	LdBPK_352270.1_1	-	-	ok

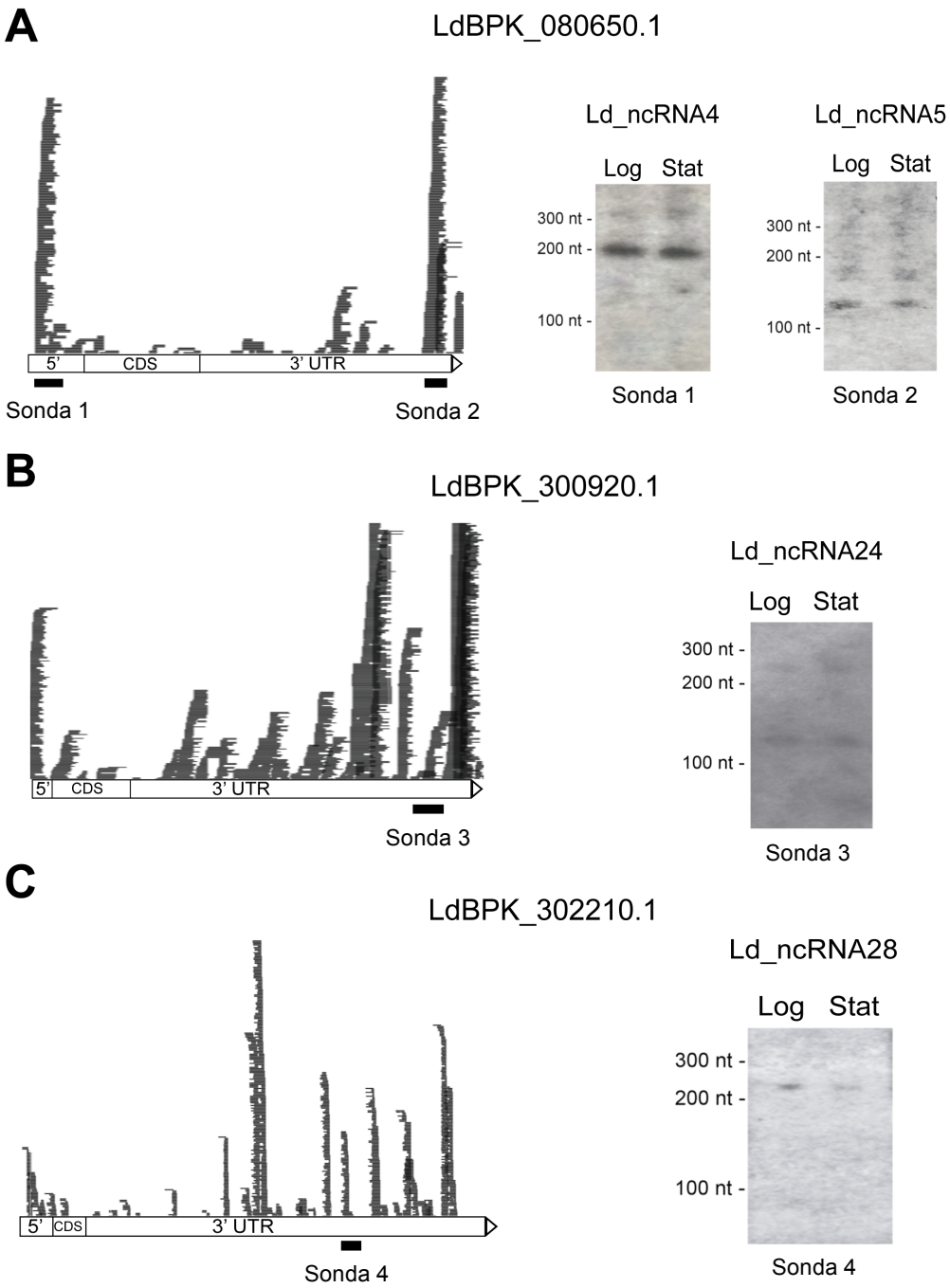
Ld_ncRNA36	330	916035	916364	3' UTR	LdBPK_352270.1_2	ok	-	-
Ld_ncRNA37	113	1990550	1990662	intergênico	LdBPK_355080.1_LdBPK_355090.1	ok	-	ok

Informações obtidas analisando manualmente as regiões destacadas como tendo maior densidade de *reads*. Os IDs (identificadores) destacados em negrito indicam as regiões selecionadas para verificação experimental.

Para comprovar a existência dos pequenos transcritos preditos pelo RNA-seq das 37 regiões com alta densidade de *reads*, 5 foram escolhidas para análise experimental (tabela 02 – IDs em negrito), via *Northern blotting*. Os critérios para a escolha incluíram a presença das seguintes caraterísticas na mesma região das *reads* mapeadas: CICS ou EST (*expressed sequence tag*), sítio de SL e localização na 3'UTR ou 5'UTR.

Na figura 10 estão apresentadas quatro regiões-exemplo com potenciais pequenos ncRNAs. Os primeiros experimentos de *Northern blotting* foram feitos e 4 transcritos pequenos foram confirmados com o tamanho similar ao predito pelo sequenciamento. Pela densidade de leituras, é possível sugerir que esses transcritos existam como pequenos RNAs, independentes dos mRNAs dos quais parecem emergir (Figura 10). Os quatro exemplos já analisados por *northern blotting* indicam que, neste caso, o nível de expressão nos diferentes estágios analisados é similar; exceção feita ao gene LdBPK_302210, que parece estar presente em níveis mais altos na fase logarítmica (Figura 10-C); resultado a ser confirmado pela repetição do experimento.

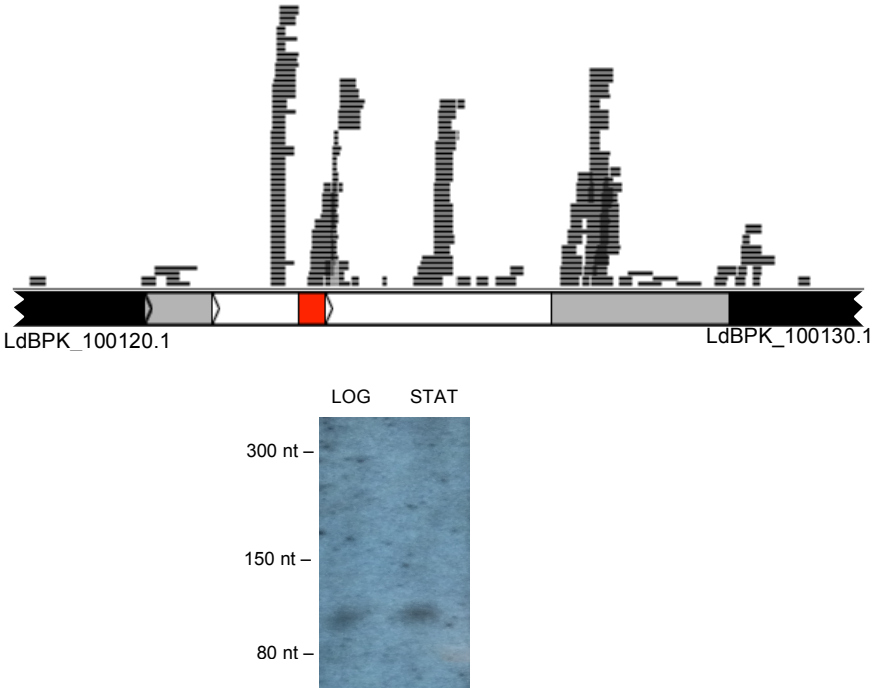
Figura 10. Exemplos de ncRNAs putativos derivados de UTRs em *L. donovani*.



Frações de RNA não polissômicas de promastigotas de *L. donovani* em fase log e estacionária foram extraídas e bibliotecas de RNA-seq foram construídas. As imagens foram obtidas no Artemis. Paineis A-C: Representação em escala de três *loci* diferentes: A) LdBPK_080650.1, B) LdBPK_300920.1 e C) LdBPK_302210.1. As caixas pretas indicam a região alvo da sonda desenvolvida e utilizada nos experimentos de *Northern blotting*. O *Northern blotting* foi feito utilizando RNA extraído de diferentes estágios de desenvolvimento de *L. donovani* e hibridizados com as sondas correspondentes (sondas disponíveis em [82]). O RNA foi fracionado em gel de poliacrilamida 8% e 7M de uréia. LOG representa fase logarítmica (3° dia) e STAT representa fase estacionária (6° dia) de culturas de promastigotas axênicos. Fonte: autoral

Além desses exemplos de ncRNAs putativos em UTRs de genes anotados, um candidato de região intergênica também apresentou resultado positivo, quando avaliado por *northern blotting* (Figura 11).

Figura 11. ncRNA putativo intergênico.



A fração de RNA não polissômica de promastigotas de *L. donovani* em fase log e estacionária de culturas axênicas foi extraída. As bibliotecas de RNA-seq foram contruídas e processadas conforme a seção 2.2.2. As imagens foram geradas com o programa Artemis [86]. Representação em escala da região intergênica entre os genes LdBPK_100120.1 e LdBPK_100130.1. As caixas pretas representam as CDS, as caixas cinzas indicam as UTRs, as linhas pretas empilhadas representam as reads mapeadas e a caixa vermelha indica o ncRNA putativo intergênico. O *Northern blotting* foi feito utilizando RNA extraído de diferentes estágios de desenvolvimento de *L. donovani* e hibridizados com a sonda descrita em [82]. RNAs fracionados em gel de poliacrilamida 8% e 7M de uréia. LOG representa fase logarítmica (3° dia) e STAT representa fase estacionária (6° dia) de culturas de promastigotas axênicos. Fonte: autoral

Os 37 ncRNAs identificados foram caracterizados por diversos parâmetros. Além de aspectos básicos relacionados à descoberta de ncRNAs como parâmetros de mapeamento de *reads*, posição no genoma, tamanho da sequência, presença de sítios de SL, e conservação também foram realizadas predições computacionais de algumas características. Dentre as predições já realizadas está a de estrutura secundária (programa RNAcon); que identificou 30 sequências como possíveis ncRNAs, com estruturas similares a introns catalíticos, rRNAs, SRPs (*Signal*

Recognition Particle) e IRES (*Internal Ribosome Entry Site*). Outra predição que considera bancos de dados de domínios de ncRNAs conhecidos (programa RNAspace) identificou domínios presentes em 9 regiões, entre eles: estruturas tRNA-símile; telomerase-símile; RnaseP (Ribonuclease P); snoRNA (*small nucleolar* RNA); SRP e IRES. A terceira predição, que avalia o potencial codificador de uma sequência (programa PORTRAIT), identificou 27 regiões preditas como tendo mais de 50% de probabilidade de serem ncRNAs. O programa snoscan encontrou somente 1 candidato como possível snoRNA (Tabela 3).

Tabela 3. Caracterização dos ncRNAs identificados.

ncRNA ID	PORTRAIT (%)	RNASpace	RNAcon	snoscan
Ld_ncRNA1	87.41	-	-	-
Ld_ncRNA2	30.98	-	-	-
Ld_ncRNA3	71.62	-	catalytic intron	-
Ld_ncRNA4	11.67	IRES	-	-
Ld_ncRNA5	85.68	-	catalytic intron	-
Ld_ncRNA6	15.50	-	catalytic intron	-
Ld_ncRNA7	70.56	-	rRNA	-
Ld_ncRNA8	41.85	-	catalytic intron	-
Ld_ncRNA9	72.33	-	IRES	-
Ld_ncRNA10	87.03	-	rRNA	-
Ld_ncRNA11	50.00	-	-	-
Ld_ncRNA12	94.41	-	catalytic intron	-
Ld_ncRNA13	84.39	-	rRNA	-
Ld_ncRNA14	21.89	-	catalytic intron	-
Ld_ncRNA15	89.43	-	rRNA	-
Ld_ncRNA16	82.51	IRES	IRES	-
Ld_ncRNA17	94.67	-	catalytic intron	-
Ld_ncRNA18	93.53	snoRNA	rRNA	-
Ld_ncRNA19	92.00	-	IRES	-
Ld_ncRNA20	97.09	-	rRNA	-
Ld_ncRNA21	-	snoRNA	catalytic intron	-
Ld_ncRNA22	8.10	-	rRNA	-
Ld_ncRNA23	95.24	-	rRNA	-
Ld_ncRNA24	38.61	tRNA-like structure	catalytic intron	ok
Ld_ncRNA25	10.95	-	-	-
Ld_ncRNA26	81.61	-	catalytic intron	-
Ld_ncRNA27	62.33	Telomerase-like	-	-
Ld_ncRNA28	75.79	-	rRNA	-
Ld_ncRNA29	59.30	Rnase P	IRES	-
Ld_ncRNA30	25.51	-	signal recognition particle	-
Ld_ncRNA31	84.72	-	rRNA	-

Ld_ncRNA32	79.93	-	rRNA	-
Ld_ncRNA33	91.63	-	catalytic intron	-
Ld_ncRNA34	86.89	-	-	-
Ld_ncRNA35	65.78	signal recognition particle	rRNA	-
Ld_ncRNA36	66.09	IRES	catalytic intron	-
Ld_ncRNA37	93.67	-	rRNA	-

Também foram feitas buscas em bancos de dados de ncRNAs conhecidos: fRNAdb, NONCODE, lncRNAdb. Nenhuma das sequências analisadas tiveram similaridade com estes bancos de dados, o que pode dever-se a prováveis diferença de sequência primária entre os ncRNAs de *Leishmania* e os dos organismos disponíveis nos bancos de dados. A combinação dessas diferentes predições trazem maior confiabilidade às regiões de mapeamento das *reads* e para a análise de prospecção dos dados.

2.3.1.1.2 Análise de correlação entre as réplicas – versão 1

Os dados da primeira versão do sequenciamento da fração citoplasmática foram utilizados na análise de expressão diferencial. A análise se inicia com a distribuição da contagem bruta de *reads*. Foram observadas grande diferença (variação) no números *reads* entre os estágios de desenvolvimento e entre as réplicas de cada estágio.

Outro ponto investigado foi a semelhança/diferença entre as réplicas utilizando a distância BCV em um gráfico MDS. A análise demonstrou uma diferença maior entre as réplicas de um mesmo estágio de desenvolvimento do que entre os estágios analisados. Por esse motivo, a análise de diferença de expressão não foi realizada, já que o resultado encontrado não seria informativo.

2.3.1.1.3 Conservação dos ncRNAs identificados

Um dos indicativos de que candidatos a ncRNAs possuam papel funcional é a conservação de sequência entre diferentes subgêneros de *Leishmania*. Por isso, essa característica também foi considerada na análise do ncRNAs putativos, sendo que encontramos esses candidatos inteira ou parcialmente conservados em pelo menos 3 espécies de *Leishmania* (Tabela 4).

Tabela 4. Grau de conservação de sequência dos ncRNAs putativos em diferentes espécies de tripanossomatideos.

ncRNA ID	*L. major*	*L. infantum*	*L. amazonensis*	*L. braziliensis*	*L. tarentolae*	*L. enriettii*	*T. brucei*
Ld_ncRNA1	ok	ok	ok	ok	ok	-	-
Ld_ncRNA2	ok	ok	ok	(204/586)	(223/586)	(55/586)	-
Ld_ncRNA3	ok	ok	ok	(234/402)	(377/402)	-	-
Ld_ncRNA4	ok	ok	ok	-	-	-	-
Ld_ncRNA5	ok	ok	ok	(50/235)	(46/235)	(45/235)	(45/235)
Ld_ncRNA6	ok	ok	ok	(46/415)	(58/415)	(47/415)	(53/415)
Ld_ncRNA7	ok	ok	ok	-	-	-	-
Ld_ncRNA8	ok	ok	ok	-	(294/334)	-	-
Ld_ncRNA9	ok	ok	ok	(32/235)	(138/235)	(34/235)	-
Ld_ncRNA10	ok	ok	ok	-	(58/100)	-	-
Ld_ncRNA11	(255/505)	ok	ok	-	(122/505)	(39/505)	-
Ld_ncRNA12	ok	ok	ok	(270/410)	(251/410)	-	-
Ld_ncRNA13	ok	ok	-	-	(57/154)	-	-
Ld_ncRNA14	(253/343)	ok	ok	(44/343)	(95/343)	(63/343)	(28/343)
Ld_ncRNA15	ok	ok	ok	-	-	-	-
Ld_ncRNA16	ok	ok	ok	(44/199)	(86/199)	(46/199)	-
Ld_ncRNA17	ok	ok	ok	(79/298)	(256/298)	-	-
Ld_ncRNA18	(103/127)	ok	ok	-	-	-	-
Ld_ncRNA19	ok	ok	ok	(63/325)	ok	-	-
Ld_ncRNA20	ok	ok	ok	ok	ok	-	-
Ld_ncRNA21	ok	ok	ok	(39/64)	(40/64)	-	-
Ld_ncRNA22	ok	ok	ok	(160/667)	(406/667)	-	-
Ld_ncRNA23	ok	ok	ok	-	-	-	-
Ld_ncRNA24	ok	ok	ok	-	(157/250)	-	-
Ld_ncRNA25	ok	ok	ok	-	-	-	-
Ld_ncRNA26	(240/280)	ok	ok	-	-	-	-
Ld_ncRNA27	ok	ok	ok	-	(66/253)	-	-
Ld_ncRNA28	ok	ok	ok	-	(280/388)	-	-
Ld_ncRNA29	(517/550)	ok	ok	(36/550)	(50/550)	(41/550)	(32/550)
Ld_ncRNA30	ok	ok	ok	(97/190)	(111/190)	-	-
Ld_ncRNA31	ok	ok	ok	(53/116)	(68/116)	(44/116)	-
Ld_ncRNA32	ok	ok	ok	-	-	-	-
Ld_ncRNA33	ok	ok	ok	-	-	-	-
Ld_ncRNA34	ok	ok	ok	-	(112/208)	-	-
Ld_ncRNA35	ok	ok	ok	(92/141)	(90/141)	-	-
Ld_ncRNA36	ok	ok	ok	-	(205/330)	(31/330)	-

"ok" indica conservação de sequência (até 20 nucleotídeos de diferença), "-" indica que nenhuma *hit* foi encontrado. Para similaridades parciais, a notação indica o número de nucleotídeos similares. Como exemplo: (165/197) significa que 165 nucleotídeos de uma sequência de *L. donovani* são similares ao total de 197 nucleotídeos da sequência de referência.

A conservação desses transcritos em diferentes subgêneros de *Leishmania* é sugestivo de pressão seletiva e consequentemente de potencial papel funcional. A conservação observada confere mais robustez à hipótese de que os ncRNA identificados computacionalmente sejam de fato ncRNAs com papeis regulatórios.

2.3.1.2 RNA-seq versão 2

Várias modificações foram inseridas nessa 'versão 2 do RNA-seq'. Para diminuir a quantidade de *reads* sequenciadas de rRNAs, o protocolo de preparação das bibliotecas foi modificado (Seção 2.2.3). Nessa direção, optamos por trabalhar somente com a fração citoplasmática, fazendo a comparação entre estágios de desenvolvimento.

Nesta segunda versão do sequenciamento, ao invés de empregar culturas de fase logarítmica e estacionária como fonte dos RNPs foram utilizados os três estágios de desenvolvimento de *L. donovani* em cultura axênica: promastigotas procíclicos, promastigotas metacíclicos e amastigotas [81]. Foram obtidas 8.438.954 de *reads*.

O alinhamento das *reads* com o genoma de referência de *L. donovani* indicou uma diminuição do número de *reads* mapeadas na região dos genes de rRNAs quando comparado com a primeira versão do sequenciamento. Apesar do sucesso na diminuição das *reads* de rRNAs, o mapeamento revelou um baixo número de *reads* mapeadas para as bibliotecas da fase procíclica (~19%) e metacíclica (~15%). Por essa razão, foi realizada uma análise para verificação de possível contaminação durante a preparação das bibliotecas ou durante o sequenciamento.

Essa análise revelou um grande número de *reads* de DNA mitocondrial representadas nas duas bibliotecas com baixo número de *reads* mapeadas no genoma nuclear do parasito (aproximadamente 90% dos *matches* encontrados). Possivelmente, esse cenário foi decorrência das condições de fracionamento celular empregadas para recuperação das RNPs-não polissômicas. Como resultado, a cobertura do transcriptoma ficou comprometida, sendo necessária uma nova alternativa para preparação das bibliotecas.

3. ANÁLISE DO RNA-SEQ DE *Leishmania braziliensis*

3.1 VISÃO GERAL

Considerando os problemas e soluções encontrados na preparação das bibliotecas de *L. donovani* e o tempo decorrido, decidimos utilizar inicialmente bibliotecas de RNA total sequenciadas em NextSeq-Illumina para aumentar a quantidade de dados gerados. Decidimos utilizar *Leishmania braziliensis* nesta nova etapa pois nesse período, Natália M. Teles (aluna de doutorado no laboratório) otimizou as condições de purificação de promastigotas metacíclicos com uma linhagem de *L. braziliensis* empregando o protocolo de colchão de ficoll e também teve sucesso na obtenção de amastigotas axênicos de forma reprodutível (Seção 3.2.1.2). A troca da espécie do parasito, de *L. donovani* para *L. braziliensis*, se deu por sua importância médica no Brasil, e particularmente em função das diferenças genéticas que incluem a existência de maquinaria de RNAi, ausentes na primeira, que permite usar ferramentas mais versáteis de genética reversa para as análises funcionais a serem conduzidas posteriormente. A nova metodologia foi testada em experimentos piloto empregando um kit comercial (*ribo-zero epidemiology* - Illumina) para exclusão dos rRNAs da população a ser sequenciada e usando um sequenciador de capacidade mais baixa (e econômica) o MiSeq-Illumina, como descrito a seguir.

3.2 MATERIAIS E MÉTODOS

Todos os arquivos de resultados citados estão disponíveis sob solicitação (akcruz@fmrp.usp.br).

3.2.1 Preparação das bibliotecas para identificação de ncRNAs

3.2.1.1 Fracionamento

A aluna de doutorado Natália M. Teles avaliou diferentes abordagens para seleção de frações específicas dos RNAs e deleção dos rRNAs das amostras. No primeiro teste foi avaliado o kit comercial (*mirVana miRNA Isolation*) que tem a finalidade de selecionar uma fração altamente enriquecida de RNAs pequenos (≤200 nucleotídeos), e que também permite a recuperação dos RNAs acima de 200 nt. Dessa forma, o teste foi realizado com duas amostras, uma de RNA total de

Leishmania major já preparada (Felipe F. Castro, aluno de doutorado no laboratório) e enriquecida de RNAs pequenos e uma amostra de *Leishmania braziliensis* enriquecida de RNAs maiores (acima de 200 nt – Natália M. Teles). No intuito de minimizar a quantidade de rRNAs nas amostras de RNA>200 e RNA<200 foi avaliada a eficiência de depleção de rRNA pelo kit *ribo-zero epidemiology* (Illumina). Embora seja específico para depleção de rRNAs de bactérias tivemos a informação que seria eficiente para depleção dos rRNAs de *Leishmania* (comunicação pessoal de S. M. Beverley). Após a depleção do rRNA, bibliotecas das amostras de RNA<200 e RNA>200 foram construídas e validadas utilizando o kit *TruSeq Stranded Total RNA*, seguindo as instruções do fabricante (Bibliotecas geradas e sequenciadas em parceria com o Centro de Medicina Genômica - HCFMRP).

3.2.1.2 RNA-total

O RNA total de cada um dos três estágios de vida de *L. braziliensis* foi extraído pelo kit de extração *Direct-zolTM* (*Zymo Research*). O experimento foi realizado em triplicata biológica e as amostras de RNA foram avaliadas quanto à quantidade e qualidade, e posteriormente foram submetidas a tratamento com o kit *ribo-zero epidemiology* (Illumina) para depleção de rRNA. A construção das bibliotecas das 9 amostras foi feita com o kit TruSeq® Stranded Total RNA (Bibliotecas geradas e sequenciadas em parceria com o Centro de Medicina Genômica - HCFMRP).

3.2.2 Processamento das bibliotecas

O processamento das bibliotecas foi similar ao utilizado anteriormente (Seção 2.2.4), ou seja, utilizamos cutadapt e FastQC para remoção de adaptadores e verificação da qualidade, respectivamente. As *reads* foram mapeadas com o genoma de referência de *L. braziliensis* linhagem 2903 (MHOM/BR/75/M2903) - versão 30 (TriTrypDB), utilizando o programa Bowtie2. Os parâmetros utilizados no alinhamento foram *–N 1* que permite apenas 1 *mismatch* no alinhamento e *--local* que não exige que a *read* alinhe de ponta a ponta. O programa Picard (https://broadinstitute.github.io/picard/) foi utilizado para converter arquivos sam para bam, além do agrupamento, ordenação e indexação desses arquivos. Os mapeamentos foram visualizados e analisados com o programa IGV [107].

3.2.3 Correlação entre as réplicas

A primeira etapa da análise foi a contagem bruta do número de *reads* das 9 bibliotecas sequenciadas com o programa featureCounts [108]. Foram utilizados o gráfico MDS (*Multidimensional Scaling*) e a distância BCV (*Biological Coefficient Variation*) como explicado na seção 2.2.5.

3.2.4 Análise de expressão diferencial

Para a análise da expressão diferencial foi utilizado o ambiente R [87] com a biblioteca DESeq do Bioconductor [88; 109]. Essa biblioteca é baseada em um modelo de distribuição binomial negativa para identificar expressão diferencial. Foi desenvolvido um *script* R (disponível sob solicitação - akcruz@fmrp.usp.br) que faz a leitura de um arquivo com a contagem do número de *reads* por região (contagem realizada com featureCounts), remove previamente os genes com contagem de *reads* igual a zero e realiza a normalização e a geração dos gráficos. Foram considerados diferencialmente expressos os genes com um p-valor ajustado < 0,05.

3.2.4.1 Análise dos genes espécie-específicos de *L. braziliensis*

Foram identificados em 2007 os genes espécie-específicos de *L. braziliensis* em comparação com *L. major* e *L. infantum* [41]. Para identificar a presença desses genes entre aqueles diferencialmente expressos em cada um dos estágios de desenvolvimento, foi necessária a atualização dos IDs dos genes disponibilizados pelos autores. De posse da lista de IDs, foi feita a busca de suas sequências no NCBI (*National Center for Biotechnology Information*). Essas sequências de nucleotídeos formaram o banco de dados para a busca de similaridade (BLAST). Os *hits* dos 49 genes espécie-específicos com as CDSs anotadas de *L. braziliensis* (linhagem 2903 – TriTrypDB – versão 30) foram analisados manualmente.

3.2.5 Enriquecimento *Gene Ontology*

Após a análise de expressão diferencial, foi feita a análise de enriquecimento de termos *Gene Ontology* (GO). O projeto GO tem como objetivo padronizar a representação dos genes e produtos gênicos (http://www.geneontology.org/). O Projeto GO desenvolveu três vocabulários estruturados e controlados (ontologias) que descrevem os produtos gênicos em termos de seus processos biológicos,

componentes celulares e funções moleculares de maneira independente da espécie e em diferentes níveis, dependendo da profundidade de conhecimento requerida.

Na análise de enriquecimento foi criado um *script* R (disponível sob solicitação - akcruz@fmrp.usp.br) que utiliza a biblioteca GOseq do pacote Bioconductor [110]. Os conjuntos de genes diferencialmente expressos foram analisados separadamente, de acordo com a comparação de dois estágios de desenvolvimento de *L. braziliensis* (promastigota procíclica versus promastigota metacíclica, promastigota metacíclica versus amastigota e amastigota versus promastigota procíclica) e da regulação positiva ou negativa. Dessa maneira foram gerados três conjuntos de genes regulados positivamente e três regulados negativamente. Um limite de p-valor < 0,05 foi utilizado.

3.2.6 Determinação dos limites das UTRs

A informação sobre tamanho das regiões de 5'UTR e 3'UTR para as CDSs de *L. braziliensis* não está disponível em bancos de dados de domínio público. Como não há dados de *ribosome profilling* para *L. braziliensis*, outra metodologia teve que ser desenvolvida para identificar essas regiões. Para auxiliar na definição dos limites das UTRs foram consideradas as coordenadas das CDSs anotadas no arquivo gff (*General Feature Format*) de *L. braziliensis* versão 30 (TriTrypDB). Foram incluídas as UTRs com tamanho máximo de 7.500 nt, seguindo a metodologia desenvolvida por Dillon e colaboradores [56].

3.2.6.1 5'UTR

Para a definição do tamanho das 5'UTR, utilizamos a biblioteca construída para identificar o sítio de SL em promastigotas procíclicas de *L. braziliensis*, descrita na seção 2.2.8 (colaboração de Peter Myler). O mapeamento foi realizado com Bowtie2 utilizando o parâmetro *–N 1* que permite apenas 1 *mismatch* no alinhamento. O arquivo de mapeamento bam foi gerado com o programa Picard e logo após foi feita a contagem da cobertura por nucleotídeo com o programa genomecoverageBed do pacote de ferramentas BEDTools [111].

Foi desenvolvido um *script* Perl para considerar como possível sítio de SL regiões com cobertura mínima de 10 *reads*. Depois de identificados os possíveis

sítios SL, outro *script* Perl interligou cada CDS ao sítio de SL *upstream* mais próximo e com maior cobertura (sítio principal), definindo assim a região de 5'UTR.

3.2.6.2 3'UTR

Para a definição das 3'UTRs, executamos novamente o mapeamento sem o parâmetro *--local*, impedindo que a *read* fosse mapeada caso alguma extremidade não fosse similar à referência. Dessa maneira, foram selecionadas as *reads* não mapeadas, formando um conjunto de *reads* candidatas a terem cauda PoliA. Agrupamos este resultado de todas as 9 bibliotecas sequenciadas para que um maior número de *reads* com cauda poliA fosse considerado. Foi realizada a seleção e *trimagem* de sequências com no mínimo quatro "A"s seguidos utilizando o programa cutadapt. Essas *reads* selecionadas foram novamente mapeadas no genoma de referência depois da remoção da sequência de "As". Metodologia similar foi utilizada por Dillon e colaboradores para *L. major* [56].

O arquivo de mapeamento bam foi gerado com o programa Picard e logo após foi feita a contagem da cobertura por nucleotídeo com o programa genomecoverageBed do pacote de ferramentas BEDTools. Foi desenvolvido um *script* Perl para definir como possível sítio de poliadenilação regiões com cobertura mínima de 10 *reads*. Depois de identificadas essas regiões, outro *script* Perl foi criado para interligar cada CDS ao sítio de poliadenilação *downstream* mais próximo e com maior cobertura (sítio principal), definindo assim a região de 3'UTR.

3.3 RESULTADOS E DISCUSSÃO

3.3.1 Teste MiSeq versão 1

O primeiro teste foi feito utilizando Ribo-Zero na preparação de duas amostras de RNA: *Leishmania major* enriquecido com RNAs<200nt e *L. braziliensis* enriquecidos com RNAs>200nt. O número de *reads* esperado era de aproximadamente 25 milhões mas a quantidade de clusters que passaram no filtro de pureza de sinal (Illumina) foi baixa, por isso, somente 13 milhões de *reads* tiveram a etapa de chamada de bases. Assim, o sequenciamento resultou em somente 13 milhões de *reads*, sendo 43% da biblioteca enriquecida em RNAs menores que 200nt e 55% da biblioteca enriquecida com RNAs maiores que 200nt.

As *reads* resultantes apresentaram boa qualidade, sendo a maioria com qualidade acima de 30 (Figura 12).

Figura 12. Porcentagem de *reads* com qualidade maior que 30.

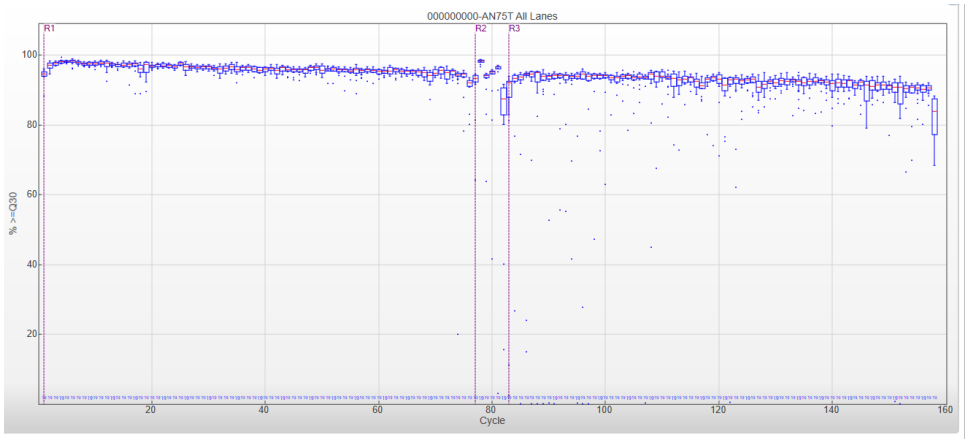
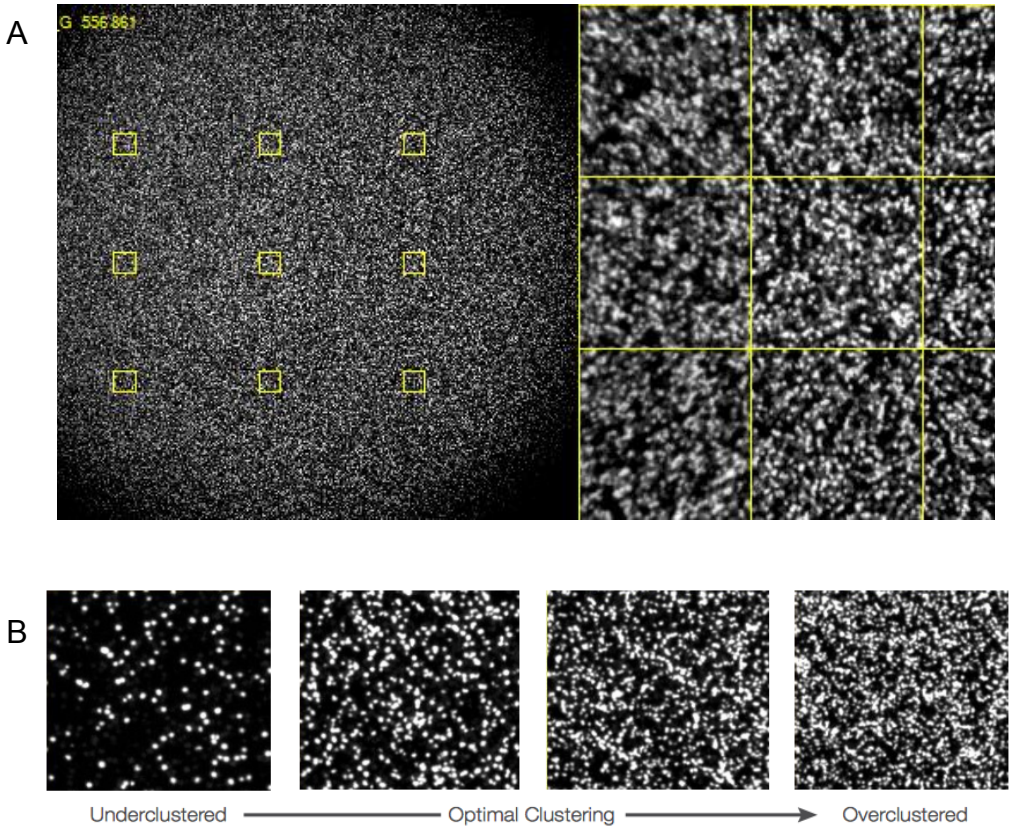


Figura gerada pelo programa SAV (*Sequencing Analysis Viewer*) – Illumina. Fonte: autoral

O baixo número de clusters que passaram pelo filtro de pureza de sinal poderia ter-se originado da sobreposição de clusters. Examinamos as imagens dos clusters gerados e constatamos o *overclustering* da *flow cell* (http://support.illumina.com/content/dam/illumina-marketing/documents/products/other/miseq-overclustering-primer-770-2014-038.pdf). Apresentamos as imagens dos clusters gerados durante o sequenciamento (Figura 13-A) em comparação com a referência disponibilizada pela Illumina (Figura 13-B).

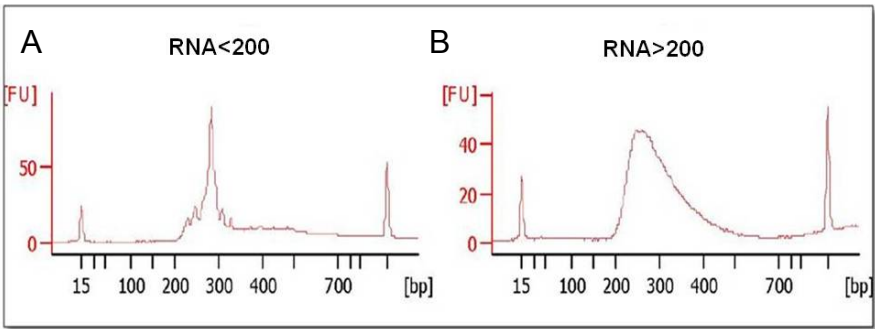
Figura 13. Imagem dos clusters gerados durante o sequenciamento.



A) Imagem da primeira versão do sequenciamento-teste feito no MiSeq. B) Imagem de referência utilizada pela Illumina. Fonte: autoral e de http://support.illumina.com/content/dam/illumina-marketing/documents/products/other/miseq-overclustering-primer-770-2014-038.pdf

O evento de *overclustering* pode ter ocorrido devido a dificuldade de quantificação da biblioteca enriquecida de RNAs<200nt, pois a quantificação por RT-qPCR considera o tamanho dos fragmentos e no gráfico do Bioanalyzer observamos a diferença entre as duas bibliotecas sequenciadas (Figura 14). Na biblioteca enriquecida de RNAs<200nt o padrão de distribuição de tamanho apresenta três picos proeminentes, o maior em torno de 280 pb e os menores em torno de 225 pb e 300 pb (Figura 14-A), já que a etapa de fragmentação não é realizada segundo o protocolo descrito pela Illumina. Sendo assim, a quantidade da amostra colocada no sequenciador deve ter sido subestimada.

Figura 14. Distribuição de tamanho da biblioteca de cDNA na etapa de validação.



Bibliotecas das amostras de *Leishmania*, construída após depleção de rRNAs por ribo-zero epidemiology. A) RNA<200: RNA menor que 200 nt; B) RNA>200: RNA maior que 200 nt (fração recuperada). Fonte: Natália M. Teles

As *reads* resultantes foram mapeadas em seus respectivos genomas de referência (*L. major* e *L. braziliensis*, ambos na versão 28 do TriTrypDB), sendo observada uma grande porcentagem de *reads* mapeadas nas duas amostras (Tabela 5).

Tabela 5. Informações da primeira versão do sequenciamento teste feito no MiSeq.

		LbrM2903		LmjF	
Biblioteca	Número de *paired-end reads*	*% reads* mapeadas	*% reads* cromossomo 6	*% reads* mapeadas	*% reads* cromossomo 27
RNA<200nt	11231754	-	-	97,52	95,56
RNA>200nt	14355126	97,57	11,60	-	-

O maior cluster de RNAs ribossômicos localiza-se no cromossomo 27 de *L. major* e, segundo consta no genoma anotado, no cromossomo 6 de *L. braziliensis*. Podemos observar que a biblioteca de *L. braziliensis* teve uma grande diminuição de *reads* mapeadas no cromossomo 6 (Tabela 03). A biblioteca de *L. braziliensis* estava enriquecida em RNAs>200nt e nela houve uma eficiente depleção dos rRNAs grandes (28S-beta, 28S-alpha e 18S). Na biblioteca de *L. major*, construída com a fração de RNA<200nt observamos a preponderância de rRNAs; 95% das *reads* mapeadas alinharam no cromossomo 27. Concluímos assim, que a depleção dos pequenos rRNAs (5S e 5.8S) é pouco eficiente, e confirma padrões de depleção obtidos em outros organismos (http://www.illumina.com/products/rrna-removal-kit-species-compatibility.html).

3.3.2 Teste MiSeq versão 2

Concluímos que trabalhar com qualquer das duas frações obtidas a partir da purificação com o kit mirVana implicaria em vieses indesejáveis; na fração RNAs<200nt temos uma abundância de pequenos rRNAs, que não são depletados, que inviabilizaria a detecção de outros RNAs pequenos e na fração dos RNAs>200nt, embora com depleção dos rRNAs, há uma sub-representação dos RNAs menores. Em virtude das dificuldades e vieses das abordagens avaliadas e do tempo decorrido, optamos por partir de amostras de RNA total do parasito, realizando as etapas de depleção de rRNA e fragmentação para construção da biblioteca. Sendo assim, utilizando Ribo-Zero, foram construídas três bibliotecas representativas do transcriptoma dos diferentes estágios de desenvolvimento de *L. braziliensis*: promastigota procíclica, promastigota metacíclica e amastigota, todos axênicos. Foram gerados aproximadamente 23 milhões de *reads* no Illumina-MiSeq, sendo a grande maioria com qualidade acima de 30 (Figura 15).

Figura 15. Porcentagem de *reads* com qualidade maior que 30.

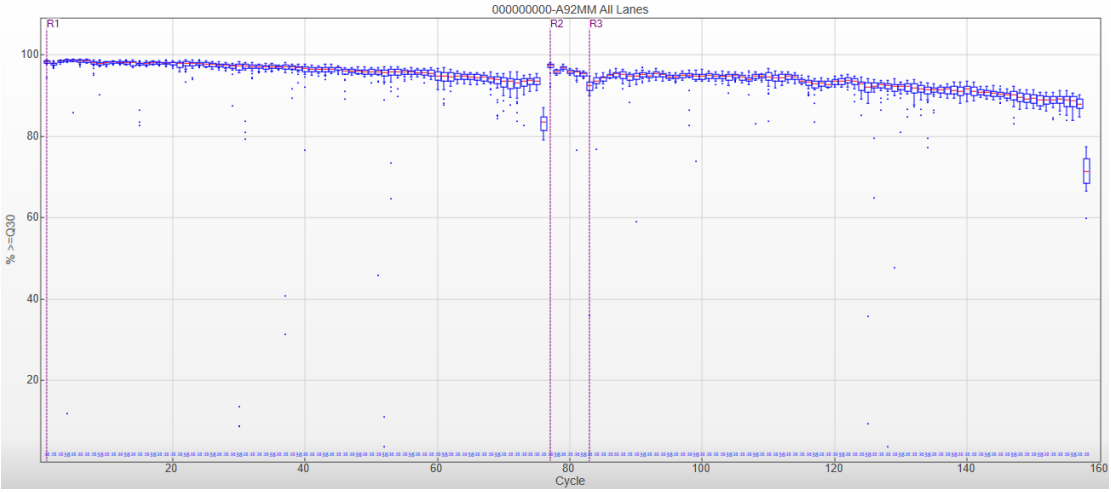


Figura gerada pelo programa SAV (*Sequencing Analysis Viewer*) – Illumina. Fonte: autoral

Foi alto o percentual (entre 94% e 97%) de *reads* mapeadas no genoma de referência (*Leishmania braziliensis* 2903 – versão 30, TriTrypDB) (Tabela 6). Os rRNAs representaram <31% do total de *reads* como estimado pela porcentagem de *reads* mapeadas no cromossomo 6. Dada a cobertura possível com a plataforma Illumina NextSeq® e a porcentagem de rRNAs nas bibliotecas concluímos que o

sequenciamento de 9 bibliotecas nessa plataforma geraria o número suficiente de *reads* para detecção de ncRNAs em *Leishmania*.

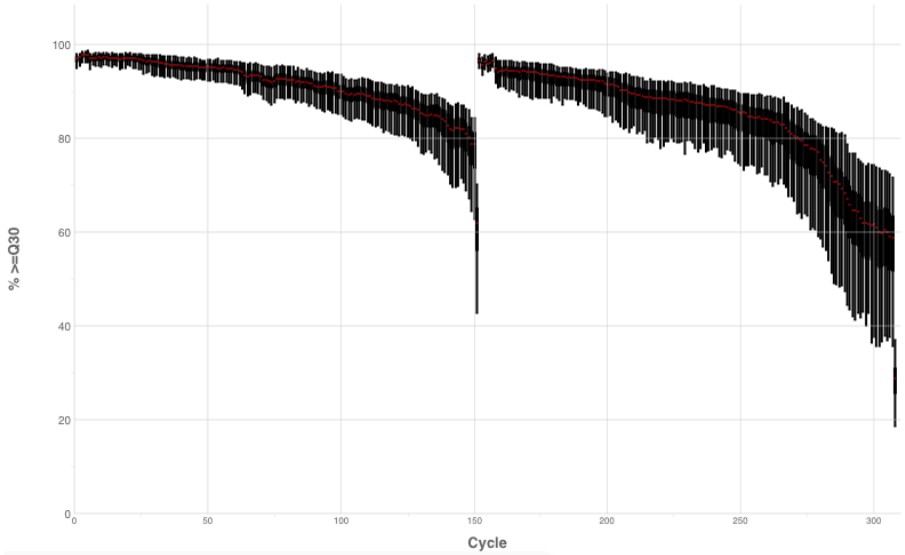
Tabela 6. Informações da segunda versão do sequenciamento teste feito no MiSeq.

Biblioteca	Número de *paired-end reads*	LbrM2903	
		% reads mapeadas	*% reads* cromossomo 6
Amastigota	12130010	94,04	22,51
Metacíclica	18479420	97,88	23,33
Procíclica	15576192	97,88	31,00

3.3.3 Sequenciamento das bibliotecas de RNA total – NextSeq

Após a verificação da representação relativa de rRNAs, nas amostras de RNA total dos três estágios do ciclo de vida de *L. braziliensis*, realizamos o RNAseq na plataforma NextSeq-Illumina. O sequenciamento gerou aproximadamente 677 milhões de *paired-end reads* (Tabela 7) sendo 91% da *read* 1 e 83% da *read* 2 com qualidade acima de 30 (Figura 16).

Figura 16. Porcentagem de *reads* com qualidade maior que 30 por ciclo.

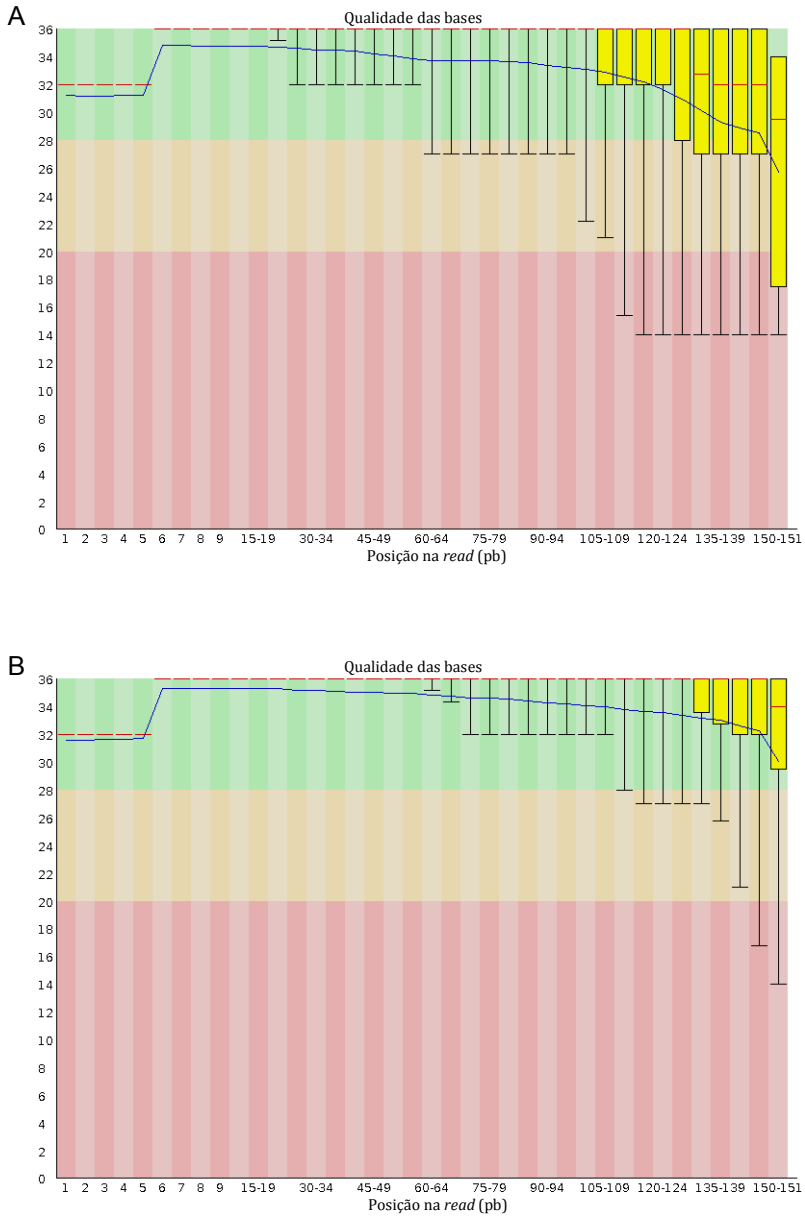


A *read* 1 é obtida nos primeiros 150 ciclos e a *read* 2 nos últimos 150 ciclos. Figura obtida no BaseSpace-Illumina. Fonte: autoral

O detalhamento da qualidade das *reads* de cada biblioteca foi observado para avaliar a necessidade de eliminação de parte das sequências. Como exemplo, a

figura 17 mostra a qualidade por base das duas *reads* da primeira réplica de promastigota procíclica, sendo que as outras bibliotecas apresentaram perfil semelhante. A extremidade final da *read* 2 foi a que apresentou bases com qualidade abaixo de 20 à partir da posição 110. Não foi necessário remover essas bases pois no mapeamento das sequências seria utilizado um parâmetro que permite o não alinhamento das extremidades das *reads* (Seção 3.2.2).

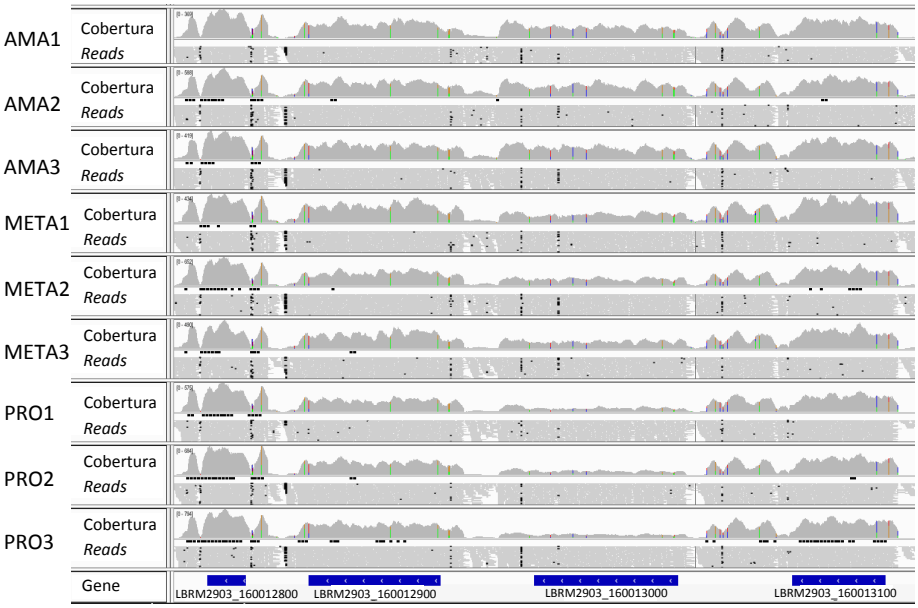
Figura 17. Qualidade das *reads* por base.



A) *Read* 1; B) *Read* 2. Exemplo da primeira réplica de promastigota procíclica de *L. braziliensis*. Figura obtida com o programa FASTQC. Fonte: autoral

O mapeamento sobre o genoma de referência (*Leishmania braziliensis* 2903 – versão 30, TriTrypDB) (Figura 18) mostrou mais de 98% das *reads* mapeadas (Tabela 07).

Figura 18. Alinhamento das bibliotecas por estágio de desenvolvimento de *L. braziliensis*.



Trecho do cromossomo 16 de *L. braziliensis* entre os genes LBRM2903_160012800 e LBRM2903_160013100. A esquerda estão identificadas as três réplicas de promastigota procíclica (cobertura e *reads* de PRO1, PRO2 e PRO3), promastigota metacíclica (cobertura e *reads* de META1, META2 e META3) e amastigota (cobertura e *reads* de AMA1, AMA2 e AMA3). Figura obtida com o programa IGV. Fonte: autoral

As *reads* mapeadas no cromossomo 6 de *L. braziliensis* (cromossomo com o maior cluster de rRNAs) variaram entre 20% e 32% por biblioteca (Tabela 7).

Tabela 7. Informações do sequenciamento no NextSeq dos três estágios de desenvolvimento de *L. braziliensis*.

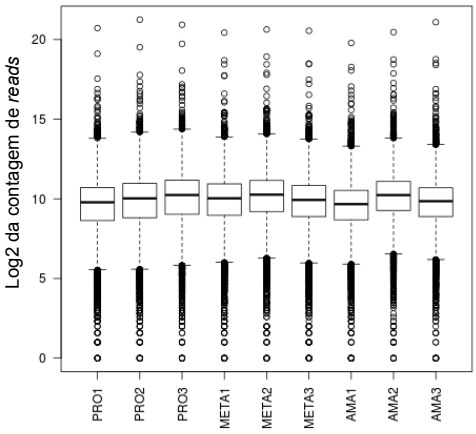
				Bowtie2 - LbrM2903	
Biblioteca	**Replica**	**Total de *reads* *paired-end***	**% *reads* por biblioteca**	**% *reads* mapeadas**	**% *reads* cromossomo 6**
Procíclica	1	69458528	10,25	98,5	28,46
Procíclica	2	84539426	12,47	98,54	30,48
Procíclica	3	90215594	13,31	98,66	25,41
Metacíclica	1	72520332	10,70	98,83	22,81
Metacíclica	2	84021672	12,40	98,84	23,01
Metacíclica	3	66830070	9,86	99,01	23,62
Amastigota	1	54228698	8,00	98,95	20,40
Amastigota	2	81757314	12,06	98,76	22,21
Amastigota	3	74281704	10,96	99,18	32,29

Sendo assim, a depleção de rRNAs e a quantidade de *reads* geradas alcançaram o resultado esperado, possibilitando a continuação das análises para identificação e caracterização dos ncRNAs de *Leishmania*.

3.3.3.1 Correlação entre as réplicas das bibliotecas de RNA total

Para verificar a correlação entre as réplicas calculamos a distribuição da contagem bruta de *reads*. Foi possível observar a equivalência na quantidade de *reads* geradas entre cada estágio de desenvolvimento e também entre as réplicas de cada estágio (Figura 19).

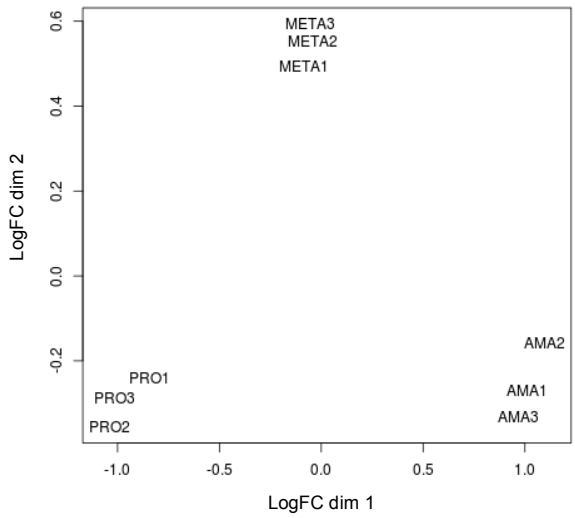
Figura 19. Contagem de reads por CDSs em cada biblioteca sequenciada.



No eixo x estão representadas as 9 bibliotecas analisadas: PRO1 (procíclico réplica 1), PRO2 (procíclico réplica 2), PRO3 (procíclico réplica 3), META1 (metacíclico réplica 1), META2 (metacíclico réplica 2), META3 (metacíclico réplica 3), AMA1 (amastigota réplica 1), AMA2 (amastigota réplica 2), AMA3 (amastigota réplica 3). O eixo y representa o log2 da contagem de *reads* por CDS. Fonte: autoral

Outro ponto investigado foi a semelhança/diferença entre as réplicas utilizando a distância BCV em um gráfico MDS. A análise demonstrou a concordância entre as réplicas de um mesmo estágio de desenvolvimento e a diferença entre os estágios analisados, possibilitando a análise de expressão diferencial (Figura 20).

Figura 20. Distância entre as réplicas das bibliotecas.

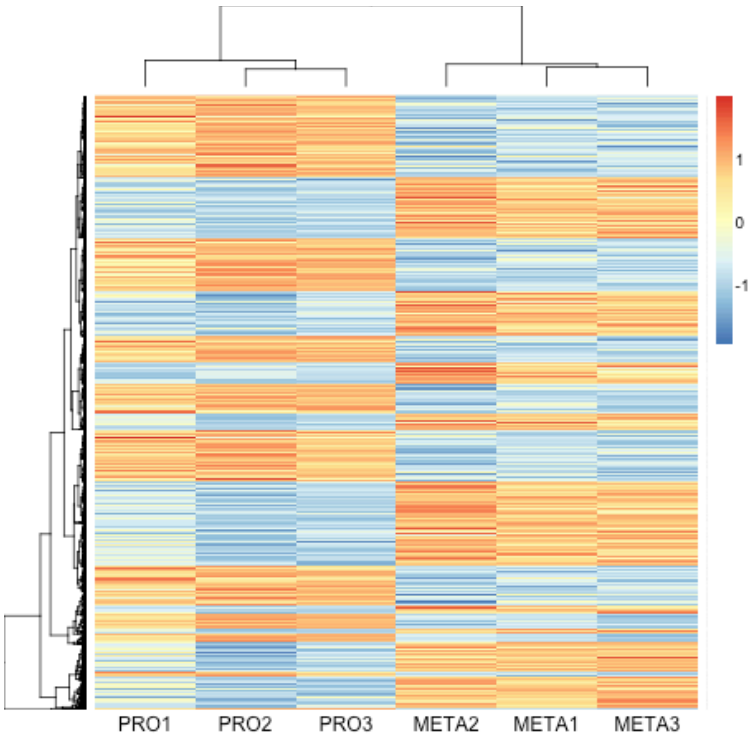


Os pontos estão representados em logFC (*log fold change*): diferença de expressão entre os transcritos (CDSs) de cada biblioteca. Os eixos x e y representam as dimensões 1 e 2, derivadas de uma análise MDS. As bibliotecas analisadas foram: PRO1 (procíclico réplica 1), PRO2 (procíclico réplica 2), PRO3 (procíclico réplica 3), META1 (metacíclico réplica 1), META2 (metacíclico réplica 2), META3 (metacíclico réplica 3), AMA1 (amastigota réplica 1), AMA2 (amastigota réplica 2), AMA3 (amastigota réplica 3). Fonte: autoral

3.3.3.2 Expressão diferencial

A análise de expressão diferencial foi realizada comparando o nível de expressão dos genes entre os seguintes estágios de desenvolvimento de *L. braziliensis*: promastigota procíclico versus promastigota metacíclico (PROvsMETA), promastigota metacíclico versus amastigota (METAvsAMA) e amastigota versus promastigota procíclico (AMAvsPRO). Do total de 9.269 genes codificadores de proteínas anotados, foram identificados 5.689 genes (61%) diferencialmente expressos (DE, p-valor ajustado < 0,05) na comparação PROvsMETA, quando empregamos os parâmetros *default* do programa DESeq (Figura 21). Ao utilizarmos um filtro de *fold change* (FC) maior ou igual a 2, encontramos 256 genes (2,8%) DE na mesma comparação (p-valor ajustado < 0,05). Para a comparação PROvsMETA, observamos nos extremos de nível de expressão diferencial genes com FC de até ~3,0 vezes, entre os sobre-representados, e de até ~5,2 vezes, entre os sub-representados.

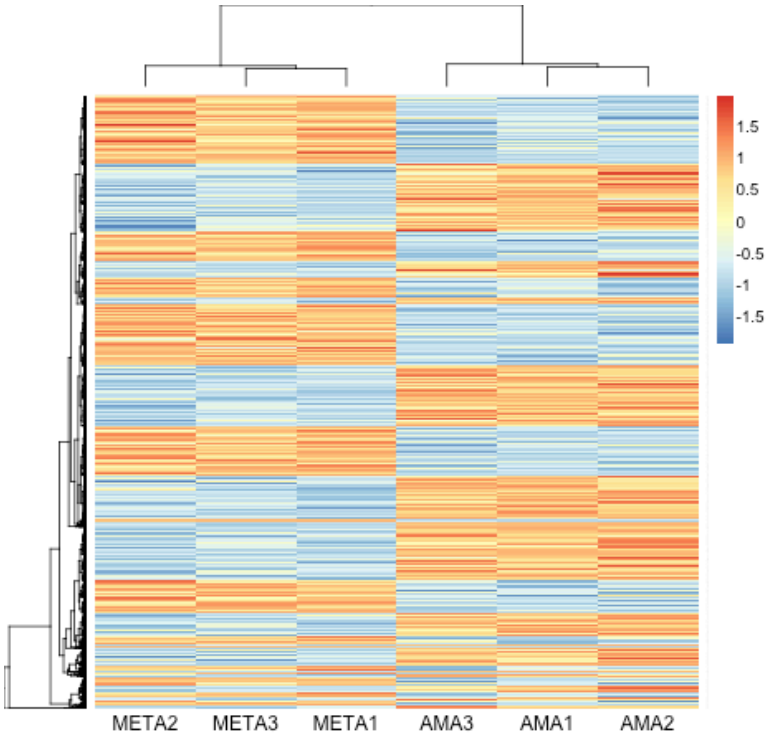
Figura 21. *Heatmap* dos genes diferencialmente expressos para PROvsMETA.



PRO1, PRO2 e PRO3 representam as réplicas de promastigotas procíclicos e META1, META2 e META3 as réplicas de metacíclicos. Fonte: autoral

O resultado da comparação METAvsAMA revelou 4.856 genes (52%) no conjunto daqueles DE, empregando os parâmetros *default* do programa DESeq. Considerando um filtro de FC maior ou igual a 2, encontramos 264 genes (2,8%) DE na mesma comparação (p-valor ajustado < 0,05). Nos extremos de nível de expressão diferencial detectamos genes com FC de até ~5,7 vezes para os sobre-representados e ~15,1 vezes para os sub-representados (Figura 22).

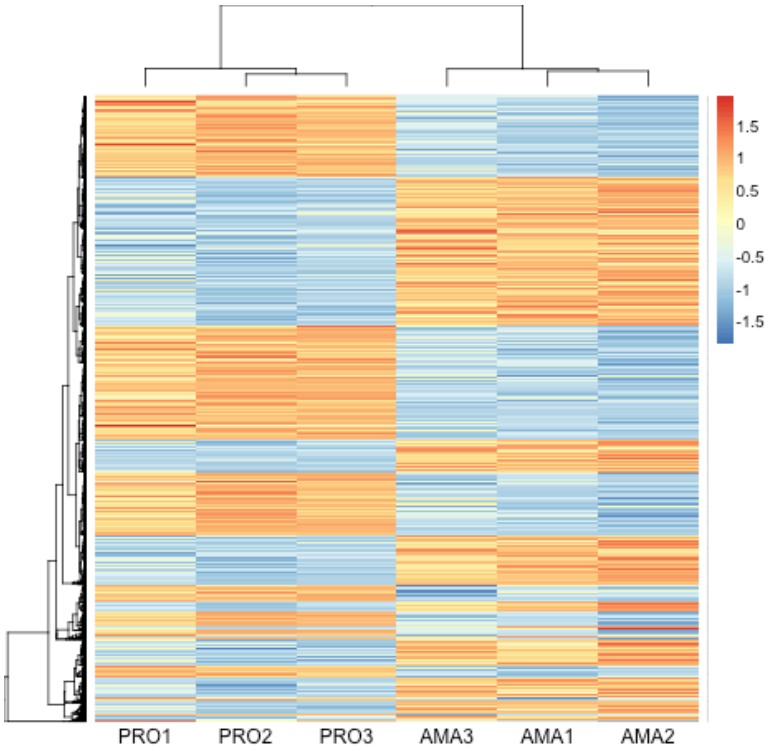
Figura 22. *Heatmap* dos genes diferencialmente expressos para METAvsAMA.



META1, META2 e META3 representam as réplicas de promastigotas metacíclicos e AMA1, AMA2 e AMA3 as réplicas de amastigotas. Fonte: autoral

Na comparação AMAvsPRO foram identificados 6.576 genes (71%) DE, empregando os parâmetros *default* do programa DESeq. Considerando um filtro de FC maior ou igual a 2, encontramos 813 genes (8,8%) DE na mesma comparação (p-valor ajustado < 0,05). Nos extremos de nível de expressão diferencial detectamos genes com FC de até ~43,7 vezes para os genes sobre-representados e de até ~3,9 vezes para os sub-representados (Figura 23).

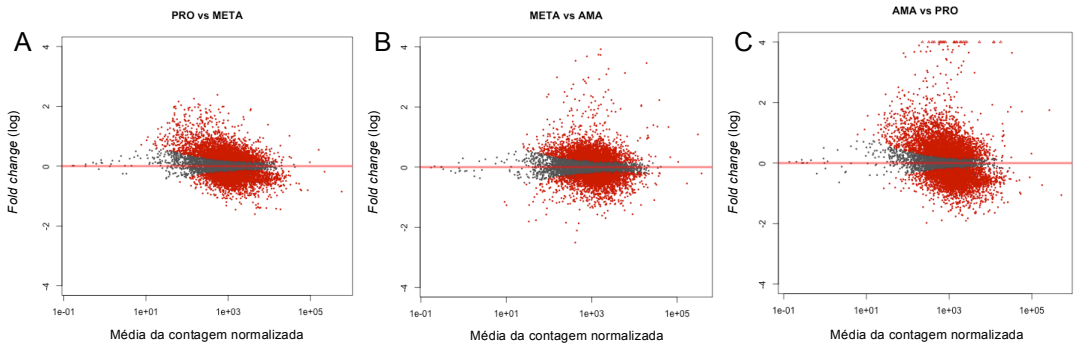
Figura 23. *Heatmap* dos genes diferencialmente expressos para AMAvsPRO.



AMA1, AMA2 e AMA3 representam as réplicas de amastigotas e PRO1, PRO2 e PRO3 as réplicas de promastigotas procíclicos. Fonte: autoral

Os genes DE foram visualizados utilizando gráficos MA (MA *plots*) que mostram a relação entre a média de expressão (média da contagem de *reads* normalizada) e *fold change* para cada gene (Figura 24).

Figura 24. Genes diferencialmente expressos na comparação entre os estágios de desenvolvimento de *L. braziliensis*.



A) PROvsMETA; B) METAvsAMA e C) AMAvsPRO. Os gráficos MA mostram a relação entre média de expressão (média da contagem normalizada) e *fold change*. Os pontos em vermelho representam os genes expressos em níveis significativamente diferentes entre os estágios de desenvolvimento comparados com um p-valor ajustado < 0,05. Fonte: autoral

Nos gráficos MA podemos observar que entre os estágios de amastigota e promastigota procíclica encontramos as maiores diferenças (maiores FCs) ao passo que entre os promastigotas procíclicos e metacíclicos encontramos as menores diferenças (menores FCs).

A tabela com todos os resultados da análise de expressão diferencial pode ser obtida sob solicitação (akcruz@fmrp.usp.br). Os transcritos com as maiores diferenças (os 25 primeiros genes) nas comparações entre os estágios de desenvolvimento, tanto sobre- quanto sub-representados estão apresentados nas Tabelas 8, 9 e 10. Na comparação PROvsMETA, o gene sobre-representado com maior FC foi o LBRM2903_350028900 (*chaperonin HSP60, mitochondrial precursor*) e o gene sub-representado com maior FC foi LBRM2903_340010300 (*proteophosphoglycan ppg3, putative*) (Tabela 8).

Tabela 8. Genes diferencialmente expressos entre promastigotas procíclico e metacíclico de *Leishmania braziliensis*: lista dos 25 genes apresentando DE mais marcante na análise comparativa.

CDS ID	Descrição	Fold Change
Regulados positivamente		
LBRM2903_350028900	chaperonin HSP60, mitochondrial precursor	3,04
LBRM2903_350029000	TCP-1/cpn60 chaperonin family, putative	2,74
LBRM2903_130008100	Alba, putative	2,72
LBRM2903_190008300	histone H2B	2,68
LBRM2903_190008200	histone H2B	2,64
LBRM2903_210012700	ATPase subunit 9, putative, pseudogene	2,64
LBRM2903_330025200	cyclophilin 4, putative	2,63
LBRM2903_350029100	chaperonin HSP60, mitochondrial precursor	2,61
LBRM2903_320039900	nucleoside diphosphate kinase b	2,61
LBRM2903_290025200	histone H2A	2,57
LBRM2903_190024500	hypothetical protein, conserved	2,49
LBRM2903_300014000	co-chaperone GrpE, putative	2,46
LBRM2903_210009400	hypothetical protein, conserved	2,40
LBRM2903_290025300	histone H2A	2,39
LBRM2903_080016500	hypothetical protein, conserved	2,32
LBRM2903_190009900	Mitochondrial carrier protein, putative	2,31
LBRM2903_300039600	hypothetical protein, conserved	2,30
LBRM2903_310012200	hypothetical protein, conserved	2,16
LBRM2903_350051700	S-adenosylhomocysteine hydrolase	2,14
LBRM2903_250025000	hypothetical protein, conserved	2,13
LBRM2903_300013900	hypothetical protein, conserved	2,12
LBRM2903_200013700	hypothetical protein, conserved	2,11
LBRM2903_260031800	hypothetical protein, conserved	2,11
LBRM2903_080016100	hypothetical protein	2,11
LBRM2903_350005300	Core histone H2A/H2B/H3/H4, putative	2,11
Regulados negativamente		
LBRM2903_340010300	proteophosphoglycan ppg3, putative	5,22
LBRM2903_310013800	hypothetical protein, conserved	4,77
LBRM2903_310036600	hypothetical protein, conserved	4,47
LBRM2903_080013700	Amastin surface glycoprotein, putative	4,21
LBRM2903_200055700	amastin-like surface protein, putative	4,12
LBRM2903_260019600	Cytochrome b5-like Heme/Steroid binding domain containing protein, putative	4,11
LBRM2903_210019000	hypothetical protein, conserved	4,09
LBRM2903_200055800	Amastin surface glycoprotein, putative	3,99
LBRM2903_200014600	hypothetical protein	3,98
LBRM2903_330023400	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain), putative	3,97

LBRM2903_200014900	Amastin surface glycoprotein, putative	3,96
LBRM2903_200077800	n-ethylmaleimide reductase-like protein	3,92
LBRM2903_150010600	hypothetical protein, conserved	3,92
LBRM2903_000007500	hypothetical protein	3,89
LBRM2903_230017100	RNA-binding protein, putative	3,88
LBRM2903_200055900	Amastin surface glycoprotein, putative	3,83
LBRM2903_240023300	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)/RNA recognition motif (a.k.a. RRM, RBD, or RNP domain), putative	3,79
LBRM2903_200018500	amastin-like surface protein, putative	3,78
LBRM2903_080005100	Amastin surface glycoprotein, putative	3,75
LBRM2903_200014800	Amastin surface glycoprotein, putative	3,73
LBRM2903_290021300	RNA binding protein, putative	3,71
LBRM2903_340049000	4E-interacting protein, putative	3,67
LBRM2903_080013600	Amastin surface glycoprotein, putative	3,67
LBRM2903_080005200	Amastin surface glycoprotein, putative	3,64
LBRM2903_310020200	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain), putative	3,64

Na comparação METAvsAMA, o genes sobre-representado com maior FC foi o LBRM2903_330039700 (*hypothetical protein, conserved*) e o sub-representado com maior FC foi o LBRM2903_180009600 (Amastin surface glycoprotein, putative) (Tabela 9).

Tabela 9. Genes diferencialmente expressos entre promastigota metacíclico e amastigota de *Leishmania braziliensis*: lista dos 25 genes apresentando DE mais marcante na análise comparativa.

CDS ID	Descrição	Fold Change
Regulados positivamente		
LBRM2903_330039700	hypothetical protein, conserved	5,68
LBRM2903_190017400	ATG8/AUT7/APG8/PAZ2, putative	4,31
LBRM2903_230020500	hypothetical protein	3,65
LBRM2903_190018000	Autophagy protein Atg8 ubiquitin like, putative	3,60
LBRM2903_030011500	hypothetical protein, conserved	3,50
LBRM2903_330023400	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain), putative	3,40
LBRM2903_270029100	D-lactate dehydrogenase-like protein	3,24
LBRM2903_240029500	3-hydroxy-3-methylglutaryl-CoA synthase, putative	3,06
LBRM2903_250023700	hypothetical protein, conserved	2,98
LBRM2903_350048600	EF-hand domain pair, putative	2,98
LBRM2903_290017100	5-histidylcysteine sulfoxide synthase, putative	2,96
LBRM2903_330035500	Radial spokehead-like protein, putative	2,87
LBRM2903_300005000	hypothetical protein	2,73
LBRM2903_330035100	hypothetical protein, conserved	2,71
LBRM2903_040007000	surface antigen-like protein	2,66
LBRM2903_180011300	RNA binding protein, putative	2,59
LBRM2903_330040900	hypothetical protein, conserved	2,55
LBRM2903_200059700	Protein of unknown function (DUF2946), putative	2,54
LBRM2903_200068300	hypothetical protein	2,47
LBRM2903_260023700	hypothetical protein, conserved	2,47
LBRM2903_330018700	hypothetical protein, conserved	2,47
LBRM2903_330044200	delta-12 fatty acid desaturase	2,44
LBRM2903_200068700	hypothetical protein, conserved	2,42
LBRM2903_200068100	hypothetical protein	2,41
LBRM2903_200043400	HECT-domain (ubiquitin-transferase), putative	2,40
Regulados negativamente		
LBRM2903_180009600	Amastin surface glycoprotein, putative	15,14
LBRM2903_000007500	hypothetical protein	13,35
LBRM2903_290021100	Amastin surface glycoprotein, putative	13,25
LBRM2903_000007600	hypothetical protein	12,83
LBRM2903_180009400	hypothetical protein, conserved	11,58
LBRM2903_200018900	Amastin surface glycoprotein, putative	10,99
LBRM2903_200018500	amastin-like surface protein, putative	9,63
LBRM2903_200056100	amastin-like surface protein, putative	8,75
LBRM2903_200018700	amastin-like surface protein, putative	8,49
LBRM2903_180009500	hypothetical protein, conserved	7,91
LBRM2903_000005600	Amastin surface glycoprotein, putative	7,88
LBRM2903_200055700	amastin-like surface protein, putative	7,59
LBRM2903_200016600	Amastin surface glycoprotein, putative	7,52

LBRM2903_000006600	amastin-like surface protein, putative	7,23
LBRM2903_200055900	Amastin surface glycoprotein, putative	6,93
LBRM2903_200055800	Amastin surface glycoprotein, putative	6,73
LBRM2903_200018600	Amastin surface glycoprotein, putative	6,09
LBRM2903_200018300	hypothetical protein	5,93
LBRM2903_200018200	hypothetical protein	5,66
LBRM2903_000005300	Amastin surface glycoprotein, putative	5,65
LBRM2903_100024200	amastin-like protein	5,36
LBRM2903_130019600	amastin-like surface protein-like protein	5,32
LBRM2903_080013700	Amastin surface glycoprotein, putative	5,15
LBRM2903_000007200	Amastin surface glycoprotein, putative	5,13
LBRM2903_200018800	Amastin surface glycoprotein, putative	5,05

Na comparação AMAvsPRO, o gene sobre-representado com maior FC foi o LBRM2903_000007500 (*hypothetical protein*) e o sub-representado com maior FC foi o LBRM2903_330039700 (*hypothetical protein, conserved*) (Tabela 10).

Tabela 10. Genes diferencialmente expressos entre amastigota e promastigota procíclico de *Leishmania braziliensis*: lista dos 25 genes apresentando DE mais marcante na análise comparativa.

CDS ID	Descrição	Fold Change
Regulados positivamente		
LBRM2903_000007500	hypothetical protein	43,67
LBRM2903_200018500	amastin-like surface protein, putative	39,23
LBRM2903_200055700	amastin-like surface protein, putative	34,69
LBRM2903_200055800	Amastin surface glycoprotein, putative	34,02
LBRM2903_180009600	Amastin surface glycoprotein, putative	32,70
LBRM2903_200018700	amastin-like surface protein, putative	32,53
LBRM2903_200055900	Amastin surface glycoprotein, putative	30,75
LBRM2903_200018900	Amastin surface glycoprotein, putative	28,66
LBRM2903_200056100	amastin-like surface protein, putative	26,33
LBRM2903_290021100	Amastin surface glycoprotein, putative	23,84
LBRM2903_080013700	Amastin surface glycoprotein, putative	23,63
LBRM2903_000007600	hypothetical protein	23,52
LBRM2903_000006600	amastin-like surface protein, putative	23,22
LBRM2903_000007200	Amastin surface glycoprotein, putative	21,75
LBRM2903_180009400	hypothetical protein, conserved	21,59
LBRM2903_000005600	Amastin surface glycoprotein, putative	19,69
LBRM2903_080013600	Amastin surface glycoprotein, putative	19,08
LBRM2903_200016600	Amastin surface glycoprotein, putative	18,89
LBRM2903_200014900	Amastin surface glycoprotein, putative	18,87
LBRM2903_200055600	Amastin surface glycoprotein, putative	17,48
LBRM2903_080005200	Amastin surface glycoprotein, putative	17,34
LBRM2903_200014800	Amastin surface glycoprotein, putative	16,89
LBRM2903_200014700	Amastin surface glycoprotein, putative	16,62
LBRM2903_080005100	Amastin surface glycoprotein, putative	16,17
LBRM2903_130019600	amastin-like surface protein-like protein	15,24
Regulados negativamente		
LBRM2903_330039700	hypothetical protein, conserved	3,94
LBRM2903_330043900	h1 histone-like protein	3,75
LBRM2903_270023100	hypothetical protein	3,71
LBRM2903_140017900	p25-alpha, putative	3,55
LBRM2903_350048600	EF-hand domain pair, putative	3,39
LBRM2903_330035500	Radial spokehead-like protein, putative	3,31
LBRM2903_270029100	D-lactate dehydrogenase-like protein	3,27
LBRM2903_330011900	paraflagellar rod component, putative	3,21
LBRM2903_140012700	small myristoylated protein-3, putative	3,16
LBRM2903_240029500	3-hydroxy-3-methylglutaryl-CoA synthase, putative	3,16
LBRM2903_350028900	chaperonin HSP60, mitochondrial precursor	3,14
LBRM2903_180020200	heat shock protein 110, putative	3,13
LBRM2903_140014000	small myristoylated protein-3, putative	3,11
LBRM2903_140013900	small myristoylated protein-3, putative	3,04
LBRM2903_290017100	5-histidylcysteine sulfoxide synthase, putative	3,04
LBRM2903_030011600	Uroporphyrinogen-III synthase HemD, putative	3,03
LBRM2903_330012600	sterol C-24 reductase, putative	3,00
LBRM2903_230020500	hypothetical protein	2,99
LBRM2903_180022900	hypothetical protein, conserved	2,96
LBRM2903_330034900	hypothetical protein, conserved	2,94
LBRM2903_330018700	hypothetical protein, conserved	2,93

LBRM2903_330008500	Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase, putative	2,91
LBRM2903_330035900	carboxypeptidase, putative	2,87
LBRM2903_300039100	3-hydroxy-3-methylglutaryl-CoA reductase	2,81
LBRM2903_300034800	hypothetical protein, conserved	2,80

Vale ressaltar que as análises de genes diferencialmente expressos são baseadas em abundância de transcritos e que a modulação do nível de proteína não obrigatoriamente acompanha os níveis dos seus transcritos [112]. Além disso, os resultados apresentados foram obtidos com culturas axênicas que podem apresentar resultados diferentes de um parasito em seu ambiente natural [113].

Os resultados da análise de expressão diferencial das três comparações de estágio de desenvolvimento (PROvsMETA, METAvsAMA e AMAvsPRO), apresentaram um elevado número de genes codificadores de proteínas hipotéticas (43 somente entre as apresentadas na Tabelas 8, 9 e 10). Esses genes precisam ser melhor investigados para se elucidar possíveis funções e a importância da diferença de expressão entre os estágios de desenvolvimento. A presença de genes codificadores de proteínas hipotéticas é comumente encontrado em outros transcriptomas de tripanossomatídeos [56; 57; 114].

Analisando os resultados de expressão diferencial, observamos a presença de múltiplas histonas (H2B, H2A e H4) que estão sub-representadas em metacíclicos (sobre-representadas na comparação PROvsMETA) (Tabela 8). Dillon e colaboradores, na análise do transcriptoma de *L. major*, encontraram resultado semelhante e correlacionam positivamente esses transcritos à metaciclogênese [56]. Soto e colaboradores fazem descrição semelhante da modulação de Histonas em *L. infantum* [115], ao passo que em *L. enriettii*, o transcrito de H2B foi encontrado em níveis mais altos em promastigotas do que em amastigotas [116]. Em *L. donovani*, o transcrito de H2A mostrou-se mais abundante em promastigotas em fase logarítimica do que em estácionária, sugerindo sua regulação negativa quando as promastigotas estão deixando o ciclo proliferativo [117]. O promastigota procíclico é a forma proliferativa extracelular de *Leishmania* no inseto vetor [31; 32], e já foi descrito que a biossíntese de histonas nos tripanossomatídeos está ligada à síntese de DNA [118; 119]. Segundo Dillon e colaboradores, a diminuição dos transcritos de histonas quando o parasito entra na fase não proliferativa (metacíclicos), sugere um modo de regulação dependente de ciclo celular [56], sendo que essa mesma regulação já foi demonstrada em outros eucariotos [120; 121].

O transcrito sobre-representado com o maior FC na comparação PROvsMETA foi a chaperonina HSP60 (Tabela 8). A HSP60, entre outras possíveis funções descritas [122], está principalmente associada ao processo de enovelamento de proteínas e na importação de substrato para a matriz mitocondrial, além disso, como uma chaperonina, está envolvida em homeostase e sobrevivência [123]. Empregando um raciocínio teleológico, é coerente com as demandas de células em proliferação, que a HSP60 esteja mais expressa em amastigotas e em promastigotas procíclicos comparativamente à forma não proliferativa (metacíclicos), como sugerem nossos resultados de sobre-representação de seu transcrito (Tabela 8 e tabela completa dos genes DE, disponível sob solicitação - akcruz@fmrp.usp.br).

Os transcritos da ATPase subunidade 9 também foram encontrados sobre-representados em promastigotas procíclicos, tanto na comparação PROvsMETA (Tabela 8) quanto na AMAvsPRO (2,28 vezes - Tabela com os resultados completos de genes DE, disponível sob solicitação - akcruz@fmrp.usp.br).

Transcritos de algumas das isoformas de amastina foram identificados como sobre-representados em promastigotas metacíclicos na comparação PROvsMETA. Apesar da maioria das amastinas serem específicas de amastigotas, algumas delas são predominantemente de promastigotas estacionária-tardia ou expressas em ambos os estágios [124]. Para as amastigotas, os transcritos de amastinas foram a aqueles que se apresentaram com as maiores diferenças (25 genes com maior DE), tanto na comparação METAvsAMA quanto na AMAvsPRO, chegando a estar 39,23 vezes mais expressa do que em promastigotas procíclicas (Tabelas 9 e 10). Em *L. infantum* os transcritos de amastina se apresentam até 13,9 vezes mais abundantes nas amastigotas [124], e em *T. cruzi* os transcritos de amastinas apresentaram-se de 50 a 85 vezes mais abundantes nas amastigotas do que em epimastigotas ou tripomastigotas [125]. Essas discrepâncias de DE podem ter diversas origens; podem ser reais, e portanto características de cada espécie ou podem dever-se ao fato de fazermos comparações entre abordagens experimentais distintas (microarranjo e RNAseq) que por vieses e limitações próprias sugerem perfis distintos [126]. As amastinas são pequenas proteínas de superfície, específicas de kinetoplastídeos e são consideradas fatores de virulência e essenciais para a multiplicação dos amastigotas nas células do hospedeiro mamífero. Apesar disso, a função das amastinas continua desconhecida, presume-se que essas proteínas interagem com as moléculas do hospedeiro ou agem como transportadores de membrana [127; 128].

Alguns estudos indicam que as amastinas estão associadas ao tráfego de íons pela membrana para ajustar o pH citoplasmático do parasito [124; 129].

O transcrito da ppg3 (Proteophosphoglycan 3) foram os mais abundantes (maior FC - 5,22) em promastigotas metacíclicos na comparação PROvsMETA (Tabela 8) e também apresentou uma sobre-representação em amastigota na comparação AMAvsPRO (Tabela completa dos genes DE, disponível sob solicitação - akcruz@fmrp.usp.br). O gene ppg3 codifica uma das glicoproteínas constituintes do fPPG (*filamentous proteophosphoglycan*), um componente principal do PSG (*Promastigote Secretory Gel*). O PSG é produzido pelo parasito dentro do intestino do inseto vetor que contribui para um bloqueio físico para a alimentação por sangue do hospedeiro e subsequente transmissão do parasito [130; 131]. PPGs de superfície são importantes na interação parasito-macrófago, elas aparentemente se ligam nas células do hospedeiro para modular a função do macrófago durante os estágios iniciais da infecção [132]. O aumento da expressão do gene da PPG em amastigotas ainda é desconhecido [133].

Transcritos de genes relacionados a autofagia também foram encontrados mais abundantes em promastigotas metacíclios (comparação METAvsAMA – Tabela 9). A autofagia, o processo pelo qual a célula pode degradar suas própria proteínas, é importante para a degradação de proteínas e organelas durante a diferenciação celular e para a defesa contra condições de pouco nutriente (estresse nutricional) [134]. Já foi descrito que para *Leishmania* a autofagia é essencial para o processo de diferenciação entre PRO, META e AMA e para virulência [135]. Transcritos relacionados a autofagia também foram encontrados em maior abundância em formas “móveis” de outras espécies de *Leishmania*, derivadas do inseto vetor [114] e durante as primeiras horas de infecção [136].

Também foram encontrados sobre-representados os transcritos de vários genes relacionados a RBPs (*RNA binding proteins*) em promastigotas metacíclicos, tanto na comparação PROvsMETA quanto em METAvsAMA (Tabelas 8 e 9). As RBPs possuem funções importantes na regulação da expressão gênica, por estaram envolvidas em processamento pós-transcricional como regulação de *splicing*, estabilidade de mRNA, transporte de mRNA e modulação de taxa de iniciação de tradução [137]. Notamos portanto, que é possível que esse aumento dos transcritos de um conteúdo grande de RBPs esteja associado à intensa modulação da expressão

gênica nos metacíclicos de *L. braziliensis*. Este mesmo perfil de expressão foi encontrado em *L. major* [56].

3.3.3.2.1 Genes espécie-específicos

Em 2007 Smith e colaboradores analisaram comparativamente os genomas de *L. major*, *L. infantum* e *L. braziliensis* e identificaram 49 genes espécie específicos em *L. braziliensis*; eram 10 genes com função predita e 39 anotados como hipotéticos [41]. Para fazer a correlação entre os IDs dos genes disponibilizados por aqueles autores como *L. braziliensis*-específicos (linhagem 2904) e os IDs atuais de *L. braziliensis* (linhagem 2903), fizemos busca por similaridade de sequência e identificamos 16 genes com função predita/definida e 32 classificados como codificadores de proteínas hipotéticas (Tabela disponível sob solicitação - akcruz@fmrp.usp.br). Trinta e nove desses genes foram encontrados como diferencialmente expressos entre os estágios de desenvolvimento, sendo 25 classificados como codificadores de proteínas hipotéticas e 14 deles com função predita/definida (Tabela 11).

Tabela 11. Genes identificados como específicos de *L. braziliensis* em 2007 [41] e diferencialmente expressos.

CDS ID	Descrição	PROvsMETA Sobre	METAvsAMA Sobre	AMAvsPRO Sobre	PROvsMETA Sub	METAvsAMA Sub	AMAvsPRO Sub	Ortólogos com o subgênero *Viannia*
LBRM2903_110008400	argonaute like protein (pseudogene)	1,2	1,23	-	-	-	1,48	*L. braziliensis* MHOM/BR/75/M2904 - argonaute-like protein (LbrM.11.0360) / *L. major* - argonaute-like protein (pseudogene) (LmjF.11.0570)*
LBRM2903_220013600	Arrestin (or S-antigen), N-terminal domain containing protein, putative	-	1,38	-	-	-	1,3	-
LBRM2903_280015300	EF hand, putative	-	1,18	1,22	1,43	-	-	-
LBRM2903_260012900	Glutathione peroxidase, putative	-	-	2,99	1,85	1,58	-	-
LBRM2903_260006200	Phosphoribosyl transferase domain containing protein, putative	-	-	-	1,24	-	-	-
LBRM2903_240021600	Protein of unknown function (DUF563), putative	-	-	1,24	-	1,17	-	-
LBRM2903_160010900	Reverse transcriptase (RNA-dependent DNA polymerase), putative	-	-	3,36	1,46	2,12	-	-
LBRM2903_110019600	hypothetical protein	-	1,18	-	-	-	1,28	-

LBRM2903_200006800	hypothetical protein	-	1,23	-	-	-	-	-
LBRM2903_230011600	hypothetical protein	-	1,22	-	-	-	1,23	-
LBRM2903_250014200	hypothetical protein	1,32	-	-	-	-	1,33	*Crithidia fasciculata* strain Cf-CI - Alpha/beta hydrolase family, putative (CFAC1_220014800)
LBRM2903_250019400	hypothetical protein	-	-	1,59	1,33	1,18	-	*L. braziliensis* MHOM/BR/75/M2904 - Ribonuclease 3, putative (LbrM.25.1020)
LBRM2903_260005000	hypothetical protein	-	1,26	-	-	-	1,23	-
LBRM2903_280005300	hypothetical protein	-	-	1,37	1,2	1,14	-	*L. braziliensis* MHOM/BR/75/M2904 - TATE DNA transposons (LbrM.11.1160)
LBRM2903_280018400	hypothetical protein	-	-	2,73	2,19	1,23	-	-
LBRM2903_290005300	hypothetical protein	-	-	1,63	1,23	1,31	-	*L. braziliensis* MHOM/BR/75/M2904 - TATE DNA transposons (LbrM.11.1160)
LBRM2903_300032300	hypothetical protein	-	-	1,27	1,21	-	-	-
LBRM2903_310033200	hypothetical protein	-	-	1,78	1,91	-	-	-
LBRM2903_320047000	hypothetical protein	-	-	1,93	-	-	-	*L. panamensis* MHOM/COL/81/L13 - zinc finger containing protein, putative (LPAL13_320043800)
LBRM2903_320047100	hypothetical protein	-	-	3,17	1,85	-	-	*L. panamensis* MHOM/COL/81/L13 - zinc finger containing protein, putative (LPAL13_320043800)
LBRM2903_330006800	hypothetical protein	1,34	-	-	-	-	1,39	*L. braziliensis* MHOM/BR/75/M2904 - RNA Interference Factor 5, putative (LbrM.33.0190)
LBRM2903_330043600	hypothetical protein	-	1,73	-	1,15	-	1,5	*T. congolense* IL3000 - reverse transcriptase (RNA-dependent DNA polymerase), putative (TcIL3000_0_23600)
LBRM2903_340027000	hypothetical protein	1,46	-	-	-	-	1,32	-
LBRM2903_340045900	hypothetical protein	-	-	1,3	-	-	-	*L. braziliensis* MHOM/BR/75/M2904 – galactokinase (LbrM.34.3650)
LBRM2903_160011000	hypothetical protein	-	-	3,24	-	2,63	-	LBRM2903_160010900 (Reverse transcriptase)
LBRM2903_060019200	hypothetical protein	-	1,24	-	1,3	-	-	-
LBRM2903_130021900	hypothetical protein	-	-	1,14	1,14	-	-	-
								Ortólogos com o subgênero *Leishmania*
LBRM2903_040019400	Methyltransferase FkbM domain containing protein, putative	1,11	-	-	-	-	-	*L. major* strain Friedlin - hypothetical protein (LmjF.04.1165)
LBRM2903_100017800	serine/threonine-protein kinase, putative	1,13	-	1,42	-	1,59	-	*L. major* strain Friedlin - serine/threonine-protein kinase, putative (LmjF.10.0830)
LBRM2903_350084600	SET domain containing protein, putative	-	-	-	-	-	1,15	*L. major* strain Friedlin - hypothetical protein (LmjF.36.6785)
LBRM2903_350077800	target of rapamycin (TOR) kinase 1, putative	-	-	1,35	-	1,24	-	*L. major* strain Friedlin - Phosphatidylinositol 3-kinase tor1 (LmjF.36.6320)
LBRM2903_350021100	Transmembrane amino acid transporter protein, putative	-	1,17	-	1,15	-	-	*L. major* strain Friedlin - amino acid permease 24, putative (LmjF.10.0710)
LBRM2903_300031000	tyrosine/dopa decarboxylase, putative (pseudogene)	-	-	2,33	1,5	1,53	-	*L. major* strain Friedlin - tyrosine/dopa decarboxylase, putative (pseudogene) (LmjF.30.2500)
LBRM2903_220006600	zinc finger containing protein, putative	-	-	2,19	2,04	-	-	*L. major* strain Friedlin - hypothetical protein, conserved (LmjF.22.0130)
LBRM2903_260010200	hypothetical protein	1,17	1,18	-	-	-	1,38	*L. major* strain Friedlin - hypothetical protein, conserved (pseudogene) (LmjF.26.0520)
LBRM2903_200010600	hypothetical protein	-	1,18	-	-	-	1,23	*L. major* strain Friedlin - beta galactofuranosyl transferase (LmjF.25.0010)
LBRM2903_230018000	hypothetical protein	-	-	1,14	-	-	-	*L. major* strain Friedlin - hypothetical protein (LmjF.23.0845)

LBRM2903_270018100	hypothetical protein	1,12	-	-	-	-	1,18	*L. major* strain Friedlin - hypothetical protein (LmjF.27.1125)
LBRM2903_330039700	hypothetical protein	-	5,68	-	1,47	-	3,94	*L. major* strain LV39c5 - hypothetical protein, conserved (LMJLV39_330040400)

Informações sobre ortólogos obtidas no TriTrypDB. Consideramos os genes como específicos do subgênero *Viannia* ou com ortólogos no subgênero *Leishmania*. *Apesar ter sido encontrado um ortólogo em *L. major* (subgênero *Leishmania*), tem um comentário do anotador esclarecendo que é um pseudogene.

Dadas as atualizações de genomas e transcriptomas de várias espécies e linhagens de *Leishmania*, 12 desses genes têm ortólogos descritos no subgênero *Leishmania* (TriTrypDB , Tabela 11 – ortólogos com o subgênero *Leishmania*).

Entre os genes específicos do subgênero *Viannia* e diferencialmente expressos está a Argonauta (LBRM2903_110008400) que é uma endonuclease envolvida na clivagem dirigida de mRNAs (induzida por *double-stranded* RNA), sendo uma componente da maquinaria de RNAi [15]. O gene da linhagem 2903 é ortólogo ao da linhagem 2904 (LbrM.11.0360) que foi descrito como contendo os domínios típicos de argonauta (PAZ e PIWI), essenciais para a TbAGO1 funcional (Argonauta de *T. brucei*) [41; 138]. Embora o LmjF.11.0570 de *L. major*, esteja anotado também como ortólogo, ele é um pseudogene (site TriTrypDB, comentário de anotador). Já foi demonstrado em *L. major* e *L. infantum*, em uma análise das regiões sintênicas, a existência de remanescentes da Argonauta, o que favorece o entendimento de que o RNAi foi perdido no subgênero *Leishmania* [41]. A Argonauta foi encontrada sobre-representada na comparação PROvsMETA e METAvsAMA e sub-representada em AMAvsPRO, o que indica a maior expressão em PRO, diminuindo em META e reduzindo ainda mais em AMA.

Cenário inverso, ou seja, aumento da expressão conforme a evolução do ciclo de vida (PRO → META → AMA), ocorre com a Glutationa peroxidase (LBRM2903_260012900) e a transcriptase reversa (RNA-dependent DNA polymerase) (LBRM2903_160010900). A glutationa peroxidase foi descrita como componente importante dos mecanismos protetivos do parasito contra os efeitos tóxicos do óxido nítrico presente nos macrófagos [139], o que indica o motivo do aumento de transcritos de glutationa peroxidase em metacíclicos e amastigotas.

O mesmo padrão de aumento da expressão conforme evolução do ciclo ocorre com a proteínas hipotéticas LBRM2903_250019400 (ortóloga a LbrM.25.1020 - *L. braziliensis* 2904 - Ribonuclease 3); LBRM2903_280005300 (ortóloga a LbrM.11.1160 - *L. braziliensis* 2904 - *TATE DNA transposons*) e LBRM2903_290005300 (ortóloga a LbrM.11.1160 - *L. braziliensis* 2904 - *TATE DNA*

transposons). Os TATEs (*Telomeric Associated Transposable Elements*) foram descritos em *L. braziliensis* em 2007 [40]. Esses elementos transponíveis, que podem afetar a integridade do genoma, devem ser regulados negativamente pela maquinaria de RNAi [140], por isso o cenário inverso de regulação ao da Argonauta (descrito acima) pode ser real, e intrigante o fato de haver uma modulação ao longo do desenvolvimento.

3.3.3.3 Enriquecimento de classes segundo *Gene Ontology*

A análise de enriquecimento de termos *Gene Ontology* (GO) foi realizada com os genes expressos diferencialmente nos estágios de desenvolvimento de *Leishmania braziliensis*. Foram analisados separadamente os genes *upregulated* e *downregulated*. Incluímos as três classes como definidas pelo GO: processo biológico, componente celular e função molecular. A seguir apresentamos a lista dos termos GO enriquecidos nos genes diferencialmente expressos (DE), em tabelas em que são comparados os três estágios analisados, dois a dois.

Ao comparar a expressão diferencial entre PROvsMETA, surgiram como categorias GO sobre-representadas em procíclicos atividades relacionadas a proliferação celular (*DNA replication*, *translation*, *translational elongation*, *protein folding* e *positive regulation of cell proliferation*), metabolismo de proteínas e energia (*nucleoside-triphosphatase activity*, *ATP synthesis*, *GTP biosynthetic process*, *UTP biosynthetic process* e *GTP biosynthetic process*) e processos celulares (*nucleosome assembly*) (Tabela 12). Como promastigota procíclico é a forma proliferativa extracelular de *Leishmania*, os processos celulares enriquecidos são coerentes com essa atividade proliferativa.

Tabela 12. Classes GO enriquecidas para genes sobre-representados em PROvsMETA.

Ontologia	ID categoria	p-valor	Função
Processo biológico	GO:0006412	5,79E-50	translation
	GO:0006457	1,13E-09	protein folding
	GO:0006334	4,12E-09	nucleosome assembly
	GO:0006414	2,41E-06	translational elongation
	GO:0044267	6,27E-06	cellular protein metabolic process
	GO:0015991	1,86E-05	ATP hydrolysis coupled proton transport
	GO:0051603	1,19E-03	proteolysis involved in cellular protein catabolic process
	GO:0007017	1,28E-03	microtubule-based process
	GO:0006888	4,92E-03	ER to Golgi vesicle-mediated transport
	GO:0015986	5,33E-03	ATP synthesis coupled proton transport
	GO:0006265	1,53E-02	DNA topological change
	GO:0045454	1,88E-02	cell redox homeostasis
	GO:0006614	2,27E-02	SRP-dependent cotranslational protein targeting to membrane
	GO:0030163	2,44E-02	protein catabolic process
	GO:0006165	3,41E-02	nucleoside diphosphate phosphorylation

	GO:0006183	3,41E-02	GTP biosynthetic process
	GO:0006228	3,41E-02	UTP biosynthetic process
	GO:0006241	3,41E-02	CTP biosynthetic process
	GO:0006952	3,59E-02	defense response
	GO:0008284	3,59E-02	positive regulation of cell proliferation
Componente celular	GO:0005840	3,12E-46	ribosome
	GO:0000786	1,51E-10	nucleosome
	GO:0015935	1,57E-04	small ribosomal subunit
	GO:0005634	9,89E-04	nucleus
	GO:0005839	1,19E-03	proteasome core complex
	GO:0005875	1,28E-03	microtubule associated complex
	GO:0019773	2,76E-03	proteasome core complex, alpha-subunit complex
	GO:0015934	3,24E-03	large ribosomal subunit
	GO:0030127	7,62E-03	COPII vesicle coat
	GO:0031514	8,50E-03	motile cilium
	GO:0016272	1,32E-02	prefoldin complex
	GO:0005737	2,00E-02	cytoplasm
	GO:0033180	2,74E-02	proton-transporting V-type ATPase, V1 domain
	GO:0005853	3,39E-02	eukaryotic translation elongation factor 1 complex
	GO:0033177	3,57E-02	proton-transporting two-sector ATPase complex, proton-transporting domain
	GO:0030286	4,57E-02	dynein complex
Função molecular	GO:0003735	2,67E-52	structural constituent of ribosome
	GO:0004298	3,21E-06	threonine-type endopeptidase activity
	GO:0051082	5,07E-04	unfolded protein binding
	GO:0046961	8,95E-04	proton-transporting ATPase activity, rotational mechanism
	GO:0003746	9,10E-04	translation elongation factor activity
	GO:0015078	1,94E-03	hydrogen ion transmembrane transporter activity
	GO:0005509	3,01E-03	calcium ion binding
	GO:0046933	5,33E-03	proton-transporting ATP synthase activity, rotational mechanism
	GO:0003755	7,71E-03	peptidyl-prolyl cis-trans isomerase activity
	GO:0005516	1,35E-02	calmodulin binding
	GO:0003918	2,23E-02	DNA topoisomerase type II (ATP-hydrolyzing) activity
	GO:0003677	2,63E-02	DNA binding
	GO:0000049	3,12E-02	tRNA binding
	GO:0004550	3,41E-02	nucleoside diphosphate kinase activity
	GO:0017111	3,51E-02	nucleoside-triphosphatase activity
	GO:0019843	3,54E-02	rRNA binding
	GO:0004252	4,18E-02	serine-type endopeptidase activity
	GO:0000062	4,37E-02	fatty-acyl-CoA binding

Para esse mesmo par (PROvsMETA), nos procíclicos estão em destaque como sub-representados GOs ligados a fosforilação (*protein phosphorylation*, *protein tyrosine phosphatase activator activity*, *protein dephosphorylation*, *protein tyrosine/serine/threonine phosphatase activity*) e sinalização (*modulation by symbiont of host nitric oxide-mediated signal transduction*, *modulation by symbiont of host protein kinase-mediated signal transduction*, *protein kinase activity*, *protein tyrosine/serine/threonine phosphatase activity* e *protein kinase binding*) (Tabela 13). Resultados similares foram encontrados em *L. major* [56].

Tabela 13. Classes GO enriquecidas para genes sub-representados em PROvsMETA.

Ontologia	ID categoria	p-valor	Função
Processo biológico	GO:0006468	5,49E-13	protein phosphorylation
	GO:0007155	4,70E-12	cell adhesion
	GO:0044081	2,45E-09	modulation by symbiont of host nitric oxide-mediated signal transduction
	GO:0075130	2,45E-09	modulation by symbiont of host protein kinase-mediated signal transduction
	GO:0055085	6,69E-07	transmembrane transport
	GO:0006508	3,44E-04	proteolysis
	GO:0006470	6,09E-03	protein dephosphorylation
	GO:0000079	1,59E-02	regulation of cyclin-dependent protein serine/threonine kinase activity

	GO:0008643	1,66E-02	carbohydrate transport
	GO:0009435	1,83E-02	NAD biosynthetic process
	GO:0043087	2,77E-02	regulation of GTPase activity
	GO:0019752	2,82E-02	carboxylic acid metabolic process
	GO:0006629	4,91E-02	lipid metabolic process
	GO:0005975	4,93E-02	carbohydrate metabolic process
Componente celular	GO:0042025	2,45E-09	host cell nucleus
	GO:0016020	1,80E-08	membrane
	GO:0016021	2,58E-04	integral component of membrane
	GO:0000139	1,66E-02	Golgi membrane
Função molecular	GO:0004672	2,59E-12	protein kinase activity
	GO:0008160	2,45E-09	protein tyrosine phosphatase activator activity
	GO:0004222	2,94E-07	metalloendopeptidase activity
	GO:0046872	1,56E-06	metal ion binding
	GO:0005524	2,95E-03	ATP binding
	GO:0016773	3,25E-03	phosphotransferase activity, alcohol group as acceptor
	GO:0003993	4,54E-03	acid phosphatase activity
	GO:0005543	7,32E-03	phospholipid binding
	GO:0008138	7,64E-03	protein tyrosine/serine/threonine phosphatase activity
	GO:0008270	9,31E-03	zinc ion binding
	GO:0030170	1,22E-02	pyridoxal phosphate binding
	GO:0046873	1,57E-02	metal ion transmembrane transporter activity
	GO:0019901	1,59E-02	protein kinase binding
	GO:0005351	1,66E-02	sugar:proton symporter activity
	GO:0008234	2,64E-02	cysteine-type peptidase activity
	GO:0016831	2,82E-02	carboxy-lyase activity
	GO:0016772	3,22E-02	transferase activity, transferring phosphorus-containing groups
	GO:0005096	4,09E-02	GTPase activator activity
	GO:0004519	4,28E-02	endonuclease activity
	GO:0004553	4,39E-02	hydrolase activity, hydrolyzing O-glycosyl compounds

Para a comparação METAvsAMA, as classes GO enriquecidas para os genes DE sobre-representados em promastigotas metacíclicos, observamos a presença de processos celulares ligados a motilidade celular (*microtubule-based movement*, *microtubule-based process*, *motile cilium*, *microtubule motor activity* e *microtubule binding*) e atividade de proteólise (*proteolysis involved in cellular protein catabolic process*, *proteasome core complex*, *alpha-subunit complex e proteasome core complex*) (Tabela 14).

Tabela 14. Classes GO enriquecidas para genes sobre-representados em METAvsAMA.

Ontologia	ID categoria	p-valor	Função
Processo biológico	GO:0006412	9,53E-04	translation
	GO:0006096	1,25E-03	glycolytic process
	GO:0006952	3,23E-03	defense response
	GO:0008284	3,23E-03	positive regulation of cell proliferation
	GO:0051603	6,81E-03	proteolysis involved in cellular protein catabolic process
	GO:0009165	7,88E-03	nucleotide biosynthetic process
	GO:0051258	8,07E-03	protein polymerization
	GO:0055114	9,55E-03	oxidation-reduction process
	GO:0007018	1,03E-02	microtubule-based movement
	GO:0042759	2,24E-02	long-chain fatty acid biosynthetic process
	GO:0007017	2,43E-02	microtubule-based process
	GO:0006108	2,56E-02	malate metabolic process
	GO:0030163	2,67E-02	protein catabolic process
	GO:0006694	2,68E-02	steroid biosynthetic process
	GO:0000302	2,77E-02	response to reactive oxygen species
	GO:0006099	3,13E-02	tricarboxylic acid cycle
	GO:0006260	3,24E-02	DNA replication
Componente	GO:0005737	3,20E-06	cytoplasm
	GO:0005840	1,12E-03	ribosome

	GO:0031514	1,44E-03	motile cilium
	GO:0019773	2,32E-03	proteasome core complex, alpha-subunit complex
	GO:0005839	6,81E-03	proteasome core complex
	GO:0033180	7,25E-03	proton-transporting V-type ATPase, V1 domain
	GO:0043234	8,07E-03	protein complex
	GO:0005875	2,43E-02	microtubule associated complex
	GO:0015935	4,66E-02	small ribosomal subunit
Função molecular	GO:0004298	3,31E-05	threonine-type endopeptidase activity
	GO:0003735	1,63E-04	structural constituent of ribosome
	GO:0005509	5,79E-04	calcium ion binding
	GO:0005516	3,93E-03	calmodulin binding
	GO:0004749	7,88E-03	ribose phosphate diphosphokinase activity
	GO:0019239	9,29E-03	deaminase activity
	GO:0003777	1,03E-02	microtubule motor activity
	GO:0000287	2,47E-02	magnesium ion binding
	GO:0050661	3,22E-02	NADP binding
	GO:0016491	3,50E-02	oxidoreductase activity
	GO:0008017	4,35E-02	microtubule binding

Considerando os genes DE sub-representados na comparação METAvsAMA (Tabela 15), observamos o enriquecimento de termos GO ligados a sinalização (*modulation by symbiont of host nitric oxide-mediated signal transduction*, *modulation by symbiont of host protein kinase-mediated signal transduction*, *protein phosphorylation* e *protein kinase activity*) e transporte (*transmembrane transport*, *phosphate ion transport*, *cation transmembrane transporter activity* e *inorganic phosphate transmembrane transporter activity*). Além disso, temos sobre-representados em amastigota (METAvsAMA) termos GO associados com atividade de ligação a DNA (DNA *binding*) e nucleossomos (*nucleosome*, *nucleosome assembly*), possivelmente refletindo a proliferação celular dos amastigotas, mas vale lembrar que amastigotas axênicos apresentam maior taxa de proliferação celular do que os amastigotas intracelulares [141].

Tabela 15. Classes GO enriquecidas para genes sub-representados em METAvsAMA.

Ontologia	ID categoria	p-valor	Função
Processo biológico	GO:0055085	4,10E-05	transmembrane transport
	GO:0006071	4,36E-03	glycerol metabolic process
	GO:0006629	5,06E-03	lipid metabolic process
	GO:0010038	7,34E-03	response to metal ion
	GO:0006334	1,77E-02	nucleosome assembly
	GO:0006468	4,05E-02	protein phosphorylation
	GO:0044081	4,18E-02	modulation by symbiont of host nitric oxide-mediated signal transduction
	GO:0075130	4,18E-02	modulation by symbiont of host protein kinase-mediated signal transduction
	GO:0046939	4,31E-02	nucleotide phosphorylation
	GO:0006308	4,31E-02	DNA catabolic process
	GO:0006310	4,48E-02	DNA recombination
	GO:0006817	4,51E-02	phosphate ion transport
Componente celular	GO:0016020	5,65E-03	membrane
	GO:0016021	9,15E-03	integral component of membrane
	GO:0000786	1,34E-02	nucleosome
	GO:0005634	2,64E-02	nucleus
	GO:0042025	4,18E-02	host cell nucleus
Função molecular	GO:0004806	5,19E-03	triglyceride lipase activity
	GO:0008270	1,05E-02	zinc ion binding
	GO:0015250	1,31E-02	water channel activity
	GO:0003747	1,36E-02	translation release factor activity

	GO:0004672	3,07E-02	protein kinase activity
	GO:0008324	3,13E-02	cation transmembrane transporter activity
	GO:0008160	4,18E-02	protein tyrosine phosphatase activator activity
	GO:0003677	4,23E-02	DNA binding
	GO:0005315	4,51E-02	inorganic phosphate transmembrane transporter activity

As classes GO enriquecidas considerando os genes DE sobre-representados em amastigotas comparativamente aos promastigotas procíclicos (AMAvsPRO, Tabela 16) repetem o resultado acima, que também encontrou enriquecido nos amastigotas processos celulares que envolvem transporte (*transmembrane transport*, *phosphate ion transport*, *carbohydrate transport*, *transporter activity*, *inorganic phosphate transmembrane transporter activity* e *metal ion transmembrane transporter activity*) e sinalização (*modulation by symbiont of host nitric oxide-mediated signal transduction*, *modulation by symbiont of host protein kinase-mediated signal transduction* e *protein kinase C-activating G-protein coupled receptor signaling pathway*). É possível, a título de especulação, que esse processos celulares podem estar relacionados ao contato parasita-hospedeiro.

Tabela 16. Classes GO enriquecidas para genes sobre-representados em AMAvsPRO.

Ontologia	ID categoria	p-valor	Função
Processo biológico	GO:0007155	3,96E-14	cell adhesion
	GO:0044081	2,74E-10	modulation by symbiont of host nitric oxide-mediated signal transduction
	GO:0075130	2,74E-10	modulation by symbiont of host protein kinase-mediated signal transduction
	GO:0006468	1,11E-08	protein phosphorylation
	GO:0055085	5,66E-04	transmembrane transport
	GO:0006508	8,50E-04	proteolysis
	GO:0006629	6,26E-03	lipid metabolic process
	GO:0006817	9,81E-03	phosphate ion transport
	GO:0006071	1,20E-02	glycerol metabolic process
	GO:0007205	1,58E-02	protein kinase C-activating G-protein coupled receptor signaling pathway
	GO:0043087	1,95E-02	regulation of GTPase activity
	GO:0008643	2,49E-02	carbohydrate transport
Componente celular	GO:0042025	2,74E-10	host cell nucleus
	GO:0016020	8,62E-10	membrane
	GO:0016021	1,89E-03	integral component of membrane
	GO:0000139	2,49E-02	Golgi membrane
Função molecular	GO:0008160	2,74E-10	protein tyrosine phosphatase activator activity
	GO:0004222	1,22E-08	metalloendopeptidase activity
	GO:0004672	1,49E-08	protein kinase activity
	GO:0008270	2,40E-06	zinc ion binding
	GO:0046872	2,50E-04	metal ion binding
	GO:0016773	2,56E-03	phosphotransferase activity, alcohol group as acceptor
	GO:0004386	8,82E-03	helicase activity
	GO:0005315	9,81E-03	inorganic phosphate transmembrane transporter activity
	GO:0004143	1,58E-02	diacylglycerol kinase activity
	GO:0005524	2,21E-02	ATP binding
	GO:0015250	2,30E-02	water channel activity
	GO:0005351	2,49E-02	sugar:proton symporter activity
	GO:0004806	2,55E-02	triglyceride lipase activity
	GO:0046873	2,63E-02	metal ion transmembrane transporter activity
	GO:0003747	2,81E-02	translation release factor activity
	GO:0005215	2,91E-02	transporter activity
	GO:0005096	3,02E-02	GTPase activator activity
	GO:0008234	4,10E-02	cysteine-type peptidase activity
	GO:0005543	4,85E-02	phospholipid binding

Para a comparação AMAvsPRO, as classes GO enriquecidas nos genes DE sub-representados (Tabela 17) apresentam majoritariamente os mesmos processos celulares sobre-representados em PRO em outras comparações, ou seja, relacionados à proliferação celular. Também encontraram-se sub-representados em AMA os GOs relacionados à motilidade celular (*microtubule-based process* e *microtubule-based movement*), que refletem a mobilidade da forma flagelada e a maior taxa de crescimento dos promastigotas procíclicos. Resultados similares foram encontrados em *L. donovani* [142].

Tabela 17. Classes GO enriquecidas para genes sub-representados em AMAvsPRO.

Ontologia	ID categoria	p-valor	Função
Processo biológico	GO:0006412	3,92E-29	translation
	GO:0006457	2,26E-07	protein folding
	GO:0006334	4,38E-07	nucleosome assembly
	GO:0044267	3,24E-05	cellular protein metabolic process
	GO:0015991	6,37E-04	ATP hydrolysis coupled proton transport
	GO:0006414	8,35E-04	translational elongation
	GO:0051603	2,33E-03	proteolysis involved in cellular protein catabolic process
	GO:0007017	2,39E-03	microtubule-based process
	GO:0007018	9,58E-03	microtubule-based movement
	GO:0015986	9,71E-03	ATP synthesis coupled proton transport
	GO:0006413	1,46E-02	translational initiation
	GO:0051258	2,20E-02	protein polymerization
	GO:0006284	4,40E-02	base-excision repair
	GO:0006165	4,80E-02	nucleoside diphosphate phosphorylation
	GO:0006183	4,80E-02	GTP biosynthetic process
	GO:0006228	4,80E-02	UTP biosynthetic process
	GO:0006241	4,80E-02	CTP biosynthetic process
	GO:0006952	4,91E-02	defense response
	GO:0008284	4,91E-02	positive regulation of cell proliferation
	GO:0030163	4,97E-02	protein catabolic process
Componente celular	GO:0005840	7,73E-25	ribosome
	GO:0000786	4,12E-08	nucleosome
	GO:0005737	2,25E-06	cytoplasm
	GO:0015935	2,50E-05	small ribosomal subunit
	GO:0030286	4,36E-04	dynein complex
	GO:0005839	2,33E-03	proteasome core complex
	GO:0005875	2,39E-03	microtubule associated complex
	GO:0031514	3,25E-03	motile cilium
	GO:0019773	4,96E-03	proteasome core complex, alpha-subunit complex
	GO:0015934	6,38E-03	large ribosomal subunit
	GO:0005634	8,27E-03	nucleus
	GO:0005829	1,23E-02	cytosol
	GO:0005643	2,17E-02	nuclear pore
	GO:0043234	2,20E-02	protein complex
	GO:0016272	2,29E-02	prefoldin complex
	GO:0030126	3,68E-02	COPI vesicle coat
	GO:0033180	4,33E-02	proton-transporting V-type ATPase, V1 domain
	GO:0005853	4,79E-02	eukaryotic translation elongation factor 1 complex
Função molecular	GO:0003735	2,45E-31	structural constituent of ribosome
	GO:0004298	1,13E-05	threonine-type endopeptidase activity
	GO:0051082	5,67E-05	unfolded protein binding
	GO:0046961	2,02E-03	proton-transporting ATPase activity, rotational mechanism
	GO:0003743	2,46E-03	translation initiation factor activity
	GO:0005509	5,47E-03	calcium ion binding
	GO:0003777	9,58E-03	microtubule motor activity
	GO:0046933	9,71E-03	proton-transporting ATP synthase activity, rotational mechanism
	GO:0003746	2,08E-02	translation elongation factor activity

GO:0015078	2,30E-02	hydrogen ion transmembrane transporter activity
GO:0005516	3,01E-02	calmodulin binding
GO:0005506	3,18E-02	iron ion binding
GO:0019239	4,08E-02	deaminase activity
GO:0008536	4,17E-02	Ran GTPase binding
GO:0000049	4,61E-02	tRNA binding
GO:0004550	4,80E-02	nucleoside diphosphate kinase activity
GO:0019843	4,88E-02	rRNA binding

3.3.4 Definição dos limites das UTRs

A definição dos limites das UTRs de genes codificadores é relevante. Além de permitir a anotação estrutural dos genes, essas regiões são importantes em processos pós-transcricionais de regulação de expressão gênica do parasito por conterem cis-elementos, como sítios ligantes de proteínas, que compõem a maquinaria que direciona a regulação pós-transcricional dos mRNAs [22]. Ressaltamos ainda, como demonstramos recentemente, que a definição das UTRs ganha relevância para compreender processos de regulação de expressão gênica por potencialmente abrigarem ncRNAs regulatórios [82].

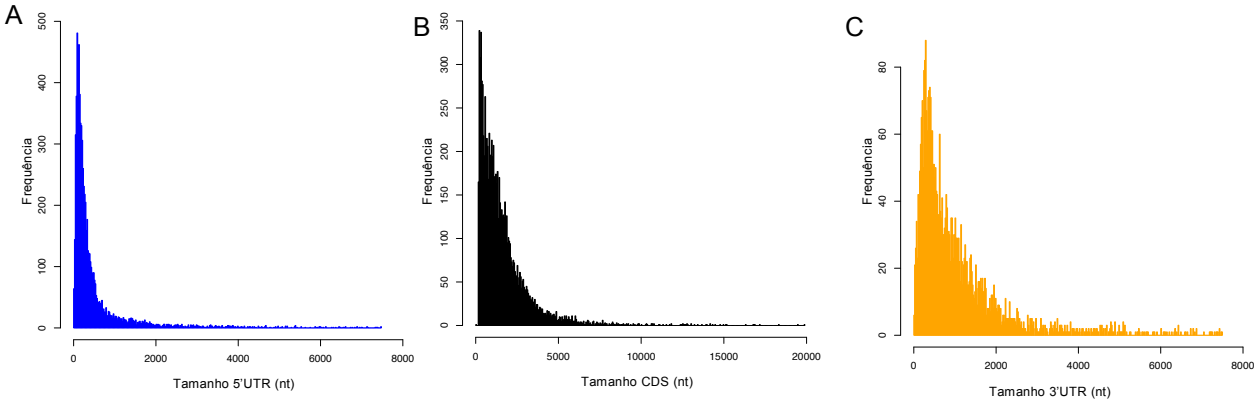
Para a predição dos limites das 3'UTRs foram utilizados os dados do sequenciamento de RNA total de *L. braziliensis*. Todas as réplicas dos três estágios de desenvolvimento foram agrupadas totalizando um conjunto de 677.853.338 *reads* das 9 amostras sequenciadas. Desse total, 1.834.437 (~0,003%) continham evidência de poliadenilação e foram mapeadas no genoma de referência (*L. braziliensis* 2903 – versão 30, TriTryDB). 82,65% das *reads* mapearam no genoma de referência e foram utilizadas para a definição dos possíveis sítios de poliadenilação. Com os sítios poliA definidos, cada CDS anotada foi ligada ao sítio poliA à jusante mais próximo e com maior cobertura (sítio principal). Dessa maneira, a região de 3'UTR foi definida considerando a coordenada do códon de parada da CDS e a coordenada do sítio de poliA. Foi possível identificar a 3'UTR de 3.568 CDSs (38%) com tamanho variando de 3 a 7.487 nt (Figura 25 – C) e mediana de 692 nt.

A definição dos limites das 5'UTRs foi feita utilizando uma biblioteca construída para identificar o sítio aceptor de SL (*Spliced Leader*) em promastigotas procíclicas de *L. braziliensis*, descrita na seção 2.2.8. Do total de 51.968.546 reads, 73,44% mapearam no genoma de referência (*L. braziliensis* 2903 – versão 30, TriTryDB) e foram utilizadas na definição dos sítios aceptores de SL. Após a definição desses sítios, cada CDS anotada foi ligada ao sítio aceptor à montante

mais próximo e com maior cobertura (sítio principal). Assim, a região de 5'UTR foi definida considerando a coordenada do códon de início da CDS e a coordenada do sítio SL. Foi possível identificar a 5'UTR de 7.494 CDSs (81%) com tamanho variando de 3 a 7.480 nt (Figura 25 – A) e mediana de 228 nt.

As coordenadas de início e fim das CDSs anotadas na versão 30 do genoma de *L. braziliensis* 2903 (TriTrypDB) foram utilizadas para calcular o tamanho das CDSs. A mediana do tamanho das CDSs foi de 1.176 nt, variando de 49 a 19.893 nt (Figura 25 - B).

Figura 25. Distribuição do tamanho das 5'UTRs, CDSs e 3'UTRs de *Leishmania braziliensis*.



A) 5'UTR; B) CDS e C) 3'UTR. Os tamanhos estão representados em nucleotídeos (nt). Fonte: autoral

Estudo similar foi feito para *Leishmania mexicana* e nela o tamanho mediano da 5'UTR, CDS e 3'UTR foi definido como 242 nt, 1395 nt e 584 nt, respectivamente [141]. Dillon e colaboradores, em estudo similar com *Leishmania major*, definiram o tamanho mediano para CDSs de 1.241 nt com variação entre 64 e 52.178 nt e as medianas do tamanho para 5'UTR e 3'UTR como sendo 233 e 517 nt quando o sítio primário foi considerado [56]. *Leishmania major* é a espécie de *Leishmania* com o genoma melhor montado e anotado mas, apesar dos problemas de montagem de anotação no genoma de *L. braziliensis*, os resultados encontrados das medianas dos tamanhos de CDS, 5'UTR e 3'UTR foram similares.

Sendo assim, a estrutura dos genes de *L. braziliensis* representada pelos valores medianos de cada elemento resulta em um mRNA de 2.096 nt, com a 5'UTR, CDS e 3'UTR correspondentes a 11, 56 e 33% do tamanho, respectivamente (228 nt da 5'UTR, 1.176 nt da CDS e 692 nt da 3'UTR).

4. *NON-CODING* RNAS EM *Leishmania braziliensis*

4.1 VISÃO GERAL

Neste capítulo estão descritas a identificação e análise dos ncRNAs putativos em *Leishmania braziliensis*. Para a definição dos ncRNAs putativos, utilizamos os dados gerados pelo sequenciamento de RNA total (descrito na seção 3.2.1.2) dos três estágios de desenvolvimento do parasito. Desenvolvemos uma metodologia para identificar possíveis transcritos em regiões não CDS ou na fita complementar à da CDS anotada no genoma. Além disso, os ncRNAs putativos identificados foram analisados para verificar a possibilidade de serem CDSs não anotadas no genoma, examinando a presença de domínios conhecidos de proteínas. A predição de características presentes em RNAs não codificadores também foi considerada para aumentar a robustez dos resultados. Foram considerados como ncRNAs putativos de *L. braziliensis* os candidatos que não apresentaram similaridade com domínios proteicos conhecidos e que tiveram pelo menos uma predição de ncRNA positiva. Após a identificação dos ncRNAs foi realizada a análise de expressão diferencial entre os três estágios de vida do parasito. Por último, foi feita a análise de prospecção dos dados dos ncRNAs identificados para encontrar possíveis padrões nos ncRNAs de *L. braziliensis*.

4.2 MATERIAIS E MÉTODOS

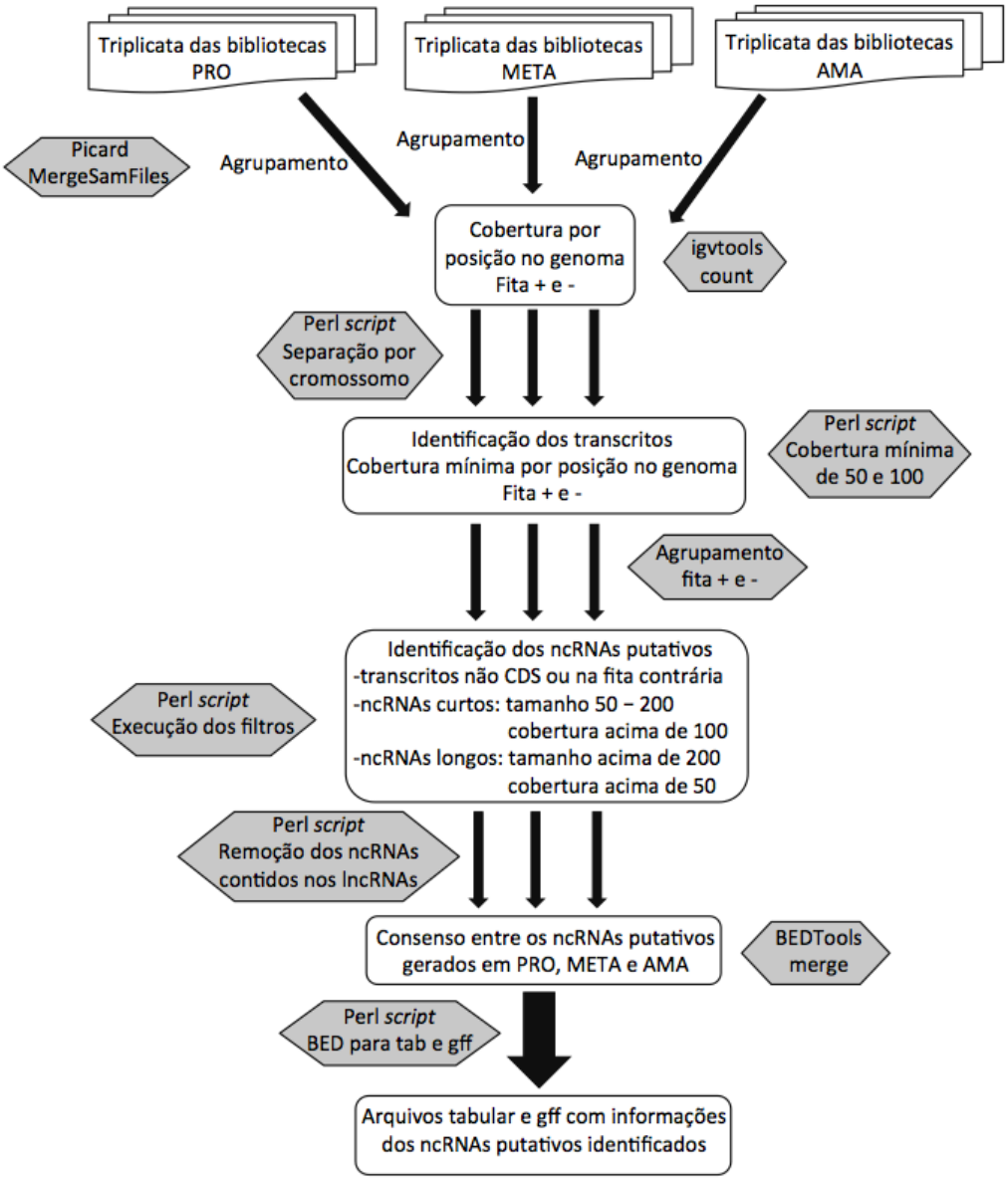
As bibliotecas de RNA total de *L. braziliensis* foram processadas e mapeadas no genoma de referência como descrito na seção 3.2.2. Todos os *scripts* de desenvolvidos e resultados citados estão disponíveis sob solicitação (akcruz@fmrp.usp.br).

4.2.1 Identificação dos ncRNAs putativos

O processo de identificação dos ncRNAs putativos inicia-se com a definição do conjunto dos transcritos mapeados pelo sequenciamento do RNA total dos três estágios de vida de *Leishmania braziliensis*: promastigota procíclico (PRO), promastigota metacíclico (META) e amastigota (AMA). Como a transcrição em *Leishmania* ocorre em blocos policistrônicos, a partir dos quais, pós transcricionalmente, o processamento (*splicing* e poliadenilação) leva à geração dos mRNAs maduros, a utilização de programas adequados aos organismos que fazem

transcrição monocistrônica, mostram-se inadequados para aplicação nos tripanossomatideo, pois frequentemente unidades policistrônicas são identificadas como um único transcrito. Para minimizar este problema, identificamos os transcritos de maneira alternativa, e empregamos os parâmetros como cobertura mínima (número de *reads* mapeadas), tamanho e localização de um transcrito para predição dos ncRNAs putativos (Figura 26).

Figura 26. Associação de abordagens usadas para identificação de ncRNAs putativos curtos e longos.



Foram utilizados os dados gerados pelo sequenciamento de RNA total de *Leishmania braziliensis*. Os programas Picard, IGVTools e BEDTools foram utilizados juntamente com *scripts* Perl para a identificação de ncRNAs putativos. PRO, META e AMA significam: promastigota procíclica, promastigota metacíclica e amastigota, respectivamente.O formato de arquivo gff é *General Feature Format*. Fonte: autoral

A definição dos transcritos inicia-se com o agrupamento das triplicatas dos três estágios de desenvolvimento do parasito para obtenção do maior número de *reads* possível, aumentando a chance de ncRNAs pouco expressos serem identificados. Esse agrupamento foi realizado pelo programa Picard (*MergeSamFiles*). Para definir os limites dos transcritos foi necessária a obtenção da cobertura de *reads* por posição no genoma e para isso, utilizamos o programa igvtools count [107] com os parâmetros *--strand first* (considera a fita original do transcrito) e *--windowSize 1* (cálculo da cobertura para cada posição no genoma). Um *script* Perl foi desenvolvido para dividir o arquivo de cobertura por cromossomo (*split_wig_IGVtools.pl*). Os arquivos de cobertura foram analisados por scripts Perl que definiram os limites dos transcritos na fita positiva (+) e na fita negativa (-) dos cromossomos (*identify_transcripts_pos.pl* e *identify_transcripts_neg.pl*) em dois conjuntos: um com regiões com cobertura mínima de 50 *reads* e outro com cobertura mínima de 100 *reads*. Os transcritos identificados tanto na fita + quanto na fita – foram agrupados em um único arquivo separados por cromossomo.

Foram considerados dois grupos de ncRNAs putativos, os curtos e os longos. Os transcritos de tamanho entre 50 e 200 com cobertura mínima (por posição) de 100 *reads* foram considerados ncRNAs curtos e aqueles de tamanho maior que 200 e cobertura mínima de 50 reads foram considerados ncRNAs longos (lncRNAs). Tanto os ncRNAs quanto os lncRNAs putativos não poderiam estar em regiões de CDSs conhecidas a não ser quando transcritos a partir da fita complementar à da CDS. O *script* Perl *identify_possible_ncRNAs.pl* foi desenvolvido para aplicação desses filtros nos transcritos identificados. A metodologia aplicada possibilitou a identificação de ncRNAs curtos internos a lncRNAs e para solucionar este problema, foi desenvolvido o *script* Perl *identify_overlap_nc-lnc.pl*. Ao final desta etapa, foram identificados os ncRNAs putativos curtos e longos para PRO, META e AMA, sendo gerados arquivos tabulares e no formato gff (*script* Perl *ncRNAs-tab_to_gff.pl*). O consenso das regiões de ncRNAs entre PRO, META e AMA foi gerado considerando a maior região de sobreposição entre os três estágios de vida. O programa *BEDTools merge* foi utilizado para gerar o consenso, com os parâmetros *–s* (agrupa somente regiões na mesma fita) e *–o distinct* (regiões não duplicadas). Foram gerados arquivos finais tabulares e no formato gff (*script* Perl *ncRNAs-tab_to_gff.pl*), disponíveis sob solicitação (akcruz@fmrp.usp.br).

4.2.2 Caracterização dos ncRNAs putativos

Depois de identificados os ncRNAs putativos, foi realizada a busca por domínios proteicos conhecidos nessas regiões, evitando assim, que possíveis CDSs não anotadas no genoma fossem consideradas ncRNAs putativos. Uma busca por similaridade de sequência nas 6 *frames* de leitura (blastx) foi feita contra o banco de dados Pfam, que contém um grande conjunto de famílias proteicas [143]. O blastx foi executado (parâmetro de *e-value* menor que 10^{-6}) no cluster do LCCA (Laboratório de Computação Científica Avançada) da USP. Os ncRNAs putativos que apresentaram *hit* com o Pfam não foram considerados na análise.

Para aumentar a confiabilidade de que as sequências identificadas são candidatos a ncRNAs foram utilizados alguns preditores de características específicos de RNAs não codificadores. Foram utilizados os programas PORTRAIT, RNAcon e snoscan descritos na seção 2.2.11. O programa RNASpace não foi incluído pois a execução em larga escala só é possível enviando as sequências para os autores da ferramenta e quando requisitamos a execução com os ncRNAs putativos, a resposta foi que seria possível somente em junho de 2017. Foram incluídos o programa tRNAscan-SE, que combina o resultado de três programas preditores de tRNA, pós-processando os resultados para extrair o que há de melhor de cada um dos preditores [144] e o programa ptRNApred que faz a predição de ncRNAs envolvidos em modificação pós-transcricional e replicação de DNA [145]. O ptRNApred inicialmente identifica se a sequência é ou não um RNA pós-transcricional e depois identifica sua possível família (snRNA, snoRNA, RNase P, RNase MRP, Y RNA e RNA telomerase). A ferramenta utiliza um algoritmo baseado em SVM (*Support Vector Machine*) e considera para predição a sequência, a estrutura e as propriedade termodinâmicas das sequências analisadas.

Ao final do processo, foi criada a pontuação para cada ncRNA putativo que toma cada predição positiva como um ponto. Os candidatos sem nenhuma pontuação foram desconsiderados da análise.

4.2.3 Conservação dos ncRNAs putativos

Para a análise de conservação dos ncRNAs identificados foi utilizada a mesma metodologia e os mesmos parâmetros descritos para a busca em *L. donovani* (Seção 2.2.12). Foi feita uma busca por similaridade de sequência com os

genomas de *L. major* Friedlin, *L. infantum* JPCM5, *L. amazonensis* MHOMBR71973M2269, *L. donovani* BPK282A1, *L. tarentolae* ParrotTarII, *L. enriettii* LEM3045, *T. brucei* TREU927 (TriTrypDB – versão 29). Um *hit* foi considerado positivo com um *e-value* $<10^{-5}$.

4.2.4 Expressão diferencial dos ncRNAs putativos

Para a análise da expressão diferencial dos ncRNAs foi utilizada a mesma metodologia descrita para análise de DE em *L. braziliensis* (da Seção 3.2.4). Nessa metodologia empregamos o ambiente R [87] e a biblioteca DESeq do Bioconductor [88; 109]. A contagem das *reads* por ncRNA identificado foi feita com o programa FeatureCounts [108]. Foi desenvolvido um *script* R (disponível sob solicitação - akcruz@fmrp.usp.br) que remove previamente os ncRNAs com contagem de *reads* igual a zero e realiza a normalização e a geração de heatmaps complexos que integram informações de diferença de expressão e as características dos ncRNAs putativos. Foi utilizada a função R aheatmap (pacote NMF), com métrica de distância = *manhattan* e método de agrupamento = *complete linkage*. Foram considerados diferencialmente expressos os ncRNAs com um p-valor ajustado < 0,05.

4.2.5 Busca por padrão

Empregamos métodos de aprendizado não supervisionado para a busca de padrões nos ncRNAs putativos de *Leishmania braziliensis*. Essa abordagem inclui 5 passos principais: 1-) preparação dos dados; 2-) avaliação do potencial de agrupamento dos dados; 3-) definição do número ideal de grupos; 4-) agrupamento dos dados e 5-) validação do agrupamento realizado. Todas as etapas foram realizadas em ambiente R.

4.2.5.1 Representação vetorial dos ncRNAs

Parte da preparação dos dados para a busca por padrão envolve a representação vetorial dos ncRNAs, que se dá pela seleção de determinadas características básicas e preditas. Utilizando características informativas para análises de prospecção de dados estas foram classificadas como variáveis binárias para facilitar a busca por padrões. Embora essa categorização seja aceita e as

metodologias escolhidas sejam adequadas no contexto, é possível prever a existência de vieses não detectáveis. *Scripts* Perl foram desenvolvidos para filtrar a informação necessária e categorizar os resultados dos preditores e características básicas dos ncRNAs putativos (*Scripts* disponíveis sob solicitação - akcruz@fmrp.usp.br).

Para a informação da fita do ncRNA foi criada a categoria 1 (fita +) e 2 (fita -). A característica tamanho foi categorizada de duas formas: Tamanho_v1: categoria 1 (entre 50-100nt), 2 (entre 100-200nt), 3 (entre 200-500nt) e 4 (>500nt); Tamanho_v2: categoria 1 (entre 50-200nt) e 2 (>200nt).

Os resultados do programa PORTRAIT foram categorizados como 0 (maior probabilidade de ser codificador), 1 (maior probabilidade de ser não codificador) e NA (Não se aplica = sequências sem predição, menores que 80nt ou região de gap no genoma). Os programas snoscan e tRNAscan-SE tiveram duas categorias: 0 para predição negativa e 1 para predição positiva.

O programa ptRNApred teve duas versões diferentes de categorização, a primeira (ptRNApred_v1) mais detalhada, de acordo com a família predita: 0 para predição negativa, 1 – snRNA, 2 – snoRNA, 3 – RnaseP, 4 – RNaseMRP, 5 – YRNA e 6 – telomerase-(fração RNA). A versão mais simples da categorização do ptRNApred (ptRNApred_v2) considerou 0 para predição negativa e 1 para predição positiva.

O programa RNAcon foi categorizado de três formas diferentes, sendo que a primeira considerou uma categoria diferente para cada classe do RNA (RNAcon_v1): 0 – mRNA, 1 – intron-gp-1, 2 – intron-gp-2, 3 – IRES, 4 – 5S-rRNA, 5 – SSU-rRNA5, 6 – HACA-BOX, 7 – 5.8S-rRNA, 8 – RIBOZYME, 9 – MIRNA, 10 – tRNA, 11 – tmRNA, 12 – SECIS, 13 – RIBOSWITCH, 14 – LEADER, 15 – CD-BOX, 16 – 6S-RNA, 17 – SRP e 18 – T-box. A segunda forma (RNAcon_v2) categorizou os ncRNAs como 0 para predição negativa de non-coding pelo RNAcon e 1 para predição positiva. A terceira forma (RNAcon_v3) agrupou classes similares de RNA: 0 – mRNA, 1 – intron-gp-1 e intron-gp-2, 2 – IRES, 3 – 5S-rRNA, SSU-rRNA5 e 5.8S-rRNA, 4 – HACA-BOX e CD-BOX, 5 – RIBOZYME, 6 – MIRNA, 7 – tRNA e tmRNA, 8 – SECIS, 9 – RIBOSWITCH, 10 – LEADER, 11 – 6S-RNA, 12 – SRP e 13 – T-box.

Assim, foi gerado um vetor descritivo para cada ncRNA com variáveis escalares. A tabela 18 apresenta um exemplo da representação vetorial que avaliamos.

Tabela 18. Representação vetorial dos ncRNAs.

ncRNA ID	Fita	Tamanho_v1	Tamanho_v2	PORTRAIT	ptRNApred_v1	ptRNApred_v2	snoscan	RNAcon_v1	RNAcon_v2	RNAcon_v3	tRNAscan-SE
LbrM2903_01_lncRNA1	2	3	2	1	3	1	0	1	1	1	0
LbrM2903_01_lncRNA2	2	3	2	1	3	1	0	19	0	0	0
LbrM2903_01_ncRNA1	2	2	1	0	1	1	0	3	1	2	0
LbrM2903_01_lncRNA3	2	3	2	1	3	1	0	1	1	1	0
LbrM2903_01_ncRNA2	2	1	1	NA	2	1	0	4	1	3	0

Exemplo de representação vetorial que será utilizada na análise de prospecção de dados.

Para possibilitar a utilização de técnicas de prospecção de dados cada ncRNA putativo identificado, e posteriormente anotado, foi descrito na forma de um vetor. A representação vetorial de todos os ncRNAs armazenados é uma tabela, na qual as linhas representam cada um dos ncRNAs e as colunas representam as características preditas. É possível criar dezenas de representações vetoriais diferentes, cada uma sendo mais ou menos informativa. Dessa forma, o poder informativo de cada representação vetorial pode ser aferido quando as técnicas de prospecção de dados são empregadas.

Após a importação desses dados no ambiente R, realizamos a remoção de ncRNAs com alguma informação ausente, pois estes casos inserem um viés importante nas análises de agrupamento. Posteriormente, realizamos a normalização dos dados (em cada observação subtraímos a média e dividimos pelo desvio padrão).

4.2.5.2 Agrupamento

Após a preparação dos dados, foi avaliado seu potencial de agrupamento utilizando a estatística de Hopkins (função R get_clust_tendency – pacote factoextra), cujo resultado indica que quanto mais próximo de zero, maior o potencial de agrupamento dos dados [146].

A etapa de agrupamento possui duas estratégias principais: métodos de particionamento (*k-means* e PAM) e agrupamento hierárquico. Sendo assim, os três métodos foram testados; *k-means*, que utiliza a média entre os elementos de um grupo para calcular a distância com um próximo elemento a ser agrupado [147]; PAM (*Partitioning Around Medoids*), que utiliza como centro dos agrupamentos os 'medóides' (elementos que proporcionam a menor soma de dissimilaridades no grupo) [148] e agrupamento hierárquico que é um método alternativo ao k-means que não necessita de um número definido de grupos [149].

Para a etapa de definição do número ideal de grupos foi utilizado o teste estatístico GAP, proposto por R. Tibshirani em 2001 [150], método que compara a variância entre cada um dos elementos de um grupo com o seu valor esperado em uma distribuição de referência nula (distribuição sem agrupamentos claros). Utilizamos máximo de 50 grupos avaliados e bootstrap de 100x. A análise e a visualização dos resultados (gráfico Elbow) foram feitas utilizando a função fviz_nbclust (pacote factoextra).

Após a definição do número ideal de grupos, foi feita a etapa de agrupamento. Para *k-means* foi utilizada a função em R kmeans (pacote stats), identificando o número de grupos desejado. Para PAM foi utilizada a função em R pam (pacote cluster), identificando o número de grupos desejado. Para clusterização hierárquica foi utilizada a função em R hclust (pacote stats - realiza agrupamento aglomerativo), com métrica de distância = *manhattan* (mesma utilizada pelo algortimo PAM em uma das suas etapas) e método de ligação = *complete linkage*. Para identificar as características predominantes dentro de cada grupo gerado pelo PAM, foi utilizada a função R aggregate (pacote stats) que identifica o número de ocorrências de cada categoria de todas as variáveis para cada um dos grupos.

Ao final do processo, foi avaliada a qualidade dos gupos gerados com a análise *silhouette* [151], que estima quão bem um elemento de um conjunto de dados está agrupado aos seus semelhantes. Nessa análise visual, quanto mais próximo de 1 a 'largura' da silhueta de um elemento, maior a sua semelhança aos seus vizinhos e, portanto, melhor o seu agrupamento. Foi utilizada a função R silhouette (pacote cluster).

As versões dos pacotes R utilizados estão no arquivo session_info_R.txt disponível sob solicitação (akcruz@fmrp.usp.br).

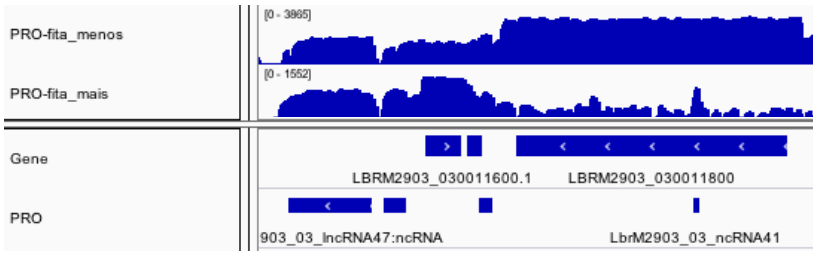
4.3 RESULTADOS E DISCUSSÃO

4.3.1 Identificação dos ncRNAs putativos

A identificação dos ncRNAs putativos de *L. braziliensis* foi feita utilizando dados de sequenciamento de RNA total. Os principais desafios eram identificar individualmente os transcritos (picos de *reads*) candidatos a ncRNAs e excluir aqueles com maior chance de serem CDSs não anotadas no genoma. O processo de identificação dos ncRNAs putativos tem início com a definição dos transcritos fita específico, com base na cobertura do genoma por nucleotídeo. Definimos dois limites mínimos iniciais para identificar os transcritos: cobertura mínima de 50 e 100 *reads* por nucleotídeo. Dessa maneira, foram identificados para promastigota procíclico (PRO) 18.991 transcritos na fita + e 17.111 transcritos na fita - com cobertura mínima de 50 e 20.017 transcritos na fita + e 18.769 transcritos na fita - com cobertura mínima de 100. Para promastigota metacíclico (META) os resultados mostraram 18.576 na fita + e 16.987 transcritos na fita - com cobertura mínima de 50 e 20.481 na fita + e 18.133 transcritos na fita - com cobertura mínima de 100. Em amastigota (AMA) foram identificados 16.715 transcritos na fita + e 14.830 na fita - com cobertura mínima de 50 e 20.825 transcritos na fita + e 18.058 na fita - com cobertura mínima de 100 (arquivos de resultados disponíveis sob solicitação - akcruz@fmrp.usp.br).

Foram considerados dois tipos de ncRNAs, os curtos (<200 nt) e os longos (>200 nt). Devido ao menor tamanho dos ncRNAs curtos, definimos que transcritos < 200 nt deveriam apresentar uma cobertura mínima 100 *reads* por nucleotídeo. Para os transcritos candidatos a ncRNAs longos (lncRNAs) definimos uma cobertura mínima de 50 *reads* por nucleotídeo. Além disso, os transcritos não poderiam estar em regiões de CDS se estivessem na mesma fita (Figura 27).

Figura 27. ncRNA localizado interno a uma CDS mas na fita contrária.

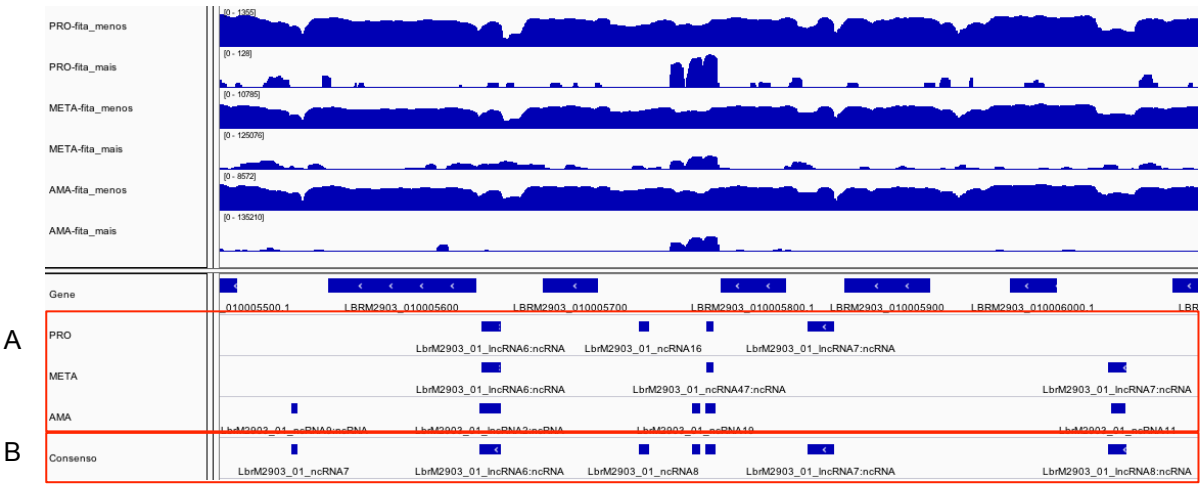


Região do cromossomo 03 de *Leishmania braziliensis* com um ncRNA (LbrM2903_03_ncRNA41) identificado em promastigota procíclica sobreposto a uma CDS mas na fita contrária. Na figura estão representadas as coberturas na fita + e na fita - (PRO-fita_mais e PRO-fita_menos), os genes anotados (Gene) e os ncRNAs putativos (PRO). Figura gerada no IGV. Fonte: autoral

Os ncRNAs curtos contidos em ncRNAs longos foram removidos. Foram identificados 9.415, 9.629 e 9.717 ncRNAs putativos para PRO, META e AMA, respectivamente (Figura 28 – A).

Para a anotação e análise foi gerado um consenso dos ncRNAs putativos entre os três estágios de desenvolvimento analisados (Figura 28 - B).

Figura 28. ncRNA putativos estágio específicos e consenso.



Região do cromossomo 01 de *Leishmania braziliensis*. A) ncRNAs putativos estágio específicos (PRO, META e AMA); B) ncRNAs putativos consenso entre os três estágios de desenvolvimento. Na figura também estão representadas as coberturas na fita + e na fita – para os três estágios de desenvolvimento e os genes anotados (Gene). Figura gerada no IGV. Fonte: autoral

Ao final, foram identificados 12.050 ncRNAs putativos para *Leishmania braziliensis*, sendo 4.073 ncRNAs curtos e 7.977 ncRNAs longos (Tabela 19).

Tabela 19. Número de ncRNAs identificados por cromossomo de *L. braziliensis*.

Cromossomo	# ncRNAs curtos	# ncRNAs longos	Total
LbrM2903_01	43	73	116
LbrM2903_02	39	105	144
LbrM2903_03	42	83	125
LbrM2903_04	71	124	195
LbrM2903_05	55	128	183
LbrM2903_06	80	139	219
LbrM2903_07	62	139	201
LbrM2903_08	66	139	205
LbrM2903_09	83	103	186
LbrM2903_10	91	201	292
LbrM2903_11	65	145	210

LbrM2903_12	37	87	124
LbrM2903_13	63	143	206
LbrM2903_14	100	156	256
LbrM2903_15	74	121	195
LbrM2903_16	77	170	247
LbrM2903_17	68	159	227
LbrM2903_18	86	213	299
LbrM2903_19	104	172	276
LbrM2903_20.1	218	423	641
LbrM2903_20.2	101	182	283
LbrM2903_21	91	149	240
LbrM2903_22	64	169	233
LbrM2903_23	108	194	302
LbrM2903_24	112	163	275
LbrM2903_25	105	197	302
LbrM2903_26	116	187	303
LbrM2903_27	155	274	429
LbrM2903_28	128	288	416
LbrM2903_29	136	265	401
LbrM2903_30	198	338	536
LbrM2903_31	179	432	611
LbrM2903_32	198	303	501
LbrM2903_33	190	362	552
LbrM2903_34	259	489	748
LbrM2903_35	305	550	855
Scaffolds	104	412	516
Total	4073	7977	12050

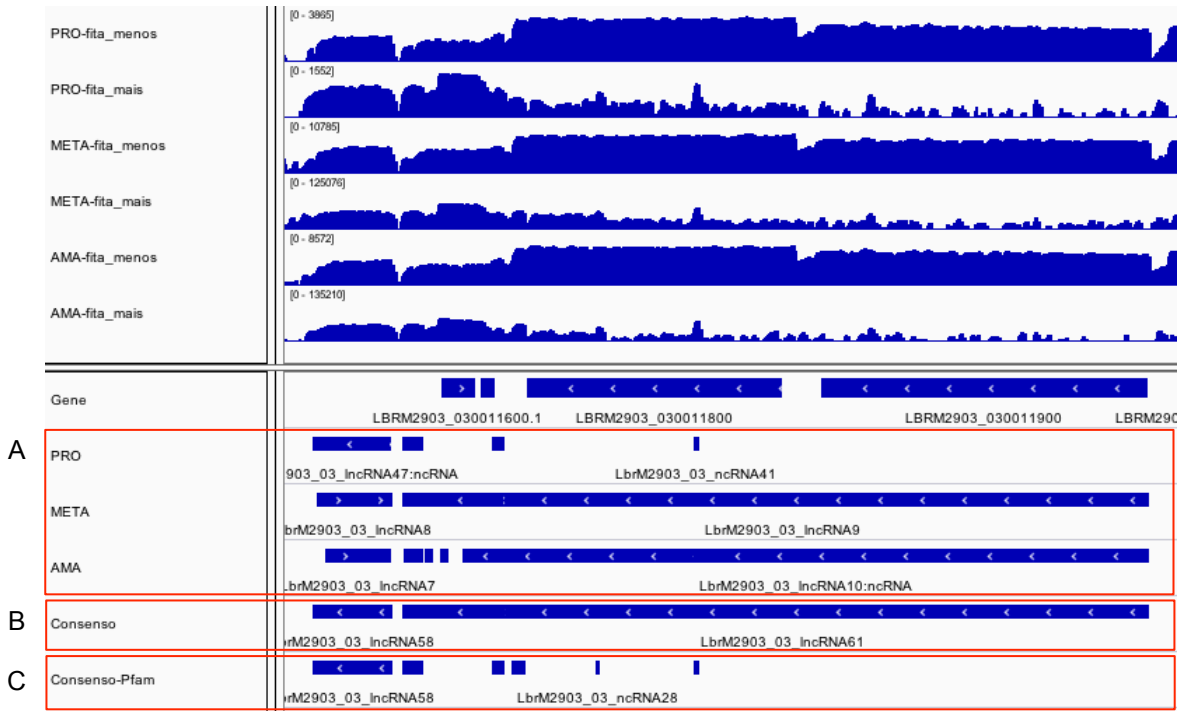
Portanto, *L. braziliensis* possui 9.269 genes codificadores de proteínas e 12.050 ncRNAs putativos. Comparando-se com outros organismos, notamos uma equivalência na proporção codificadores de proteínas versus não codificadores de proteína. Em *Homo sapiens*, há 20.310 genes codificadores de proteínas e 22.529 ncRNAs (GRCh38 - http://www.ensembl.org/Homo_sapiens/Info/Annotation - acessado em maio de 2017). Para *Caenorhabditis elegans*, estão anotados 20.362 genes codificadores de proteínas e 24.719 ncRNAs (WBcel235 - http://www.ensembl.org/Caenorhabditis_elegans/Info/Annotation - acessado em maior de 2017).

4.3.2 Caracterização dos ncRNAs putativos

Após a predição dos ncRNAs putativos, foram identificados os transcritos que poderiam ser RNAs codificadores não anotados. Essa análise considerou a

presença de domínios proteicos conhecidos (bando de dados Pfam) nos ncRNAs putativos. Do total de 12.050 ncRNAs putativos, 659 tiveram similaridade significativa com domínios proteicos e foram removidos da análise, restando 11.391 ncRNAs putativos (Figura 29).

Figura 29. Consenso de ncRNAs putativos após filtro com o Pfam.



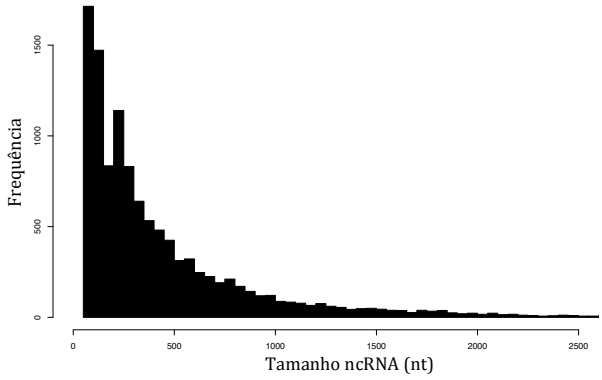
Região do cromossomo 03 de *Leishmania braziliensis*. A) ncRNAs putativos estágio específicos (PRO, META e AMA); B) ncRNAs putativos consenso entre os três estágios de desenvolvimento. C) Consenso dos ncRNAs putativos após a remoção dos candidatos com similaridade a algum domínio protéico (Pfam). Na figura estão representadas as coberturas na fita + e na fita – para os três estágios de desenvolvimento (painel superior) e os genes anotados (Gene). Figura gerada no IGV. Fonte: autoral

Os programas PORTRAIT, RNAcon, ptRNApred, snoscan e tRNAscan-SE foram utilizados para predizer características relacionadas a RNAs não codificadores, cada um com sua especificidade (detalhes na seção 2.2.11). Foram preditos como possíveis ncRNAs 7.086 (62%) candidatos pelo programa PORTRAIT, o RNAcon apresentou uma predição positiva para 7.942 (69%) candidatos e o ptRNApred para 10.216 (89%). O snoscan identificou 242 candidatos como possíveis snoRNAs e o programa tRNAscan-SE fez a predição positiva para 27 transcritos como candidatos a tRNAs.

Foram selecionados os candidatos com pelo menos uma predição positiva para o restante das análises, totalizando 11.376 ncRNAs putativos em *L.*

braziliensis. Esses ncRNAs putativos apresentaram como mediana de tamanho 281 nt, sendo o tamanho mínimo de 50 nt e máximo de 9.960 nt (Figura 30).

Figura 30. Distribuição dos tamanhos dos ncRNAs putativos.



Para a montagem do gráfico foram considerados somente os ncRNAs de tamanho máximo de 2.500 nt. Gráfico feito em ambiente R. Fonte: autoral

Do total de 11.376 ncRNAs putativos identificados em *L. braziliensis*, 4.044 foram ncRNAs curtos e 7.332 ncRNAs longos. Com relação a direção de transcrição, 5.877 ncRNAs putativos estão na fita + e 5.499 da fita -.

4.3.3 Conservação dos ncRNAs putativos

A análise de conservação dos ncRNAs putativos, indicativa de pressão seletiva, foi feita para investigar a possível relevância funcional dos candidatos. As sequências dos ncRNAs putativos de *Leishmania braziliensis* foram comparadas às sequências genômicas de sete tripanossomatídeos, e os resultados foram: *L. amazonensis*: 3.942 (35%) ncRNAs conservados; *L. donovani*: 4.554 (40%) ncRNAs conservados; *L. enriettii*: 3.233 (28%) ncRNAs conservados; *L. infantum*: 4.598 (40%) ncRNAs conservados; *L. major*: 4.629 (41%) ncRNAs conservados; *L. tarentolae*: 3.032 (27%) ncRNAs conservados e *T. brucei*: 756 (7%) ncRNAs conservados.

A conservação foi um critério utilizado na pontuação dos ncRNAs putativos para seleção dos candidatos com maiores chances de serem ncRNAs funcionais. A tabela com a pontuação final está disponível sob solicitação (akcruz@fmrp.usp.br).

4.3.4 Expressão diferencial dos ncRNAs putativos

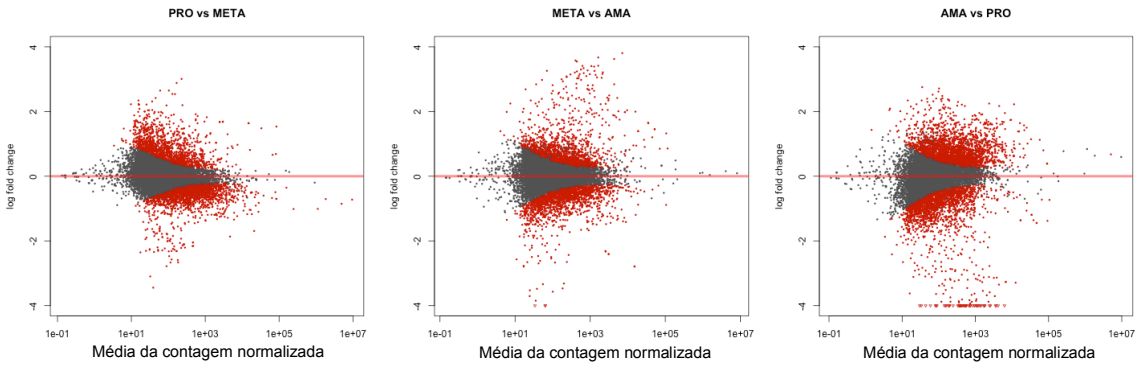
A análise de expressão diferencial comparou o nível de expressão dos 11.376 ncRNAs entre os seguinte estágios de desenvolvimento de *L. braziliensis*: promastigota procíclico versus promastigota metacíclico (PROvsMETA), promastigota metacíclico versus amastigota (METAvsAMA) e amastigota versus promastigota procíclico (AMAvsPRO). Foram identificados (p-valor ajustado < 0,05) 3.266 ncRNAs (29%) diferencialmente expressos (DE) na comparação PROvsMETA, empregando as condições *default* do programa DESeq. Utilizando um filtro de FC maior ou igual a 2, foram identificados 625 ncRNAs (5,5%) DE na mesma comparação. Nos extremos de nível de expressão diferencial detectamos genes com FC de até ~11,1 vezes para os sobre-representados e de até ~8,0 vezes para os sub-representados.

O resultado da comparação METAvsAMA revelou 3.058 ncRNAs (27%) no conjunto dos DE, empregando as condições *default* do programa DESeq. Considerando FC maior ou igual a 2, foram identificados 730 ncRNAs (6,4%) DE na mesma comparação. Nos extremos de nível de expressão diferencial detectamos genes com FC de até ~21,4 vezes para os sobre-representados e de até ~14,3 vezes para os sub-representados.

Na comparação AMAvsPRO foram identificados 4.380 ncRNAs (38%) DE utilizando os parâmetros *default* do programa DESeq. Considerando FC maior ou igual a 2, foram identificados 1448 ncRNAs (12,7%) DE na mesma comparação. Nos extremos de nível de expressão diferencial detectamos genes com FC de até ~7,1 vezes para os ncRNAs sobre-representados e de até ~44,7 vezes para os sub-representados.

Os transcritos não codificadores DE foram visualizados utilizando gráficos MA que mostram a relação entre a média de expressão (média da contagem de *reads* normalizada) e *fold change* para cada ncRNA (Figura 31).

Figura 31. ncRNAs diferencialmente expressos na comparação entre os estágios de desenvolvimento de *Leishmania braziliensis*.



A) PROvsMETA; B) METAvsAMA e C) AMAvsPRO. Os gráficos MA mostram a relação entre média de expressão (média da contagem normalizada) e *fold change*. Os pontos em vermelho representam os ncRNAs expressos em níveis significativamente diferentes entre os estágios de desenvolvimento comparados com um p-valor ajustado < 0,05. Fonte: autoral

Observando os gráficos MA, notamos novamente a maior semelhança na comparação PROvsMETA (menores FCs). E a comparação AMAvsPRO com o maior número de ncRNAs diferencialmente expressos.

A tabela com todos os resultados da análise de expressão diferencial pode ser obtida sob solicitação (akcruz@fmrp.usp.br). Os ncRNAs com as maiores diferenças (os 25 primeiros) nas comparações entre os estágios de desenvolvimento, tanto sobre- quanto sub-representados estão apresentados nas Tabelas 20, 21 e 22.

Na comparação PROvsMETA foram encontrados 1.847 ncRNAs sobre-representados, sendo o candidato com maior FC o *long non-coding* 2 do *Scaffold* 110 (LBRM2903_SCAF000110_lncRNA2) de *L. braziliensis* (Tabela 20). Considerando os sub-representados, foram identificados 1.421 ncRNAs DE, sendo o candidato com maior FC o *long non-coding* 56 do cromossomo 34 (LbrM2903_34_lncRNA56) de *L. braziliensis* (Tabela 20).

Tabela 20. ncRNAs diferencialmente expressos na comparação PROvsMETA.

Sobre-representados	
ncRNA ID	*FoldChange*
LBRM2903_SCAF000110_lncRNA2	11,07
LbrM2903_31_lncRNA196	8,69
LBRM2903_SCAF000172_lncRNA2	6,96
LBRM2903_SCAF000675_lncRNA1	6,44
LBRM2903_SCAF000344_lncRNA2	6,42

LBRM2903_SCAF000675_lncRNA2	6,13
LBRM2903_SCAF000344_lncRNA1	6,02
LBRM2903_SCAF000172_lncRNA1	5,62
LBRM2903_SCAF000344_ncRNA1	5,19
LBRM2903_SCAF000232_lncRNA1	5,13
LBRM2903_SCAF000001_lncRNA2	5,06
LBRM2903_SCAF000518_lncRNA1	5,04
LBRM2903_SCAF000173_lncRNA1	5,03
LBRM2903_SCAF000173_lncRNA2	5,02
LBRM2903_SCAF000694_lncRNA3	5,02
LBRM2903_SCAF000288_lncRNA3	5,00
LBRM2903_SCAF000288_lncRNA2	4,95
LBRM2903_SCAF000694_lncRNA2	4,90
LBRM2903_SCAF000679_lncRNA1	4,88
LBRM2903_SCAF000197_lncRNA1	4,87
LBRM2903_SCAF000227_ncRNA1	4,83
LBRM2903_SCAF000101_lncRNA1	4,80
LBRM2903_SCAF000550_lncRNA2	4,75
LBRM2903_SCAF000705_lncRNA1	4,72
LBRM2903_SCAF000670_lncRNA2	4,67
Sub-representados	
ncRNA ID	*FoldChange*
LbrM2903_34_lncRNA56	8,02
LbrM2903_34_lncRNA58	7,36
LbrM2903_34_lncRNA48	6,27
LbrM2903_34_lncRNA46	6,26
LbrM2903_34_lncRNA45	5,82
LbrM2903_34_lncRNA61	5,79
LbrM2903_33_lncRNA271	5,35
LbrM2903_06_ncRNA22	5,11
LBRM2903_SCAF000169_ncRNA1	4,98
LbrM2903_33_lncRNA76	4,73
LbrM2903_28_ncRNA99	4,73
LbrM2903_34_ncRNA23	4,67
LbrM2903_35_lncRNA323	4,67
LbrM2903_33_lncRNA269	4,66
LbrM2903_34_lncRNA57	4,59
LbrM2903_34_lncRNA55	4,55
LbrM2903_11_lncRNA31	4,27
LbrM2903_20.1_ncRNA49	4,16
LbrM2903_34_lncRNA54	4,13
LbrM2903_31_lncRNA203	4,12
LbrM2903_34_lncRNA62	4,12
LbrM2903_34_ncRNA193	4,02
LbrM2903_34_lncRNA377	4,01
LbrM2903_22_lncRNA69	3,99

LbrM2903_11_ncRNA57	3,98

A comparação METAvsAMA encontrou 1.614 ncRNAs sobre-representados, sendo o candidato com maior FC o *non-coding* 57 do cromossomo 11 (LbrM2903_11_ncRNA57) de *L. braziliensis* (Tabela 21). Considerando os sub-representados, foram identificados 1.446 ncRNAs DE, sendo o candidato com maior FC o *long non-coding* 162 do cromossomo 20.1 (LbrM2903_20.1_lncRNA162) de *L. braziliensis* (Tabela 21).

Tabela 21. ncRNAs diferencialmente expressos na comparação METAvsAMA.

Sobre-representados	
ncRNA ID	*FoldChange*
LbrM2903_11_ncRNA57	21,41
LbrM2903_07_ncRNA15	18,42
LBRM2903_SCAF000555_ncRNA1	17,60
LbrM2903_30_ncRNA198	15,38
LbrM2903_09_ncRNA2	14,15
LbrM2903_19_ncRNA11	13,16
LbrM2903_23_ncRNA3	12,45
LbrM2903_34_ncRNA99	12,02
LbrM2903_23_ncRNA1	11,74
LbrM2903_13_ncRNA1	11,10
LbrM2903_34_ncRNA98	10,81
LbrM2903_25_lncRNA2	7,68
LbrM2903_25_lncRNA1	7,64
LbrM2903_06_ncRNA22	7,33
LbrM2903_33_lncRNA76	7,23
LbrM2903_16_ncRNA1	6,72
LbrM2903_25_lncRNA142	6,67
LbrM2903_31_ncRNA107	6,62
LbrM2903_33_lncRNA289	6,44
LbrM2903_28_ncRNA99	6,36
LbrM2903_33_lncRNA290	6,34
LbrM2903_29_lncRNA259	6,06
LbrM2903_33_ncRNA163	5,92
LbrM2903_17_lncRNA158	5,85
LbrM2903_33_lncRNA269	5,82
Sub-representados	
ncRNA ID	*FoldChange*
LbrM2903_20.1_lncRNA162	14,27
LbrM2903_29_lncRNA94	13,98
LbrM2903_29_lncRNA93	12,75
LBRM2903_SCAF000477_lncRNA1	12,30

LbrM2903_08_lncRNA80	11,96
LbrM2903_08_lncRNA78	11,72
LbrM2903_20.1_lncRNA163	11,07
LbrM2903_20.1_lncRNA156	10,67
LBRM2903_SCAF000549_lncRNA1	10,64
LbrM2903_08_lncRNA79	10,61
LBRM2903_SCAF000639_lncRNA1	10,46
LbrM2903_20.1_lncRNA157	10,14
LBRM2903_SCAF000603_lncRNA1	9,86
LBRM2903_SCAF000156_ncRNA1	9,72
LbrM2903_20.1_lncRNA145	9,65
LbrM2903_20.1_lncRNA161	9,65
LBRM2903_SCAF000207_lncRNA1	9,54
LBRM2903_SCAF000153_lncRNA1	9,46
LbrM2903_20.1_lncRNA138	9,25
LbrM2903_20.1_lncRNA154	9,25
LBRM2903_SCAF000028_lncRNA1	9,10
LbrM2903_20.1_lncRNA153	8,95
LBRM2903_SCAF000652_lncRNA1	8,90
LBRM2903_SCAF000301_lncRNA1	8,90
LBRM2903_SCAF000275_ncRNA1	8,88

A comparação AMAvsPRO encontrou 1.929 ncRNAs sobre-representados, sendo o candidato com maior FC o *long non-coding* 162 do cromossomo 20.1 (LbrM2903_20.1_lncRNA162) de *L. braziliensis* (Tabela 22). Considerando os sub-representados, foram identificados 2.453 ncRNAs DE, sendo o candidato com maior FC o *non-coding* 2 do *Scaffold* 1 (LBRM2903_SCAF000001_ncRNA2) de *L. braziliensis* (Tabela 22).

Tabela 22. ncRNAs diferencialmente expressos na comparação AMAvsPRO.

Sobre-representados	
ncRNA ID	*FoldChange*
LbrM2903_20.1_lncRNA162	44,66
LBRM2903_SCAF000275_ncRNA1	36,99
LBRM2903_SCAF000477_lncRNA1	33,75
LbrM2903_20.1_lncRNA161	32,92
LbrM2903_20.1_lncRNA157	32,23
LbrM2903_20.1_lncRNA163	31,23
LbrM2903_20.1_lncRNA145	30,50
LBRM2903_SCAF000652_lncRNA1	30,34
LBRM2903_SCAF000549_lncRNA1	30,31
LBRM2903_SCAF000345_lncRNA1	30,26
LBRM2903_SCAF000515_ncRNA1	30,15

LBRM2903_SCAF000156_ncRNA1	30,13
LbrM2903_20.1_lncRNA154	29,97
LBRM2903_SCAF000153_lncRNA1	28,42
LBRM2903_SCAF000430_lncRNA1	27,56
LBRM2903_SCAF000336_lncRNA1	27,28
LBRM2903_SCAF000028_lncRNA1	26,69
LBRM2903_SCAF000639_lncRNA1	26,30
LBRM2903_SCAF000138_lncRNA1	26,09
LbrM2903_20.1_lncRNA156	25,98
LbrM2903_20.1_lncRNA158	25,45
LbrM2903_20.1_lncRNA153	25,09
LBRM2903_SCAF000207_lncRNA1	25,02
LbrM2903_20.1_lncRNA149	24,74
LbrM2903_20.1_lncRNA421	24,67
Sub-representados	
ncRNA ID	*FoldChange*
LBRM2903_SCAF000001_ncRNA2	7,15
LbrM2903_06_lncRNA77	6,63
LbrM2903_31_lncRNA198	6,20
LBRM2903_SCAF000518_lncRNA1	5,89
LbrM2903_06_lncRNA78	5,81
LBRM2903_SCAF000344_lncRNA1	5,50
LBRM2903_SCAF000518_lncRNA2	5,50
LbrM2903_33_lncRNA289	5,35
LBRM2903_SCAF000118_lncRNA1	5,34
LBRM2903_SCAF000675_lncRNA2	5,33
LBRM2903_SCAF000675_lncRNA1	5,28
LbrM2903_31_lncRNA196	5,28
LbrM2903_33_lncRNA290	5,24
LbrM2903_33_lncRNA42	5,21
LBRM2903_SCAF000344_lncRNA2	5,20
LbrM2903_35_lncRNA281	5,09
LbrM2903_33_ncRNA164	5,04
LbrM2903_27_ncRNA95	4,96
LBRM2903_SCAF000101_lncRNA1	4,88
LbrM2903_20.1_lncRNA201	4,77
LbrM2903_27_ncRNA155	4,77
LbrM2903_14_lncRNA98	4,71
LbrM2903_33_ncRNA13	4,68
LbrM2903_33_ncRNA165	4,64
LbrM2903_08_lncRNA93	4,57

Os ncRNAs diferencialmente expressos foram pontuados positivamente para definição de uma priorização dos candidatos a serem analisados

experimentalmente. A tabela com a pontuação final está disponível sob solicitação (akcruz@fmrp.usp.br).

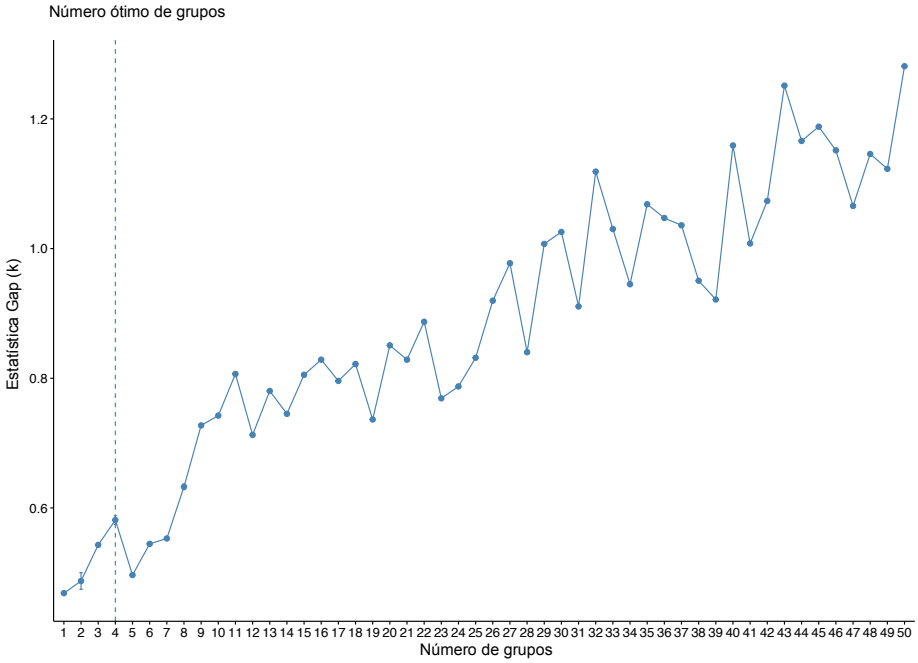
4.3.5 Busca por padrão

A busca por padrões nos ncRNAs identificados foi feita com os dados representados vetorialmente por variáveis escalares (Seção 4.2.5.1). Foram testadas algumas representações vetorias e a que inicialmente se mostrou como melhor para o agrupamento dos dados foi a que empregou as seguintes características: Fita, Tamanho_v1, PORTRAIT, ptRNApred_v1 e RNAcon_v3 (descritos na seção 4.2.5.1). Os transcritos sem os parâmetros mínimos exigidos pelo preditor para sua classificação foram removidos da análise, pois estes casos inserem um viés nos agrupamentos.

Obtivemos um valor de 0.257 para a estatística de Hopkins que avalia o potencial de agrupamento dos dados (quanto mais próximo de zero, maior o potencial de agrupamento).

A análise para a definição do número ideal de clusters, indicou que quatro é potencialmente a quantidade de grupos adequados para o conjunto de dados de ncRNAs de *L. braziliensis* (Figura 32).

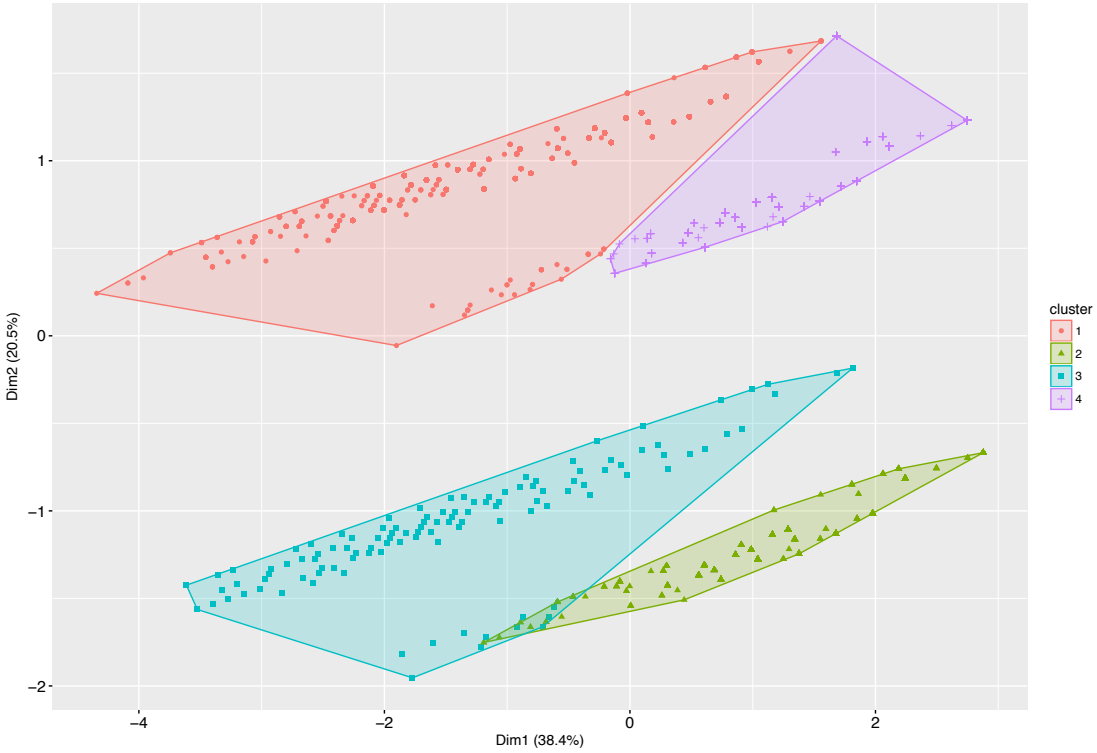
Figura 32. Gráfico *Elbow* gerado para o conjunto de dados de ncRNAs utilizando estatística gap.



Fonte: autoral

Conhecendo a quantidade ideal de grupos para o nosso conjunto de dados, foi realizado o agrupamento. Inicialmente utilizamos o método *k-means*, mas os grupos gerados estavam malformados (figura disponível sob solicitação - akcruz@fmrp.usp.br). Um dos fatores que pode influenciar os agrupamentos gerados pelo método *k-means* são 'valores extremos'. Justamente por trabalhar com médias, *k-means* é sensível a estes valores. Outro método utilizado foi o PAM (*Partitioning Around Medoids*), menos sensível a *outliers* por não utilizar a média como centro dos agrupamentos e sim medóides (vide seção 4.2.5.2). A figura 33 apresenta o agrupamento realizado com o método PAM para os ncRNAs de *L. braziliensis*.

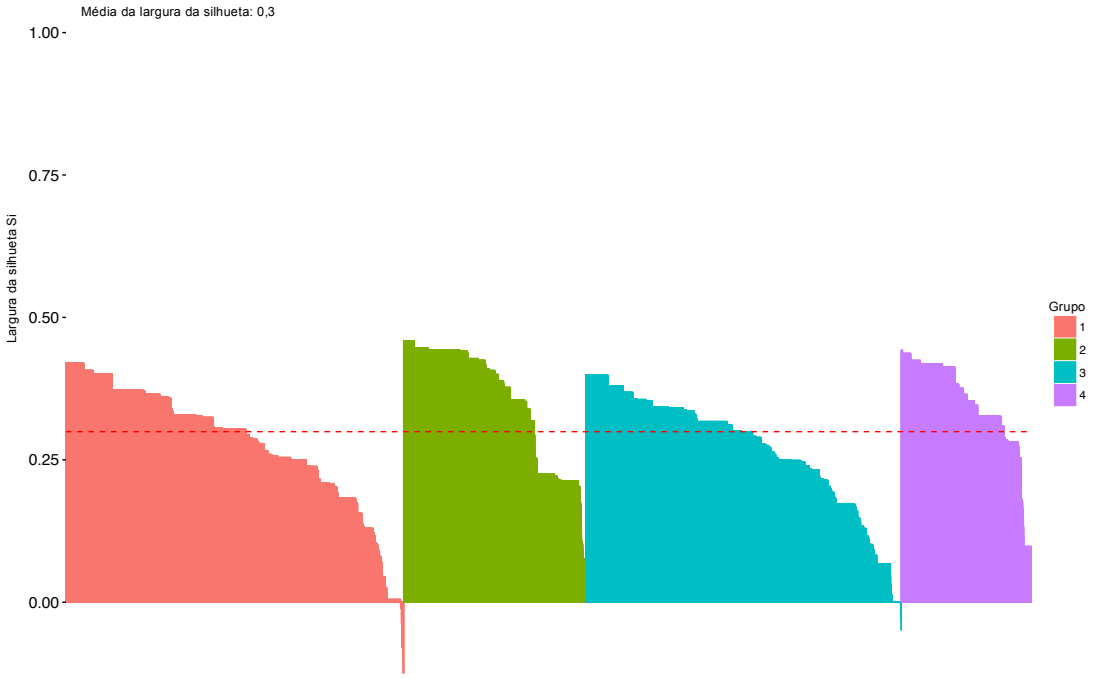
Figura 33. Agrupamento do conjunto de dados de ncRNAs de *L. braziliensis*.



Legenda de cores para cada cluster apresentada à direita. Fonte: autoral

Avaliamos a qualidade dos quatro grupos gerados pelo método PAM com a análise *silhouette* (Figura 34). Observamos que os quatro grupos apresentaram agrupamento consistente, com poucos elementos demonstrando dissimilaridade aos grupos.

Figura 34. Avaliação da qualidade dos grupos gerados pela PAM.



Legenda de cores para cada agrupamento apresentada à direita. Fonte: autoral

Os 4 grupos propostos foram analisados para identificar características predominantes dentro de cada um deles (Tabela 23).

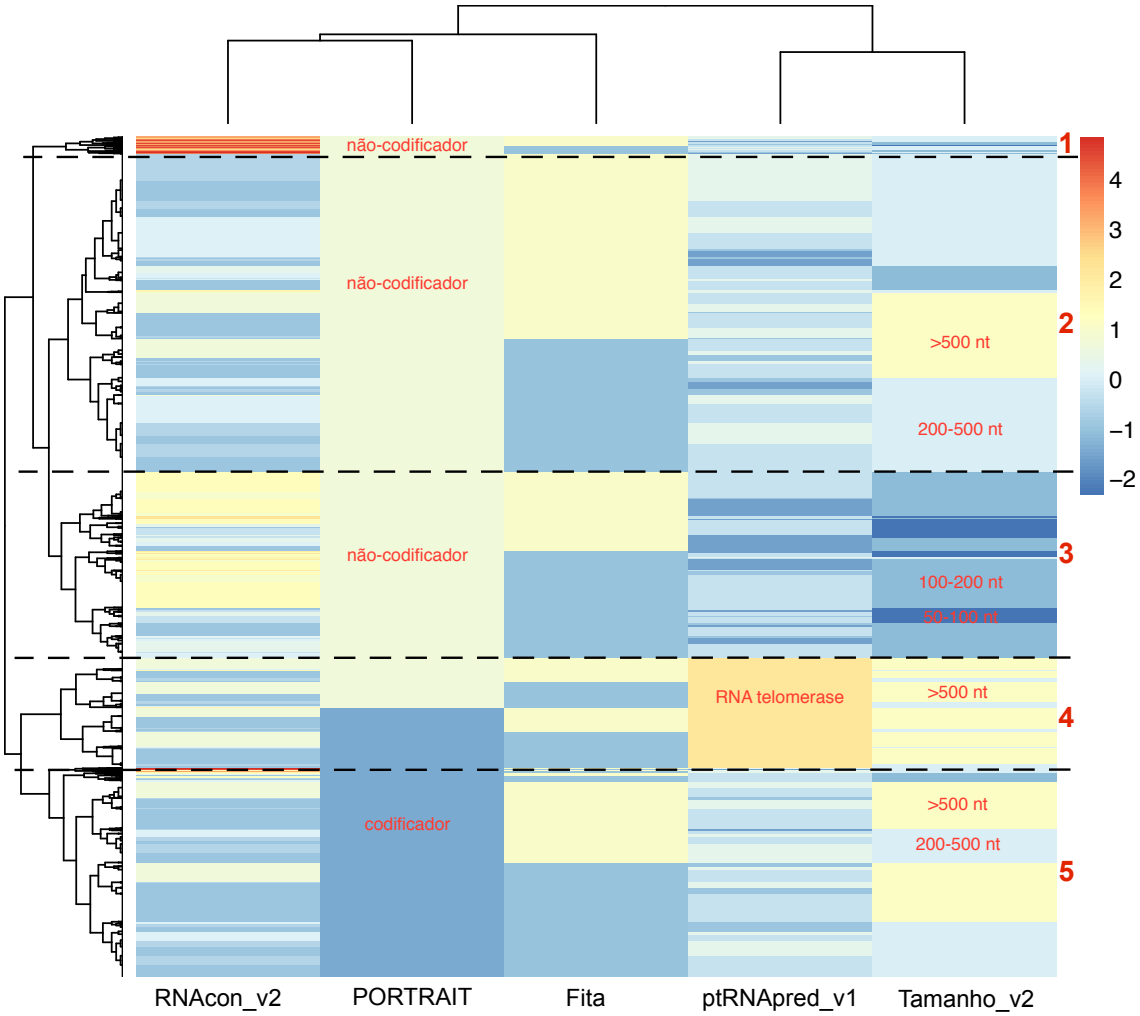
Tabela 23. Características predominantes dos grupos formados pela PAM.

Grupo	Tamanho_v2 (nt)	PORTRAIT	ptRNApred_v1	RNAcon_v2
1	> 200	codificador	snoRNA	mRNA
2	> 200	não-codificador	snoRNA	IRES
3	> 200*	codificador	RNA telomerase	SSU-rRNA
4	< 200	não-codificador	snoRNA	5.8S-rRNA

*Grupo exclusivo de ncRNAs putativos > 200 nt.

O último método de agrupamento testado foi o hierárquico que não considera um número de clusters definidos previamente. O conjunto de dados foi agrupado em 5 grupos principais (Figura 37). Observamos que as características Tamanho_v2 e prtRNApred_v1 se agrupam, assim como RNAcon_v2 e PORTRAIT e esses dois são agrupados a característica Fita.

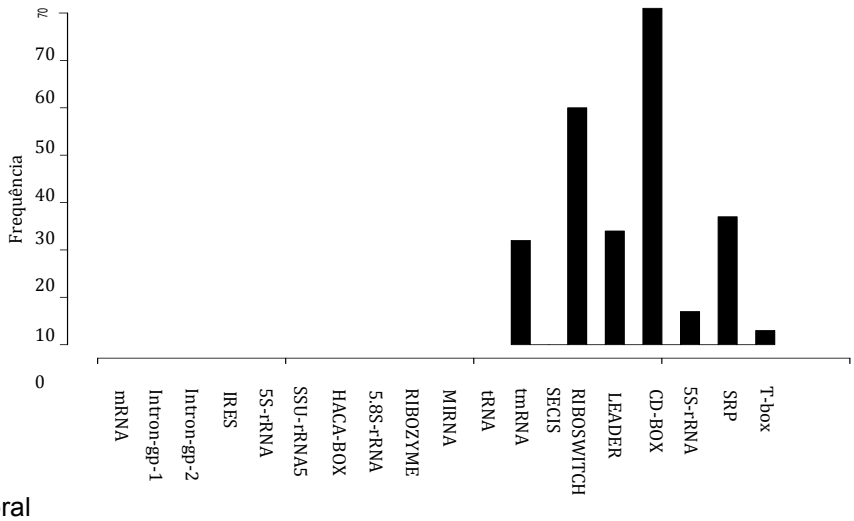
Figura 35. Clusterização hierárquica.



Os números em vermelho representam os agrupamentos considerados. Características que se destacaram dentro de cada grupo foram escritas em vermelho. Fonte: autoral

O primeiro grupo (Figura 35 - 1) apresenta RNAs preditos como não codificadores (PORTRAIT) e funções preditas como tmRNA, RIBOSWITCH, LEADER, CD-BOX, 5S-rRNA, SRP e T-box (RNAcon_v2 – Figura 36).

Figura 36. Frequência RNAcon para o primeiro grupo do agrupamento hierárquico.



Fonte: autoral

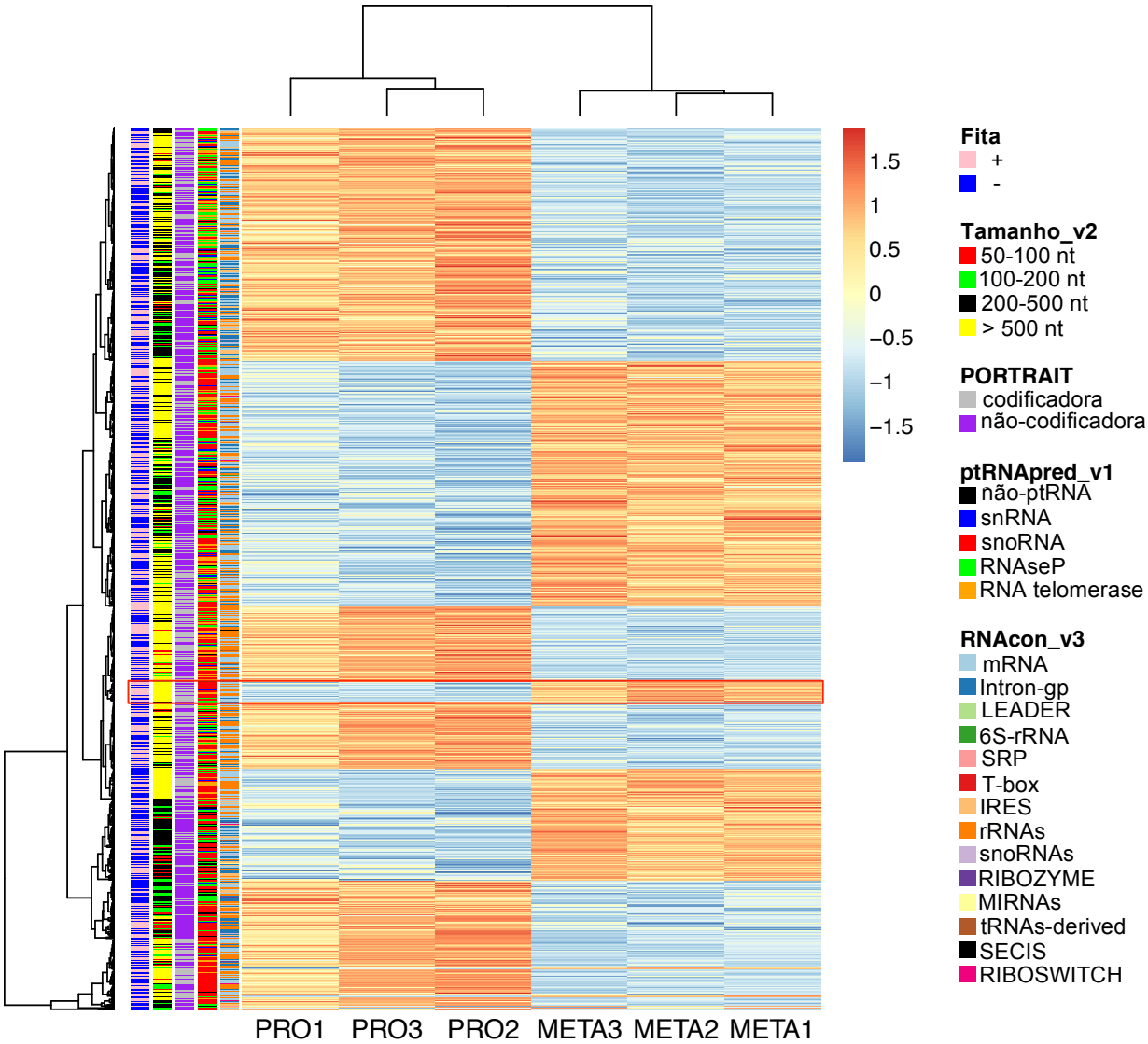
O segundo grupo (Figura 35 - 2) apresenta RNAs preditos como ncRNAs (PORTRAIT) e com tamanho maior que 200 nt (ncRNAs longos). O terceiro grupo (Figura 35 – 3) contém os RNAs preditos como ncRNAs (PORTRAIT) e com tamanho < 200 nt (ncRNAs curtos), sendo que este agrupamento também foi realizado pelo método PAM (Tabela 23 - Grupo 4). O quarto agrupamento (Figura 35 – 4) apresenta RNAs sendo preditos como RNA telomerase (ptRNApred) e maiores de 500 nt, sendo que este agrupamento também foi realizado pelo método PAM (Tabela 37 - Grupo 3). O quinto agrupamento (Figura 35 – 5) contém RNAs preditos como codificadores de proteína (PORTRAIT) e maiores de 200 nt.

A direção de transcrição (Fita) esteve presente em todos os agrupamentos de maneira equilibrada em suas duas versões (fita + e fita -). Essa característica pode ser modificada para considerar a direção do bloco policistrônico em que está localizado o ncRNA putativo, identificando se está na fita *sense* ou *antisense* ao PTU, podendo assim surgir novos padrões.

4.3.5.1 *Heatmaps* complexos – ncRNAs diferencialmente expressos

Heatmaps complexos integraram dados de expressão diferencial e dados das características dos ncRNAs de *L. braziliensis* nas três comparações de estágio de desenvolvimento: PROvsMETA (Figura 37), METAvsAMA (Figura 38) e AMAvsPRO (Figura 39).

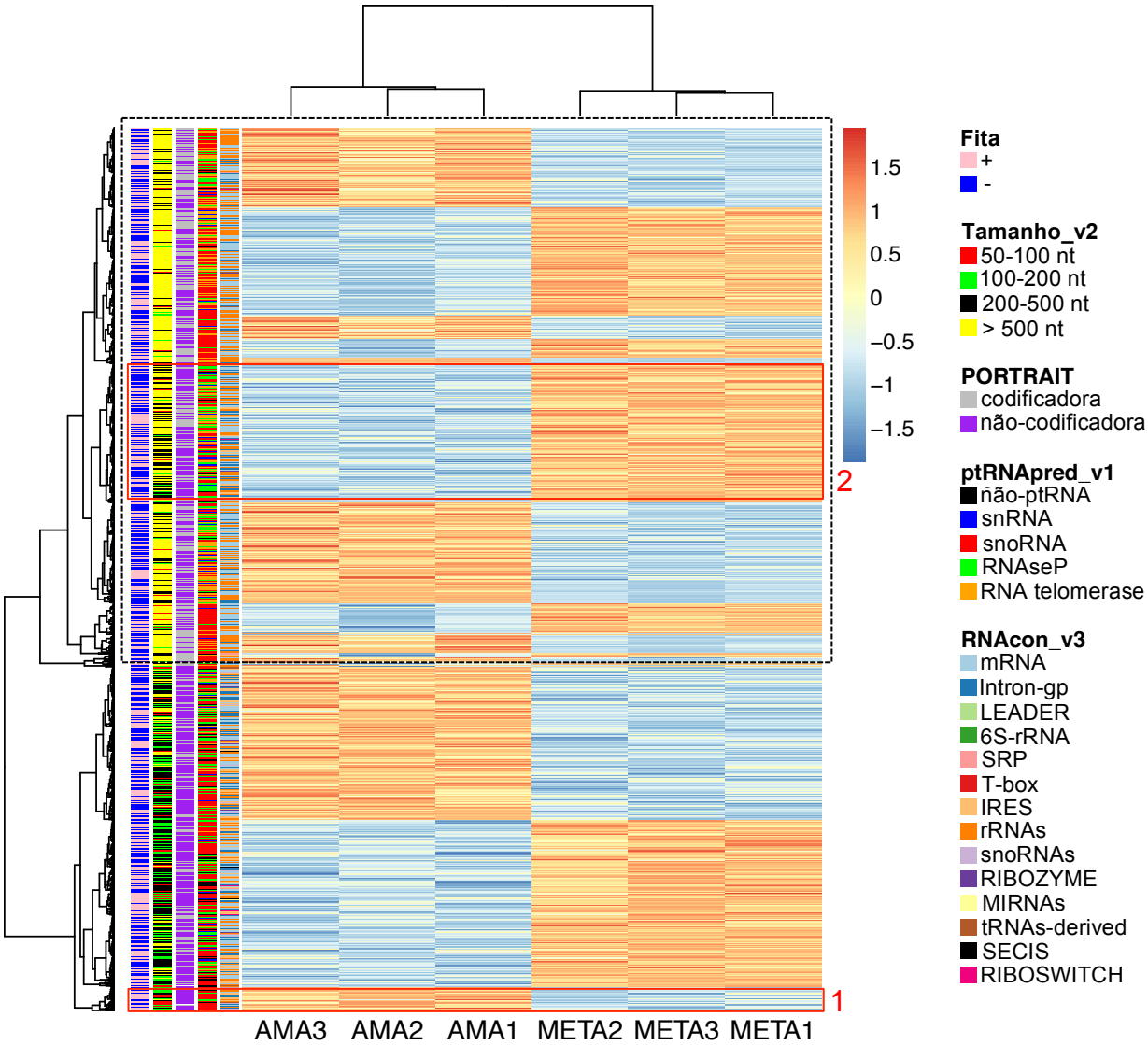
Figura 37. *Heatmap* complexo integrando dados de expressão diferencial (PROvsMETA) e características preditas dos ncRNAs de *L. braziliensis*.



O quadro vermelho destaca ncRNAs longos (> 500 nt) sub-representados em promastigotas procíclicos. Fonte: autoral

Considerando a diferença de expressão (PROvsMETA) e o agrupamento das características de ncRNAs putativos, encontramos um grupo de ncRNAs longos (> 500 nt) sub-representados em promastigotas procíclicos (quadro vermelho).

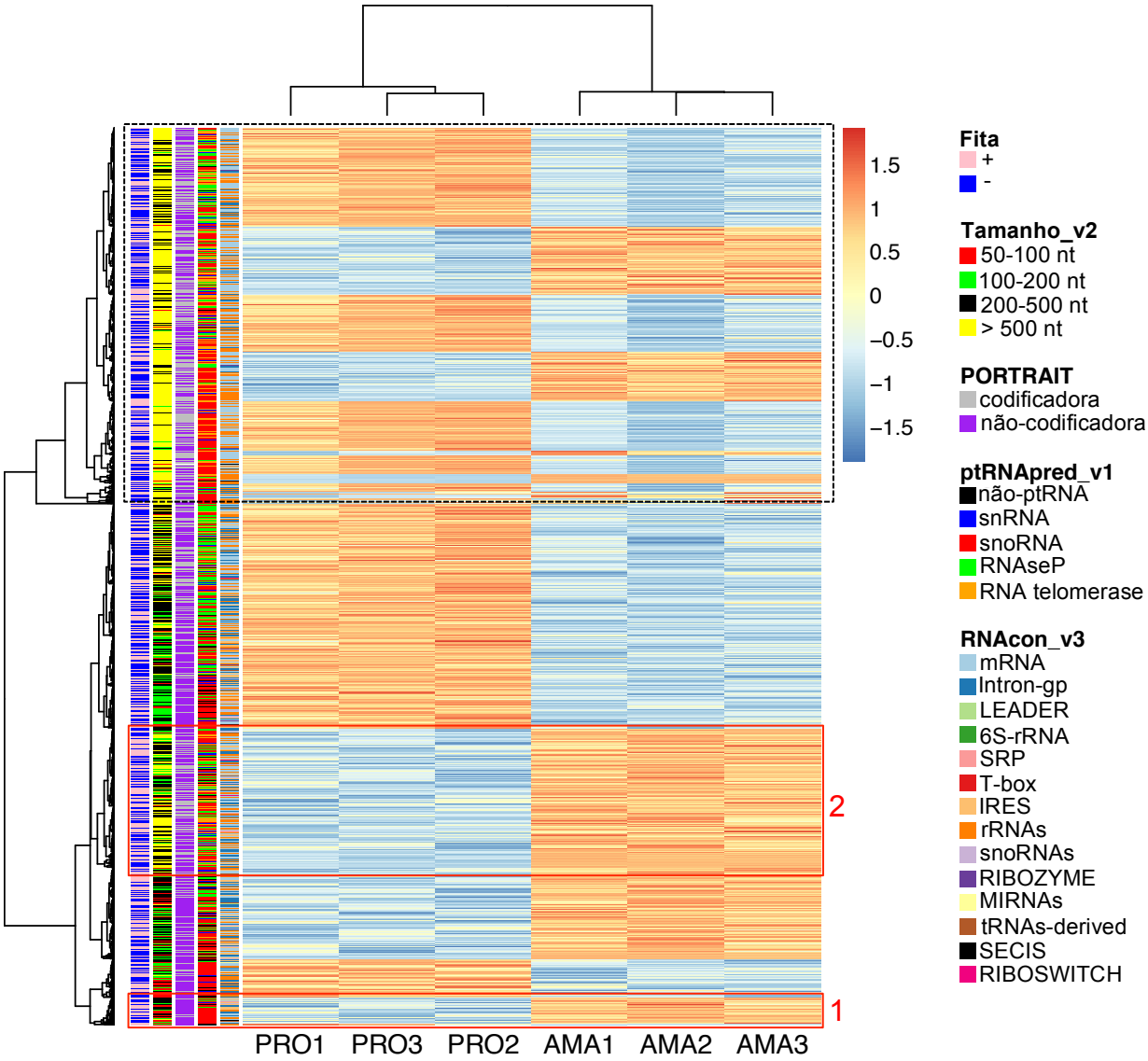
Figura 38. *Heatmap* complexo integrando dados de expressão diferencial (METAvsAMA) e características preditas dos ncRNAs de *L. braziliensis*.



O quadro pontilhado sugere ncRNAs agrupados por tamanho. O quadro vermelho 1 destaca ncRNAs entre 50 e 100 nt sobre-representados em amastigotas. O quadro vermelho 2 destaca ncRNAs entre 200 e 500 nt sub-representado em amastigotas. Fonte: autoral

Notamos a divisão por tamanho (> 200 nt) (quadro pontilhado) mas esse padrão não apresentou relação com a diferença de expressão entre os estágios. Um padrão que podemos observar é o enriquecimento de ncRNAs entre 50 e 100 nt sobre-representados em amastigotas com relação aos promastigotas metacíclicos (quadro vermelho – 1) e um outro grupo entre 200 e 500 nt sub-representado em amastigotas na comparação METAvsAMA (quadro vermelho – 2).

Figura 39. *Heatmap* complexo integrando dados de expressão diferencial (AMAvsPRO) e características preditas dos ncRNAs de *L. braziliensis*.



O quadro pontilhado sugere ncRNAs agrupados por tamanho. O quadro vermelho 1 destaca ncRNAs entre 50 e 100 nt sub-representados em promastigotas procíclicos. O quadro vermelho 2 destaca ncRNAs > 500 nt sub-representados em procíclicos. Fonte: autoral

Notamos a divisão por tamanho (> 200 nt) e com maior número de candidatos preditos como codificadores pelo PORTRAIT (quadro pontilhado) mas esse padrão não apresentou relação com a diferença de expressão entre os estágios. Um padrão que podemos observar é o enriquecimento de ncRNAs entre 50 e 100 nt sub-representados em promastigotas procíclicos com relação aos amastigotas (quadro vermelho – 1) e um outro grupo > 500 nt também sub-representados em procíclicos na comparação AMAvsPRO (quadro vermelho – 2).

De maneira geral, notamos que as predições funcionais dos ncRNAs não revelaram nenhum padrão indicador de correlações funcionais, isso pode ser real ou pode ser devido aos preditores utilizados. Sendo assim, novos preditores de função devem ser adicionados nas análises, juntamente com características de composição das sequências.

5. CONSIDERAÇÕES FINAIS

Avanços nas tecnologias de sequenciamento e nas análises computacionais foram um importante marco no estudo de vários aspectos biológicos de diferentes organismos, inclusive para os parasitos do gênero *Leishmania*. O primeiro genoma de uma espécie do parasito, *Leishmania major*, finalizado em 2005 [38] foi um marco para compreensão de sua organização genética e proporcionou uma mudança e ampliação marcantes das possibilidades de exploração da genética do organismo. Em 2013 [70], o sequenciamento de RNA em larga escala, ampliou nosso conhecimento em relação a expressão gênica desses parasitos.

Os tripanossomatídeos apresentam transcrição policistrônica com regulação da expressão gênica majoritariamente pós-transcricional [52; 53]. O controle da expressão gência em nível pós-transcricional pode ocorrer em vários níveis, inclusive envolvendo elementos de RNA especificamente envolvidos em processos regulatórios pós-transcricionais [21]. Isso indica a especial importância desses RNAs não codificadores regulatórios para os tripanossomatídeos [24].

Neste trabalho, utilizando dados de sequenciamento de RNA em larga escala e ferramentas computacionais, identificamos ncRNAs putativos em *Leishmania*. Na primeira versão do sequenciamento, empregamos um estratégia na construção das bibliotecas para evitar os mRNAs, os ncRNAs clássicos e os RNAs livres (fração possivelmente enriquecida para produtos de degradação) de *L. donovani*. Com esse subtranscriptoma e após a construção de bibliotecas e sequenciamento com alguns problemas sérios de representação, identificamos 37 ncRNAs putativos, sendo 4 deles confirmados por experimentos de *Northern blot*. Apesar desses resultados promissores, o poder informativo do sequenciamento foi baixo, impedindo qualquer análise estatística robusta. Uma avaliação cuidadosa foi necessária para identificar os problemas que ocorreram e assim, propusemos melhorias no processo de preparação de bibliotecas e na estratégia de sequenciamento. Realizamos testes em pequena escala e validamos as soluções propostas.

Enquanto o processo de avaliação de diferentes estratégias e validação das bibliotecas com subtranscriptomas era feito, decidimos sequenciar o RNA total para a análise de expressão diferencial dos genes e identificação dos ncRNAs mais abundantes. Decidimos utilizar *Leishmania braziliensis* nesta nova etapa, pois naquele momento os processos de diferenciação *in vitro*, para obtenção de metacíclicos e amastigotas, haviam sido otimizados (Natalia M. M. Teles) também

para essa espécie. A *L. braziliensis* tem uma vantagem significativa sobre as espécies do subgênero *Leishmania*, como a *L. donovani*, que é a presença e funcionalidade da maquinaria de RNAi. Tornando-se assim possível o uso de ferramentas mais versáteis de genética reversa para as análises funcionais futuras.

Dessa forma, neste trabalho foi analisado pela primera vez o transcriptoma de *L. braziliensis*, com excelente cobertura e de forma a comparar os transcriptomas dos principais estágios de vida, promastigotas procíclicos, metacíclicos e os amastigotas. Conduzimos a análise de expressão diferencial durante a diferenciação do parasito fazendo comparações sempre entre dois dados estágios. Foram encontrados 61%, 52% e 71% de genes diferencialmente expressos na comparação de promastigota procíclica versus metacíclica, promastigota metacíclica versus amastigota e amastigota versus promastigotas procíclicas, respectivamente. Essa diferença de expressão entre os estágios de desenvolvimento era esperada devido aos diferentes ambientes hostis do ciclo de vida do parasito e resultados similares foram encontrados na análise de transcriptoma de outras espécies de *Leishmania* [56; 141; 142]. Com os dados de sequenciamento de RNA-total também pudemos definir as 5'UTR de 81% das CDSs anotadas e a 3'UTR de 38% das CDSs anotadas, sendo que os resultados se assemelham aos obtidos para *L. major* [56] e *L. mexicana* [141].

Para a definir os ncRNAs putativos, o principal desafio foi identificar os picos de *reads* (*reads* mapeadas) candidatos a ncRNAs. Os dados utilizados foram de RNA total e a abundante presença de *reads* referentes a mRNAs dificultou a identificação dos ncRNAs putativos. Obtivemos o máximo de detalhamento possível do cenário de mapeamento calculando a cobertura de *reads* fita específica para cada nucleotídeo. De posse dessa informação, decidimos que os candidatos a ncRNAs longos (>200 nt) deveriam ter uma cobertura mínima por nucleotídeo de 50 *reads* e os candidatos a ncRNAs curtos (<200 nt) uma cobertura mínima, por nucleotídeo, de 100 *reads*. Exigimos uma cobertura maior para os ncRNAs curtos para aumentar a confiabilidade dos resultados, já que o tamanho da região era menor. Os transcritos identificados foram considerados candidatos a ncRNAs curtos ou longos quando estavam fora da região de CDS ou na fita contrária a ela. Os ncRNAs putativos identificados foram analisados com preditores que avaliam diferentes características de ncRNAs; aqueles candidatos sem nenhuma predição positiva, de acordo com os preditores empregados, foram excluídos do processo. Dessa forma, foram identificados 11.376 ncRNAs putativos em *L. braziliensis*, sendo

4.044 ncRNAs curtos e 7.332 ncRNAs longos. Destes, entre 27% e 41% são conservados em outras espécies de *Leishmania*. Além disso, e talvez o mais importante para os nossos propósitos, as ferramentas computacionais escolhidas, os filtros e a *pipeline* proposta nos permitiram alcançar com êxito nossos objetivos que eram, mais do que identificar potenciais ncRNAs novos em *Leishmania*, também identificar possíveis ncRNAs regulados diferencialmente ao longo do desenvolvimento (diferença de expressão entre os estágios). Estes transcritos são potencialmente os mais interessantes, ou aqueles que prioritariamente devam ser investigados funcionalmente, já que pode significar a busca/identificação por/de fatores responsáveis por promover o sucesso do parasito em uma nova etapa do desenvolvimento, particularmente aquelas que envolvem a interação do protozoário com o mamífero hospedeiro. O impedimento do processo que leva ao sucesso do estabelecimento da infecção e a proliferação do parasito é objetivo central da investigação.

Os resultados obtidos com as bibliotecas contruídas com o subtranscriptoma, tanto para os ncRNAs de *L. donovani* [82] quanto nos testes em menor escala para os ncRNAs de *L. braziliensis*, nos sugerem, que selecionar um subtranscriptoma que exclua grande parte dos RNAs associados a poliribossomos é de fato uma abordagem eficiente pois evidencia, ou permite detectar, ncRNAs internos a mRNAs e esses são perdidos pela abundância relativa dos RNAs associados e dos não associados aos poliribossomos. A ‘descoberta’ de potenciais RNAs pequenos e internos a mRNAs, que podem ser elementos regulatórios relevantes, fica evidenciada no perfil de picos ncRNAs pequenos descobertos no subtranscriptoma mas não revelados com a análise dos transcritos do RNAtotal. De forma que nossa hipótese é que os dois transcriptomas; RNA total e subtranscriptoma/fração não polissômica, sejam abordagens complementares que devem permitir ampliar a descoberta de novos ncRNAs de *Leishmania*. Nesse sentido, progredimos com a otimização, alcançamos um processo reprodutível e confiável (feito pelo pós-doutor Leandro Dias) e as bibliotecas com o subtranscriptoma dos três estágios de desenvolvido de *L. braziliensis* já foram sequenciadas no NextSeq-Illumina e serão integradas nas análises, no futuro próximo.

Assim, além do ineditismo na identificação e descrição de ncRNAs em *Leishmania* padrões foram descortinados com o processo de análise proposto e executado. Esse trabalho pode assim tornar-se uma contribuição significativa para

compreender melhor os processos de regulação de expressão gênica em *Leishmania* e oferecerá à comunidade um conjunto grande e importante de informações sobre a organização genética do parasito, diferenças genéticas e regulatórias ao longo do desenvolvimento, e um grande banco de dados do transcriptoma de forma global.

5.1 PRINCIPAIS CONTRIBUIÇÕES

As principais contribuições deste trabalho foram:

1. Desenvolvimento de *scripts* específicos e de um *pipeline* de análise de dados de sequenciamento para identificação computacional de RNAs não codificadores em *Leishmania*;
2. Identificação computacional de ncRNAs derivados de UTR em *L. donovani*;
3. Otimização de parâmetros para extração e fracionamento de RNA para a preparação das bibliotecas, além de definir a escala de sequenciamento para possibilitar a identificação de ncRNAs de *Leishmania*;
4. Análise de expressão gênica diferencial entre os três estágios de desenvolvimento de *L. braziliensis*;
5. Identificação computacional de ncRNAs de *L. braziliensis* utilizando dados de sequenciamento de RNA total que permitiu estimar a população e a diversidade desses ncRNAs putativos;
6. Análise de expressão diferencial dos ncRNAs identificados;
7. Identificação de padrões quanto ao perfil de ncRNAs putativos de *Leishmania*.

5.2 PERSPECTIVAS

As perspectivas detes trabalho incluem:

1. Associar os dados obtidos com a análise do subtranscriptoma (RNAs associados a proteínas em frações não polissômicas) ao conjunto de dados obtidos pelo sequenciamento de RNA total, para identificação de ncRNAs menos abundantes e/ou internos a regiões de UTR ou CDS; A análise com essas abordagens complementares aumentará o impacto e abrangência desse trabalho.
2. Analisar um novo sequenciamento com bibliotecas enriquecidas de RNAs capturados pela cauda PoliA+, melhorando assim a definição dos limites das 3’UTRs;

3. Adicionar informações sobre proteínas hipotéticas identificadas como diferencialmente expressas;
4. Selecionar dentre os ncRNAs identificados e diferencialmente expressos os mais promissores para análises funcionais;
5. Avaliar o cenário de ncRNAs em regiões de SSR (*Strand Switch Region*) de *L. braziliensis*;
6. Avaliar a direção dos ncRNAs identificados (*sense* ou *antisense*) em relação a direção de transcrição do bloco policistrônico em que está localizado;
7. Automatizar do *pipeline* de identificação de ncRNAs em *Leishmania*;
8. Melhorar a anotação gênica de *L. braziliensis*, identificando genes anotados de maneira equivocada ou não anotados.

PRODUÇÃO CIENTÍFICA DURANTE O DOUTORADO

1. Primeira autoria compartilhada com o aluno de doutorado Felipe F. Castro.

Castro FF, **Ruy PC**, Zeviani KN, Santos RF, Toledo JS, Cruz AK. Evidence of putative non-coding RNAs from *Leishmania* untranslated regions. Molecular and Biochemical Parasitology; 2017; doi: 10.1016/j.molbiopara.2017.04.002. (Anexo 01)

2. Colaboração

Velasco-Cruz AA, Alves-Ferreira EVC, Milbratz-Moré G, Chahud F, **Ruy PC**, Duarte MIS, and Cruz AK. Case Report: Sclerosing Orbital Inflammation Caused by *Leishmania braziliensis*. Am J Trop Med Hyg; 2016; doi:10.4269/ajtmh.16-0389.

3. Colaboração

Alves-Ferreira EV, Toledo JS, Oliveira AH, Ferreira TR, **Ruy PC**, Pinzan CF, Santos RF, Boaventura V, Rojo D, López-Gonzálvez Á, Rosa JC, Barbas C, Barral-Netto M, Barral A, Cruz AK. Differential Gene Expression and Infection Profiles of Cutaneous and Mucosal *Leishmania braziliensis* Isolates from the Same Patient. PLoS Negl Trop Dis; 2015; doi: 10.4269/ajtmh.16-0389.

4. Artigo mestrado

Ruy PC, Torrieri R, Toledo JS, Alves VS, Cruz AK, Ruiz JC. Intrinsically disordered proteins (IDPs) in trypanosomatids. BMC Genomics 2014; doi: 10.1186/1471-2164-15-1100.

5. Primeira autoria compartilhada com o pós-doutor Fabiano S. Pais

Pais FS, **Ruy PC**, Oliveira G, Coimbra R. Assessing the efficiency of multiple sequence alignment programs. Algorithms for Molecular Biology, 2014; doi: 10.1186/1748-7188-9-4.

REFERÊNCIAS*

1 MATTICK, J. S.; MAKUNIN, I. V. Non-coding RNA. **Hum Mol Genet,** v. 15 Spec No 1, p. R17-29, Apr 2006. ISSN 0964-6906. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/16651366 >.

2 WILUSZ, J. E.; SUNWOO, H.; SPECTOR, D. L. Long noncoding RNAs: functional surprises from the RNA world. **Genes Dev,** v. 23, n. 13, p. 1494-504, Jul 2009. ISSN 1549-5477. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/19571179 >.

3 MATTICK, J. S. The genetic signatures of noncoding RNAs. **PLoS Genet,** v. 5, n. 4, p. e1000459, Apr 2009. ISSN 1553-7404. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/19390609 >.

4 TAFT, R. J.; PHEASANT, M.; MATTICK, J. S. The relationship between non-protein-coding DNA and eukaryotic complexity. **Bioessays,** v. 29, n. 3, p. 288-99, Mar 2007. ISSN 0265-9247. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/17295292 >.

5 AMARAL, P. P.; MATTICK, J. S. Noncoding RNA in development. **Mamm Genome,** v. 19, n. 7-8, p. 454-92, Aug 2008. ISSN 0938-8990. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/18839252 >.

6 BARTEL, D. P. MicroRNAs: target recognition and regulatory functions. **Cell,** v. 136, n. 2, p. 215-33, Jan 2009. ISSN 1097-4172. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/19167326 >.

7 WATERS, L. S.; STORZ, G. Regulatory RNAs in bacteria. **Cell,** v. 136, n. 4, p. 615-28, Feb 2009. ISSN 1097-4172. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/19239884 >.

8 MATTICK, J. S. Non-coding RNAs: the architects of eukaryotic complexity. **EMBO Rep,** v. 2, n. 11, p. 986-91, Nov 2001. ISSN 1469-221X. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/11713189 >.

9 CECH, T. R.; STEITZ, J. A. The noncoding RNA revolution-trashing old rules to forge new ones. **Cell,** v. 157, n. 1, p. 77-94, Mar 2014. ISSN 1097-4172. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/24679528 >.

10 MORRIS, K. V.; MATTICK, J. S. The rise of regulatory RNA. **Nat Rev Genet,** v. 15, n. 6, p. 423-37, Jun 2014. ISSN 1471-0064. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/24776770 >.

11 FARAZI, T. A.; JURANEK, S. A.; TUSCHL, T. The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members.

* De acordo com a Associação Brasileira de Normas Técnicas (ABNT NBR 6023).

Development, v. 135, n. 7, p. 1201-14, Apr 2008. ISSN 0950-1991. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/18287206 >.

12 MERCER, T. R.; DINGER, M. E.; MATTICK, J. S. Long non-coding RNAs: insights into functions. **Nat Rev Genet,** v. 10, n. 3, p. 155-9, Mar 2009. ISSN 1471-0064. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/19188922 >.

13 HOLLEY, R. W. et al. STRUCTURE OF A RIBONUCLEIC ACID. **Science,** v. 147, n. 3664, p. 1462-5, Mar 1965. ISSN 0036-8075. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/14263761 >.

14 WATERHOUSE, P. M.; GRAHAM, M. W.; WANG, M. B. Virus resistance and gene silencing in plants can be induced by simultaneous expression of sense and antisense RNA. **Proc Natl Acad Sci U S A,** v. 95, n. 23, p. 13959-64, Nov 1998. ISSN 0027-8424. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/9811908 >.

15 FIRE, A. et al. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. **Nature,** v. 391, n. 6669, p. 806-11, Feb 1998. ISSN 0028-0836. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/9486653 >.

16 ROSENBLOOM, K. R. et al. ENCODE whole-genome data in the UCSC Genome Browser: update 2012. **Nucleic Acids Res,** v. 40, n. Database issue, p. D912-7, Jan 2012. ISSN 1362-4962. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/22075998 >.

17 DINGER, M. E. et al. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. **PLoS Comput Biol,** v. 4, n. 11, p. e1000176, Nov 2008. ISSN 1553-7358. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/19043537 >.

18 EDDY, S. R. Non-coding RNA genes and the modern RNA world. **Nat Rev Genet,** v. 2, n. 12, p. 919-29, Dec 2001. ISSN 1471-0056. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/11733745 >.

19 GOTTESMAN, S. Micros for microbes: non-coding regulatory RNAs in bacteria. **Trends Genet,** v. 21, n. 7, p. 399-404, Jul 2005. ISSN 0168-9525. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/15913835 >.

20 WASSARMAN, K. M.; ZHANG, A.; STORZ, G. Small RNAs in Escherichia coli. **Trends Microbiol,** v. 7, n. 1, p. 37-45, Jan 1999. ISSN 0966-842X. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/10068996 >.

21 RABANI, M.; KERTESZ, M.; SEGAL, E. Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. **Proc Natl Acad Sci U S A,** v. 105, n. 39, p. 14885-90, Sep 2008. ISSN 1091-6490. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/18815376 >.

22 CLAYTON, C.; SHAPIRA, M. Post-transcriptional regulation of gene expression in trypanosomes and leishmanias. **Mol Biochem Parasitol,** v. 156, n. 2, p. 93-101, Dec 2007. ISSN 0166-6851. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/17765983 >.

23 HAILE, S.; PAPADOPOULOU, B. Developmental regulation of gene expression in trypanosomatid parasitic protozoa. **Curr Opin Microbiol,** v. 10, n. 6, p. 569-77, Dec 2007. ISSN 1369-5274. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/18177626 >.

24 MAO, Y.; NAJAFABADI, H. S.; SALAVATI, R. Genome-wide computational identification of functional RNA elements in Trypanosoma brucei. **BMC Genomics,** v. 10, p. 355, Aug 2009. ISSN 1471-2164. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/19653906 >.

25 MURRAY, H. W. et al. Antagonizing deactivating cytokines to enhance host defense and chemotherapy in experimental visceral leishmaniasis. **Infect Immun,** v. 73, n. 7, p. 3903-11, Jul 2005. ISSN 0019-9567. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/15972476 >.

26 REITHINGER, R. et al. Cutaneous leishmaniasis. **Lancet Infect Dis,** v. 7, n. 9, p. 581-96, Sep 2007. ISSN 1473-3099. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/17714672 >.

27 WHO. World Health Organization - Leishmaniasis: magnitude of the problem. 2009. Disponível em: < http://www.who.int/leishmaniasis/burden/magnitude/burden_magnitude/en/index.html >. Acesso em: December.

28 OUELLETTE, M.; DRUMMELSMITH, J.; PAPADOPOULOU, B. Leishmaniasis: drugs in the clinic, resistance and new developments. **Drug Resist Updat,** v. 7, n. 4-5, p. 257-66, 2004 Aug-Oct 2004. ISSN 1368-7646. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/15533763 >.

29 HANDMAN, E. Leishmaniasis: current status of vaccine development. **Clin Microbiol Rev,** v. 14, n. 2, p. 229-43, Apr 2001. ISSN 0893-8512. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/11292637 >.

30 CRUZ, A. K. et al. Current treatment and drug discovery against Leishmania spp. and Plasmodium spp.: a review. **Curr Drug Targets,** v. 10, n. 3, p. 178-92, Mar 2009. ISSN 1873-5592. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/19275555 >.

31 BATES, P. A. Transmission of Leishmania metacyclic promastigotes by phlebotomine sand flies. **Int J Parasitol,** v. 37, n. 10, p. 1097-106, Aug 2007. ISSN 0020-7519. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/17517415 >.

32 TEIXEIRA, M. J. et al. Chemokines in host-parasite interactions in leishmaniasis. **Trends Parasitol,** v. 22, n. 1, p. 32-40, Jan 2006. ISSN 1471-

4922. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/16310413 >.

33 MCCONVILLE, M. J.; HANDMAN, E. WITHDRAWN: The molecular basis of Leishmania pathogenesis. **Int J Parasitol**, Jun 2007. ISSN 1879-0135. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/19004440 >.

34 NADERER, T.; MCCONVILLE, M. J. The Leishmania-macrophage interaction: a metabolic perspective. **Cell Microbiol,** v. 10, n. 2, p. 301-8, Feb 2008. ISSN 1462-5822. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/18070117 >.

35 CHAPPUIS, F. et al. Visceral leishmaniasis: what are the needs for diagnosis, treatment and control? **Nat Rev Microbiol,** v. 5, n. 11, p. 873-82, Nov 2007. ISSN 1740-1534. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/17938629 >.

36 MARTÍNEZ-CALVILLO, S. et al. Transcription of Leishmania major Friedlin chromosome 1 initiates in both directions within a single region. **Mol Cell,** v. 11, n. 5, p. 1291-9, May 2003. ISSN 1097-2765. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/12769852 >.

37 _____. Transcription initiation and termination on Leishmania major chromosome 3. **Eukaryot Cell,** v. 3, n. 2, p. 506-17, Apr 2004. ISSN 1535-9778. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/15075279 >.

38 IVENS, A. C. et al. The genome of the kinetoplastid parasite, Leishmania major. **Science,** v. 309, n. 5733, p. 436-42, Jul 2005. ISSN 1095-9203. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/16020728 >.

39 LLANES, A. et al. The genome of Leishmania panamensis: insights into genomics of the L. (Viannia) subgenus. **Sci Rep,** v. 5, p. 8550, 2015. ISSN 2045-2322. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/25707621 >.

40 PEACOCK, C. S. et al. Comparative genomic analysis of three Leishmania species that cause diverse human disease. **Nat Genet,** v. 39, n. 7, p. 839-47, Jul 2007. ISSN 1061-4036. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/17572675 >.

41 SMITH, D. F.; PEACOCK, C. S.; CRUZ, A. K. Comparative genomics: from genotype to disease phenotype in the leishmaniases. **Int J Parasitol,** v. 37, n. 11, p. 1173-86, Sep 2007. ISSN 0020-7519. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/17645880 >.

42 NGÔ, H. et al. Double-stranded RNA induces mRNA degradation in Trypanosoma brucei. **Proc Natl Acad Sci U S A,** v. 95, n. 25, p. 14687-92, Dec 1998. ISSN 0027-8424. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/9843950 >.

43 ROBINSON, K. A.; BEVERLEY, S. M. Improvements in transfection efficiency and tests of RNA interference (RNAi) approaches in the protozoan parasite Leishmania. **Mol Biochem Parasitol,** v. 128, n. 2, p. 217-28, May 2003. ISSN 0166-6851. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/12742588 >.

44 DAROCHA, W. D. et al. Tests of cytoplasmic RNA interference (RNAi) and construction of a tetracycline-inducible T7 promoter system in Trypanosoma cruzi. **Mol Biochem Parasitol,** v. 133, n. 2, p. 175-86, Feb 2004. ISSN 0166-6851. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/14698430 >.

45 LYE, L. F. et al. Retention and loss of RNA interference pathways in trypanosomatid protozoans. **PLoS Pathog,** v. 6, n. 10, p. e1001161, 2010. ISSN 1553-7374. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/21060810 >.

46 SHLOMAI, J.; ZADOK, A. Kinetoplast DNA minicircles of trypanosomatids encode for a protein product. **Nucleic Acids Res,** v. 12, n. 21, p. 8017-28, Nov 1984. ISSN 0305-1048. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/6095184 >.

47 WALDER, J. A. et al. The 35-nucleotide spliced leader sequence is common to all trypanosome messenger RNA's. **Science,** v. 233, n. 4763, p. 569-71, Aug 1986. ISSN 0036-8075. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/3523758 >.

48 AGABIAN, N. Trans splicing of nuclear pre-mRNAs. **Cell,** v. 61, n. 7, p. 1157-60, Jun 1990. ISSN 0092-8674. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/2142018 >.

49 STURM, N. R.; SIMPSON, L. Kinetoplast DNA minicircles encode guide RNAs for editing of cytochrome oxidase subunit III mRNA. **Cell,** v. 61, n. 5, p. 879-84, Jun 1990. ISSN 0092-8674. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/1693097 >.

50 FLINN, H. M.; SMITH, D. F. Genomic organisation and expression of a differentially-regulated gene family from Leishmania major. **Nucleic Acids Res,** v. 20, n. 4, p. 755-62, Feb 1992. ISSN 0305-1048. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/1371863 >.

51 LEBOWITZ, J. H. et al. Coupling of poly(A) site selection and trans-splicing in Leishmania. **Genes Dev,** v. 7, n. 6, p. 996-1007, Jun 1993. ISSN 0890-9369. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/8504937 >.

52 BOOTHROYD, J. C.; CROSS, G. A. Transcripts coding for variant surface glycoproteins of Trypanosoma brucei have a short, identical exon at their 5' end. **Gene,** v. 20, n. 2, p. 281-9, Dec 1982. ISSN 0378-1119. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/7166234 >.

53 BEVERLEY, S. M. Protozomics: trypanosomatid parasite genetics comes of age. **Nat Rev Genet,** v. 4, n. 1, p. 11-9, Jan 2003. ISSN 1471-0056. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/12509749 >.

54 DE GAUDENZI, J. G. et al. Gene expression regulation in trypanosomatids. **Essays Biochem,** v. 51, p. 31-46, 2011. ISSN 1744-1358. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/22023440 >.

55 ULLU, E.; MATTHEWS, K. R.; TSCHUDI, C. Temporal order of RNA-processing reactions in trypanosomes: rapid trans splicing precedes polyadenylation of newly synthesized tubulin transcripts. **Mol Cell Biol,** v. 13, n. 1, p. 720-5, Jan 1993. ISSN 0270-7306. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/8417363 >.

56 DILLON, L. A. et al. Transcriptomic profiling of gene expression and RNA processing during Leishmania major differentiation. **Nucleic Acids Res,** v. 43, n. 14, p. 6799-813, Aug 2015. ISSN 1362-4962. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/26150419 >.

57 KOLEV, N. G. et al. The transcriptome of the human pathogen Trypanosoma brucei at single-nucleotide resolution. **PLoS Pathog,** v. 6, n. 9, p. e1001090, 2010. ISSN 1553-7374. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/20838601 >.

58 SIEGEL, T. N. et al. Genome-wide analysis of mRNA abundance in two life-cycle stages of Trypanosoma brucei and identification of splicing and polyadenylation sites. **Nucleic Acids Res,** v. 38, n. 15, p. 4946-57, Aug 2010. ISSN 1362-4962. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/20385579 >.

59 EL-SAYED, N. M. et al. Comparative genomics of trypanosomatid parasitic protozoa. **Science,** v. 309, n. 5733, p. 404-9, Jul 2005. ISSN 1095-9203. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/16020724 >.

60 MCKEE, A. E.; SILVER, P. A. Systems perspectives on mRNA processing. **Cell Res,** v. 17, n. 7, p. 581-90, Jul 2007. ISSN 1748-7838. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/17621309 >.

61 MANIATIS, T.; REED, R. An extensive network of coupling among gene expression machines. **Nature,** v. 416, n. 6880, p. 499-506, Apr 2002. ISSN 0028-0836. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/11932736 >.

62 HERSHEY, J. W.; SONENBERG, N.; MATHEWS, M. B. Principles of translational control: an overview. **Cold Spring Harb Perspect Biol,** v. 4, n. 12, Dec 2012. ISSN 1943-0264. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/23209153 >.

63 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. ISSN 1471-2164. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/18510761 >.

64 SMITH, M.; BRINGAUD, F.; PAPADOPOULOU, B. Organization and evolution of two SIDER retroposon subfamilies and their impact on the Leishmania genome. **BMC Genomics,** v. 10, p. 240, 2009. ISSN 1471-2164. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/19463167 >.

65 BRINGAUD, F. et al. Members of a large retroposon family are determinants of post-transcriptional gene expression in Leishmania. **PLoS Pathog,** v. 3, n. 9, p. 1291-307, Sep 2007. ISSN 1553-7374. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/17907803 >.

66 VASCONCELOS, E. J. et al. In silico identification of conserved intercoding sequences in Leishmania genomes: unraveling putative cis-regulatory elements. **Mol Biochem Parasitol,** v. 183, n. 2, p. 140-50, Jun 2012. ISSN 1872-9428. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/22387760 >.

67 KRATZ, A.; CARNINCI, P. The devil in the details of RNA-seq. **Nat Biotechnol,** v. 32, n. 9, p. 882-4, Sep 2014. ISSN 1546-1696. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/25203036 >.

68 WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics. **Nat Rev Genet,** v. 10, n. 1, p. 57-63, Jan 2009. ISSN 1471-0064. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/19015660 >.

69 MORTAZAVI, A. et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. **Nat Methods,** v. 5, n. 7, p. 621-8, Jul 2008. ISSN 1548-7105. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/18516045 >.

70 RASTROJO, A. et al. The transcriptome of Leishmania major in the axenic promastigote stage: transcript annotation and relative expression levels by RNA-seq. **BMC Genomics,** v. 14, p. 223, 2013. ISSN 1471-2164. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/23557257 >.

71 KAPLER, G. M.; ZHANG, K.; BEVERLEY, S. M. Nuclease mapping and DNA sequence analysis of transcripts from the dihydrofolate reductase-thymidylate synthase (R) region of Leishmania major. **Nucleic Acids Res,** v. 18, n. 21, p. 6399-408, Nov 1990. ISSN 0305-1048. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/2243782 >.

72 NGÔ, H. et al. Double-stranded RNA induces mRNA degradation in Trypanosoma brucei. **Proc Natl Acad Sci U S A,** v. 95, n. 25, p. 14687-92, Dec 1998. ISSN 0027-8424. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/9843950 >.

73 ULLU, E.; TSCHUDI, C.; CHAKRABORTY, T. RNA interference in protozoan parasites. **Cell Microbiol,** v. 6, n. 6, p. 509-19, Jun 2004. ISSN 1462-5814. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/15104593 >.

74 MERCER, T. R.; DINGER, M. E.; MATTICK, J. S. Long non-coding RNAs: insights into functions. **Nat Rev Genet,** v. 10, n. 3, p. 155-9, Mar 2009. ISSN 1471-0064. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/19188922 >.

75 MARRAFFINI, L. A.; SONTHEIMER, E. J. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. **Science,** v. 322, n. 5909, p. 1843-5, Dec 2008. ISSN 1095-9203. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/19095942 >.

76 DUMAS, C. et al. A novel class of developmentally regulated noncoding RNAs in Leishmania. **Eukaryot Cell,** v. 5, n. 12, p. 2033-46, Dec 2006. ISSN 1535-9778. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/17071827 >.

77 IRIBAR, M. P.; TOSI, L. R.; CRUZ, A. K. A processed short transcript of Leishmania, ODD1. **Mol Biochem Parasitol,** v. 127, n. 2, p. 205-8, Apr 2003. ISSN 0166-6851. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/12672530 >.

78 REDERSTORFF, M. et al. RNPomics: defining the ncRNA transcriptome by cDNA library generation from ribonucleo-protein particles. **Nucleic Acids Res,** v. 38, n. 10, p. e113, Jun 2010. ISSN 1362-4962. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/20150415 >.

79 MICHAELI, S. et al. RNA-seq analysis of small RNPs in Trypanosoma brucei reveals a rich repertoire of non-coding RNAs. **Nucleic Acids Res,** v. 40, n. 3, p. 1282-98, Feb 2012. ISSN 1362-4962. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/21976736 >.

80 LAMBERTZ, U. et al. Small RNAs derived from tRNAs and rRNAs are highly enriched in exosomes from both old and new world Leishmania providing evidence for conserved exosomal RNA Packaging. **BMC Genomics,** v. 16, p. 151, 2015. ISSN 1471-2164. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/25764986 >.

81 GOYARD, S. et al. An in vitro system for developmental and genetic studies of Leishmania donovani phosphoglycans. **Mol Biochem Parasitol,** v. 130, n. 1, p. 31-42, Aug 2003. ISSN 0166-6851. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/14550894 >.

82 FREITAS CASTRO, F. et al. Evidence of putative non-coding RNAs from Leishmania untranslated regions. **Mol Biochem Parasitol,** v. 214, p. 69-74, Apr 2017. ISSN 1872-9428. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/28385563 >.

83 ASLETT, M. et al. TriTrypDB: a functional genomic resource for the Trypanosomatidae. **Nucleic Acids Res,** v. 38, n. Database issue, p. D457-62, Jan 2010. ISSN 1362-4962. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/19843604 >.

84 LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. **Nat Methods,** v. 9, n. 4, p. 357-9, Apr 2012. ISSN 1548-7105. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/22388286 >.

85 LI, H. et al. The Sequence Alignment/Map format and SAMtools. **Bioinformatics,** v. 25, n. 16, p. 2078-9, Aug 2009. ISSN 1367-4811. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/19505943 >.

86 RUTHERFORD, K. et al. Artemis: sequence visualization and annotation. **Bioinformatics,** v. 16, n. 10, p. 944-5, Oct 2000. ISSN 1367-4803. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/11120685 >.

87 R CORE TEAM. **R: A Language and Environment for Statistical Computing** Vienna, Austria: R Foundation for Statistical Computing 2014.

88 ANDERS, S.; HUBER, W. Differential expression analysis for sequence count data. **Genome Biol,** v. 11, n. 10, p. R106, 2010. ISSN 1465-6914. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/20979621 >.

89 ANDERS, S.; PYL, P. T.; HUBER, W. HTSeq--a Python framework to work with high-throughput sequencing data. **Bioinformatics,** v. 31, n. 2, p. 166-9, Jan 2015. ISSN 1367-4811. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/25260700 >.

90 ALTSCHUL, S. et al. Basic local alignment search tool. **J Mol Biol,** v. 215, n. 3, p. 403-10, Oct 1990. ISSN 0022-2836. Disponível em: < http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve& >.

91 INGOLIA, N. T. et al. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. **Science,** v. 324, n. 5924, p. 218-23, Apr 2009. ISSN 1095-9203. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/19213877 >.

92 _____. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. **Nat Protoc,** v. 7, n. 8, p. 1534-50, Aug 2012. ISSN 1750-2799. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/22836135 >.

93 JENSEN, B. C. et al. Extensive stage-regulation of translation revealed by ribosome profiling of Trypanosoma brucei. **BMC Genomics,** v. 15, p. 911, 2014. ISSN 1471-2164. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/25331479 >.

94 IRIBAR, M. P.; CRUZ, A. K. Base compositional bias in trans-spliced sequences of unknown function in Leishmania major. **Exp Parasitol,** v. 100,

n. 1, p. 1-5, Jan 2002. ISSN 0014-4894. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/11971647 >.

95 ALMEIDA, R. et al. From genomes to vaccines: Leishmania as a model. **Philos Trans R Soc Lond B Biol Sci,** v. 357, n. 1417, p. 5-11, Jan 2002. ISSN 0962-8436. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/11839176 >.

96 DONELSON, J. E.; GARDNER, M. J.; EL-SAYED, N. M. More surprises from Kinetoplastida. **Proc Natl Acad Sci U S A,** v. 96, n. 6, p. 2579-81, Mar 1999. ISSN 0027-8424. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/10077550 >.

97 PANWAR, B.; ARORA, A.; RAGHAVA, G. P. Prediction and classification of ncRNAs using structural information. **BMC Genomics,** v. 15, p. 127, 2014. ISSN 1471-2164. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/24521294 >.

98 SATO, K. et al. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. **Bioinformatics,** v. 27, n. 13, p. i85-93, Jul 2011. ISSN 1367-4811. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/21685106 >.

99 CROS, M. J. et al. RNAspace.org: An integrated environment for the prediction, annotation, and analysis of ncRNA. **RNA,** v. 17, n. 11, p. 1947-56, Nov 2011. ISSN 1469-9001. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/21947200 >.

100 ARRIAL, R. T.; TOGAWA, R. C.; BRIGIDO, M. E. M. Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: case study of the pathogenic fungus Paracoccidioides brasiliensis. **BMC Bioinformatics,** v. 10, p. 239, 2009. ISSN 1471-2105. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/19653905 >.

101 LOWE, T. M.; EDDY, S. R. A computational screen for methylation guide snoRNAs in yeast. **Science,** v. 283, n. 5405, p. 1168-71, Feb 1999. ISSN 0036-8075. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/10024243 >.

102 KIN, T. et al. fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. **Nucleic Acids Res,** v. 35, n. Database issue, p. D145-8, Jan 2007. ISSN 1362-4962. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/17099231 >.

103 XIE, C. et al. NONCODEv4: exploring the world of long non-coding RNA genes. **Nucleic Acids Res,** v. 42, n. Database issue, p. D98-103, Jan 2014. ISSN 1362-4962. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/24285305 >.

104 AMARAL, P. P. et al. lncRNAdb: a reference database for long noncoding RNAs. **Nucleic Acids Res,** v. 39, n. Database issue, p. D146-51, Jan 2011. ISSN 1362-4962. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/21112873 >.

105 BÜHLER, M. et al. TRAMP-mediated RNA surveillance prevents spurious entry of RNAs into the Schizosaccharomyces pombe siRNA pathway. **Nat Struct Mol Biol,** v. 15, n. 10, p. 1015-23, Oct 2008. ISSN 1545-9985. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/18776903 >.

106 WESTERMANN, A. J.; GORSKI, S. A.; VOGEL, J. Dual RNA-seq of pathogen and host. **Nat Rev Microbiol,** v. 10, n. 9, p. 618-30, Sep 2012. ISSN 1740-1534. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/22890146 >.

107 THORVALDSDÓTTIR, H.; ROBINSON, J. T.; MESIROV, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. **Brief Bioinform,** v. 14, n. 2, p. 178-92, Mar 2013. ISSN 1477-4054. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/22517427 >.

108 LIAO, Y.; SMYTH, G. K.; SHI, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. **Bioinformatics,** v. 30, n. 7, p. 923-30, Apr 2014. ISSN 1367-4811. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/24227677 >.

109 GENTLEMAN, R. C. et al. Bioconductor: open software development for computational biology and bioinformatics. **Genome Biol,** v. 5, n. 10, p. R80, 2004. ISSN 1465-6914. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/15461798 >.

110 YOUNG, M. D. et al. Gene ontology analysis for RNA-seq: accounting for selection bias. **Genome Biol,** v. 11, n. 2, p. R14, 2010. ISSN 1474-760X. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/20132535 >.

111 QUINLAN, A. R.; HALL, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. **Bioinformatics,** v. 26, n. 6, p. 841-2, Mar 2010. ISSN 1367-4811. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/20110278 >.

112 DE ALMEIDA-BIZZO, J. H. et al. Characterization of the pattern of ribosomal protein L19 production during the lifecycle of Leishmania spp. **Exp Parasitol,** v. 147, p. 60-6, Dec 2014. ISSN 1090-2449. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/25290356 >.

113 HOLZER, T. R.; MCMASTER, W. R.; FORNEY, J. D. Expression profiling by whole-genome interspecies microarray hybridization reveals differential gene expression in procyclic promastigotes, lesion-derived amastigotes, and axenic amastigotes in Leishmania mexicana. **Mol Biochem Parasitol,** v. 146, n. 2, p. 198-218, Apr 2006. ISSN 0166-6851. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/16430978 >.

114 INBAR, E. et al. The Transcriptome of Leishmania major Developmental Stages in Their Natural Sand Fly Vector. **MBio,** v. 8, n. 2, Apr 2017. ISSN 2150-7511. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/28377524 >.

115 SOTO, M. et al. Cell-cycle-dependent translation of histone mRNAs is the key control point for regulation of histone biosynthesis in Leishmania infantum. **Biochem J,** v. 379, n. Pt 3, p. 617-25, May 2004. ISSN 1470-8728. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/14766017 >.

116 GENSKE, J. E. et al. Structure and regulation of histone H2B mRNAs from Leishmania enriettii. **Mol Cell Biol,** v. 11, n. 1, p. 240-9, Jan 1991. ISSN 0270-7306. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/1986223 >.

117 GALANTI, N. et al. Histone genes in trypanosomatids. **Parasitol Today,** v. 14, n. 2, p. 64-70, Feb 1998. ISSN 0169-4758. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/17040700 >.

118 SOTO, M. et al. Histone synthesis in Leishmania infantum is tightly linked to DNA replication by a translational control. **Biochem J,** v. 346 Pt 1, p. 99-105, Feb 2000. ISSN 0264-6021. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/10657245 >.

119 SABAJ, V. et al. Histone synthesis in Trypanosoma cruzi. **Exp Cell Res,** v. 236, n. 2, p. 446-52, Nov 1997. ISSN 0014-4827. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/9367629 >.

120 GERBAULET, S. P. et al. Downregulation of histone H4 gene transcription during postnatal development in transgenic mice and at the onset of differentiation in transgenically derived calvarial osteoblast cultures. **J Cell Biochem,** v. 49, n. 2, p. 137-47, Jun 1992. ISSN 0730-2312. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/1400621 >.

121 STEIN, J. L. et al. Control of cell cycle regulated histone genes during proliferation and differentiation. **Int J Obes Relat Metab Disord,** v. 20 Suppl 3, p. S84-90, Mar 1996. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/8680483 >.

122 SARKAR, S.; LAKHOTIA, S. C. The Hsp60C gene in the 25F cytogenetic region in Drosophila melanogaster is essential for tracheal development and fertility. **J Genet,** v. 84, n. 3, p. 265-81, Dec 2005. ISSN 0022-1333. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/16385159 >.

123 SAIBIL, H. Chaperone machines for protein folding, unfolding and disaggregation. **Nat Rev Mol Cell Biol,** v. 14, n. 10, p. 630-42, Oct 2013. ISSN 1471-0080. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/24026055 >.

124 ROCHETTE, A. et al. Characterization and developmental gene regulation of a large gene family encoding amastin surface proteins in Leishmania spp. **Mol Biochem Parasitol,** v. 140, n. 2, p. 205-20, Apr 2005. ISSN 0166-6851. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/15760660 >.

125 TEIXEIRA, S. M. et al. A differentially expressed gene family encoding "amastin," a surface protein of Trypanosoma cruzi amastigotes. **J Biol Chem,** v. 269, n. 32, p. 20509-16, Aug 1994. ISSN 0021-9258. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/8051148 >.

126 ZHAO, S. et al. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. **PLoS One,** v. 9, n. 1, p. e78644, 2014. ISSN 1932-6203. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/24454679 >.

127 DE PAIVA, R. M. et al. Amastin Knockdown in Leishmania braziliensis Affects Parasite-Macrophage Interaction and Results in Impaired Viability of Intracellular Amastigotes. **PLoS Pathog,** v. 11, n. 12, p. e1005296, Dec 2015. ISSN 1553-7374. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/26641088 >.

128 ROCHETTE, A. et al. Whole-genome comparative RNA expression profiling of axenic and intracellular amastigote forms of Leishmania infantum. **Mol Biochem Parasitol,** v. 165, n. 1, p. 32-47, May 2009. ISSN 0166-6851. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/19393160 >.

129 RAFATI, S. et al. Amastin peptide-binding antibodies as biomarkers of active human visceral leishmaniasis. **Clin Vaccine Immunol,** v. 13, n. 10, p. 1104-10, Oct 2006. ISSN 1556-6811. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/17028214 >.

130 ROGERS, M. E. et al. Transmission of cutaneous leishmaniasis by sand flies is enhanced by regurgitation of fPPG. **Nature,** v. 430, n. 6998, p. 463-7, Jul 2004. ISSN 1476-4687. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/15269771 >.

131 BATES, P. A. Leishmania sand fly interaction: progress and challenges. **Curr Opin Microbiol,** v. 11, n. 4, p. 340-4, Aug 2008. ISSN 1369-5274. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/18625337 >.

132 PIANI, A. et al. Leishmania major proteophosphoglycan is expressed by amastigotes and has an immunomodulatory effect on macrophage function. **Microbes Infect,** v. 1, n. 8, p. 589-99, Jul 1999. ISSN 1286-4579. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/10611735 >.

133 DEPLEDGE, D. P. et al. Comparative expression profiling of Leishmania: modulation in gene expression between species and in different host genetic backgrounds. **PLoS Negl Trop Dis,** v. 3, n. 7, p. e476, Jul 2009. ISSN 1935-2735. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/19582145 >.

134 REGGIORI, F.; KLIONSKY, D. J. Autophagosomes: biogenesis from scratch? **Curr Opin Cell Biol,** v. 17, n. 4, p. 415-22, Aug 2005. ISSN 0955-0674. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/15978794 >.

135 BESTEIRO, S. et al. Endosome sorting and autophagy are essential for differentiation and virulence of Leishmania major. **J Biol Chem,** v. 281, n. 16, p. 11384-96, Apr 2006. ISSN 0021-9258. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/16497676 >.

136 FERNANDES, M. C. et al. Dual Transcriptome Profiling of Leishmania-Infected Human Macrophages Reveals Distinct Reprogramming Signatures. **MBio,** v. 7, n. 3, May 2016. ISSN 2150-7511. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/27165796 >.

137 KISHORE, S.; LUBER, S.; ZAVOLAN, M. Deciphering the role of RNA-binding proteins in the post-transcriptional control of gene expression. **Brief Funct Genomics,** v. 9, n. 5-6, p. 391-404, Dec 2010. ISSN 2041-2657. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/21127008 >.

138 SHI, H.; TSCHUDI, C.; ULLU, E. Functional replacement of Trypanosoma brucei Argonaute by the human slicer Argonaute2. **RNA,** v. 12, n. 6, p. 943-7, Jun 2006. ISSN 1355-8382. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/16611939 >.

139 ROMÃO, P. R. et al. Glutathione and the redox control system trypanothione/trypanothione reductase are involved in the protection of Leishmania spp. against nitrosothiol-induced cytotoxicity. **Braz J Med Biol Res,** v. 39, n. 3, p. 355-63, Mar 2006. ISSN 0100-879X. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/16501815 >.

140 ATAYDE, V. D. et al. The structure and repertoire of small interfering RNAs in Leishmania (Viannia) braziliensis reveal diversification in the trypanosomatid RNAi pathway. **Mol Microbiol,** v. 87, n. 3, p. 580-93, Feb 2013. ISSN 1365-2958. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/23217017 >.

141 FIEBIG, M.; KELLY, S.; GLUENZ, E. Comparative Life Cycle Transcriptomics Revises Leishmania mexicana Genome Annotation and Links a Chromosome Duplication with Parasitism of Vertebrates. **PLoS Pathog,** v. 11, n. 10, p. e1005186, Oct 2015. ISSN 1553-7374. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/26452044 >.

142 SAXENA, A. et al. Analysis of the Leishmania donovani transcriptome reveals an ordered progression of transient and permanent changes in gene expression during differentiation. **Mol Biochem Parasitol,** v. 152, n. 1, p. 53-65, Mar 2007. ISSN 0166-6851. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/17204342 >.

[143] FINN, R. D. et al. The Pfam protein families database: towards a more sustainable future. **Nucleic Acids Res,** v. 44, n. D1, p. D279-85, Jan 2016. ISSN 1362-4962. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/26673716 >.

[144] LOWE, T. M.; EDDY, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. **Nucleic Acids Res,** v. 25, n. 5, p. 955-64, Mar 1997. ISSN 0305-1048. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/9023104 >.

[145] GUPTA, Y. et al. ptRNApred: computational identification and classification of post-transcriptional RNA. **Nucleic Acids Res,** v. 42, n. 22, p. e167, Dec 2014. ISSN 1362-4962. Disponível em: < https://http://www.ncbi.nlm.nih.gov/pubmed/25303994 >.

[146] LAWSON, R. G.; JURS, P. C. New index for clustering tendency and its application to chemical problems. **J. Chem. Inf. Comput. Sci.,** v. 30, p. pp 36–4, 1990.

[147] MACQUEEN, J. **Some methods for classification and analysis of multivariate observations**. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability: Oakland, CA, USA., 1967. 281-297 p.

[148] KAUFMAN, L.; ROUSSEEUW, P. J. Partitioning around medoids (program pam). **Finding groups in data: an introduction to cluster analysis**, p. 68-125, 1990. ISSN 0470316802.

[149] WARD JR, J. H. Hierarchical grouping to optimize an objective function. **Journal of the American statistical association,** v. 58, n. 301, p. 236-244, 1963. ISSN 0162-1459.

[150] TIBSHIRANI, R.; WALTHER, G.; HASTIE, T. Estimating the number of clusters in a data set via the gap statistic. **Journal of the Royal Statistical Society: Series B (Statistical Methodology),** v. 63, p. 411-423, 2001.

[151] ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of computational and applied mathematics,** v. 20, p. 53-65, 1987. ISSN 0377-0427.

[152] ABDI, H.; WILLIAMS, L. J. Principal component analysis. **Wiley interdisciplinary reviews: computational statistics,** v. 2, n. 4, p. 433-459, 2010. ISSN 1939-0068.

ANEXOS

Molecular & Biochemical Parasitology xxx (2017) xxx-xxx

Contents lists available at ScienceDirect

Molecular & Biochemical Parasitology

journal homepage: www.elsevier.com

Short communication

Evidence of putative non-coding RNAs from *Leishmania* untranslated regions

Felipe Freitas de Castro [1], Patricia de C ssia Ruy[1], Karina Nogueira Zeviani, Ramon de Freitas Santos, Juliano Sim es Toledo, Angela Kaysel Cruz

Department of Cell and Molecular Biology, Ribeirão Preto Medical School, University of São Paulo, Brazil

ARTICLE INFO

Keywords:
Leishmania
Non-coding RNA
ncRNA
UTR-Transcript
UTR-derived RNAs
UTR-associated RNA
uaRNA

ABSTRACT

Non-coding RNAs (ncRNAs) are regulatory elements present in a wide range of organisms, including trypanosomatids. ncRNAs transcribed from the untranslated regions (UTRs) of coding genes have been described in the transcriptomes of several eukaryotes, including *Trypanosoma brucei*. To uncover novel putative ncRNAs in two *Leishmania* species, we examined a *L. major* cDNA library and a *L. donovani* non-polysomal RNA library. Using a combination of computational analysis and experimental approaches, we classified 26 putative ncRNA in *L. major*, of these, 5 arising from intergenic regions and 21 from untranslated regions. In *L. donovani*, we classified 37 putative ncRNAs, of these, 7 arising from intergenic regions, and 30 from UTRs. Our results suggest, for the first time, that UTR-transcripts may be a common feature in the eukaryote *Leishmania* similarly to those previously shown in *T. brucei* and other eukaryotes.

In trypanosomatids, genes are organized into divergent or convergent bidirectional polycistronic units [1]. Unlike prokaryotes, trypanosomatid polycistrons contain hundreds of functionally unrelated genes. Trypanosomatid pre-mRNAs are processed by trans-splicing of a capped spliced-leader RNA to a splice acceptor site of a gene, coupled with the polyadenylation of the upstream gene [2]. Genetic organization in trypanosomatids brings the control of gene expression mainly to the post-transcriptional level [3,4]. In many different organisms, non-coding RNAs (ncRNAs) are linked to mRNA processing, mRNA stability and emerges as key players in a variety of regulatory processes, such as DNA replication, chromosome maintenance, transcriptional regulation, translation, transcript and protein stability, and the translocation of regulatory proteins (reviewed in [5,6]). During the last decades, several ncRNAs have been discovered and classified in known ncRNA families, using computational and experimental approaches [7 9].

Mercer and co-workers demonstrated the expression of distinct RNAs from untranslated regions (UTRs) in the transcriptome of humans, mice and flies and called these transcripts UTR-associated ncRNAs (uaRNAs), but the biogenesis mechanisms and biological functions remain a challenge [10]. In *Trypanosoma brucei*, whole transcriptome analyses revealed putative ncRNAs that have unambiguous alternative processing and generate, besides the full mRNA transcript, an UTR-internal transcript. Kolev and co-workers also found snoRNAs initially produced as long mRNA precursors, and these precursors may contain one or more snoRNA sequences within UTRs [11]. Subsequently, Michaeli and co-workers expanded the number of snoRNAs inside 3 UTRs and described novel putative ncRNAs processed by trans-splicing and polyadenylation [12]. Despite the evidence for the presence of several UTR-internal transcripts in *T. brucei* [11,12], to none of them a functional role has been assigned.

In *Leishmania major* species, a high-throughput RNA sequencing (RNA-seq) effort identified 10,285 transcripts; of these, approximately 18% are novel genes that, according to the authors, might be ncRNAs. This study represents the first annotation of the *L. major* genome to delimit the translated and untranslated regions of each gene [13].

This published *L. major* transcriptome prompted us to revisit previously explored clones from a cDNA library exhibiting unusual properties for protein encoding genes. These clones were investigated because they were too short and their GC content was too low to encode an Open Reading Frame compared to the average *Leishmania* protein coding sequence. We previously hypothesized that these odd transcripts were functional ncRNAs [14].

Corresponding author. Present address: Department of Cell and Molecular Biology, Ribeir o Preto Medical School, University of S o Paulo, Av. Bandeirantes, no. 3900, CEP 14049-900, Ribeir o Preto, SP, Brazil.

Email address: akcruz@fmrp.usp.br, angelakaysel@gmail.com (A.K. Cruz)

[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.molbiopara.2017.04.002
Received 7 October 2016; Received in revised form 29 March 2017; Accepted 1 April 2017
Available online xxx
0166-6851/ © 2016 Published by Elsevier Ltd.

UNCORRECTED PROOF

To define putative ncRNA among the so called *ODD* transcripts, we aligned 68 cDNA clones against the annotated *L. major* genome (Sanger Institute, version 6) by current transcriptome data (Fig. 1A) [13]. Ap-

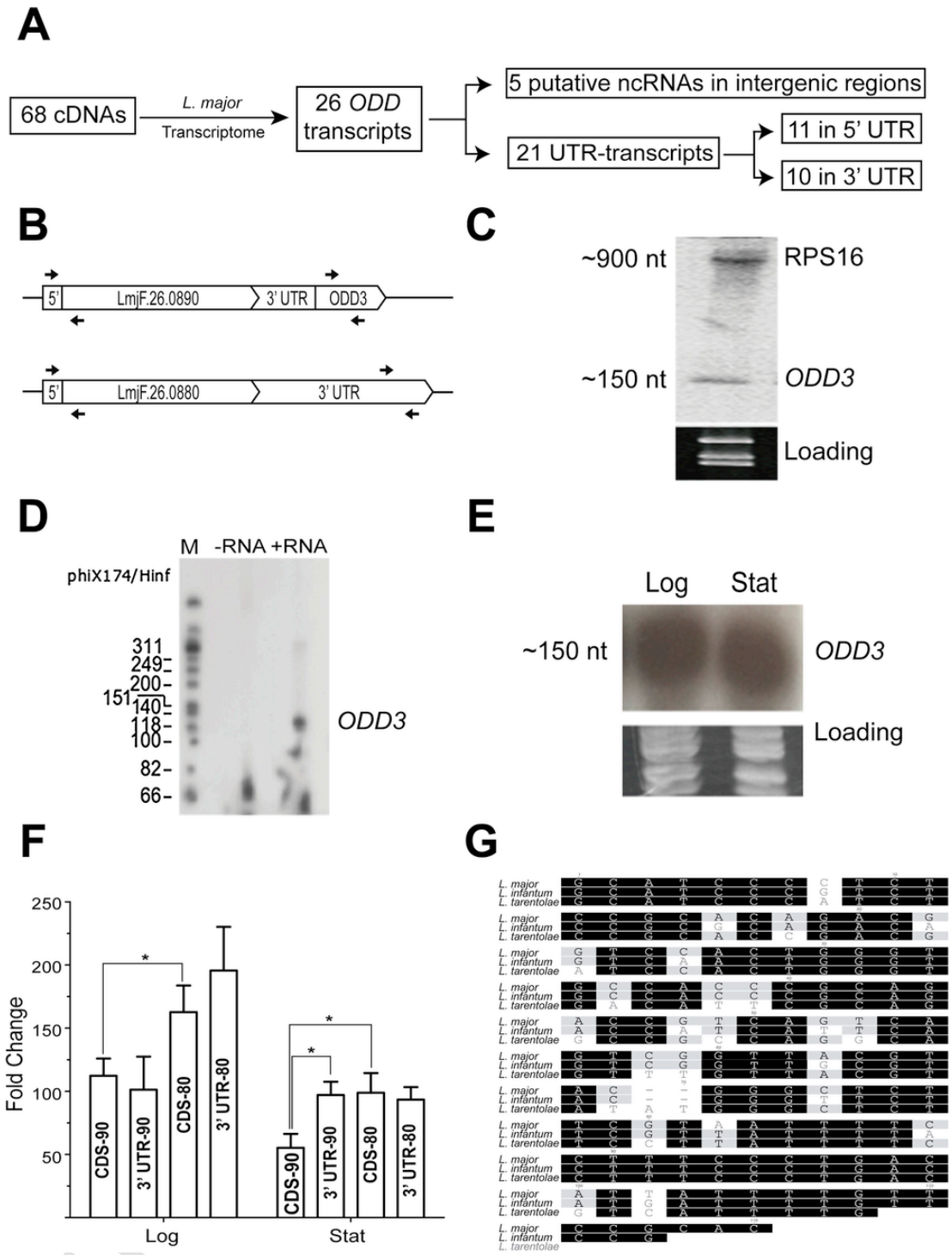


Fig. 1. *ODD3* is a UTR-transcript from LmjF.26.0890 gene. (A) Scheme of the *in silico* analysis to classify cDNAs as putative ncRNAs from *L. major*. (B) Representation of RPS16 genes. The primers used in RT-qPCR (Table S4) are indicated by black arrows (map not drawn to scale). (C) Northern blot of *L. major* total RNA fractionated on an 8% polyacrylamide 7 M urea gel using a specific antisense oligonucleotide (ODD3_antisense1) as probe (Table S5). (D) Autoradiography containing the products of the primer extension experiment using ODD3_antisense1 primer (Table S6) and fractionated in a 4 10% polyacrylamide gel. The molecular size marker (phiX174/*Hinf*) and corresponding fragments are presented (in base pairs) on the left side of the autoradiography. M (molecular size marker), −RNA (reaction performed without RNA) and +RNA (reaction performed with RNA). (E) Northern blotting analysis was performed using RNA extracted from different developmental stages of *L. major*. Log represents the logarithmic phase (3rd day), and Stat represents the stationary phase (6th day) of axenic promastigote culture. RNA was resolved on 8% polyacrylamide 7 M urea gels and was hybridized with ODD3_antisense2 primer probe (Table S5). In panels (C) and (E), total RNA stained with ethidium bromide was used as loading control (Loading). (F) Comparative expression analysis of LmjF.26.0890 (CDS-90), its 3 UTR (3 UTR-90), LmjF.26.0880 (CDS-80) and its 3 UTR (3 UTR-80), using RT-qPCR. Position of the primers used are depicted as black arrows in panel B and their sequence are presented on Table S4. Log represents the logarithmic phase (3rd day), and Stat represents the stationary phase (6th day) of axenic promastigote culture. Bars represent the mean SD from 3 independent experiments. Two-way ANOVA and Bonferroni post-tests were used for statistical analysis. Asterisks indicate statistically significant differences between samples ($p < 0.05$). (G) DNA sequence alignment of *ODD3* in *L. major*, *L. infantum* and *L. tarentolae*. Black boxes mean similarity of 100% (3/3) and gray boxes 75% (2/3).

proximately, 58% were excluded from the list of putative ncRNAs because they were aligned to coding DNA sequence (CDS), and 4% could not be aligned with the genome using the study parameters. These unaligned sequences might be part of unassembled regions of the *L. major* genome or might be technical artifacts. Among the 26 putative ncRNAs, 5 aligned in intergenic regions pointed by Rastrojo and co-workers [13] as novel genes and putative ncRNAs (Table S1). Interestingly, 21 mapped to untranslated regions: 11 were in 5 UTRs, and 10 were in 3 UTRs (two of which were overlapping sequences of a single 3 UTR) (Table 1).

The first of these putative ncRNAs to be investigated, *ODD1*, was previously detected by Northern blotting, as two distinct bands: a large

Table 1
xxx.

Organism	Putative ncRNA ID	Predicted Length (nt)	Annotation	Gene	ncRNA Score	Conservation Score
L. major	Lm_ncRNA1	369	3 UTR	LmjF.01.0110 (AA741618)	*	+ + + +
	Lm_ncRNA2 (ODD1)	224	5 UTR	LmjF.06.0390 (W88361)	*	+ + + +
	Lm_ncRNA4[a]	357	3 UTR	LmjF.06.1300 (AI034537)		+ + +
	Lm_ncRNA5[a]	439	3 UTR	LmjF.06.1300 (AI034621)		+ + +
	Lm_ncRNA7	273	5 UTR	LmjF.12.1250 (AA680652)	*	+ +
	Lm_ncRNA8	163	5 UTR	LmjF.18.0440 (AA741783)	**	+
	Lm_ncRNA9[b]	146	5 UTR	LmjF.19.0270 (AI034878)	**	+ +
	Lm_ncRNA10	344	5 UTR	LmjF.26.0240 (AA741583)	*	+ +
	Lm_ncRNA11 (ODD3)	114	3 UTR	LmjF.26.0890 (AA728081)	**	+ +
	Lm_ncRNA12	264	3 UTR	LmjF.27.1805 (AI034514)	*	+ + + +
	Lm_ncRNA13	197	5 UTR	LmjF.30.2200 (AA680639)	*	+ +
	Lm_ncRNA15	350	3 UTR	LmjF.31.3190 (AI034613)	**	+
	Lm_ncRNA17	325	3 UTR	LmjF.32.0890 (AI034648)	**	+ + + +
	Lm_ncRNA18	306	3 UTR	LmjF.34.0010 (AA728115)	**	+ + + +
	Lm_ncRNA19	300	5 UTR	LmjF.34.0770 (AA060773)	**	+ + +
	Lm_ncRNA20	83	5 UTR	LmjF.34.1170 (AA741590)	**	+ +
	Lm_ncRNA21	432	5 UTR	LmjF.34.3380 (AI034769)		+
	Lm_ncRNA22	463	3 UTR	LmjF.35.0010 (AI034487)	**	+
	Lm_ncRNA23	316	5 UTR	LmjF.35.1080 (AA680793)	**	+ + + +
	Lm_ncRNA24	394	3 UTR	LmjF.35.2440 (AI034539)	*	+ +
	Lm_ncRNA25	206	5 UTR	LmjF.35.2880 (AA680637)	**	+ +
L. donovani	Ld_ncRNA3	402	3 UTR	LdBPK_061360.1	**	+ + +
	Ld_ncRNA4	280	5 UTR	LdBPK_080650.1_1	*	+
	Ld_ncRNA5	235	3 UTR	LdBPK_080650.1_2	**	+ + + + +
	Ld_ncRNA8	334	5 UTR	LdBPK_100130.1_1	*	+ +
	Ld_ncRNA9	235	3 UTR	LdBPK_100130.1_2	**	+ + + +
	Ld_ncRNA10	100	3 UTR	LdBPK_100130.1_3	**	+ +
	Ld_ncRNA11	505	3 UTR	LdBPK_100130.1_4	-	+ + +
	Ld_ncRNA12	410	3 UTR	LdBPK_101440.1	**	+ + +
	Ld_ncRNA13	154	5 UTR	LdBPK_151130.1	**	+ +
	Ld_ncRNA14	343	3 UTR	LdBPK_151150.1	*	+ + + + +
	Ld_ncRNA15	163	3 UTR	LdBPK_200030.1	**	+
	Ld_ncRNA16	199	3 UTR	LdBPK_201320.1_1	***	+ + + +
	Ld_ncRNA17	298	3 UTR	LdBPK_201320.1_2	**	+ + +
	Ld_ncRNA20	126	3 UTR	LdBPK_211160.1	**	+ + +
	Ld_ncRNA21	64	5 UTR	LdBPK_211170.1	**	+ + +
	Ld_ncRNA22	667	3 UTR	LdBPK_211290.1	*	+ + +
	Ld_ncRNA23	136	3 UTR	LdBPK_212160.1	**	+
	Ld_ncRNA24	250	3 UTR	LdBPK_300920.1	***	+ +
	Ld_ncRNA25	334	5 UTR	LdBPK_302200.1		+
	Ld_ncRNA26	280	3 UTR	LdBPK_302210.1_3	**	+
	Ld_ncRNA27	253	3 UTR	LdBPK_302210.1_4	**	+ +
	Ld_ncRNA28	388	3 UTR	LdBPK_302210.1_5	**	+ +
	Ld_ncRNA29	550	3 UTR	LdBPK_302210.1_6	***	+ + + + +
	Ld_ncRNA30	190	3 UTR	LdBPK_302220.1	*	+ + +
	Ld_ncRNA31	116	3 UTR	LdBPK_340010.1	**	+ + + +
	Ld_ncRNA32	100	3 UTR	LdBPK_343430.1_1	**	+
	Ld_ncRNA33	83	3 UTR	LdBPK_343430.1_2	**	+
	Ld_ncRNA34	208	3 UTR	LdBPK_343440.1	*	+ +
	Ld_ncRNA35	141	5 UTR	LdBPK_352270.1_1	***	+ + +
	Ld_ncRNA36	330	3 UTR	LdBPK_352270.1_2	***	+ + +

L. major nucleotide sequence data is deposited in GenBank database under the accession numbers shown in column Gene in brackets. *L. donovani* raw sequencing data is available in SRA (Sequence Read Archive) database using accession number SRP090024.

a − cDNAs (AI034537 and AI034621) have similar coordinates in the same 3 UTR; b- cDNA expands from intergenic to the 5 UTR; ncRNA Score: Each symbol "*" means positive prediction in one of the computational tools (PORTRAIT, RNAcon, RNASpace and snoscan); Conservation Score: Symbol "+" means sequence conservation inside *Leishmania* subgenus, each "+" means sequence conservation in another genus (*T. brucei*) or subgenus. See supplementary materials for details.

UNCORRECTED PROOF

(3800 nt) and a small (400 nt) transcript. Interestingly, the short transcript was polyadenylated and more abundant in the cytoplasm, suggesting that the longer RNA might be a precursor. The *ODD1* gene is located on chromosome 6 as a single-copy gene [15]. Using recent transcriptome annotation, *ODD1* (Lm_ncRNA2) was shown to be within the 5 UTR of the LmjF.06.0390 gene (coding for a hypothetical protein) (Table 1). This information suggests that *ODD1* is a 5 UTR-transcript in *L. major*.

Next, we tested another putative ncRNA; the Lm_ncRNA11 transcript (named *ODD3*) was mapped in the terminal portion of the 3 UTR of LmjF.26.0890, one of the two copies of the ribosomal protein S16 (RP-S16) genes (Fig. 1B and Table 1). The CDSs of both copies are identical and the UTRs are divergent (Fig. S1).

A Northern blot of *L. major* total RNA using an *ODD3* antisense-specific probe revealed a large (900 nt) and a small (150 nt) RNA band (Fig. 1C). These features suggest that *ODD3* arises from one of the RP-S16 genes in *L. major* species. To further validate that *ODD3* is expressed from LmjF.26.0890 3 UTR, a primer extension assay was performed. *ODD3* expression was detected with an *ODD3* internal primer in a poly(A)-enriched RNA fraction (Fig. 1D).

Subsequently, we analyzed *ODD3* expression profile in axenic *L. major* promastigotes. RNA levels from log- and stationary-phase promastigotes was evaluated using Northern blotting and RT-qPCR. Similar amounts of the *ODD3* were observed between the log-phase and stationary-phase promastigotes, as measured by Northern blot using the specific antisense probe (Fig. 1E). No differences in 3 UTR of LmjF.26.0890 (*ODD3*) expression levels were observed in log- and stationary-phase promastigotes by RT-qPCR (Fig. 1F), reinforcing northern blotting result.

We investigated the relative expression of the LmjF.26.0880 and LmjF.26.0890 full-length transcripts using unique sequences of their 5 and 3 UTRs. The RT-qPCR result indicates that LmjF.26.0890 is less abundant than LmjF.26.0880 (Fig. 1F). More importantly, we observed a significant difference between the LmjF.26.0890 transcript level and its 3 UTR (*ODD3*) in the stationary-phase (Fig. 1F). In opposition, no difference of expression level was found between 3 UTR and LmjF.26.0880 gene in log- and stationary-phase promastigotes (Fig. 1F). There are several possible explanations for these findings, including that *ODD3*, a putative ncRNA, is more stable than the RPS16 mRNA in stationary-phase, which need to be further investigated.

In addition, we conducted a BLAST analysis to analyze *ODD3* sequence conservation in *Leishmania* species from different subgenera and *T. brucei*. *ODD3* sequence was found conserved in strains from subgenera *Leishmania* and *Sauroleishmania* suggesting selective pressure (Fig. 1G, Tables 1 and S2).

To categorize *ODD3* under one of the known ncRNA classes we used different computational tools. PORTRAIT [16], an algorithm for identifying non-coding RNAs on a transcriptome context using *ab initio* methods, ranked *ODD3* as a ncRNA with a probability score of 92.5%. RNAcon, using structural information [17] predicted a 5S rRNA-like structure and RNAspace [18] did not find any specific domain on *ODD3*. Snoscan [19] failed to identify *ODD3* as a snoRNAs (Tables 1 and S3), and so did a BLAST analysis against the *L. major* snoRNAs previously published [20].

Thus, our results suggest that *ODD3* may be a functional ncRNA; it has been detected in northern blotting and RT-qPCR during promastigote s axenic growth, computational analysis ranked the transcript with a high probability score to be a ncRNA and the sequence is highly conserved in different species of the subgenus *Leishmania* and it is also conserved in the subgenus *Sauroleishmania.* Nevertheless, we still lack evidence of a possible biological function, to be examined.

To further investigate expression of ncRNAs from UTRs in *Leishmania*, we extended the study to *Leishmania donovani* using transcriptomic data previously generated in our laboratory. Using RNA-seq technology, libraries from *L. donovani* promastigotes were prepared to enrich the fraction of non-polysomal RNAs bound to proteins (ribonucleoproteins, Fig. S2). A similar methodology was applied to *Trypanosoma brucei* to successfully reveal a rich repertoire of non-coding RNAs [12].

Six paired-end libraries corresponding to three independent replicates from log- and stationary-phase *L. donovani* promastigotes were prepared and sequenced using an Illumina MiSeq. A total of 1.7 10^6 reads was reported; each replicate number of reads varied from 1.2 to 6.4 10^5 (SRP090024). The *L. donovani* reference genome (TriTrypDB − version 6.0) [21] was used to align the reads (see Supplementary Material for a full description of the experiment). We selected non-CDS regions and searched for those with higher density of reads (see supplementary materials for details). The *L. donovani* UTRs were mapped using the transcriptome data from *L. major* [13] (Fig. 2).

We chose 37 putative ncRNAs, 7 mapped to intergenic regions and 30 mapped to UTRs of annotated protein genes: 6 were in 5 UTRs, and 24 were in 3 UTRs (Tables 1 and S1). Five of these were detected by Northern blotting; 4 were putative UTR-transcripts (Fig. 2), and 1 (Ld_ncRNA6) was in an intergenic region from chromosome 10 (Fig. S3). UTR-transcripts were observed in both the 5 and 3 UTRs of the LdBPK_080650.1 gene (Ld_ncRNA4 and Ld_ncRNA5, respectively), which codes for a hypothetical protein (Fig. 2A); in the 3 UTR of LdBPK_300920.1 gene (Ld_ncRNA24) which codes for amastin (Fig. 2B) and another putative ncRNA was identified in the 3 UTR of the LdBPK_302210.1 gene (Ld_ncRNA28), which codes for a hypothetical protein (Fig. 2C). UTR-transcripts were detected in log- and stationary-phase promastigotes by Northern blots, and the estimated sizes of the predominant bands are compatible with the *in silico* predicted sizes (Fig. 2).

We conducted an *in silico* analysis of the 63 putative ncRNAs from *L. major* and *L. donovani* using the same bioinformatics approaches. The used algorithms, as expected, classified differently the putative ncRNAs (Tables 1 and S3). Interestingly, RNAcon and RNASpace classified some of the transcripts as catalytic RNAs. According to RNAcon, 18 transcripts were classified as catalytic introns and RNASpace identifies 1 of the transcripts as Ribonuclease P. Another group of transcripts were classified as IRES (Internal Ribosome Entry Site) (8 with RNAcon and 3 with RNASpace). Snoscan classified only 4 transcripts as snoRNA and RNASpace identified 3 different transcripts within this class (Table S3). Of note, none of the transcripts, under BLAST analysis, shared similarity with those identified in *L. major* by Eliaz and cols [21]. RNAcon and RNASpace failed to classify 19 and 52 transcripts, respectively. PORTRAIT classified 45 transcripts as putative ncRNAs.

We hypothesize that the UTR-transcripts might be ncRNAs for; (i) they were detected in Northern blots in different culture conditions (log- and stationary-phase, Figs. 1 and 2), (ii) they are conserved in different *Leishmania* species (Tables 1 and S2) and, (iii) they received at least one positive prediction as an ncRNA using three different programs developed for this purpose (RNAcon, RNAspace, PORTRAIT and snoscan) (Tables 1 and S3).

We have no information so far on the mechanisms of biogenesis of the UTR-transcripts. Not necessarily all of UTR-transcripts have the same biogenesis mechanism. Some of these RNAs might be the result of transcription initiation sites internal to PTUs, might be by-products of the polycistronic nature of transcription in the parasite or might be alternative processing products, which may or may not be functional. Nevertheless, the whole group of *L. major* putative ncRNA came from a cDNA library designed to rescue trans-spliced and polyadenylated transcripts [22], suggesting they are alternative processing products. Some of them might be self-splicing RNAs as indicated by the conducted computational analysis. The observed conservation of the transcripts in

UNCORRECTED PROOF

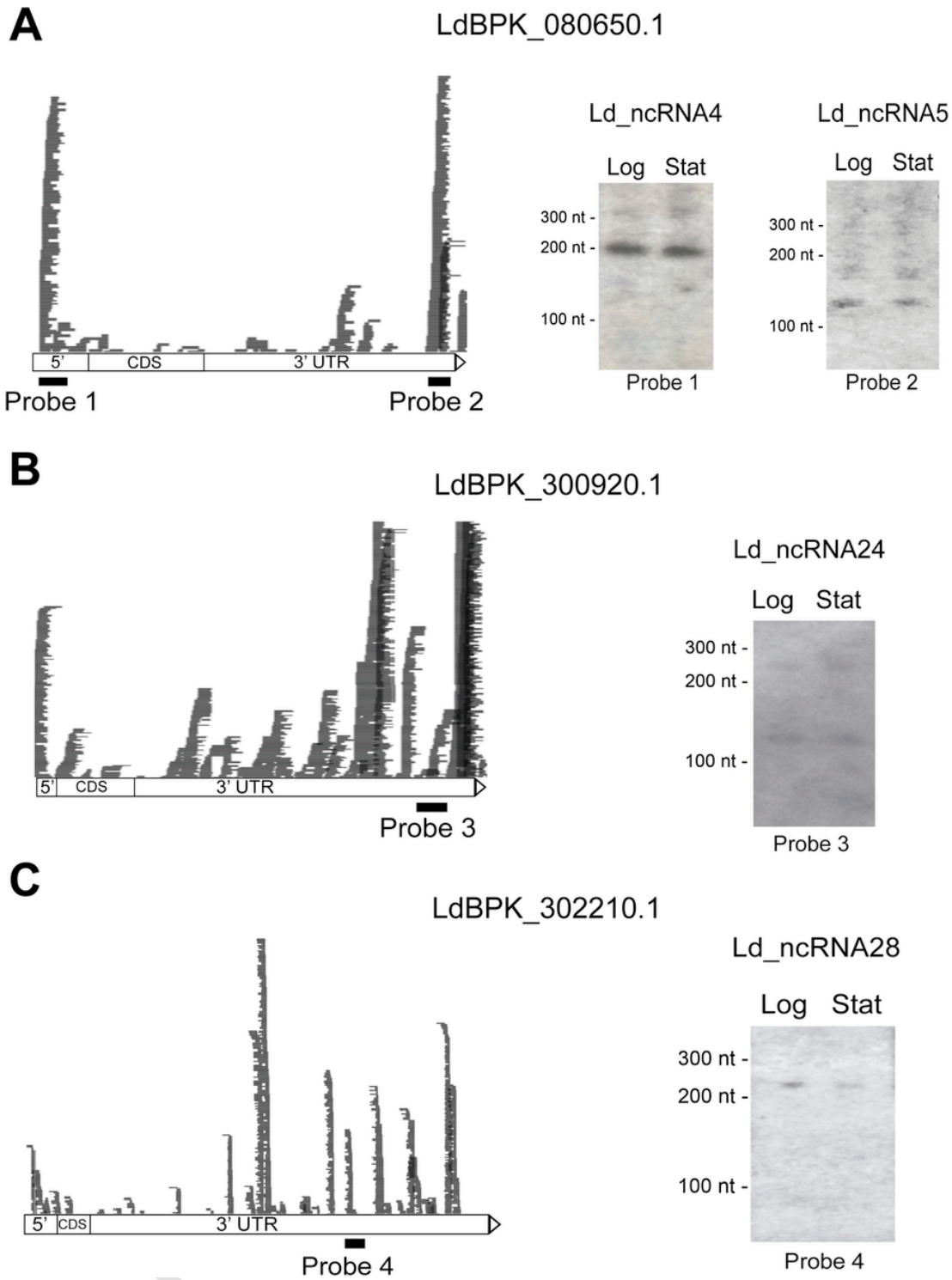


Fig. 2. Identification of putative UTR-transcripts in *L. donovani*. Non-polysomal RNA fractions from *L. donovani* promastigotes in log- and stationary-phase of axenic culture were extracted, and RNA-seq libraries were constructed and processed as described in the supplementary material. Images were generated with Artemis [23]. Panels A C: Representation to scale of three different loci: (A) LdBPK_080650.1, (B) LdBPK_300920.1 and (C) LdBPK_302210.1. The CDSs and UTRs are indicated in the panels, the stacked black lines represent the mapped reads, and the black boxes indicate the region targeted by the designed probes used in the Northern blotting experiments. Probes 1 4 are indicated in the panels and were designed as described in the Supplementary Material (Table S5 and text). The Northern blot was performed using RNA extracted from different developmental stages of *L. donovani*. Log represents the logarithmic phase (3rd day), and Stat represents the stationary phase (6th day) of axenic promastigote culture. RNA was resolved on 8% polyacrylamide 7 M urea gels.

Leishmania from different subgenera might be suggestive of selective pressure and putative functional roles.

In conclusion, our results suggest that UTR-transcripts may be a common feature in the early-branching eukaryote *Leishmania* similarly to those previously shown in *T. brucei* [11,12] and other eukaryotes [10]. Further investigation must be conducted with the herein described UTR-transcripts to understand biogenesis mechanisms and biological function.

Acknowledgments

We thank Viviane Ambr sio and T nia Paula Aquino Defina for technical assistance. We thank Dr. Stephen M. Beverley for providing *L. donovani* BOB. We thank Jeziel Dener Damasceno for fruitful discussion. We thank the Department of Cell and Molecular Biology for support. This work was supported by FAPESP (2009/14184-0 and 2013/50219-9) and CNPq. FFC (2012/17025-3), RFS (2012/16749-8) and KCNZ (2006/01252-0) were supported by FAPESP fellowships, PCR was supported by CAPES, and JST was supported by CNPq (151599/2008-4) fellowships.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.molbiopara.2017.04.002.

References

[1] S. Mart nez-Calvillo, S. Yan, D. Nguyen, M. Fox, K. Stuart, P.J. Myler, Transcription of Leishmania major Friedlin chromosome 1 initiates in both directions within a single region, Mol. Cell 11 (5) (2003) 1291–1299.
[2] J.H. LeBowitz, H.Q. Smith, L. Rusche, S.M. Beverley, Coupling of poly(A) site selection and trans-splicing in Leishmania, Genes Dev. 7 (6) (1993) 996–1007.
[3] A.C. Ivens, C.S. Peacock, E.A. Worthey, L. Murphy, G. Aggarwal, M. Berriman, E. Sisk, M.A. Rajandream, E. Adlem, R. Aert, A. Anupama, Z. Apostolou, P. Attipoe, N. Bason, C. Bauser, A. Beck, S.M. Beverley, G. Bianchettin, K. Borzym, G. Bothe, C.V. Bruschi, M. Collins, E. Cadag, L. Ciarloni, C. Clayton, R.M. Coulson, A. Cronin, A.K. Cruz, R.M. Davies, J. De Gaudenzi, D.E. Dobson, A. Duesterhoeft, G. Fazelina, N. Fosker, A.C. Frasch, A. Fraser, M. Fuchs, C. Gabel, A. Goble, A. Goffeau, D. Harris, C. Hertz-Fowler, H. Hilbert, D. Horn, Y. Huang, S. Klages, A. Knights, M. Kube, N. Larke, L. Litvin, A. Lord, T. Louie, M. Marra, D. Masuy, K. Matthews, S. Michaeli, J.C. Mottram, S. M ller-Auer, H. Munden, S. Nelson, H. Norbertczak, K. Oliver, S. O'neil, M. Pentony, T.M. Pohl, C. Price, B. Purnelle, M.A. Quail, E. Rabbinowitsch, R. Reinhardt, M. Rieger, J. Rinta, J. Robben, L. Robertson, J.C. Ruiz, S. Rutter, D. Saunders, M. Sch fer, J. Schein, D.C. Schwartz, K. Seeger, A. Seyler, S. Sharp, H. Shin, D. Sivam, R. Squares, S. Squares, V. Tosato, C. Vogt, G. Volckaert, R. Wambutt, T. Warren, H. Wedler, J. Woodward, S. Zhou, W. Zimmermann, D.F. Smith, J.M. Blackwell, K.D. Stuart, B. Barrell, et al., The genome of the kinetoplastid parasite, Leishmania major, Science 309 (5733) (2005) 436–442.
[4] C.E. Clayton, Life without transcriptional control? From fly to man and back again, EMBO J. 21 (8) (2002) 1881–1888.
[5] J.S. Mattick, The functional genomics of noncoding RNA, Science 309 (5740) (2005) 1527–1528.
[6] K.V. Morris, J.S. Mattick, The rise of regulatory RNA, Nat. Rev. Genet. 15 (6) (2014) 423–437.
[7] C. Wang, L. Wei, M. Guo, Q. Zou, Computational approaches in detecting noncoding RNA, Curr. Genom. 14 (6) (2013) 371–377.
[8] D. Veneziano, G. Nigita, A. Ferro, Computational approaches for the analysis of ncRNA through deep sequencing techniques, Front. Bioeng. Biotechnol. 3 (2015) 77.
[9] A. H ttenhofer, J. Vogel, Experimental approaches to identify non-coding RNAs, Nucleic Acids Res. 34 (2) (2006) 635–646.
[10] T.R. Mercer, D. Wilhelm, M.E. Dinger, G. Sold , D.J. Korbie, E.A. Glazov, V. Truong, M. Schwenke, C. Simons, K.I. Matthaei, R. Saint, P. Koopman, J.S. Mattick, Expression of distinct RNAs from 3' untranslated regions, Nucleic Acids Res. 39 (6) (2011) 2393–2403.
[11] N.G. Kolev, J.B. Franklin, S. Carmi, H. Shi, S. Michaeli, C. Tschudi, The transcriptome of the human pathogen Trypanosoma brucei at single-nucleotide resolution, PLoS Pathog. 6 (9) (2010) e1001090.
[12] S. Michaeli, T. Doniger, S.K. Gupta, O. Wurtzel, M. Romano, D. Visnovezky, R. Sorek, R. Unger, E. Ullu, RNA-seq analysis of small RNPs in Trypanosoma brucei reveals a rich repertoire of non-coding RNAs, Nucleic Acids Res. 40 (3) (2012) 1282–1298.
[13] A. Rastrojo, F. Carrasco-Ramiro, D. Mart n, A. Crespillo, R.M. Reguera, B. Aguado, J.M. Requena, The transcriptome of Leishmania major in the axenic promastigote stage: transcript annotation and relative expression levels by RNA-seq, BMC Genom. 14 (2013) 223.
[14] M.P. Iribar, A.K. Cruz, Base compositional bias in trans-spliced sequences of unknown function in Leishmania major, Exp. Parasitol. 100 (1) (2002) 1–5.
[15] M.P. Iribar, L.R. Tosi, A.K. Cruz, A processed short transcript of Leishmania, ODD1, Mol. Biochem. Parasitol. 127 (2) (2003) 205–208.
[16] R.T. Arrial, R.C. Togawa, M.e.M. Brigido, Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: case study of the pathogenic fungus Paracoccidioides brasiliensis, BMC Bioinf. 10 (2009) 239.
[17] B. Panwar, A. Arora, G.P. Raghava, Prediction and classification of ncRNAs using structural information, BMC Genom. 15 (2014) 127.
[18] M.J. Cros, A. de Monte, J. Mariette, P. Bardou, B. Grenier-Boley, D. Gautheret, H. Touzet, C. Gaspin, RNAspace.org: an integrated environment for the prediction, annotation, and analysis of ncRNA, RNA 17 (11) (2011) 1947–1956.
[19] T.M. Lowe, S.R. Eddy, A computational screen for methylation guide snoRNAs in yeast, Science 283 (5405) (1999) 1168–1171.
[20] D. Eliaz, T. Doniger, I.D. Tkacz, V.K. Biswas, S.K. Gupta, N.G. Kolev, R. Unger, E. Ullu, C. Tschudi, S. Michaeli, Genome-wide analysis of small nucleolar RNAs of Leishmania major reveals a rich repertoire of RNAs involved in modification and processing of rRNA, RNA Biol. 12 (11) (2015) 1222–1255.
[21] M. Aslett, C. Aurrecoechea, M. Berriman, J. Brestelli, B.P. Brunk, M. Carrington, D.P. Depledge, S. Fischer, B. Gajria, X. Gao, M.J. Gardner, A. Gingle, G. Grant, O.S. Harb, M. Heiges, C. Hertz-Fowler, R. Houston, F. Innamorato, J. Iodice, J.C. Kissinger, E. Kraemer, W. Li, F.J. Logan, J.A. Miller, S. Mitra, P.J. Myler, V. Nayak, C. Pennington, I. Phan, D.F. Pinney, G. Ramasamy, M.B. Rogers, D.S. Roos, C. Ross, D. Sivam, D.F. Smith, G. Srinivasamoorthy, C.J. Stoeckert, S. Subramanian, R. Thibodeau, A. Tivey, C. Treatman, G. Velarde, H. Wang, TriTrypDB: a functional genomic resource for the Trypanosomatidae, Nucleic Acids Res. 38 (2010) D457–62. Database issue.
[22] R. Almeida, B.J. Gilmartin, S.H. McCann, A. Norrish, A.C. Ivens, D. Lawson, M.P. Levick, D.F. Smith, S.D. Dyall, D. Vetrie, T.C. Freeman, R.M. Coulson, I. Sampaio, H. Schneider, J.M. Blackwell, Expression profiling of the Leishmania life cycle: cDNA arrays identify developmentally regulated genes present but not annotated in the genome, Mol. Biochem. Parasitol. 136 (1) (2004) 87–100.
[23] T. Carver, S.R. Harris, M. Berriman, J. Parkhill, J.A. McQuillan, Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data, Bioinformatics 28 (4) (2012) 464–469.

UNCORRECTED PROOF

ELSEVIER LICENSE
TERMS AND CONDITIONS

May 25, 2017

This Agreement between Ms. Patricia Ruy ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	4116060960466
License date	May 25, 2017
Licensed Content Publisher	Elsevier
Licensed Content Publication	Molecular and Biochemical Parasitology
Licensed Content Title	Evidence of putative non-coding RNAs from Leishmania untranslated regions
Licensed Content Author	Felipe Freitas Castro,Patricia C. Ruy,Karina Nogueira Zeviani,Ramon Freitas Santos,Juliano Simões Toledo,Angela Kaysel Cruz
Licensed Content Date	Jun 1, 2017
Licensed Content Volume	214
Licensed Content Issue	n/a
Licensed Content Pages	6
Start Page	69
End Page	74
Type of Use	reuse in a thesis/dissertation
Portion	full article
Format	both print and electronic
Are you the author of this Elsevier article?	Yes
Will you be translating?	No
Order reference number	
Title of your thesis/dissertation	in silico analysis of ncRNAs with RNA-seq data in Leishmania.
Expected completion date	Jun 2017
Estimated size (number of pages)	130
Elsevier VAT number	GB 494 6272 12
Requestor Location	Ms. Patricia Ruy Rua Magda Perona Frossard, 120 - ap 205 Ribeirão Preto, 14026-596 Brazil Attn: Ms. Patricia Ruy
Total	0.00 USD

Terms and Conditions

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.
3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:
"Reprinted from Publication title, Vol /edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."
4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.
5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier Ltd. (Please contact Elsevier at permissions@elsevier.com). No modifications can be made to any Lancet figures/tables and they must be reproduced in full.
6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.
7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.
8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.
9. Warranties: Publisher makes no representations or warranties with respect to the licensed material.

10. Indemnity: You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.
11. No Transfer of License: This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.
12. No Amendment Except in Writing: This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).
13. Objection to Contrary Terms: Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.
14. Revocation: Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to Elsevier and/or Copyright Clearance Center for denied permissions.

LIMITED LICENSE

The following terms and conditions apply only to specific license types:
15. **Translation**: This permission is granted for non-exclusive world **English** rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article.
16. **Posting licensed content on any Website**: The following terms and conditions apply as follows: Licensing material from an Elsevier journal: All content posted to the web site must maintain the copyright information line on the bottom of each image; A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx or the Elsevier homepage for books at http://www.elsevier.com; Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.
Licensing material from an Elsevier book: A hyper-text link must be included to the Elsevier homepage at http://www.elsevier.com . All content posted to the web site must maintain the copyright information line on the bottom of each image.

Posting licensed content on Electronic reserve: In addition to the above the following clauses are applicable: The web site must be password-protected and made available only to bona fide students registered on a relevant course. This permission is granted for 1 year only. You may obtain a new license for future website posting.
17. **For journal authors:** the following clauses are applicable in addition to the above:
Preprints:
A preprint is an author's own write-up of research results and analysis, it has not been peer-

reviewed, nor has it had any other value added to it by a publisher (such as formatting, copyright, technical enhancement etc.).
Authors can share their preprints anywhere at any time. Preprints should not be added to or enhanced in any way in order to appear more like, or to substitute for, the final versions of articles however authors can update their preprints on arXiv or RePEc with their Accepted Author Manuscript (see below).
If accepted for publication, we encourage authors to link from the preprint to their formal publication via its DOI. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help users to find, access, cite and use the best available version. Please note that Cell Press, The Lancet and some society-owned have different preprint policies. Information on these policies is available on the journal homepage.
Accepted Author Manuscripts: An accepted author manuscript is the manuscript of an article that has been accepted for publication and which typically includes author-incorporated changes suggested during submission, peer review and editor-author communications.
Authors can share their accepted author manuscript:

- immediately
    - via their non-commercial person homepage or blog
    - by updating a preprint in arXiv or RePEc with the accepted manuscript
    - via their research institute or institutional repository for internal institutional uses or as part of an invitation-only research collaboration work-group
    - directly by providing copies to their students or to research collaborators for their personal use
    - for private scholarly sharing as part of an invitation-only work group on commercial sites with which Elsevier has an agreement
- After the embargo period
    - via non-commercial hosting platforms such as their institutional repository
    - via commercial sites with which Elsevier has an agreement

In all cases accepted manuscripts should:

- link to the formal publication via its DOI
- bear a CC-BY-NC-ND license - this is easy to do
- if aggregated with other manuscripts, for example in a repository or other site, be shared in alignment with our hosting policy not be added to or enhanced in any way to appear more like, or to substitute for, the published journal article.

Published journal article (JPA): A published journal article (PJA) is the definitive final record of published research that appears or will appear in the journal and embodies all value-adding publishing activities including peer review co-ordination, copy-editing, formatting, (if relevant) pagination and online enrichment.
Policies for sharing publishing journal articles differ for subscription and gold open access articles:
Subscription Articles: If you are an author, please share a link to your article rather than the full-text. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help your users to find, access, cite, and use the best available version.
Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.
If you are affiliated with a library that subscribes to ScienceDirect you have additional

private sharing rights for others' research accessed under that agreement. This includes use for classroom teaching and internal training at the institution (including use in course packs and courseware programs), and inclusion of the article for grant funding purposes.
Gold Open Access Articles: May be shared according to the author-selected end-user license and should contain a CrossMark logo, the end user license, and a DOI link to the formal publication on ScienceDirect.
Please refer to Elsevier's posting policy for further information.
18. **For book authors** the following clauses are applicable in addition to the above: Authors are permitted to place a brief summary of their work online only. You are not allowed to download and post the published electronic version of your chapter, nor may you scan the printed edition to create an electronic version. **Posting to a repository:** Authors are permitted to post a summary of their chapter only in their institution's repository.
19. **Thesis/Dissertation**: If your license is for use in a thesis/dissertation your thesis may be submitted to your institution in either print or electronic form. Should your thesis be published commercially, please reapply for permission. These requirements include permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis and include permission for Proquest/UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission. Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

Elsevier Open Access Terms and Conditions
You can publish open access with Elsevier in hundreds of open access journals or in nearly 2000 established subscription journals that support open access publishing. Permitted third party re-use of these open access articles is defined by the author's choice of Creative Commons user license. See our open access license policy for more information.
Terms & Conditions applicable to all Open Access articles published with Elsevier:
Any reuse of the article must not represent the author as endorsing the adaptation of the article nor should the article be modified in such a way as to damage the author's honour or reputation. If any changes have been made, such changes must be clearly indicated.
The author(s) must be appropriately credited and we ask that you include the end user license and a DOI link to the formal publication on ScienceDirect.
If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source it is the responsibility of the user to ensure their reuse complies with the terms and conditions determined by the rights holder.
Additional Terms & Conditions applicable to each Creative Commons user license:
CC BY: The CC-BY license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article and to make commercial use of the Article (including reuse and/or resale of the Article by commercial entities), provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by/4.0.
CC BY NC SA: The CC BY-NC-SA license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article, provided this is not done for commercial purposes, and that the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. Further, any new works must be made available on the same conditions. The full details of the license are available at http://creativecommons.org/licenses/by-nc-sa/4.0.

CC BY NC ND: The CC BY-NC-ND license allows users to copy and distribute the Article, provided this is not done for commercial purposes and further does not permit distribution of the Article if it is changed or edited in any way, and provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, and that the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by-nc-nd/4.0. Any commercial reuse of Open Access articles published with a CC BY NC SA or CC BY NC ND license requires permission from Elsevier and will be subject to a fee.
Commercial reuse includes:

- Associating advertising with the full text of the Article
- Charging fees for document delivery or access
- Article aggregation
- Systematic distribution via e-mail lists or share buttons

Posting or linking by commercial companies for use by customers of those companies.

20. **Other Conditions**:

v1.9

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.