UNIVERSIDADE FEDERAL DE OURO PRETO
Núcleo de Pesquisas em Ciências Biológicas
Programa de Pós-Graduação em Biotecnologia

Victor Fernandes de Oliveira

RNAS NÃO CODIFICADORES LONGOS EM *Schistosoma mansoni*:
Predição e perfil de expressão diferencial no estágio evolutivo de verme adulto

Ouro Preto
2018

Victor Fernandes de Oliveira

RNAS NÃO CODIFICADORES LONGOS EM *Schistosoma mansoni*: Predição e perfil de expressão diferencial no estágio evolutivo de verme adulto

Tese de doutorado submetida ao programa de Pós-Graduação em Biotecnologia do Núcleo de Pesquisa em Ciências Biológicas da Universidade Federal de Ouro Preto, como parte integrante dos requisitos para obtenção do título de Doutor em Biotecnologia.

Orientadora: Profª. Drª. Renata Guerra de Sá Cota

Área de concentração Genômica e Proteômica.

Ouro Preto

O41r Oliveira, Victor Fernandes de .
RNAs Não Codificadores Longos em Schistosoma mansoni: [manuscrito]: Predição e perfil de expressão diferencial no estágio evolutivo de verme adulto / Victor Fernandes de Oliveira. - 2018.
129f.: il.: color; grafs; tabs; mapas.

Orientador: Profa. Dra. Renata Guerra de Sá Cota.

Tese (Doutorado) - Universidade Federal de Ouro Preto. Pró-Reitoria de Pesquisa e Pós-Graduação. Núcleo de Pesquisas em Ciências Biológicas. Programa de Pós-Graduação em Biotecnologia.
Área de Concentração: Genômica e Proteômica.

1. Schistosoma mansoni. 2. Regulação de expressão gênica. 3. Bioinformática. I. Cota, Renata Guerra de Sá. II. Universidade Federal de Ouro Preto. III. Titulo.

CDU: 579.258

Catalogação: www.sisbin.ufop.br

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA

UFOP

ATA DE DEFESA DE TESE

Aos 16 dias do mês de março do ano de 2018, às 13:30 horas, nas dependências Núcleo de Pesquisas em Ciências Biológicas (Nupeb), foi instalada a sessão pública para a defesa de tese do doutorando Victor Fernandes de Oliveira, sendo a banca examinadora composta pela Profa. Renata Guerra de Sá Cota (Presidente - UFOP), pelo Prof. Henrique Cota de Freitas (Membro - Externo), pelo Prof. Leandro Marcio Moreira (Membro - UFOP), pelo Prof. Matheus de Souza Gomes (Membro - Externo), pelo Prof. William de Castro Borges (Membro - UFOP). Dando início aos trabalhos, a presidente, com base no regulamento do curso e nas normas que regem as sessões de defesa de tese, concedeu ao doutorando Victor Fernandes de Oliveira 40 minutos para apresentação do seu trabalho intitulado "Rnas Não Codificadores Longos em Schistosoma Mansoni: Predição e Perfil de Expressão Diferencial no Estágio Evolutivo de Verme Adulto", na área de concentração: Genômica e Proteômica. Terminada a exposição, a presidente da banca examinadora concedeu, a cada membro, um tempo para perguntas e respostas ao candidato sobre o conteúdo da tese. Dando continuidade, ainda de acordo com as normas que regem a sessão, a presidente solicitou aos presentes que se retirassem do recinto para que a banca examinadora procedesse à análise e decisão, anunciando, a seguir, publicamente, que o doutorando foi aprovado por unanimidade, sob a condição de que a versão definitiva da tese deva incorporar todas as exigências da banca, devendo o exemplar final ser entregue no prazo máximo de 60 (sessenta) dias à Coordenação do Programa. Para constar, foi lavrada a presente ata que, após aprovada, vai assinada pelos membros da banca examinadora e pelo doutorando. Ouro Preto, 16 de março de 2018.

Presidente: Renata Guerra de Sá Cota

Membro: Henrique Cota de Freitas

Membro:

Membro:

Membro: William de Castro Borges

Doutorando: Victor Fernandes de Oliveira

Este trabalho foi desenvolvido no Laboratório de Bioquímica e Biologia Molecular NUPEB/ICEB/UFOP em Ouro Preto, Minas Gerais, e no Max Planck Institut für molekulare Genetik em Berlim, Alemanha, com apoio financeiro da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

Dedico esta tese aos meus pais, Alvimar e Zenaide, fontes de força, amor, carinho e apoio incondicional perante todas as dificuldades.

AGRADECIMENTOS

A Deus, por sempre ter guiado meus passos durante toda a minha vida.

Aos meus pais, Alvimar e Zenaide, pelo apoio, carinho e amor. Obrigado por tudo. Amo vocês!;

Aos meus irmãos Isis e Igor pelo apoio, amizade e companheirismo em vários momentos dessa etapa;

À Prof.ª Dr.ª Renata Guerra de Sá Cota, que me deu a oportunidade de conhecer e trabalhar com a biologia molecular durantes todos esses anos desde a graduação até doutorado. Obrigado pelos ensinamentos práticos e teóricos, pelo carinho, paciência, e sobretudo, pela orientação;

Ao Bruno pelo companheirismo e amizade durante todos esses anos do mestrado e doutorado e sobretudo, pelo apoio incondicional nos momentos mais difíceis;

Ao Wagner pela amizade construída durante todos esses anos na academia desde o primeiro dia da minha vida na UFOP;

Aos meus amigos, Ester, Fran, Diogo e Allan, por compartilhar momentos únicos e muito especiais em Ouro Preto;

A todos os amigos que estão presentes ou que passaram pelo laboratório (LBBM): Lauro, Regina, Talita, Viviano, Grazi, Daiane, Isabela, Luiza, Monica, Roberta V., Pollyana, Karina, Sávio, Natália, Matheus, Camila, Érica, Leandro, Raquel, Isabel, Carrol, Nayara M., Soraya, Kelvin, Cíntia, Walmir... Obrigado! Essa trajetória foi mais agradável com a ajuda de vocês;

Aos demais amigos da graduação, e do mestrado;

Aos professores doutores do laboratório Elísio e Leandro pela amizade, aprendizado e convívio agradável;

Aos técnicos Ezequiel e Eduardo pelo auxílio;

Ao professor William pela amizade e ensinamentos no trabalho;

Aos demais professores da UFOP que participaram da minha formação acadêmica;

Aos amigos da Alemanha em especial Laura, Naty, Fernanda, Lisa e Sebastian;

E por todos aqueles que participaram de forma direta ou indireta na realização deste trabalho.

“Many places I have been
Many sorrows I have seen
But I don't regret
Nor will I forget
All who took the road with me

[...]

To these memories I will hold
With your blessing I will go
To turn at last to paths that lead home
And though where the road then takes me
I cannot tell
We came all this way
But now comes the day
To bid you farewell

I bid you all a very fond farewell”

(BOYD, 2014)

RESUMO

A esquistossomose é considerada uma doença debilitante de grande impacto socioeconômico no mundo causada por platelmintos do gênero *Schistosoma*. Uma das principais espécies identificadas nesse gênero é o parasita *Schistosoma mansoni*, que apresenta um ciclo de vida bastante complexo. Acredita-se que a complexidade dos programas de diferenciação e desenvolvimento observado entre os diferentes estágios evolutivos e ambientes onde o parasita vive seja dependente da regulação da expressão gênica. Os RNAs não codificantes representam uma das principais classes de moléculas que potencialmente controlam a regulação gênica nos níveis: epigenético, transcricional, pós-transcricional e traducional. Esses RNAs são divididos em dois grandes grupos, os RNAs pequenos não codificantes de proteína e os RNAs longos não codificantes (lncRNAs), o foco desse projeto. Os lncRNAs correspondem aos transcritos com mais de 200 nucleotídeos e que não codificam proteínas. Particularmente, estão envolvidos em diversas funções regulatórias celulares, como regular a transcrição, induzir o remodelamento da cromatina e modificações em histonas, originar siRNAs endógenos, modular atividades de proteínas, alterar a localização celular de proteínas, ser precursores de pequenos RNAs, entre outros. A hipótese desse trabalho é que o *S. mansoni* expresse um conjunto de lncRNAs de forma diferencial através da reprogramação de sua expressão gênica em determinados estágios evolutivos. Para investigá-la foi utilizado um conjunto de dados de RNA-seq disponível para a fase adulta do parasito objetivando: (i) a montagem de um *pipeline* para identificar e caracterizar lncRNAs a partir de dados de RNA-seq com alta confiança; (ii) classificar os novos lncRNAs preditos utilizando a anotação Gene Ontology; (iii) analisar a expressão de um conjunto de 20 lncRNAs em cercária, esquistossômulos com 3,5h de cultivo in vitro, machos, fêmeas, casal e ovos; (iv) analisar a expressão de lncRNAs em parasitos resistentes ao praziquantel e (v) analisar a expressão de lncRNAs em fígado de camundongo infectado e não infectado com *S. mansoni*. Foram identificados 170 novos Sm-lncRNAs com termos de ontologia relacionados ao metabolismo, transporte e biossíntese. Quinze dos preditos lncRNAs mostraram expressão diferencial nos estágios avaliados, bem como entre machos e fêmeas. Alguns apresentaram expressão diferencial em parasitos com resistência ao praziquantel e em fígado de camundongos infectados. Esses achados abrem novas perspectivas para estudos funcionais focados em resistência a essa droga e desenvolvimento de biomarcadores específicos para a esquistossomose.

Palavras-chave: *S. mansoni*, lncRNAs, regulação gênica, RNA-seq.

ABSTRACT

Schistosomiasis is a debilitating disease of great socioeconomic impact in the world caused by flatworms of the genus Schistosoma. One of the main species identified in this genus is the parasite *Schistosoma mansoni*, which presents a very complex life cycle. The differentiation and development programs complexity observed among the different evolutionary stages and environments where the parasite lives depend on the gene expression. Non-coding RNAs represent one of the major classes of molecules that potentially control gene regulation at levels: epigenetic, transcriptional, post-transcriptional, and translational. These RNAs are divided into two large groups, the small non-coding RNAs and the long non-coding RNAs (lncRNAs), the focus of this study. The lncRNAs correspond to transcripts with more than 200 nucleotides and are not responsible for protein coding. In particular, they are involved in several cellular regulatory functions, such as regulating transcription, inducing chromatin remodeling and histone modifications, originating endogenous siRNAs, modulating protein activities, altering cell localization of proteins, being precursors of small RNAs and others functions. This study hypothesis is that *S. mansoni* expresses a set of lncRNAs differentially by reprogramming their gene expression at certain evolutionary stages. To investigate it, a set of available RNA-seq data was used for the adult phase of the parasite aiming at: (i) assembling a pipeline to identify and characterize lncRNAs from highly reliable RNA-seq data; (ii) classify the novel lncRNAs predicted using the Gene Ontology annotation; (iii) to analyze the expression of a set of 20 lncRNAs in cercariae, schistosomules with 3.5 h of *in vitro* culture, males, females, couple and eggs; (iv) to analyze the expression of lncRNAs in praziquantel resistant parasites and (v) to analyze the expression of lncRNAs in the liver of infected and uninfected *S. mansoni* mice. We identified 170 new Sm-lncRNAs with terms of ontology related to metabolism, transport and biosynthesis. Fifteen of them showed differential expression in the evaluated stages, as well as between males and females. Some showed differential expression in parasites with resistance to praziquantel and in liver of infected mice. These findings highlight new perspectives for functional studies focused on resistance to this drug and specific biomarkers development for schistosomiasis.

Keywords: *S. mansoni*, lncRNAs, gene expression regulation, RNA-seq.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

aa: aminoácidos;

BAM: *Binary version of Sequence Alingment /Map*;

CAPES: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior;

cDNA: DNA complementar;

CEUA: Comissão de Ética no Uso de Animais;

CP: Classificador do Ponto de Corte Ideal;

CPAT: *Coding Potential Assessment Tool*;

CPqRR/FIOCRUZ: Centro de Pesquisa René Rachou, Fundação Oswaldo Cruz

EMT-3,5h: Esquistossômulos mecanicamente transformados *in vitro* de 3,5h;

ENCODE: *Encyclopedia of DNA Elements;*

eRNA: *Enhancer* RNA;

EST: *Expressed Sequence Tag;*

FAPEMIG: Fundação de Amparo à Pesquisa do Estado de Minas Gerais;

FAPESP: Fundação de Amparo à Pesquisa do Estado de São Paulo;

FMRP-USP: Faculdade de Medicina da Universidade de São Paulo, Campus Ribeirão Preto;

FPKM: *Fragments Per Kilobase of Transcript per Million Mapped;*

GO: *Gene Ontology*;

HPRT1: *Hypoxanthine Phosphoribosyltransferase 1;*

IGV: *Integrative Genomics Viewer*;

LBBM: Laboratório de Bioquímica e Biologia Molecular;

LEP: Laboratório de Enzimologia e Proteômica;

LE-PZQ: Vermes adultos resistentes ao PZQ;

lncRNA: *Long noncoding* RNA;

LRT: *Long Terminal Repeat*;

Mac OS: *Macintosh Operating System;*

Mb: Mega bases;

MEG: *Micro Exon Gene;*

miRNA: microRNA;

MPIMG: *Max Planck Institut für molekulare Genetik;*

NCBI: *National Center for Biotechnology Information*;

ncRNA: *noncoding* RNA;

nt: Nucleotídeos;

OMS: Organização Mundial de Saúde;

ORESTES: *Open Reading Frame ESTs;*

ORF: *Open Reading Frame;*

pb: Pares de bases;

PCR: *Polymerase Chain Reaction;*

PERL: *Practical Extraction and Report Language*;

piRNA: *Piwi-interacting* RNA;

PZQ: Praziquantel;

qRT-PCR: *Quantitative Reverse Transcription PCR;*

RNA-seq: Sequenciamento de RNA;

RPKM: *Reads Per Kilobase of Transcript per Million Mapped;*

RPMI: Meio *Roswell Park Memorial Institute;*

SAM: *Sequence Alignment/Map;*

SmeIF4E: *Eukaryotic translation initiation factor 4E*;

TIGR: *The Institute for Genomic Research*;

TLR: *Tool-Like Receptors;*

UFOP: Universidade Federal de Ouro Preto;

WHO: *World Health Organization.*

SUMÁRIO

1 INTRODUÇÃO

1.1 Esquistossomose

Acredita-se que a ocorrência da doença esquistossomose em humanos seja milenar. Evidências apontam que o contato humano com os ovos do parasito tenha acontecido há pelo menos 4.000 anos através de sua detecção em múmias chinesas e egípcias. Recentemente, pesquisadores evidenciaram a presença de ovos do parasito próximo aos ossos do quadril de uma múmia encontrada no norte da Síria e datada de aproximadamente 6.200 anos (ANASTASIOU et al., 2014).

A primeira descrição da doença foi feita pelo patologista alemão Theodor Maximilian Bilharz em 1851 ao identificar vermes adultos da espécie *Schistosoma haematobium* durante autópsias feitas em camponeses do Egito (BEAVER, 1984). A partir daí a esquistossomose foi descrita também como bilharziose em sua homenagem. No ano de 1902, o escocês Patrick Manson, ao realizar exames microscópicos, verificou que possivelmente existiriam duas espécies de Bilharzia. Ele descreveu que os ovos apresentavam diferenças morfológicas na posição da espícula e na sua localização no sítio da infecção, sendo um encontrado somente no reto e o outro podendo ser encontrado no reto e na bexiga. (MANSON, 1902). Posteriormente no ano de 1907, essa nova espécie, que apresentava diferenças na morfologia dos ovos, foi finalmente descrita por Louis Westenra Sambon (SAMBON, 1907) e nomeada de *Schistosoma mansoni* em homenagem a Patrick Manson. Somente no ano de 1908, o médico brasileiro Manuel Augusto Pirajá da Silva publicou um artigo do primeiro registro de esquistossomose no Brasil. Além dessa descoberta, ele foi a primeira pessoa a descrever efetivamente a morfologia dos ovos, dos vermes adultos e dos miracídios, diferenciando definitivamente o *S. mansoni* das outras espécies. Em 1912, ele publicou ainda um novo trabalho em que descreveu o estágio da cercária na esquistossomose (FALCÃO, 2008).

A esquistossomose humana é uma doença negligenciada considerada como a segunda parasitose tropical de maior impacto socioeconômico do mundo, segundo a Organização Mundial de Saúde (OMS), sendo superada somente pela malária (W.H.O., 2002). Pelo menos 261 milhões

de pessoas estão infectadas em todo o mundo, e mais de 800 milhões vivem em áreas endêmicas com risco de infecção (STEINMANN et al., 2006; W.H.O., 2015). A transmissão da esquistossomose é encontrada em mais de 78 países e territórios, sendo prevalente em áreas tropicais e subtropicais, incluindo a África, América Central, Sudeste da Ásia, China e Brasil (W.H.O., 2015) (Figura 1).

Atualmente mais de 20 espécies de parasitos do gênero *Schistosoma* já foram identificados (LAWTON; JOHNSTON; ROLLINSON, 2011). Entretanto, as principais espécies conhecidas que infectam o homem são: *S. mansoni*, *S. hematobium*, *Schistosoma japonicum*, *Schistosoma intercalatum*, *Schistosoma mekongi* e *Schistosoma guineenses.*

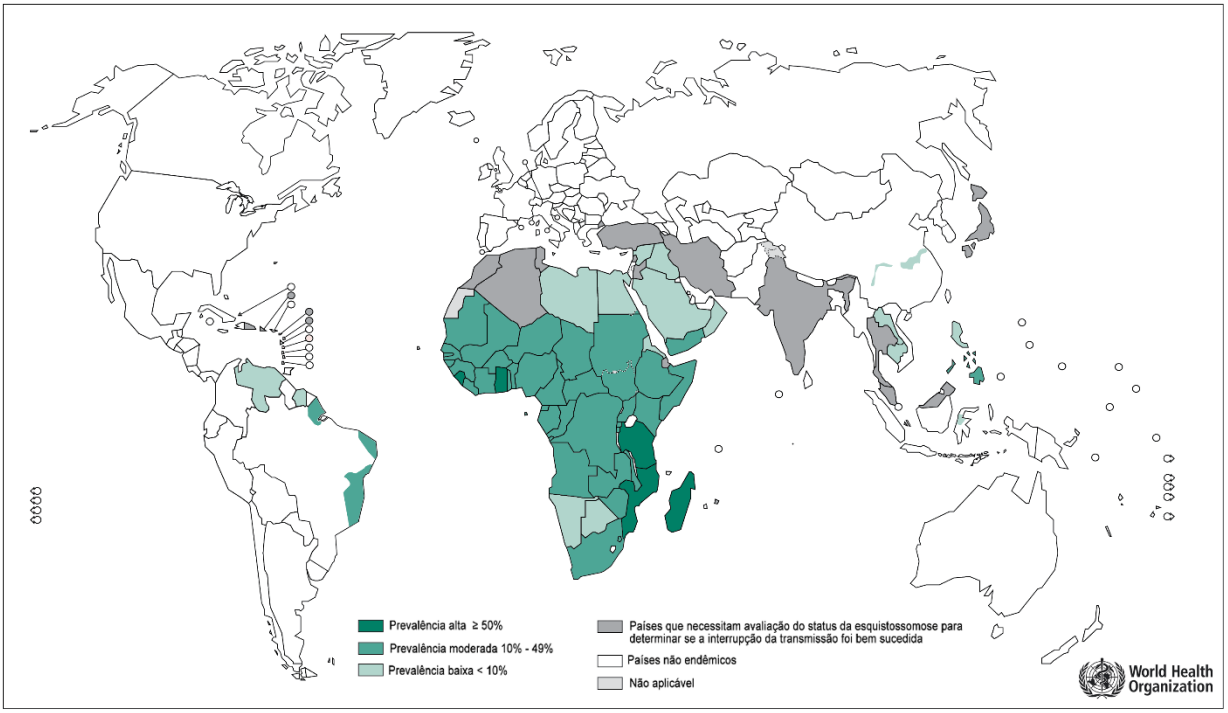


Figura 1: Distribuição mundial da esquistossomose. Distribuição da esquistossomose em todo o mundo, 2012. Adaptado de WHO (2014).

No cenário nacional, acredita-se que a introdução dessa doença aconteceu por meio do tráfico negreiro originário da costa ocidental da África ingressando principalmente pelos portos de Recife e Salvador e se expandindo pelo Brasil devido aos movimentos migratórios (Figura 2) (BRASIL, 2014). Os primeiros registros do *S. mansoni* foram realizadas pelo brasileiro Pirajá da Silva a partir de 1908, mas somente em 1950, que Pellon e Teixeira evidenciaram, através de um

inquérito nacional, a existência da esquistossomose mansônica em 612 das 877 localidades pesquisadas na região Nordeste, no Estado de Minas Gerais e posteriormente em mais seis estados (FREITAS, 1972). Esse Inquérito Nacional da Esquistossomose foi realizado em 11 estados brasileiros, utilizando o método de sedimentação das fezes. Foram examinadas 440.784 amostras de fezes de escolares de 7 a 14 anos em todo o Brasil, sendo 162.176 delas proveniente de Minas Gerais.

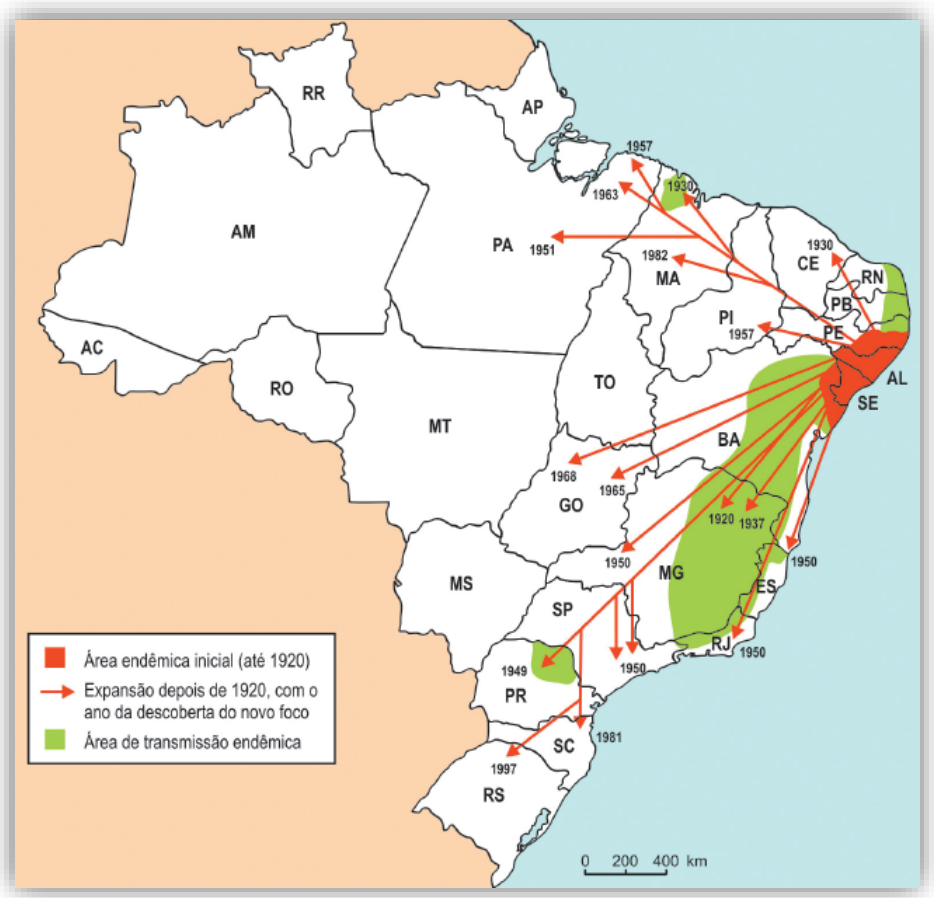


Figura 2: Representação da expansão da esquistossomose mansônica no Brasil. Adaptado de Brasil (2014).

Os dados mais recentes do Ministério da Saúde do Brasil demostram que a doença está presente em pelo menos 19 estados brasileiros e no Distrito Federal, dentre os quais destacam-se: Pernambuco, Alagoas, Bahia, Rio Grande do Norte, Paraíba, Sergipe, Espírito Santo e Minas Gerais (Figura 3). Nos demais estados, Pará, Maranhão, Piauí, Ceará, Rio de Janeiro, São Paulo,

Rondônia, Santa Catarina, Paraná, Rio Grande do Sul, Goiás e no Distrito Federal, a transmissão ocorre em focos específicos, não atingindo grandes áreas (BRASIL, 2009). Cerca de 7 milhões de pessoas podem estar vivendo com esquistossomose no Brasil e mais de 25 milhões vivendo em áreas de risco, sendo a maioria em áreas isoladas, sem diagnóstico e tratamento (BRASIL, 2009; SCHOLTE et al., 2014).

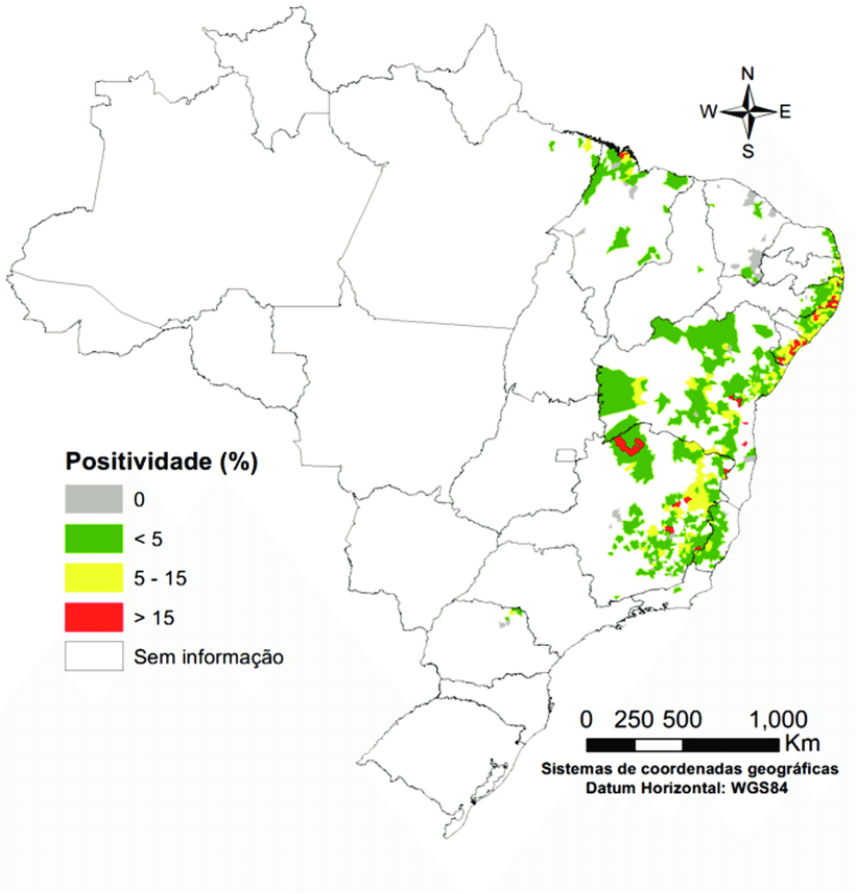


Figura 3: Distribuição da esquistossomose mansônica no Brasil. Percentual de positividade em inquéritos coproscópicos - Brasil 2012 (BRASIL, 2014).

1.2 Ciclo de vida do *S. mansoni* e patogenia da doença

O *S. mansoni* é um parasito platelminto pertencente a classe Trematoda, família Schistosomatidae e do gênero *Schistosoma*. Esses parasitos são considerados tremadotas atípicos, pois os seus estágios larvais de cercária e miracídio, assim como o estágio adulto, apresentam

dimorfismo sexual. Além disso, o seu ciclo de vida é considerado heteroxênico, ou seja, alterna um estágio em um hospedeiro invertebrado (fase assexuada) e em um hospedeiro vertebrado (fase sexual) (NEVES, 2012; REY, 2001). Dentre as espécies da família Schistosomatidae, os hospedeiros definitivos podem ser mamíferos, aves, e, ocasionalmente répteis; os hospedeiros intermediários geralmente são caramujos de ambientes aquáticos (KOLAROVA, 2007). Entretanto, para a espécie *S. mansoni*, os hospedeiros definitivos podem ser algumas espécies de mamíferos, incluindo o homem. Os hospedeiros intermediários correspondem a 37 espécies de caramujos do gênero *Biomphalaria*, cuja as principais espécies encontradas no Brasil são *Biomphalaria glabrata*, *Biomphalaria straminea* e *Biomphalaria tenagophila* (DOS SANTOS CARVALHO; JANNOTTI-PASSOS; CALDEIRA, 2008; NEVES, 2012; TEODORO et al., 2010).

O ciclo de vida do *S. mansoni* inicia-se quando ocorre a liberação das fezes contaminadas pelo hospedeiro definitivo (Figura 4). Nas fezes estão presentes os ovos do parasito e, quando esses atingem a água, estímulos do ambiente, como a temperatura, luz e pH, fazem com que esses ovos eclodam e liberem a forma larval infectante, o miracídio (NEVES, 2012). O miracídio consegue nadar ativamente no ambiente aquático até ser atraído ao seu hospedeiro intermediário. Além de apresentar uma superfície corporal ciliada, possui glândulas adesivas e um conjunto de espículas necessários para sua penetração na homocele do caramujo. Ao atingir o tegumento do caramujo, os miracídios diferenciam-se em um aglomerado de células germinativas denominadas esporocistos primários. Essa é a etapa onde ocorre a reprodução assexuada em que os esporocistos primários são submetidos a processos de multiplicação, dobrando o seu tamanho. Dentro do esporocisto primário, numerosas células germinativas entram em multiplicação ocorrendo a formação de esporocistos secundários que se inicia a partir do 14º dia após a penetração dos miracídios. A última geração de células embrionárias origina-se dos esporocistos anteriores sendo chamados de esporocistos terciários. Finalmente existe um hipótese que os esporocistos podem se diferenciar em novas gerações de esporocistos ou na terceira geração de larvas, denominadas de cercárias (NEVES, 2012; WILSON, R. A., 1979).

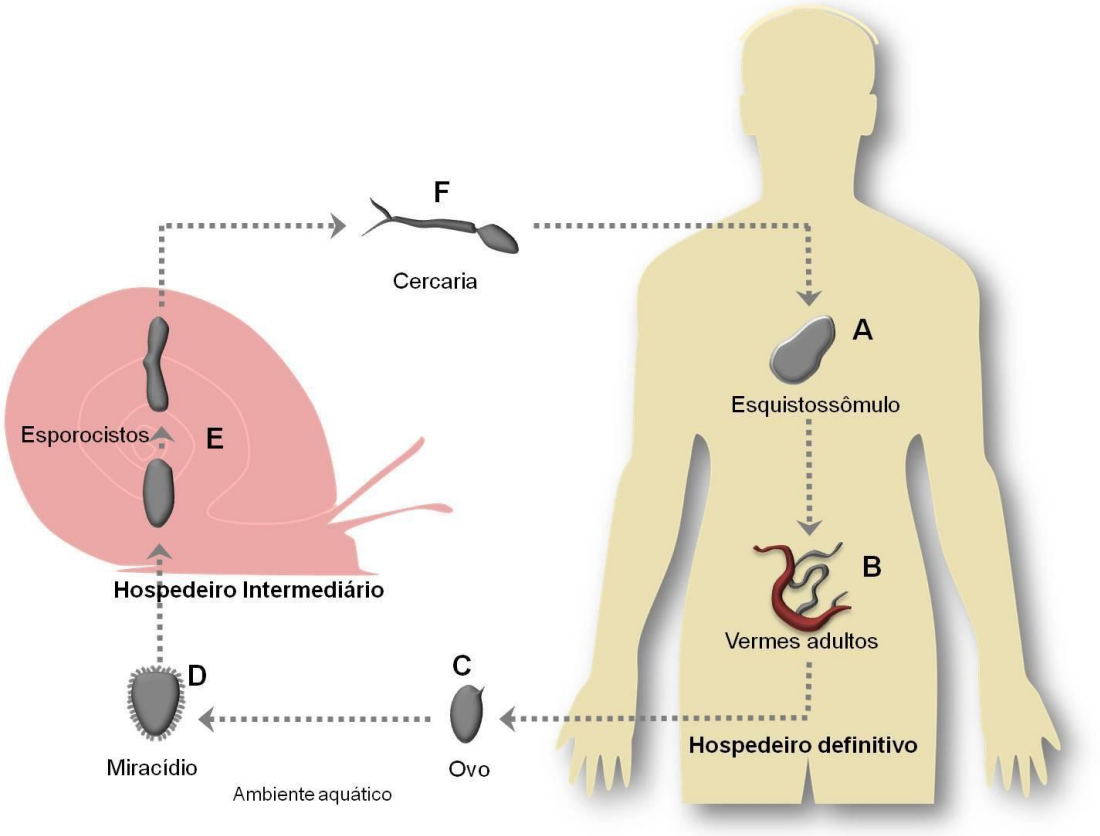


Figura 4: Ciclo de vida do *S. mansoni*. No hospedeiro definitivo, os esquistossômulos (A) dão origem aos vermes adultos (B), os quais se acasalam e produzem ovos (C) que são liberados no ambiente aquático. Os ovos maduros eclodem liberando os miracídios (D), que penetram nos caramujos, originando os esporocistos (E). Os esporocistos se diferenciam em cercárias (F) que são liberadas em água, infectando o hospedeiro vertebrado e fechando o ciclo (Imagem: Marcela Pereira Costa).

Uma vez formadas, as cercárias, devido a estímulos externos, como temperatura e luz, abandonam o hospedeiro invertebrado (LUTZ, 1919). Caramujos infectados podem liberar cercárias por toda sua vida, ou seja, por um ano. No ambiente aquático, as cercárias sobrevivem até 48 horas, nadando ativamente através da movimentação da cauda bifurcada até encontrar um hospedeiro definitivo. Elas geralmente permanecem agrupadas em águas rasas onde podem penetrar em vários vertebrados, mas se desenvolvem somente nos seus hospedeiros definitivos corretos. Ao entrar em contato com o vertebrado, a cercária contrai-se, fixa-se fortemente em seu tegumento e adota na pele, adotando uma posição paralela à mesma. Geralmente essa fixação é feita entre os folículos pilosos com o auxílio de duas ventosas e de uma substância mucoproteica

secretada por suas glândulas acetabulares. Nesse processo de penetração, a cercaria acaba por perder sua cauda. (HAAS et al., 2002; MCKERROW; SALTER, 2002).

Poucas horas após a infecção, o corpo cercariano começa a sofrer uma série de modificações bioquímicas e morfológicas em adaptação às condições fisiológicas do hospedeiro, culminando em sua transformação no estágio de esquistossômulos. Por volta de 24 a 72 horas, os esquistossômulos atravessam a epiderme e derme do hospedeiro e aproximadamente no quarto dia atingem ao sistema linfático e posteriormente chegando nos pulmões. Ao longo do trajeto, os esquistossômulos sofrem um grande alongamento do seu corpo sem sofrer divisões celulares. A partir do oitavo dia, os parasitos migram para o sistema porta-hepático onde se alimentam do sangue do hospedeiro e iniciam a divisão celular. A fase adulta de macho e fêmea é atingida por volta de 25-28 dias após a penetração (CLEGG, 1965). Ao se acasalarem, inicia-se o processo de ovoposição realizado pelas fêmeas no qual podem ser liberados cerca de 400 ovos por dia. Cerca de 50% dos ovos são eliminados nas fezes atingindo o ambiente externo e fechando o ciclo, sendo o restante permanecendo nas paredes dos capilares e vênulas (NEVES, 2012; WILSON, R. A., 1979).

Vários fatores relacionam a patologia da esquistossomose, como por exemplo: a cepa do parasito, carga parasitária, idade, estado nutricional e a resposta imunitária do parasito. Dentre esses fatores, os considerados mais importantes são a resposta do sistema imune do indivíduo infectado e a carga parasitária devido a correlação direta entre eles e a sintomatologia. Dentre todos os estágios do ciclo, o principal responsável pela patogenia são os ovos. Sendo assim, quanto maior a carga parasitária, maior será a quantidade de ovos liberadas e consequentemente os danos ao indivíduo. Os ovos quando não são liberados nas fezes, podem ficar retidos nas paredes de órgãos, como o fígado e baço provocando uma reação inflamatória granulomatosa. Essas lesões são as principais responsáveis pelas complicações clinicas observadas. Esses granulomas podem desenvolver em 3 fases denominadas: (I) fase necrótica, (II) fase produtiva ou (III) reação histiocitária e fase de cura ou fibrótica. Em casos mais graves do desenvolvimento desses granulomas, ocorre a formação da hepatoesplenomegalia e da ascite podendo levar o paciente ao óbito (BOROS, 1989; NEVES, 2012).

1.3 Genoma e transcrissoma do *S. mansoni*

Os estudos genômicos do *S. mansoni* iniciaram em 1982 com experimentos de extrações de DNA pelo pesquisador Andrew Simpson. Nesses trabalhos, ele estimou que o parasito possuísse um genoma de aproximadamente 270 megabases (Mb) agrupado em 8 pares de cromossomos, sendo divididos em sete autossômicos e um sexual. Baseando-se nesse tamanho e na classificação evolutiva, ele ainda estimou que o *S. mansoni* possuísse cerca de 15 a 20 mil genes (ALI et al., 1991; SIMPSON; SHER; MCCUTCHAN, 1982). O parasito macho foi identificado como homogamético e denominado ZZ, enquanto a fêmea, como heterogamético e denominada ZW (SHORT; MENZEL; PATHAK, 1979).

Muitas iniciativas foram realizadas pelo Programa de Pesquisas em Doenças Tropicais da OMS as quais levaram a criação de projetos como o Projeto Genoma do *S. mansoni*. O primeiro projeto iniciado em 1992, foi mediado por Sergio Pena e Andrew Simpson. Esses pesquisadores identificaram e caracterizaram novos genes do parasito, entre outras características do genoma. Até o ano de 2000, cerca de 17.000 *Expressed Sequence Tags* (ESTs) foram geradas pelas bibliotecas de DNA complementar (cDNA) de diferentes estágios do parasito (FRANCO et al., 1995; FRANCO et al., 1997).

No ano de 2001, programas de sequenciamento de transcrissoma foram iniciados em larga escala gerando mais 180.000 ORESTES (*Open Reading Frames ESTs*) e ESTs dos estágios de vermes adultos, ovos, miracídios, esporocistos cercárias e esquistossômulos, correspondendo cerca de 92% do genoma expresso. Esses dados foram ainda agrupados em 30.000 contigs (conjuntos não redundantes de sequências expressas) com uma estimativa que o genoma tivesse cerca de 14.000 genes. Financiado pela FAPESP, os dados finais desses projetos foram publicados no final de 2003. (VERJOVSKI-ALMEIDA et al., 2003)

No ano de 2009, uma parceria entre os Wellcome Trust Sanger Institute e The Institute for Genomic Research (TIGR) e outros laboratórios no mundo, finalizaram a primeira versão do genoma do *S. mansoni*, sendo disponibilizada no banco de dados do GeneDB (http://www.genedb.org) (BERRIMAN et al., 2009).

Atualmente o genoma do *S. mansoni* não está devidamente fechado devido as suas características intrínsecas, como: ser composto de 40% de elementos repetitivos fragmentados em *scaffolds*; apresenta regiões intrônicas grandes; presença de *Micro Exons Genes* (MEGs) que aumentam a variabilidade dos seus transcritos por splicing alternativo, entre outras características (BERRIMAN et al., 2009; PROTASIO et al., 2012). Segundo a última anotação publicada em 2012, o genoma do *S. mansoni* ainda é considerado altamente fragmentado, pois apresenta 380 Mb com 885 *scaffolds* (PROTASIO et al., 2012). Apesar disso, cerca de 81% de suas bases foram organizadas nos 8 cromossomos presentes no parasito. Mais de 45% dos genes preditos foram modificados resultando na redução de seu número total de 11.807 para 10.852. Nessa notação cerca de 500 genes foram considerados novos e mais de 1600 foram removidos por apresentarem baixa acurácia ou predição incorreta. Dentre os motivos para essas modificações, os pesquisadores verificaram eventos de trans-splicing em pelo menos 11% dos genes preditos e possíveis erros de sequenciamento nas montagens anteriores (PROTASIO et al., 2012).

Os dados recentes de sequenciamentos de nova geração, tanto do genoma como do transcrissoma do *S. mansoni*, contribuíram para diversos estudos acerca da interação do parasito com o hospedeiro, a interação macho e fêmea, dentre outros aspectos da biologia do parasito. O sequenciamento de RNA (RNA-seq), por exemplo, é uma metodologia que possibilitou a descoberta sistemática de unidades de transcrição. No caso de um genoma tão complexo como o do *S. mansoni*, essa metodologia permitiu desvendar o transcrissoma nos diversos estágios ou tecidos específicos do parasito (PROTASIO et al., 2012; WILSON, R. A. et al., 2015). Além disso, a anotação de novos genes descritos podem ainda ser utilizados para a identificação de alvos para novas drogas ou em novos métodos diagnósticos para a doença (OLIVEIRA, G.; FRANCO; VERJOVSKI-ALMEIDA, 2008; VERJOVSKI-ALMEIDA et al., 2003).

1.4 Regulação da expressão gênica em *S. mansoni*

Diversos estudos envolvendo o transcrissoma das principais espécies do gênero *Schistosoma* tem demostrado que esses parasitos apresentam uma regulação da expressão gênica bastante complexa (ALMEIDA et al., 2012; LIU et al., 2006; PARKER-MANUEL et al., 2011;

SUN, J. et al., 2014; VERJOVSKI-ALMEIDA et al., 2003; WANG, X. et al., 2015). Essa complexidade se deve ao fato do seu ciclo biológico ser bastante peculiar, onde o parasito sofre diversas mudanças morfológicas e bioquímicas decorrente de sua passagem por diferentes ambientes. Durante a evolução, esses parasitos desenvolveram adaptações que permitiram a sobrevivência dentro de hospedeiros, como mamíferos e moluscos, assim como no ambiente aquático. Acredita-se que todas essas adaptações estejam relacionadas com mecanismos de regulação dos seus genes nos diversos estágios de seu complexo ciclo de vida que podem ocorrer em vários níveis: transcricional, pós-transcricional, traducional e pós-traducional (HAN et al., 2009).

Várias técnicas moleculares são utilizadas para entender os mecanismos envolvidos no processo da transcrição das células, como, por exemplo, as metodologias de microarranjos de DNA (HINTON et al., 2004). Apesar dessa técnica ter sido considerada uma das mais eficientes para a determinação do padrão de expressão gênica em nível celular, algumas limitações acompanham as demandas atuais de análises, como a especificidade do arranjo para cada isolado, densidade e qualidade variáveis dos *spots* analisados, entre outros (BLOOM et al., 2009). Com o surgimento das técnicas de sequenciamento de nova geração, novos protocolos e metodologias especificas para a análise do transcrissoma passaram a ser utilizadas, destacando-se dentre elas a tecnologia do RNA-seq (CHU; COREY, 2012; NAGALAKSHMI; WAERN; SNYDER, 2010). Essa técnica permite análises do transcrissoma de genomas com uma resolução muito maior do que aquela disponível no sequenciamento de Sanger e os métodos baseados em microarranjos. No método RNA-Seq, os DNAs complementares (cDNAs), gerados a partir do RNA de interesse, são sequenciados utilizando tecnologias de terceira geração. As leituras obtidas podem então ser alinhadas a partir de um genoma de referência ou, de maneira *ab initio*, construídas a partir de um mapa de todos os transcritos do genoma. RNA-Seq tem sido usado com sucesso para quantificar com precisão os níveis de transcrição, confirmar ou revisar previamente anotados 5 'e 3' de genes, e identificação de exons e íntros (NAGALAKSHMI et al., 2010).

A utilização de análises de bioinformática em conjunto com a técnica de RNA-seq elucidaram vários mecanismos envolvidos na evolução, desenvolvimento, metabolismo, interação parasito-hospedeiro e evasão do sistema imune, por exemplo (ALMEIDA et al., 2012). Estudos utilizando-as evidenciaram padrões de expressão diferenciais em diversos genes durante todas as

fases no ciclo de vida do parasito (Figura 5) (ALMEIDA et al., 2012; HAN et al., 2009; PARKER-MANUEL et al., 2011; SUN, J. et al., 2014).

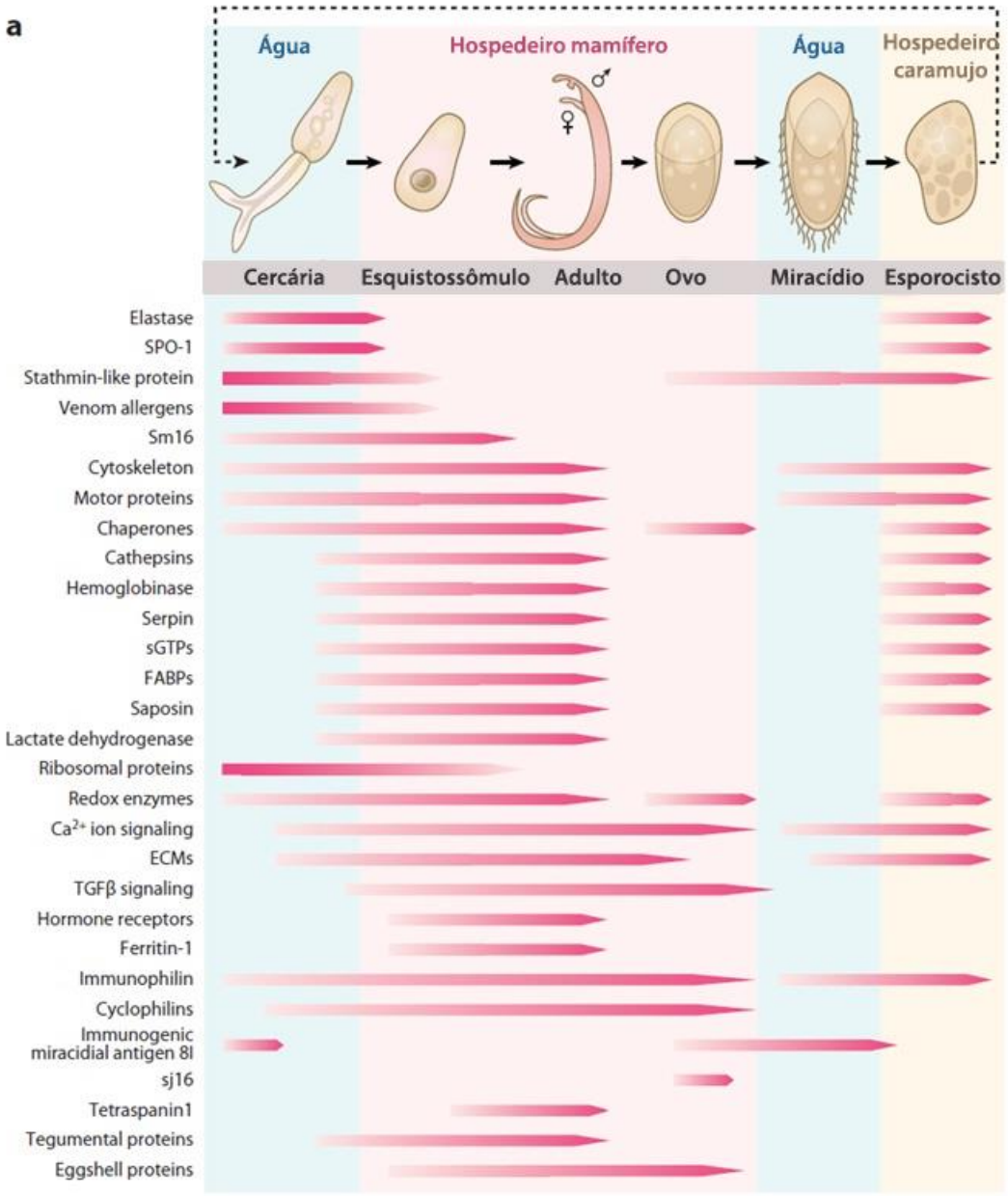


Figura 5: Expressão gênica em *S. mansoni.* Expressão gênica diferenciada entre os estágios baseado nos projetos transcrissoma do parasito. Adaptado de HAN et al (2009).

Nesses estudos foram identificadas várias proteínas envolvidas na resposta ao estresse, invasão da pele do hospedeiro, formação da membrana do parasito, degradação de hemoglobina e

na evasão do sistema imune durante a transição de cercária até o estágio de verme adulto. A maioria delas são consideradas estágio-específicas, ou seja, são expressas de acordo com a necessidade do parasito, como, por exemplo, a elastase, uma serina-protease, que é responsável pela digestão da elastina presente na cauda das cercárias para sua locomoção, proteínas da casca do ovo sintetizadas pelos vermes adultos para sua formação e proteínas do tegumento no processo de transição de esquistossômulos até vermes adultos. Essas adaptações da regulação gênica permitem um ótimo desenvolvimento do parasito no hospedeiro até a sua reprodução completando o ciclo (HAN et al., 2009).

Dentre as diversas moléculas responsáveis pela regulação gênica no *S. mansoni*, as mais conhecidas e estudadas atuam no nível pós-transcricional. Nesse sentido, destacam-se os RNAs não codificantes, um grupo de moléculas que não codificam proteínas (BLIGNAUT, 2012). São divididos em duas classes: (i) os RNAs constitutivos, compostos pelos RNAs ribossomais e pequenos RNAs nucleares e (ii) RNAs reguladores, como os microRNAs (miRNAs), RNAs circulares, *Piwi-interacting* RNAs (piRNAs), RNAs longos não codificantes (lncRNAs) e outros pequenos RNAs de interferência (CECH; STEITZ, 2014).

1.5 RNAs longos não codificantes

Os lncRNAs são um grupo de moléculas de RNA transcritos que não codificam proteínas. Distinguem-se dos outros RNAs não codificantes pela característica principal de apresentar mais de 200 nucleotídeos de tamanho. Na maioria dos casos, eles não apresentam *Open Reading Frames* (ORFs), sequências que poderiam ser traduzidas em proteínas, tendem a ser mais curtos que os RNAs mensageiros (mRNAs), apresentam éxons longos e em menor número que os mRNAs, níveis baixos de expressão e baixa conservação filogenética (CABILI et al., 2011; DERRIEN et al., 2012).

A classificação dos RNAs ‘longos’ como uma classe de moléculas é bastante controversa e complexa, pois é uma construção arbitrária proveniente de *cutoffs* de tamanho que os separam dos outros RNAs não codificantes menores, como os microRNAs (CAO, 2014; ST LAURENT; WAHLESTEDT; KAPRANOV, 2015). Além disso, já existem relatos de que algumas dessas

moléculas identificadas e caracterizadas possam apresentar ORFs funcionais em sua estrutura que podem ser traduzidas em pequenos peptídeos (BAZZINI et al., 2014; GASCOIGNE et al., 2012; JUNTAWONG et al., 2014).

Apesar da biogênese da maioria dos lncRNAs ainda permanecer desconhecida, sabe-se que eles são tipicamente sintetizados pela mesma maquinaria transcricional dos mRNAs, sendo esse fato evidenciado por análises de interações da RNA polimerase II e modificações de histonas associadas com a sua transcrição e alongamento.

Eles são classificados de acordo com a sua proximidade com os genes que codificam proteínas no genoma. Dessa forma, são geralmente divididos em cinco categorias: *sense*, *anti-sense*, bidirecional, intergênico e intrônico. Alguns pesquisadores hipotetizam que algumas dessas moléculas seriam na realidade mRNAs ainda não processados pela maquinaria de *splicing* devido a presença de éxons nelas (MERCER et al., 2008). Uma outra hipótese sugere que algumas sequências de lncRNAs seriam precursores de RNAs não codificantes (ncRNAs) pequenos que apresentam funções regulatórias definidas, assim como os miRNAs. No entanto, outras características da estrutura dos lncRNAs revelam diferenças substanciais dentre essas duas classes de moléculas, os lncRNAs e os mRNAs. Geralmente, os lncRNAs, em relação aos mRNAs, são menores, poucos apresentam isoformas no *splicing* alternativo, apresentam um conteúdo de GC muito menor e consequentemente uma menor probabilidade de formar estruturas secundárias estáveis, as suas ORFs quando presentes são menores, códons de iniciação inexistentes ou ineficientes para transcrição, entre outras características (BLIGNAUT, 2012; WANG, C. et al., 2017; YANG; ZHANG, 2015).

Por muito tempo, os lncRNAs foram considerados não funcionais e atualmente a sua presença e importância vem sendo bastante debatida (VAN BAKEL; HUGHES, 2009; VAN BAKEL et al., 2010). Com as recentes tecnologias de sequenciamento, estudos estimaram que mais de 70-90% do genoma dos mamíferos são transcritos e, que apenas 3% do DNA nuclear, seria traduzido em proteínas (CARNINCI et al., 2005; DJEBALI et al., 2012). Dentre esses transcritos, várias classes de RNAs pequenos e longos foram estabelecidas, sendo que para humanos cerca de 15.900 lncRNAs já foram identificados (KASHI et al., 2016) (Figura 6). Muitos desses lncRNAs de humanos e de outras espécies já estão depositados nos principais bancos de dados específicos para essas sequências como o lncRNAdb, NONCODE, LNCipedia, fRNAdb e NRED.

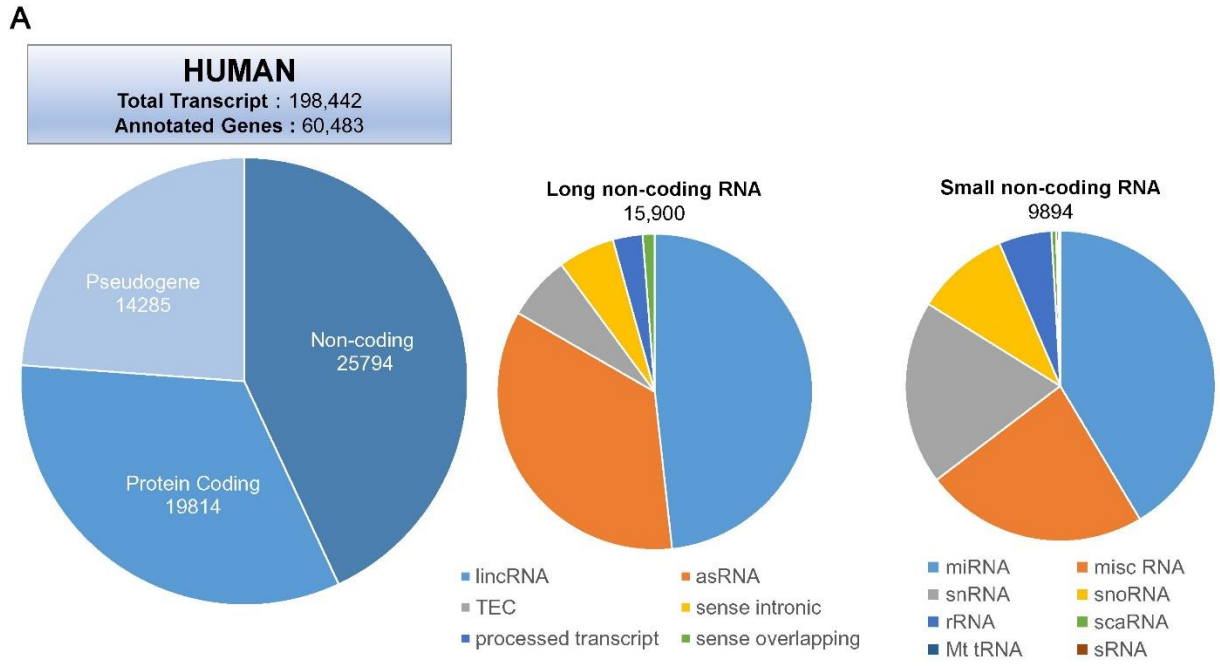


Figura 6: Anotação atual do genoma humano. A última anotação do genoma humano no GENCODE, versão 22, classificam os principais RNAs longos e os pequenos RNAs representando 25.794 transcritos. Adaptado de KASHI et al (2016).

Os lncRNAs desempenham funções críticas na regulação de diversos processos celulares, como na regulação da transcrição, no processamento pós-transcricional, no remodelamento da cromatina, no controle epigenético da cromatina entre outras funções moleculares e celulares (Figura 7) (CHEN; CARMICHAEL, 2010; PONTING; OLIVER; REIK, 2009; WILUSZ; SUNWOO; SPECTOR, 2009). Além disso, vários estudos descreveram o envolvimento direto dessas moléculas em diversas patologias. Neoplasias, doenças neurodegenerativas, cardiovasculares e parasitoses são alguns de seus exemplos (SUN, M.; KRAUS, 2015; WAPINSKI; CHANG, 2011).

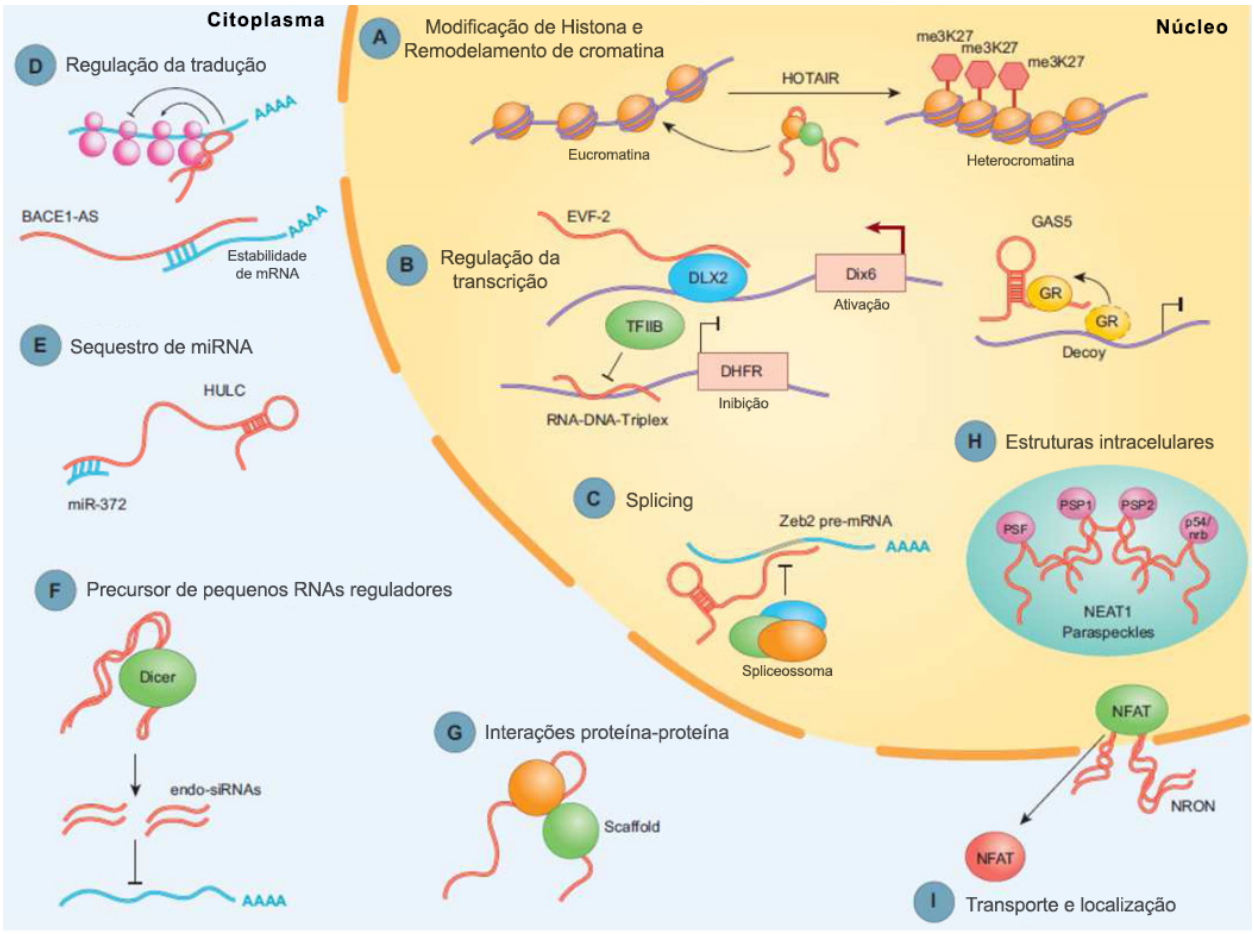


Figura 7: Funções dos lncRNAs. Os lncRNAs podem apresentam diversas funções como: (A) Modificação de histona e o remodelamento da cromatina, (B) Regulação da transcrição, (C) Splicing, (D) Regulação da tradução, (E) Sequestro de miRNAs, (F) Precursor de pequenos RNAs reguladores, (G) Interações proteína-proteína (H) Estruturas intracelulares, (I) Transporte e localização. Adaptado de (KARLSSON; BACCARELLI, 2016).

Os mecanismos moleculares precisos do funcionamento dos lncRNAs ainda permanecem desconhecidos, pois apresentam diferentes sequências, domínios e estruturas que podem interagir de forma integrada no tempo e no espaço. Algumas hipóteses sobre esses processos incluem interações de um tríplex RNA:DNA:DNA, híbridos de RNA:DNA, híbridos de RNA:RNA e interações entre RNA-Proteínas (RINN; CHANG, 2012). Essas interações podem ocorrer de maneira que os lncRNAs respondam a diversos sinais moleculares. Rinn e Chang (2012) verificaram que os lncRNAs *Air, Xist* e *COOLAIR* inativam ou silenciam os mRNAs. Podem ainda se ligar a *enhancers* e regular a síntese de mRNAs, como o *HOTAIR* e o *HOTTIP* (WANG, K. C.; CHANG, 2011), formar estruturas em *scaffold* funcionando como adaptadores de complexos

proteicos e atuar como reguladores *Cis* e *Trans* dependendo da região que ele exerce a função (LI, Z.; RANA, 2012).

Recentemente foram publicados estudos que relataram a expressão de muitos lncRNAs no sistema imune e sua participação em funções importantes do desenvolvimento das células imunológicas (SATPATHY; CHANG, 2015). Guttman e colaboradores foram os responsáveis pela primeira identificação da participação dos lncRNAs na regulação da resposta imune inata em 2009 (GUTTMAN et al., 2009). Nele, foi demonstrado que o lincRNA-COX2 regulou negativamente a expressão de grupos distintos de genes inflamatórios em camundongo, além de ter a sua expressão induzida por receptores do tipo Toll (TLR: *Tool-like receptors*) (CARPENTER et al., 2013; GUTTMAN et al., 2009). Desde então, vários outros lncRNAs foram identificados e relacionados com o controle da expressão de genes imunológicos por experimentos de *microarray* e RNA-seq, como Lethe, PACER, THRIL e NEAT1 (ZHANG, Y.; CAO, 2015).

Esses estudos relataram ainda que diferentes tipos de patógenos, incluindo vírus, bactérias e parasitos, induzem lncRNAs funcionais a controlar e regular infecções em hospedeiros (SCARIA; PASHA, 2012; TYCOWSKI et al., 2015). Muitos desses lncRNAs auxiliam hospedeiros no processo de resistência a algumas infecções virais, enquanto outros lncRNAs, favorecem a sobrevivência e a invasão dos patógenos no hospedeiro. O lncRNA-TARE é um exemplo de um lncRNA que é expresso pelo próprio parasito *Plasmodium falciparum* após a sua replicação. Ele desempenha um papel importante na manutenção dos telômeros do *P. falciparum* e na regulação de genes de virulência, entre outros processos no cromossomo do parasita (BROADBENT et al., 2011). Outros lncRNAs também já foram identificados regulando a virulência desse parasito presentes em íntrons na fita *antisense* dos genes da família *var*. Uma vez expressos, esses lncRNAs são integrados na cromatina e promovem a regulação da ativação de genes *var* que foram eventualmente silenciados (AMIT-AVRAHAM et al., 2015).

As descobertas recentes dessas funções dos lncRNAs abrem novas perspectivas de estudo para interações entre patógeno-hospedeiro beneficiando não só o entendimento dos mecanismos dos lncRNAs, mas também a detecção de marcadores para o diagnóstico, tratamento, prognóstico e prevenção de doenças.

1.6 Métodos computacionais para identificação de lncRNAs

Nos últimos anos, os estudos computacionais integrados com as metodologias de sequenciamento de nova geração vêm se intensificando e gerando dados do genoma e transcrissoma em larga escala. Entretanto, o repertório transcricional completo de um organismo não é predito apenas pelas sequências genômicas, mas sim por protocolos que sequenciam, analisam e montam transcritos que são expressos, como na metodologia do RNA-seq. Em teoria, o RNA-seq pode ser utilizado para construir mapas completos dos transcrissomas em todos os organismos e tipos celulares (WANG, Z.; GERSTEIN; SNYDER, 2009). Nesse repertório estão presentes sequências que são transcritas e não são traduzidas, como, por exemplo, os lncRNAs.

Para analisar dados provenientes de RNA-seq, são utilizados programas divididos basicamente em 3 grandes grupos. O primeiro grupo aborda metodologias para realizar o alinhamento utilizando como referência algum genoma ou transcrissoma (mapeamento de *reads*). O segundo aborda metodologias para identificar transcritos, genes e suas isoformas expressas (Reconstrução do transcrissoma). O terceiro e último grupo apresenta métodos para a estimativa dos genes e abundância das isoformas, além de métodos para a análise da expressão diferencial entre amostras (Quantificação da expressão) (ILOTT; PONTING, 2013). Devido as melhorias contínuas na quantidade e qualidade da geração de dados no RNA-Seq, há uma grande variabilidade no nível de precisão e eficiência de ferramentas computacionais disponíveis. A maior parte dos genomas apresentarem características intrínsecas em cada espécie o que acarreta um constante crescimento do desenvolvimento de algoritmos que realizam essas análises, cujo objetivo é personalizar cada vez mais os resultados.

Após a realização do mapeamento e da montagem de um transcrissoma, utilizando ou não um genoma de referência, a identificação de lncRNAs segue a partir de critérios para filtragem e caracterização computacional dessas sequências baseando principalmente em seus potenciais codificantes e características específicas de cada genoma (ILOTT; PONTING, 2013).

1.6.1 Análise do tamanho dos transcritos

Para a caracterização dos lncRNAs, o primeiro critério que deve ser seguido é a análise do tamanho dos transcritos. O *cuttoff* principal estabelecido nos protocolos para um transcrito não codificante ser considerado longo é apresentar uma sequência com tamanho maior ou igual a 200 pb (SUN, K. et al., 2014; SUN, L. et al., 2012; ZHANG, Y. C. et al., 2014). Esse filtro elimina possíveis classes de transcritos que também não são codificantes, como os miRNAs, piRNAs, que poderiam estar presentes no resultado final.

1.6.2 Análise do tamanho das ORFs

O segundo critério para identificar um lncRNA é avaliar as ORFs que podem estar presentes nos transcritos. Diferentemente dos genes codificantes de proteínas, os códons de iniciação e de término tendem a estar distribuídos randomicamente nos lncRNAs. Como resultado, o tamanho das suas ORFs são menores e raramente superam a 100 códons. Baseado nesse princípio, outra maneira de discriminar os lncRNAs é utilizar um *cutoff* que considera transcritos com ausência de ORFs ou a presença delas com no máximo 100 códons (PAULI et al., 2012). Entretanto, já foram identificados que alguns lncRNAs podem apresentar ORFs completas ou maiores que 100 códons (CLAMP et al., 2007), dificultando a diferenciação entre os lncRNAs e os mRNAs.

1.6.3 Homologia com proteínas conhecidas

É impossível saber precisamente se uma ORF predita em um lncRNA será transcrita, e, muito menos, se a proteína ou peptídeo gerado será funcional sem realizar ensaios de homologia com bancos de proteínas preditas. Para isso, a terceira etapa utilizada na maioria dos *pipelines* para as análises das ORFs consiste em alinhar essas ORFs com bancos, como Pfam ou SwissProt, ou

com programa específicos, removendo assim os transcritos que apresentam proteínas preditas e funcionais (ILOTT; PONTING, 2013; ZHANG, Y. C. et al., 2014).

1.6.4 Normalização e quantificação da expressão

Os dados finais obtidos em uma montagem de transcrissoma precisam ser convertidos em uma medida quantitativa da expressão do gene. Cada sequenciamento pode ter diferentes variantes que irão influenciar o número de fragmentos mapeados entre amostras. Nesse processo, é necessário a normalização dos dados de modo a evitar que os genes que parecem ser diferencialmente expressos apenas como um resultado da presença de mais sequências em um quadro, quando comparados a outro (MARIONI et al., 2008; ROBINSON, M. D.; OSHLACK, 2010). A abordagem mais utilizada para normalizar os dados de RNA-Seq é calcular valores com as medidas métricas *Reads Per Kilobase of Transcript per Million Mapped* (RPKM) e *Fragments Per Kilobase of Transcript per Million Mapped* (FPKM) (MORTAZAVI et al., 2008).

2 JUSTIFICATIVA, RELEVÂNCIA E HIPÓTESE

Desde o ano de 2007, o LBBM vem realizando estudos de predição e caracterização de ncRNAs em *S. mansoni*, sobretudo os miRNAs, o principal grupo estudado. Inicialmente demonstramos que as proteínas envolvidas na biogênese dos miRNAs em *S. mansoni* são conservadas e diferencialmente expressas e que seus respectivos genes são em média 3 vezes mais transcritos em cercárias, esquistossômulos com 3,5h e 24h de cultivo *in vitro* e em miracídios, quando comparados a vermes adultos ou os estágios de machos e fêmeas isolados. Esses dados corroboraram com nossa hipótese inicial de que durante as fases larvais, a regulação da expressão gênica ao nível pós-transcricional permitiria uma rápida adaptação do parasito aos diferentes ambientes e hospedeiros. Estudos posteriores identificaram 105 miRNAs, sendo 35 conservados e 70 não conservados. Desse conjunto de miRNAs conservados, observamos uma expressiva conservação entre os eucariotos, incluindo camundongos e humanos (DE SOUZA GOMES et al., 2011). Esses microRNAs também apresentam um padrão de expressão diferencial entre as fases larvais e adultas, reforçando a hipótese de que mecanismos pos-transcricionais tem um papel chave nos mecanismos de adaptação do parasito a seus hospedeiros. Concomitantemente a essas análises, também ficou claro que íntrons podem gerar miRNAs em *S. mansoni*, como, por exemplo, o miR190, reforçando que vias não canônicas de biogênese de miRNA também são funcionais em *S. mansoni*.

Outros grupos de pesquisa também vem contribuindo para identificar e analisar as funções dos miRNAs que ainda permanecem desconhecidas no gênero *Schistosoma* (COPELAND et al., 2009; OLIVEIRA, K. C. et al., 2011; ZHU, L.; LIU; CHENG, 2014). Analisando todos os dados publicados em *S. mansoni* sobre ncRNA disponíveis até o momento, parece-nos claro que os miRNAs não explicam *per se* um papel epigenético direto, como o descrito para muitos organismos. Além desses estudos, estão sendo realizadas predições do repertório de outros tipos de ncRNAs que permanecem poucos explorados ou desconhecidos nesse gênero (CAI et al., 2013; NOWACKI et al., 2015).

Os lncRNAs representam um grupo de ncRNAs que vem sendo extensivamente estudados devido a sua grande importância na regulação de genomas e transcrissomas das espécies. Entretanto, esse grupo ainda foi muito pouco estudado no gênero *Schistosoma*, contando apenas

com alguns trabalhos que mensuram a sua existência por experimentos de RNA-seq associados a análises de bioinformática (ALMEIDA et al., 2012; CAI et al., 2013; COPELAND et al., 2009). Eles podem desempenhar um papel essencial, uma vez que estão envolvidos em vários mecanismos regulatórios, ou seja, epigenético, transcricional, pós-transcricional e traducional, como brevemente exemplificado abaixo:

(i) Os lncRNAs constituem uma rede de moduladores epigenéticos capazes de recrutar, orientar e formar plataformas de complexos de ribonucleoproteínas em locais genômicos específicos. Por exemplo, 20% de lncRNAs humanos são capazes de recrutar repressores multimoleculares, ativadores ou complexos de remodelação da cromatina (PRC1 / PRC2 por ANRIL, LSD1 / PRC2 por HOTAIR) cujas unidades (EZH2, EED, BMI1, SUZ12, CBX7, Corest e JARID1) são capazes de alterar o código de histona e perfil de metilação (ZAPATA et al., 2017);

(ii) lncRNAs podem estar associados à promotores e a alguns *enhanced* RNAs (eRNAs) podem regular diretamente a transcrição de genes alvo pela ativação da transcrição;

(iii) lncRNAs também podem atuar ao nível pós-transcricional e estão amplamente envolvidas na biogênese e estabilidade de mRNAs, pois participam dos mecanismos de *splicing* alternativo, tráfico, localização celular direta, além de promoverem a degradação de mRNAs. Os lncRNAs também formam estruturas semelhantes a esponjas para evitar a ligação de miRNAs aos seus respectivos mRNAs alvo, como, por exemplo, CDR1-as / CIRS-7 e RNA circular esponja para miR-7);

(iv) lncRNAs podem ligar-se a fatores gerais de tradução e assim inibir a síntese proteica e interagir diretamente com os ribossomos;

(v) Finalmente os lncRNAs podem ser candidatos a biomarcadores e assim ter um grande potencial biotecnológico.

Nesse sentido, a hipótese central desse trabalho é que os lncRNAs representam a maior parte dos transcritos preditos em *S. mansoni* e que o repertório dos lncRNAs estágio-específicos atuem como um segundo código genético, que é responsável pela regulação da expressão gênica estágio-específica e pela adaptação rápida do parasito nos ambientes aquáticos, hospedeiros vertebrados e invertebrados.

Os resultados obtidos nesse projeto serão futuramente explorados quanto ao seu potencial biotecnológico em uma das linhas de pesquisas de nosso laboratório, *Desenvolvimento de um sistema de detecção de esquistossomose humana em fase aguda*. Essa linha está vinculada ao Programa de Incentivo à Inovação da UFOP, edição 2014, que visa o fomento da tradução de tecnologias acadêmicas em estudos capazes de gerar inovações tecnológicas de produtos e processos. De maneira geral, os diversos kits para diagnóstico da esquistossomose em estudo atualmente apresentam importantes problemas no que tange aos elevados índices de reação cruzadas com outras parasitoses, pois são fundamentados essencialmente em metodologias imunoenzimáticas. Nossa linha de pesquisa difere das demais por explorar alvos moleculares espécie específicos e, portanto, contribui com uma abordagem inovadora ao processo de diagnóstico da fase aguda da doença através do enfoque da detecção de ncRNA específicos de *S. mansoni*, miRNA e lncRNA. Esse método diagnóstico também traz o benefício do acompanhamento com relação à eficácia do tratamento realizado a partir da identificação do parasita no paciente, bem como a extensão da lesão hepática causada pelo verme. Além disso, por utilizar plasma ou soro dos indivíduos, as amostras são facilmente obtidas, tornando sua utilização menos dispendiosa e desgastante ao paciente. Por fim, o uso dessa ferramenta possibilitaria um acompanhamento mais eficaz à resposta terapêutica empregada e agilizaria o processo de diagnóstico de forma a viabilizar um tratamento mais precoce.

3 OBJETIVOS

3.1 Objetivo geral

Esse trabalho tem como objetivo geral identificar e caracterizar os lncRNAs de *S. mansoni* no estágio de verme adulto, além de investigar as interações de expressão gênica existentes entre esses transcritos.

3.2 Objetivos específicos

- Estabelecer um *pipeline* computacional para identificar os lncRNAs de *S. mansoni* a partir de um banco público de RNA-seq;
- Identificar e caracterizar um conjunto robusto de lncRNAs de *S. mansoni* no estágio de verme adulto;
- Classificar os novos lncRNAs preditos utilizando a anotação Gene Ontology atribuindo funções preditas;
- Analisar a expressão de um conjunto de lncRNAs de verme adulto nos estágios de cercária, esquistossômulos e ovos do *S. mansoni*;
- Analisar a expressão de um conjunto de lncRNAs de verme adulto pareados e não pareados em parasitos resistentes ao praziquantel;
- Analisar a expressão de um conjunto de lncRNAs de verme adulto em fígados de camundongos infectados e não infectados com *S. mansoni.*

4 MATERIAIS E MÉTODOS

4.1 Análises *in silico*

As análises *in silico* desse trabalho foram realizadas nos servidores disponíveis da Universidade Federal de Ouro Preto (UFOP) e do instituto Max Planck Institut für molekulare Genetik (MPIMG) em Berlim na Alemanha.

4.1.1 Análise inicial do *pipeline* utilizando dados de RNA-Seq

A montagem do inicial do transcrissoma foi realizada de um *pipeline* descrito a seguir desenhado especificamente para o *S. mansoni*. Nele, foram implementadas várias etapas com filtros e ferramentas que auxiliaram na predição final de lncRNAs. O primeiro segmento nesse *pipeline* é representado na Figura 8A e o segundo na Figura 8B.

4.1.1.1 Sequências do genoma de S. mansoni

A partir da última versão disponível do genoma de *S. mansoni* (Genome Assembly v5.2), foram recuperados os arquivos *Schistosoma_mansoni_v5.2.fa* e *Schistosoma_mansoni_v5.2.gff* presentes no banco de dados GeneDB (http://www.genedb.org) para a posterior utilização no *pipeline* desenvolvido. Esses dois arquivos ficaram hospedados nos servidores como base para todas as predições de lncRNAs independente da amostra de RNA-seq do parasito.

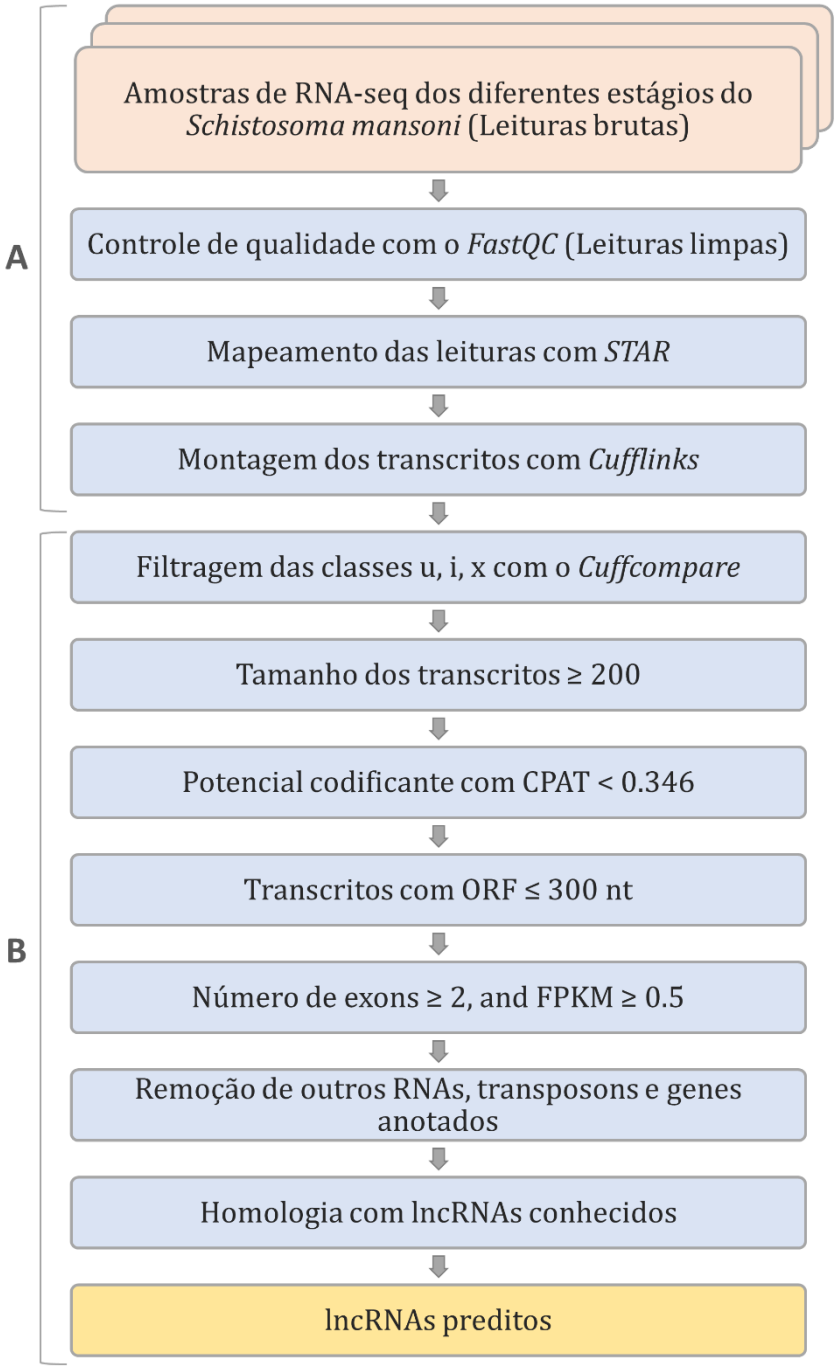


Figura 8: *Pipeline* computacional integrativo para a identificação de lncRNAs em *S. mansoni*. (A) Análise inicial do *pipeline*. Os dados RNA-seq brutos foram pré-processados, alinhados com o *STAR* e montados com *Cufflinks* no modo *ab initio*. (B) Predição de lncRNAs. Os transcritos gerados foram submetidos a várias etapas contendo filtros e predizendo um conjunto final de lncRNAs.

4.1.1.2 Sequências de RNA-seq

Para identificar os potenciais RNAs longos não codificantes no *S. mansoni*, um conjunto de amostras de sequencias de RNA-seq (PROTASIO et al., 2012) depositadas no banco ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) foram recuperadas com o número de acesso E-MTAB-451 e posteriormente analisadas para utilização. A descrição das sequências está apresentada pela Tabela 1.

Essas sequências foram obtidas pela extração do RNA total dos parasitos nos diferentes estágios e o sequenciamento utilizando o protocolo experimental da empresa Illumina para posterior processamento de imagens e de obtenção das bases (PROTASIO et al., 2012).

Tabela 1: Amostras das reads de RNA-seq de *S. mansoni*.

Número do acesso	Número experimento	Estágio	Tempo
ERR022872	ERX009279	cerc10a	4h
ERR022873	ERX009276	adult2	7 semanas
ERR022874	ERX009284	somule1	3h
ERR022875	ERX009282	tail1	0
ERR022876	ERX009281	somule1	3h
ERR022877	ERX009278	cerc12	4h
ERR022878	ERX009277	cerc13	4h
ERR022879	ERX009274	somule2	3h
ERR022880	ERX009285	somule3	24h
ERR022881	ERX009280	somule4	24h
ERR022882	ERX009283	somule5	24h
ERR022883	ERX009275	somule6	24h

Legenda: *Cerc* (Cercária), *adult* (Verme adulto), *somule* (Esquistossômulo) e *tail* (Cauda).

Para as seguintes análises no *pipeline* estabelecido nesse trabalho, somente 10 pares de amostras de RNA-seq foram escolhidas para serem utilizadas (Tabela 2).

Tabela 2: Amostras de RNA-seq escolhidas para a análise.

Cercária	Esquistossômulos 3h	Esquistossômulos 24h	Verme Adulto
ERR022872.1	ERR022874.1	ERR022880.1	ERR022873.1
ERR022872.2	ERR022874.2	ERR022880.2	ERR022873.2
ERR022877.1	ERR022876.1	ERR022881.1	-
ERR022877.2	ERR022876.2	ERR022881.2	-
ERR022878.1	ERR022879.1	ERR022882.1	-
ERR022878.2	ERR022879.2	ERR022882.2	-

4.1.1.3 Processamento e controle de qualidade das amostras

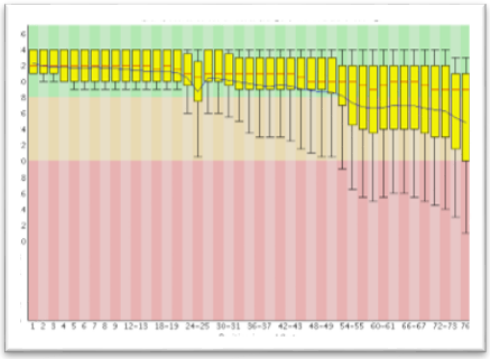
As análises da qualidade das amostras de RNA-seq foram inicialmente realizadas pelo programa FastQC na versão 0.11.4 (ANDREWS, 2010). A sua programação é baseada na linguagem Java e está disponível gratuitamente sob o GPLv3. Essa ferramenta cria um relatório abrangente na qual avalia a composição e qualidade da sequência analisada. Ela pode operar tanto como um aplicativo autônomo interativo ou em um modo não-interativo que é ideal para integrar e adequar a um *pipeline* de sequenciamento.

Apesar do FastQC analisar 12 parâmetros, foram priorizados os seguintes parâmetros para todas as amostras: estatísticas básicas, qualidade da sequência por base, conteúdo de GC por sequência, distribuição do tamanho da sequência, níveis de duplicação da sequência e sequencias adaptadoras (Figura 9). A verificação desses parâmetros é essencial para manter a qualidade posterior dessas sequencias que serão utilizadas posteriormente.

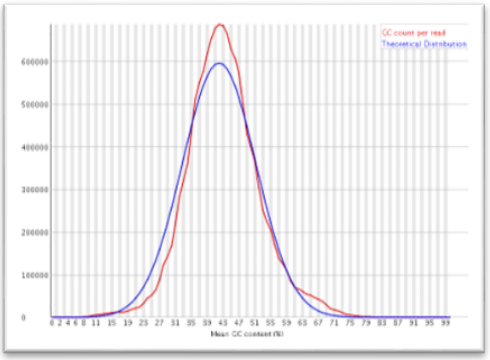
Estatísticas básicas

Measure	Value
Filename	ERR022873_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	10521255
Sequences flagged as poor quality	0
Sequence length	76
%GC	43

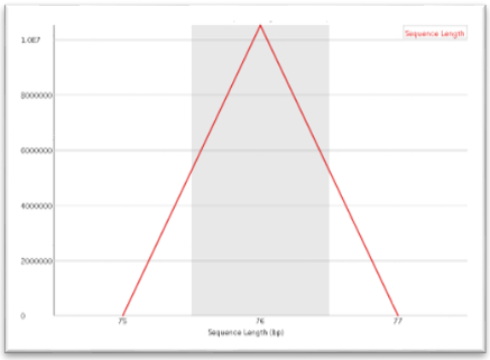
Qualidade da sequência por base



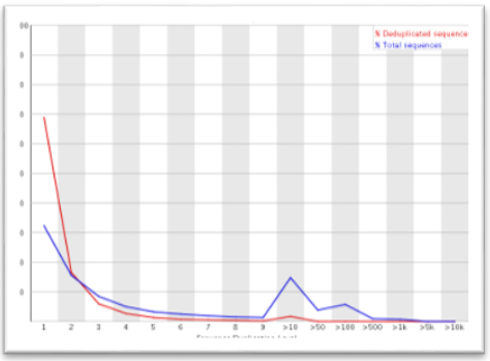
Conteúdo de GC por sequência



Distribuição do tamanho da sequência



Níveis de Duplicação de Sequência



Sequências adaptadoras

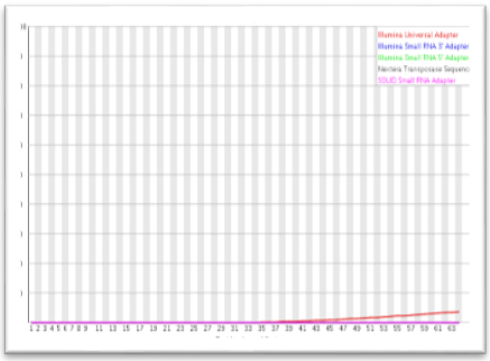


Figura 9: Principais parâmetros analisados no *FastQC*. Os gráficos presentes na figura representam os principais parâmetros analisados pelo FastQC na amostra *ERR022873_1.fastq* de verme adulto, como as estatísticas básicas da sequência, qualidade da sequência por base, conteúdo de GC por sequência, distribuição do tamanho da sequência, níveis de duplicação da sequência e sequencias adaptadoras.

Após a análise de qualidade, foram removidos os possíveis adaptadores utilizados para ancoragem dos oligonucleotídeos iniciadores provenientes do sequenciamento pelo Illumina. Essa remoção é de extrema importância pois essas sequencias podem ser incorporadas nas futuras montagens dos transcritos gerando erros graves. Para isso foi utilizado o programa *Trimmomatic* versão 0.32 (BOLGER; LOHSE; USADEL, 2014) incluindo a opção ILLUMINACLIP que corta os adaptadores e outras sequencias específicas do Illumina.

Após a remoção dos possíveis adaptadores, o segundo passo foi a remoção das sequências consideradas de baixa qualidade utilizando novamente o *Trimmomatic*. Nesse processo, foi utilizado o índice de qualidade *Phred* que é uma medida que avalia as bases geradas pelo sequenciamento automatizado de DNA (EWING; GREEN, 1998; EWING et al., 1998). Somente foram consideradas para esse trabalho as bases com uma qualidade *Phred* maior ou igual a 20 (*Phred* 20 considera 1 erro a cada 100 bases) (EWING; GREEN, 1998).

Dentre os parâmetros disponíveis para executar o *Trimmomatic* foram utilizados HEADCROP:15 (corta o número especificado de bases no início da leitura), LEADING:20 (corta as bases do início de uma leitura se estiver abaixo do limiar de qualidade), TRAILING:20 (corta as bases no final de uma leitura se estiver abaixo do limiar de qualidade), SLIDINGWINDOW:5:20 (Executa uma abordagem de corte de janela deslizante avaliando a qualidade da extremidade 5’ da leitura e realizando o corte estiver abaixo do limiar de qualidade) e MINLEN:50 (Remove a leitura se ela estiver com tamanho menor que o especificado).

4.1.1.4 Mapeamento das reads

Depois de verificar a qualidade dos arquivos no *FastQC* e remover todas as sequências de adaptadores ou de baixa qualidade que poderiam estar presentes, foi feito o mapeamento das *reads* utilizando o genoma de referência do *S. mansoni* na ferramenta *STAR* (DOBIN et al., 2013) . Para a utilização desse programa, foi necessário realizar o indexamento do genoma para o formato compatível com o *STAR*.

Uma vez que a indexação foi realizada, esse arquivo pode ser utilizado para o mapeamento de todas as *reads* de RNA-seq com o genoma do *S. mansoni*. Nessa etapa o *STAR* gera uma saída de mapeamento utilizando os nomes de arquivos fixos recomendando-se a execução do mesmo sempre de um diretório novo. Como todas as amostras presentes são consideradas *paried-end reads* os comandos específicos para essa opção foram definidos.

4.1.1.5 Formatação das extensões

A ferramenta *SAMtools* na versão 1.2 (LI, H. et al., 2009) foi utilizada para a conversão de todos os arquivos de extensão SAM gerados pelo *STAR* para extensão *Binary version of Sequence Alingment /Map* (BAM) e a sua respectiva ordenação. Ele fornece ainda utilitários para processamento de arquivos *Sequence Alignment/Map* (SAM) tais como classificação, indexação, ordenamento entre outros que podem ser utilizados dependendo da amostra e da análise. *SAMtools* é implementado em linguagem C e é um código fonte aberto.

O arquivo SAM é delimitado por tabulações e consiste de milhões de linhas, onde cada linha representa uma *read* e suas informações de alinhamento. As informações necessárias obtidas nesta linha são compostas da identificação de leitura e a posição inicial do mapeamento. Essa etapa é necessária para a posterior utilização na ferramenta *Cufflinks*.

4.1.1.6 Reconstrução dos transcritos

A reconstrução e identificação dos transcritos foi realizada com o programa *Cufflinks* (TRAPNELL et al., 2012), a partir da montagem dos alinhamentos em um conjunto parcimonioso de transcritos. Nele ainda é possível estimar a abundância e a expressão diferencial dos transcritos a partir de ferramentas dentro da última versão do *Cufflinks 2.2.1*.

Duas abordagens para a reconstrução dos transcritos são possíveis no *Cufflinks*. A primeira é realizada pela utilização do arquivo BAM gerado anteriormente e um arquivo GFF do genoma

inicial do *S. mansoni.* Nessa abordagem é quantificado principalmente o nível de expressão relativa dos genes codificantes de proteínas. Na segunda abordagem, ele realiza uma identificação *De novo* de transcritos, ou seja, sem a utilização do arquivo GFF do genoma. É uma ferramenta muito útil para analisar organismos pouco estudados e para identificar raros ou novos transcritos em um organismo.

Para esse trabalho, a segunda abordagem denominada *de novo* foi utilizada seguindo as opções padrões e omitindo o comando "-G refseq.gtf ". Sendo assim, novos transcritos foram montados para o genoma de *S. mansoni* para cada amostra de RNA-seq estudada.

O Cufflinks, apresenta ainda outras ferramentas adicionais que podem ser utilizadas dependendo da necessidade das amostras, denominadas de: Cuffmerge, Cuffquant, Cuffdiff, Cuffnorm, Cuffcompare. A ferramenta Cuffmerge foi utilizada nesse trabalho nas amostras em triplicata de cercária, esquistossômulos 3h, esquistossômulos 24h sendo combinadas para maximizar a sensibilidade da predição dos transcritos (Figura 10). No estágio de verme adulto essa ferramenta não foi utilizada pois não haviam amostras em triplicata para realizar a análise. Cuffquant é uma ferramenta que gera arquivos no formato CXB que funcionam como entrada para o Cuffdiff e/ou Cuffnorm. Outra ferramenta utilizada nesse trabalho foi o Cuffdiff que verifica a expressão diferencial enquanto a Cuffnorm cria expressões normalizadas e tabela de contas para serem utilizadas no programa R. Cuffdiff foi executado para gerar arquivos compatíveis para algumas análises posteriores do programa R (IHAKA; GENTLEMAN, 1996).

Finalmente a última ferramenta denominada Cuffcompare é utilizada para comparar os resultados gerados pela montagem *ab initio* com a montagem já publicada no genoma referência no formato GFF ou GTF. Essa ferramenta foi utilizada em todos os estágios, inclusive verme adulto com a finalidade de identificar transcritos que não tem potencial codificantes e que foram submetidos aos filtros seguintes.

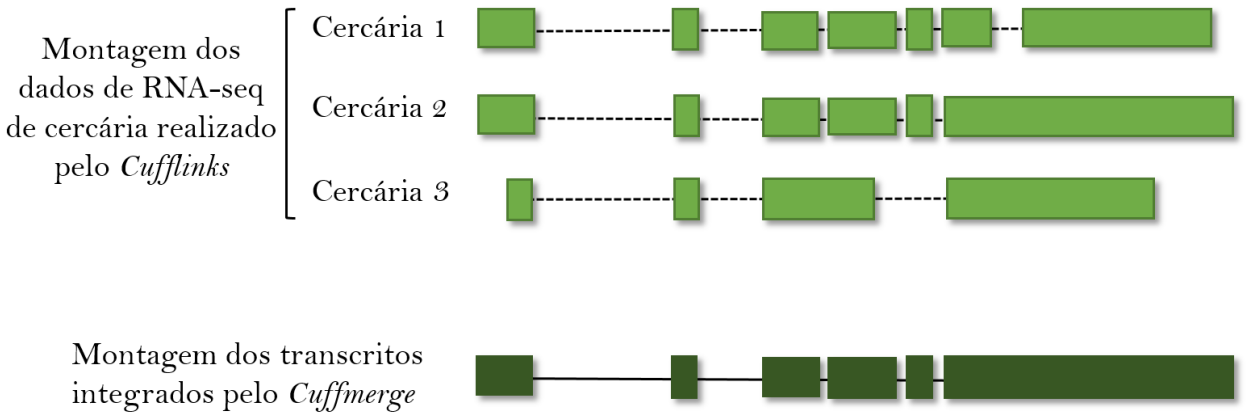


Figura 10: Esquema da montagem realizada pelo Cuffmerge. Exemplo da fusão dos transcritos gerados pelo Cufflinks em triplicata para o estágio de cercária gerando um consenso final pelo Cuffmerge.

4.1.1.7 Cálculos estatísticos e visualizações gráficas

Para a geração dos gráficos e análises estatísticas, foi utilizado o programa R versão 3.4.2 (IHAKA; GENTLEMAN, 1996) em associação com o programa RStudio versão 1.0.136 (http://www.rstudio.org). Nele podem ser realizadas análises e cálculos estatísticos além de gerar visualizações gráficas. É um programa livre originado de um projeto do GNU no qual podem ser instalados inúmeros pacotes para análises de amostras. Os gráficos foram gerados a partir do CummeRbund, um pacote do Bioconductor que é executado no R. Nele ainda está inserido diversos pacotes como *ggplot2*, *reshape2*, *fastcluster* entre outros, que auxiliam na interpretação e na plotagem dos dados.

A ferramenta *Integrative Genomics Viewer* (IGV) (ROBINSON, J. T. et al., 2011) foi utilizada para visualização e exploração de alto desempenho de conjuntos de dados genômicos grandes e integrados comparando os transcritos gerados. Assim, os arquivos dos transcritos obtidos em cada amostra dos diferentes estágios foram visualizados e analisados, comparando com o genoma e transcrissoma de *S. mansoni* anotados e depositados no Gene DB (Figura 11).

Figura 11: Visualização dos transcritos no IGV. Os transcritos originados dos quatro estágios pelo Cufflinks foram visualizados no programa IGV para análises futuras necessárias.

4.1.2 Predição dos lncRNAs em *S. mansoni*

Após as primeiras etapas que levaram a montagem dos novos transcritos pela ferramenta Cufflinks (Figura 8A), o arquivo final foi submetido a uma série filtros no *pipeline* a seguir para predição de lncRNAs no *S. mansoni* (Figura 8B). A maioria dos filtros dessas etapas foram compostos de ferramentas já descritas ou algoritmos próprios escritos na linguagem de programação Perl e Python específicos para o genoma do *S. mansoni*.

O primeiro passo foi utilizar o Cuffcompare para comparar os resultados gerados pela montagem dos transcritos de maneira *ab initio* com as anotações de transcritos conhecidos presentes no GeneDB no formato GTF. Como resultado do Cuffcompare, todas as montagens detectadas como novos transcritos foram classificados em 12 categorias diferentes (Tabela 3) de acordo com sua localização em comparação com os genes de referência do genoma de *S. mansoni* (TRAPNELL et al., 2012). Somente foram mantidas as três seguintes classes: transcrito intergênico

desconhecido (u), um fragmento que sobrepõe inteiramente dentro de um intron de referência (i) e sobreposição exônica com referência na cadeia oposta (x).

Tabela 3: Códigos das classes do Cuffcompare (Adaptado de (TRAPNELL et al., 2012).

Prioridade	Código	Descrição
1	=	*Match* completo com uma cadeia da introns
2	c	Contido
3	j	Isoforma potencialmente nova (fragmento): pelo menos uma junção de união é compartilhada com uma transcrição de referência
4	e	Fragmento de exon simples sobrepondo a um exon de referência e com menos 10 pb de um intron de referência, indicando um possível fragmento de pré-mRNA.
5	i	Um fragmento que sobrepõe inteiramente dentro de um intron de referência
6	o	Sobreposição exônica genérica com um transcrito de referência
7	p	Possível fragmento atuante da polimerase (dentro de 2 mil bases de um transcrito de referência)
8	r	Repetir. É determinado examinando a sequência de referência mascarada e aplicada a transcritos onde pelo menos 50% das bases são minúsculas
9	u	Transcrito intergênico desconhecido
10	x	Sobreposição exônica com referência na cadeia oposta
11	s	Um intron do fragmento sobrepõe um intron de referência na cadeia oposta (provavelmente devido a erros de mapeamento de leitura)
12	.	(apenas arquivo rastreamento, indica múltiplas classificações)

O segundo passo foi utilizado um algoritmo em Perl para extrair as sequências com um tamanho igual ou maior do que 200 nucleotídeos sendo uma das principais características dos lncRNAs conhecidos (DINGER et al., 2008; FRITH et al., 2006). Essa etapa é muito importante pois promove a remoção principalmente dos pequenos RNAs não codificantes que podem estar presentes na análise.

Para a realização do terceiro passo a ferramenta Coding Potential Assessment Tool (CPAT) foi utilizada. Ela foi escolhida dentre vários outras ferramentas para calcular o potencial codificante de transcritos desconhecidos com base em resultados mais eficientes na sensibilidade, especificidade, e velocidade de processamento nas amostras (WANG, L. et al., 2013). Esse programa se baseia em um método sem alinhamento que avalia o potencial codificante dos transcritos. Para predizer esse potencial codificante dessas moléculas, o CPAT utiliza um modelo de regressão logística construído com quatro recursos de sequência: (i) tamanho da ORF, (ii) cobertura da ORF, (iii) estatística de Fickett e (iv) frequência do frame em hexâmero. Somente os transcritos classificados pelo CPAT como não codificantes seguiram adiante gerando um novo arquivo FASTA. Para o nosso filtro proposto, foi utilizado o modelo logístico estabelecido para moscas (espécies do gênero *Drosophila*) como classificador do ponto de corte ideal (CP) de 0,39 sendo que CP ≥ 0,39 indicou uma sequência codificante, CP <0,39 indicou uma sequência não codificante (WANG, L. et al., 2013).

O quarto passo consistiu na extração somente dos transcritos que apresentaram ORF putativa menor que 300 nt utilizando o servidor OrfPredictor (MIN et al., 2005). Os transcritos com potencial de codificar proteínas geralmente apresentam ORFs maiores do que 300 nt de tamanho, gerando proteínas maiores ou iguais a 100 aa (aminoácidos) (CLAMP et al., 2007; DINGER et al., 2008).

Na quinta etapa, dois filtros foram aplicados juntos para melhor eficiência do processo uma vez que eles estão relacionados. Um ponto de corte foi estabelecido nos valores de FPKM com base na distribuição do nível de expressão dos lncRNAs. Somente os transcritos com valores de FPKM ≥ 2 seguiram sendo filtrados de acordo com o número de exons presentes. Para o segundo filtro dessa etapa somente os que apresentaram um número de exons ≥ 2 permaneceram em um novo arquivo pois lncRNAs com somente um exon são muito raros e mais comuns em sequências de mRNAs. (DERRIEN et al., 2012; ZHAN et al., 2016)

O sexto e sétimo passos foram de extrema importância na predição dos lncRNAs. Eles foram realizados manualmente utilizando os bancos de dados do *National Center for Biotechnology Information* (NCBI) (https://ncbi.nlm.nih.gov/) e GeneDB (http://genedb.org) como referência assim como sequências de lncRNAs do RNAcentral (http://rnacentral.org/) e de outros trabalhos com lncRNAs identificados (COPELAND et al., 2009; VASCONCELOS et al., 2017). Nesta

etapa, foram removidos outros tipos de RNA, transposons e genes anotados que possuíam homologia em espécies do gênero *Schistosoma*. Essas buscam evolveram a utilização da ferramenta BLAST versão 2.7.1 pelo NCBI e utilizando localmente (ALTSCHUL et al., 1997; CAMACHO et al., 2009).

4.1.3 Análises de enriquecimento do Gene Ontology

Para avaliar as possíveis funções dos genes nas proximidades dos lncRNAs preditos, utilizou-se a análise de enriquecimento do Gene Ontology (GO) (http://geneontology.org/). Nessa análise, todos os genes vizinhos em até 50 mil bases a montante e a jusante dos lncRNAs foram selecionados e, em seguida, suas funções foram avaliadas (OROM et al., 2010). Uma lista foi obtida com todos os termos GO encontrados e plotados com auxílio dos programas R, versão 3.4.2, e RStudio, versão 1.0.136, sendo o valor $p \leq 0{,}05$ considerado como significativamente enriquecido.

4.2 Análises *in vitro*

Os experimentos *in vitro* realizados foram padronizados nos últimos anos para o parasito *S. mansoni* no Laboratório de Bioquímica e Biologia Molecular (LBBM) da UFOP.

Todos os experimentos envolvendo animais foram autorizados pela Comissão de Ética no Uso de Animais da UFOP (CEUA-UFOP) sob o número de protocolo 2011/55. Além disso, todos os procedimentos realizados seguiram rigorosamente as diretrizes nacionais e internacionais para o manejo e cuidado de animais de laboratório.

4.2.1 Obtenção dos parasitos em diferentes estágios

Os diferentes estágios evolutivos do *S. mansoni*, linhagem LE, foram utilizados para análise nesse estudo. Os vermes adultos, vermes adultos resistentes ao praziquantel (PZQ) e as cercárias foram obtidos no Moluscário do Centro de Pesquisa René Rachou, Fundação Oswaldo Cruz (CPqRR/FIOCRUZ) em Belo Horizonte, Minas Gerais. Já os esquistossômulos mecanicamente transformados *in vitro* de 3,5h (EMT-3,5h) e os ovos foram cultivados no LBBM da UFOP. Esses experimentos foram realizados no LBBM em colaboração com o Laboratório de Enzimologia e Proteômica (LEP) da UFOP.

Os vermes adultos foram obtidos através da perfusão do sistema porta-hepático de camundongos da linhagem *Swiss Webster*, infectados por via subcutânea com aproximadamente 100 cercárias, após 50 dias (BASCH, 1981; SMITHERS; TERRY, 1965). Após sua coleta foram armazenados em tubo de 1,5 mL a -80ºC para posterior utilização.

Para a obtenção dos vermes adultos resistentes ao PZQ (LE-PZQ), uma porção dos vermes adultos foi submetida a três tratamentos com esse fármaco. Cada tratamento foi administrado em 5 dias consecutivos com intervalo de uma semana para a seleção dos parasitos menos suscetíveis. Os parasitos selecionados foram induzidos *in vivo* em camundongos tratados por 45 dias com 400 mg/kg de PZQ após a infecção (COUTO et al., 2011). Em seguida, os LE-PZQ foram obtidos por perfusão hepática em camundongos após 50 dias de infecção e lavados no meio *Roswell Park Memorial Institute* (RPMI) 1640 (Sigma Chemical Co.). O material obtido após a coleta foi rapidamente congelado em nitrogênio líquido e armazenado a -80 ° C.

Os ovos foram obtidos de amostras contendo, cada uma, 10 fígados de camundongos da linhagem *Swiss Webster* com 50 dias de infecção. Cada amostra foi triturada e homogeneizada com auxílio de um liquidificador doméstico em solução tampão de 200 mL (0,06 M de Na2HPO4, 0,0033 M de KH2PO4, pH 8,3), contendo 20 mg de tripsina (Invitrogen, Carlsbad, CA, EUA), e mantidos por 3h a temperatura de 37ºC em banho-maria. Esse produto sofreu tamisação em peneiras de malhas de 300 e 180 μm em solução salina isotônica seguida por 5 minutos de decantação (CASTRO-BORGES, 2005). Esse passo foi realizado por várias vezes até não se retirar mais ovos do material. Os ovos coletados foram transferidos para tubos de 1,5 mL e congelados em nitrogênio líquido a -80°C para posterior utilização.

As cercárias foram obtidas após 27 a 31 dias de infecção por miracídios no molusco *B. glabrata*. Cada caramujo foi colocado em contato com cerca de 10 miracídios em presença de luz.

As impurezas e o sobrenadante, que continham as cercárias, do meio foram retirados. O material obtido foi alocado em tubos de 1,5 mL e congelados em nitrogênio líquido a -80°C. para posterior utilização.

Para a obtenção dos esquistossômulos, a metodologia de Harrop e Wilson (1993) foi seguida. Inicialmente, o recipiente com as cercárias foram colocados em gelo para sedimentação e remoção das impurezas. Em seguida, o material foi transferido para tubos do tipo Falcon de 15 mL e ressuspendido em 10 mL do meio RPMI 1640 (Invitrogen^{TM}) filtrado. A suspensão foi deixada em gelo por 10 minutos. Esse processo de lavagem foi realizado por três vezes para remoção do máximo possível de impurezas.

Cada tubo foi separado com aproximadamente 200.000 cercárias e cinco mL de RPMI 1640, sendo, em seguida, agitados vigorosamente em vórtex, sob velocidade máxima, durante 90 segundos para separação mecânica da cauda do corpo cercariano. Após a ruptura, todo volume foi transferido para um novo recipiente de cultura composto por RPMI 1640, 1% de penicilina cristalina G (1000 UI/mL) e estreptomicina (1000 µg/mL) e incubado em estufa de CO_2 5% a 37°C por três horas e meia. Posteriormente, para sedimentação dos esquistossômulos e remoção das caudas presentes no sobrenadante, várias lavagens foram realizadas com um intervalo de quarto minutos entre cada em um fluxo laminar, sendo o procedimento acompanhado em um microscópio de luz invertida. O precipitado com apenas o corpo cercariano foi considerado como os EMT-3,5h de cultivo *in vitro* (Figura 12). Estes foram separados e utilizados para cultivo de um dia ou armazenados em tubos de 1,5 mL congelados em nitrogênio líquido e mantidos a -80°C.

Figura 12: Esquistossômulos cultivados *in vitro:* Os esquistossômulos foram cultivados por 3,5 horas em estufa de CO_2 5% a 37ºC em meio 169 (Adaptado de (DE SOUZA GOMES, 2008)).

Os EMT-3,5h foram cultivados em placas de seis poços contendo 8 mL de meio 169 (Tabela 4) suplementado com 10% de soro fetal bovino (Gibco) e 1% de penicilina cristalina G (1000 UI/mL) e estreptomicina (1000 μg/mL) por poço.

Tabela 4: Meio 169 com seus componentes e concentração.

Composição	Concentração
Hidrolisado de lactoalbumina	0,1%
Hidrocortisona	1×10^{-6} M
Hipoxantina	5×10^{-7} M
Triiodotironina (T3)	2×10^{-7} M
Serotonina	1×10^{-6} M
Glicose	0,1%
Meio mínimo vitamina	0,5%
Meio Schneider	5,0%
HEPES	20 mM
RPMI 1640 (Invitrogen^{TM})	q.s.p. 500 mL

4.2.2 Análise da expressão gênica por qRT-PCR

Para a expressão relativa por PCR quantitativa em Tempo Real (qRT-PCR) do conjunto de lncRNAs selecionados, foram utilizadas as etapas já padronizadas no LBBM e brevemente descritas abaixo.

4.4.2.1 Extração de RNA total

Aproximadamente 100 mg de vermes adultos, cercárias, EMT-3,5h e ovos foram utilizados para extração de RNA total utilizando o kit RNA total (SV total RNA Isolation System - PromegaT™) seguindo o protocolo do fabricante.

As amostras biológicas foram homogeneizadas em 1 mL de TRIzol Reagent® (Invitrogen™) com auxílio de um homogeneizador tipo politron. Posteriormente, os homogeneizados foram transferidos para tubos do tipo Eppendorf de 1,5 mL e incubados por 5 minutos à temperatura ambiente para completa dissociação dos complexos de nucleoproteínas. A seguir foram adicionados 0,2 mL de clorofórmio (Sigma) para cada 1,0 mL de TRIzol Reagent® (Invitrogen™) homogeneizando vigorosamente com auxílio de um vórtex. Posteriormente, as amostras foram centrifugadas por 12 minutos a 12.000 g à temperatura ambiente. A fase aquosa foi transferida para um novo tubo seguido da adição de volume equivalente de etanol 95% v/v, preparado com água livre de RNAses, e homogeneizado por inversão do tubo por três vezes para precipitação do RNA. A seguir, o RNA total foi purificado com o kit SVRNA System, conforme instrução do boletim técnico. A qualidade das preparações foi avaliada em gel de agarose/formaldeído (Figura 13) e a pureza e quantificação dos RNAs foram determinadas utilizando o aparelho Nano Vue Spectrophotometer (GEHealthcare) avaliando as relações entre os comprimentos de onda 260nm/280nm e 260nm/230nm indicativos de pureza da amostra. Todas a amostras de RNA extraídas foram armazenadas em alíquotas em tubos de 0,2 mL e mantidos a -80°C até o momento do uso.

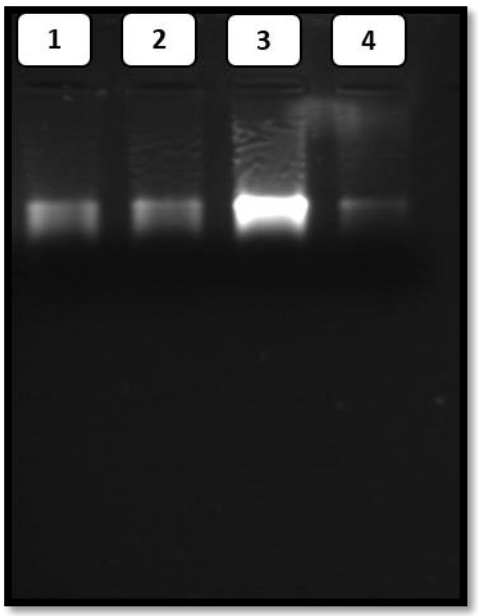


Figura 13: RNA total obtido do estágio de verme adulto**.** Foto do gel de agarose na concentração de 3% sendo as quatro canaletas representadas com RNA total de diferentes amostras de vermes adultos de *S. mansoni*.

4.2.2.2 Extração de RNA total de fígado de camundongos C57BL/6 (WT)

Foram realizadas ao todo 13 extrações de RNA com os fígados camundongos C57BL/6 (WT), sendo 10 infectados e 3 não infectados. Esses experimentos foram realizados em colaboração com o Prof. Dr. Vanderlei Rodrigues, do Laboratório de Biologia Molecular de Parasitas da Faculdade de Medicina da Universidade de São Paulo, Campus Ribeirão Preto (FMRP-USP), conforme protocolo CEUA-FMRP 35/2015.

Para esta etapa do projeto foi padronizado a utilização de 100-120mg de fígado obtido do lóbulo esquerdo do fígado. O RNA total foi extraído utilizado o kit PROMEGA SV total, como descrito em 4.2.2.1.

4.2.2.3 Oligonucleotídeos iniciadores

Para o conjunto de 15 lncRNAs em *S. mansoni* (Sm-lncRNAs) preditos e escolhidos para a expressão nesse projeto, foram desenhados pares de oligonucleotídeos iniciadores específicos para cada um de acordo com a sequência de interesse com auxílio do programa *Gene Runner*, versão *6.5.48*, conforme consta na Tabela 5. Além disso, foram desenhados pares oligonucleotídeos iniciadores para um Retrotransposon *Long Terminal Repeat* (LRT) e para os genes constitutivos/normalizadores *Eukaryotic translation initiation factor 4E* em *S. mansoni* (*SmeIF4E*) e *Hypoxanthine Phosphoribosyltransferase 1* (HPRT1) na espécie *Mus musculus* (Tabela 5).

Tabela 5: Sequências dos oligonucleotídeos iniciadores dos 15 Sm-lncRNAs e dos genes normalizadores.

Nome do lncRNA	ID do Transcrito	Sequência do oligonucleotídeo	Tamanho do produto (pb)
Sm-lncRNA 1	TCONS_00001011	F: 5'- AAGGGATGAGTTGACTGC -3'	119
		R: 5'- ACACGAAGACACCTATGACC -3'	
Sm-lncRNA 2	TCONS_00012347	F: 5'- AGACAATGCGATGCCGTTAG -3'	97
		R: 5'- TTTGGAACTCGTCAGCTAGG -3'	
Sm-lncRNA 3	TCONS_00013257	F: 5'- TCACATCTCGCAACTCAG -3'	103
		R: 5'- TCACATCTCGCAACTCAG -3'	
Sm-lncRNA 4	TCONS_00003599	F: 5'- TTTCGACACGGCAACTGATC -3'	99
		R: 5'- GCCGATTCAGTGTAGCAAAG -3'	
Sm-lncRNA 5	TCONS_00000625	F: 5'- GATCGAGCTGTAACTGCAC -3'	92
		R: 5'- GATCCACATCCATATGAGTG -3'	
Sm-lncRNA 6	TCONS_00001840	F: 5'- GACTGTTGGAAGAGGAAATG -3'	82
		R: 5'- GAGGATTTAAGCGACCATTG -3'	
Sm-lncRNA 7	TCONS_00009100	F: 5'- CCGATGAGATGCGTATAG -3'	134
		R: 5'- GCAACACAGTGAGGTAGAG -3'	
Sm-lncRNA 8	TCONS_00009852	F: 5'- CCACACAGGTAGTTCAGC -3'	111
		R: 5'- GAATCACTTGCACTTCGC -3'	

Sm-lncRNA 9	TCONS_00009849	F: 5'- CTGTGAGAATGGTGGATG -3'	83
		R: 5'- ACGTTTATGAGCCGTAGC -3'	
Sm-lncRNA 10	TCONS_00009851	F: 5'- GTGATATGCCCGGACAAAG -3'	108
		R: 5'- TTGAACGAGCAGCTGGAC -3'	
Sm-lncRNA 11	TCONS_00012478	F: 5'- CCTCGTGTTTGTGCTTTG -3'	82
		R: 5'- GGAATGTGATTGCCTAGTCG -3'	
Sm-lncRNA 12	TCONS_00010393	F: 5'- GCACTTGACACTAACCAGG -3'	125
		R: 5'- GGAGCTGTTCACTCATTG -3'	
Sm-lncRNA 13	TCONS_00010903	F: 5'- TTCCCTCCAGACTATGATCC -3'	144
		R: 5'- CACGTATTGCACCTGATG -3'	
Sm-lncRNA 14	TCONS_00011021	F: 5'- GTTGAAGAAGGTGAGTGC -3'	124
		R: 5'- GTGGAGGACTTGGAGATAC -3'	
Sm-lncRNA 15	TCONS_00013835	F: 5'- CCATGCAAGTGTGATCCG -3'	145
		R: 5'- GTGGGATTATCAGCTGCAGG -3'	
Retrotransposon	LTR- Retrotransposon	F: 5'- GGGTGCATCAGAGTAATC -3'	124
		R: 5'- ACTTGATCCGCATACTCC -3'	
SmeIF4E	Smp_001500	F: 5'- TGTTCCAACCACGGTCTCG -3'	89
		R: 5'- TCGCCTTCCAATGCTTAGG -3'	
MmuHPRT1	NM_013556.2	F: 5'- GCAGACTTTCCTTGGATC -3'	115
		R: 5'- CAACACTTCGAGAGGTCC -3'	

4.2.2.4 Síntese dos cDNAs

Para a realização das expressões dos transcritos por qRT-PCR, foi necessário realizar a síntese da primeira fita do cDNA de cada uma das amostras de RNA totais extraídas. Para isso, foi utilizado 1 μg de RNA total extraído de cada amostra e o reagentes Kit *High Capacity RT-PCR System* (AppliedBiosystems), seguindo as recomendações dadas pelo fabricante. Para cada 1μg de RNA total extraído foram utilizados: 2 μL de tampão da reação, 2μL de *primers* randômicos, 10 mM de dNTPs, 1,0 μL de Transcriptase reversa Multiscribe, 1 μL de inibidor de RNAse e água livre de nuclease para um volume final de 10 μL. Homogeneizou-se o tubo pipetando para cima e para baixo 2 vezes para completa homogeneização do RNA total na mistura. Posteriormente, o tubo contendo a mistura de reagentes e o RNA total foi incubado em termociclador (ThermoHybaid Px2) seguindo o seguinte programa: 10 minutos a 25º C, 120 minutos a 37º C para produção do

cDNA, 85°C por 5 minutos para inativação da enzima e por fim 4ºC. A amostra de cDNA foi estocada a -80°C até o momento do uso.

4.2.2.5 Expressão relativa dos lncRNAs por qRT-PCR

A análise da expressão dos lncRNAs nesse estudo foi realizada a partir da técnica qRT-PCR. As reações foram realizadas utilizando o kit *Power SYBR® Green Master Mix* (Thermo Fisher Scientific) utilizando 2µL dos iniciadores na concentração de 300 nM, 5µL de SYBR Green, 3µL de cDNA diluído por cinco vezes em água livre da enzima DNAse resultando em um volume final de 10 µL de reação por poço da placa. Todos os ensaios foram realizados em triplicata técnica e biológica para todos os transcritos de *S. mansoni* analisados, com o normalizador *SmeIF4E* e o Retrotrasposon-LRT presentes na mesma placa que os lncRNAs avaliados. O mesmo foi realizado para o normalizador *MumHPRT1* nos experimentos realizados em fígados de camundongos. Os valores do *threshold* foram fixados em 0,2 e os do *baseline* ajustados para 3-15 ciclos.

As análises foram feitas pelo método de quantificação relativa da expressão gênica (ΔCq), que permite quantificar diferenças no nível de expressão de um alvo específico entre as diferentes amostras. Os níveis dos genes alvos foram normalizados pelos níveis do controle endógeno (Gene normalizador). Os resultados foram alcançados pela fórmula aritmética $2^{-\Delta Cq} = 2^{-(Cq\ Gene\ -\ Cq\ Gene\ Normalizador\ eIF4E)}$. Todas a reações de qRT-PCR realizadas foram conduzidas conforme programação contida no aparelho ABI 7300 (Applied Biosystems).

4.2.2.6 Curva de eficiência dos oligonucleotídeos iniciadores

A eficiência da amplificação de todos os oligonucleotídeos iniciadores utilizados foi analisada através da realização de 5 diluições seriadas (1:4) de uma amostra de cDNA do estágio cercária. Este ensaio foi realizado em triplicatas com concentração inicial de *primer* de 300 nM.

Os dados obtidos geraram um gráfico de regressão linear onde o eixo X representa o log das concentrações de cDNA e o eixo Y o valor de Cq.

A eficiência da amplificação é determinada pelo *slope* da curva representada pela fórmula: Eficiência= $[10^{(-1/slope)} - 1] \times 100$ (Figura 14). Essa eficiência foi considerada apropriada para avaliação da expressão gênica, quando obtidos valores da reação acima de 95% e abaixo de 105%. Valores próximos, mas fora do estabelecido foram recalculados na fórmula aritmética da normalização citada anteriormente para adequar aos valores finais nos gráficos das expressões.

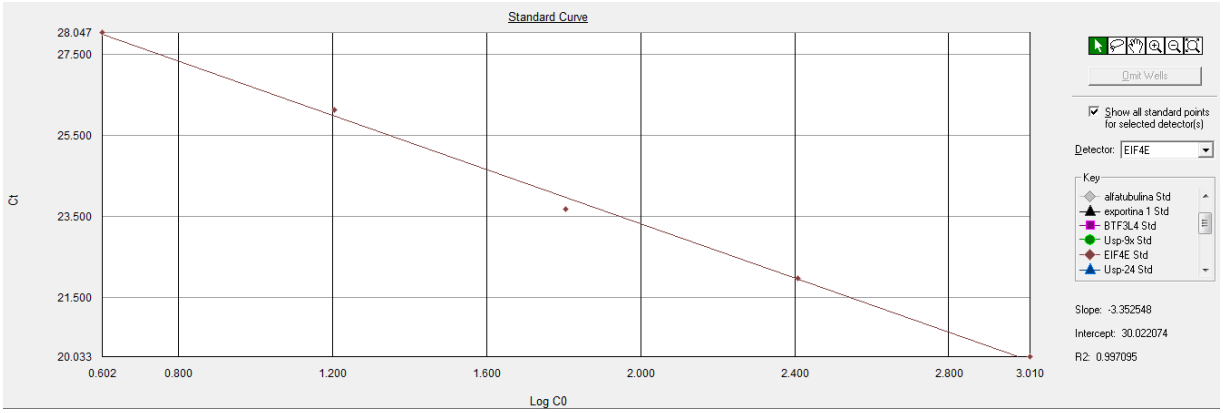


Figura 14: Exemplo da curva padrão referente ao gene constitutivo *SmeIF4E*. No eixo X estão representados os valores de Log da concentração de cDNA e no eixo Y os valores de Cq correspondes a cada diluição realizada. O coeficiente de linearidade e de *slope* estão representados na figura. Para a realização da curva de eficiência, foram utilizadas amostras de cDNA de cercária em uma diluição seriada de 4 vezes.

Os valores do limite (*Threshold*) do qRT-PCR foram fixados em 0,2 e os do *baseline* ajustados para 3-15 ciclos como representados na Figura 15.

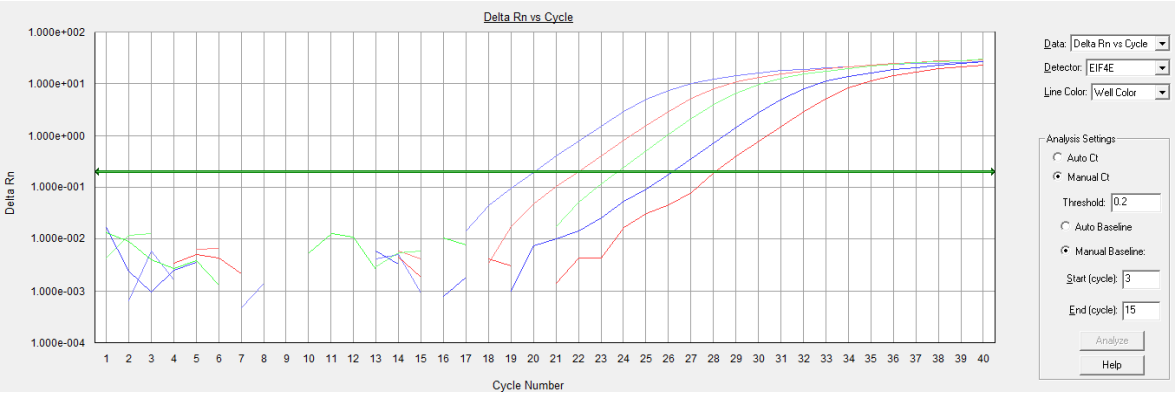


Figura 15: Gráfico de amplificação referente a curva de eficiência do gene *SmeIF4E*. No eixo X está representado o valor dos ciclos da qRT-PCR e no eixo Y os valor de Delta Rn. Amostras de cDNA de cercária foram utilizadas em uma diluição seriada de 4 vezes.

4.2.2.7 Análise da curva de dissociação dos produtos amplificados

Para analisar a curva de dissociação dos produtos amplificados da qRT.PCR, é necessário adicionar essa etapa na programação da amplificação. Ela tem como o objetivo observar a presença de possíveis amplificações inespecíficas no produto. Ao final dos 40 ciclos da programação da qRT-PCR para os lncRNAs, a temperatura foi elevada gradualmente de 60ºC para 95ºC, mantendo-se por 15s em cada temperatura, durante o qual é feito a leitura da emissão de fluorescência.

O gráfico resultante permite verificar se há um ou mais produtos de PCR presentes em cada reação devido a diferenças de temperatura de dissociação da sequência (*Melting temperature* - Tm). Neste caso a temperatura desejável foi entre 75°C e 80°C como exemplificado na Figura 16.

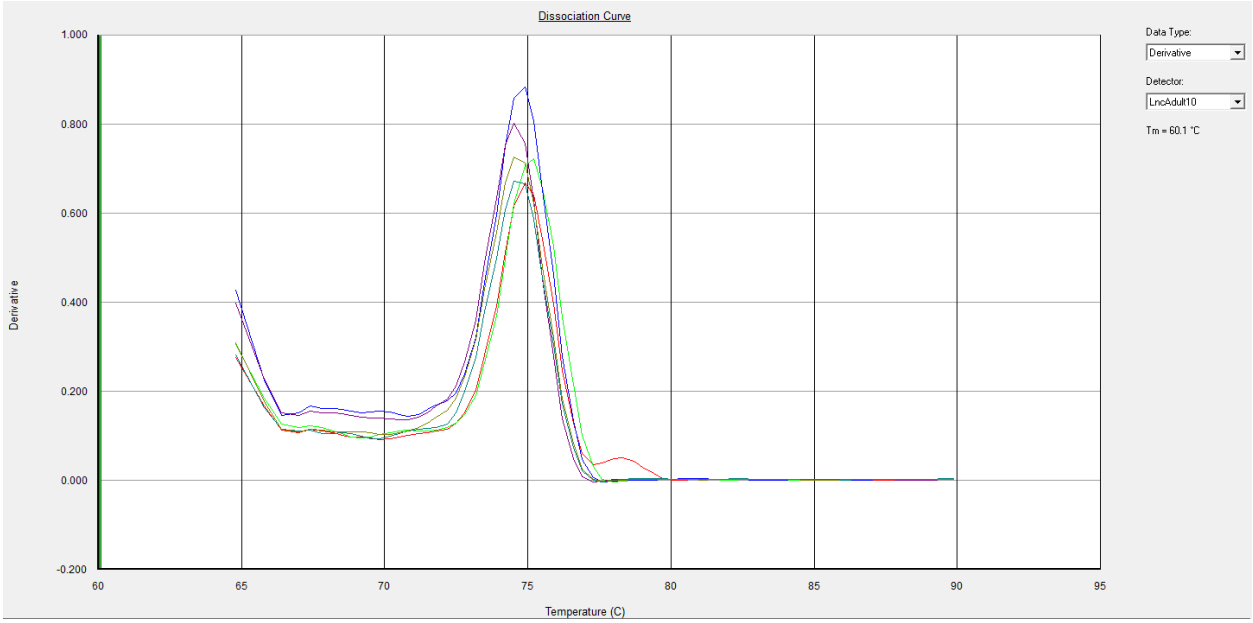


Figura 16: Curva de dissociação referente ao *Sm-lncRNA 10*. No eixo Y está representado a derivada do valor emitido pela florescência e no eixo X a temperatura de dissociação do produto gerado pela qRT-PCR que no caso é 75 °C.

4.2.2.8 Análise da qualidade dos produtos amplificados

Todos produtos amplificados na qRT-PCR juntamente com o padrão de peso molecular de *100pb DNA Ladder* (Invitrogen) foram analisados em gel de agarose com uma concentração de 1,2% (90 volts por aproximadamente 80 minutos). A finalidade dessa metodologia foi confirmar a presença de uma única banda com peso molecular esperado para cada um dos produtos da qRT-PCR obtidos (Figura 17).

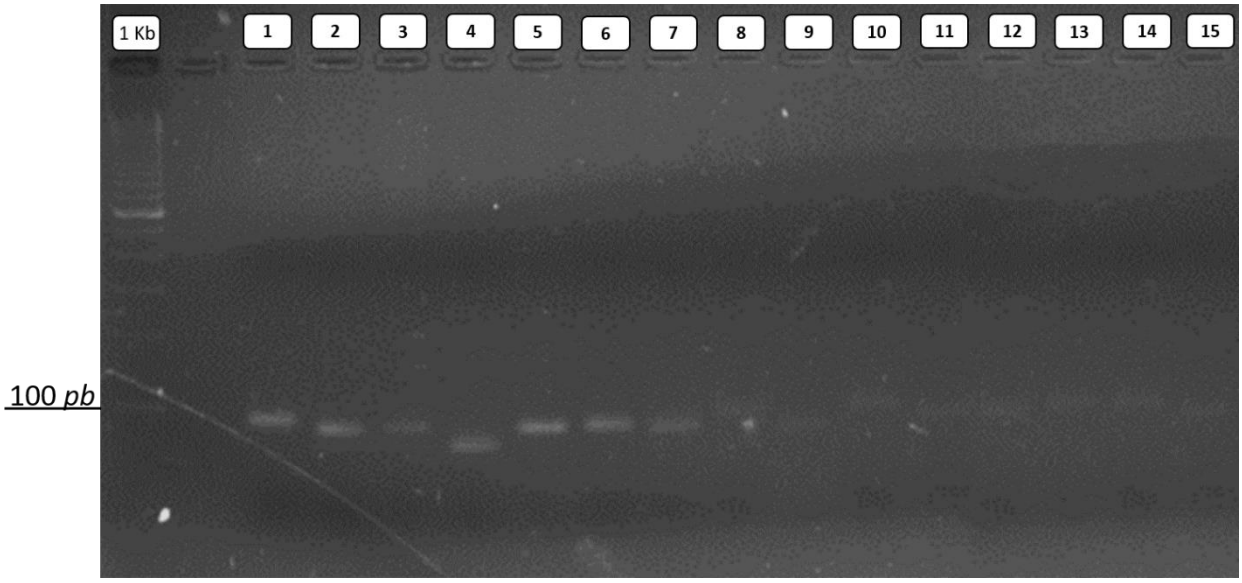


Figura 17: Gel de agarose dos lncRNAs**:** Gel de agarose na concentração de 1.2% com marcador de peso molecular de 1Kb (*Plus DNA Ladder Thermo Fisher*). Nas canaletas de 1 a 15 estão representados os produtos da qRT-PCR esperados do conjunto de 15 lncRNAs do 1 ao 15 respectivamente.

4.2.2.9 Análises estatísticas

As expressões relativas dos lncRNAs nos estágios do parasito foram comparadas utilizando a análise da variância por ONE-WAY (com pós-teste de Tukey). Foram consideradas estatisticamente significantes expressões de transcritos com o valor de $p < 0.05$ e $p < 0.001$. Essas análises estatísticas foram realizadas utilizando o programa GraphPad Prism 7 (San Diego, CA, USA).

5 RESULTADOS E DISCUSSÕES

5.1 Análises *in silico*

O *pipeline* descrito na metodologia foi executado para todos os 4 estágios analisados. Ao longo desse processo foram testados vários outros programas diferentes a fim de avaliar principalmente a sensibilidade e rapidez das análises nas amostras que serão discutidas nessa sessão. Entretanto somente para o estágio de verme adulto o *pipeline* foi finalizado incluindo as duas etapas finais e mais importantes (remoção de outros RNAs, transposons, genes anotados e homologia com lncRNAs conhecidos). A escolha desse estágio se deve principalmente a importância dele no ciclo, estando presente no hospedeiro definitivo e a viabilidade para validar experimentalmente um conjunto de lncRNAs preditos.

5.1.1 - Análise inicial do *pipeline*

5.5.1.1 Processamento das amostras

Antes de utilizar sequências adquiridas a partir de sequenciadores de última geração, é fundamental avaliar inicialmente a qualidade das bibliotecas de sequencias bem como o desempenho do sequenciamento. Antes mesmo da realização de um alinhamento, montagem de um genoma ou de um transcrissoma, devemos remover as bases com baixa qualidade e os adaptadores de sequências para não comprometer as análises posteriores com essa biblioteca (LEVIN et al., 2010; ZHOU et al., 2013).

Todas as 10 amostras (3 pares no estágio de cercária, 3 pares em esquistossômulos 3.5h, 3 pares em esquistossômulos 24h e 1 par de verme adulto) foram analisadas em relação a qualidade do sequenciamento seguindo inicialmente o controle executado pelo FastQC. Somente foram considerados sequências com uma qualidade aceitável para as análises, ou seja, resultados com o

valor de Phred maiores que 20. Esse valor é considerado o limiar para prosseguir com análises de montagem ou mapeamento (COCK et al., 2010).

Todas as amostras tiveram que ser processadas pelo Trimommatic seguindo a metodologia descrita pois apresentavam sequências com qualidades menores que Phred 20 e adaptadores utilizados no sequenciamento pelo Illumina. Além disso, a qualidade do início e do fim das bibliotecas foram avaliadas e em todas elas foram removidas as bases necessárias das leituras pois estavam abaixo do limiar de qualidade estabelecido. Após esse processamento inicial, somente as amostras que apresentaram condições ideais e foram incluídas nos parâmetros citados anteriormente (Figura 8) foram utilizadas.

A seguir está representado um exemplo da qualidade encontrada em uma biblioteca obtida a partir do estágio evolutivo de cercária denominada *ERR022877*. Nelas podem ser observadas a qualidade da biblioteca bruta (Figura 18) e da biblioteca processada (Figura 19). Todos os outros estágios seguiram o mesmo processamento padrão encontrado no estágio de cercária, sendo que os detalhes de cada amostra foram representados na Tabela 6.

Tabela 6: Bibliotecas de RNA-seq analisadas e processadas nos estágios do *S. mansoni*.

Biblioteca de RNA-seq	**Estágio**	**Total de sequências**	**Tamanho das sequências**	**%GC**
ERR022872_1 e 2	Cercária	16846390	76	44
ERR022872_1 e 2 Processadas	Cercária	10009831	50-61	41
ERR022877_1 e 2	Cercária	30777230	76	41
ERR022877_1 e 2 Processadas	Cercária	25733476	50-61	40
ERR022878_1 e 2	Cercária	21874383	76	41
ERR022878_1 e 2 Processadas	Cercária	20011735	50-61	40
ERR022874_1 e 2	Esquistossômulos 3h	7048165	76	46
ERR022874_1 e 2 Processadas	Esquistossômulos 3h	5067689	50-61	41
ERR022876_1 e 2	Esquistossômulos 3h	23727384	76	44
ERR022876_1 e 2 Processadas	Esquistossômulos 3h	19563051	50-61	40
ERR022879_1 e 2	Esquistossômulos 3h	29314489	76	38
ERR022879_1 e 2 Processadas	Esquistossômulos 3h	24198330	50-61	37
ERR022880_1 e 2	Esquistossômulos 24h	25308306	76	38
ERR022880_1 e 2 Processadas	Esquistossômulos 24h	22052213	50-61	37
ERR022881_1 e 2	Esquistossômulos 24h	25220643	76	38
ERR022881_1 e 2 Processadas	Esquistossômulos 24h	21205874	50-61	36
ERR022882_1 e 2	Esquistossômulos 24h	22248179	76	37
ERR022882_1 e 2 Processadas	Esquistossômulos 24h	17326584	50-61	35
ERR022873_1 e 2	Verme adulto	10521255	76	43
ERR022873_1 e 2 Processadas	Verme adulto	6371002	50-61	40

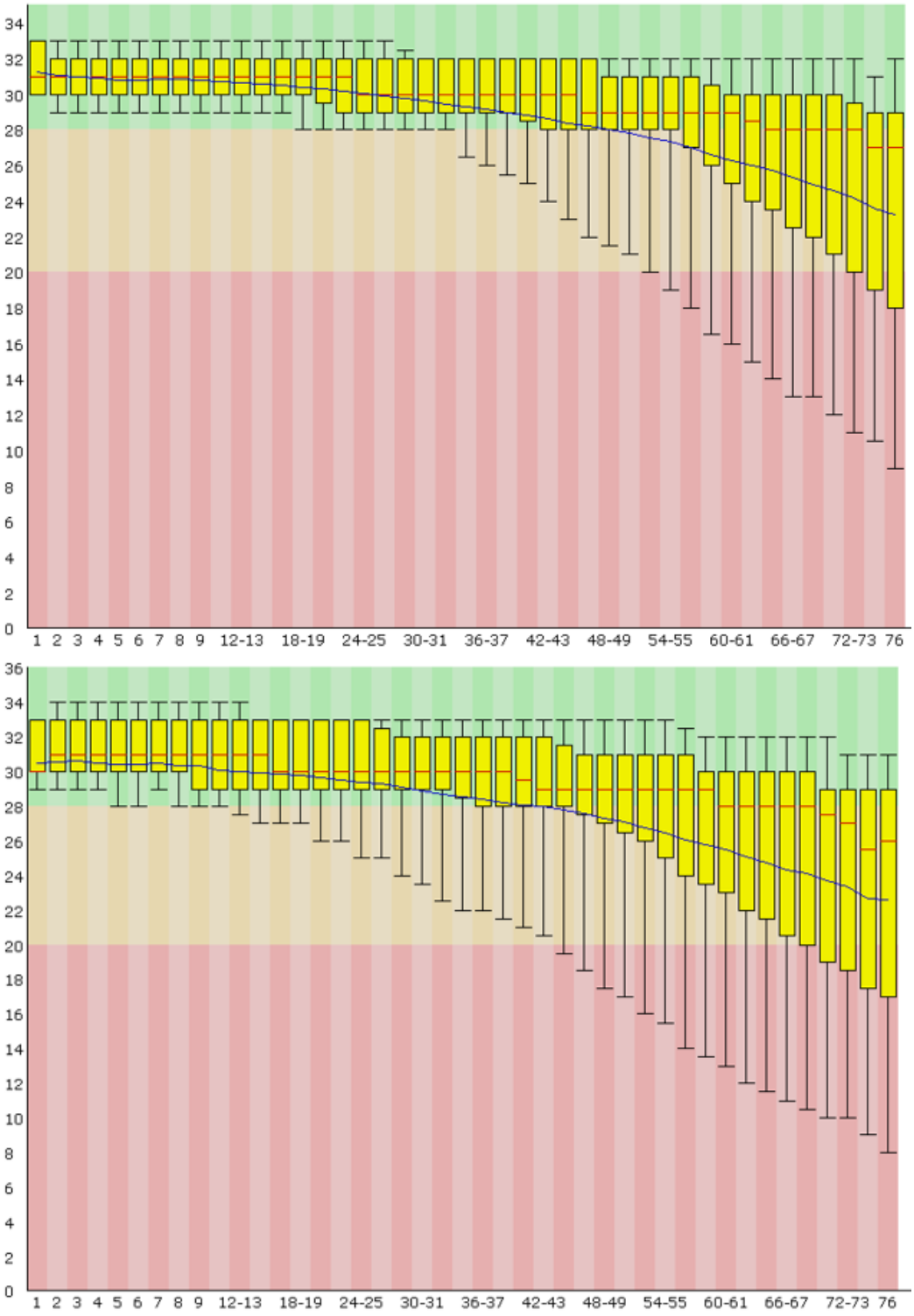


Figura 18: Exemplo da qualidade encontrada nas leituras brutas pareadas da biblioteca *ERR022877* no estágio de cercária. No eixo X está representado a posição nas leituras em pares de bases e no eixo Y os valores das qualidades em todas as bases.

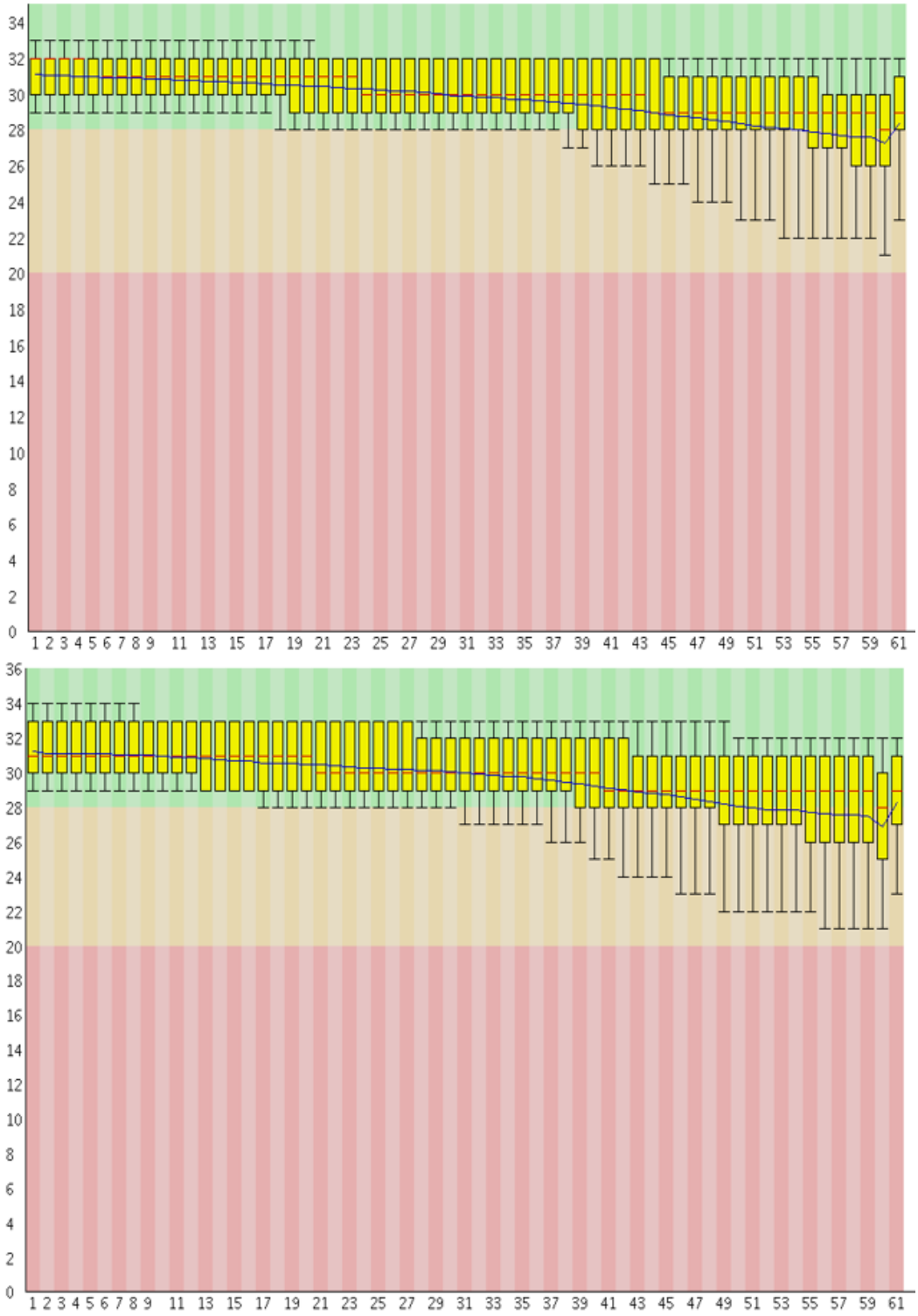


Figura 19: Exemplo da qualidade encontrada nas leituras processadas e pareadas da biblioteca *ERR022877* no estágio de cercária. No eixo X está representado a posição nas leituras em pares de bases e no eixo Y os valores das qualidades em todas as bases.

5.5.1.2 Mapeamento das leituras

Atualmente existem várias ferramentas que conseguem realizar o mapeamento de transcrissomas delimitando efetivamente onde os íntrons e os éxons estão presentes. Esses alinhamentos podem ser realizados na presença ou ausência de um genoma de referência, dependendo da disponibilidade do genoma da espécie analisada ou do tipo da análise realizada. Dentre os principais programas utilizados, podemos citar: *TopHat*, *GSNAP*, *MapSplice* e *STAR* (DOBIN et al., 2013; KIM et al., 2013; TRAPNELL; PACHTER; SALZBERG, 2009; WANG, K. et al., 2010; WU, T. D.; NACU, 2010).

Inicialmente a realização do mapeamento das leituras processadas foi realizada pela ferramenta TopHat que já é bem descrita para esse tipo de análise (TRAPNELL et al., 2009). Entretanto, devido à baixa velocidade de processamento e falta de compatibilidade com outros processos futuros utilizados no *pipeline,* o TopHat foi substituído por uma nova ferramenta denominada *STAR* que além de ser mais nova e mais rápida foi mais eficiente atendendo as necessidades do *pipeline* estabelecido (DOBIN et al., 2013).

A ferramenta *STAR* (*Spliced Transcripts Alignment to a Reference*) vem se destacando das demais devido a sua acurácia e velocidade do mapeamento realizado, sendo possível a finalização de vários projetos em larga escala, otimizando grandes conjuntos de dados de transcrissoma como por exemplo no projeto da criação do ENCODE no qual foi gerado mais de 80 bilhões de *reads* sequenciados pelo sistema Illumina (DJEBALI et al., 2012). Por esse motivo, ela foi escolhida e utilizada nesse trabalho devido a sua alta acurácia e velocidade de processamento em relação as outras, sendo composta por um algoritmo especializado em mapear amostras RNA-seq (ENGSTROM et al., 2013). Ela se mostrou bastante eficiente realizando os mapeamentos das amostras de RNA-seq nos diferentes estágios do *S. mansoni* em aproximadamente 3h-4h.

Os resultados obtidos por esse mapeamento apresentaram uma média de 66% de alinhamento único das leituras em relação ao genoma referência (Tabela 7). Esse fato pode ser explicado devido a última versão atualizada do genoma ainda não estar completamente finalizada, o que impossibilita muitas leituras alinharem. Outro fato que dificulta o alinhamento dessas leituras é a presença de muitos *contigs* e *scaffolds* que ainda não estão alocados em cromossomos no genoma do *S. mansoni*. Sequências com tamanhos muito pequenos também tiveram grande

influência no percentual das leituras não mapeadas. Entretanto esses resultados foram considerados satisfatórios para a posterior reconstrução dos transcritos seguindo o *pipeline* descrito.

Tabela 7: Dados das leituras mapeadas pela ferramenta *STAR*.

Estágio	**Amostra**	**Leituras iniciais**	**Leituras únicas mapeadas**	**Leituras múltiplas mapeadas**	**Leituras não mapeadas**
Cercária	ERR022872	10009831	59.08 %	9.91 %	16.03 %
Cercária	ERR022877	25733476	75.02 %	7.85 %	14.09%
Cercária	ERR022878	20011735	71.20 %	11.22 %	13.19 %
Esquistossômulos 3.5h	ERR022874	5067689	61.98 %	5.23 %	22.45 %
Esquistossômulos 3.5h	ERR022876	19563051	68.45 %	8.09 %	17.01 %
Esquistossômulos 3.5h	ERR022879	24198330	70.12 %	6.50 %	18.98 %
Esquistossômulos 24h	ERR022880	22052213	59.22 %	3.02 %	30.44 %
Esquistossômulos 24h	ERR022881	21205874	65.23 %	5.88 %	28.77 %
Esquistossômulos 24h	ERR022882	17326584	57.93 %	6.95 %	29.46 %
Verme adulto	ERR022873	6371002	74.11 %	8.07 %	12.09 %

5.5.1.3 Reconstrução dos transcritos

A etapa de reconstrução do transcrissoma foi realizada pelo Cufflinks (TRAPNELL et al., 2012; TRAPNELL et al., 2010), um programa livre escrito em C++ que o utilizou o mapeamento gerado pelo STAR, alinhando as leituras obtidas e alocando-as dentro de possíveis transcritos. Apesar de ser um programa que demanda bastante memória no sistema, ele foi executado sem problemas em sistemas Linux e *Macintosh Operating System* (Mac OS), permitindo arquivos de entrada com extensões do tipo BAM e SAM (LI, H. et al., 2009) utilizados.

A reconstrução realizada nesse trabalho conseguiu a montagem de aproximadamente 30 mil transcritos para cada um dos estágios de cercária, esquistossômulos 3,5h e esquistossômulos 24h. Isso foi possível devido ao uso das ferramentas Cuffmerge presente no Cufflinks que realizou a integração das 3 amostras em triplicata para cada estágio em uma amostra final (Figura 20). Para o estágio de verme adulto esse número foi somente de 15.776 transcritos pois somente uma amostra foi utilizada não necessitando assim o uso da ferramenta Cuffmerge.

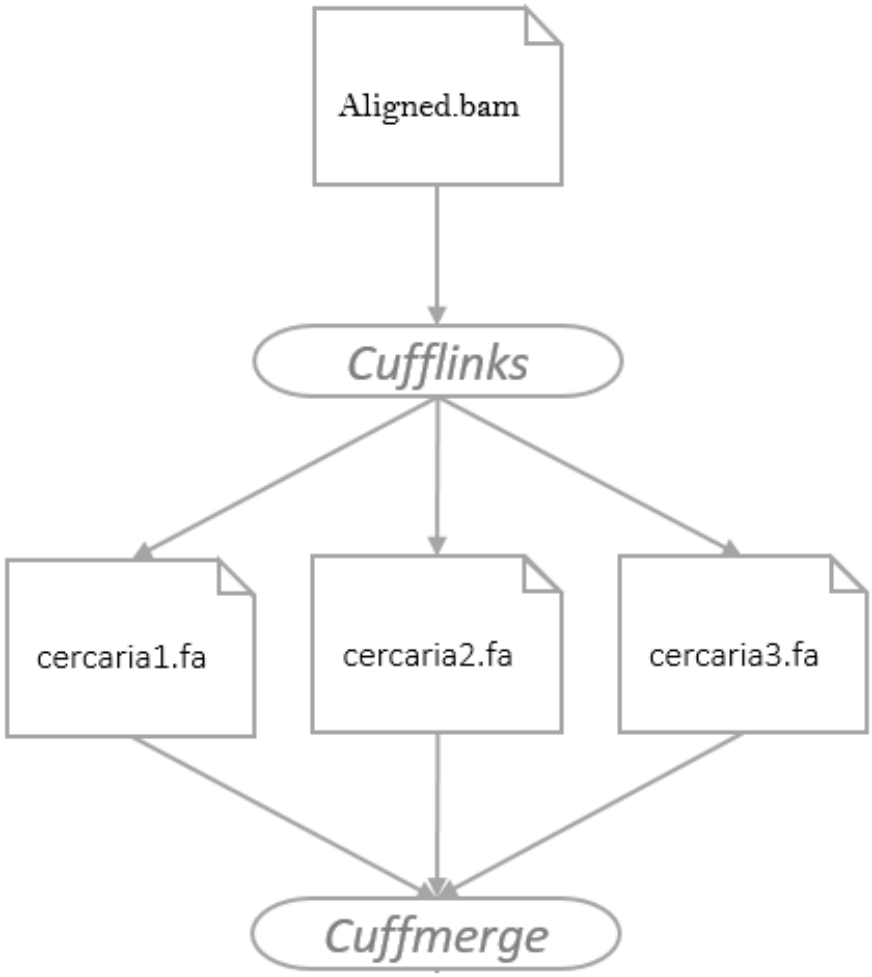


Figura 20: Ferramentas utilizadas no Cufflinks. Após o mapeamento, o arquivo de saída *Aligned.bam* de cada triplicada de cercária foi submetido ao Cufflinks e integrado pelo Cuffmerge em um arquivo único.

5.1.2 Predição de lncRNAs em *S. mansoni*.

A elaboração desse *pipeline* para predição de lncRNAs no genoma do *S. mansoni* foi de extrema importância devido à falta de metodologias adaptadas e específicas para o gênero *Schistosoma*. As metodologias atuais geralmente são espécie-específicas e seguem as características intrínsecas de cada genoma aumentando a sensibilidade e a especificidade da geração dos dados (ILOTT; PONTING, 2013; ZHAO et al., 2014). Cada metodologia assim, segue um fluxograma que melhor se adapta as etapas e ferramentas utilizadas. A Tabela 8 resume os principais trabalhos até o ano de 2013 realizados envolvendo a predição de lncRNAs a partir de dados de RNA-seq. Fica evidente que cada trabalho seguiu um fluxograma onde foram adaptadas as etapas e programas utilizados em cada espécie incluindo humanos, camundongos, *Zebrafish*, *Drosophila melanogaster*, *Caenorhabditis elegans*, e *Gallus gallus domesticus* (ILOTT; PONTING, 2013). O genoma do *S. mansoni* por exemplo apresenta características particulares, uma vez que parte da fração repetitiva do seu DNA consiste em genes ribossômicos repetidos em tandem, dos quais existem 500-1000 cópias por genoma que representam 1,8-3,6% do DNA total e sequências não ribossômicas repetitivas que compreendem mais de 2,0% do DNA total (SIMPSON et al., 1982).

Nos últimos 4 anos, essas metodologias para a predição de lncRNAs sofreram ainda mais alterações devido a publicações de novas ferramentas mais rápidas, sensíveis e específicas para cada genoma (LUO et al., 2017; ZHANG, Y. et al., 2017). Foram ainda criados servidores específicos para predição de lncRNAs em alguns genomas específicos, facilitando as análises de dados de RNA-seq gerados em larga escala (HOU et al., 2016; LIAO et al., 2011).

Tabela 8: Metodologias de 9 estudos para predição de lncRNAs utilizando dados de RNA-seq. Adaptado de (ILOTT; PONTING, 2013).

Nine studies that have predicted lncRNAs using RNA-seq. Library type, sequencing strategies, alignment and assembly algorithms and the number of lncRNAs discovered are provided.

Organism	Cell/tissue	Selection	SE/PE	Read length (bp)	Raw reads	Mapped	Stranded	Alignment	Assembly	Coding potential	Number of lncRNA loci
Human	[a]Tier I cell lines Tier II cell lines Tier III cell lines	pA+, rRNAd, pA−, rRNAd, DSN	PE	76	95	76–89%	✓	STAR (CSHL) or TopHat (CalTech)	Cufflinks	Mass spectrometry	33686 Single-exon 7518 multi-exon
Human	[b]22 tissues/cells	pA+	PE, SE	50, 75	[d]>175	–	×	TopHat	Cufflinks and Scripture	PhyloCSF/Pfam alignment	4662 stringent multi-exonic 8195 non-stringent multi-exonic
Mouse	Embryonic stem cells, lung fibroblasts, neuronal progenitor cells	pA+	PE	76	118	61%	×	TopHat	Scripture	ORF length, CSF	1749 multi-exonic
Mouse	Cortical layers 1–3, 4, upper 5, lower 5, 6, and 6b	pA+	PE	50	55	75%	×	TopHat	Cufflinks	CPC	1879 multi-exonic
Mouse	Foetal liver (*Elf1* Knockout)	pA+	PE	76	26	84%	×	TopHat	Cufflinks	ORF length, phyloCSF	308 multi-exonic
Zebrafish	[c]Embryonic development	pA+	PE	76	260	80%	✓	TopHat	Cufflinks and Scripture	PhyloCSF, BLASTX/P, ORF length	1133 multi- and single-exon
Drosophila melanogaster	Embyros, L1 larvae, L2 larvae, L3 larvae, pupae, Adult	pA+	PE, SE	75	135	64%	×	TopHat	Cufflinks	CPC, phyloCSF	1119 multi-exonic
C. elegans	Embryonic, larval, dauer and adult	pA+	PE, SE	36	[d]>240	–	✓/×	TopHat	Cufflinks	CPC	1145 multi- and single-exonic
Chicken	Skeletal muscle	rRNAd	PE	75	14	80%	×	TopHat	Cufflinks	ORF length, CSF	281 multi- and single-exon

Abbreviations: pA+, poly-A+ selection; pA−, poly-A− selection; rRNAd, rRNA depletion; DSN, double stranded nuclease normalisation; PE, paired-end; SE, single-end; CPC, coding potential calculator; CSF, codon substitution frequency; ORF, open reading frame.

[a] Tier I cell lines: K562, GM12878, H1-hES cells (H1-hESC); Tier II cell lines: HUVEC, HepG2, Hela-S3; Tier III cell lines: NHEK, MCF7, AGO4450, SK-N-SH + Retinoic Acid, A549, HSMM, NHLF, HMEC, BJ.

[b] Human lung fibroblasts (hLF), hLF2, foreskin fibroblasts, brain, liver, placenta, testes, adipose, adrenal gland, HeLa, breast, colon, kidney, heart, lung, lymph node, ovary, prostate, skeletal muscle, white blood cells, thyroid.

[c] Two–four cell, 1000 cell, dome, shield, bud, 28 hpf, 48 hpf, and 120 hpf.

[d] Where the number of raw reads could not be determined from the publication, the number of aligned reads is provided in the raw reads column.

Para esse estudo, todos os transcritos obtidos na análise inicial do *pipeline* (Figura 21A) foram submetidos a predição de lncRNAs em *S. mansoni*. Nessa fase, várias etapas incluindo filtros de remoções e seleções foram realizados por algoritmos próprios e programas específicos que em conjunto realizaram a predição desses lncRNAs. Esses filtros foram aplicados em todos os estágios gerando um número de transcritos preditos respectivamente para cada etapa (Figura 21B). Como já explicado inicialmente, somente para o estágio de vermes adultos foram aplicados os dois últimos filtros finais e determinantes para a otimização do processo.

Os transcritos gerados pelo *Cufflinks* foram submetidos a primeira etapa de filtragem e seleção dos lncRNAs preditos. Ela foi realizada pela ferramenta *Cuffcompare* que classificou esses transcritos de acordo com sua localização genômica e removeu montagens sobrepostas as sequências anotadas no genoma de referência, incluindo transcritos codificantes reconstruídos e transcritos codificantes conhecidos. Foram selecionadas somente as classes u, i e x, determinadas pelo *Cuffcompare* levando a remoção de 50% ou mais dos transcritos montados nos diferentes estágios. Essas classes foram escolhidas de acordo com resultados de outros *pipelines* de predição de lncRNAs estabelecidos. Neles foram avaliados entre outras coisas, a viabilidade dessas sequências serem possíveis lncRNAs sendo incluídas nas análises (LI, H. et al., 2016; LI, T. et al., 2012; ZHU, B. et al., 2017).

A segunda etapa foi submeter os transcritos restantes a um filtro de comprimento mantendo somente sequências com o tamanho ≥ 200 nt. Esse filtro foi essencial para remover possíveis classes de RNAs pequenos das análises. Como já esperado, no geral poucos transcritos foram removidos nessa etapa corroborando com *pipelines* de outros estudos para predição de lncRNAs (BOERNER; MCGINNIS, 2012; SUN, L. et al., 2012; ZHANG, Y. C. et al., 2014).

Após serem submetidos ao filtro de comprimento, os transcritos foram analisados em relação aos seus potenciais codificantes. Para essa análise a terceira e quarta etapa foram implementadas com o programa *CPAT* e um filtro para predição de ORFs ≤ 300 nt. Novamente como esperado, poucos transcritos foram removidos em cada uma das etapas nos diferentes estágios. Entretendo, esses resultados em conjunto apresentam uma remoção significativa uma vez que as duas ferramentas utilizam predições de ORFs de maneiras diferentes ao longo das sequências.

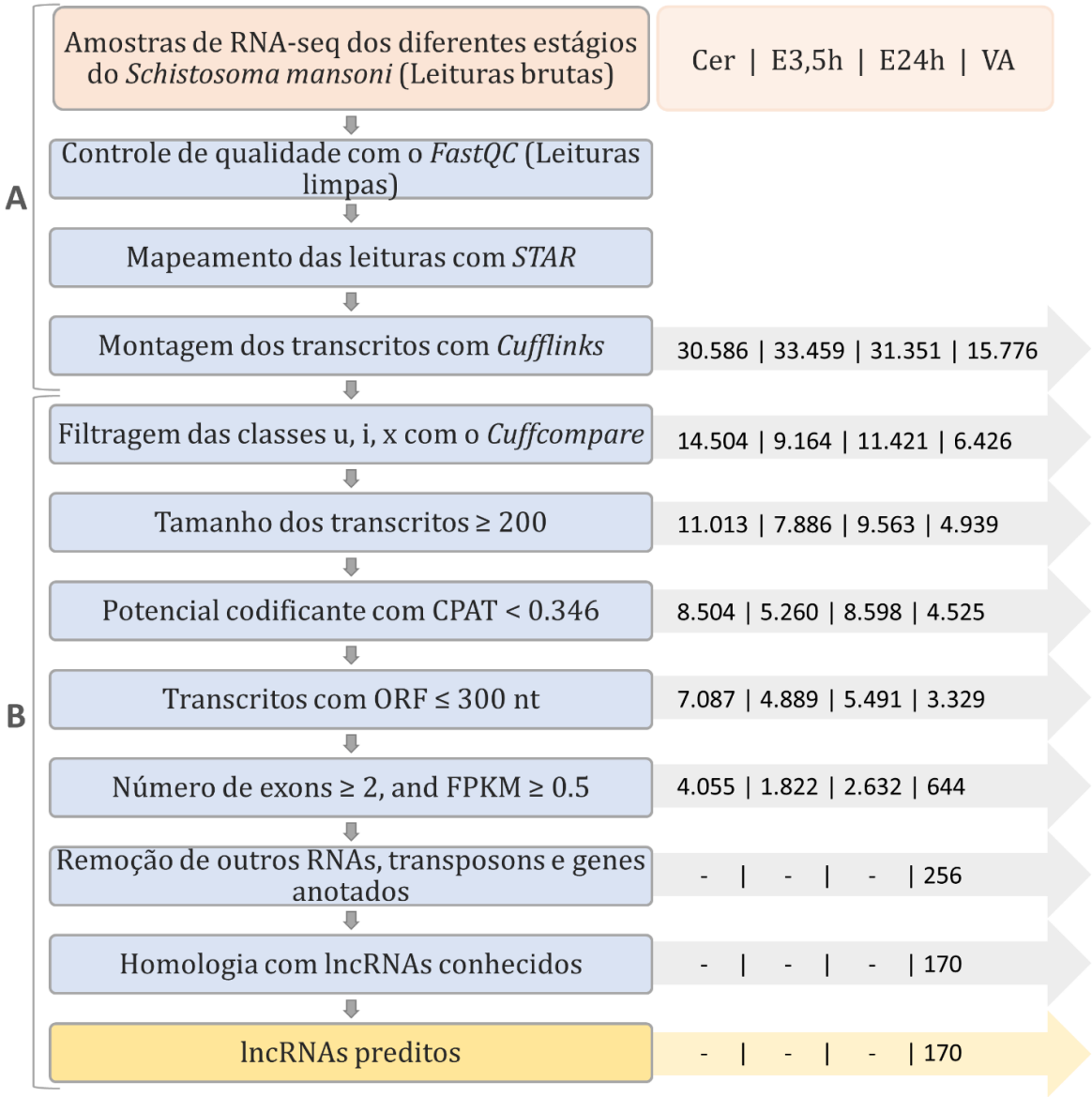


Figura 21: *Pipeline* para predição dos lncRNAs em *S. mansoni*. (A) Análise inicial do *pipeline*. (B) Predição dos lncRNAs. Os resultados obtidos em cada etapa estão indicados nas setas a direita nas amostras de Cercária (Cer), Esquistossômulos 3,5h (E3,5h), Esquistossômulos 24h (E24h) e Verme Adulto (VA) respectivamente.

Para a quinta etapa foram implementadas duas categorias: transcritos com número de exons ≥ 2 e FPKM ≥ 0,5. Nela aproximadamente 3.000 transcritos foram removidos em cada estágio, gerando um número de candidatos mais robustos. Para o estágio de cercária foram preditos 4.055 transcritos, 1.822 em esquistossômulos 3,5h, 2.632 em esquistossômulos 24h e 644 em verme adulto.

Uma vez que este estudo se concentrou nos transcritos longos não codificantes, as duas últimas etapas foram realizadas com uma curadoria manual dos dados. Essas etapas foram realizadas somente para o estágio de verme adulto. A sexta etapa removeu outros tipos de RNAs, transposons, e transcritos já anotados. Cerca de 35-40% dos transcritos apresentaram similaridade significativa com transposons já identificados no genoma, 5% de outros RNAs, e 15% de genes homólogos anotados em outras espécies do gênero *Schistosoma*. Para esses transcritos com homologia para genes anotados foram identificadas sequências nas espécies: *Schistosoma rodhaini, Schistosoma curassoni*. *S. japonicum, S. haematobium.*

Vale ressaltar que os retrotransposons contribuem fortemente para a diversidade de sequências e alta complexidade dos lncRNAs. Mais de dois terços das sequências maduras de lncRNAs (75 e 68% de humano e rato, respectivamente) têm em pelo menos uma inserção parcial de retrotransposon em sua sequência. No total, 29.519 sequências com características funcionais derivadas do elemento de transposição (TSS, sítios poliA e *splicing*) foram identificados no GENCODE v13 (DINGER et al., 2008; KAPUSTA et al., 2013). Estima-se que 45% do genoma de *S. mansoni* seja composto por elementos de transposição (PROTASIO et al., 2012). Curiosamente, a função da maioria dos retrotransposons permanece mal compreendida. Muitos parecem regular a transcrição e a manutenção da estrutura da heterocromatina e provavelmente representam um conjunto de elementos reguladores, como, por exemplo, promotores alternativos (MUÑOZ-LÓPEZ; GARCÍA-PÉREZ, 2010).

Seguindo essas características do genoma, nesse trabalho optamos por adicionar a sétima e última etapa ao *pipeline*, a pesquisa por homologia dos transcritos remanescentes com outros lncRNAs descritos em outros trabalhos. Nessa busca foram encontrados e removidos 86 (33,5 %) lncRNAs com similaridade maior ou igual a 80 % com as sequencias do trabalho de Vasconcelos (VASCONCELOS et al., 2017). Ao final desse *pipeline* foi predito um conjunto robusto de 170 lncRNAs em *S. mansoni*.

5.1.3 Análise dos genes alvos dos lncRNAs

As análises de enriquecimento do Gene Ontology (GO) para os genes alvo da vizinhança foram feitas em três aspectos diferentes, sendo o processo biológico (BP) (Figura 22), a função molecular (MF) (Figura 23) e o componente celular (CC) (Figura 24). Os 10 aspectos de GO mais significantes foram representados incluindo 389 genes envolvidos na BP, 555 genes envolvidos em MF, 768 genes envolvidos em CC.

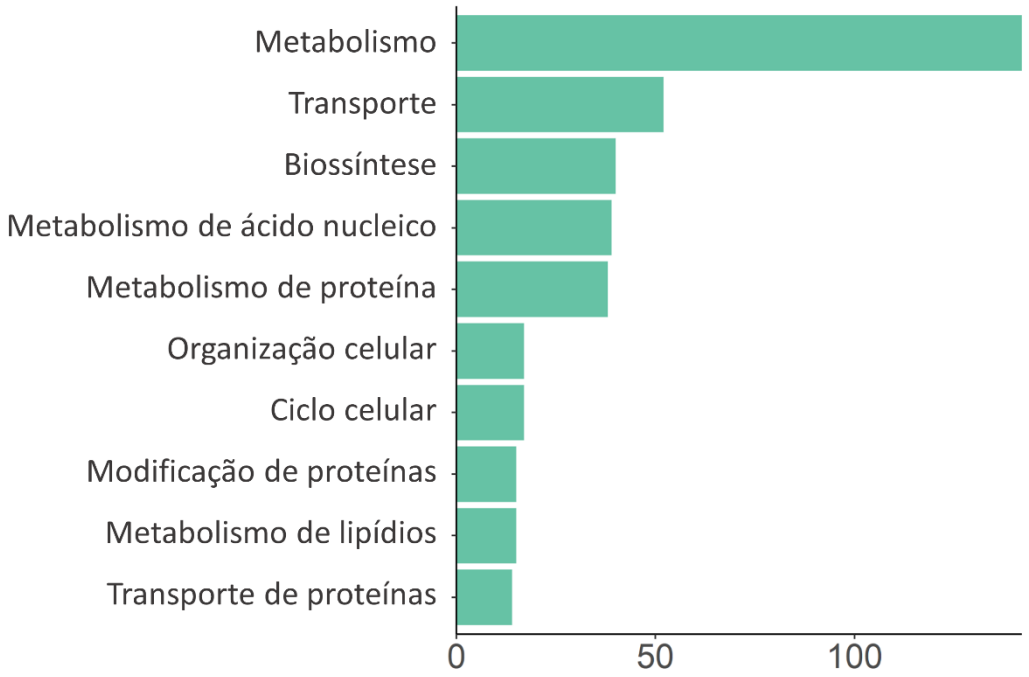


Figura 22: Análise de enriquecimento dos processos biológicos para os genes alvos dos lncRNAs. As 10 categorias mais frequentes foram calculadas a partir do enriquecimento dos genes.

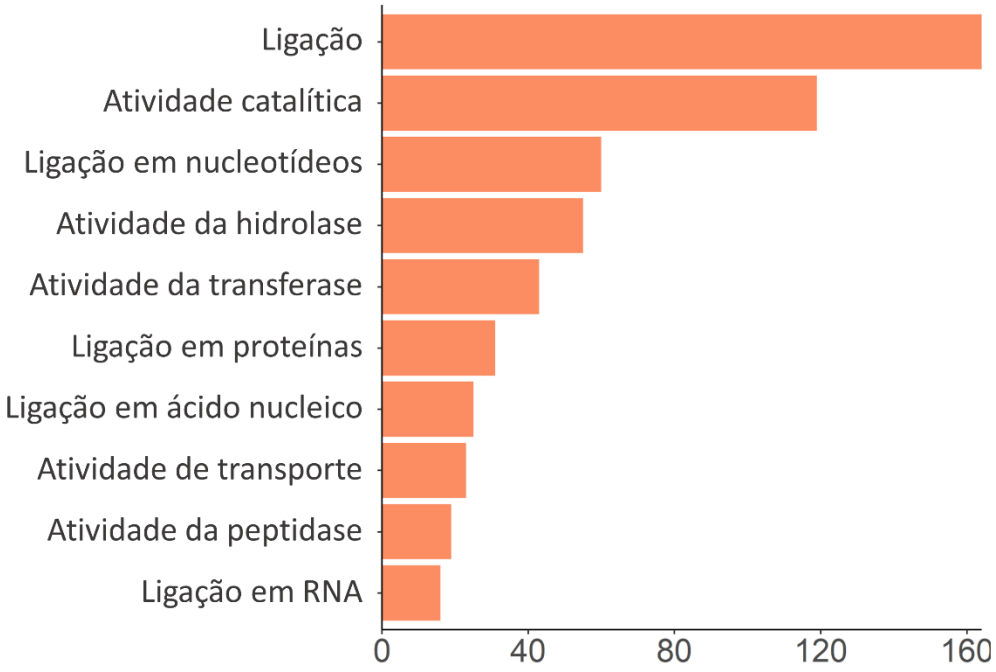


Figura 23: Análise de enriquecimento das funções moleculares para os genes alvos dos lncRNAs. As 10 categorias mais frequentes foram calculadas a partir do enriquecimento dos genes.

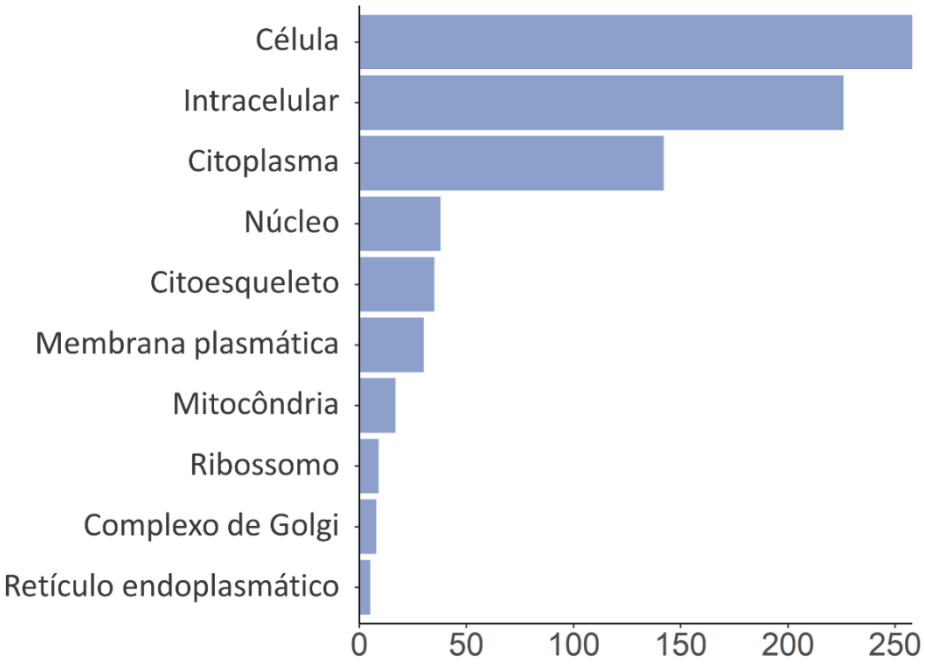


Figura 24: Análise de enriquecimento dos componentes celulares para os genes alvos dos lncRNAs. As 10 categorias mais frequentes foram calculadas a partir do enriquecimento dos genes.

A maior parte dos genes encontrados no termo de processos biológicos foram relacionados ao metabolismo, transporte e biossíntese. Para genes no termo função molecular foram predominantemente encontradas proteínas de ligação incluindo atividade catalítica e ligação a nucleotídeos. Quanto aos componentes celulares, os termos mais enriquecidos foram na célula, na região intracelular e no citoplasma. Foi feita uma tabela detalhada dos 15 lncRNA expressados, incluindo os genes de codificação vizinhos e suas respectivas entradas GO (Tabela 9).

Tabela 9: Conjunto do 15 Sm-lncRNAs expressos com a localização dos genes codificantes vizinhos.

LncRNA	ID do Transcrito	Tamanho	Vizinhança genômica	Gene de *S. mansoni*
Sm-lncRNA 1	TCONS_00001011	570	20861 pb fita 5' : Complexo nucleolar de levedura putativo	Smp_154980
			36025 pb fita 3' : Proteína hipotética	Smp_086060
Sm-lncRNA 2	TCONS_00012347	556	6900 pb fita 5' : Proteína hipotética	Smp_193940
			1354 pb fita 3' : Proteína de fidelidade de transmissão cromossômica 8	Smp_130740
Sm-lncRNA 3	TCONS_00013257	742	13223 pb fita 5' : Domínio protéico putativo de pdz and lim	Smp_022340
			5551 pb fita 3' : DNAj putativo homologo da subunidade B mebro 2,6,8	Smp_022330
Sm-lncRNA 4	TCONS_00003599	346	6057 pb fita 3' : Histona Lisina N metiltransferase MLL5	Smp_246410
Sm-lncRNA 5	TCONS_00000625	202	15028 pb fita 5' : Subunidade proteasoma beta 1 (família T01)	Smp_025800
			3974 pb fita 3' : Produto de proteína sem nome	Smp_200530
Sm-lncRNA 6	TCONS_00001840	278	889 pb fita 5' : Ribonucleoproteína nucleolar U3 pequena	Smp_102820
			308 pb fita 3' : Receptor de fator de crescimento de fibroblastos a	Smp_175590
Sm-lncRNA 7	TCONS_00009100	659	7028 pb fita 5' : Histona putativa H2A	Smp_130880
			9619 pb fita 3' : Cadeia E da ATP sintase	Smp_015980
Sm-lncRNA 8	TCONS_00009852	1157	5857 pb fita 5' : rRNA	Smp_sma.5s-14.1
			2437 pb fita 3' : Supressor da actina (sac)	Smp_060420
Sm-lncRNA 9	TCONS_00009849	2098	27717 pb fita 5' : Proteína hipotética	Smp_193460
			53702 pb fita 3' : rRNA	Smp_sma.5s-14.1
Sm-lncRNA 10	TCONS_00009851	875	27717 pb fita 5' : Proteína hipotética	Smp_193460
			53702 pb fita 3' : rRNA	Smp_sma.5s-14.1
Sm-lncRNA 11	TCONS_00012478	1591	47352 pb fita 5' : Proteina S9 ribossomal 40s putativa	Smp_180000

			-	
Sm-lncRNA 12	TCONS_00010393	237	9284 pb fita 5' : Resistência à toxina difteria de resistência 2, dph2 putativa	Smp_174680
			4506 pb fita 3' : Enzima conjugadora da ubiquitina E2 J1	Smp_174670
Sm-lncRNA 13	TCONS_00010903	1166	1571 pb fita 5' : Proteína de radiação radial flagelada putativa 3	Smp_170010
			32039 pb fita 3' : Fator de ubiquitinação E4a putativo	Smp_030780
Sm-lncRNA 14	TCONS_00011021	2280	7796 pb fita 5' : Proteína de alérgeno de veneno (VAL) proteína 7	Smp_199890
			42906 pb fita 3' : Proteína hipotética	Smp_156240
Sm-lncRNA 15	TCONS_00013835	664	8674 pb fita 3' : Ubiquinol citocromo C redutase	Smp_061870
			-	

5.1.4 Características do lncRNAs preditos

Para determinar as características dos lncRNAs de *S. mansoni*, foram analisados os seguintes aspectos: a localização genômica, o número de exons, tamanho dos transcritos, log_2 FPKM (Figura 25). Neste conjunto de dados, foi identificado que a maioria dos lncRNAs foram encontrados no cromossomo 1, sexual (ZW) e nos *scaffolds* que ainda não estão integrados ao genoma. Esse mesmo padrão de localização genômica foi observado para os lncRNAs de origem intrônica (Figura 25A). No geral, a maioria desses lncRNAs preditos foram localizados em regiões intergências, corroborando com outros estudos (ETEBARI; FURLONG; ASGARI, 2015; KHEMKA et al., 2016). Estima-se que as funções e localizações celulares de muitos lncRNAs estejam associadas à sua origem no genoma (CHEN, 2016).

Os lncRNAs apresentaram ainda poucos exons por transcrito (2-3) (Figura 25B), sendo a maioria desses transcritos entre 200 pb e 2.000 pb (Figura 25C). Dentre as características descritas em outros lncRNAs, essa faixa de tamanho encontrada é a que apresenta maior abundância dessas moléculas (WANG, F. et al., 2014) .

Finalmente, a maioria destes lncRNAs apresentaram os valores de expressão de FPKM altos quando comparados com valores de muitos mRNAs conhecidos (Figura 25D). Em conjunto esses dados corroboram com estudos que avaliaram as características de lncRNAs, caracterizando

com possíveis transcritos funcionais (KHEMKA et al., 2016; LEI et al., 2017; WU, Y. et al., 2016).

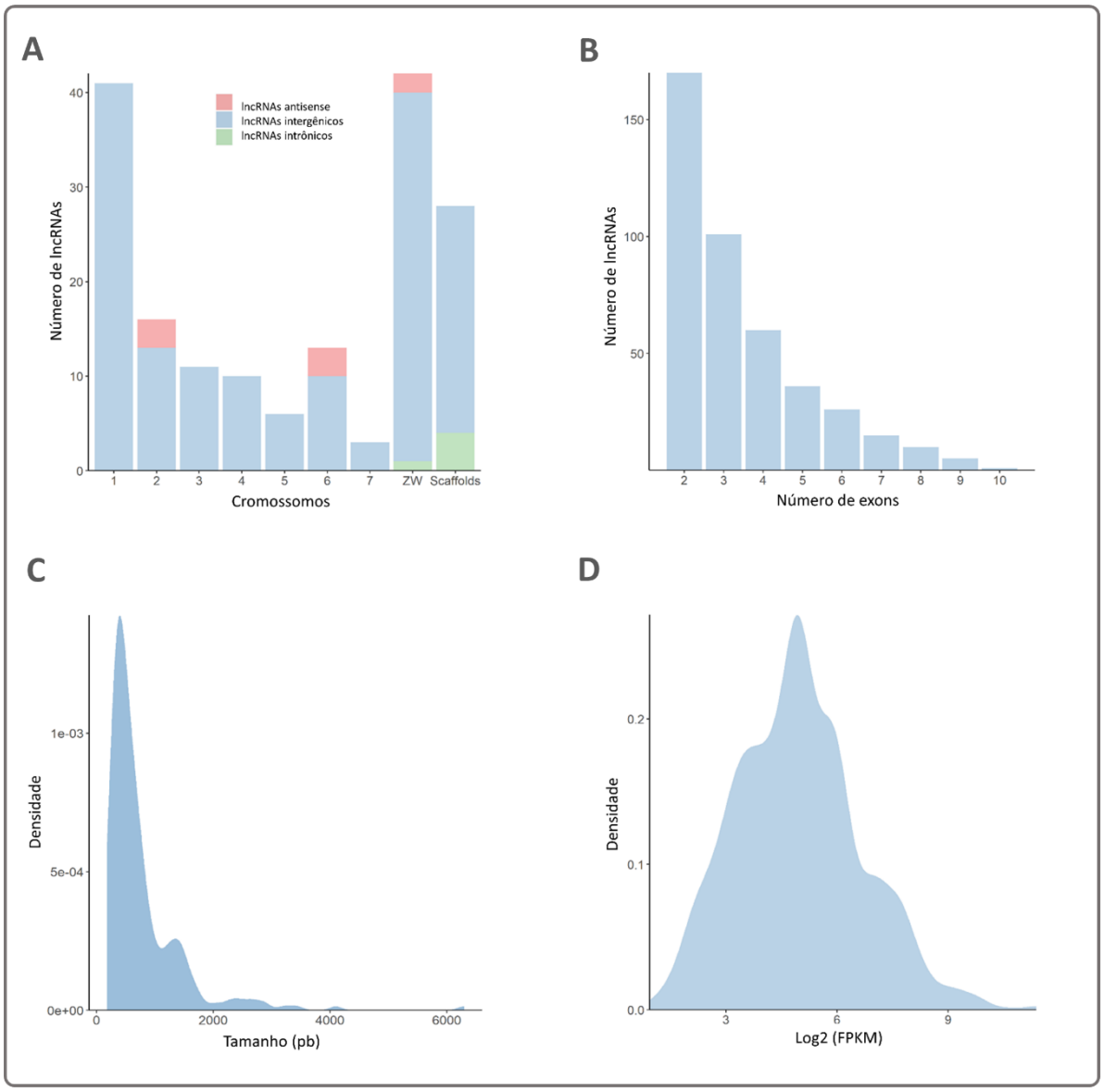


Figura 25: Características dos lncRNAs preditos no estágio de verme adulto em *S. mansoni*. (A) Localização genômica dos lncRNAs, (B) número de exons por transcrito, (C) tamanho dos transcritos, (D) expressão do log2 (FPKM).

5.2 Análises *in vitro*

5.2.1 Validação dos lncRNAs por qRT-PCR

Com o intuito de estender as análises dos 170 novos lncRNAs preditos, foram selecionados 15 Sm-lncRNAs e um retrotransposon (LTR) como referência para a expressão seguindo a metodologia de qRT-PCR. Para essas análises o gene *Sm-eIF4* foi utilizado como normalizador nos estágios de *S. mansoni* e *MmuHPRT1* nos fígados dos camundongos.

5.2.1.1 Expressão nos estágios de vermes adultos normais e resistentes ao praziquantel

O *S. mansoni* apresenta características típicas e interessantes, dentre elas se destacam: (i) faz parte da única classe de trematódeos com vida dioica; (ii) o pareamento permanente do macho com a fêmea é essencial para a maturação sexual do parasito fêmea e a produção contínua de ovos e (iii) possui a capacidade de viver durante décadas no hospedeiro mamífero humano (COLLEY et al., 2014; POPIEL; BASCH, 1984). Vários grupos demonstraram que diversos genes, envolvidos com as mais diversas vias metabólicas apresentam sua expressão aumentada ou diminuída em resposta a separação dos casais de parasitos, entretanto, o efeito sobre a expressão dos lncRNAs não é conhecido (ANDERSON et al., 2015; FITZPATRICK et al., 2005; GREVELDING; SOMMER; KUNZ, 1997).

A Figura 26 demonstra que os Sm-lncRNAs 1, 2, 4 foram mais expressos em fêmeas separadas em comparação ao casal ou parasitos machos e uma diminuição na expressão dos Sm-lncRNAs 3, 7, 9, 11, 13, 14 e 15 em fêmea. Já os Sm-lncRNAS 6, 8, 12 não foram afetados pela separação do casal. Vale ressaltar que apesar de abordagens genômicas e transcrissômicas serem aplicadas ao *S. mansoni* desde 1992, até o momento há somente uma publicação focada no tema lncRNAs (VASCONCELOS et al., 2017) e portanto, para efeitos de comparação dos dados obtidos nesse trabalho utilizamos os já descritos para mRNAs. Esses dados são corroborados e descritos por vários autores (ANDERSON et al., 2015; CAI et al., 2016; KINCAID-SMITH et al., 2018;

LU et al., 2017), reforçando a hipótese de que os mecanismos de regulação da expressão dos lncRNAs também são afetados pelo pareamento de forma similar ao já descrito para mRNAs. Eles também são corroborados por Vasconcelos (2017) e abrem perspectivas para um estudo aprofundado, visto que sugerem fortemente que o controle da expressão gênica envolvem lncRNAs em nível pós-transcricional. Este mecanismo é crucial em muitos organismos, e em *S. mansoni* também parece ser uma etapa regulatória fundamental.

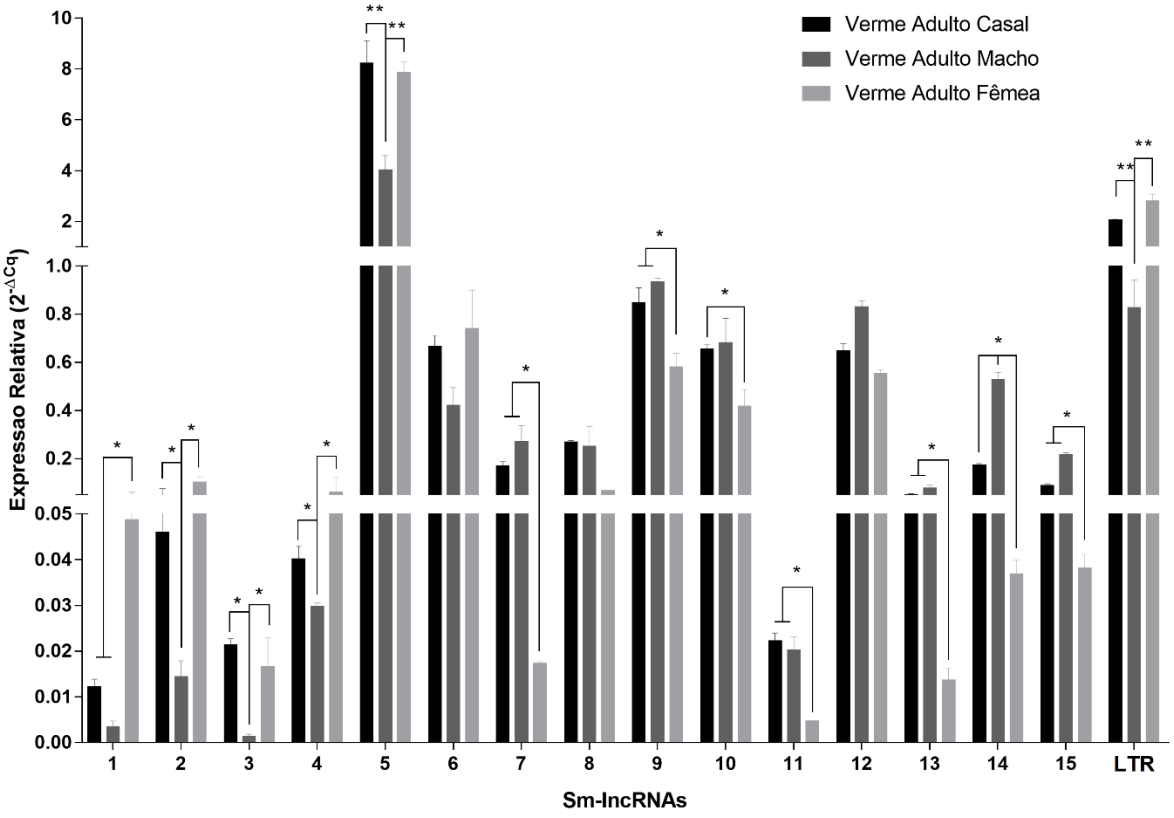


Figura 26: Expressão relativa dos 15 Sm-lncRNAs no estágio de verme adulto no *S. mansoni*. A expressão relativa da qRT-PCR foi determinada pelo método do $2^{-\Delta Cq}$, utilizando como gene constitutivo o *Sm-eIF4E*. O Retrotransposon (LTR) foi utilizado como medida comparativa na análise. As análises estatísticas foram realizadas utilizando o teste de ANOVA *one-way* com pós-teste de Tukey (**p-valor* ≤ 0.05; ***p-valor* ≤ 0.001).

Também foi nosso objetivo avaliar o perfil de expressão desse conjunto de lncRNA em parasitos resistentes ao praziquantel. Embora nas últimas décadas, uma série de fármacos tenham sido utilizados para o tratamento da esquistossomose, apenas o PZQ é amplamente empregado. Apesar de eficaz contra todas as espécies de esquistossomose que infectam humanos (HAGAN et al., 2004) e ser relativamente barato e fácil de usar, o PZQ não fornece uma cura, já que os esquistossômulos jovens são relativamente resistentes aos efeitos anti-helmínticos da droga. Embora o PZQ traga alívio para os pacientes tratados, os parasitos jovens que escapam da eliminação durante o tratamento, amadurecem e começam a liberar ovos. Este mecanismo de resistência é preocupante, pois sob essa pressão ineficaz, a resistência a medicamentos também pode surgir em humanos, como ocorre em modelo murino (ARAGON et al., 2009; GÖNNERT; ANDREWS, 1977; PICA-MATTOCCIA; CIOLI, 2004; SABAH et al., 1986; YOU; MCMANUS; GOBERT, 2015).

Como pode ser observado pelo resultado apresentado pela Figura 27, foi observado um perfil de expressão gênica similar, sendo os Sm-lncRNAs diferencialmente expressos em casal de parasitos e influenciados pelo pareamento dos vermes. Podemos observar um aumento significativo na expressão para os Sm-lncRNA 4, 5, 7, 8, 9, 13, 14, 15 e LTR nos parasitos machos em comparação aos casais e aos parasitos fêmeas. Vale ressaltar o aumento significativo na expressão do lncRNA 5 nos parasitos machos e fêmeas separados, em relação ao casal. Essas diferenças podem ser melhor avaliadas na Figura 28. A diminuição na expressão dos lncRNAs em comparação aos vermes sensíveis ao PZQ reforça a hipótese do envolvimento dessas moléculas em mecanismos regulatórios importantes para a sobrevivência do parasita. Até o momento não se conhece exatamente o efeito do PZQ na expressão gênica global de vermes adultos. Os dados da literatura são bem controversos, entretanto, Sanchez el al (2017) demonstraram que após quatro dias consecutivos de tratamento com PZQ houve um aumento significativo na expressão dos transportadores tipo ABC: ABCB1-1, B8, C1-1, C1-2, G1 e G2 em esquistossômulos jovens. Nenhum gene transportador ABC mostrou uma redução no nível de transcrição em vermes adultos expostos a PZQ por 4 dias consecutivos, reforçando a hipótese de que a ação principal do PZQ é sobre os esquistossômulos jovens. Desta forma, os próximos experimentos, consequentes desses resultados serão avaliar a expressão desse conjunto de lncRNAs e esquistossômulos jovens, recuperados de camundongos infectados após 25-28 de infecção, expostos ou não ao PZQ para uma melhor compreensão dos resultados obtidos até o momento.

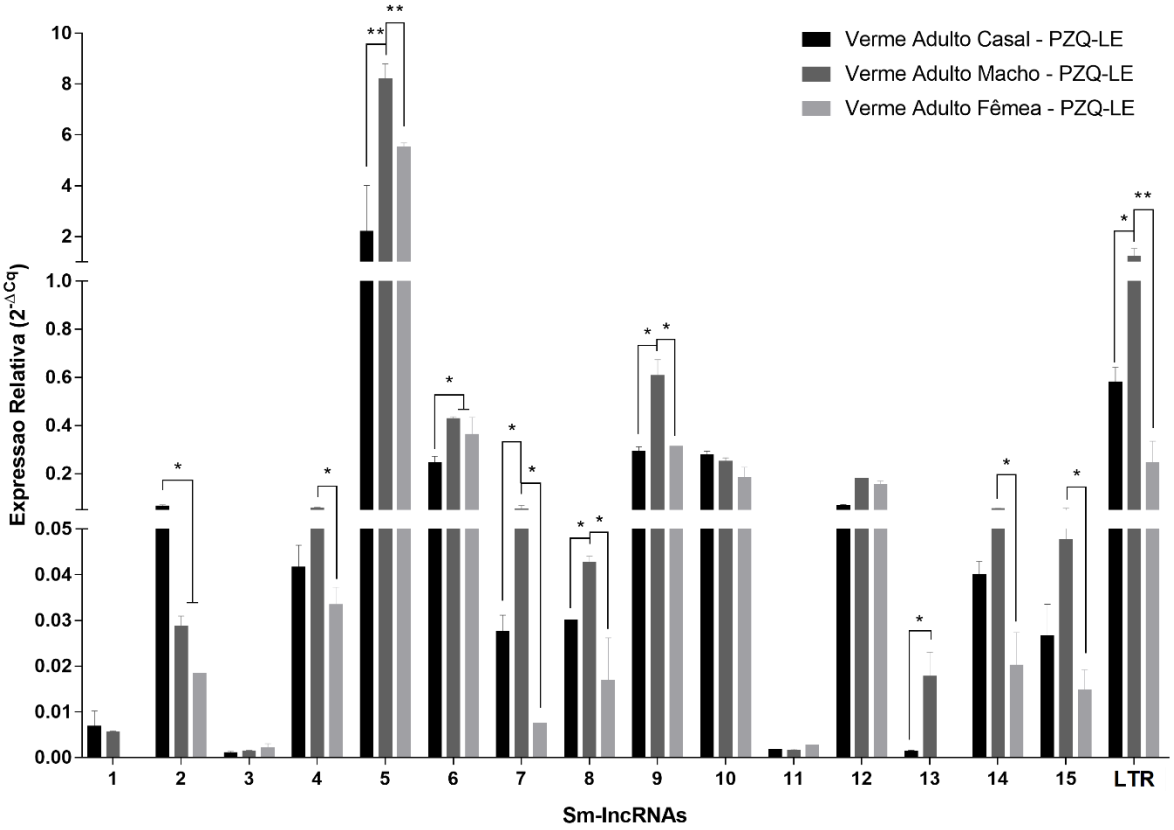


Figura 27: Expressão relativa dos 15 Sm-lncRNAs no estágio de verme adulto resistente ao praziquantel no *S. mansoni*. A expressão relativa da qRT-PCR foi determinada pelo método do $2^{-\Delta Cq}$, utilizando como gene constitutivo o *Sm-eIF4E*. O Retrotransposon (LTR) foi utilizado como medida comparativa. As análises estatísticas foram realizadas utilizando o teste de ANOVA *one-way* com pós-teste de Tukey (**p-valor* ≤ 0.05; ***p-valor* ≤ 0.001).

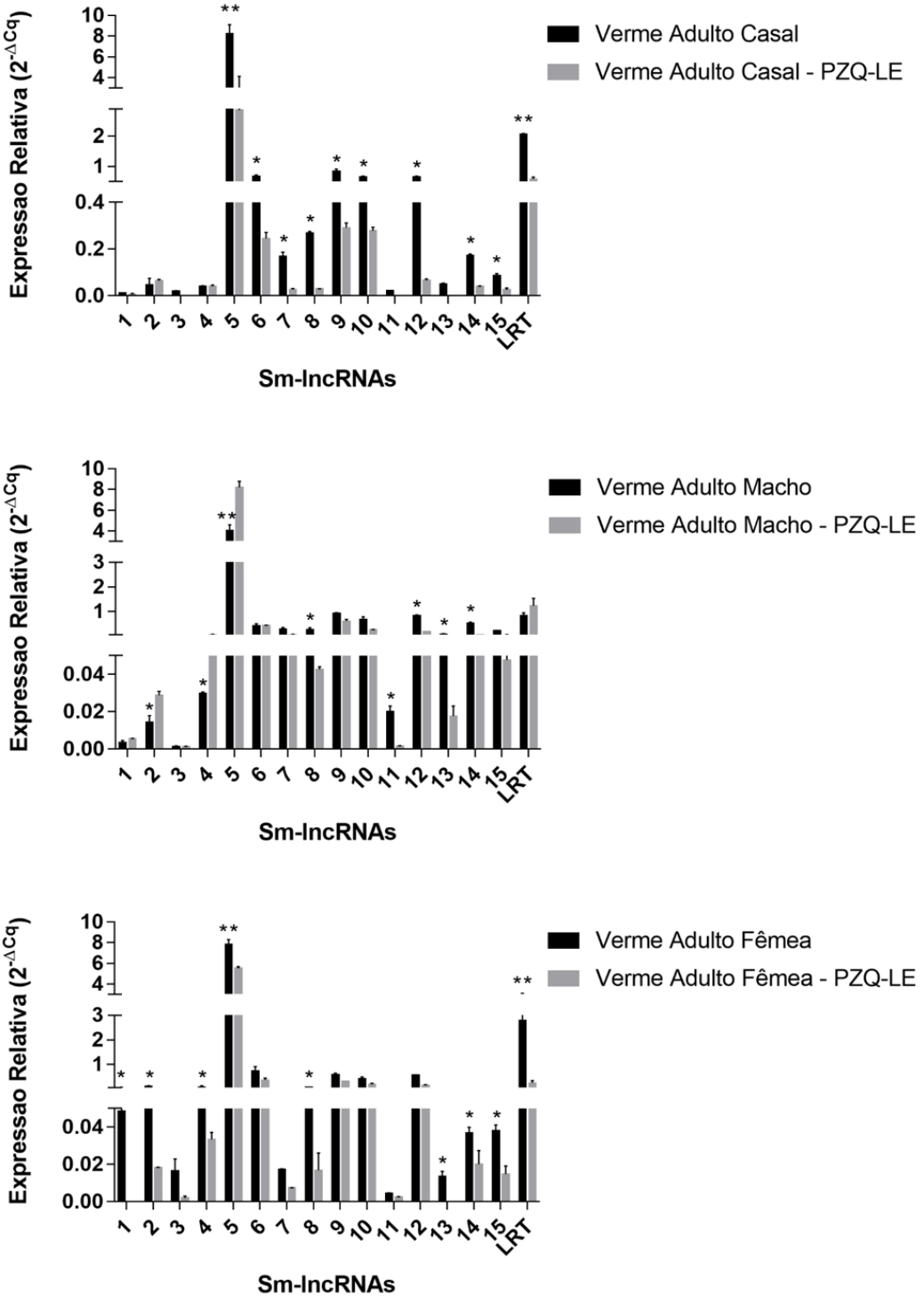


Figura 28: Expressão relativa dos 15 lncRNAs nos estágios de vermes adultos sensíveis e resistentes ao praziquantel no *S. mansoni*. A expressão relativa da qRT-PCR foi determinada pelo método do $2^{-\Delta Cq}$, utilizando como gene constitutivo o *Sm-eIF4E*. O Retrotransposon (LTR) foi utilizado como medida comparativa. As análises estatísticas foram realizadas utilizando o teste de ANOVA *one-way* com pós-teste de Tukey (**p-valor* ≤ 0.05; ***p-valor* ≤ 0.001).

5.2.1.2 Expressão nos estágios de cercária, esquistossômulos 3,5h e ovos

O *S. mansoni* apresenta um ciclo biológico com várias fases evolutivas e uma de suas características é apresentar um conjunto de genes estágio-específicos e/ou diferencialmente expressos (GROSSMAN et al., 1990). Para verificar se os Sm-lncRNAs identificados a partir de dados de RNA-seq de vermes adultos também poderiam ser expressos em outras fases evolutivas do parasito, nosso próximo objetivo foi avaliar a expressão em cercárias, esquistossômulos com 3,5h de cultivo *in vitro* e ovos. Vale ressaltar que nessa etapa do projeto decidimos focar nesses estágios pois, cercária é forma infectante da doença para o hospedeiro mamífero, esquistossômulos são as formas intermediárias ao estágio adulto do hospedeiro mamífero e como dito anteriormente a fase considerada mais importante nos mecanismos de resistência ao PZQ, e ovos devido a sua importância na patologia doença.

Como pode ser observado na Figura 29, todos os lncRNAs avaliados foram expressos em cercária, esquistossômulos e ovos. Também observamos um perfil de expressão diferencial para os Sm-lncRNAs 10, 12 e 13 que estão com níveis aumentados em cercária em relação a esquistossômulos e ovos. Foi também observado uma maior expressão dos lncRNA 1 e 6 em esquistossômulos, quando comparados com cercária e ovos. Todos os lncRNAs avaliados em ovos apresentaram uma baixa expressão quando comparados a esquistossômulos e cercária, sendo o Sm-lncRNA 11 não expresso em ovos. Os resultados apresentados nesse trabalho também são corroborados pelos descritos por Vasconcelos et al (2017) que descrevem um perfil de expressão diferencial para os lncRNAs durante o desenvolvimento de *S. mansoni*.

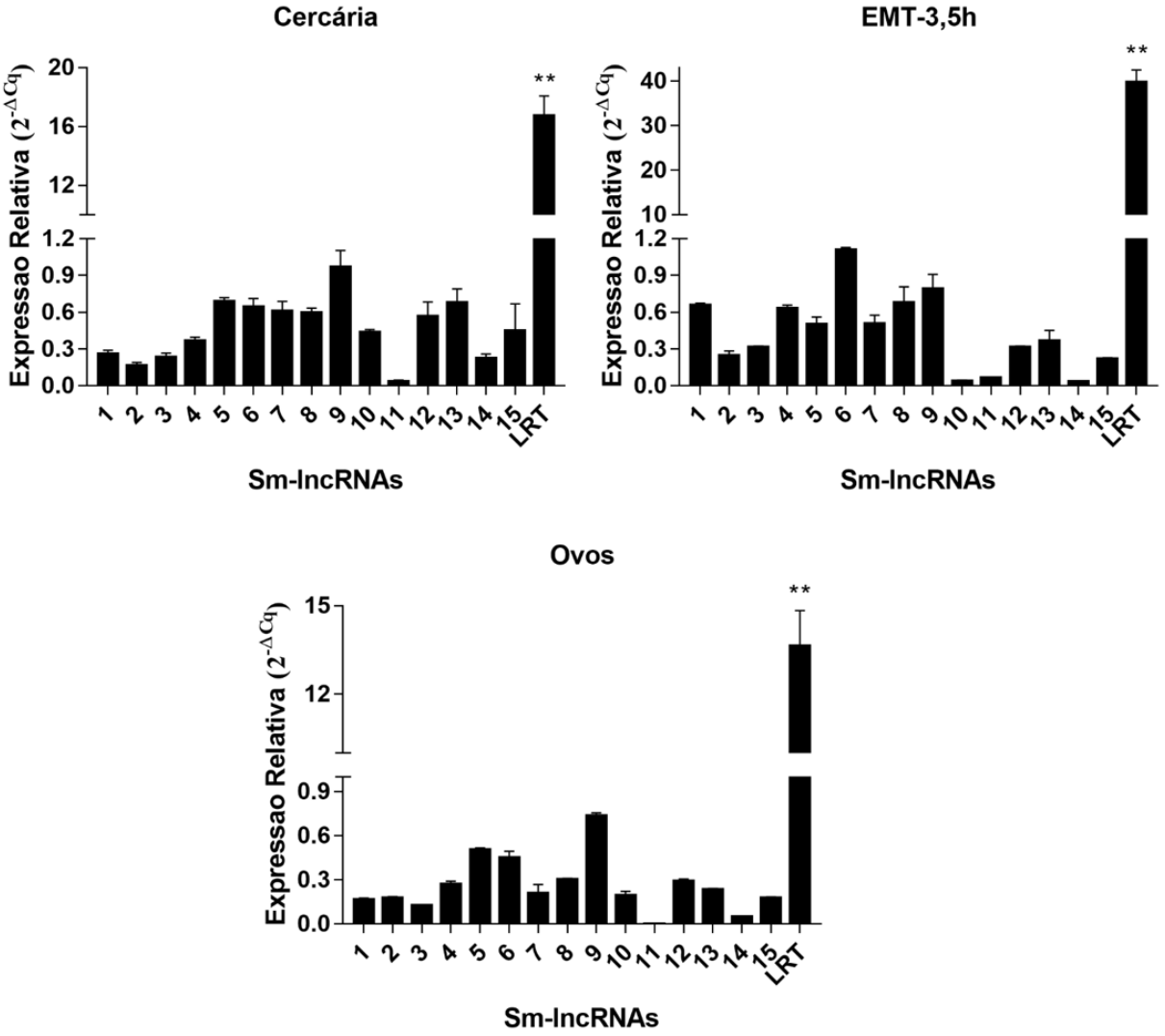


Figura 29: Expressão relativa dos 15 lncRNAs nos estágios de cercária, EMT-3,5h e ovos no *S. mansoni*. A expressão relativa da qRT-PCR foi determinada pelo método do $2^{-\Delta Cq}$, utilizando como gene constitutivo o *Sm-eIF4E*. O Retrotransposon (LTR) foi utilizado como medida comparativa. As análises estatísticas foram realizadas utilizando o teste de ANOVA *one-way* com pós-teste de Tukey. Diferente de todos os outros transcritos (*p-valor* ≤ 0.001).

5.2.1.3 Expressão em fígados de camundongos infectados e não infectados

Atualmente é consenso que os genes ortólogos de lncRNAs possuem uma baixa conservação, em termos de sequência de nucleotídeos, quando comparado os genes codificantes (ULITSKY; BARTEL, 2013). Por exemplo, 80% dos lncRNAs de ratos, incluindo os mais altamente expressos, não são detectados em eritroblastos humanos de estádios similares de desenvolvimento e a maioria dos lncRNAs de eritrócitos humanos não são detectados em camundongos. Além disso, nossas descobertas são consistentes com as descritas pelo Consórcio do Mouse ENCODE (MOUSE et al., 2012) em que, coletivamente, os ncRNA são muito mais específicos de espécies do que os genes codificantes.

Como descrito anteriormente, uma busca por homologia foi realizada utilizando a plataforma RNAcentral sugeriu que os lncRNAs identificados nesse trabalho são espécie-específico. Para testar essa hipótese, analisamos a expressão dos 15 Sm-lncRNAs em estudo em fígados de camundongos C57BL/6 não infectados. Como podemos observar na Tabela 10, não foi detectado expressão, reforçando a hipótese de que esse conjunto de candidatos a lncRNAs seria específico do gênero *Schistosoma*. Nessa tabela, todos os valores de Cq dos 15 Sm-lncRNAs foram maiores que 37 ou não detectados, caracterizando ruídos a ausência de expressão. Somente o gene utilizado como normalizador *MumHPRT1* foi expresso nesse experimento.

Continuando essa investigação, nós avaliamos a expressão desse conjunto de Sm-lncRNAs em fígado de animais C57BL/6 infectados posteriormente com 100 cercárias e eutanasiados após 7 semanas de infecção. Como pode ser observado na Figura 30, foi possível detectar a expressão dos seguintes Sm-lncRNAs: 2, 3, 7, 9, 10, 11 e 13. O LTR foi o mais expresso comparado com os demais Sm-lncRNAs. A detecção de lncRNAs do parasito no hospedeiro, abre grandes perspectivas para o melhor entendimento do papel dos lncRNAs nas respostas metabólicas consequentes da interação parasito-hospedeiro.

Tabela 10: Valores brutos de Cq da expressão dos 15 Sm-lncRNAs em fígados de camundongo não infectados.

Sm-lncRNAs	**WTI1**	**WTI2**	**WTI3**
Sm-lncRNA1	0	0	0
Sm-lncRNA2	0	0	0
Sm-lncRNA3	0	0	0
Sm-lncRNA4	37,26	37,653	38,95
Sm-lncRNA5	0	0	0
Sm-lncRNA6	0	0	0
Sm-lncRNA7	0	38,061	0
Sm-lncRNA8	0	0	0
Sm-lncRNA9	39,15	0	39,65
Sm-lncRNA10	38,47	38,124	38,30
Sm-lncRNA11	0	0	39,82
Sm-lncRNA12	0	0	0
Sm-lncRNA13	0	0	0
Sm-lncRNA14	0	0	0
Sm-lncRNA15	38,2	39,258	39,2
LTR	0	0	0
MumHPRT1	32,12	32,347	30,56

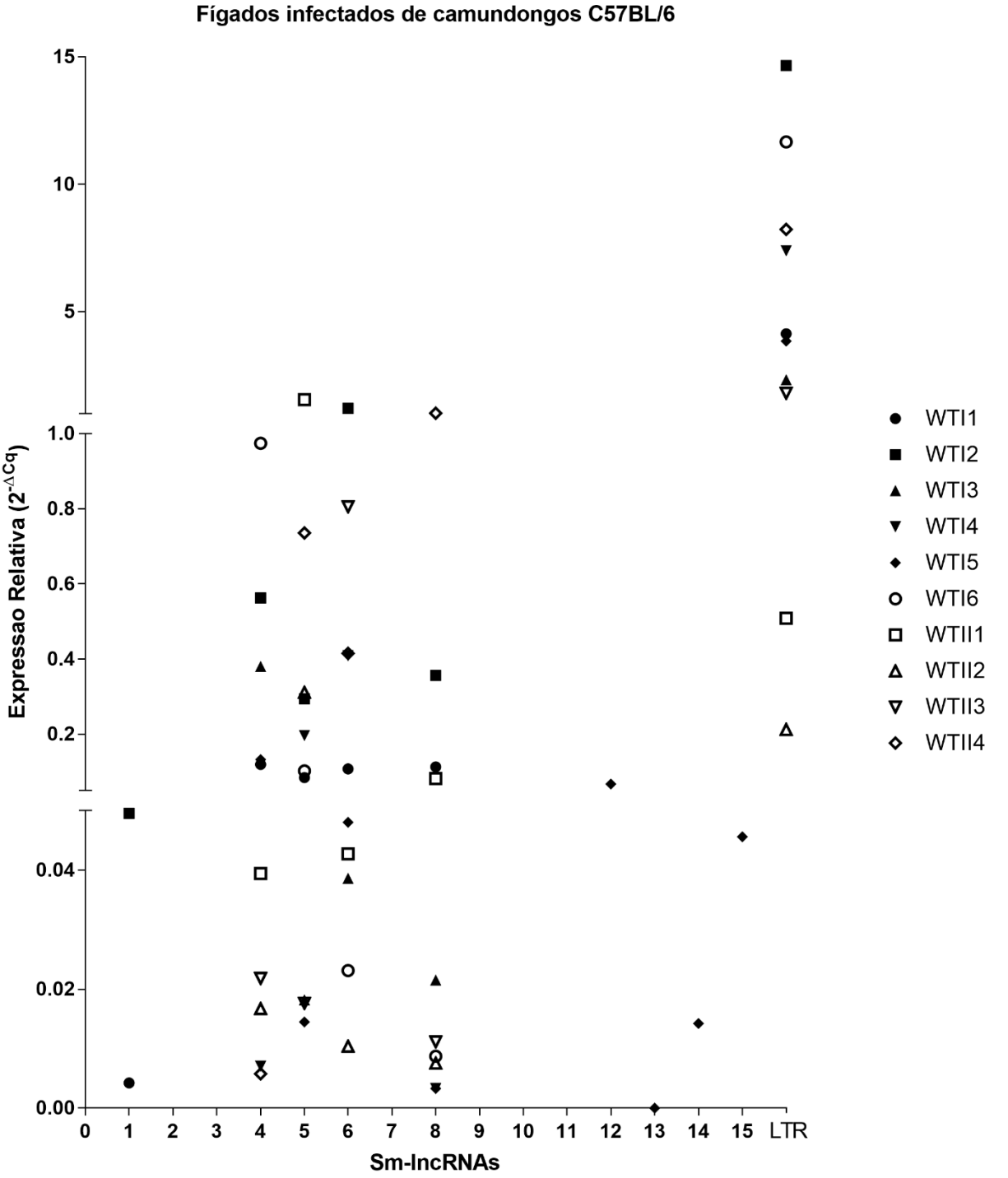


Figura 30: Expressão relativa dos 15 Sm-lncRNAs em fígados de camundongos C57BL/6 infectados. A expressão relativa da qRT-PCR foi determinada pelo método do $2^{-\Delta Cq}$, utilizando como gene constitutivo o *MmuHPRT1*. O Retrotransposon (LTR) foi utilizado como medida comparativa. As análises estatísticas foram realizadas utilizando o teste de ANOVA *one-way* com pós-teste de Tukey (*p-valor* ≤ 0.001). Estatística não representada no gráfico.

Durante muitos anos, os pesquisadores se concentraram no estudo de proteínas dos parasitos, que podem ser consideradas o primeiro nível de comunicação do parasita com o hospedeiro, seja com componentes do hospedeiro na fase aguda ou com receptores expostos nas superfícies das células hospedeiras. Desta forma, as proteínas foram consideradas como as principais efetoras dos mecanismos de evasão do sistema imune do hospedeiro. Entretanto, o desenvolvimento recente das estratégias mais robustas para o sequenciamento e análise dos dados de transcrição, as moléculas de ncRNA vem tomando destaque como reguladores centrais em doenças infecciosas. Um bom exemplo disso, são os parasitos protozoários, como os do gênero *Plasmodium*, *Toxoplasma*, *Leishmania* e *Trypanosoma*. Cada um desses parasitos utiliza mecanismos de evasão específicos para evitar serem mortos pelo sistema de defesa do hospedeiro (BAYER-SANTOS; MARINI; DA SILVEIRA, 2017). Um número crescente de estudos vem mostrando que esses agentes patogênicos podem transferir moléculas de ncRNA para as células hospedeiras para modular suas funções. Esta transferência geralmente ocorre através de vesículas extracelulares, que são pequenas vesículas de membrana segregadas pelo microrganismo (ATAYDE; TSCHUDI; ULLU, 2011; DAROCHA et al., 2004; ROBINSON, K. A.; BEVERLEY, 2003; STOCO et al., 2014; ZHENG; CAI; BRADLEY, 2013).

Está cada vez mais evidente que essas vesículas extracelulares, particularmente os exossomas, desempenham um papel fundamental na comunicação celular. Os exossomas são nanovesículas em torno de 50-100 nm de tamanho que são segregados por praticamente todas as células para facilitar a transferência, principalmente de lipídios, proteínas e espécies de RNA, mantendo os marcadores fenotípicos de sua célula de origem (ALLMANG et al., 1999; MITCHELL et al., 1997). Os exossomas se desenvolvem dentro de uma célula por brotação interna de endossomas multi-vesiculares e, portanto, contêm componentes da célula parental, como RNAs ou proteínas, que podem ser traficados para o mesmo compartimento. A descoberta de vesículas extracelulares de cinetoplastídeos, fungos e bactérias levou a teoria de que a comunicação mediada pelo exossomo poderia operar em uma plataforma de espécies cruzadas, pelo qual os exossomos derivados de parasitos poderiam interagir e potencialmente modular o sistema imune do hospedeiro (COAKLEY; MAIZELS; BUCK, 2015; COLOMBO; RAPOSO; THERY, 2014; MARCILLA et al., 2014; RAPOSO; STOORVOGEL, 2013).

Somente recentemente, os exossomos foram reconhecidos como produtos integrantes de organismos extracelulares como helmintos. Isto foi inicialmente relatado nos componentes excêntricos-secretórios dos trematódeos, *Echinostoma caproni* e *Fasciola hepatica*, que infectam o trato gastrointestinal e fígado, respectivamente (MARCILLA et al., 2012) e no nematódeo *Heligmosomoides polygyrus*, que infecta o intestino delgado (BUCK et al., 2014). Os dados dos estudos de trematódeo sugerem ainda que os exossomos derivados desses excêntricos-secretórios são capazes de atingir o ambiente do hospedeiro, pois parecem ser encontrados intactos no tegumento dos parasitos. Um suporte adicional disso é demonstrado pela internalização de exossomos de helmintos pelas células epiteliais intestinais do hospedeiro, sugerindo a capacidade de comunicação entre helmintos e mamíferos. Em 2016, Sotillo e colaboradores demonstraram que os vermes adultos de *S. mansoni* secretam vesículas extracelulares semelhantes ao exossomo, com um tamanho variando de 50 a 130 nm (SOTILLO et al., 2016). A análise proteômica dessas vesículas extracelulares mostrou inúmeros candidatos vacinais conhecidos, potenciais fatores de virulência e moléculas implicadas na alimentação. Entretanto o componente RNA presente nessas vesículas ainda permanece desconhecido (SOTILLO et al., 2016).

Além disso, o granuloma hepático é tanto uma forma de proteção hepática como a principal causa da patogênese da esquistossomose, exemplificando a complexidade da relação parasita - hospedeiro na infecção por *S. mansoni* (GRYSEELS et al., 2006; MORAIS et al., 2008; WILSON, M. S. et al., 2007; WYNN et al., 2004). Assim, baseados nos resultados apresentados na Figura 30, concluímos que a expressão dos lncRNAs de *S. mansoni* detectada no fígado de animais infectados está relacionada com a presença dos ovos. Até o momento, não existem relatos da presença de RNAs associados aos antígenos solúveis dos ovos, entretanto, essa possibilidade não pode ser descartada, principalmente após os relatos da presença de vesículas extracelulares semelhantes ao exossomo em *S. mansoni* (SOTILLO et al., 2016). Novos experimentos serão realizados para investigar essa possibilidade.

6 CONCLUSÃO

Foi realizado um estudo sistemático e abrangente que identificou um novo conjunto de 170 novos lncRNAs a partir de dados de RNA-seq em *S. mansoni*. A relevância funcional foi sugerida por análise de enriquecimento de termos do GO. Foram ainda validados por expressão 15 lncRNAs, confirmando a existência desse conjunto de moléculas, bem como, a expressão gênica diferencial em diversos estágios do parasita, e o efeito do pareamento dos vermes. O efeito do praziquantel e sua detecção em fígado de camundongos infectados foi avaliado, sugerindo que essas moléculas possuem aplicabilidade biotecnológica podendo servir como biomarcadores alternativos e ou metas de tratamento moleculares para a esquistossomose.

8 REFERÊNCIAS

ALI, P. O. et al. Sequence of a small subunit rRNA gene of Schistosoma mansoni and its use in phylogenetic analysis. **Mol Biochem Parasitol,** v. 46, n. 2, p. 201-8, Jun 1991.

ALLMANG, C. et al. The yeast exosome and human PM-Scl are related complexes of 3' --> 5' exonucleases. **Genes Dev,** v. 13, n. 16, p. 2148-58, Aug 15 1999.

ALMEIDA, G. T. et al. Exploring the Schistosoma mansoni adult male transcriptome using RNA-seq. **Exp Parasitol,** v. 132, n. 1, p. 22-31, Sep 2012.

ALTSCHUL, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Res,** v. 25, n. 17, p. 3389-402, Sep 1 1997.

AMIT-AVRAHAM, I. et al. Antisense long noncoding RNAs regulate var gene activation in the malaria parasite Plasmodium falciparum. **Proc Natl Acad Sci U S A,** v. 112, n. 9, p. E982-91, Mar 3 2015.

ANASTASIOU, E. et al. Prehistoric schistosomiasis parasite found in the Middle East. **The Lancet Infectious Diseases,** v. 14, n. 7, p. 553-554, 2014.

ANDERSON, L. et al. Schistosoma mansoni Egg, Adult Male and Female Comparative Gene Expression Analysis and Identification of Novel Genes by RNA-Seq. **PLoS Negl Trop Dis,** v. 9, n. 12, p. e0004334, Dec 2015.

ANDREWS, S. FastQC: a quality control tool for high throughput sequence data. v. Disponível online em http://www.bioinformatics.babraham.ac.uk/projects/fastqc, 2010.

ARAGON, A. D. et al. Towards an understanding of the mechanism of action of praziquantel. **Mol Biochem Parasitol,** v. 164, n. 1, p. 57-65, Mar 2009.

ATAYDE, V. D.; TSCHUDI, C.; ULLU, E. The emerging world of small silencing RNAs in protozoan parasites. **Trends Parasitol,** v. 27, n. 7, p. 321-7, Jul 2011.

BASCH, P. F. Cultivation of Schistosoma mansoni in vitro. I. Establishment of cultures from cercariae and development until pairing. **J Parasitol,** v. 67, n. 2, p. 179-85, Apr 1981.

BAYER-SANTOS, E.; MARINI, M. M.; DA SILVEIRA, J. F. Non-coding RNAs in Host-Pathogen Interactions: Subversion of Mammalian Cell Functions by Protozoan Parasites. **Front Microbiol,** v. 8, p. 474, 2017.

BAZZINI, A. A. et al. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. **EMBO J,** v. 33, n. 9, p. 981-93, May 2 2014.

BEAVER, P. C., R. C. JUNG, AND E. W. CUPP. . Examination of specimens for parasites. **Clinical parasitology**, n. 9th ed. Lea & Febiger, p. p. 733-758

1984.

BERRIMAN, M. et al. The genome of the blood fluke Schistosoma mansoni. **Nature,** v. 460, n. 7253, p. 352-8, Jul 16 2009.

BLIGNAUT, M. Review of Non-coding RNAs and the epigenetic regulation of gene expression: A book edited by Kevin Morris. **Epigenetics,** v. 7, n. 6, p. 664-666, 2012.

BLOOM, J. S. et al. Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. **BMC Genomics,** v. 10, p. 221, 2009.

BOERNER, S.; MCGINNIS, K. M. Computational identification and functional predictions of long noncoding RNA in Zea mays. **PLoS One,** v. 7, n. 8, p. e43047, 2012.

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

BOROS, D. L. Immunopathology of Schistosoma mansoni infection. **Clin Microbiol Rev,** v. 2, n. 3, p. 250-69, Jul 1989.

BOYD, B. **The Last Goodbye**. The Hobbit: The Battle of the Five Armies: Decca Records 2014.

BRASIL. **Guia de Vigilância Epidemiológica**. Ministério da Saúde. Secretaria de Vigilância em Saúde. 7.ed. 2009.

_____. **Vigilância da Esquistossomose Mansoni: diretrizes técnicas**. Brasília, DF: Ministério da Saúde, 2014. 144

BROADBENT, K. M. et al. A global transcriptional analysis of Plasmodium falciparum malaria reveals a novel family of telomere-associated lncRNAs. **Genome Biol,** v. 12, n. 6, p. R56, 2011.

BUCK, A. H. et al. Exosomes secreted by nematode parasites transfer small RNAs to mammalian cells and modulate innate immunity. **Nat Commun,** v. 5, p. 5488, Nov 25 2014.

CABILI, M. N. et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. **Genes Dev,** v. 25, n. 18, p. 1915-27, Sep 15 2011.

CAI, P. et al. Comprehensive Transcriptome Analysis of Sex-Biased Expressed Genes Reveals Discrete Biological and Physiological Features of Male and Female Schistosoma japonicum. **PLoS Negl Trop Dis,** v. 10, n. 4, p. e0004684, Apr 2016.

CAI, P. et al. A deep analysis of the small non-coding RNA population in Schistosoma japonicum eggs. **PLoS One,** v. 8, n. 5, p. e64003, 2013.

CAMACHO, C. et al. BLAST+: architecture and applications. **BMC Bioinformatics,** v. 10, p. 421, Dec 15 2009.

CAO, J. The functional role of long non-coding RNAs and epigenetics. **Biol Proced Online,** v. 16, p. 11, 2014.

CARNINCI, P. et al. The transcriptional landscape of the mammalian genome. **Science,** v. 309, n. 5740, p. 1559-63, Sep 2 2005.

CARPENTER, S. et al. A long noncoding RNA mediates both activation and repression of immune response genes. **Science,** v. 341, n. 6147, p. 789-92, Aug 16 2013.

CASTRO-BORGES, W. D. **Diversidade de proteassoma 20S e perfil de ubiquitinação durante o desevolvimento do parasita Schistosoma mansoni: uma abordagem proteômica**. 2005.

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 27 2014.

CHEN, L. L. Linking Long Noncoding RNA Localization and Function. **Trends Biochem Sci,** v. 41, n. 9, p. 761-772, Sep 2016.

CHEN, L. L.; CARMICHAEL, G. G. Decoding the function of nuclear long non-coding RNAs. **Curr Opin Cell Biol,** v. 22, n. 3, p. 357-64, Jun 2010.

CHU, Y.; COREY, D. R. RNA sequencing: platform selection, experimental design, and data interpretation. **Nucleic Acid Ther,** v. 22, n. 4, p. 271-4, Aug 2012.

CLAMP, M. et al. Distinguishing protein-coding and noncoding genes in the human genome. **Proc Natl Acad Sci U S A,** v. 104, n. 49, p. 19428-33, Dec 4 2007.

CLEGG, J. A. In Vitro Cultivation of Schistosoma Mansoni. **Exp Parasitol,** v. 16, p. 133-47, Apr 1965.

COAKLEY, G.; MAIZELS, R. M.; BUCK, A. H. Exosomes and Other Extracellular Vesicles: The New Communicators in Parasite Infections. **Trends Parasitol,** v. 31, n. 10, p. 477-489, Oct 2015.

COCK, P. J. et al. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. **Nucleic Acids Res,** v. 38, n. 6, p. 1767-71, Apr 2010.

COLLEY, D. G. et al. Human schistosomiasis. **Lancet (London, England),** v. 383, n. 9936, p. 2253-2264, 04/01 2014.

COLOMBO, M.; RAPOSO, G.; THERY, C. Biogenesis, secretion, and intercellular interactions of exosomes and other extracellular vesicles. **Annu Rev Cell Dev Biol,** v. 30, p. 255-89, 2014.

COPELAND, C. S. et al. Homology-based annotation of non-coding RNAs in the genomes of Schistosoma mansoni and Schistosoma japonicum. **BMC Genomics,** v. 10, p. 464, Oct 8 2009.

COUTO, F. F. et al. Schistosoma mansoni: a method for inducing resistance to praziquantel using infected Biomphalaria glabrata snails. **Mem Inst Oswaldo Cruz,** v. 106, n. 2, p. 153-7, Mar 2011.

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.

DE SOUZA GOMES, M. **Caracterização inicial dos constituintes da maquinaria de silenciamento gênico mediada por microRNAs em Schistosoma mansoni**. 2008. 122 (Mestrado em Ciências Biológicas). NUPEB, Universidade Federal de Minas Gerais

DE SOUZA GOMES, M. et al. Genome-wide identification of novel microRNAs and their target genes in the human parasite Schistosoma mansoni. **Genomics,** v. 98, n. 2, p. 96-111, Aug 2011.

DERRIEN, T. et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. **Genome Res,** v. 22, n. 9, p. 1775-89, Sep 2012.

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.

DJEBALI, S. et al. Landscape of transcription in human cells. **Nature,** v. 489, n. 7414, p. 101-8, Sep 6 2012.

DOBIN, A. et al. STAR: ultrafast universal RNA-seq aligner. **Bioinformatics,** v. 29, n. 1, p. 15-21, Jan 1 2013.

DOS SANTOS CARVALHO, O.; JANNOTTI-PASSOS, L. K.; CALDEIRA, R. L. Importância Epidemiológica e Biologia Molecular Aplicada ao Estudo dos Moluscos do Gênero Biomphalaria. **Schistosoma mansoni & Esquistossomose: uma visão multidisciplinar**, p. 309, 2008.

ENGSTROM, P. G. et al. Systematic evaluation of spliced alignment programs for RNA-seq data. **Nat Methods,** v. 10, n. 12, p. 1185-91, Dec 2013.

ETEBARI, K.; FURLONG, M. J.; ASGARI, S. Genome wide discovery of long intergenic non-coding RNAs in Diamondback moth (Plutella xylostella) and their expression in insecticide resistant strains. **Scientific Reports,** v. 5, p. 14642, 09/28/online 2015.

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

EWING, B. et al. Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment. **Genome Res,** v. 8, n. 3, p. 175-185, // 1998.

FALCÃO, E. C. Pirajá da Silva : o incontestável descobridor do “schistosoma mansoni”. v. Série I. , n. História da Saúde no Brasil, p. XX p. : il., 2008.

FITZPATRICK, J. M. et al. An oligonucleotide microarray for transcriptome analysis of Schistosoma mansoni and its application/use to investigate gender-associated gene expression. **Mol Biochem Parasitol,** v. 141, n. 1, p. 1-13, May 2005.

FRANCO, G. R. et al. Identification of new Schistosoma mansoni genes by the EST strategy using a directional cDNA library. **Gene,** v. 152, n. 2, p. 141-7, Jan 23 1995.

FRANCO, G. R. et al. Evaluation of cDNA libraries from different developmental stages of Schistosoma mansoni for production of expressed sequence tags (ESTs). **DNA Res,** v. 4, n. 3, p. 231-40, Jun 30 1997.

FREITAS, C. A. Situação atual da esquistossomose no Brasil. **Revista Brasileira de Malariologia e Doenças Tropicais,** v. 24, p. 03-63, 1972.

FRITH, M. C. et al. Discrimination of non-protein-coding transcripts from protein-coding mRNA. **RNA Biol,** v. 3, n. 1, p. 40-8, Jan-Mar 2006.

GASCOIGNE, D. K. et al. Pinstripe: a suite of programs for integrating transcriptomic and proteomic datasets identifies novel proteins and improves differentiation of protein-coding and non-coding genes. **Bioinformatics,** v. 28, n. 23, p. 3042-50, Dec 1 2012.

GÖNNERT, R.; ANDREWS, P. Praziquantel, a new broad-spectrum antischistosomal agent. **Zeitschrift für Parasitenkunde,** v. 52, n. 2, p. 129-150, January 01 1977.

GREVELDING, C. G.; SOMMER, G.; KUNZ, W. Female-specific gene expression in Schistosoma mansoni is regulated by pairing. **Parasitology,** v. 115, n. 6, p. 635-640, 1997.

GROSSMAN, Z. et al. Schistosoma mansoni: Stage-specific expression of muscle-specific genes. **Experimental Parasitology,** v. 70, n. 1, p. 62-71, 1990/01/01/ 1990.

GRYSEELS, B. et al. Human schistosomiasis. **Lancet,** v. 368, n. 9541, p. 1106-18, Sep 23 2006.

GUTTMAN, M. et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. **Nature,** v. 458, n. 7235, p. 223-7, Mar 12 2009.

HAAS, W. et al. Recognition and invasion of human skin by Schistosoma mansoni cercariae: the key-role of L-arginine. **Parasitology,** v. 124, n. Pt 2, p. 153-67, Feb 2002.

HAGAN, P. et al. Schistosomiasis control: keep taking the tablets. **Trends Parasitol,** v. 20, n. 2, p. 92-7, Feb 2004.

HAN, Z. G. et al. Schistosoma genomics: new perspectives on schistosome biology and host-parasite interaction. **Annu Rev Genomics Hum Genet,** v. 10, p. 211-40, 2009.

HARROP, R.; WILSON, R. A. Protein synthesis and release by cultured schistosomula of Schistosoma mansoni. **Parasitology,** v. 107 (Pt 3), p. 265-74, Sep 1993.

HINTON, J. C. et al. Benefits and pitfalls of using microarrays to monitor bacterial gene expression during infection. **Curr Opin Microbiol,** v. 7, n. 3, p. 277-82, Jun 2004.

HOU, M. et al. AnnoLnc: a web server for systematically annotating novel human lncRNAs. **BMC Genomics,** v. 17, n. 1, p. 931, Nov 16 2016.

IHAKA, R.; GENTLEMAN, R. R: A Language for Data Analysis and Graphics. **Journal of Computational and Graphical Statistics,** v. 5, n. 3, p. 299-314, 1996.

ILOTT, N. E.; PONTING, C. P. Predicting long non-coding RNAs using RNA sequencing. **Methods,** v. 63, n. 1, p. 50-9, Sep 1 2013.

JUNTAWONG, P. et al. Translational dynamics revealed by genome-wide profiling of ribosome footprints in Arabidopsis. **Proc Natl Acad Sci U S A,** v. 111, n. 1, p. E203-12, Jan 7 2014.

KAPUSTA, A. et al. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. **PLoS Genet,** v. 9, n. 4, p. e1003470, Apr 2013.

KARLSSON, O.; BACCARELLI, A. A. Environmental Health and Long Non-coding RNAs. **Curr Environ Health Rep,** v. 3, n. 3, p. 178-87, Sep 2016.

KASHI, K. et al. Discovery and functional analysis of lncRNAs: Methodologies to investigate an uncharacterized transcriptome. **Biochim Biophys Acta,** v. 1859, n. 1, p. 3-15, Jan 2016.

KHEMKA, N. et al. Genome-wide analysis of long intergenic non-coding RNAs in chickpea and their potential role in flower development. **Scientific Reports,** v. 6, p. 33297, 09/15

05/19/received

08/22/accepted 2016.

KIM, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. **Genome Biol,** v. 14, n. 4, p. R36, Apr 25 2013.

KINCAID-SMITH, J. et al. Parent-of-Origin Dependent Gene Expression in Male and Female Schistosome Parasites. **Genome Biol Evol**, Feb 12 2018.

KOLAROVA, L. Schistosomes causing cercarial dermatitis: a mini-review of current trends in systematics and of host specificity and pathogenicity. **Folia Parasitol (Praha),** v. 54, n. 2, p. 81-7, Jun 2007.

LAWTON, S., HIRAI, HIROHISA, IRONSIDE, JOE,; JOHNSTON, D.; ROLLINSON, D. Genomes and geography: genomic insights into the evolution and phylogeography of the genus Schistosoma. **Parasites & Vectors,** v. 4, n. 1, p. 131, 2011.

LEI, D. et al. Genome-Wide Analysis of mRNA and Long Noncoding RNA Profiles in Chronic Actinic Dermatitis. **Biomed Res Int,** v. 2017, p. 7479523, 2017.

LEVIN, J. Z. et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. **Nat Methods,** v. 7, n. 9, p. 709-15, Sep 2010.

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

LI, H. et al. Genome-wide long non-coding RNA screening, identification and characterization in a model microorganism Chlamydomonas reinhardtii. **Sci Rep,** v. 6, p. 34109, Sep 23 2016.

LI, T. et al. Identification of long non-protein coding RNAs in chicken skeletal muscle using next generation sequencing. **Genomics,** v. 99, n. 5, p. 292-8, May 2012.

LI, Z.; RANA, T. M. Molecular mechanisms of RNA-triggered gene silencing machineries. **Acc Chem Res,** v. 45, n. 7, p. 1122-31, Jul 17 2012.

LIAO, Q. et al. ncFANs: a web server for functional annotation of long non-coding RNAs. **Nucleic Acids Res,** v. 39, n. Web Server issue, p. W118-24, Jul 2011.

LIU, F. et al. New perspectives on host-parasite interplay by comparative transcriptomic and proteomic analyses of Schistosoma japonicum. **PLoS Pathog,** v. 2, n. 4, p. e29, Apr 2006.

LU, Z. et al. A gene expression atlas of adult Schistosoma mansoni and their gonads. **Sci Data,** v. 4, p. 170118, Aug 22 2017.

LUO, H. et al. Identification and function annotation of long intervening noncoding RNAs. **Brief Bioinform,** v. 18, n. 5, p. 789-797, Sep 1 2017.

LUTZ, A. O. Schistosomum mansoni e a schistosomose segundo observações feitas no Brasil. **Mem Inst Oswaldo Cruz**, n. 11, p. 121-144, 1919.

MANSON, P. Report of a Case of Bilharzia from the West Indies. **Br Med J,** v. 2, n. 2190, p. 1894-5, Dec 20 1902.

MARCILLA, A. et al. Extracellular vesicles in parasitic diseases. **J Extracell Vesicles,** v. 3, p. 25040, 2014.

MARCILLA, A. et al. Extracellular vesicles from parasitic helminths contain specific excretory/secretory proteins and are internalized in intestinal host cells. **PLoS One,** v. 7, n. 9, p. e45974, 2012.

MARIONI, J. C. et al. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. **Genome Res,** v. 18, n. 9, p. 1509-17, Sep 2008.

MCKERROW, J. H.; SALTER, J. Invasion of skin by Schistosoma cercariae. **Trends Parasitol,** v. 18, n. 5, p. 193-5, May 2002.

MERCER, T. R. et al. Specific expression of long noncoding RNAs in the mouse brain. **Proc Natl Acad Sci U S A,** v. 105, n. 2, p. 716-21, Jan 15 2008.

MIN, X. J. et al. OrfPredictor: predicting protein-coding regions in EST-derived sequences. **Nucleic Acids Res,** v. 33, n. Web Server issue, p. W677-80, Jul 1 2005.

MITCHELL, P. et al. The exosome: a conserved eukaryotic RNA processing complex containing multiple 3'-->5' exoribonucleases. **Cell,** v. 91, n. 4, p. 457-66, Nov 14 1997.

MORAIS, C. N. et al. Cytokine profile associated with chronic and acute human schistosomiasis mansoni. **Mem Inst Oswaldo Cruz,** v. 103, n. 6, p. 561-8, Sep 2008.

MORTAZAVI, A. et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. **Nat Methods,** v. 5, n. 7, p. 621-8, Jul 2008.

MOUSE, E. C. et al. An encyclopedia of mouse DNA elements (Mouse ENCODE). **Genome Biol,** v. 13, n. 8, p. 418, Aug 13 2012.

MUÑOZ-LÓPEZ, M.; GARCÍA-PÉREZ, J. L. DNA Transposons: Nature and Applications in Genomics. **Current Genomics,** v. 11, n. 2, p. 115-128, 09/13/received

11/18/revised

12/01/accepted 2010.

NAGALAKSHMI, U.; WAERN, K.; SNYDER, M. RNA-Seq: a method for comprehensive transcriptome analysis. **Curr Protoc Mol Biol,** v. Chapter 4, p. Unit 4 11 1-13, Jan 2010.

NEVES, D. P. **Parasitologia humana**. Atheneu, 2012. ISBN 9788538802204.

NOWACKI, F. C. et al. Protein and small non-coding RNA-enriched extracellular vesicles are released by the pathogenic blood fluke Schistosoma mansoni. **J Extracell Vesicles,** v. 4, p. 28665, 2015.

OLIVEIRA, G.; FRANCO, G.; VERJOVSKI-ALMEIDA, S. The Brazilian contribution to the study of the Schistosoma mansoni transcriptome. **Acta Trop,** v. 108, n. 2-3, p. 179-82, Nov-Dec 2008.

OLIVEIRA, K. C. et al. Non-coding RNAs in schistosomes: an unexplored world. **An Acad Bras Cienc,** v. 83, n. 2, p. 673-94, Jun 2011.

OROM, U. A. et al. Long noncoding RNAs with enhancer-like function in human cells. **Cell,** v. 143, n. 1, p. 46-58, Oct 1 2010.

PARKER-MANUEL, S. J. et al. Gene expression patterns in larval Schistosoma mansoni associated with infection of the mammalian host. **PLoS Negl Trop Dis,** v. 5, n. 8, p. e1274, Aug 2011.

PAULI, A. et al. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. **Genome Res,** v. 22, n. 3, p. 577-91, Mar 2012.

PICA-MATTOCCIA, L.; CIOLI, D. Sex- and stage-related sensitivity of Schistosoma mansoni to in vivo and in vitro praziquantel treatment. **Int J Parasitol,** v. 34, n. 4, p. 527-33, Mar 29 2004.

PONTING, C. P.; OLIVER, P. L.; REIK, W. Evolution and functions of long noncoding RNAs. **Cell,** v. 136, n. 4, p. 629-41, Feb 20 2009.

POPIEL, I.; BASCH, P. F. Reproductive development of female Schistosoma mansoni (Digenea: Schistosomatidae) following bisexual pairing of worms and worm segments. **J Exp Zool,** v. 232, n. 1, p. 141-50, Oct 1984.

PROTASIO, A. V. et al. A systematically improved high quality genome and transcriptome of the human blood fluke Schistosoma mansoni. **PLoS Negl Trop Dis,** v. 6, n. 1, p. e1455, Jan 2012.

RAPOSO, G.; STOORVOGEL, W. Extracellular vesicles: exosomes, microvesicles, and friends. **J Cell Biol,** v. 200, n. 4, p. 373-83, Feb 18 2013.

REY, L. Parasitologia: parasitos e doenças parasitárias do homem nas Américas e na África. **Guanabara** v. n1, p. p.856, 2001.

RINN, J. L.; CHANG, H. Y. Genome regulation by long noncoding RNAs. **Annu Rev Biochem,** v. 81, p. 145-66, 2012.

ROBINSON, J. T. et al. Integrative Genomics Viewer. **Nature biotechnology,** v. 29, n. 1, p. 24-26, 2011.

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.

ROBINSON, M. D.; OSHLACK, A. A scaling normalization method for differential expression analysis of RNA-seq data. **Genome Biol,** v. 11, n. 3, p. R25, 2010.

SABAH, A. A. et al. Schistosoma mansoni: chemotherapy of infections of different ages. **Exp Parasitol,** v. 61, n. 3, p. 294-303, Jun 1986.

SAMBON, L. W. Remarks on Schistosoma mansoni. **The American Journal of Tropical Medicine and Hygiene,** v. 22, n. 10, p. 303-304, 1907.

SANCHEZ, M. C. et al. Effect of praziquantel on the differential expression of mouse hepatic genes and parasite ATP binding cassette transporter gene family members during Schistosoma mansoni infection. **PLoS Negl Trop Dis,** v. 11, n. 6, p. e0005691, Jun 2017.

SATPATHY, A. T.; CHANG, H. Y. Long noncoding RNA in hematopoiesis and immunity. **Immunity,** v. 42, n. 5, p. 792-804, May 19 2015.

SCARIA, V.; PASHA, A. Long Non-Coding RNAs in Infection Biology. **Front Genet,** v. 3, p. 308, 2012.

SCHOLTE, R. G. et al. Predictive risk mapping of schistosomiasis in Brazil using Bayesian geostatistical models. **Acta Trop,** v. 132, p. 57-63, Apr 2014.

SHORT, R. B.; MENZEL, M. Y.; PATHAK, S. Somatic chromosomes of Schistosoma mansoni. **J Parasitol,** v. 65, n. 3, p. 471-3, Jun 1979.

SIMPSON, A. J.; SHER, A.; MCCUTCHAN, T. F. The genome of Schistosoma mansoni: isolation of DNA, its size, bases and repetitive sequences. **Mol Biochem Parasitol,** v. 6, n. 2, p. 125-37, Aug 1982.

SMITHERS, S. R.; TERRY, R. J. The infection of laboratory hosts with cercariae of Schistosoma mansoni and the recovery of the adult worms. **Parasitology,** v. 55, n. 4, p. 695-700, Nov 1965.

SOTILLO, J. et al. Extracellular vesicles secreted by Schistosoma mansoni contain protein vaccine candidates. **Int J Parasitol,** v. 46, n. 1, p. 1-5, Jan 2016.

ST LAURENT, G.; WAHLESTEDT, C.; KAPRANOV, P. The Landscape of long noncoding RNA classification. **Trends Genet,** v. 31, n. 5, p. 239-51, May 2015.

STEINMANN, P. et al. Schistosomiasis and water resources development: systematic review, meta-analysis, and estimates of people at risk. **Lancet Infect Dis,** v. 6, n. 7, p. 411-25, Jul 2006.

STOCO, P. H. et al. Genome of the avirulent human-infective trypanosome--Trypanosoma rangeli. **PLoS Negl Trop Dis,** v. 8, n. 9, p. e3176, Sep 2014.

SUN, J. et al. Transcriptome profilings of female Schistosoma japonicum reveal significant differential expression of genes after pairing. **Parasitol Res,** v. 113, n. 3, p. 881-92, Mar 2014.

SUN, K. et al. Sebnif: an integrated bioinformatics pipeline for the identification of novel large intergenic noncoding RNAs (lincRNAs)--application in human skeletal muscle cells. **PLoS One,** v. 9, n. 1, p. e84500, 2014.

SUN, L. et al. Prediction of novel long non-coding RNAs based on RNA-Seq data of mouse Klf1 knockout study. **BMC Bioinformatics,** v. 13, p. 331, Dec 13 2012.

SUN, M.; KRAUS, W. L. From discovery to function: the expanding roles of long noncoding RNAs in physiology and disease. **Endocr Rev,** v. 36, n. 1, p. 25-64, Feb 2015.

TEODORO, T. M. et al. Occurrence of Biomphalaria cousini (Mollusca: Gastropoda) in Brazil and its susceptibility to Schistosoma mansoni (Platyhelminths: Trematoda). **Mol Phylogenet Evol,** v. 57, n. 1, p. 144-51, Oct 2010.

TRAPNELL, C.; PACHTER, L.; SALZBERG, S. L. TopHat: discovering splice junctions with RNA-Seq. **Bioinformatics,** v. 25, n. 9, p. 1105-11, May 1 2009.

TRAPNELL, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. **Nat Protoc,** v. 7, n. 3, p. 562-78, Mar 01 2012.

TRAPNELL, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. **Nat Biotechnol,** v. 28, n. 5, p. 511-5, May 2010.

TYCOWSKI, K. T. et al. Viral noncoding RNAs: more surprises. **Genes Dev,** v. 29, n. 6, p. 567-84, Mar 15 2015.

ULITSKY, I.; BARTEL, D. P. lincRNAs: genomics, evolution, and mechanisms. **Cell,** v. 154, n. 1, p. 26-46, Jul 3 2013.

VAN BAKEL, H.; HUGHES, T. R. Establishing legitimacy and function in the new transcriptome. **Brief Funct Genomic Proteomic,** v. 8, n. 6, p. 424-36, Nov 2009.

VAN BAKEL, H. et al. Most "dark matter" transcripts are associated with known genes. **PLoS Biol,** v. 8, n. 5, p. e1000371, May 2010.

VASCONCELOS, E. J. R. et al. The Schistosoma mansoni genome encodes thousands of long non-coding RNAs predicted to be functional at different parasite life-cycle stages. **Sci Rep,** v. 7, n. 1, p. 10508, Sep 5 2017.

VERJOVSKI-ALMEIDA, S. et al. Transcriptome analysis of the acoelomate human parasite Schistosoma mansoni. **Nat Genet,** v. 35, n. 2, p. 148-57, Oct 2003.

W.H.O. Map: Distribution of schistosomiasis, worldwide, 2012. **The World Health Organization**, 2014.

W.H.O. Prevention and control of schistosomiasis and soil-transmitted helminthiasis. **World Health Organ Tech Rep Ser,** v. 912, p. i-vi, 1-57, back cover, 2002.

_____. Schistosomiasis Fact Sheet. **World Health Organ Tech Rep Ser**, 2015.

WANG, C. et al. LncRNA Structural Characteristics in Epigenetic Regulation. **International Journal of Molecular Sciences,** v. 18, n. 12, p. 2659, 12/08

10/27/received

11/26/accepted 2017.

WANG, F. et al. Characteristics of long non-coding RNAs in the Brown Norway rat and alterations in the Dahl salt-sensitive rat. **Sci Rep,** v. 4, p. 7146, Nov 21 2014.

WANG, K. et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. **Nucleic Acids Res,** v. 38, n. 18, p. e178, Oct 2010.

WANG, K. C.; CHANG, H. Y. Molecular mechanisms of long noncoding RNAs. **Mol Cell,** v. 43, n. 6, p. 904-14, Sep 16 2011.

WANG, L. et al. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. **Nucleic Acids Res,** v. 41, n. 6, p. e74, Apr 1 2013.

WANG, X. et al. Transcriptome Bioinformatical Analysis of Vertebrate Stages of Schistosoma japonicum Reveals Alternative Splicing Events. **PLoS One,** v. 10, n. 9, p. e0138470, 2015.

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.

WAPINSKI, O.; CHANG, H. Y. Long noncoding RNAs and human disease. **Trends Cell Biol,** v. 21, n. 6, p. 354-61, Jun 2011.

WILSON, M. S. et al. Immunopathology of schistosomiasis. **Immunol Cell Biol,** v. 85, n. 2, p. 148-54, Feb-Mar 2007.

WILSON, R. A. **Introdução à Parasitologia**. 1979.

WILSON, R. A. et al. The Schistosome Esophagus Is a 'Hotspot' for Microexon and Lysosomal Hydrolase Gene Expression: Implications for Blood Processing. **PLoS Negl Trop Dis,** v. 9, n. 12, p. e0004272, Dec 2015.

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 01 2009.

WU, T. D.; NACU, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. **Bioinformatics,** v. 26, n. 7, p. 873-81, Apr 1 2010.

WU, Y. et al. Systematic Identification and Characterization of Long Non-Coding RNAs in the Silkworm, Bombyx mori. **PLoS One,** v. 11, n. 1, p. e0147147, 2016.

WYNN, T. A. et al. Immunopathogenesis of schistosomiasis. **Immunol Rev,** v. 201, p. 156-67, Oct 2004.

YANG, J.-R.; ZHANG, J. Human Long Noncoding RNAs Are Substantially Less Folded than Messenger RNAs. **Molecular Biology and Evolution,** v. 32, n. 4, p. 970-977, 01/21 2015.

YOU, H.; MCMANUS, D. P.; GOBERT, G. N. Current and prospective chemotherapy options for schistosomiasis. **Expert Opinion on Orphan Drugs,** v. 3, n. 2, p. 195-205, 2015/02/01 2015.

ZAPATA, J. C. et al. The Human Immunodeficiency Virus 1 ASP RNA promotes viral latency by recruiting the Polycomb Repressor Complex 2 and promoting nucleosome assembly. **Virology,** v. 506, p. 34-44, 03/21 2017.

ZHAN, S. et al. Genome-wide identification and characterization of long non-coding RNAs in developmental skeletal muscle of fetal goat. **BMC Genomics,** v. 17, p. 666, Aug 22 2016.

ZHANG, Y.; CAO, X. Long noncoding RNAs in innate immunity. **Cell Mol Immunol**, Aug 17 2015.

ZHANG, Y. et al. A Review on Recent Computational Methods for Predicting Noncoding RNAs. **Biomed Res Int,** v. 2017, p. 9139504, 2017.

ZHANG, Y. C. et al. Genome-wide screening and functional analysis identify a large number of long noncoding RNAs involved in the sexual reproduction of rice. **Genome Biol,** v. 15, n. 12, p. 512, Dec 3 2014.

ZHAO, Y. et al. Computational methods to predict long noncoding RNA functions based on co-expression network. **Methods Mol Biol,** v. 1182, p. 209-18, 2014.

ZHENG, Y.; CAI, X.; BRADLEY, J. E. microRNAs in parasites and parasite infection. **RNA Biol,** v. 10, n. 3, p. 371-9, Mar 2013.

ZHOU, Q. et al. QC-Chain: fast and holistic quality control method for next-generation sequencing data. **PLoS One,** v. 8, n. 4, p. e60234, 2013.

ZHU, B. et al. Genome-wide identification of lncRNAs associated with chlorantraniliprole resistance in diamondback moth Plutella xylostella (L.). **BMC Genomics,** v. 18, n. 1, p. 380, May 15 2017.

ZHU, L.; LIU, J.; CHENG, G. Role of microRNAs in schistosomes and schistosomiasis. **Front Cell Infect Microbiol,** v. 4, p. 165, 2014.

9 ANEXOS

Anexo A. Algoritmo utilizado no *pipeline* para predição dos lncRNAs em *S. mansoni*.

```
import os
import sys
import time

path = "/Users/victorfernandes/project"
#path = sys.argv[1]
#

#for arg in sys.argv:
 #   print arg

#argv gets information from the command line such as options, input filenames, foldernames etc.
reads = sys.argv[1]
genomeschisto = sys.argv[2]
gfffile = sys.argv[3]

#lists the files in a folder, reads is the name/path of the folder
listfiles = os.listdir(reads)

#for ist in listfiles:
#    print "STAR --inputfile ist", ist
   #time.sleep(10)

print os.path.exists(path)
#exit(-1)
#os.mkdir("/Users/victorfernandes/project/test")
if not os.path.exists(path):
  os.makedirs("/Users/victorfernandes/project/test/basti")
#STAR -option input_file

print os.getcwd()

#genomefasta = "/Users/victorfernandes/project/genome/"
genomename = "Schistosoma_mansoni_v5.2.fa"
#reads = "/Users/victorfernandes/project/genome/reads/"
read1_name = "ERR022873_1.fastq"
read2_name = "ERR022873_2.fastq"
```

```
#Paths programs
starfolder = "/Users/victorfernandes/project/programs/STAR-STAR_2.4.0j/bin/MacOSX_x86_64/STAR"
cufflinksfolder = "/Users/victorfernandes/project/programs/cufflinks-2.2.1.OSX_x86_64/"


#Results
mappingSTAR = "/Users/victorfernandes/project/results/mappingSTAR/"
genomestar = "/Users/victorfernandes/project/results/genomestar/"

read1 = os.path.join(reads,read1_name)
read2 = os.path.join(reads,read2_name)
#genomeschisto = os.path.join(genomefasta,genomename)
print "read1: ", read1

if os.path.exists(read1):
  print "OK"
else:
  print "not OK"

runThreadN = 4

#exit(-1)
#os.execlp("echo", 'name','time')

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


#Some decision about whether to do shrimp or not


#1 - Create a fasta genome using STAR
command = "{0} --runMode genomeGenerate --genomeDir {1} --genomeFastaFiles {2} --runThreadN {3}".format(starfolder, genomestar, genomeschisto, runThreadN)
print "command1: ", command
os.system(command)
command = "/Users/victorfernandes/project/programs/STAR-STAR_2.4.0j/bin/MacOSX_x86_64/STAR --runMode genomeGenerate --genomeDir {0} --outFileNamePrefix /Users/victorfernandes/project/genome/genomestar --genomeFastaFiles {1} --runThreadN {2}".format(genomestar, genomeschisto, runThreadN)
print "command1: ", command
os.execvp("STAR",  ["STAR", command])

#Test here for looping
```

```
#print genomeschisto

#print "starfolder=", starfolder
#print "genomeschisto=", genomeschisto
#print os.path.join(reads,listfiles[i])
#print os.path.join(reads,listfiles[i+1])
#print "mappingstar=", mappingSTAR
#print "runThreadN=",runThreadN

i=0
while(i<len(listfiles)):
   if(listfiles[i]==".DS_Store"):
      i=i+1
#3 – Trimming (Trimmomatic)
java -jar /Users/victor/Doc/Aplicativos/Trimmomatic-0-2.36/trimmomatic-0.36.jar PE -
phred33 /Users/victor/Doc/Stages/Adult/Raw_Reads/ERR022873_1.fastq
/Users/victor/Doc/Stages/Adult/Raw_Reads/ERR022873_2.fastq
/Users/victor/Doc/Stages/Adult/Trimmed_Reads/trimmed_ERR022873_1P.fastq
/Users/victor/Doc/Stages/Adult/Trimmed_Reads/trimmed_ERR022873_1U.fastq
/Users/victor/Doc/Stages/Adult/Trimmed_Reads/trimmed_ERR022873_2P.fastq
/Users/victor/Doc/Stages/Adult/Trimmed_Reads/trimmed_ERR022873_2U.fastq
HEADCROP:15 LEADING:20 TRAILING:20 SLIDINGWINDOW:5:20 MINLEN:50


   # 2 - Mapping the genome with STAR
   command = "{0} --genomeDir {1} --readFilesIn {2} {3} --outFileNamePrefix {4} --
runThreadN {5} --outSAMstrandField intronMotif".format(starfolder, genomestar ,
os.path.join(reads,listfiles[i]), os.path.join(reads,listfiles[i+1]), mappingSTAR,runThreadN)
   print "command2: ", command
   #os.system(command)

   # 3 - Convert the file SAM to BAM
   command ="samtools view -bS
/Users/victorfernandes/project/results/mappingSTAR/Aligned.out.sam >
/Users/victorfernandes/project/results/sam2bam/Aligned.out.bam"
   print "command3: ", command
   #os.system(command)

   command = "samtools sort
/Users/victorfernandes/project/results/sam2bam/Aligned.out.bam
/Users/victorfernandes/project/results/sam2bam/File1" #File1 is the output file.
   print "command3.1: ", command
   #os.system(command)

   command = "samtools index /Users/victorfernandes/project/results/sam2bam/File1.bam"
```

```
    print "command3.2: ", command
    #os.system(command)

    # 4 - Cufflinks
    command = "{0}cufflinks /Users/victorfernandes/project/results/sam2bam/File1.bam -o
/Users/victorfernandes/project/results/cufflinks/".format(cufflinksfolder)
    print "command4: ", command
    #os.system(command)

    #cufflinks -o /home/olivei_a/Testread1/cufflinks Victor1.bam

/home/victor/Aplicativos/cufflinks-2.2.1.Linux_x86_64/cuffcompare -r
/home/victor/Doc/ReadsProtasio/Genome/GenomeFasta/Schistosoma_mansoni_v5.2.gff -V -
R /home/victor/Doc/Trimmed_Reads/Results/Adult/Cufflinks/transcripts_adult.gtf

#Single-exon

/home/victor/Aplicativos/cufflinks-2.2.1.Linux_x86_64/gffread -w
/home/victor/Doc/Trimmed_Reads/Results/Adult/Transcripts/transcripts_adult_single_exo
n.fa -g
/home/victor/Doc/ReadsProtasio/Genome/GenomeFasta/Schistosoma_mansoni_v5.2.fa
/home/victor/Doc/Trimmed_Reads/Results/Adult/Cuffcompare/cuffcompare_class_i_u_x.gt
f


    # 5 - Extracting transcript sequences (gffread tools from cufflinks)
    command = "/Users/victorfernandes/project/programs/cufflinks-
2.2.1.OSX_x86_64/gffread -w
/Users/victorfernandes/project/results/transcriptsfasta/transcripts.fa -g
/Users/victorfernandes/project/genome/Schistosoma_mansoni_v5.2.fa
/Users/victorfernandes/project/results/cufflinks/transcripts.gtf"
    print "command5: ", command
    #os.system(command)

    # 6 - Sequences => 200 pb
    command = "perl cat.pl
/Users/victorfernandes/project/results/transcriptsfasta/transcripts.fa >
/Users/victorfernandes/project/results/transcripts-long/transcripts-long.fa"
    print "command6: ", command
    #os.system(command)

    # 7 - CPAT (Coding Potential Assessment Tool)
    command = "/home/victor/Aplicativos/CPAT-1.2.2/bin/cpat.py -g
/home/victor/Doc/Trimmed_Reads/Results/Adult/Transcripts/transcripts_adult_long.fa -d
/home/victor/Aplicativos/CPAT-1.2.2/prebuilt_models/fly_logitModel.RData -x
/home/victor/Aplicativos/CPAT-1.2.2/prebuilt_models/fly_Hexamer.tab -o
/home/victor/Doc/Trimmed_Reads/Results/Adult/Transcripts/transcripts_adult_lncRNAs.f
```

```
a"

   # 8 - Non-coding transcripts
   command = "perl -lane 'if($F[-1]==1){system \"grep -A 1  $F[0]
/Users/victorfernandes/project/results/transcripts-long/transcripts-long.fa\";}'
/Users/victorfernandes/project/results/cpc/ff.feat
>/Users/victorfernandes/project/results/non-coding-transcripts/lncRNAs.fa"
   print "command8: ", command
   #os.system(command)
   #perl -lane 'if($F[-1]==1){system "grep -A 1  $F[0] ../transcripts-long/transcripts-long.fa";}'
ff.feat >non-coding-transcripts

9 – ORF predictor
   command = " perl fa2oneline.pl orf_adult > orf_adult_oneline.fa"
   print "command9: ", command
   #os.system(command)
   i = i +2

exit(1)
```

Anexo B - Algoritmos utilizados individualmente

#Algoritmo 1 - Tabela 1 do Cuffcompare

```
filesNames=['genes.fpkm_tracking_adult.fpkm_tracking','multi_exon_adult.fa','multi_exo
n_adult.gtf']


f1=open(filesNames[2])
f2=open(filesNames[0])
out=open('Tabela1.csv', 'w')


lines1=f1.readlines()
lines2=f2.readlines()

D={}
for line in lines2:
```

```
    f=line.split('\t')
    D[f[0]]=f[9]

def processLine(line):
    fields1 = line.split('\t')
    fields1 = fields1[:-1] + fields1[-1].split(';')[:-1]
    print fields1

    c1=fields1[11].split('"')[-2]
    # print c1
    c2=fields1[9].split('"')[-2]
    # print c2
    c3=fields1[8].split('"')[-2]
    # print c3
    c4=fields1[0].split('.')[-1]
    # print c4
    c5=fields1[10].split('"')[-2]
    # print c5
    c6=fields1[12].split('"')[-2]
    # print c6
    if len(c6)>1:
        c6=fields1[13].split('"')[-2]
    key='.'.join(c1.split('.')[:-1])
    c7=D[key]
    cols=(c1,c2,c3,c4,c5,c6,c7)
    return ','.join(cols)

headers=('Transcript Old ID','Transcript New ID','Gene ID','Genome Location','Exon
number','Class Code','FPKM','\n')
out.write(','.join(headers))
for line in lines1:
    row=processLine(line)+'\n'
    out.write(row)




f1.close()
f2.close()
out.close()

print('<END>')
```

#Algoritmo 2 - Tabela 2 do Cuffcompare

```
from Bio import SeqIO

records=SeqIO.parse("multi_exon_adult.fa", "fasta")

out=open('Tabela2.csv','w')

out.write(','.join(('Transcript New ID', 'Length', '\n')))
for r in records:
   out.write(','.join((r.id, str(len(r)), '\n')))
   # print r.id, len(r)

out.close()

print('<END>')
```

#Algoritmo 3 - Tabela 3 do Cuffcompare

```
f1=open('multi_exon_adult.gtf')
D={}
lines=f1.readlines()

for line in lines:
   fields=line.split('\t')
   fields=fields[:-1]+fields[-1].split(';')
   c1=fields[9].split('"')[-2]
   if c1 in D.keys():
      D[c1].append( (int(fields[3]), int(fields[4])) )
   else:
      D[c1] = [(int(fields[3]), int(fields[4]))]


out=open('Tabela3.csv','w')
headers=','.join(('Transcript New ID','Exon size','Intron size','\n'))
out.write(headers)

keys=D.keys()
keys.sort()

for k in keys:
   # print k, D[k]
   exons=[]
   introns=[]
   for ex in D[k]:
      exons.append( ex[1]-ex[0] )
   for i in range( 1, len(D[k]) ):
      introns.append( D[k][i][0] - D[k][i-1][1] )
```

```
    I=';'.join([ str(x) for x in introns ])
    E=';'.join([ str(x) for x in exons ])
    out.write( ','.join([k, '"'+E+'"', '"'+I+'"', '\n']) )


f1.close()
out.close()
print('<END>')
```

#Algoritmo 4 - Filtrar arquivos da tabela referencia

```
from Bio import SeqIO

lines = set([x for x in open('256.csv').read().split('\n')[1:-1]])

fasta = SeqIO.parse("644.fa", "fasta")

records = list()
for rec in fasta:
    if rec.id in lines:
        records.append(rec)

SeqIO.write(records, "fasta_filtrado.fa", "fasta")

print('END')
```

#Algoritmo 5 - Filtrar resultados das ORFs

```
from collections import defaultdict

fasta = defaultdict(list)
with open('ORFS') as file_one:
    for line in file_one:
        if line.startswith(">"):
            seqInfo = line
            seqAA = next(file_one).rstrip()
            if len(seqAA) <= 100:
                seqName = seqInfo.strip(">\n")
                fasta[seqName].append(seqAA)

tags = set()
with open('ORFS_edited', 'w') as file_one:
    for k, v in fasta.items():
```

```
        tags.add( k.split()[0] )
        file_one.write('>' + k + '\n')
        file_one.write(v[0] + '\n')

fasta = dict()
with open('NonCodingSequencesgff3.fa') as file_one:
    for line in file_one:
        if line.startswith(">"):
            seqInfo = line.strip(">\n")
            seqName = seqInfo.split()[0]
            if seqName in tags:
                fasta[seqInfo] = next(file_one).rstrip()

with open('NonCodingSequencesgff3_edited_ORFS.fa', 'w') as file_one:
    for k, v in fasta.items():
        file_one.write('>' + k + '\n')
        file_one.write(v + '\n')

tags = set()
with open('NOORF') as file_one:
    for line in file_one:
        seqName = line.strip().split()[0]
        tags.add(seqName)

fasta = dict()
with open('NonCodingSequencesgff3.fa') as file_one:
    for line in file_one:
        if line.startswith(">"):
            seqInfo = line.strip(">\n")
            seqName = seqInfo.split()[0]
            if seqName in tags:
                fasta[seqInfo] = next(file_one).rstrip()

with open('NonCodingSequencesgff3_edited_NOORF.fa', 'w') as file_one:
    for k, v in fasta.items():
        file_one.write('>' + k + '\n')
        file_one.write(v + '\n')

print('END')
```

#Algoritmo 6 - Filtrar resultados dos resultados do CPAT

```
from sets import Set
# ================================================
```

```
# ==============      CONFIG     ===================
# ======================================================

fastafile = 'transcripts_cercarie.fa'
outputfile = 'output2'

aboveThresholdFile = 'nonCodingSequences.fa'
belowThresholdFile = 'codingSequences.fa'




# ======================================================
# =================  SCRIPT   ======================
# ======================================================

threshold = 0.1

fin1 = None
fin2 = None
fout1 = None
fout1 = None

try:

        fin1 = open(outputfile, 'r')
        fin2 = open(fastafile, 'r')

        fout1 = open(belowThresholdFile, 'w')
        fout2 = open(aboveThresholdFile, 'w')

        codingSet = Set()
        nonCodingSet = Set()

        lines = fin1.readlines()
        cont = 0
        for line in lines[1:]:
                fields = line.split()
                value = float(fields[-1])
                if value >= threshold:
                        codingSet.add(fields[0])
                else:
                        nonCodingSet.add(fields[0])
```

```
        lines = iter(fin2.readlines())

        cont = 0

        line = next(lines)
        while 1:
                if line.startswith('>') and len(line) > 3:
                        ID = line[1:].split()[0]
                        if ID in nonCodingSet:
                                fout1.write(line)
                                line = next(lines)
                                while not line.startswith('>'):
                                        fout1.write(line)
                                        line = next(lines)
                        elif ID in codingSet:
                                fout2.write(line)
                                line = next(lines)
                                while not line.startswith('>'):
                                        fout2.write(line)
                                        line = next(lines)
                        else:
                                line = next(lines)
                else:
                        line = next(lines)


except StopIteration, e:
        pass
finally:
        if fin1:
                fin1.close()
        if fin2:
                fin2.close()
        if fout1:
                fout1.close()
        if fout2:
                fout2.close()

print 'END'
```

Anexo C. Sequências do conjunto de 15 Sm-lncRNAs preditos e validados.

>Sm-lncRNA 1 TCONS_00001011 gene=XLOC_000494

TGCGGATCACGCGGTTCCCTTCCTGTCCCAGCGTCCCATTTGCTCTTATTCTGCGAGT
TTTGGTTAGACGAGTGCCTATAAATCACAGCGAGTCGATTGCGAAGATTGTTTGGAA
ACATATCCAATGAAAATTGTAGTTAAATACTAGGAGTCGTCCTACCGGGAGATTCAC
TTATTGACTCGGAATACCCCAATAATATAGTTCTGTTTGAAGATGCTGACAAAGTGC
AATCTTCTAACCACATTAAGTAATAACGCAAGCATGTTTCGGATGCGGTTCTAACCC
TCCAAATCCAAAATGTTACCTCAGGATTAACCTGCGTTCGAGTCTGAGTTAGTCATA
GGGAGTGAATTATTCGGACGTATCGACCGCCTCACTTATCTTGGAAGTGTCATCAGC
AGTGGTGTTTGAAGAAATCTCAGCACGAATTCAAAGGGTTTTATTGGTATTTGCCAA
CTTGCTGTGTGGGGAAGATATCGTTTATACTTGCTTGATGAGTACATTTATCAAGTG
ATCATAATTTTTTCTGCTTTGACAATCCTAATTTTCCGAGAGTTTATTATTATGTA

>Sm-lncRNA 2 TCONS_00012347 gene=XLOC_007663

GGCAAAGTTGACAAAATCGGCTTGGAGGTCCCACTCTGTAATTATGTAGCAAGCGAT
AAAATCTGGCGCGATAACTGCATTAAGGAGGCGAGCGCAGCTAAAGCATGGTCCAA
GAATTGGAGCTTCCTTACTTTAACCCCTCGGGAGCTCCTAAAAGATGAACTCCACGA
ACTGATTGATCCAAACAGAAAACCTATTGAAATCCCACAATACTTGAAGGTGGCCG
AAGCAGTACCTATCTCGGCCTATATAAAGGTGGAACCTTCGCCGAAGCCCATCCCGC
AAACAACAAGTAGAATGATTGGATGGAGATCTGGACTGCCTCAGTACAAACTAGAC
AAGTATGAAGTAGCAAAAAGACCTCAAGGTTCGCTTCTTAAGAGGTTTAACTGGCC
AATAGAAGCTTTATACTGATGTTAAATTTGACTTGGGTAGTGGATTGGAAGCGCCAG
TCTCATAAGTCCCTGAGAGAGATATGTGAAAAACCCGTCCTTCGTACGTAAGAGTAC
TATGCCTCCATTTCTCCCATAGTTCAAAATTGACATTGAAGTTCAAT

>Sm-lncRNA 3 TCONS_00013257 gene=XLOC_008107

ATATTCTCCAAGCCTCATCAATACTTCATCATCTAGTTAGCAGGGCCCGGGAGAATG
GATTAGAATAACCTACCGTGTCACTGTGTCTTGACCTGCATTCCGTTTACCGAATAA
GGTGAACTCGTGATCTCATCAAACATATCAATAACAAACCACCAATTTCCATCAAAT
ACCCGAAATGCTAAACCAAAGGGGCGGGATAAAAGCTAGCTTGAAATTGAAGTTCA
GACTTTTTCATTTCAGCGACCCTGAGAGGCTTGACTTAGGGCAGTCATTCCATGATT
GAAGTGCCTCAGTTCAGTGATGCTTTTAAAGGATGGCTGTAATTAGTAATATGTGTG
CCGTTGGCGGAGCTGCATGTCTCAGAAGACACCAAACAGTGCCATTTATTGTACGTT
ATTTCCTGCCTTTCAGCTAGAGAAAACTTATGTGTTGAGCTTTTTAATTTTCGGCCTT
TGGTTTGTAGTAGTGGTGGCACTGTGTTAAACACGTCTGAAAGTATTTTGACAGACC
TAACCGATACTTCTTGGCAGGCGATCCTATCTCAAAGGTTCTTGACATGTGGTGAAT
TCAGAGATTTCTTGAACTTTTTCCAGCATATAAGGAAAGCATAATTCTGATGGTTGC
AAATGACATGTACTCATGACATCCACAAACAAAGTAAGCACTCCTTTTTTCCTGTGA
TGTTTCTCAATAACAATTATGTGGAACTTTCAGGTCCAGCACATCATGCTTGTAGGG
G

>Sm-lncRNA 4 >TCONS_00003599 gene=XLOC_002172

CCTTCTAAGATTTTCCCACCTTTTGTACACCCATCCAGGACACAAATTATCTTTAGGC
CAGAATCTAACCAAACGAGTGGTGAATATTATATATAATGTTGTTTTTTTCATTTTTT
AAAACTTTAGTTGGGAAGGATGTGTTAATTGAACTGAAGAATGACCTATGTATATGT
GGTACGTTACACTCGGTCGACCAGTACCACAACTTCAAACTCACCGATATTAGCGTA
ACTGATCCTGAAAAACACCCTCATATGTTGTCCGTAAAAAACTGTTTCATAAGAGGG
TCTATTGTGAGATATGTCCAGCTCCCTGCTGATGAATGCGATACGATTGAATTA

>Sm-lncRNA 5 TCONS_00000625 gene=XLOC_000312

ACAATTTCGTACATTTGAGCTAATACAATGTCTGACTGCTGTGAAGGAAAATGTGGT
TGCGGATCGAGCTGTAACTGCACCTCTGGCACCTGTAAATGTGATGGCAAATGTTCG
GGATCCAAATAACCAAACACACTCATATGGATGTGGATCAAGAAATCAACGAGTAG
GAGTACAGTTTCGACAGCTTGTCAACACAA

>Sm-lncRNA 6 TCONS_00001840 gene=XLOC_000922

ACCAAATTGAGCAGAAACTTAAATCATTAAACCCACTTCATATCGAAATAGTTGACT
TTTCAGATGGATGCGGTCTGAAATTTGATGTAAAAGTCGTCTCACAAGAATTCGAGG
GTAAATCGCTTGTTGACAGACACAGACTGTTGGAAGAGGAAATGAAATCAGTACAT
GCCCTTACTCTAAAAACGCTAGCTCCTTCACAATGGTCGCTTAAATCCTCCTAACAC
CCTTATCATCGGGCTCGATAATTGGAAATTGTCATATATTTTTGTAAATAA

>Sm-lncRNA 7 TCONS_00009100 gene=XLOC_005569

GCCGCAAAATTTCAAACCCCAGTTTTGTAACGTTAGCAACACAGTGAGGTAGAGAC
GACTTTTGCACTAAAATCACTGATTAATTGGTCGGTTGAATTAATACATTGGACTGA
GAAGTTACAGTTCGTTGGAGGCCAATTGAGCCGTGAATCTATACGCATCTCATCGGG
GGGATTGACTCCAACGACAATTGCAGACAAGTGATCGAGTTACACACTACAGTTTAT
TTATCCTCTGGACTAAACTTCTCACCTGCTGATGTCGTGTTCACAGATTGCAGTAGAG
TGCACTAGAGAACGATTTAGGCTTGTATATCCCCATCCTAACGTTCTTTATCAGTGG
AGGGCAATACATATGTTTTAGTCGAGGAAATGCATTTGCAATATATCGATGTATTTT
TGCTTCCTTGTTTGCAACTCTACTCATTTTGGATTATATGTGGATAATTCTACGTGAT
CCAAAATTTCCAGACTTTCTATCCTGGTGTTTGGATGCATCTCAGTGGGCTTTGTTAA
GCGCAACTGTATGTTATATAATATTGGCGTGTAATGTGGCGTGCATATCTAGACGAA
ATGAGGAATCAAACAACGAAAGTAAGTATTACCAATCAATAAAGCTAAACTATGAG
GGCGATCCAGTGATTATTATGCAAACCAATT

>Sm-lncRNA 8 TCONS_00009852 gene=XLOC_006131

AGGAATATGCACTTGGGTTCTTAATAACGCCACATTAAGGATATCTCAATTACAAGT
TACTCTTTTTGTTCTTTTATTGATTTGATATAATCTGCACATATGTTGCTAGATTATAG
TTTATCACACTTAGAAATTTGTCATTTATTACATGAACCTCATTTTTACGTCTAGTTTT
TTCAGTGTTATGATTTGTTTTCTGCTTTTTGATGCTTTAATTTATTCATTCAAAAGTAG
TGAGTAAATTCCAACAAAATGGGGAAGTTGGGCATCCTAATTCAGATGATGAGGTA
CATCCAAATTCATCAGCATTATATAACCACAGCTATGAATTGTCTCTTGAGCCAACC

CATGATGTTATACATCATTTGCCGAATAAAAATCTGTTTGTAGATGATACAATGCAC
GTGATATGTGAAGATGCTAGTACTACTACCAACAAAAATAATCGGAAACACTCATT
ATGTGTTGAATTGGTATCATCACTTTCAGCACCGCCACCTGAAGATGATGGCGTTAT
GTGCAATCGTCCATTGATGAATATAAATCGTAAATTTTCTATTCAGAATAATCAACA
TACCTATGATAACAATGGGATTATGAATAACAAAGCATGCCTTTAGCACCAACAAAT
ATACAATCATATATCCAAACAAACAGGCTCGTATAAACATAAATAGAATGAATACA
ATGATTTGTGTTCGGTTTCTCCCTCCCCTATTTCTTCTTGTAATATTTTATTATAAACA
CTTGATTTGTTCGTTTCAAATTGTTAATATCTTGATTTATTTTTCTGATTGAGATTTGT
GTCAAGAACAATACTGTTATTATCTCTGTGTTTTTTGGATGACATCAGAGTGATACTT
GTTTTTCTTGTAATGTGTTTCCTCTATTTTTTCTGAATAAAAAACCGTTAATTATTATT
TTCAAGATGAAAAATATTTTCCCTTATCATCAATTACCCTCTACCGAAAAACCACAC
AGGTAGTTCAGCCAAGTAGTAGTAACTGTTGTTGGTGCAAATGGGACTAGGAATTTC
AAAACAATTTATCCTCTCTGTAATTCAATTGCGAAGTGCAAGTGATTCGTCTTTGAA
AAATATTAAAATTATCAGTTTCATTCCTTTATTCAACTACACCTTGTGTCTACCATAA
TTAAAA

>Sm-lncRNA 9 TCONS_00009849 gene=XLOC_006129

TCTTGAGTTTAGGCCTTACGCGCCTGCTTTGTTTTCGTATCCTTCGATGCATATTTTAC
GTCCTGTATTTCATGCCTTCTGTCGTTATAACCTATAGTATACGCAAACCTCAAAGTT
CGTTTTACCTAGAAATCAGATATGCACATCAAGAAAAGACTTCGCCGCTACATAAGT
AACATTATAAGCTTTACTAATTCCCACACTCGAAACACGCTTCGTGATCGATCTTCTG
ATTTTTATGCAAATGAAGTTACAGATGAATTGGTTGATCTCATGCGTAAAGTTTGGC
TACCGATTGTTGGTCCTGGTTATGACATCATAGTTTTACATATTCATAAATCTCGAGG
AACATCAAATTTGGGCATCAGCATAGAAGGTGTGACCTATGTGGCAGAACCAGATT
TCGATCATGAAAATTTACCATCTACATTGGTTACAACTAATAGTCAAATTGAAGATA
AATCTAATCCAAGTAGTAAACTGTTGCCAAATGGTATGGCAAAATCTGCCCATTCAA
ATGATATCAAGTATAGTGATATACCTTCACGCCATTTCGTTCAATATATTGTACCTGA
TGGTCTGATCGGTAGTTTGGGTGTTGTGCAAAAGGGTGATGAGTTACTTCAGGCTAA
TGGTCATCGTATACATGGTACAACGCATACTAGTACTCTTCGTTATTTACGCAATTTA
CCATCACGCATTGAATTAGTCTTTGCCCGTAGAAAGTCTACATATGAGAATGGAGAT
GATGACGTGTTTAGTATTACGGGTGGTAAGGATCAGTTGATTGATGTCGTAGGGGAA
TCACTACTTGAAGTTGCTGCTAGTGAAATCGGTTCCGTAGATACTGGCTCATCTGTG
AGAATGGTGGATGCTTACGATAAAGTTGATCGGTCAAATTATTCCAGTCCAGTATCA
CCTGCTACGGCTCATAAACGTGTCACTGAATGGATTCGAAAATCTCAAGGAGATTTA
AGCGCGAATATTCACTATCCATCTTCACCTGAATCTTCTGTAACTATGCAAACTAAT
AATAAATCAGAAAGATTTAGTTGTGATTTAAAACATGCACAATCGTACAAAATTCAC
CCCCTCCATAACTATATAAATCCGAATTACACTAGTAAATTGCAACAACAGTATCTA
GTAAATACAATGAAAAACGATCGTAAATATTCATTACCAAGTACAGTACAGCCGTC
AGAAACACAAATCCATTCTAGGACTTTAGGGCGTTTGCCAATATCAAGATGTAGTAA
AGTGTGTATAGGAAAACAATCATCTACTGAACAACACAATCATCGCCGTTGTCAAAC
TCTACCACATTATCATCATAATCGACCAAATCGTATTGATCCTCGTTACGGCTTTAAA
CGACCATGTTGGTCTTCAGTGCCTTTAATTATTCAACTGAATAAAACATCTCATGGTT
TTGGATTTAGTATAGCTGAATATGAGGAACTACCTGTTACTGACTTAGAAGGCAATA

CACTCCGTAAAGCTTATTCACTTGATAGAAGGAGTTCTAATATGATTGAATCTGATC
GTCGATCATTTCATTCCTATTATTCTACCGGTTCGACCATTGCATCATTATCATCGTT
ACCTGGTAAAAATTCCAAATCTTCCAGTGGATTAAAACGAAGATCCACGTGGTCTAG
TCGATCCAAAGCACACGGTATCTTATTAGTTGACAGTTTAACTCCAGGTGGTATTGC
ACAGTTGGATGGACGTATATCAATCGGTGACAGGCTTTTATTTGTTAATGATAAAAA
TTTAATGAAATCTAGTGTATTTGAGGCCGCTAATACACTAAAGTCCCTTCCTAATGG
ACCGTGTTTAATTGGTATTGCAAAAATGCAGTTAGAATCAAACGAAACAGATGAAA
TTCACGAAAAAAACCAGCAACAATTATTACTGCCATACCCTGTTGTCTCTCCTAGAC
CTTCTGTTATTGGAATGTTTCCAGATACTGGTGAAATGAAAAAGATCAGTGTTAAGT
CTGATCTATCATATCCGCTACACGTAAACATTGGTGACAGCACTATCGATAATGGTG
ATAATAATACGATTATGGTAAGTATTATAACATGGTGTT

>Sm-lncRNA 10 TCONS_00009851 gene=XLOC_006130

TCAGACGGTTTTGGTATTTTCATCGTTAATTTGAGTCCAAACAATGAACCTGGTGTAT
TCGTTAGCGAAATCCGTCCAAACAGCCCAGCATCACAACAAGGAGTTCTTAGACCTC
ATGATCGTATACTAGCTATTGATGGCCAACTTCAATGTGATTATGAAAGTACACTGG
AATTGTTGCAAAAATCAAGAAAATCTGTACGTCTAACCATCGGTAGACAAATACCTT
ATCACAGTGACTCACAACAAATTCAACAAATTGGACATGTAACTACAACAGTGGAT
AATCCAGACAATGATATAAAACATAAATTACCCATTATACCTGGGATTCCTGCTACA
GTTACTCTGTACAAGACTGACGGAGGTCTTGGATTCTCCATTGTCGGTGGAAGTGAC
ACAGTTTTGAGTAATATTTTGGTCCATGAAGTACACTCTGGTGGAGCTGCTGCACGT
GATGGTCGGCTTCAAGTTGGTGACCGTTTATTAGCTGTCAATGGAATCGATCTTCGT
GAAGCTACTCAGAAAGATGCTACAAAGATTATACGAACAGCTGATGACTGTATCCA
ATTGGTTGTGTACAGAGATCCTGAACCACAATATATAAATCAAGGTGTATTCGAGTG
TCATAGTGTTCACTTGAAACGTGATATGCCCGGACAAAGTTTTGGTCTGTCATTGATT
GGTCGACCTCATTATTCAACCGGCACGGCTATTGGCGGCATTATAGAGAATAGTCCA
GCTGCTCGTTCAAATCTTTTAGAAGTTGGTGATATAATTTTAGAAATTAATGGTTGG
GATATGCGTTTGGCTAAATCAGATGAAGTAGTCAATTTATTAAAGAATGCACATAAC
GATGTGAAATTATTGATTGG

>Sm-lncRNA 11 TCONS_00012478 gene=XLOC_007735

CGGTGGGCAACTTGGAGGAGGTCTGACCTTTTTACAGTGCCTTGCATGATGAATGAA
TCTGTCTGTTCTGTCTTTCGTGACAAACTAAAGCGTTTACTTCACTGGGGTCCCATCT
CTACGATCTGGATTATATTTTTTATTACACTCACGAGTTTATACAGTACCTTACATGT
TGCCCCTCCCTGTTCATCACTTCTTGGTCTGTTTCTATCTACGTGCATATTTAGTTCCT
TTTATATGATTCTGAAGTCTTACCTTTGTGCTGTATTTGTTGGCCCAGGATTTGTTCCT
CTGGGTTGGAGGCCGTGTGACCCTTCAGCTGAACAGAAGCTTCAGTTCTGTAATGTT
TGTAGAGGTTTTAAGCCCCCACGAGCCCATCATTGTCGAGCTTGCAACCGGTGCATA
ATGAAAATGGACCATCATTGTCCCTGGATAAACACATGTTGCGGTCATCTTAACCAT
AAATATTTTTTGATATTTTTGTTGTTTGCTCCATTTGGATGCATCACCTCTTGTATAGC
ACTATTGCTTTCTATATATCAAAGTCCAGCAATCCTGTGGGTGTTGCACTGTCTGTTG
GGATTTTAGCCATTTTCCAGTTGAAAGCTGTTGCAAGGAATCAAACGGGAATAGAAT
CATGGATCGTAGCGAAAATTGATTGAATCAGTGGCTTGTCTATATGTGAAGTTTCAT

CACTGTGACTATTTTAATCTATTTAGTCTCATTATTTTTGAAGGCAAACGTTTGGCGA
AAAGATGTAGGGGAAAAAAAACCATTTCGCTATCCATATGACCTTGGAAAGATCGG
AAATTTTCAGCAGATTTTTCTGTGGTCTGGTAAAGTCTTAGGAGATGGATACTACTG
GCCTGTTGTGAAAGGCTGCACTCAGTATGACTTGACTTTGGAACAGATCTATCAAAA
AAGGTTGAAGCAGAAAATTCAGCGCACATTCAAAATTACCCGAAATTATGATGGGA
GTCGGTGTTTGTGTTTTCGTTATGGATGTCTAACTGCGATACGTTCCCCATGTTTTGA
AGAACCCAGAATACCTGTACGTGTCGGTGATGTTCTTATGGTGACAAGAGGAACAA
AGTACTGGATTTATGGTCATTTGGTCCCTTCTGAGAGTTTTGGTGATTTTTCAGATTC
TGTCGAAACTCGTGGATGGGTTCCTCGTGTTTGTGCTTTGGAAGTCGGTTTCAAGCA
CAAAAACGACAAGTTCTCTCTTAAAAACGACTAGGCAATCACATTCCATCGTTCAGA
GGACGATCCGTTTTCTATTTTGTAAATTACTACAGTGTTCGTACTAATTATCGTTACT
TTACTCACTTTTGTAATTTCTTTTCTTGTTCTGTTCTTTCATTTTATTTTCCTATTTACT
TTCTCTTATCTAAAAGATTGAGTGGGATGAAATTATTCTAATAACACAAACTGAACA
TCTCAGTCTAAGATTTGTCTAGATGTATTGTTTACCATCATTCTAACCACCAATCTTA
TGAAGTGACCCAAATGTTAGAGGAGCGTACTAGTAGTTGGTGTTTGTTTAACACAAG
CTTTTCATATAG

>Sm-lncRNA 12 TCONS_00010393 gene=XLOC_006480

TGCACTTGACACTAACCAGGCTGATGTTAAGCAAAATCCAGAAGAGATTCAAGACT
TGACTGCATCCATTTCAGTTATGCTGAAAAATATCCAGGAATCATTCAATACAATGA
GTGAACAGCTCCTAGAAAAAATAGATGATTTATCAAAGAGAGTAGATGATGTTGAA
AAGAATATTGGTGAAATTATCAGTAGTCTTGATGATGATGGCGGTAATGAATAATAA
TAATAAACGTC

>Sm-lncRNA 13 TCONS_00010903 gene=XLOC_006766

AATAGAAAAGTGCATTCTATCTCAATCTGGTACTTTTCTTCTGGTTGAGTTTTAATAA
TGACTGAACGCAAAGTCCTAAATAAATATTTCCCTCCAGACTATGATCCTTCAAAAA
TACCTCGTCTTAGAAGAGGTGACCGGCGTAAACAGTTTAATATTAGAACAATGGCTC
CATTTAATATGAGGTGACTTCATTTACAGTGTTGTCATTAGCATCAGGTGCAATACG
TGCAATGGTTATATATACAAAGCTAAAAAGTTCAATTCGCGAATGGAAACTGCCGA
AAATGTAGATTACTTGGGCCTAAGACATTATCGATTTTACATCCGATGTCCTTTATGC
TGTGCTGAGATTATTTGGCGTACTGATTTAGAAAGTGGCGACTATGTTCTAGAAAGT
GGGGCTAAAAGAAACTTCGAAGCATTAAAAACAGCGGAAGAACTAGAAGCTAAAC
GTCAAGCTGAAGAGGAAGAAGAACTTGCTAATAATCCAATGAAATTATTGGAAAAA
AGAACTGATCAAAGCAAACAAGAAATGGAAATGGTCGAAGTAATTGAAGACTTGAA
ACAGTTAAATCAACGTCAAGCCACCATGGAAGCAGATCATGTTCTTTTGAGGCAAAT
GTGGCGAGAAGAGGAAGCCCTCAAGGAGGCTGAAAAAGAAGAAGATGATGCACTA
ATAAAGGAGTTACTCGCTTCAAAATCCGAGCAAGTCCATTCATTACCATTGGAGTTT
GGAGGTGAAAATTCCTCTAGACCAATCAAGATGCCAGGTTCATGACGTTTTATCTAA
TTGTTAATATTTCTCTAGCTTTGAAAGATTTGATGAAAGCCAATAAAAAGGAACCAG
TAGAAATTGGGAAATCTTACCTAAAGCGACAACTTCAAGGTGTTTTACGGGTGCAAA
AGAGGCCAGTATCAGAAACTAAAATTCCAAAGCTTGATGGTGATTCTGAAGCATCA
GCCATGAATAACATAAATGTTTCAAATTCCTCAACGAATAATTCGTTAGCAAATGAT

AAGTCAAAAGATAGTAATGTTAATCAACAAGTGTCATGTACCAGTGATGATAACAA
TACTAGTACTCAATTGTATCAACCATTACCTGGAATGGTTTATTCAGACCATAGTGA
TAGTGATTGAGAGGGAAAAGACAGACAATTGTA

>Sm-lncRNA 14 TCONS_00011021 gene=XLOC_006840

TTCAGGATGATCAGATCTACTGGTAGTTTACCGGAATCATCCACAACTTTTAATAAA
CCAAACTGTGACAGAGACGGCCAACGCAATACAGTTAACAGTTTGAATAATGTTGTT
GTTGTTACAGTACCATCAGTTAATCTGACAAATCCATCTACATCAGTGATTGTTGTTA
GTCAATCGACTACTACGATGTCTAGTTCTAATCGTAATTCTCGAAATAATTTTAATTA
CAAATGTTCATGCAAAGAGTTAGCGTCTCATCCTTCTCATCGTTCCAATCACTATTCA
TCTGGTGTTCATTCCTCACATTTTCAACATCATGGTAATCATCATCATCATCATCACC
ATCATCACTATAATCATCCTGGTCATAAAGAAAATCACACCTCTCATAGTGTTGTTG
AAAGTAGTAATTCATCTCCCACCAGTGGGACAAGTTCAAGTCATCATTATTCTTCAC
ATTATTTTACATCTAATCAAACTGTTCCAACTTTTCAACACAACATAAAGCATAATTC
ATTCCCCATTAAATGGAAACCAGCGATAAGTAGACACTTACAAGAGTTCATGATCA
GTCAATGGGAACAGTTTCATGCACATCTGTTACATTCGAATAATATTGTTACAAAAT
CCTGCTGTACCGGAGCAACAACAATAAATACTGGCACACTGACAACAGTAGCCATT
AACAATCAGCTTACCACATCTGGTGATTCTACTAATAATAATGACAGCCCATCATCA
TCAAATGACATGACTTCATCAAGAACTAGTTTATCTGATAATTCTTCGTCTGGACAA
ATTGCCAACACCCAATCTAATAATAATGGTAGTAATAATAATAGTGGTAATAATTCA
CCTTCGATAATTGAATCAAATATTATTGGTCGTAGTAGCGGTAGTAGAGTGAATGAT
TTACCTGTCACACTAGCTACTATTACTACTACTATTAGTAATACAACACATGGACCC
GGTAGTAATAATGTCAGTAGTAACACTTGTGGTGTTGCGATGAGCAGTTGTAATAGT
TATCCTAGACGTAAACAAAGTTGTGGATATGTTACAAACCAAATGGGGAAGGGTAT
GAATACAGGTTCATTGGGATGGAAAAGAAATCAAACTTCAGGAAGAAAAATCTCTT
CACCTGGGAATTTTAACAATCGATATCCTGTAAATTACTTCAGTAATCGAAACTCTT
CATTCATTATATCTACTACTACTACTACTACTACTACTACTGCTACATCAAATAGTGT
GCTTCCTGTCGTCACGTATATTTCATCATCTATAACCACCACAACCATCACAACAAC
AACAACAACTCCAACAGTAGTGAATTCATCAAACATTGATCAAATGAGTATCACCA
CTGTAACCATAAATAATCAATCCAATTTGCAGAGTTCAAATGTGATTGATTTTAAAA
CGAATACTTCAGTACCCTTACTATTGTCATCATCCTCAGAGGGTCGTATACAACAGA
AAGTTTCTTCTTCTCTTCCTGAATCCAATTCTATTTTTAAAATTGATATTTACAACTCT
TCGACTGTAACGAATTCAATAATAAATTCACAAAAATTGGAAATAAAAACAGAGGA
TGTTGATAAAGTTTCAGATACGAAAATTCATGTTAACAACAACTGTGTTGTTATAGC
ACAAAAACATTTGTCTGATTTACACCTTACCAGTGGTGATAATAACAATACAATTGA
TAAGTTGAAAGAAGATATTCCAATTGATGAATCGACGAAAAATTCATGTATTACTTT
ATCCAATTCATCATTATTATTATTGCCTGAAGAAGAAAGAACATTAACCATTGCATC
AATTTTGAAAGTTGATCAAGAAATTGGAAGTTCTATTAGCTGTGGAAAAGTTGAAGA
AGGTGAGTGCTACTGGACTCCGGATCCTCCAGCCTCTCCTGAAAATAACACTACCCC
ACATCATGAAAATTATTCCAGTAGTTATCATTCAACTAGTGTATCTCCAAGTCCTCCA
CAACTATCAGTTTCATTACCATGCTCAAAAGATTTGCACACTACTGCCTATCATCCAC
CGCTTAGTATTCGCACAGAATTATCTTCAATGCTTGATTTTCGTCGATTCAATCAACC
CAGAGATTCTCCGCTTATCTCATCATCTAGGTTAGAAAGCCTCAATGATGTCGATTA

AAAGTTTAATAAGAATTTTTTTTTCTATTGTCTATTATAATCACCAATATTTAATTGC
CTCCAGGTTTTCTTTATTTCAAAGAACTAAAACCCTCCTTACTTTTATTTT

>Sm-lncRNA 15 TCONS_00013835 gene=XLOC_008390

AGGCAAAGACTATTGGGTTCGAGTACTTGACAGTATTTCACACTCAGACATTTCGCC
ACCTGCTTTGGTTTAAACAGCTACATAGTACCGCTGGTCCCCATGCAAGTGTGATCC
GAAGCCGAATACATAAACCTGTACATGGTAAATGAGTGTAATAACTTGATTGTGTTT
TGTAATGGATACATCATAAAGGCATTCTGTTTTTGTTTTGCAGATGTTTTGCCTGCAG
CTGATAATCCCACTGCTTTTTTTACCGAAACCATCTAACCCGGCCGCAGATATCTATC
ATGCAATATATATACTTTTATTTCTTGTTTGTTTCTTCTTAGAGCGTAAACCATGTGG
AATATGTGCTATTATTTTGTTTATTTTCATTATTCTTCCATGTTATAGTTCCTTAGATA
ATTTATGTATTTTCACCACATGCAGTAAAGGCCGAACTTTGACGTAGTGTCTGACGT
ATATATATATTCTGACTTGACCATTGTTATGGTGTAATGTAAATATTGTTCCCCGAGT
AGAGAACCCTTCTCTTGCTTTCATTTGAAATTTCCGAGATTTCATGATGGATTAGCCG
ATTTTTACGAGTTGGTATGGATATTTACACCAACCTTTGTGCGCATAGTTTTTCAATT
TACAATAGATTTTTGGTGACTATATACCT