Universidade de São Paulo
Instituto de Matemática e Estatística
Programa Interunidades de Pós-graduação em Bioinformática

LUCAS FERREIRA MACIEL

Identificação e anotação de RNAs longos não-codificantes em *Schistosoma mansoni* e *Schistosoma japonicum*

São Paulo
2020

LUCAS FERREIRA MACIEL

Identificação e anotação de RNAs longos não-codificantes em *Schistosoma mansoni* e *Schistosoma japonicum*

Versão Corrigida

Dissertação apresentada ao Programa Interunidades de Pós-graduação em Bioinformática para obtenção do título de Mestre em Ciências pelo Instituto de Matemática e Estatística da Universidade de São Paulo.

Área de concentração: Bioinformática

Orientador: Prof. Dr. Sergio Verjovski-Almeida

São Paulo

2020

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

FICHA CATALOGRÁFICA

Maciel, Lucas Ferreira
M152 Identificação e anotação de RNAs longos não-codificantes em Schistosoma mansoni e Schistosoma japonicum / Lucas Ferreira Maciel, orientador Sergio Verjovski-Almeida. -- São Paulo : 2020.
135 p.

Dissertação (Mestrado) - Universidade de São Paulo
Orientador: Prof. Dr. Sergio Verjovski-Almeida

Programa Interunidades de Pós-Graduação em Bioinformática
Área de concentração: Bioinformática

1. Parasitologia. 2. RNAs longos não- codificantes. 3. Análise de co-ex - pressão. 4.Transcriptômica. 5. RNA-seq. I. Verjovski-Almeida, Sergio, orientador. II. Universidade de São Paulo. III.Título.

CDD: 616.96

Elaborada pelo Serviço de Informação e Biblioteca Carlos Benjamin de Lyra do IME-USP, pela Bibliotecária Maria Lucia Ribeiro CRB-8/2766

Nome: MACIEL, Lucas Ferreira

Título: Identificação e anotação de RNAs longos não-codificantes em *Schistosoma mansoni* e *Schistosoma japonicum*

Dissertação apresentada ao Programa Interunidades de Pós-graduação em Bioinformática para obtenção do título de Mestre em Ciências pelo Instituto de Matemática e Estatística da Universidade de São Paulo.

Aprovado em: 23/03/2020

Banca Examinadora

Prof. Dr. Eduardo Moraes Rego Reis
Instituição: IQ-USP
Julgamento: Aprovado

Prof. Dr. Thiago Motta Venancio
Instituição: UEFN
Julgamento: Aprovado

Profa. Dra. Júlia Pinheiro Chagas da Cunha
Instituição: IB(ICB) - USP
Julgamento: Aprovado

AGRADECIMENTOS

Agradeço a minha namorada Renata, por ter me apoiado e estado presente desde o momento em que tomei a decisão de prestar a prova para ingressar no programa até o final desta dissertação, mesmo com a longa distância em uma parte deste período. Você tem uma participação muito importante neste trabalho.

Às minhas Avós Ana e Terezinha, por terem me criado e educado com tanto amor e carinho, vocês foram uma das maiores motivações de todo meu esforço.

Ao meu Pai e minha Mãe, por sempre colocarem os estudos como a coisa mais importante, e por entenderem e apoiarem a carreira que escolhi seguir, apesar de ainda não saberem o que eu faço exatamente.

Ao meu amigo David, que colaborou de maneira direta para a produção dos resultados apresentados nesta dissertação e por todas nossas muitas conversas e divagações ao longo desses anos, acadêmicas e não acadêmicas. Agradeço também a Daisy e a Adriana pela amizade e parceria que formamos ao longo destes anos. Vocês, junto com os demais colegas de laboratório, tornaram a caminhada mais fácil.

Ao Prof Dr. Sergio Verjovski-Almeida, por toda confiança depositada no meu trabalho, pela liderança científica e por nos dar excelentes condições de trabalho mesmo em tempos nebulosos para a ciência brasileira.

A Universidade de São Paulo e ao Instituto Butantan por fornecerem a estrutura necessária para o desenvolvimento de pesquisas.

Por fim, agradeço a Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) por todo auxílio e financiamento fornecidos para a execução deste projeto (2018/19591-2).

Pode se encontrar a felicidade mesmo nas horas mais sombrias, se a pessoa se lembrar de acender a luz.

Alvo Dumbledore

RESUMO

MACIEL, L. F. **Identificação e anotação de RNAs longos não-codificantes em *Schistosoma mansoni* e *Schistosoma japonicum*.** 2020. 135 pp. Dissertação (Mestrado em Ciências) – Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2020.

Os RNAs longos não-codificantes (lncRNAs) (> 200 nt) são expressos em níveis inferiores aos dos RNAs mensageiros (mRNAs) e, em todas as espécies modelo eucarióticas onde foram caracterizados, são transcritos a partir de milhares de diferentes *loci* genômicos. Em humanos, cerca de quatro dezenas de lncRNAs foram estudados em detalhes e demonstraram desempenhar papéis importantes na regulação da transcrição, agindo em conjunto com fatores de transcrição e marcas epigenéticas para modular os programas de ativação e repressão da transcrição gênica tecido-específica. Trabalhos anteriores identificaram cerca de 10.000 e 3.000 lncRNAs em *Schistosoma mansoni* e *Schistosoma japonicum*, respectivamente. Estes são os vermes que causam a esquistossomose, uma importante doença tropical negligenciada. O número limitado de bibliotecas de sequenciamento de RNA (RNA-seq) que haviam sido previamente avaliadas, juntamente com o uso de versões antigas e incompletas dos genomas de *S. mansoni* e *S. japonicum* e suas anotações no transcriptoma de codificação de proteínas, dificultaram a identificação de todos os lncRNAs expressos nesses parasitas. Aqui foram utilizadas 66 bibliotecas de RNA-seq de *S. japonicum* de vermes inteiros em diferentes estágios do ciclo de vida e 633 bibliotecas de RNA-seq de *S. mansoni* de vermes inteiros de diferentes estágios, de tecidos isolados, de populações de células e de células únicas para identificar lncRNAs. Todas as bibliotecas foram obtidas do domínio público. Um *pipeline* foi desenvolvido e utilizado para o mapeamento de transcritos contra os genomas e a montagem dos genes. Um conjunto de 16.583 transcritos de *S. mansoni* e 12.291 transcritos de *S. japonicum* foram identificados e anotados como lncRNAs; as análises de identidade de sequencia e sintenia dos genes entre as espécies demonstram que alguns lncRNAs possuem conservação de sintenia, mesmo quando há falta de conservação da sequência. Análises de redes de co-expressão gênica ponderadas foram realizadas para ambas as espécies e lncRNAs com expressão dinâmica através do desenvolvimento do parasita foram identificados. Esses lncRNAs são co-expressos com genes codificadores de proteínas associados a várias importantes vias biológicas, como reprodução, replicação, metabolismo de medicamentos e desenvolvimento do sistema nervoso. Este

trabalho abre caminho para a futura caracterização funcional do papel dos lncRNAs nos esquistossomos.

Palavras-chave: Parasitologia, RNAs longos não-codificantes, análise de co-expressão, transcriptômica, RNA-seq.

ABSTRACT

MACIEL, L. F. **Long non-coding RNAs identification and annotation in *Schistosoma mansoni* and *Schistosoma japonicum*.** 2020. 135 pp. Dissertation (Master in Sciences) – Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2020.

Long non-coding RNAs (lncRNAs) (>200 nt) are expressed at levels lower than those of the messenger RNAs (mRNAs), and in all eukaryotic model species where they have been characterized, they are transcribed from thousands of different genomic loci. In humans, some four dozen lncRNAs have been studied in detail, and they have been shown to play important roles in transcriptional regulation, acting in conjunction with transcription factors and epigenetic marks to modulate the tissue-type specific programs of transcriptional gene activation and repression. Previous works have identified around 10,000 and 3,000 lncRNAs in *Schistosoma mansoni* and *Schistosoma japonicum*, respectively. These are flatworms that cause schistosomiasis, an important neglected tropical disease. The limited number of RNA-sequencing (RNA-seq) libraries that had been previously assessed, together with the use of old and incomplete versions of *S. mansoni* and *S. japonicum* genomes and their protein-coding transcriptome annotations, have hampered the identification of all lncRNAs expressed in these parasites. Here 66 *S. japonicum* RNA-seq libraries from whole worms at different life-cycle stages and 633 *S. mansoni* RNA-seq libraries from whole worms at different stages, from isolated tissues, from cell-populations, and from single-cells were used to identify lncRNAs. All libraries were obtained from the public domain. A pipeline was developed and used for transcripts mapping to the genomes and gene assembly. A set of 16,583 *S. mansoni* and 12,291 *S. japonicum* lncRNA transcripts were identified and annotated; gene sequences identity and synteny analyses between the species demonstrate that some lncRNAs have synteny conservation even when there is a lack of sequence conservation. Weighted gene co-expression network analyses were performed for both species and lncRNAs with dynamic expression through parasite development were identified. These lncRNAs are co-expressed with protein-coding genes associated with several important biological pathways such as reproduction, replication, drug metabolism, and nervous system development. This work paves the way towards future functional characterization of the role of lncRNAs in schistosomes.

Keywords: Parasitology, long non-coding RNAs, co-expression analysis, transcriptomics, RNA-seq.

SUMÁRIO

1 INTRODUÇÃO.....15

1.1 ESQUISTOSSOMOSE.....15
1.2 RNAS LONGOS NÃO-CODIFICANTES EM EUCARIOTOS.....18
1.3 RNAS LONGOS NÃO-CODIFICANTES EM *SCHISTOSOMA SPP.*.....20
1.4 REANÁLISE DE *DATASETS* DE RNA-SEQ EM *S. MANSONI* E *S. JAPONICUM* COM FOCO EM LNCRNAS.....22

2 OBJETIVOS.....25

2.1 OBJETIVO GERAL.....25
OBJETIVOS ESPECÍFICOS.....25
2.2.....25

3 MANUSCRITOS.....26

3.1 CAPÍTULO I – *S. MANSONI*.....26
Abstract.....27
Introduction.....28
Material and Methods.....29
Results.....36
Discussion.....43
Data Availability Statement.....46
References.....46
Tables and Figures.....53
Supplementary Figures.....68
3.2 CAPÍTULO II - *S. JAPONICUM*.....78
Abstract.....79
Introduction.....80
Results.....81
Discussion.....88
Methods.....89
Data Availability.....91
References.....92
Tables and Figures.....97

4 CONCLUSÕES E PERSPECTIVAS**.....107**

5 REFERÊNCIAS.....109

ANEXOS.....115

1 INTRODUÇÃO

1.1 Esquistossomose

Esquistossomose é uma das principais doenças tropicais negligenciadas, causada por vermes da classe *Trematoda* e gênero *Schistosoma*, com estimativas de mais de 250 milhões de pessoas crônica ou agudamente infectadas, e responsável por 200 mil mortes anualmente apenas na África Subsaariana (Who, 2015). As três principais espécies do gênero que infectam os humanos são *S. haematobium*, presente no continente Africano, *S. japonicum*, prevalente na Indonésia, China e sudeste Asiático, e *S. mansoni*, distribuído no continente Africano e América Latina (Cdc, 2018) **(Figura 1)**. Na América, estima-se que 1-3 milhões de pessoas estão infectadas com *S. mansoni* e 25 milhões vivem em áreas de risco, sendo o Brasil e Venezuela os países mais afetados (Zoni *et al.*, 2016). A prevalência destas doenças está intrinsecamente relacionada a fatores econômicos e ambientais (Gomes Casavechia *et al.*, 2018).

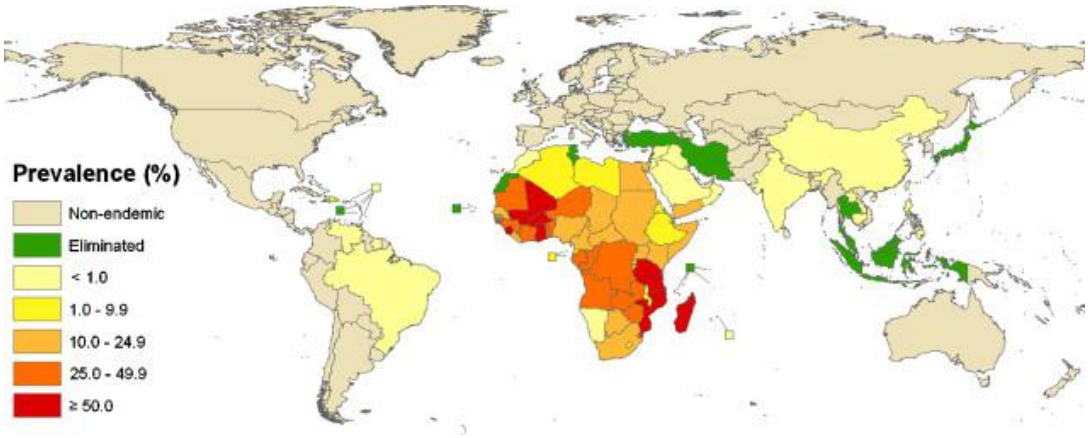


Figura 1 – Distribuição global da esquistossomose. Distribuição global estratificada de acordo com estimativas da prevalência da esquistossomose nos países e indicada pela escala de cores. Fonte: Utzinger *et al.* (2011).

Como representado na **Figura 2**, o ciclo de vida destes parasitas é complexo, sendo dividido em diversos estágios: 1) os ovos ao serem excretados do corpo humano nas fezes entram em contato com a água, e em condições ideais eclodem e liberam miracídios; 2) estes miracídios então nadam até encontrar caramujos, que são seus hospedeiros intermediários, penetrando o seu tecido; 3) durante o período em seu hospedeiro intermediário se desenvolvem para esporocistos por divisão assexuada e posteriormente se tornam cercarias; 4) As cercarias então são liberadas do caramujo na água, onde podem entrar em contato com os humanos, seu hospedeiro definitivo, e penetrar ativamente a pele ou mucosas; 5) Ao penetrar

a pele humana estes organismos perdem sua cauda e transformam-se em esquistossômulos; 6) esquistossômulos por meio da circulação sanguínea são capazes de migrar e se alocar em diversos tecidos, que são nichos espécie-específicos. No caso de *S. mansoni* e *S. japonicum*, os pulmões são os primeiros órgãos colonizados, posteriormente passando pelo coração, e migrando para o fígado onde alimentam-se, diferenciam-se sexualmente e atingem a fase adulta; 7) organismos adultos então fazem uma migração retrógrada, contra a circulação do sangue, para as veias mesentéricas, local em que ocorre o pareamento entre machos e fêmeas, que irão produzir centenas de ovos diariamente. Estes ovos podem migrar através da barreira endotelial e da parede dos intestinos, quando caem no lúmen intestinal e são liberados no ambiente junto com as fezes, e caso entrem em contato com a água poderão iniciar um novo ciclo (Wilson, 2009; Collins *et al.*, 2011; Gray *et al.*, 2011; Cdc, 2012). Muitos destes ovos, em lugar de excretados, são carreados pela circulação sanguínea para longe do sítio de oviposição, acumulando em outros tecidos como o hepático, onde causam as imunopatologias típicas da doença, como granulomas fibróticos do fígado (Hoffmann *et al.*, 2002; Burke *et al.*, 2009).

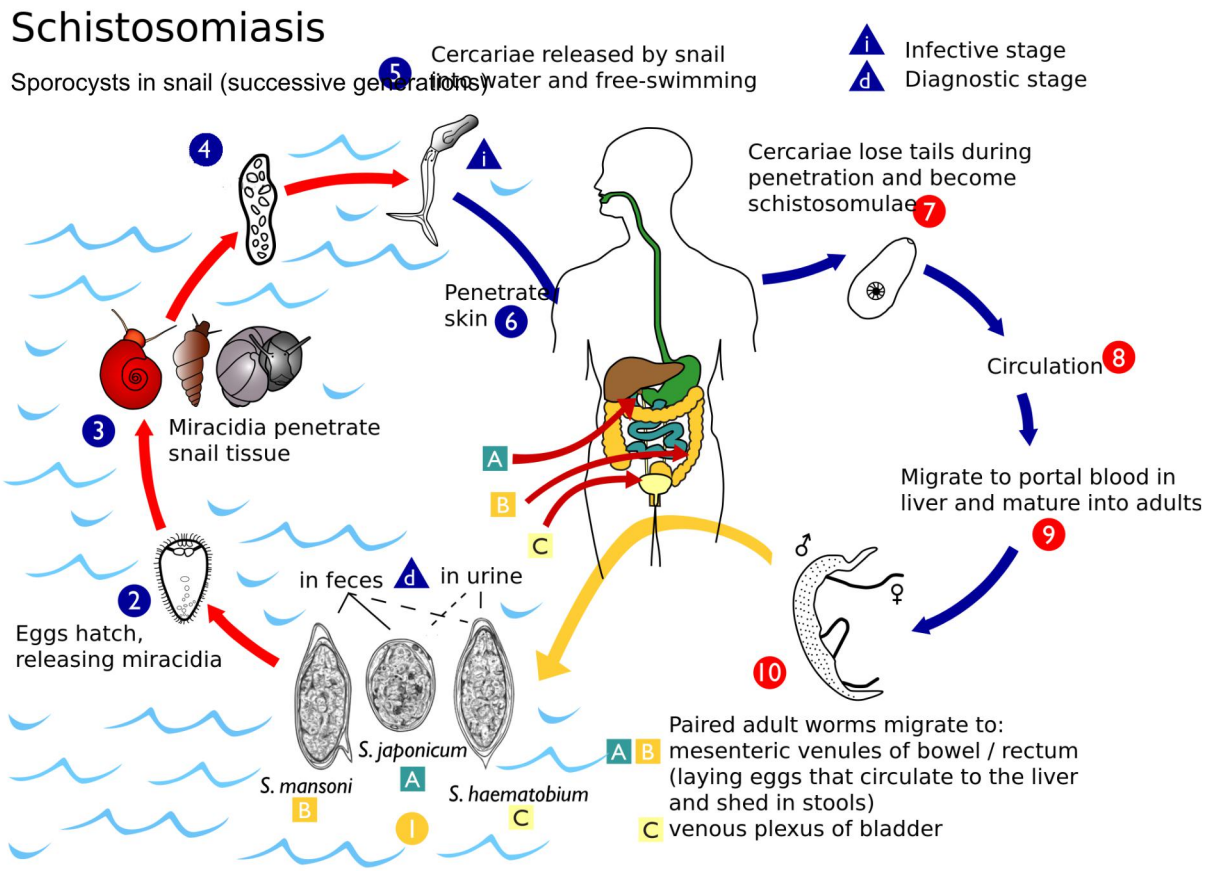


Figura 2 – Ciclo de vida de parasitas do gênero *Schistosoma.* Esquema representando o ciclo de vida dos parasitas *S. japonicum* (A), *S. mansoni* (B) e *S. haematobium* (C) incluindo todos os estágios de desenvolvimento intra e extra-hospedeiros. Fonte: Cdc (2012).

Existe apenas uma droga eficiente, barata e amplamente disponível no tratamento de esquistossomose, o Praziquantel (PZQ) (Cupit e Cunningham, 2015). Apesar de eficiente, o uso desta droga está suscetível à emergência de organismos resistentes e por isso tem-se buscado novas estratégias de tratamento. Devido a clara importância da produção de ovos tanto na transmissão do patógeno quanto para o surgimento das imunopatologias, novos alvos terapêuticos e drogas que afetem negativamente a biologia reprodutiva e oviposição têm sido amplamente estudadas e buscadas na literatura (Fitzpatrick *et al.*, 2009; Geyer *et al.*, 2011; Picard *et al.*, 2016; Geyer *et al.*, 2018).

A abordagem de *RNA-sequencing* (RNA-Seq) têm sido uma das mais aplicadas, isso devido a sua capacidade de avaliar alterações nos níveis de expressão gênica ao longo dos diferentes estágios ou condições/tratamentos, levando a uma melhor compreensão da biologia reprodutiva do organismo e seus mecanismos regulatórios, sendo estes potenciais alvos para novas estratégias terapêuticas (Lu *et al.*, 2016; Picard *et al.*, 2016; Lu *et al.*, 2017; Wang *et al.*, 2017).

Dois trabalhos recentes, um em *S. mansoni* e outro em *S. japonicum*, que buscaram entender o processo de maturação sexual dos organismos por meio de RNA-Seq, podem ser destacados. Lu *et al.* (2016) analisaram o perfil de expressão gênica de machos e fêmeas obtidos por meio de processos de infecção com *S. mansoni* pareados (infecções com parasitas de ambos os gêneros) ou não pareados (infecções apenas com parasitas machos ou fêmeas). Além de analisar os adultos inteiros, estas análises também foram realizadas com ovários e testículos dissecados e isolados de adultos obtidos nas mesmas condições de infecção. Foi demonstrado então que fêmeas não-pareadas são atrofiadas sexualmente e possuem um perfil de expressão gênica muito mais semelhante a machos, pareados ou não, quando comparado a fêmeas pareadas e maduras sexualmente. Já Wang *et al.* (2017) fizeram uma mapeamento detalhado do perfil transcricional de machos e fêmeas de *S. japonicum* em oito tempos diferentes ao longo do processo de desenvolvimento sexual, desde o pareamento até a completa maturação. Nesse caso os resultados também demonstraram que a fêmea imatura é muito mais semelhante aos machos do que as fêmeas maduras sexualmente. Além disso, ambos estudos destacam a importância de processos neurais durante o processo de maturação sexual da fêmea induzida pelo pareamento com o macho, principalmente por meio de neuropeptideos e aminas biogênicas.

Apesar de extremamente interessantes os resultados, em ambos os casos os autores não exploraram um grande potencial, os RNAs longos não-codificantes (*long non-coding RNAs*, lncRNA), que certamente devem ter sido sequenciados em suas amostras de RNA-seq, mas que não foram nem identificados, nem anotados, nem quantificados naquelas amostras.

1.2 RNAs longos não-codificantes em eucariotos

LncRNA são definidos como transcritos com mais de 200 pares de bases (pb), sem aparente potencial codificante de proteína (Cao *et al.*, 2018). O termo "aparente" potencial codificante é incluído porque já se sabe que alguns RNAs podem atuar na verdade como *dual function RNAs*, tendo funcionalidade tanto na forma de RNA longo não-codificante quanto por meio da expressão do RNA e sua tradução em um pequeno polipeptideo de até 100 aminoácidos (Nam *et al.*, 2016; Choi *et al.*, 2018). Isso torna a anotação e predição destes transcritos ainda mais complexa e aberta a debates na comunidade científica.

Os transcritos podem ser classificados de acordo com o seu tamanho, com seu mapeamento em regiões genômicas em relação a genes codificadores de proteínas (GCPs), ou em relação a outros elementos do DNA com função conhecida como *enhancers* e promotores, sequencias repetitivas, ou de acordo com a conservação da sequência e estrutura gênica entre espécies (St. Laurent *et al.*, 2015). Uma das classificações mais comuns, empregada pelo GENCODE e diversas *databases* específicas, é a associação com GCPs conhecidos, sendo divididos em três principais subtipos: (1) os lncRNAs antisenso (lnca), que mapeiam no genoma da espécie em regiões que se sobrepõem a regiões exônicas de GCPs mas na fita genômica de orientação contrária; (2) lncRNAs intrônicos senso (lncs), que mapeiam no genoma em regiões que se sobrepõem a regiões intrônicas dos GCPs na mesma orientação; e (3) lncRNAs intergênicos (*long intergenic non-coding RNA*, lincRNA), que mapeiam no genoma em regiões que não possuem qualquer sobreposição com GCPs **(Figura 3)**.

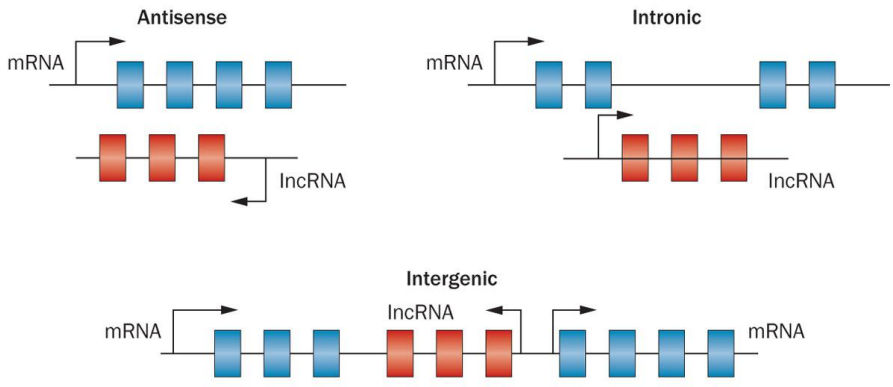


Figura 3 – Classificação de RNAs longos não-codificantes (lncRNAs). Representação esquemática dos lncRNAs, em vermelho, classificados como intergênico, antisenso e senso intrônico de acordo com sua posição relativa ao gene codificador de proteína (GCP) mais próximo, em azul. Fonte: Adapatado de Knoll *et al.* (2015).

Os mecanismos de ação e funcionalidade da vasta maioria de lncRNAs humanos recentemente descobertos ainda permanece um grande enigma, porém diversos estudos já demonstraram que lncRNAs podem atuar de diversas maneiras que estão ilustradas na **Figura 4**.

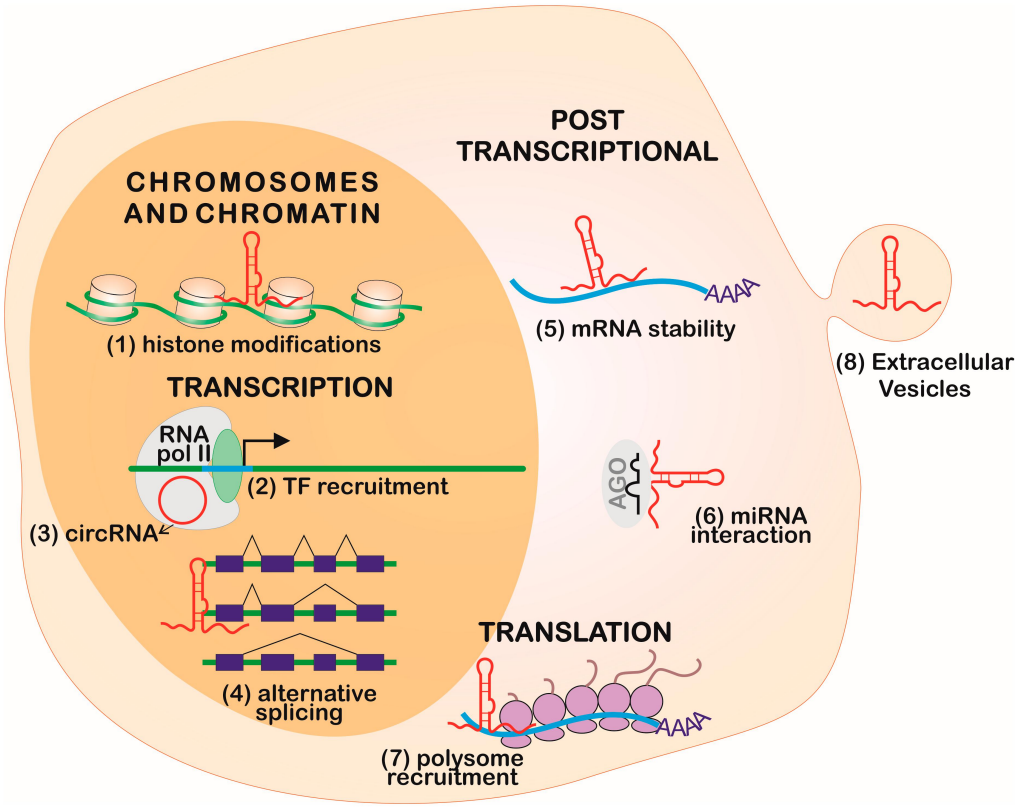


Figura 4 – Mecanismos de ação conhecidos de lncRNAs. Diferentes mecanismos de ação de lncRNAs já identificados em mamíferos, atuando no controle da expressão gênica. Fonte: Fernandes *et al.* (2019).

Um clássico exemplo de lncRNA funcional é o Xist (*X-inactive specific transcript*), um transcrito de aproximadamente 17 kilobases (kb), que atua como *scaffold* macromolecular que faz o recrutamento de proteínas formando um complexo ribonucleoproteico, e durante o processo de desenvolvimento de mamíferos fêmeas é o principal regulador do processo de silenciamento de um dos cromossomos X (Cerase *et al.*, 2015; Pintacuda *et al.*, 2017). Outra maneira comum de atuação de lncRNAs é como agentes reguladores da expressão gênica em cis, em genes vizinhos, ou em trans, de maneira mais global em genes distantes, podendo atuar, por exemplo, como *enhancers* e recrutar fatores de transcrição (Engreitz *et al.*, 2016).

Por meio destes diversos mecanismos já foi demonstrado em mamíferos que os lncRNAs participam da regulação de processos vitais como o ciclo celular (Kitagawa *et al.*, 2013), manutenção da pluripotência (Rosa e Ballarino, 2016) e reprodução (Golicz *et al.*, 2018). Em humanos a desregulação da expressão de lncRNAs também já foi associada a diversas doenças como câncer (Fang e Fullwood, 2016), Alzheimer (Zijian, 2016) e doenças cardíacas (Simona *et al.*, 2018).

Devido justamente a diversidade e complexidade da interação dos lncRNAs com GCPs, a grande maioria dos estudos envolvendo a utilização destes transcritos como alvos terapêuticos não passou dos testes com modelos celulares e animais, e a progressão para a clínica tem sido lenta (Harries, 2019). Entretanto, lncRNAs ainda representam alvos terapêuticos promissores (Matsui e Corey, 2017; Blokhin *et al.*, 2018; Harries, 2019). Como foi revisado por Matsui e Corey (2017), em camundongos modelo da síndrome de Angelman, a administração de oligonucleotídeos antisenso (*antisense oligonucleotides*, ASOs), que tinham como alvos para degradação o lncRNA Ube3a-ATS, reverteram parcialmente os déficits cognitivos associados a doença (Meng *et al.*, 2014). ASOs que alvejavam o lncRNA SAMMSON também foram utilizados com resultados positivos para o tratamento de melanoma (Leucci *et al.*, 2016).

Em parasitas, o *knock-down* de um lncRNA antisenso em *Plasmodium falciparum* induziu a repressão do gene var, levando a deleção da memória epigenética e alterando substancialmente o padrão de expressão gênica (Amit-Avraham *et al.*, 2015). LncRNAs apresentam uma menor conservação evolutiva entre espécies quando comparados a RNAs mensageiros (mRNAs) (Pang *et al.*, 2006; Johnsson *et al.*, 2014), o que dentro do campo da parasitologia poderia representar um grande potencial de alvos para terapias gênicas contra o parasita com menores chances de *off-targets* contra o hospedeiro.

1.3 RNAs longos não-codificantes em *Schistosoma spp.*

Com o crescente interesse e importância dos lncRNAs, três trabalhos recentes foram publicados com o objetivo de identificar lncRNAs em *Schistosoma spp.*. Nosso grupo foi responsável pela primeira publicação (Vasconcelos *et al.*, 2017), onde 88 bibliotecas de RNA-Seq de *S. mansoni* em 5 estágios de vida diferentes foram analisadas a partir de um *pipeline ad hoc*, tendo-se identificado 7029 lncRNAs de 2596 *loci* genômicos (Vasconcelos *et al.*, 2017).

Já o segundo trabalho (Liao, Qi *et al.*, 2018) utilizou um *pipeline* que já havia sido previamente desenvolvido para a identificação de lncRNAs em *P. falciparum* (Liao *et al.*,

2014), e a partir de 4 bibliotecas de RNA-Seq de *S. mansoni* e 2 de *S. japonicum* identificaram 3247 e 3033 lncRNAs, respectivamente. Não houve uma comparação com o primeiro trabalho de *S. mansoni* (Vasconcelos *et al.*, 2017) para se saber quantos destes lncRNAs descritos (Liao, Qi *et al.*, 2018) eram novos.

O terceiro e último trabalho (Oliveira *et al.*, 2018) utilizou 2 bibliotecas de *S. mansoni*, em um *pipeline* próprio, e foi capaz de identificar 170 novos lncRNAs transcritos (Oliveira *et al.*, 2018), quando comparado com o trabalho original de Vasconcelos *et al.* (2017).

Cada uma das publicações possui pontos fortes e pontos fracos. Vasconcelos *et al.* (2017) têm a seu favor um grande *dataset* e profundidade de sequenciamento, porém utilizou o Trinity como ferramenta de montagem, independente de mapeamento genômico, o que levou à geração de uma grande quantidade de isoformas putativas, que muitas vezes refletem apenas transcritos não maduros, com retenção de introns. Liao, Qi *et al.* (2018) e Oliveira *et al.* (2018) utilizaram ferramentas de mapeamento contra o genoma de referência e posterior reconstrução do transcrito, porém com um conjunto de dados muito limitado e de apenas um estágio, adultos. Além disso, cada trabalho definiu de maneira diferente os *thresholds* e parâmetros, como tamanho e cobertura das fases de leitura abertas (*open reading frames*, ORFs), e ferramentas diferentes foram empregadas. De tal forma, um transcrito que foi identificado como lncRNA em um *pipeline* não seria necessariamente considerado como lncRNA por outro pipeline.

Ademais, todos conjuntos haviam sido anotados contra as versões desatualizadas dos genomas de *S. mansoni* e *S. japonicum*. Em *S. mansoni*, haviam sido utilizadas a versão 5 do genoma, com uma montagem muito fragmentada, e a versão 5.2 do transcriptoma (Protasio *et al.*, 2012), que de maneira geral está incompleta, principalmente em relação às regiões não traduzidas dos transcritos (UTRs). Como um resultado disso, os transcritos identificados como lncRNAs pelos trabalhos anteriores ao serem mapeados contra a nova versão do genoma (versão 7), bem menos fragmentada e melhor anotada com a versão 7.1 do transcriptoma, demonstram que milhares destes na verdade representam pre-mRNAs parcialmente processados que estão localizados em novos GCPs, que até então não estavam anotados; estes transcritos foram classificados como sem potencial codificante devido a retenção de introns, como exemplificado pela **Figura 5**.

Este *locus* também exemplifica a redundância das anotações entre os trabalhos. Sendo assim, está clara a necessidade da obtenção de um conjunto de lncRNAs de *S. mansoni* coeso, mais confiável, que exclua possíveis formas imaturas de genes codificantes de proteínas, e

que esteja de acordo com a nova anotação dos genes codificantes de proteínas do genoma (v. 7.1) disponibilizada em agosto de 2018 pelo Sanger Institute, e ainda não publicada (https://parasite.wormbase.org/Schistosoma_mansoni_prjea36577/).

Em *S. japonicum*, as novas versões do genoma e transcriptoma foram recentemente publicadas, com melhoras relevantes de contiguidade e anotação (Luo *et al.*, 2019). Levando em consideração que o único conjunto de lncRNAs de *S. japonicum* foi anotado contra a antiga versão do genoma, utilizando apenas duas bibliotecas de um estágio, e utilizando o *pipeline* também aplicado em *S. mansoni* (Liao, Q. *et al.*, 2018), que possuia algumas dificuldades em relação a remoção de pre-mRNAs (**Figura 5**), também se faz necessário a re-anotação destes lncRNAs com um novo *pipeline* e um maior número de bibliotecas e estágios.

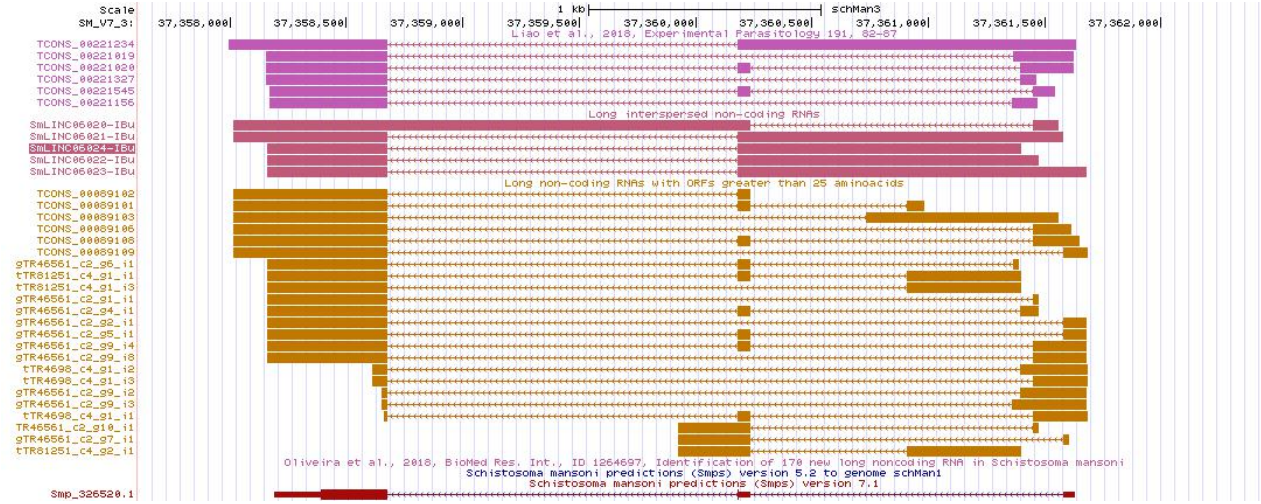


Figura 5 – Potenciais pre-mRNAs anotados como lncRNAs. *Locus* do genoma de *S. mansoni* exemplificando transcritos considerados como lncRNAs e que podem ser na verdade pre-mRNAs. As cores vermelho (a *track* mais de baixo), rosa e roxo (as duas *tracks* mais ao alto) representam, respectivamente, o gene codificador da proteína Smp_326520.1 da nova v. 7.1 do transcriptoma, e os hipotéticos lncRNAs de Vasconcelos *et al.* (2017) e Liao, Qi *et al.* (2018). Já o dourado representa transcritos que foram removidos por Vasconcelos *et al.* (2017) porque tinham ORFs maiores que 25 aa e que representavam mais de 25% da cobertura; vê-se que a isoforma madura predita pela nova anotação V.7 é uma das isoformas removidas pela análise de Vasconcelos *et al.* (2017) (a 11ª *track* de baixo para cima, nas *tracks* douradas). Fonte: Autoral, imagem gerada no UCSC-like genome browser de nosso grupo, disponível em http://genome.verjolab.usp.br/folders/geneNetwork/)

1.4 Reanálise de *datasets* de RNA-Seq em *S. mansoni* e *S. japonicum* com foco em lncRNAs

A grande maioria das bibliotecas de RNA-Seq utilizadas nas análises de ambas as espécies são provenientes de vermes adultos, trazendo assim um viés grande na identificação de lncRNAs neste estágio. Com isso, a introdução de bibliotecas de outros estágios,

principalmente em *S. japonicum,* onde até agora só haviam sido utilizados dados de adultos, pode auxiliar na anotação de um conjunto mais completo de lncRNAs.

Além disso, todas as bibliotecas utilizadas foram geradas a partir do organismo inteiro. Em mamíferos, muitos trabalhos demonstraram que, devido à grande importância da dinâmica de modificações da cromatina, muitos lncRNAs apresentam expressão tecido-específica (Washietl *et al.*, 2014; Wu *et al.*, 2016; Credendino *et al.*, 2017). Assim, as bibliotecas de sequenciamento geradas a partir de testículos e ovários de adultos pareados e não pareados de *S. mansoni*, geradas por Lu *et al.* (2016), representam uma grande fonte de novos lncRNAs, até então não analisados, com potencial papel no processo de maturação sexual e reprodução.

Os dados de sequenciamento de RNA de células únicas (*single-cell sequencing*, scRNA-Seq) a partir de células tronco de *S. mansoni* em diferentes estágios, gerados em dois trabalhos recentes (Tarashansky *et al.*, 2018; Wang *et al.*, 2018), também possuem grande potencial de identificação de lncRNAs, até agora não analisados. Apesar dos grandes desafios técnicos, o scRNA-Seq é capaz de identificar transcritos de lncRNAs raros, cuja expressão também pode ser célula-específica, como mostrado em mamíferos (Kim *et al.*, 2015; Luo *et al.*, 2015; Gawronski e Kim, 2017). A identificação e anotação de novos lncRNAs a partir dos dados de scRNA-Seq recentemente obtidos em *S. mansoni* poderia aumentar conhecimentos a respeito da participação de lncRNAs na biologia do organismo e seus mecanismos de manutenção e proliferação, similar ao que já foi mostrado para mamíferos (Kim *et al.*, 2015; Pal e Rao, 2017).

Por fim, a reanálise de dados de RNA-Seq de *Schistosoma spp.* com foco em lncRNAs pode ser um primeiro passo para a identificação de lncRNAs funcionais e um indicativo de seus papéis nos parasitas. Redes de co-expressão podem ser utilizadas para identificar a expressão dinâmica de lncRNAs e sua potencial função de acordo com os GCPs com os quais eles estão correlacionados, em um método denominado *guilt-by-association* (Lefever *et al.*, 2017). Estudos em mamíferos demonstraram que lncRNAs com expressão dinâmica ao longo do desenvolvimento possuem maior enriquecimento em marcas de funcionalidade (Sarropoulos *et al.*, 2019).

Estas análises de co-expressão serão realizadas com WGCNA (Langfelder and Horvath, 2008), uma ferramenta que através da análise de correlação e sobreposição topológica dos genes ao longo das amostras é capaz de clusterizar genes com perfil de expressão similares em grupos denominados módulos. Genes que compõe os mesmo módulos tendem a fazer parte da mesma via biológica. Sendo assim, lncRNAs que fazem parte destes módulos

provavelmente participam destas mesmas vias. Estes módulos também podem ser correlacionados a diversas variáveis, como os estágios, tecidos, gênero ou tempo de infecção.

Em *S. mansoni*, redes de co-expressão utilizando os mais diferentes estágios podem ajudar a indicar lncRNAs envolvidos na progressão do ciclo de vida do parasita, enquanto as gônadas podem ser exploradas para expressão tecido específica de lncRNAs associados a reprodução. Já em *S. japonicum*, os dados de maturação sexual gerados por Wang *et al.* (2017) são interessantes para identificação de lncRNAs com expressão dinâmica ao longo do processo de maturação sexual, também indicando o seu possível papel na biologia reprodutiva do organismo.

Assim sendo, o presente trabalho foi desenvolvido a fim de prover uma base para futuros estudos sobre o papel de lncRNAs na biologia de *S. mansoni* e *S. japonicum*, que pode eventualmente levar a identificação de potenciais alvos terapêuticos. Para isso, fizemos a anotação de um conjunto de lncRNAs para cada espécie, mais robusto e completo, em concordância com as anotações mais atualizadas dos transcriptomas, analisando *datasets* ainda não anotados para a presença de lncRNAs (gônadas e *single-cell* em *S. mansoni* e os diferentes estágios de *S. japonicum*) e a reanálise de *datasets* com foco no papel de lncRNAs ao longo da progressão dos estágios e da maturação sexual.

2 OBJETIVOS

2.1 Objetivo geral

Identificar conjuntos de lncRNAs em *S. mansoni* e *S. japonicum* mais robustos e completos, e reanalisar *datasets* até então não anotados para a presença destes lncRNAs, a fim de melhor compreender a biologia dos parasitas

2.2 Objetivos específicos

* Estabelecer *pipeline* para identificar e anotar lncRNAs a partir de dados de RNA-Seq;

* Anotar novo conjunto de lncRNAs a partir de dados públicos de sequenciamento dos mais diferentes estágios, tecidos e células-tronco de *S. mansoni*;

* Anotar novo conjunto de lncRNAs a partir de dados públicos de sequenciamento dos mais diferentes estágios de *S. japonicum*;

* Comparações de identidade e sintenia entre os dois conjuntos identificados;

* Realizar análises de co-expressão para identificar lncRNAs com expressão dinâmica ao longo da progressão dos estágios e tecido específica em *S. mansoni;*

* Reanalisar dados de scRNA-Seq para identificação de lncRNAs marcadores de populações celulares em *S. mansoni;*

* Realizar análises de co-expressão para identificar lncRNAs com expressão dinâmica ao longo da maturação sexual em *S. japonicum;*

3 MANUSCRITOS

Os métodos, resultados e discussões pertinentes a esta dissertação estão dispostos na forma de manuscritos, divididos em Capítulo I e Capítulo II, referentes aos trabalhos que foram publicados ou submetidos para publicação em periódicos internacionais.

3.1 Capítulo I – *S. mansoni*

Este capítulo é referente a identificação e anotação de lncRNAs e análises de co-expressão em *S. mansoni*. O manuscrito, intitulado *Weighted Gene Co-Expression Analyses Point to Long Non-Coding RNA Hub Genes at Different Schistosoma mansoni Life-Cycle Stages*, está disposto na forma como foi submetido e aprovado para publicação no periódico *Frontiers in Genetics* (ISSN**:** 1664-8021). O manuscrito foi escrito por mim, e revisado e aprovado pelos demais co-autores. Declaro que o trabalho aqui apresentado foi realizado por mim, exceto nas partes expressamente indicadas no texto abaixo.

O texto no formato publicado está anexado a esta dissertação (**Anexo A**) e pode ser também acessado no link https://doi.org/10.3389/fgene.2019.00823. Neste link também estão presentes os arquivos suplementares indicados no texto.

Weighted gene co-expression analyses point to long non-coding RNA hub genes at different *Schistosoma mansoni* life-cycle stages

Lucas F. Maciel [1,2], David A. Morales-Vicente[1,3], Gilbert O. Silveira[1,3], Raphael O. Ribeiro[1,3], Giovanna G. O. Olberg[1], David S. Pires[1], Murilo S. Amaral[1], Sergio Verjovski-Almeida*[1,3]

[1]Laboratório de Expressão Gênica em Eucariotos, Instituto Butantan, São Paulo, Brazil.

[2]Programa Interunidades em Bioinformática, Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, Brazil

[3]Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, São Paulo, Brazil

Abstract

Long non-coding RNAs (lncRNAs) (>200 nt) are expressed at levels lower than those of the protein-coding mRNAs, and in all eukaryotic model species where they have been characterized, they are transcribed from thousands of different genomic *loci*. In humans, some four dozen lncRNAs have been studied in detail, and they have been shown to play important roles in transcriptional regulation, acting in conjunction with transcription factors and epigenetic marks to modulate the tissue-type specific programs of transcriptional gene activation and repression. In *Schistosoma mansoni*, around ten thousand lncRNAs have been identified in previous works. However, the limited number of RNA-seq libraries that had been previously assessed, together with the use of old and incomplete versions of the *S. mansoni* genome and protein-coding transcriptome annotations, have hampered the identification of all lncRNAs expressed in the parasite. Here we have used 633 publicly available *S. mansoni* RNA-seq libraries from whole worms at different stages (n=121), from isolated tissues (n=24), from cell-populations (n=81) and from single-cells (n=407). We have assembled a set of 16,583 lncRNA transcripts originated from 10,024 genes, of which 11,022 are novel *S. mansoni* lncRNA transcripts, while the remaining 5,561 transcripts comprise 120 lncRNAs that are identical to and 5,441 lncRNAs that have gene overlap with *S. mansoni* lncRNAs already reported in previous works. Most importantly, our more stringent assembly and filtering pipeline has identified and removed a set of 4,293 lncRNA transcripts from previous publications that were in fact derived from partially processed mRNAs with intron retention. We have used weighted gene co-expression network analyses and identified 15 different gene co-expression modules. Each parasite life-cycle stage has at least one highly correlated gene

co-expression module, and each module is comprised of hundreds to thousands lncRNAs and mRNAs having correlated co-expression patterns at different stages. Inspection of the top most highly connected genes within the modules' networks has shown that different lncRNAs are hub genes at different life-cycle stages, being among the most promising candidate lncRNAs to be further explored for functional characterization.

Keywords: Parasitology, RNA-seq, Single-cell sequencing data, *Schistosoma mansoni*, long non-coding RNAs, weighted genes co-expression network analysis.

Introduction

Schistosomiasis is a neglected tropical disease, caused by flatworms from the genus *Schistosoma*, with estimates of more than 250 million infected people worldwide and responsible for 200 thousand deaths annually at the Sub-Saharan Africa (WHO, 2015). *Schistosoma mansoni*, prevalent in Africa and Latin America, is one of the three main species related to human infections (CDC, 2018). In America, it is estimated that 1-3 million people are infected by *S. mansoni* and over 25 million live in risk areas, being Brazil and Venezuela the most affected (Zoni et al., 2016). The prevalence of this disease is correlated to social-economic and environmental factors (Gomes Casavechia et al., 2018).

This parasite has a very complex life-cycle comprised of several developmental stages, with a freshwater snail intermediate-host and a final mammalian host (Basch, 1976). Recently, it has been shown that epigenetic changes are required for life cycle-progression (Roquis et al., 2018). However, little is known about the genes and molecules that drive this process through the life-cycle stages of *S. mansoni*. A better understanding of the gene expression regulation mechanisms and of their components may lead to new therapeutic targets (Batugedara et al., 2017), and one key element could be the long non-coding RNAs (lncRNAs) (Blokhin et al., 2018).

LncRNAs are defined as transcripts longer than 200 nucleotides, without apparent protein-coding potential (Cao et al., 2018). The term "apparent" is included because it is already known that some lncRNAs actually have dual function roles, being functional both as lncRNAs and through peptides shorter than 100 amino acids that they encode (Nam et al., 2016;Choi et al., 2018). In mammalians, lncRNAs regulate gene expression through different mechanisms (Bhat et al., 2016), including mediating epigenetic modifications (Hanly et al., 2018), and were shown to be important in vital processes such as cell cycle regulation

(Kitagawa et al., 2013), pluripotency maintenance (Rosa and Ballarino, 2016) and reproduction (Golicz et al., 2018).

In *S. mansoni*, the expression of lncRNAs at different life-cycle stages was first detected by our group in 2011 using microarrays (Oliveira et al., 2011). Subsequently, large-scale identification of *S. mansoni* lncRNAs has been reported in three studies from our group and from others that analyzed high-throughput RNA-sequencing (RNA-seq) data (Vasconcelos et al., 2017;Liao et al., 2018;Oliveira et al., 2018), but each of them has used a the limited number of datasets (from 4 to 88 RNA-seq libraries). Because each work used different mapping tools and parameters (Vasconcelos et al., 2017;Liao et al., 2018;Oliveira et al., 2018), and given that Liao et al. (2018) did not compare their lncRNAs with the previously published ones, part of the lncRNAs are redundant among the three reports. In addition, the lncRNAs were annotated against the old version 5.2 of the genome and protein-coding transcriptome (Protasio et al., 2012); as a result, a set of transcripts that were previously annotated as lncRNAs (Vasconcelos et al., 2017;Liao et al., 2018;Oliveira et al., 2018), seem now to represent partially processed pre-mRNAs arising from novel protein-coding genes annotated in the new version 7.1 of the transcriptome (https://parasite.wormbase.org/Schistosoma_mansoni_prjea36577/); these transcripts were previously annotated as having no coding potential due to intron retention, as exemplified in **Supplementary Figure S1**. Besides, these three works used expression data from whole parasites, while it is known from other species that lncRNAs have tissue- and cell-specific expression (Wu et al., 2016;Credendino et al., 2017).

The aim of the present work is to identify and annotate a robust and more complete set of lncRNAs that agrees with the most updated transcriptome annotation, and to analyze RNA-seq datasets still non-annotated for the presence of lncRNAs – e.g., gonads (Lu et al., 2016) and single-cell (Tarashansky et al., 2018;Wang et al., 2018) RNA-seq libraries. The goal is to provide a foundation that will enable future studies on the role of lncRNAs in *S. mansoni* biology, which could eventually identify potential new therapeutic targets.

Material and Methods

Transcripts reconstruction

In order to identify new lncRNAs, 633 publicly available RNA-seq libraries from whole worms at different stages (miracidia n=1, sporocysts n=1, cercariae n=8, schistosomula n = 11, juveniles n=9, adult males n=34, adult females n=37, and mixed adults = 20), from tissues

(testes n= 6, ovaries n= 5, posterior somatic tissues n = 3, heads n= 5 and tails n=5), from cell-populations (n=81) and from single-cells (from juveniles n=370 and mother sporocysts stem-cells n=37) were downloaded from the SRA and ENA databases (**Supplementary Table S1**). The only whole-worm stage that was not included was eggs, because there is a single RNA-seq library available in the public domain (Anderson et al., 2016), which has only 252,000 egg reads, an amount that is 4-fold lower than the minimum number of reads per library in the other whole-worm libraries that we used (namely 1 million good quality reads), being a too-low coverage for an unbiased detection of stage- or tissue-specific lncRNAs in complex organisms (Sims et al., 2014). The new versions of the genome (v 7) and transcriptome (v 7.1), which were used as reference in this study, were downloaded from the WormBase ParaSite resource (Howe et al., 2017) at https://parasite.wormbase.org/Schistosoma_mansoni_prjea36577/.

[Nota: O desenho e a construção do pipeline automatizado, que estão descritos no parágrafo abaixo, são o resultado de uma colaboração com um dos co-autores do artigo, o aluno David A. Morales-Vicente. A aplicação desse pipeline sobre os dados de RNA-Seq de *S. mansoni* e a geração dos resultados descritos na respectiva seção foram realizados por mim.]

Quality control was done with fastp v 0.19.4 (Chen et al., 2018) (default parameters), removing adapters and low-quality reads. The reads in each library were then mapped against the genome with STAR v 2.6.1c in a two-pass mode, with parameters indicated by STAR's manual as the best ones to identify new splicing sites and transcripts (Dobin et al., 2013). RSeQC v 2.6.5 (Wang et al., 2012) was used to identify RNA-Seq library strandedness to be used in transcripts reconstruction and expression levels quantification. For each library, multi mapped reads were removed with Samtools v 1.3 (Li et al., 2009) and uniquely mapped reads were used for transcript reconstruction with Scallop v 10.2 (--min_mapping_quality 255 –min_splice_boundary_hits 2) (Shao and Kingsford, 2017). A new splicing site should be confirmed at least by two reads in order to be considered. A consensus transcriptome from all libraries was built using TACO v 0.7.3 (--filter-min-length 200 –isoform-frac 0.05), an algorithm that reconstructs the consensus transcriptome from a collection of individual assemblies (Niknafs et al., 2016). As described by Niknafs et al. (2016), TACO employs change point detection to break apart complex loci and correctly delineate transcript start and

end sites, and a dynamic programming approach to assemble transcripts from a network of splicing patterns (Niknafs et al., 2016).

LncRNAs classification

In the consensus transcriptome, transcripts shorter than 200 nt, monoexonic or with exon-exon overlap with protein-coding genes from the same genomic strand were removed from the set. The coding potential of the remaining transcripts was evaluated by means of the FEELnc tool v 0.1.1 (Wucher et al., 2017) with the shuffle mode, that uses a random forest machine-learning algorithm and classifies these transcripts into lncRNAs or protein-coding genes, and also by CPC2 v 0.1 (Yang et al., 2017), that classifies through a support vector machine model using four intrinsic features. Only transcripts classified as lncRNAs by both tools were kept. ORFfinder v 0.4.3 (https://www.ncbi.nlm.nih.gov/orffinder/) was used to extract the putative longest open reading frames (ORFs); these putative peptides were then submitted to orthology-based annotation with eggNOG-mapper webtool (HMMER mapping mode) (Huerta-Cepas et al., 2017). Transcripts with no hits against the eukaryote eggNOG database were then considered as lncRNAs. If any transcript isoform was classified as a protein-coding mRNA at any step, all transcripts mapping to the same genomic *locus* were removed in order to avoid eventual pre-mRNAs. After this final step, a lncRNAs GTF file was created.

Histone marks

In order to identify histone H3 lysine 4 trimethylation (H3K4me3) and H3 lysine 27 trimethylation (H3K27me3) marks near the transcription start site (TSS) of lncRNAs we used 12 libraries of Chromatin Immunoprecipitation Sequencing (ChIP-Seq) data generated by Roquis et al. (2018) for cercariae, schistosomula and adults (**Supplementary Table S1**), which had more than 90% overall mapping rate. The reads were downloaded from the SRA database and mapped against the genome v 7 with Bowtie2 v 2.3.4.3 (Langmead and Salzberg, 2012) (parameters end-to-end,–sensitive,–gbar 4). As there are no input datasets publicly available in the SRA database for the Roquis et al. (2018) paper, we were not able to exactly reproduce the pipeline that was described in the Methods section of that paper, which used the input as a reference for peak calling. Instead, we used HOMER v 4.10 (Heinz et al., 2010) for removing multi-mapped and duplicated reads and for significant peak calling as described by Anderson et al. (2016), an approach also used by Vasconcelos et al. (2017) in the first large-

scale annotation of lncRNAs in *S. mansoni*. The number of reads in the peak should be at least 4-fold higher than in the peaks of the surrounding 10kb area and the Poisson p-value threshold cutoff was 0.0001. The lncRNAs with significant histone mark peaks within 1 kb distance upstream and downstream from their TSS were annotated. The lncRNAs with overlapping marks are shown with an intersection diagram that was plotted using the UpSetR tool v 1.3.3 (Lex et al., 2014). The Venn diagram tool at http://bioinformatics.psb.ugent.be/beg/tools/venn-diagrams was used for generating the lists of lncRNA genes belonging to each intersection set.

Co-expression networks

The lncRNAs GTF file was then added to the *S. mansoni* public protein-coding transcriptome version 7.1 GTF file and the resulting protein-coding + lncRNAs GTF was used as the reference together with the genome sequence v.7 for mapping the reads of each RNA-seq library under study, again using the STAR tool, now in the one pass mode, followed by gene expression quantification with RSEM v 1.3 (Li and Dewey, 2011). Weighted gene co-expression network analyses v 1.68 (WGCNA) (Langfelder and Horvath, 2008) were then performed in order to identify modules related to the life-cycle stages and tissues of the organism. For this purpose, only libraries from whole worms or from tissues with more than 50 % of the reads uniquely mapped were used. To reduce noise, only transcripts with expression greater than 1 transcript per million (TPM) in at least half of the libraries in one or more stages/tissues were considered. Expression levels were measured in log space with a pseudocount of 1 (log2(TPM+1)) and we set the transcript expression to zero when log2(TPM+1) < 1. For the construction of the adjacency matrix the power adjacency function for signed networks was applied with the soft-thresholding beta parameter equal to 14, which resulted in a scale-free topology model fit index R^2=0.935. The adjacency matrix was then converted to the Topological Overlap Matrix (TOM) and the dissimilarity TOM (1 – TOM) was calculated (Langfelder and Horvath, 2008).

Correlation between the modules and the stages was calculated based on the Pearson correlation coefficient between the expression levels of the transcripts belonging to each module along the stages, as suggested in the WGCNA tutorial (https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/) . As miracidia and sporocysts have only one library each, are closely related stages of development, and were clustered together as an outgroup based on their overall expression

patterns (as shown in the Results), we decided to consider both stages together as one group (miracidia/sporocysts) to calculate the correlation and p-values between modules and stages.

The Gene Trait Significance (GS) was calculated based on the correlation of an individual transcript and the trait, which in our case was always the stage of higher absolute Pearson correlation coefficient with the module where the transcript belongs. For example, for a transcript that belongs to the red module (most highly correlated with testes, see Results) the correlation was calculated between the expression of the transcript in the testes libraries and the expression of the transcript in all other non-testes libraries.

Gene Ontology (GO) Enrichment

Protein-coding genes were submitted to eggNOG-mapper (Huerta-Cepas et al., 2017) for annotation of GO terms. Based on this annotation (available at **Supplementary Table S2**), we performed GO enrichment analyses with BINGO (Maere et al., 2005). For each module we used a Hypergeometric test, the whole annotation as reference set, and FDR ≤ 0.05 was used as significance threshold.

Single-cell analyses

The expression levels were quantified in single-cell RNA-seq libraries from juveniles' stem cells (Tarashansky et al., 2018) and mother sporocysts stem cells (Wang et al., 2018) by RSEM. We used Scater v 1.10.1 (McCarthy et al., 2017) to normalize and identify high-quality single-cell RNA-Seq libraries, i.e. those that have at least 100,000 total counts and at least 1,000 different expressed transcripts, as recommended by McCarthy et al. (2017); all libraries were classified as high-quality.

Next, we used the R package Single-Cell Consensus Clustering (SC3) tool v 1.10.1 (Kiselev et al., 2017), that performs an unsupervised clustering of scRNA-seq data. Based on the clusters identified, we used the plot SC3 markers function to find marker genes based on the mean cluster expression values. These markers are highly expressed in only one of the clusters and indicates the specific-expression at the cell level. As described by Kiselev et al. (2017), the area under the receiver operating characteristic (ROC) curve is used to quantify the accuracy of the prediction. A p-value is assigned to each gene by using the Wilcoxon signed rank test. Genes with the area under the ROC curve (AUROC) > 0.85 and with p-value < 0.01 are defined as marker genes.

Ethics statement

All protocols involving animals were conducted in accordance with the Ethical Principles in Animal Research adopted by the Brazilian College of Animal Experimentation (COBEA), and the protocol/experiments have been approved by the Ethics Committee for Animal Experimentation of Instituto Butantan (CEUAIB Protocol number 1777050816).

[Nota: Todos os experimentos de bancada, que estão descritos nos tópicos "Parasite materials, RNA extraction, quantification and quality assessment, e RT-qPCR assays", não foram realizados por mim e sim por colaboradores do nosso grupo.]

Parasite materials

All parasite material was from a BH isolate of *S. mansoni* maintained by passage through golden hamster (*Mesocricetus auratus*) and *Biomphalaria glabrata* snails. Eggs were purified from livers of hamsters previously infected with *S. mansoni*, according to Dalton et al. (1997). After purification, eggs were added to 10 ml of distilled water and exposed to a bright light. Supernatant containing hatched miracidia was removed every 30 min for 2 h and replaced by fresh water. The supernatants containing the miracidia were pooled, chilled on ice and miracidia were then recovered by centrifugation at 15000 g for 20 seconds (Dalton et al., 1997). Supernatant was discarded and miracidia stored in RNAlater (Ambion) until RNA extraction.

Cercariae were collected from snails infected with 10 miracidia each. Thirty-five days after infection, the snails were placed in the dark in water and then illuminated for two hours to induce shedding. The emerging cercariae were collected by centrifugation, washed with PBS once and then stored in RNAlater (Ambion) until RNA extraction.

Schistosomula were obtained by mechanical transformation of cercariae and separation of their bodies as previously described (Basch, 1981), with some modifications. Briefly, cercariae were collected as described above and then suspended in 15 ml of M169 medium (Vitrocell, cat number 00464) containing penicillin/streptomycin, amphotericin (Vitrocell, cat number 00148). Mechanical transformation was performed by passing the cercariae ten times through a 23G needle. In order to separate schistosomula from the tails, the tail-rich supernatant was decanted and the sedimented bodies resuspended in a further 7 ml of M169 medium. The procedure was repeated until less than 1% of tails remained. The newly transformed schistosomula were maintained for 24 h in M169 medium (Vitrocell, cat number

00464) supplemented with penicillin/streptomycin, amphotericin, gentamicin (Vitrocell, cat number 00148), 2 % fetal bovine serum, 1 μM serotonin, 0.5 μM hypoxanthine, 1 μM hydrocortisone and 0.2 μM triiodothyronine at 37°C and 5% CO_2. Schistosomula cultivated for 24 h were collected, washed 3 times with PBS and stored in RNAlater (Ambion) until RNA extraction.

Adult *S. mansoni* worms were recovered by perfusion of golden hamsters that had been infected with 250 cercariae, 7 weeks previously. Approximately 200 *S. mansoni* (BH strain) adult worm pairs were freshly obtained through the periportal perfusion of hamster, as previously described (Anderson et al., 2016;Vasconcelos et al., 2017). After perfusion, the adult worm pairs were kept for 3 h at 37 °C and 5 % CO_2 in Advanced RPMI Medium 1640 (Gibco, #12633–012) supplemented with 10 % fetal bovine serum, 12 mM HEPES (4-(2-hydroxyethyl) piperazine-1-ethanesulfonic acid) pH 7.4, and 1 % penicillin/streptomycin, amphotericin (Vitrocell, cat number 00148). After 3 h of incubation, the adult worm pairs were collected, washed 3 times with PBS and stored in RNAlater (Ambion) until RNA extraction. Before the extraction of RNA from males or females, adult worm pairs were manually separated in RNAlater (Ambion) using tweezers.

RNA extraction, quantification and quality assessment

Total RNA from eggs (E), miracidia (Mi), cercariae (C) and schistosomula (S) was extracted according to Vasconcelos et al. (2017). Briefly, 100,000 eggs, 15,000 miracidia, 25,000 cercariae or 25,000 schistosomula were ground with glass beads in liquid nitrogen for 5 minutes. Then the Qiagen RNeasy Micro Kit (Cat number 74004) was used for RNA extraction and purification according to the manufacturer's instructions, except for the DNase I treatment: the amount of DNase I was doubled and the time of treatment was increased to 45 minutes.

Male (M) or female (F) adult worms were first disrupted in Qiagen RLT buffer using glass potters and pestles. RNA from males or females was then extracted and purified using the Qiagen RNeasy Mini Kit (Cat number 74104), according to the manufacturer's instructions, except for the DNase I treatment, which was the same used for egg, miracidia, cercariae and schistosomula RNA extraction.

All the RNA samples were quantified using the Qubit RNA HS Assay Kit (Q32852, Thermo Fisher Scientific) and the integrity of RNAs was verified using the Agilent RNA 6000 Pico Kit (5067-1513 Agilent Technologies) in a 2100 Bioanalyzer Instrument (Agilent

Technologies). Four biological replicates were assessed for each life cycle stage, except for schistosomula, for which three biological replicates were assessed.

RT-qPCR assays

The reverse transcription (RT) reaction was performed with 200 ng of each total RNA sample using the SuperScript IV First-Strand Synthesis System (18091050, Life Technologies) and random hexamer primers in a 20 µL final volume. The obtained complementary DNAs (cDNAs) were diluted 4 times in DEPC water and quantitative PCR was performed using 2.5 µL of each diluted cDNA in a total volume of 10 µL containing 1X LightCycler 480 SYBR Green I Master Mix (04707516001, Roche Diagnostics) and 800 nM of each primer in a LightCycler 480 System (Roche Diagnostics). Primers for selected transcripts (**Supplementary Table S3**) were designed using the Primer 3 tool (http://biotools.umassmed.edu/bioapps/primer3_www.cgi), and each real-time qPCR was run in two technical replicates. The results were analyzed by comparative Ct method (Livak and Schmittgen, 2001). Real-time data were normalized in relation to the level of expression of Smp_090920 and Smp_062630 reference genes.

Results

LncRNAs identification and annotation

Using 633 publicly available *S. mansoni* RNA-seq libraries from whole worms at different stages, from isolated tissues, from cell-populations and from single-cells (see Methods), our pipeline assembled a consensus transcriptome comprised of 78,817 transcripts, of which 7,954 were classified as intergenic lncRNAs (lincRNAs), 7,438 as antisense lncRNAs and 1,191 as sense lncRNAs, totalizing 16,583 lncRNA transcripts originated from 10,024 genes (on average, 1.65 lncRNA isoforms per lncRNA gene); the summary of all six filtering steps in the pipeline is presented in **Table 1**. With the FEELnc lncRNA classification tool (**Table 1, step 5**), the most important feature for transcripts classification was the ORF coverage (**Supplementary Figure S2A**), i.e. the fraction of the total length of the transcript that is occupied by the longest predicted ORF. In the FEELnc model training process, an optimal coding probability cutoff (0.348) was identified, which resulted in 0.962 sensitivity and specificity of mRNA classification (**Supplementary Figure S2B**). Analogous information is not provided in the output of the CPC2 classification tool (**Table 1, step 5**).

Only the lncRNAs classified as such by both prediction tools were retained in the final set (**Table 1**).

From the total set of 16,583 lncRNAs obtained here, 11,022 are novel *S. mansoni* lncRNAs, while the remaining 5,561 transcripts comprise 120 lncRNAs that are identical to and 5,441 lncRNAs that have gene overlap with *S. mansoni* lncRNAs already reported in previous works (Vasconcelos et al., 2017;Liao et al., 2018;Oliveira et al., 2018) (**Supplementary Table S4**). In particular, among the 7,029 lincRNAs previously reported by our group (Vasconcelos et al., 2017), a total of 4,368 transcripts have partial or complete sequence overlap with the lncRNAs obtained here, whereas the remaining 2,661 transcripts (37.8 %) previously annotated by Vasconcelos et al. (2017) are no longer in the present updated *S. mansoni* lncRNAs dataset.

Among the transcripts in the public dataset that were previously classified as lncRNAs (Vasconcelos et al., 2017;Liao et al., 2018;Oliveira et al., 2018) and are now excluded, a total of 4,293 were reconstructed in our assembly however they were removed from our set of lncRNAs because they were partially processed pre-mRNA transcripts that have exon-exon overlap with new protein-coding genes of version 7.1. The remaining transcripts previously classified as lncRNAs were reconstructed here but were removed by the more stringent, presently used filtering steps. We have created a track on the *S. mansoni* UCSC-like genome browser (http://schistosoma.usp.br/) where the set of 16,583 lncRNAs obtained here can be visualized and the GTF and BED files can be downloaded. In **Figure 1** we show a selected protein-coding desert genomic *locus* on chromosome 2 covering 245 kilobases, which harbors only 3 protein-coding genes, and where we identified 7 lincRNAs, 2 sense lncRNAs and 1 antisense lncRNA that were not previously described.

In order to identify the contribution from each type of RNA-Seq library to the final lncRNAs set, we used the TACO transcriptome assembler to obtain the transcriptomes of the four following groups: whole organisms, tissues, cell-populations and single-cells. The result is presented in **Figure 2** and shows that each type of sample contributed with at least 1 thousand unique lncRNAs, detected only in that group. It is worthy to mention that around 4 % of the 16,583 lncRNAs are lost when the four transcriptomes are reconstructed separately.

Almost all lncRNAs encode short canonical ORFs within their sequences, however, as described by Verheggen et al. (2017), one can evaluate if these ORFs are originated only by random nucleotide progression by comparing the relative sizes of ORFs using the reverse-complement of the sequence as a control. As presented in **Figure 3**, it is very clear that the

size distribution of *bona fide S. mansoni* mRNA ORFs (sense) from the annotated v 7.1 transcriptome is greatly shifted towards longer sizes, compared with the size distribution of random ORFs found in their reverse-complement sequences. It is also possible to observe that the size distribution of ORFs found both within the lncRNAs (sense) and within their reverse-complement sequences is very similar and is also similar to the size distribution of random ORFs in the reverse-complement sequence of mRNAs.

Histone marks at the TSS of lncRNAs as evidence of regulation

As reported earlier, cross-matching of the lncRNAs genomic coordinates with the genomic coordinates of different publicly available histone mark profiles, obtained by ChIP-Seq at different life-cycle stages, adds another layer of functionality evidence for this class of RNAs (Vasconcelos et al., 2017;Cao et al., 2018). We used data for two different histone marks obtained by Roquis et al. (2018) in cercariae, schistosomula and adult parasites, namely H3K4me3 that is generally associated with active transcription, and H3K27me3 associated to transcription repression (Barski et al., 2007). First we analyzed the histone mark profiles of H3K4me3 and H3K27me3 around the TSS of protein-coding genes through the stages and they were very similar to the ones presented by Roquis et al. (2018) (**Supplementary Figure S3)**. **Figure 4** shows that these marks are also present around the TSS of *S. mansoni* lncRNAs at the three different life-cycle stages; a comparison with Supplementary Figure S3 shows that these marks are less abundant in lncRNAs than in protein-coding genes loci, and more spread away of the lncRNAs TSSs when compared with protein-coding genes. This profile is similar to that observed by Sati et al. (2012) when comparing histone marks around the TSS of human protein-coding genes and lncRNAs. A total of 8,599 lncRNA transcripts have at least one histone modification mark within 1 kb from their TSS (**Supplementary Table S5)**, being 3,659 lincRNAs, 4,188 antisense lncRNAs and 752 sense lncRNAs. A comparison of the lists of lncRNAs having a given histone mark at their TSS at either of the three different life-cycle stages (**Figure 5**) shows that the most abundant mark is the transcriptional repressive mark H3K27me3. This mark is present at the TSS of different sets of lncRNAs at each of the three stages, with abundancies ranging from 1,334 lncRNAs with the H3K27me3 mark exclusively in schistosomula, to 1,147 lncRNAs with the mark exclusively in adults, and 1,024 lncRNAs with the mark exclusively in cercariae (**Figure 5, red**). In addition, the transcriptional activating mark H3K4me3 is present at the TSS of a different set of lncRNAs, with abundancies ranging from 740 lncRNAs with the H3K4me3 mark exclusively in

schistosomula, to 282 lncRNAs with the mark exclusively in cercariae, and 214 lncRNAs with the mark exclusively in adults (**Figure 5, green**). Interestingly, among the lncRNAs with the most abundant patterns of marks at their TSS, there are 316 lncRNAs in cercariae that have the characteristic marks of bivalent poised promoters (having both H3K4me3 and H3K27me3 marks at their TSS) (Voigt et al., 2013) (**Figure 5, blue**). This is analogous to the marks at the TSS of protein-coding genes in cercariae, where most genes have the bivalent mark (Roquis et al., 2018), indicating that lncRNAs are under a similar transcriptional regulatory program as the protein-coding genes in cercariae. **Supplementary Table S5** has a complete UpSet plot similar to that of Figure 5, showing the number of lncRNAs found in all different intersections, along with the lists of lncRNAs belonging to each intersection set.

Gene co-expression analyses

Once we identified our final lncRNAs set, we applied weighted gene co-expression network analyses (WGCNA) to integrate the expression level differences observed for lncRNAs and mRNAs among all life-cycle stages and the gonads, using all RNA-seq libraries available. The file containing expression levels (in TPM) for all transcripts in all 633 RNA-Seq libraries is available at http://schistosoma.usp.br/. After normalization and gene filtering (see Methods), 90 libraries out of the 112 from the different stages (mixed-sex adults were not included) remained in the WGCNA analyses, and 19,258 transcripts were retained (12,693 protein-coding genes and 6,565 lncRNAs).

Samples from miracidia, sporocysts, schistosomula, cercariae and gonads (testes and ovaries) were correctly clustered together by their expression correlation, based on Euclidian distance metrics (**Figure 6**). For samples from adult worms, in spite of the fact that we have one cluster branch mainly composed of females, and another mainly composed of males, there are some male samples among the female ones and vice-versa. Besides, due to the known similarity between males and juveniles (Wang et al., 2017) their samples were not well separated. It is interesting to note that immature females, that were shown to have a similar expression profile as that of males (Lu et al., 2016), are clustered here in the males-branch. As the WGCNA performs an unsupervised co-expression analysis, we decided to keep all male and female samples in the analysis, including those that are clustered apart from their main group, in order not to add a bias in the construction of modules.

We identified 15 different lncRNAs/mRNAs co-expression modules **(Figure 7)**, the sizes ranging from 215 to 3,318 transcripts (**Table 2** and **Supplementary Table S6**). The

ratio between the number of lncRNAs and mRNAs that comprise each module varies among the modules; thus, while lncRNAs comprise 86 % of the transcripts in the cyan module, only 5 % of the transcripts from the black module are lncRNAs (**Table 2**).

A Pearson correlation analysis indicates that each stage/tissue has at least one module whose genes' expression has a statistically significant positive correlation with that stage or tissue (**Figure 8**). Some stages also have modules that have a statistically significant negative correlation, such as the black module that is negatively correlated with miracidia/sporocysts. For the black module, the transcripts that compose the module have an expression in miracidia/sporocysts that is lower when compared with the overall expression of those transcripts across the other stages. The grey color represents the group of transcripts with a highly heterogeneous co-expression pattern that could not cluster into any of the 15 modules. In fact, it can be seen in Figure 8 that in this group the best correlation coefficient obtained in juveniles is lower ($|r| = 0.32$) and the p-value is much higher ($p = 0.002$) than the best parameters that were obtained in at least one stage for any module ($|r| \geq 0.51$ and $p \leq 3e\text{-}07$). Here, our choice of keeping in the WGCNA analysis those male and female samples that cluster apart from their main group (**Figure 6**) has an impact, decreasing the correlation coefficient of the modules mostly correlated to males or females (pink or turquoise, respectively) when compared with correlation coefficients in the other stages/tissues, nevertheless they still have a statistically significant high correlation.

We chose three RNA-seq library samples from each of the 9 different stages/tissues (among all the libraries under analysis) to construct a representative expression heatmap (**Figure 9**). This heatmap shows the expression across all stages of the top 50 transcripts with the highest gene module membership (GMM) to the most correlated module of each stage (as seen in **Figure 8)** (for GMM definition see WGCNA background and glossary, available at https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/) (Langfelder and Horvath, 2008). The heatmap (**Figure 9**) confirms that the top transcripts belonging to one module are more expressed in one given stage/tissue, which is the stage/tissue with which the module has the highest correlation. It is noteworthy that female library SRR5170160, which clustered inside the male group (**Figure 6**) when all filtered transcripts under analysis were used for clustering, now is correctly clustered with the other female samples (**Figure 9**) when only the top 50 transcripts with the highest GMM are considered. Also, juveniles share with adult males a similar expression pattern of the top 50

male genes, which is in line with the clustering of juveniles along with males in the analysis of **Figure 6**.

Validation of lncRNAs expression by RT-qPCR

[Nota: Todos os resultados apresentados neste tópico foram fruto do trabalho de colaboradores do nosso grupo, que são co-autores do artigo, e não foram obtidos por mim.]

We designed PCR primer pairs for a selected set of eleven lincRNAs belonging to five different modules, as determined by WGCNA, in order to detect their expression along the different *S. mansoni* life-cycle stages and to eventually validate their different expression levels at the stages. Our selection was based on the Gene Trait Significance score (GS score) (**Supplementary Table S7**) of each lncRNA in the module where it belongs, which varies from -1 to 1, using the stages as external information (see Methods). The higher the absolute value of the GS score, the more biologically significant and correlated to the stage of interest is the transcript expression. For the RT-qPCR assays we used samples from eggs (E), miracidia (Mi), cercariae (C), schistosomula (S), adult males (M) and females (F).

First, we measured the expression of five protein-coding genes that were used as stage markers (Parker-Manuel et al., 2011;Anderson et al., 2016) and we found that in our RNA samples they were more highly expressed at the predicted stages (**Supplementary Figure S4**).

Then, we tested the selected eleven lincRNAs and detected that they were expressed in at least one of the six stages that were assayed; specifically, each of six lincRNAs were more highly expressed at the stage predicted by the correlation with the modules **(Figure 10)**, at four life-cycle stages: two more highly expressed in miracidia (SmLINC158013-IBu and SmLINC123205-IBu, purple module), two in cercariae (SmLINC123474-IBu and SmLINC134196-IBu, tan module), one in schistosomula (SmLINC105065-IBu, magenta module), and one in males (SmLINC100046-IBu, turquoise module) **(Figure 10)**. In **Supplementary Figure S5** we present the values in transcripts per million reads (TPM) from the RNA-seq libraries for each of these six validated lincRNAs. Additionally, the five other lincRNAs that were tested were detected as expressed across all stages, however they were not differentially expressed as predicted by the RNA-seq (**Supplementary Figure S6**). This indicates that there is variability of lncRNAs expression between the experimental conditions and parasite strain used in our assays and those found among the dozens of samples that are publicly available.

Protein-coding genes ontology enrichment and lncRNA hub genes in the modules

Gene Ontology (GO) enrichment analyses show that the protein-coding genes belonging to the red module, which have a correlation of 0.99 with testes, are enriched with processes related to sperm motility such as cilium movement and the axoneme assembly **(Figure 11A)**. Besides, the green module, correlated with both ovaries and testes, is enriched with proteins associated with cellular replication **(Figure 11B)**. All other modules with GO enrichment, which in general are enriched with proteins associated to general metabolism, are presented in **Supplementary Figures S7-S10**. The black, cyan, midnightblue, purple and tan modules have no significantly enriched GO terms due to the small number of protein-coding genes with GO annotation within each of these modules.

All transcripts that belong to the same module are connected, however in order to better visualize this, gene co-expression networks were constructed only with the most connected genes (as determined by the adjacency threshold) (**Figure 12**), and they show, along with the correlation values presented in **Supplementary Table S7,** that some lncRNAs are hub genes from the network. **Figures 12A-B** show lncRNA hub genes in the co-expression networks from the purple and tan modules, strongly correlated with miracidia/sporocysts and cercariae life-cycle stages, respectively. In both modules, the lncRNAs represent around half of the transcripts that comprise the modules (see **Table 2**). However, there are some cases, such as in the red module, where 3/4 of the member transcripts are lncRNAs, and among the most connected genes in that co-expression network almost all are lncRNAs (**Figure 12C**). Also, in the blue module only 16 % of the member transcripts are lncRNAs, and only one is among the most connected genes in the co-expression network (**Figure 12D**). All the gene networks for all modules in a format compatible with Cytoscape are available at **Supplementary Table S8**. An adjacency cutoff threshold = 0.1 was used.

LncRNAs expressed in single cells

Finally, analyses using single-cell data from two stages, mother sporocysts stem cells and juveniles' stem cells, identified three different clusters. Cluster 1 is composed by a subgroup of juvenile's stem cells, cluster 2 is composed by all mother sporocysts stem cells, and cluster 3 is composed by a second and smaller subgroup of juveniles' stem cells (**Figure 13A**. The marker gene analyses shows for the first time in *S. mansoni* that lncRNAs have specific expression also at the single-cell level, where from the top 10 markers that allow us

to differentiate mother sporocysts stem cells from juveniles' stem cells, 8 are lncRNAs (**Figure 13B**), confirming the stage-specificity of lncRNAs also seen in whole worm analyses by WGCNA. Besides, another lncRNA was identified as marker for cluster 3 when compared to the other two clusters (**Figure 13B**).

Discussion

When the human genome was first sequenced, the vast genomic regions that lie between protein-coding genes (intergenic regions) were considered junk DNA; one decade later the Encyclopedia of DNA Elements (ENCODE) project found that 80 % of the human genome serves some biochemical purpose (Pennisi, 2012), including giving rise to the transcription of nearly ten thousand lncRNAs (Derrien et al., 2012). Although we are still at the beginning of the studies with lncRNAs, with the vast majority of their roles and mechanisms of action in humans being still unknown, it is now clear that most of the lncRNAs are transcribed from intergenic regions and are key regulators in vital processes (Kitagawa et al., 2013;Rosa and Ballarino, 2016;Golicz et al., 2018), being associated to several pathologies in humans such as cancer (Fang and Fullwood, 2016), Alzheimer's (Zijian, 2016) and cardiac diseases (Simona et al., 2018).

In *S. mansoni*, with the release in 2012 of version 5.2 of the genome and annotated transcriptome (Protasio et al., 2012), and with the accumulation until 2017 of large amounts of information on gene expression obtained through 88 publicly available RNA-seq libraries, our group decided to map the RNA-seq data and identify the lncRNAs repertoire expressed in this parasite (Vasconcelos et al., 2017); this was followed by two other papers that provided an additional set of lncRNAs (Liao et al., 2018;Oliveira et al., 2018). In the present work, by extending the analysis to 633 publicly available RNA-seq libraries, and by performing a detailed curation of the assembled transcripts, we observed that at the sequencing depth obtained with the current RNA-seq datasets, a considerable amount of partially processed pre-mRNAs is being sequenced. These pre-mRNAs give rise to assembled transcript units showing intron-retention and frequent stop codons in the retained introns, and therefore these transcripts can be mistakenly annotated as lncRNAs.

In fact, the failure to identify partially processed pre-mRNA in previous publications (Vasconcelos et al., 2017;Liao et al., 2018) may explain the report of probable protein-coding genes as lncRNAs (**Supplementary Figure S1**). Our current pipeline has removed at step 4 a total of 31,183 assembled transcripts that had partial or total exon-exon overlap on the same

genomic strand with known *S. mansoni* protein-coding genes, and this included around 14,000 assembled transcripts that represented fully processed mature protein-coding transcripts that exactly matched the annotated v 7.1 transcripts from the Wellcome Sanger Institute, as well as some 17,000 assembled transcripts that for the most part represent partially processed pre-mRNAs with intron retention; among the latter are 4,293 transcripts that were previously classified as lncRNAs (Vasconcelos et al., 2017;Liao et al., 2018;Oliveira et al., 2018) and are now excluded. With the six stringent filtering steps used in the present work, we are confident that our final set of 16,583 lncRNAs is a robust representation of the lncRNAs complement expressed in *S. mansoni*, of which 11,022 transcripts are novel lncRNAs, and 5,561 have gene overlap with lncRNAs already reported in previous works (Vasconcelos et al., 2017;Liao et al., 2018;Oliveira et al., 2018).

One question that has been raised about lncRNAs is the possibility that their function is executed through translation into short peptides, a concern that arises from the fact that almost all lncRNAs encode short canonical ORFs within their sequences (Verheggen et al., 2017); the fact that the size distribution of ORFs found within our set of lncRNAs (sense) is very similar to the size distribution of random ORFs found within their reverse-complement sequences and within the reverse-complement sequence of mRNAs suggest that the putative short ORFs from the lncRNAs identified here are indeed random ORFs, most probably not translated into short functional peptides. Nevertheless, future functional characterization in *S. mansoni* of selected lncRNAs may eventually include a search for a possible dual function role (Nam et al., 2016;Choi et al., 2018) both as lncRNA and through a translated short peptide.

Histone marks were found here at the TSS of lncRNAs, and the identification of different sets of lncRNAs that have at their TSS the transcriptional activation H3K4me3 mark, or the repressive H3K27me3 mark, when the three life-cycle stages are compared, suggests that lncRNAs expression in *S. mansoni* is regulated by an epigenetic program. This finding reinforces the hypothesis that different lncRNAs may play important roles along the parasite life-cycle, and the sets of lncRNAs identified in this analysis might be the first candidates to be explored for further functional characterization.

Gene co-expression networks correlated to the different *S. mansoni* life-cycle stages were identified by our analyses, and they pointed to sets of protein-coding genes and lncRNAs with expression most correlated to one given stage. This information provides an initial platform for prioritizing the lncRNAs to be selected for further direct functional

characterization, which will include a search for altered *S. mansoni* phenotypes upon knockdown of lncRNA candidates. In *Plasmodium falciparum*, the knockdown of antisense lncRNAs has down-regulated the active var gene, a gene related to immune evasion, erasing the epigenetic memory and substantially changing the var gene expression pattern (Amit-Avraham et al., 2015). In analogy, it is expected that characterization of lncRNAs in *S. mansoni* will help to recognize the biochemical pathways where they play a functional role, will permit to identify their interacting protein partners, and will eventually point to relevant ways of intervention in the parasite physiology.

Due to the complex and diverse mechanisms displayed by lncRNAs in regulating protein-coding genes and miRNAs, the majority of studies have not progressed beyond cell or animal models, and progression towards the clinic has been slow (Harries, 2019). Nevertheless, lncRNAs represent potentially good therapeutic targets (Matsui and Corey, 2017;Blokhin et al., 2018;Harries, 2019). As reviewed by Matsui and Corey (2017), in Angelman syndrome model mouse, the administration of antisense oligonucleotides (ASOs), which target the Ube3a-ATS lncRNA for degradation, partially reversed some cognitive defects associated with the disease in the animals (Meng et al., 2014). Also, in xenograft melanoma models the intravenous injection of ASOs targeting the lncRNA SAMMSON caused p53 activation, tumor growth suppression, decreased cell proliferation and increased apoptosis (Leucci et al., 2016). In this respect, it is noteworthy that lncRNAs are considerably less conserved between species when compared with protein-coding genes (Pang et al., 2006;Blokhin et al., 2018), and that only a few dozen ancient lncRNAs have conserved orthologs between ancient non-amniote *Xenopus* and the closest amniote Chicken model animals (Necsulea et al., 2014), which shows that lncRNAs have evolutionarily conserved gene regulatory functions however low sequence conservation across distant species (Necsulea et al., 2014). This feature reduces the chances that targeting a lncRNA in *S. mansoni*, for example with ASOs, will cause unwanted off-target effects against the mammalian host.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

LFM, MSA and SV-A conceived the project. LFM and SV-A designed the experiments and wrote the paper. LFM and DAM-V performed the *in silico* analyses. MSA, GOS, ROR and GGOO performed the wet lab experiments and analyses. LFM and SV-A analyzed and interpreted the data. DSP contributed with informatic resources.

Funding

This work was supported by the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) grant numbers 2014/03620-2 and 2018/23693-5 to SV-A. LFM, GOS and ROR received FAPESP fellowships (grant numbers 2018/19591-2, 2018/24015-0 and 2017/22379-2, respectively) and DAM-V received a fellowship from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). SV-A laboratory was also supported by institutional funds from Fundação Butantan and received an established investigator fellowship award from CNPq, Brasil.

Acknowledgments

We thank Dr. J.C. Setubal for access to the computational facilities of the Bioinformatics Laboratory of Instituto de Química, Universidade de São Paulo (USP). We also acknowledge Patricia Aoki Miyasato and Dr. Eliana Nakano, Laboratorio de Malacologia, Instituto Butantan, for maintaining the *S. mansoni* life cycle.

Data Availability Statement

The datasets analyzed in this study can be found in the SRA repository (https://www.ncbi.nlm.nih.gov/sra) and in the ENA repository (https://www.ebi.ac.uk/ena). The specific accession numbers for each and all datasets that were downloaded from these databases and used here are given in **Supplementary Table S1**.

References

Amit-Avraham, I., Pozner, G., Eshar, S., Fastman, Y., Kolevzon, N., Yavin, E., and Dzikowski, R. (2015). Antisense long noncoding RNAs regulate var gene activation in the malaria parasite *Plasmodium falciparum*. *Proc Natl Acad Sci U S A* 112**,** E982-991.

Anderson, L., Amaral, M.S., Beckedorff, F., Silva, L.F., Dazzani, B., Oliveira, K.C., Almeida, G.T., Gomes, M.R., Pires, D.S., Setubal, J.C., Demarco, R., and Verjovski-Almeida, S. (2016). *Schistosoma mansoni* Egg, Adult Male and Female Comparative Gene Expression

Analysis and Identification of Novel Genes by RNA-Seq. *PLOS Neglected Tropical Diseases* 9**,** e0004334.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* 129**,** 823-837.

Basch, P.F. (1976). Intermediate host specificity in Schistosoma mansoni. *Exp Parasitol* 39**,** 150-169.

Basch, P.F. (1981). Cultivation of *Schistosoma mansoni* In vitro. I. Establishment of Cultures from Cercariae and Development until Pairing. *The Journal of Parasitology* 67**,** 179-185.

Batugedara, G., Lu, X.M., Bunnik, E.M., and Le Roch, K.G. (2017). The Role of Chromatin Structure in Gene Regulation of the Human Malaria Parasite. *Trends in Parasitology* 33**,** 364-377.

Bhat, S.A., Ahmad, S.M., Mumtaz, P.T., Malik, A.A., Dar, M.A., Urwat, U., Shah, R.A., and Ganai, N.A. (2016). Long non-coding RNAs: Mechanism of action and functional utility. *Non-coding RNA Research* 1**,** 43-50.

Blokhin, I., Khorkova, O., Hsiao, J., and Wahlestedt, C. (2018). Developments in lncRNA drug discovery: where are we heading? *Expert Opin Drug Discov* 13**,** 837-849.

Cao, H., Wahlestedt, C., and Kapranov, P. (2018). Strategies to Annotate and Characterize Long Noncoding RNAs: Advantages and Pitfalls. *Trends in Genetics* 34**,** 704-721.

Cdc (2018). *Centers for Disease Control and Prevention, Parasites - Schistosomiasis* [Online]. Available: https://www.cdc.gov/parasites/schistosomiasis/ [Accessed 20/07/2018 2018].

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34**,** i884-i890.

Choi, S.W., Kim, H.W., and Nam, J.W. (2018). The small peptide world in long noncoding RNAs. *Brief Bioinform***,** bby055-bby055.

Credendino, S.C., Lewin, N., De Oliveira, M., Basu, S., Andrea, B., Amendola, E., Di Guida, L., Nardone, A., Sanges, R., De Felice, M., and De Vita, G. (2017). Tissue- and Cell Type-Specific Expression of the Long Noncoding RNA Klhl14-AS in Mouse *International Journal of Genomics* 2017**,** 7.

Dalton, J.P., Day, S.R., Drew, A.C., and Brindley, P.J. (1997). A method for the isolation of schistosome eggs and miracidia free of contaminating host tissues. *Parasitology* 115 (Pt 1)**,** 29-32.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J.B., Lipovich, L., Gonzalez, J.M., Thomas, M., Davis, C.A., Shiekhattar, R., Gingeras, T.R., Hubbard, T.J., Notredame, C., Harrow, J., and Guigo, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* 22**,** 1775-1789.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29**,** 15-21.

Fang, Y., and Fullwood, M.J. (2016). Roles, Functions, and Mechanisms of Long Non-coding RNAs in Cancer. *Genomics, Proteomics & Bioinformatics* 14**,** 42-54.

Golicz, A.A., Bhalla, P.L., and Singh, M.B. (2018). lncRNAs in Plant and Animal Sexual Reproduction. *Trends in Plant Science* 23**,** 195-205.

Gomes Casavechia, M.T., De Melo, G.D.a.N., Da Silva Fernandes, A.C.B., De Castro, K.R., Pedroso, R.B., Da Silva Santos, T., and Teixeira, J.J.V. (2018). Systematic review and meta-analysis on *Schistosoma mansoni* infection prevalence, and associated risk factors in Brazil. *Parasitology* 145**,** 1000-1014.

Hanly, D.J., Esteller, M., and Berdasco, M. (2018). Interplay between long non-coding RNAs and epigenetic machinery: emerging targets in cancer? *Philos Trans R Soc Lond B Biol Sci* 373**,** 20170074.

Harries, L.W. (2019). RNA Biology Provides New Therapeutic Targets for Human Disease. *Front Genet* 10**,** 205.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell* 38**,** 576-589.

Howe, K.L., Bolt, B.J., Shafie, M., Kersey, P., and Berriman, M. (2017). WormBase ParaSite - a comprehensive resource for helminth genomics. *Mol Biochem Parasitol* 215**,** 2-10.

Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., Von Mering, C., and Bork, P. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Molecular Biology and Evolution* 34**,** 2115-2122.

Kiselev, V.Y., Kirschner, K., Schaub, M.T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K.N., Reik, W., Barahona, M., Green, A.R., and Hemberg, M. (2017). SC3: consensus clustering of single-cell RNA-seq data. *Nature Methods* 14**,** 483.

Kitagawa, M., Kitagawa, K., Kotake, Y., Niida, H., and Ohhata, T. (2013). Cell cycle regulation by long non-coding RNAs. *Cellular and Molecular Life Sciences* 70**,** 4785-4794.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9**,** 559.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods* 9**,** 357-359.

Leucci, E., Vendramin, R., Spinazzi, M., Laurette, P., Fiers, M., Wouters, J., Radaelli, E., Eyckerman, S., Leonelli, C., Vanderheyden, K., Rogiers, A., Hermans, E., Baatsen, P., Aerts, S., Amant, F., Van Aelst, S., Van Den Oord, J., De Strooper, B., Davidson, I., Lafontaine, D.L.J., Gevaert, K., Vandesompele, J., Mestdagh, P., and Marine, J.-C. (2016). Melanoma addiction to the long non-coding RNA SAMMSON. *Nature* 531**,** 518.

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). UpSet: Visualization of Intersecting Sets. *IEEE Trans Vis Comput Graph* 20**,** 1983-1992.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12**,** 323.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25**,** 2078-2079.

Liao, Q., Zhang, Y., Zhu, Y., Chen, J., Dong, C., Tao, Y., He, A., Liu, J., and Wu, Z. (2018). Identification of long noncoding RNAs in *Schistosoma mansoni* and *Schistosoma japonicum*. *Exp Parasitol* 191**,** 82-87.

Livak, K.J., and Schmittgen, T.D. (2001). Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method. *Methods* 25**,** 402-408.

Lu, Z., Sessler, F., Holroyd, N., Hahnel, S., Quack, T., Berriman, M., and Grevelding, C.G. (2016). Schistosome sex matters: a deep view into gonad-specific and pairing-dependent transcriptomes reveals a complex gender interplay. *Scientific Reports* 6**,** 31150.

Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics* 21**,** 3448-3449.

Matsui, M., and Corey, D.R. (2017). Non-coding RNAs as drug targets. *Nat Rev Drug Discov* 16**,** 167-179.

Mccarthy, D.J., Campbell, K.R., Wills, Q.F., and Lun, A.T.L. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* 33**,** 1179-1186.

Meng, L., Ward, A.J., Chun, S., Bennett, C.F., Beaudet, A.L., and Rigo, F. (2014). Towards a therapy for Angelman syndrome by targeting a long non-coding RNA. *Nature* 518**,** 409.

Nam, J.W., Choi, S.W., and You, B.H. (2016). Incredible RNA: Dual Functions of Coding and Noncoding. *Mol Cells* 39**,** 367-374.

Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker, J.C., Grutzner, F., and Kaessmann, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* 505**,** 635-640.

Niknafs, Y.S., Pandian, B., Iyer, H.K., Chinnaiyan, A.M., and Iyer, M.K. (2016). TACO produces robust multisample transcriptome assemblies from RNA-seq. *Nature Methods* 14**,** 68.

Oliveira, K.C., Carvalho, M.L., Maracaja-Coutinho, V., Kitajima, J.P., and Verjovski-Almeida, S. (2011). Non-coding RNAs in schistosomes: an unexplored world. *An Acad Bras Cienc* 83**,** 673-694.

Oliveira, V.F., Moares, L.a.G., Mota, E.A., Jannotti-Passos, L.K., Coelho, P.M.Z., Mattos, A.C.A., Couto, F.F.B., Caffrey, B.E., Marsico, A., and Guerra-Sa, R. (2018). Identification of 170 New Long Noncoding RNAs in *Schistosoma mansoni*. *Biomed Res Int* 2018**,** 1264697.

Pang, K.C., Frith, M.C., and Mattick, J.S. (2006). Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends in Genetics* 22**,** 1-5.

Parker-Manuel, S.J., Ivens, A.C., Dillon, G.P., and Wilson, R.A. (2011). Gene Expression Patterns in Larval *Schistosoma mansoni* Associated with Infection of the Mammalian Host. *PLOS Neglected Tropical Diseases* 5**,** e1274.

Pennisi, E. (2012). Genomics. ENCODE project writes eulogy for junk DNA. *Science* 337**,** 1159, 1161.

Protasio, A.V., Tsai, I.J., Babbage, A., Nichol, S., Hunt, M., Aslett, M.A., De Silva, N., Velarde, G.S., Anderson, T.J., Clark, R.C., Davidson, C., Dillon, G.P., Holroyd, N.E., Loverde, P.T., Lloyd, C., Mcquillan, J., Oliveira, G., Otto, T.D., Parker-Manuel, S.J., Quail, M.A., Wilson, R.A., Zerlotini, A., Dunne, D.W., and Berriman, M. (2012). A systematically improved high quality genome and transcriptome of the human blood fluke *Schistosoma mansoni*. *PLoS Negl Trop Dis* 6**,** e1455.

Roquis, D., Taudt, A., Geyer, K.K., Padalino, G., Hoffmann, K.F., Holroyd, N., Berriman, M., Aliaga, B., Chaparro, C., Grunau, C., and Augusto, R.D.C. (2018). Histone methylation changes are required for life cycle progression in the human parasite *Schistosoma mansoni*. *PLOS Pathogens* 14**,** e1007066.

Rosa, A., and Ballarino, M. (2016). Long Noncoding RNA Regulation of Pluripotency. *Stem Cells International* 2016**,** 1797692.

Sati, S., Ghosh, S., Jain, V., Scaria, V., and Sengupta, S. (2012). Genome-wide analysis reveals distinct patterns of epigenetic features in long non-coding RNA loci. *Nucleic Acids Research* 40**,** 10018-10031.

Shao, M., and Kingsford, C. (2017). Accurate assembly of transcripts through phase-preserving graph decomposition. *Nature Biotechnology* 35**,** 1167.

Simona, G., Antonio, S.S., Yvan, D., and Fabio, M. (2018). Long Noncoding RNAs and Cardiac Disease. 29**,** 880-901.

Sims, D., Sudbery, I., Ilott, N.E., Heger, A., and Ponting, C.P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet* 15**,** 121-132.

Tarashansky, A.J., Xue, Y., Quake, S.R., and Wang, B. (2018). Self-assembling Manifolds in Single-cell RNA Sequencing Data. *bioRxiv***,** http://dx.doi.org/10.1101/364166.

Vasconcelos, E.J.R., Dasilva, L.F., Pires, D.S., Lavezzo, G.M., Pereira, A.S.A., Amaral, M.S., and Verjovski-Almeida, S. (2017). The *Schistosoma mansoni* genome encodes thousands of long non-coding RNAs predicted to be functional at different parasite life-cycle stages. *Sci Rep* 7**,** 10508.

Verheggen, K., Volders, P.-J., Mestdagh, P., Menschaert, G., Van Damme, P., Gevaert, K., Martens, L., and Vandesompele, J. (2017). Noncoding after All: Biases in Proteomics Data Do Not Explain Observed Absence of lncRNA Translation Products. *Journal of Proteome Research* 16**,** 2508-2515.

Voigt, P., Tee, W.-W., and Reinberg, D. (2013). A double take on bivalent promoters. *Genes & development* 27**,** 1318-1338.

Wang, B., Lee, J., Li, P., Saberi, A., Yang, H., Liu, C., Zhao, M., and Newmark, P.A. (2018). Stem cell heterogeneity drives the parasitic life cycle of *Schistosoma mansoni*. *Elife* 7**,** e35449.

Wang, J., Yu, Y., Shen, H., Qing, T., Zheng, Y., Li, Q., Mo, X., Wang, S., Li, N., Chai, R., Xu, B., Liu, M., Brindley, P.J., Mcmanus, D.P., Feng, Z., Shi, L., and Hu, W. (2017). Dynamic transcriptomes identify biogenic amines and insect-like hormonal regulation for mediating reproduction in *Schistosoma japonicum*. *Nature Communications* 8**,** 14693.

Wang, L., Wang, S., and Li, W. (2012). RSeQC: quality control of RNA-seq experiments. *Bioinformatics* 28**,** 2184-2185.

Who, W.H.O. (2015). *Investing to overcome the global impact of neglected tropical diseases: third WHO report on neglected tropical diseases 2015.* World Health Organization.

Wu, W., Wagner, E.K., Hao, Y., Rao, X., Dai, H., Han, J., Chen, J., Storniolo, A.M.V., Liu, Y., and He, C. (2016). Tissue-specific Co-expression of Long Non-coding and Coding RNAs Associated with Breast Cancer. *Scientific Reports* 6**,** 32731.

Wucher, V., Legeai, F., Hédan, B., Rizk, G., Lagoutte, L., Leeb, T., Jagannathan, V., Cadieu, E., David, A., Lohi, H., Cirera, S., Fredholm, M., Botherel, N., Leegwater, P.a.J., Le béguec, C., Fieten, H., Johnson, J., Alföldi, J., André, C., Lindblad-Toh, K., Hitte, C., and Derrien, T. (2017). FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Research* 45**,** e57-e57.

Yang, D.-C., Kong, L., Wei, L., Hou, M., Kang, Y.-J., Meng, Y.-Q., and Gao, G. (2017). CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Research* 45**,** W12-W16.

Zijian, Z. (2016). Long non-coding RNAs in Alzheimer's disease. *Current Topics in Medicinal Chemistry* 16**,** 511-519.

Zoni, A.C., Catalá, L., and Ault, S.K. (2016). Schistosomiasis Prevalence and Intensity of Infection in Latin America and the Caribbean Countries, 1942-2014: A Systematic Review in the Context of a Regional Elimination Goal. *PLoS Neglected Tropical Diseases* 10**,** e0004493.

Tables and Figures

Table 1. Summary of transcripts removed at each filtering step and the final set of *S. mansoni* lncRNAs.

Pipeline step number	Removed transcripts	Remaining transcripts (Genes)
1. Total assembled transcripts		**78,817 (42,337)**
2. Remove short transcripts (< 200 nt)	11	78,806
3. Remove monoexonic transcripts	27,255	51,551
4. Remove transcripts that overlap exon-exon with known Sm protein-coding genes	31,183	20,368
5. Remove transcripts with coding potential (FEELnc and/or CPC2 tools)	3,618	16,750
6. Remove transcripts with hits on eggNOG-Mapper	167	16,583
7. Total lncRNAs identified		**16,583 (10,024)**
Long intergenic non-coding RNAs		7,954
Antisense long non-coding RNAs		7,438
Sense long non-coding RNAs		1,191

Table 2. Number of transcripts per module and percentage of lncRNAs in each module.

Module color	Total number of transcripts	mRNAs	lncRNAs	% of lncRNAs	Stage of higher absolute correlation value
Black	989	940	49	5	Miracidia / Sporocysts
Blue	3,211	2,688	523	16	Juveniles
Brown	2,466	1,761	705	29	Gonads
Cyan	253	36	217	86	Ovaries
Green	1,308	841	467	35	Gonads
Greenyellow	502	417	85	17	Gonads
Magenta	748	273	475	64	Schistosomula
Midnightblue	215	43	172	80	Juveniles
Pink	840	506	334	40	Adult Females
Purple	590	274	316	54	Miracidia/Sporocysts
Red	1,254	333	921	73	Testes
Salmon	267	230	37	14	Gonads
Tan	356	158	198	56	Cercariae
Turquoise	3,318	2,470	848	26	Adult Males
Yellow	2,067	1,398	669	32	Adult Males

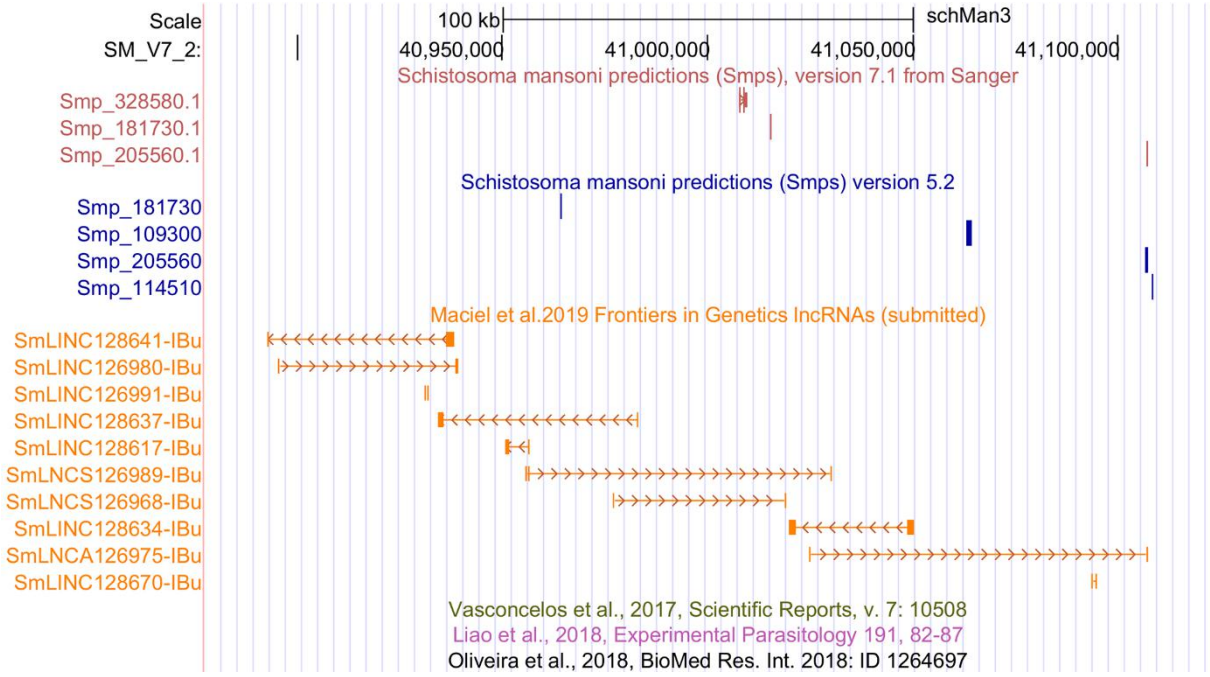


Figure 1 – Novel *S. mansoni* lncRNAs discovered in a protein-coding desert *locus*. Snapshot of a *S. mansoni* genome browser image, showing a region spanning 245 kb on chromosome 2 with coordinates SM_V7_2:40,877,676-41,122,371 (top black row). The red track (top) shows the three protein-coding genes from transcriptome version 7.1, while the blue track (middle) represents the protein-coding genes from version 5.2. The orange track (lower track) shows seven intergenic lncRNAs (SmLINCnnnnnn-IBu), two sense lncRNAs (SmLNCSnnnnnn-IBu) and one antisense lncRNA (SmLNCAnnnnnn-IBu) that were not annotated by the previously published works on lncRNAs, of which there are three empty tracks at the bottom, namely Vasconcelos et al. (2017), Liao et al. (2018) and Oliveira et al. (2018).

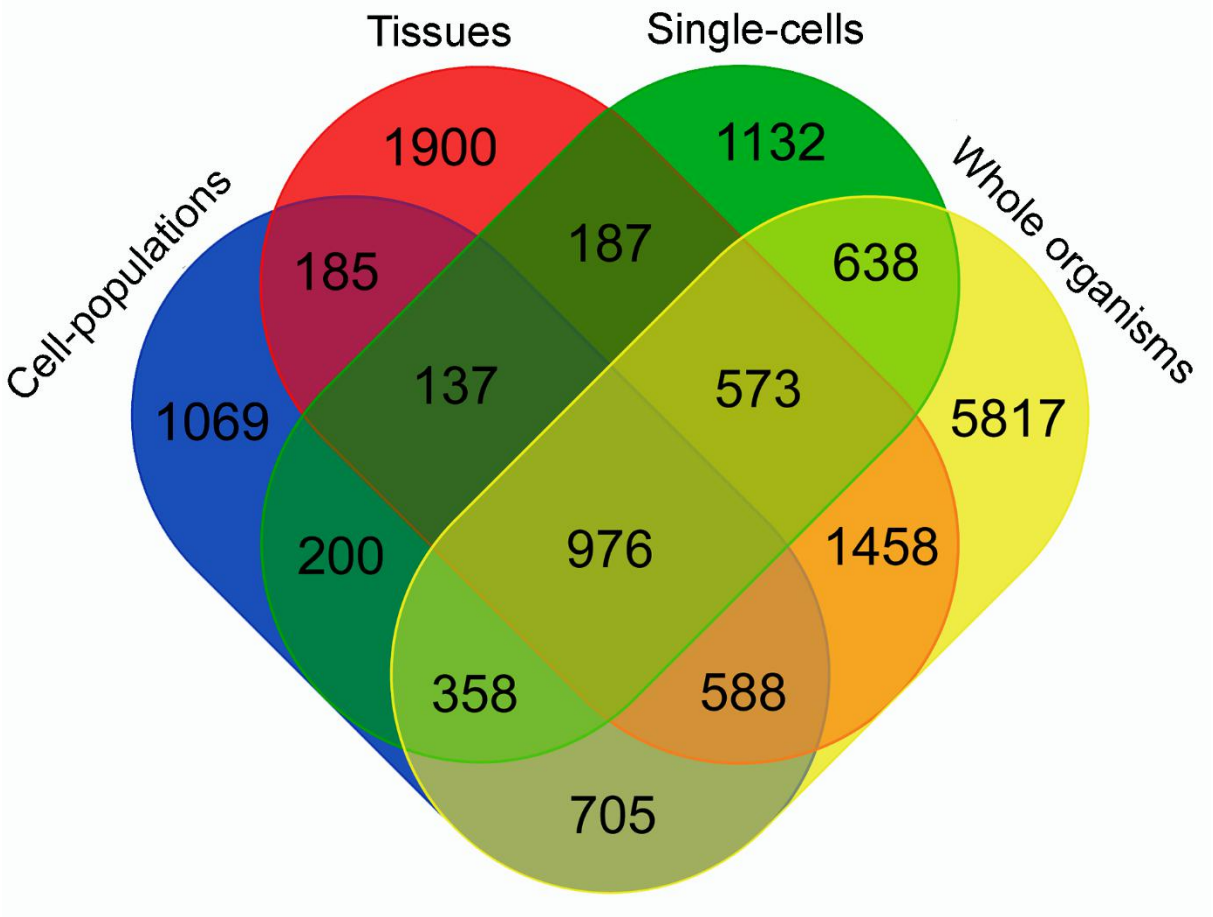


Figure 2 – Venn diagram representing the specific contribution from each type of RNA-Seq library to the *S. mansoni* lncRNAs set. TACO assembler was run separately for the RNA-Seq data from samples of four groups: whole organisms (yellow), tissues (red), cell-populations (blue) and single-cells (green), and each value indicates the number of transcripts that were reconstructed specifically with samples from groups indicated in each intersection.

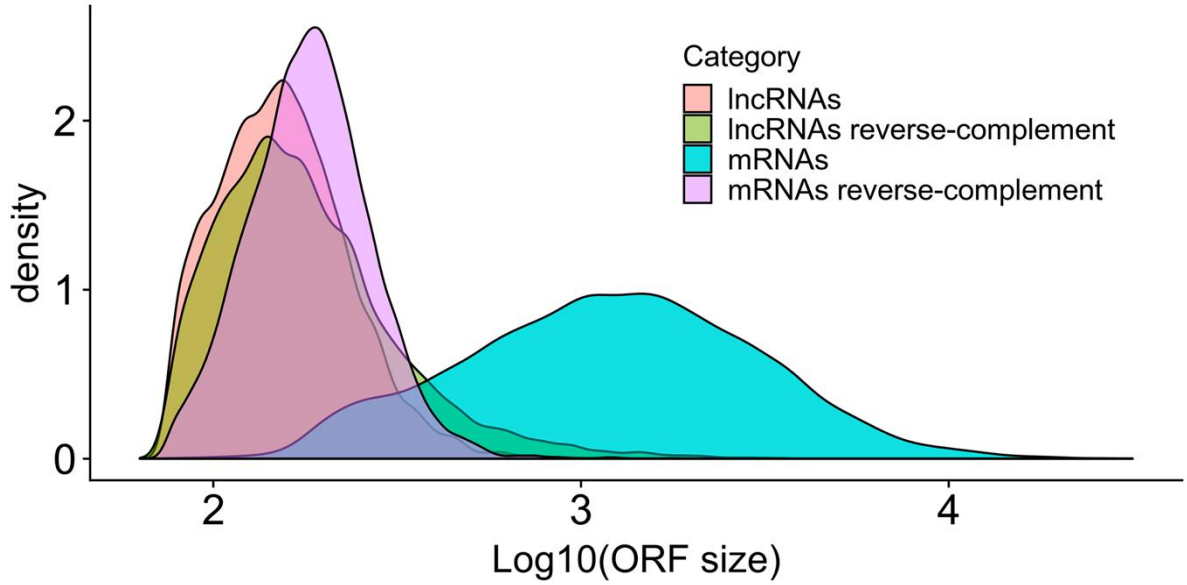


Figure 3 – Size distribution in *S. mansoni* of the longest canonical ORFs in lncRNAs and in mRNAs. The graph shows the density (y-axis) of the different sizes for the longest detected ORFs (in nucleotides, x-axis) of all lncRNAs (pink), of all mRNAs (blue) and of their reverse-complement sequences as controls (green and purple, respectively).

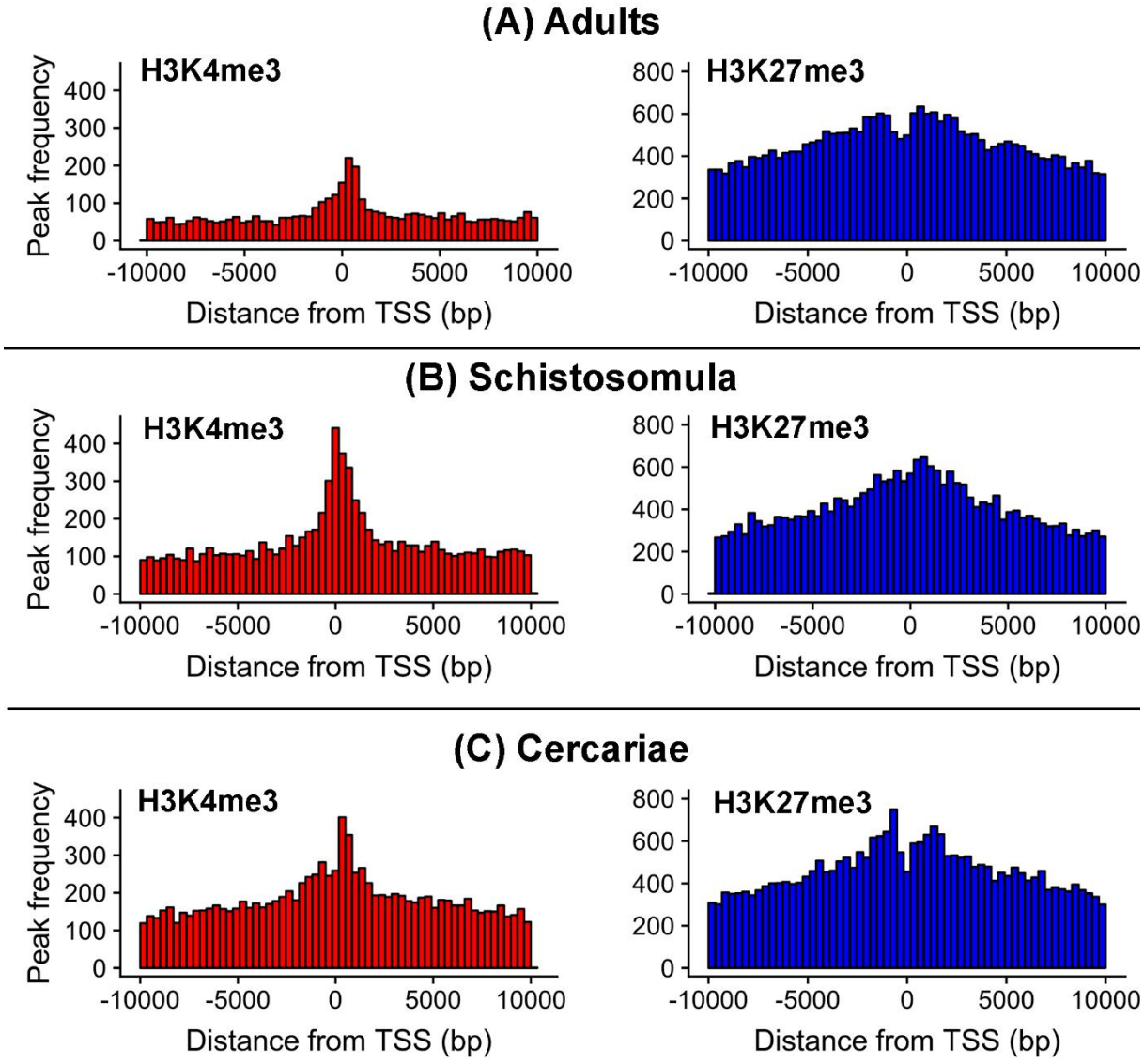


Figure 4 - Epigenetic histone marks H3K4me3 and H3K27me3 surrounding the TSS of *S. mansoni* lncRNAs. The frequency of the H3K4me3 marks (red) or of the H3K27me3 marks (blue) mapping within 10 kb around the TSS of all lncRNAs in (A) adults, (B) schistosomula and (C) cercariae was computed.

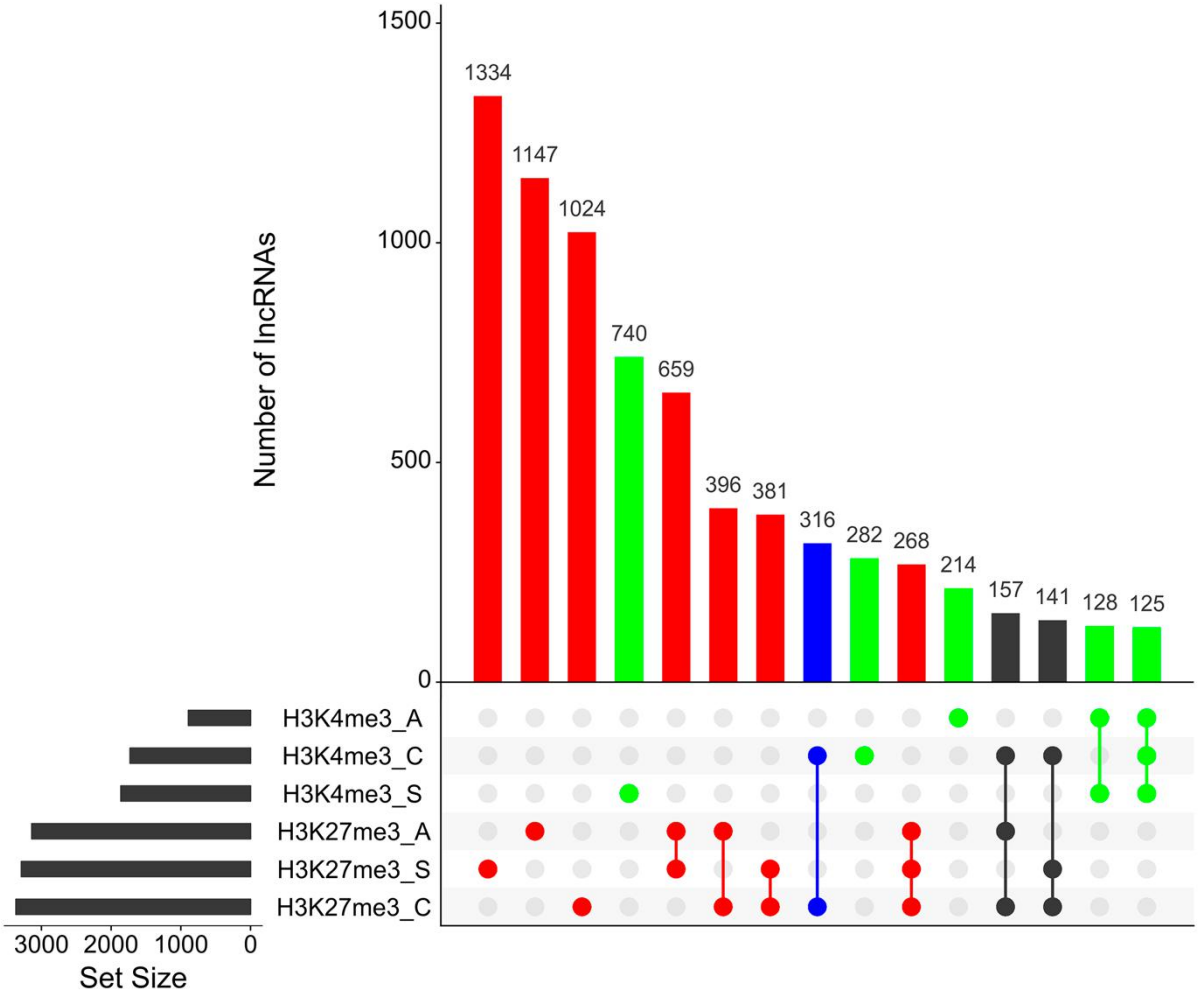


Figure 5 – Hundreds of *S. mansoni* lncRNAs have histone transcriptional activating or repressive marks at their TSS exclusively in one life-cycle stage. The UpSet intersection diagram shows the number of *S. mansoni* lncRNAs (y-axis) that have been detected in each of the intersection sets, indicated by the connected points in the lower part of the plot, as having the H3K4me3 transcriptional activating marks (green) and/or the H3K27me3 repressive marks (red) within 1 kb (upstream or downstream) from their TSS. Six histone mark datasets indicated at the bottom left were analyzed: H3K4me3_A in adults, H3K4me3_C in cercariae, H3K4me3_S in schistosomula, H3K27me3_A in adults, H3K27me3_S in schistosomula and H3K27me3_C in cercariae, and each Set Size black bar represents the number of lncRNAs that contain the indicated histone mark at the indicated stage. The top 15 most enriched intersection sets are shown; all intersection sets and the lists of lncRNAs in each intersection set are shown in Supplementary Table S5. The intersection set in blue shows the number of lncRNAs with the simultaneous H3K4me3_C/ H3K27me3_C marks at their TSS in cercariae, characteristic of poised promoters.

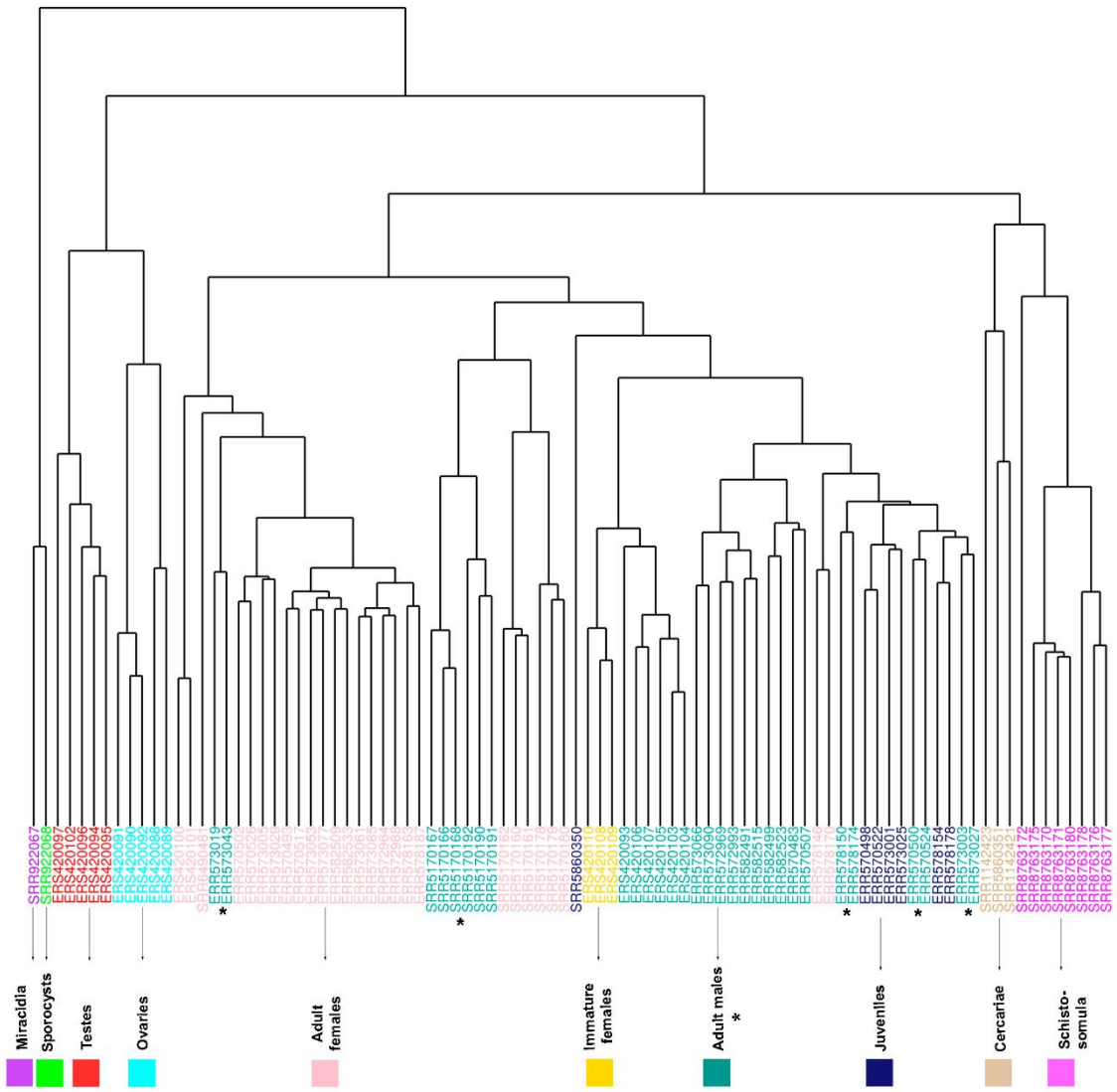


Figure 6 – RNA-seq samples clustering based on Euclidian distance according to the expression levels of all genes used for WGCNA. The expression levels of all lncRNA and mRNA genes from all RNA-seq datasets analyzed in the WGCNA were used for unsupervised clustering of the samples, including RNA-seq datasets from adult females (pink), adult males (turquoise), cercariae (tan), immature females (gold), juveniles (midnight blue), miracidia (purple), ovaries (cyan), schistosomula (magenta), sporocysts (green) and testes (red). The SRA or ENA accession number for each RNA-seq library is indicated at each leaf. The asterisks mark the adult male datasets, whose clustering pattern is the most spread one.

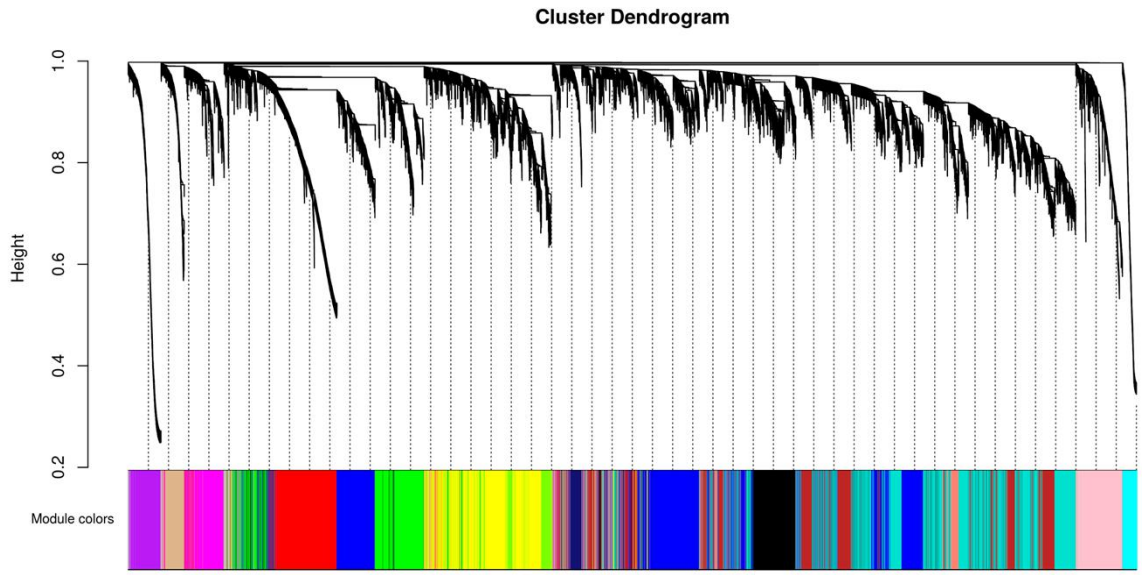


Figure 7 – Identification of gene co-expression modules among the different RNA-seq libraries analyzed with the WGCNA tool. Gene hierarchical cluster dendrogram based on a dissimilarity measure of the Topological Overlap Matrix (1 – TOM) calculated by WGCNA, together with the 15 assigned module colors.

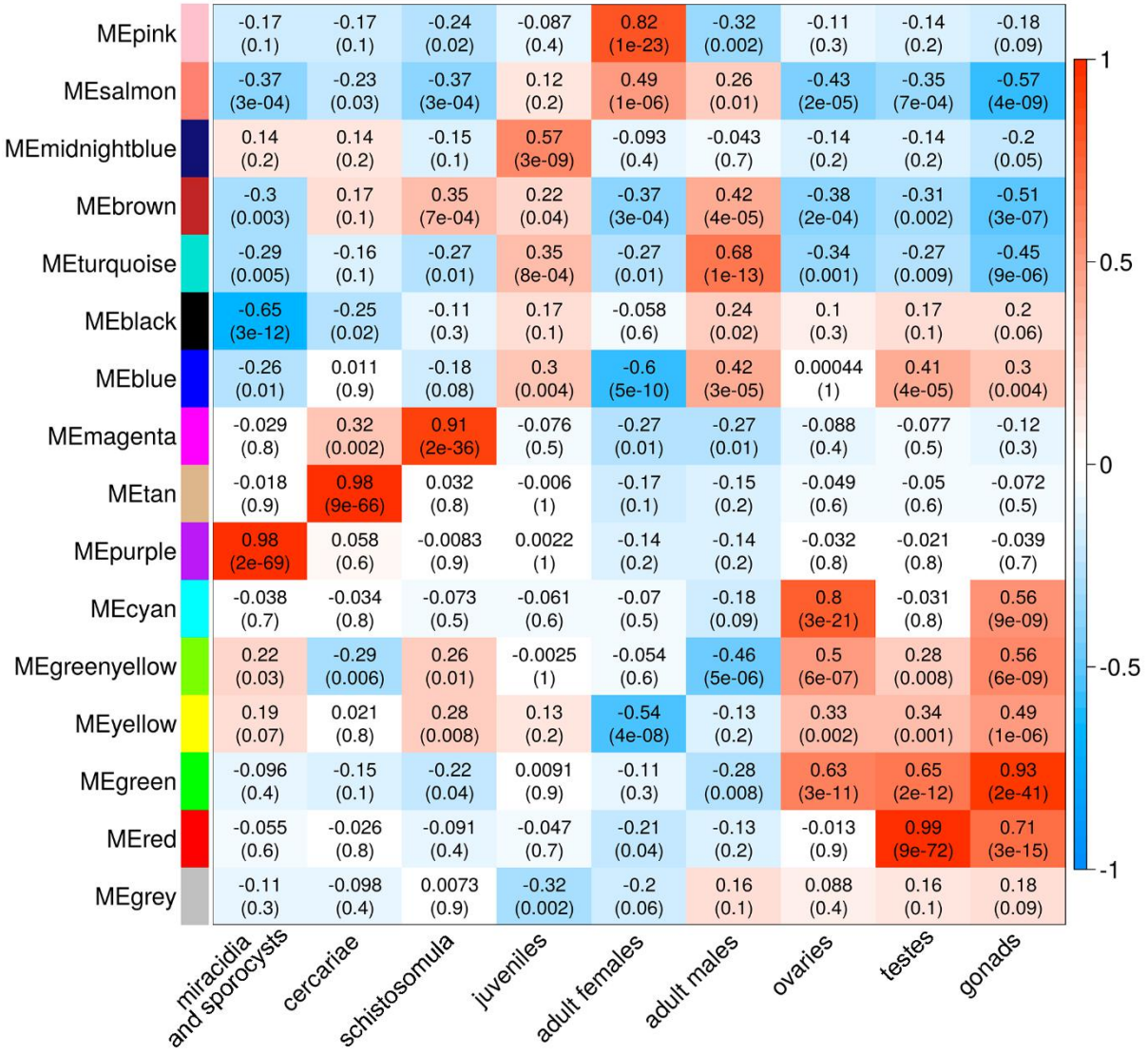


Figure 8 – Each parasite life-cycle stage (or tissue) has at least one highly correlated gene co-expression module. Each cell in the table shows the Pearson correlation (with the p-value in parenthesis) between each of the 15 co-expression modules determined by WGCNA (indicated at left) and the stages/tissues of *S. mansoni* (indicated at the bottom). The cells are colored according to the scale (at right), which is related to the Pearson correlation coefficient.

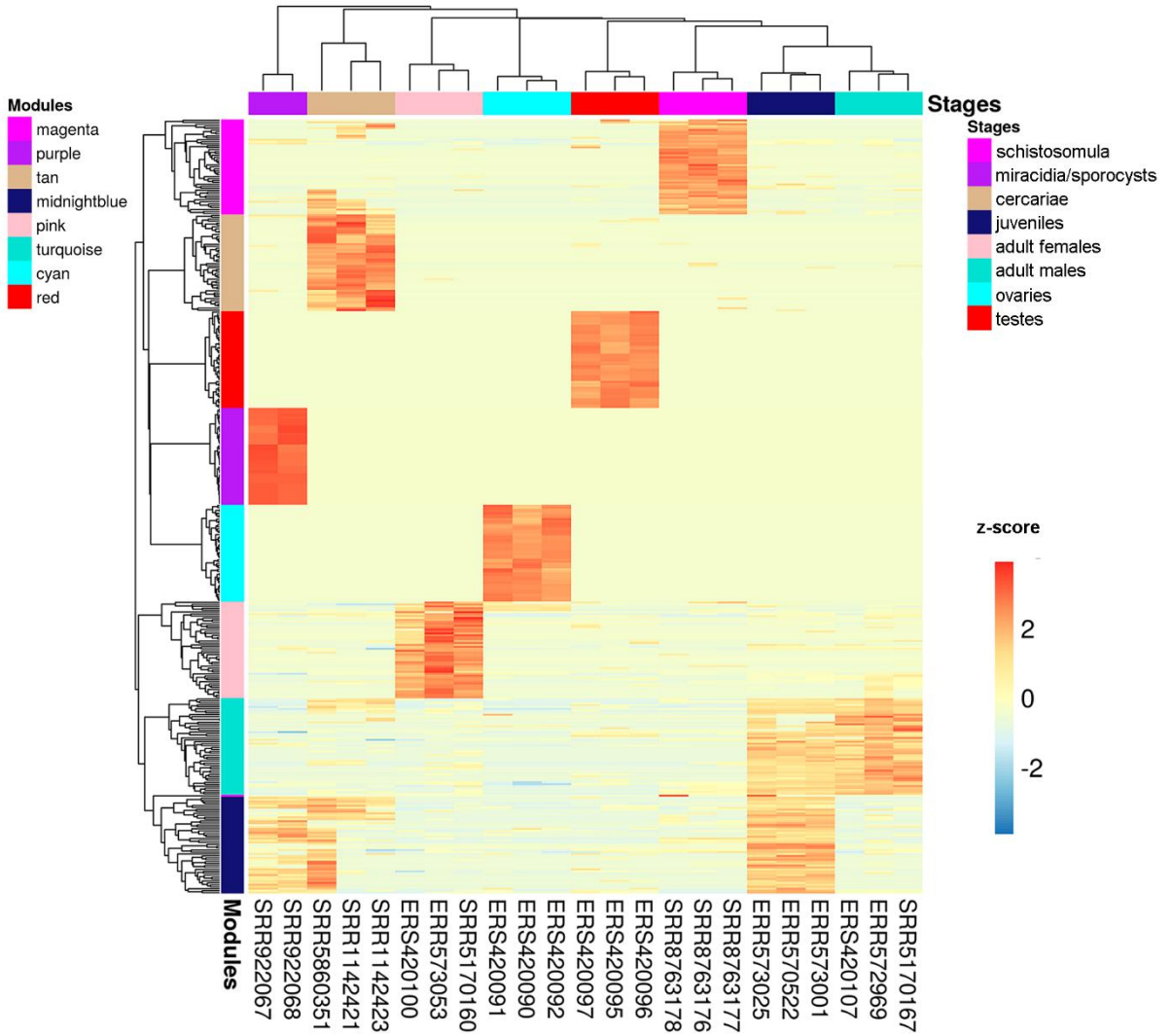


Figure 9 – Gene expression heatmap across the life-cycle stages/tissues of the parasite. Representative heatmap of gene expression levels for the top 50 genes (each in one line) with the highest GMM values from each of the eight modules (indicated at left) with the highest positive correlation to each stage/tissue (indicated at the top). Expression data from three chosen RNA-seq libraries (one in each column) were picked as representative libraries for each stage/tissue, and their SRA or ENA accession numbers are given at the bottom; for miracidia/sporocysts only two RNA-seq libraries were available. Unsupervised clustering using the Euclidean distance was performed; expression of each gene (in one line) is shown as the z-score (from -3 to 3), which is the number of standard deviations above (red) or below (blue) the mean expression value of that gene across all RNA-seq libraries; the z-score color-scale is shown on the right.

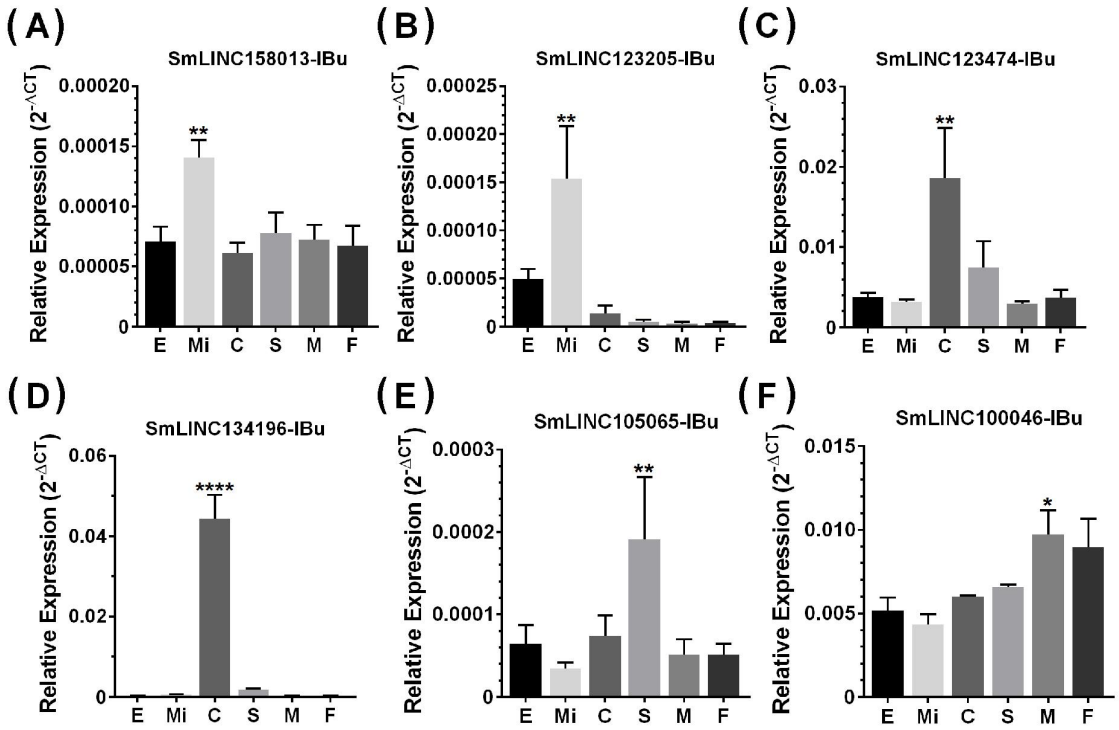


Figure 10: Confirmation by RT-qPCR of the module-specific lincRNAs relative expression. Six lincRNAs were measured at different developmental stages of *S. mansoni*. From left to right in the x-axis, lincRNAs were measured in RNA samples from eggs (E), miracidia (Mi), cercariae (C), *in vitro* mechanically transformed schistosomula cultivated for 24 h (S), adult males (M) and females (F). The lincRNAs relative gene expression was calculated against the geometric mean of two housekeeping genes: Smp_090920 and Smp_062630. **(A)** and **(B)** show SmLINC158013-IBu and SmLINC123205-IBu representing the **purple** module, specific for miracidia/sporocysts. In **(C)** and **(D)** the SmLINC123474-IBu and SmLINC134196-IBu representing the cercariae-specific **tan** module. In **(E)** the schistosomula-specific lincRNA SmLINC105065-IBu from the **magenta** module and **(F)** the adult male-specific lincRNA SmLINC100046-IBu from the **turquoise** module. Bars represent standard deviation of the mean from four biological replicates for each stage. Two technical replicates were assayed for each of the four biological replicates per stage. The ANOVA Tukey test was used to calculate the statistical significance of the expression differences among the parasite stage samples (*p-value ≤ 0.05; **p-value ≤ 0.01; ****p-value ≤ 0.0001). For clarity purposes, we show only the highest p-value obtained in the ANOVA Tukey test for expression comparisons against one another among the stages.

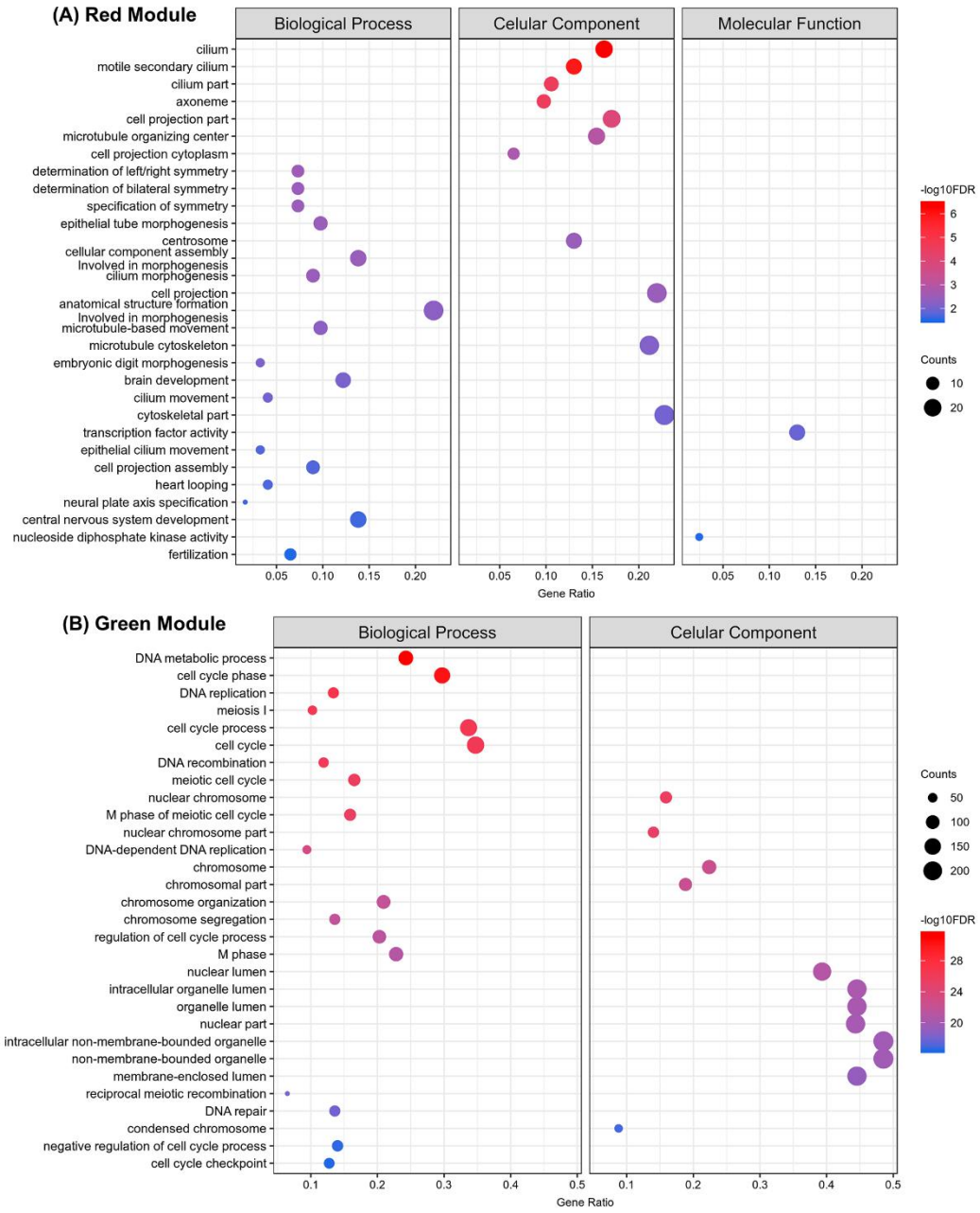


Figure 11 – Top 30 Gene Ontology most significantly enriched terms for protein-coding genes belonging to the red and green co-expression network modules. At left are the enriched GO term annotations. For the (**A**) red (testes) and the (**B**) green (gonads) modules, the enriched GOs are separately represented into the three major GO term categories, namely Biological Process, Cellular Component and Molecular Function. No Molecular Function term was significantly enriched in the green module. The size of the circles is proportional to the number of genes (counts scale on the right) in each significantly enriched GO category, and the colors show the statistical significance of the enrichment, as indicated by the -log10 FDR values (color-coded scales at the right).

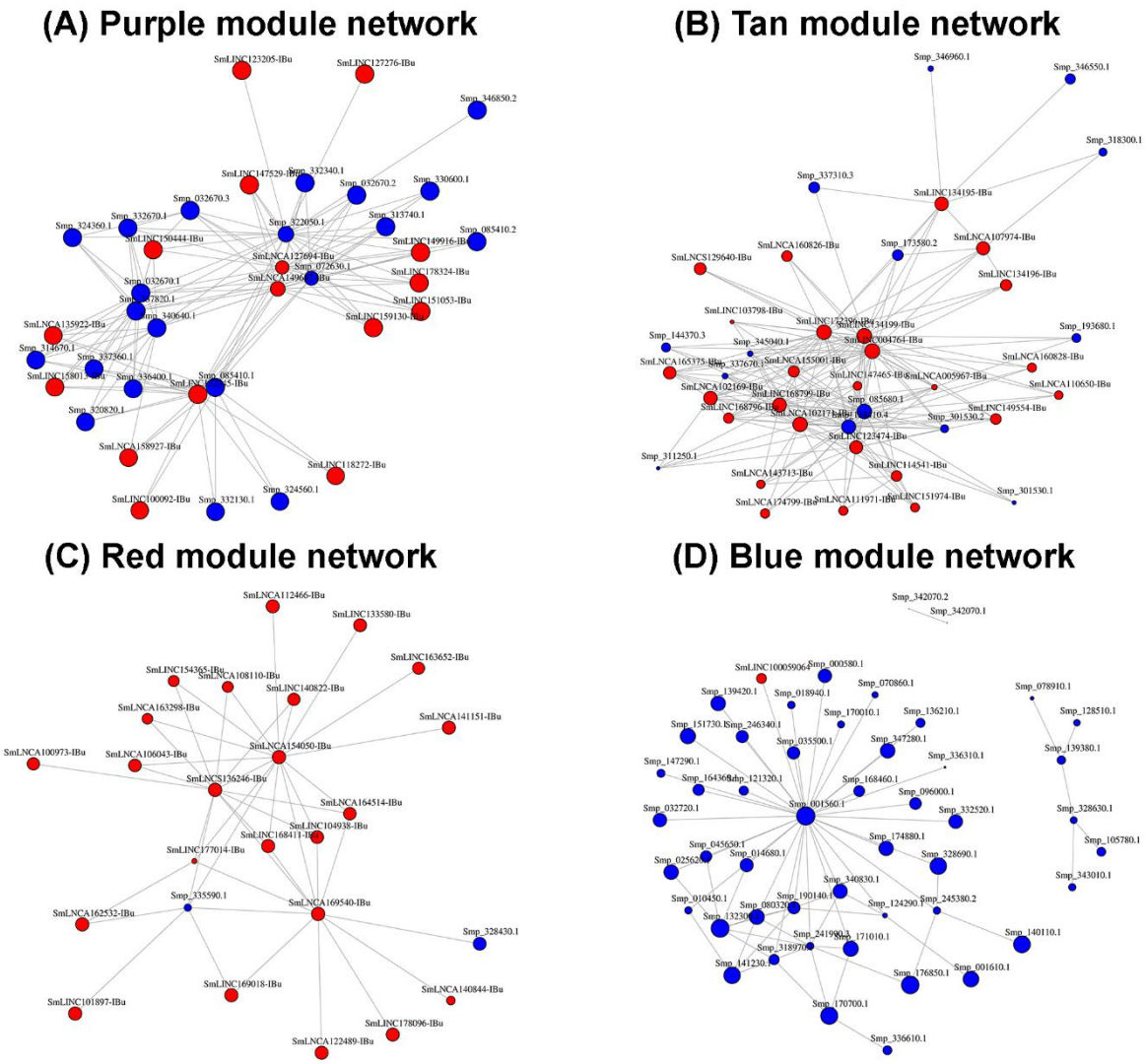


Figure 12 – Gene co-expression networks. The top most highly connected genes from the (A) purple (miracidia/sporocysts), (B) tan (cercariae), (C) red (testes) or (D) blue (females) co-expression network modules are shown. The adjacency thresholds are 0.77, 0.48, 0.58 and 0.26 respectively. Each red circle represents one lncRNA (SmLINC, SmLNCA or SmLNCS) and each blue circle represents one protein-coding gene (Smp_). Circle sizes are related to the intramodular connectivity value for each transcript.

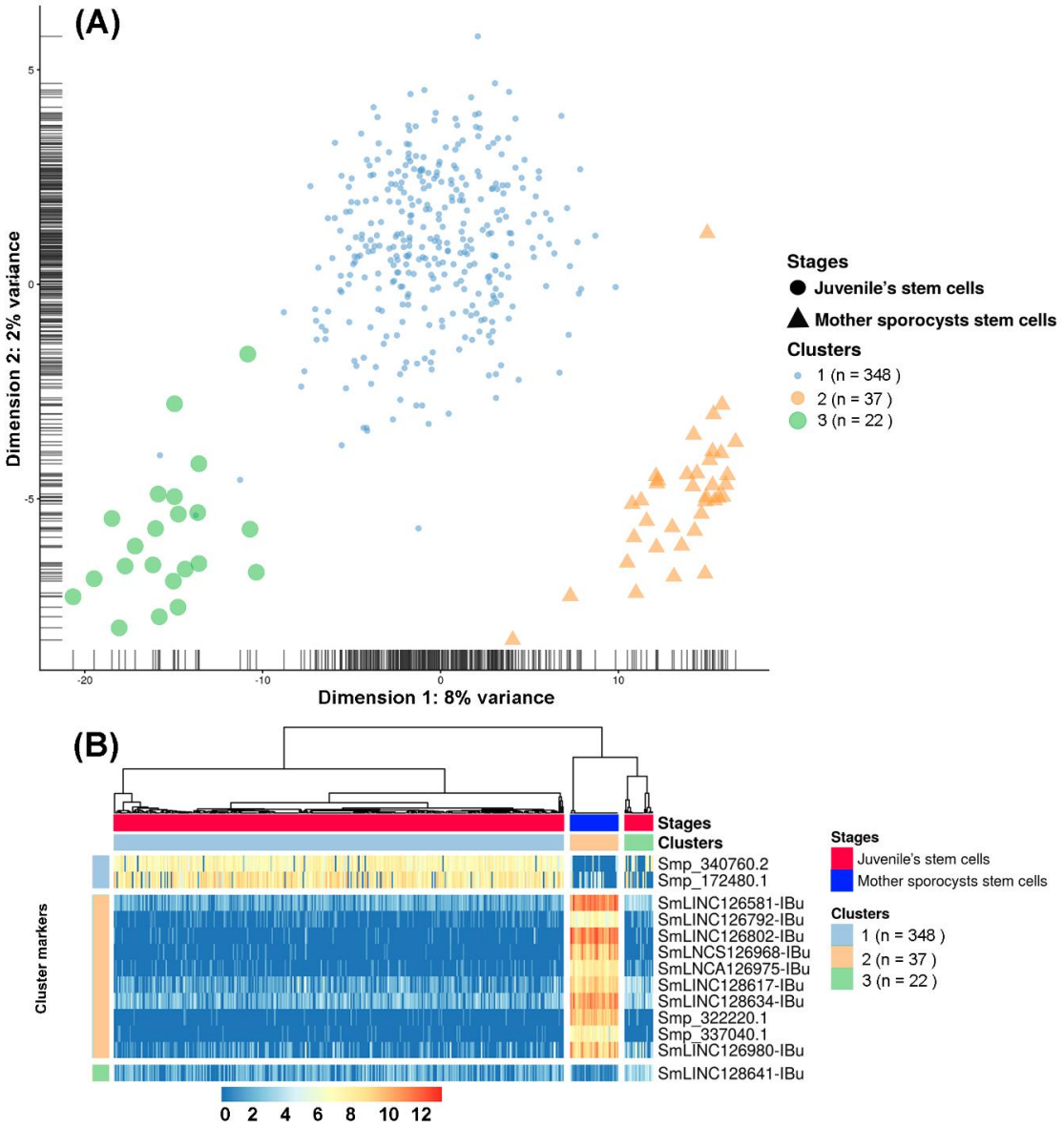
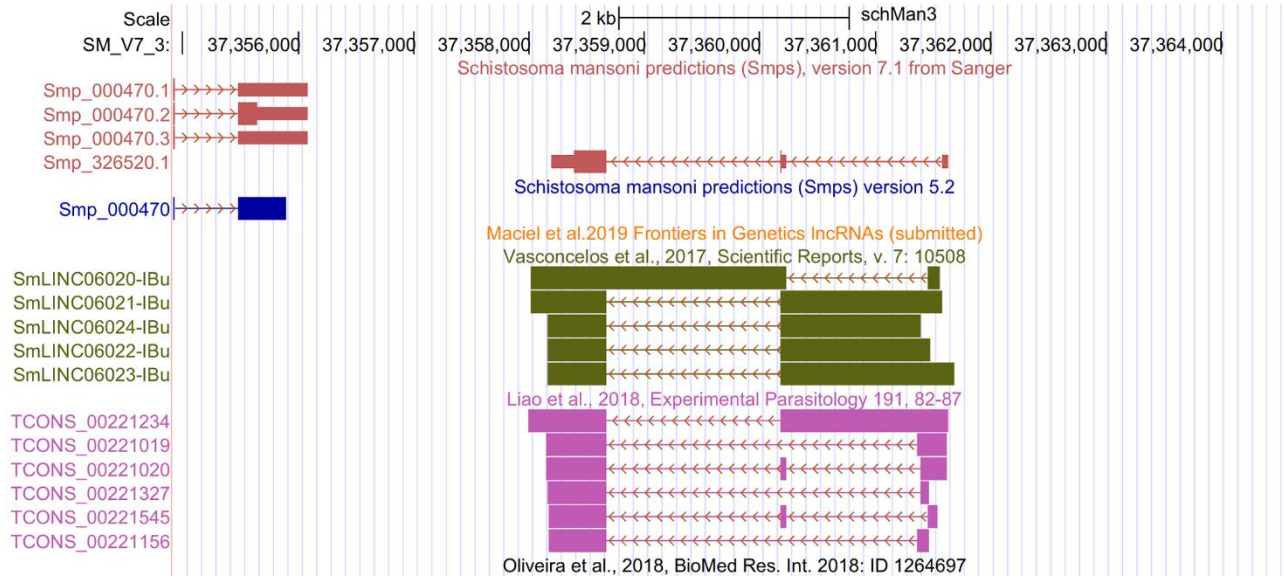
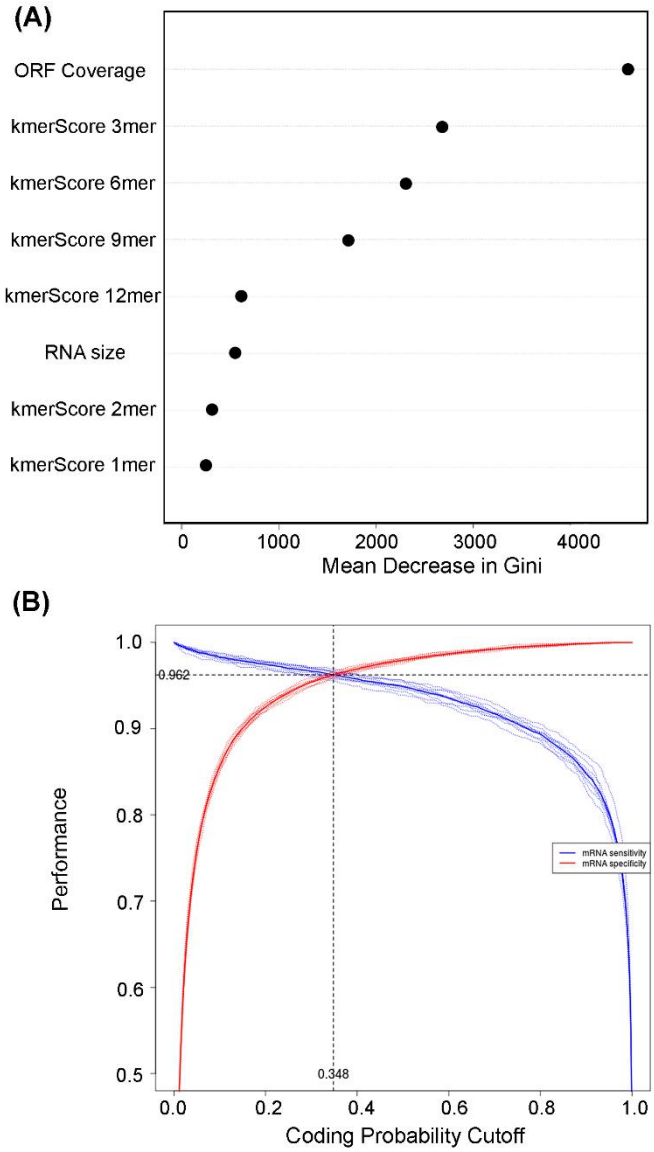


Figure 13 – Single-cell expression analysis identified three different cell population clusters when comparing *S. mansoni* juveniles' stem cells and mother sporocysts stem cells and lncRNAs as gene markers at the single-cell level. (A) Single-cell RNA-Seq data from two RNA-Seq libraries, one from juveniles' stem cells and another from mother sporocysts' stem cells, were analyzed with the SC3 tool that performed an unsupervised clustering of the cells based on the single-cell gene expression data. Principal component analysis plot, where the symbol colors and sizes indicate the three clusters identified by SC3, and the shapes indicate the two life-cycle stages from which the stem cells were isolated. The symbol size is inversely related to the number of cells that belong to the cluster. (B) In the marker-gene expression matrix (log-transformation, represented by the color scale), the statistically significant gene markers are the rows and the cells are columns. The life-cycle stage from which each cell was isolated is indicated by the color bar at the top (stages). The clusters of cells are separated by white vertical lines and are indicated by the second color bar at the top (clusters). The cluster marker genes are separated by white horizontal lines, the markers groups are indicated at left, and the names of the marker genes at right. Only the top 10 most significant marker genes are shown for cluster 2.

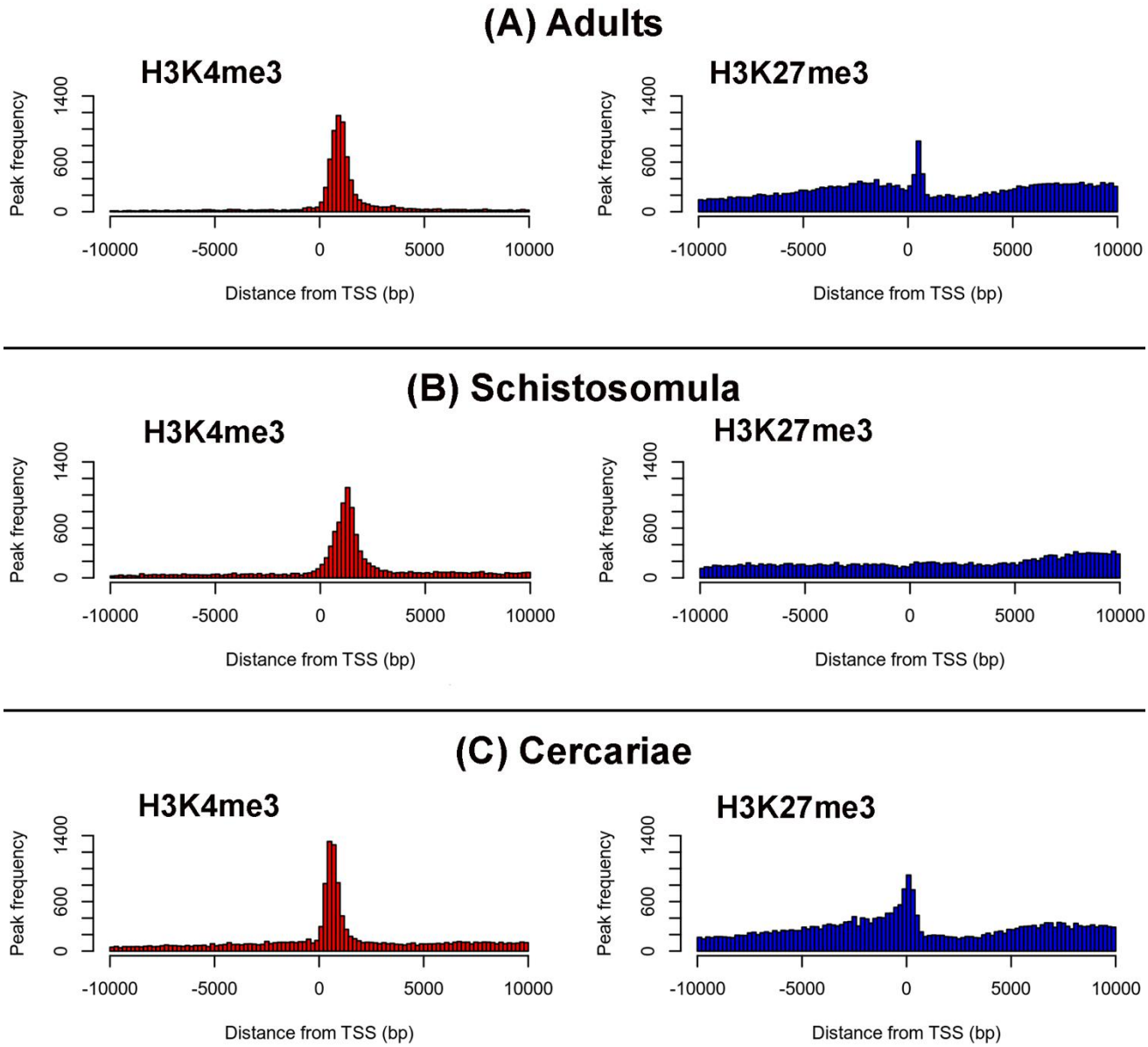
Supplementary Figures



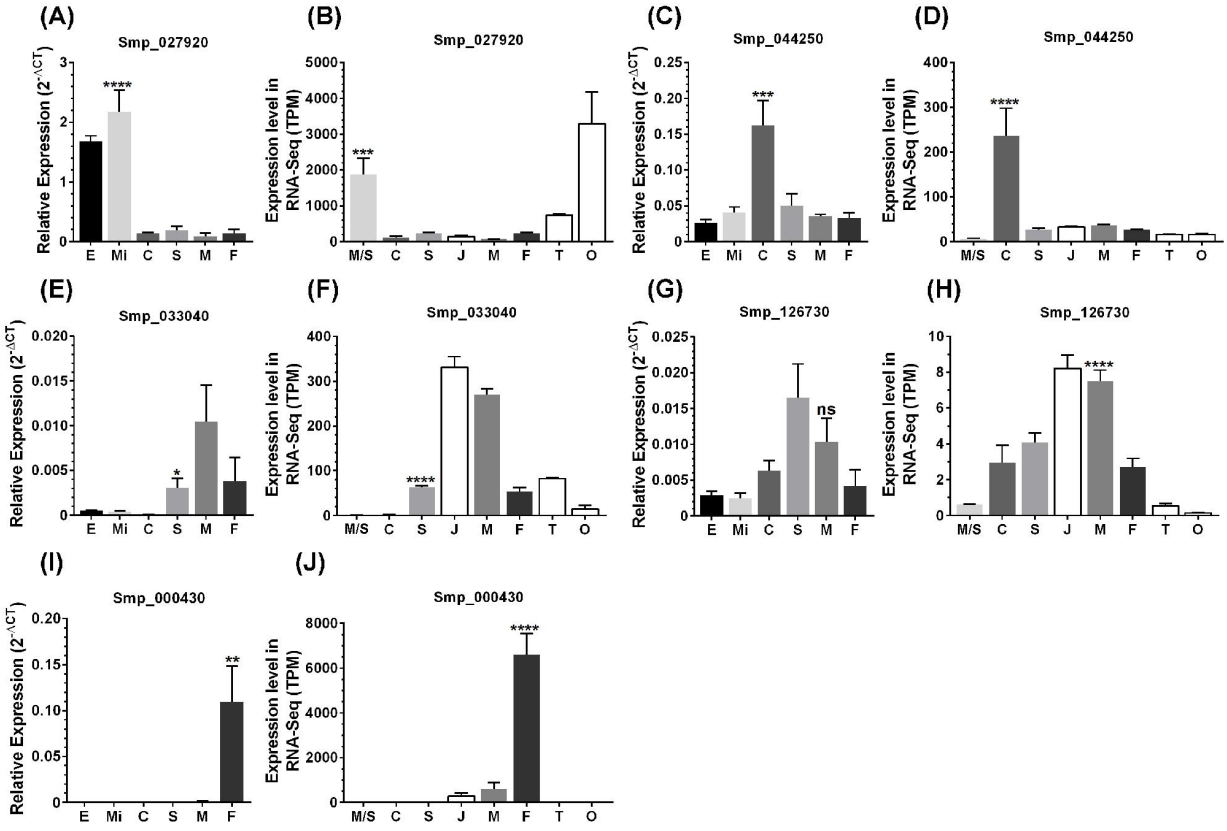
Supplementary Figure S1. Previously annotated *S. mansoni* lncRNAs are in fact partially processed pre-mRNAs in a genomic *locus* encoding a new protein-coding gene. Snapshot of a *S. mansoni* genome browser image, showing a region spanning 9.7 kb on chromosome 3 with coordinates SM_V7_3:37,354,920-37,364,636 (top black row). The red track (top) shows a novel protein-coding gene Smp_326520.1 newly annotated in transcriptome version 7.1, and not present in the old transcriptome version 5.2 (blue track, middle). The orange track (just below the blue track) represents the transcripts annotated in the present work, and it no longer shows any lncRNA in this *locus*. Partially processed pre-mRNAs with intron retention can be recognized among the five supposedly intergenic lncRNAs (SmLINC06020-IBu to SmLINC06024-IBu, grey track) that were annotated in the previous work by Vasconcelos et al. (2017), and among the six different supposedly intergenic lncRNAs (TCONS_00221xxx, pink track) that were annotated in the previous work by Liao et al. (2018). No lncRNAs were annotated in this *locus* by Oliveira et al. (2018) (empty black track at the bottom).



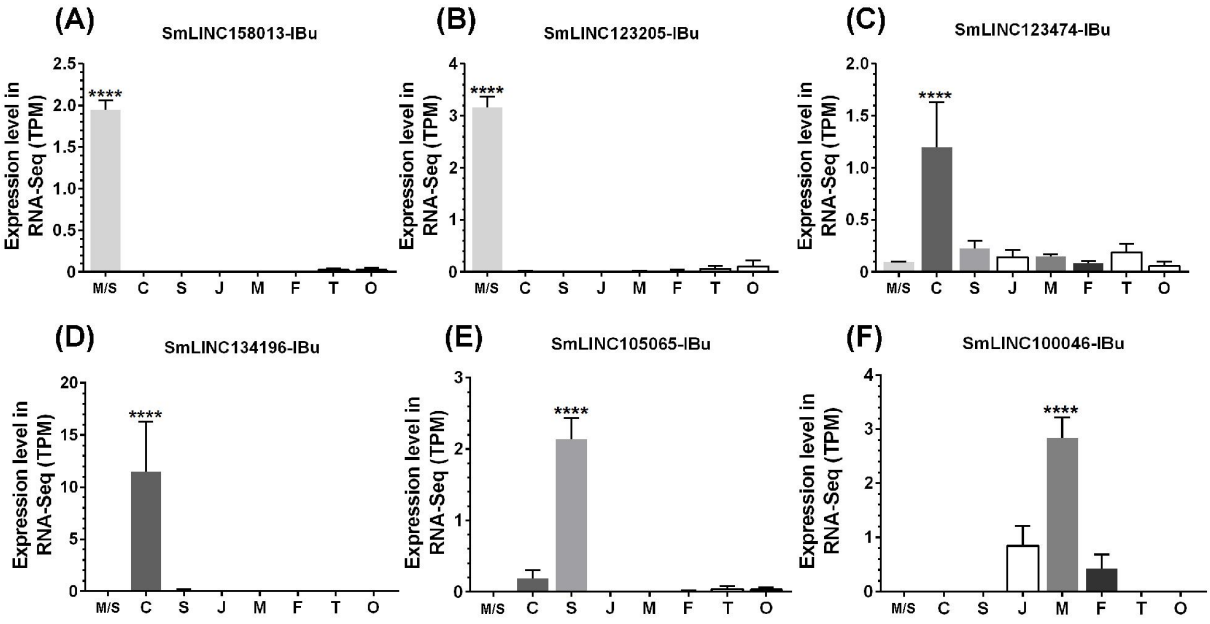
Supplementary Figure S2. Coding probability classification parameters determined by the FEELnc lncRNA classification tool. (A) The graph shows the rank of the eight mRNA sequence features (y-axis) that were used by the Random Forest machine learning FEELnc classifier algorithm to discriminate between known *S. mansoni* mRNAs and putative lncRNAs; these features were ranked based on their discriminatory potential given by the Mean Decrease in Gini metric (x-axis), which is a measure of each feature importance for sequence classification across all of the trees that make up the forest; a higher Mean Decrease in Gini indicates higher feature importance. The most important feature for transcripts classification was the ORF coverage, i.e. the fraction of the total length of the transcript that is occupied by the longest predicted ORF. (**B**) an optimal coding probability cutoff (0.348) was identified, which resulted in 0.962 sensitivity (blue) and specificity (red) of mRNA classification.



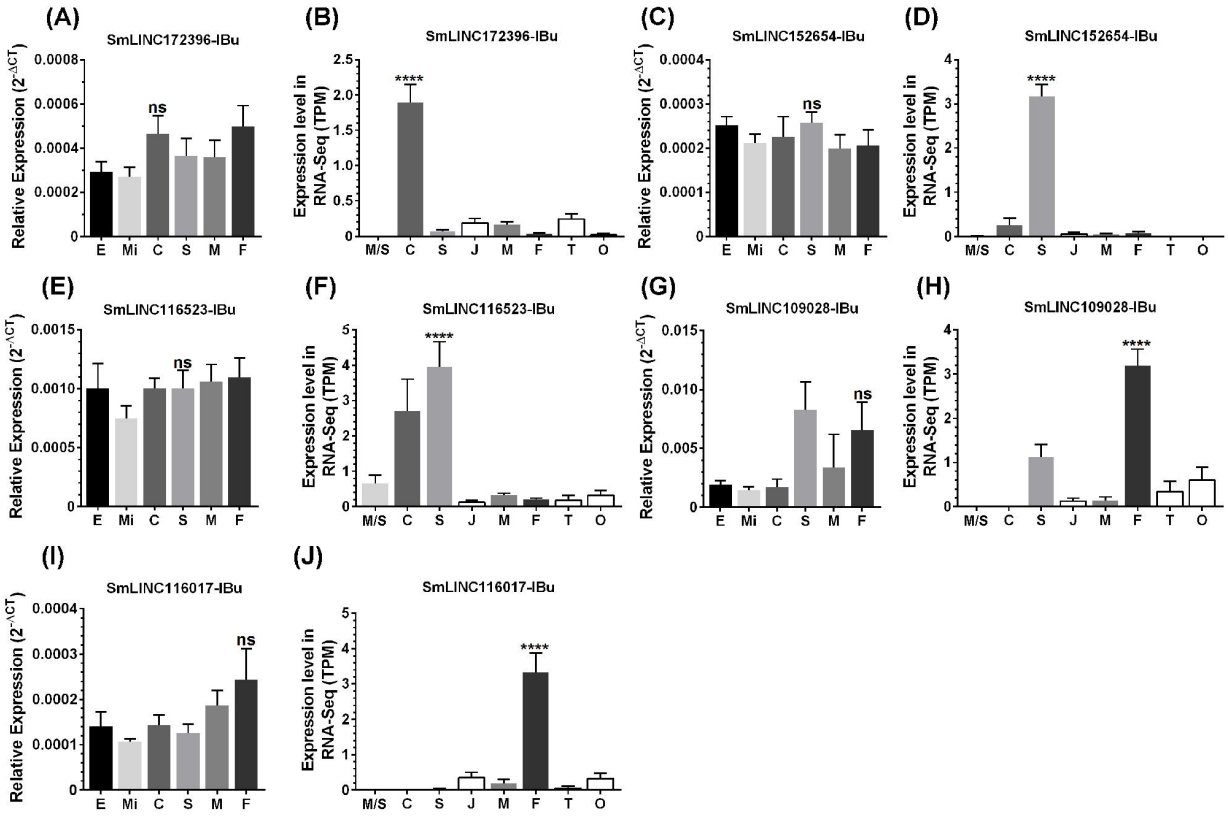
Supplementary Figure S3: Epigenetic histone marks H3K4me3 and H3K27me3 surrounding the TSS of *S. mansoni* protein-coding genes. The frequency of the H3K4me3 marks (red) or of the H3K27me3 marks (blue) mapping within 10 kb around the TSS of all protein-coding genes in (A) adults, (B) schistosomula and (C) cercariae was computed.



Supplementary Figure S4: RT-qPCR and RNA-Seq (TPM) gene expression level of protein-coding genes used as sample markers. Expression of five protein-coding genes was measured at different developmental stages of *S. mansoni* by the RT-qPCR assay, and their stage-specific expression pattern was compared with their expression as determined by the RNA-seq data. In the y-axis, lincRNAs expression levels measured by RT-qPCR (**panels A, C, E, G and I**) or determined from the RNA-seq (TPM) analysis (**panels B, D, F, H and J**) are shown at the stages indicated in the x-axis, as follows: eggs (E), miracidia (Mi), miracidia/sporocysts (M/S), cercariae (C), *in vitro* mechanically transformed schistosomula cultivated for 24 h (S), juveniles (J), adult males (M), adult females (F), and their gonads, namely testes (T) and ovaries (O). The protein-coding genes relative gene expression by RT-qPCR was calculated against the geometric mean of two housekeeping genes: Smp_090920 and Smp_062630. For each lincRNA, the RT-qPCR expression and the RNA-Seq expression results are shown in two panels side by side, as follows: **(A)** and **(B)** Smp_027920 (Tubulin), a gene that marks miracidia; **(C)** and **(D)** Smp_044250 (Metalloprotease), a gene that marks cercariae; **(E)** and **(F)** Smp_033040 (Lactate dehydrogenase), a gene that marks schistosomula; **(G)** and **(H)** Smp_126730 (5-HTR), a gene that marks adult males; and **(I)** and **(J)** Smp_000430 (Egg Shell Protein), a gene that marks adult females. Bars represent standard deviation of the mean from two to thirty biological replicates for each stage. The ANOVA Tukey test was used to calculate the statistical significance of the expression differences among the parasite stage samples (ns: p-value $\geq$ 0.05; *p-value $\leq$ 0.05; **p-value $\leq$ 0.01; **p-value $\leq$ 0.001; ****p-value $\leq$ 0.0001). For clarity purposes, we show only the highest p-value obtained from the ANOVA Tukey test for expression comparisons against one another among the stages.

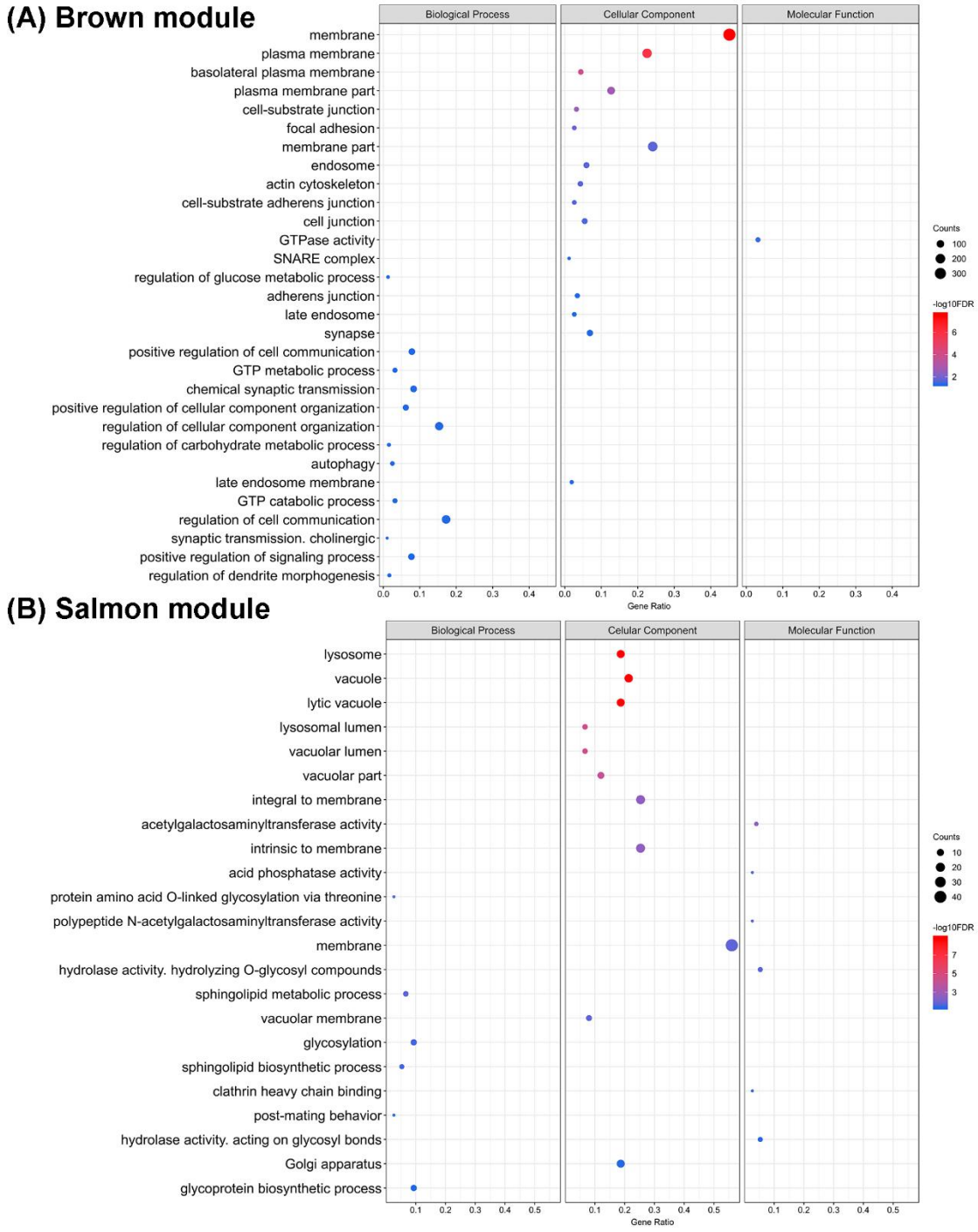


Supplementary Figure S5. RNA-seq expression levels (in transcripts per million, TPM) of the module-specific lincRNAs that had their expression confirmed by RT-qPCR. The six lincRNAs whose gene IDs are indicated at the top of each panel were selected according to their expression levels significantly higher at the given developmental stage of *S. mansoni* indicated in the x-axis, as determined by analyses of the publicly available RNA-seq libraries. The y-axis shows the lincRNA expression level in the RNA-seq assays (TPM) as determined at the stage indicated in the x-axis as follows: miracidia/sporocysts (M/S), cercariae (C), schistosomula (S), juveniles (J), adult males, (M), adult females (F), and their gonads, namely testes (T) and ovaries (O). **(A)** and **(B)** show SmLINC158013-IBu and SmLINC123205-IBu representing the **purple** module, specific for miracidia/sporocysts. **(C)** and **(D)** show SmLINC123474-IBu and SmLINC134196-IBu representing the cercariae-specific **tan** module. **(E)** shows the schistosomula-specific lincRNA SmLINC105065-IBu from the **magenta** module, and **(F)** the adult male-specific lincRNA SmLINC100046-IBu from the **turquoise** module. Bars represent standard deviation of the mean from two to thirty biological replicates for each stage. The ANOVA Tukey test was used to calculate the statistical significance of the expression differences among the parasite stage samples (****p-value $\leq$ 0.0001). For clarity purposes, we show only the highest p-value obtained in the ANOVA Tukey test for expression comparisons against one another among the stages.

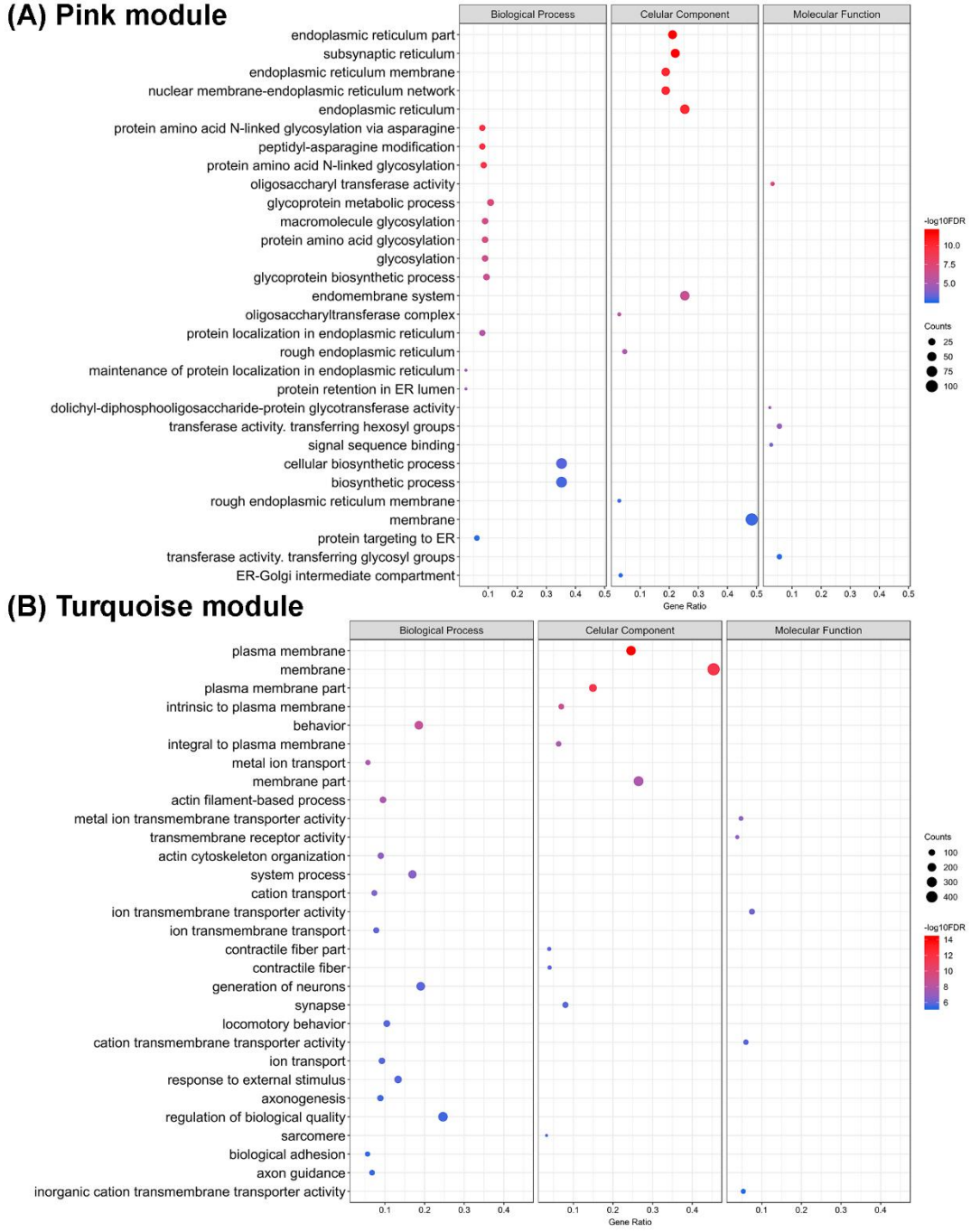


Supplementary Figure S6. RT-qPCR and RNA-Seq (TPM) expression levels of the five lincRNAs that were detected as expressed but did not have the module-specific expression confirmed by the RT-qPCR analysis. Expression of five lincRNAs was measured by RT-qPCR in the RNA from different developmental stages of *S. mansoni* and were not validated as most highly expressed in the same stage as determined from the RNA-Seq data. LincRNAs expression level in the RT-qPCR assay (**panels A, C, E, G and I**) and in RNA-Seq analysis (TPM) (**panels B, D, F, H and J**) was measured in the stages, as indicated in the x-axis: eggs (E), miracidia (Mi), miracidia/sporocysts (M/S), cercariae (C), schitosomula (S), juveniles (J), adult males (M), adult females (F), and their gonads, namely testes (T) and ovaries (O). The lincRNAs RT-qPCR relative gene expression was calculated against the geometric mean of two housekeeping genes: Smp_090920 and Smp_062630. For each lincRNA, the RT-qPCR expression and the RNA-Seq expression results are shown in two panels side by side, as follows: **(A)** and **(B)** SmLINC172396-IBu represents the **tan** module (cercariae); **(C)** and **(D)** SmLINC152654-IBu and **(E)** and **(F)** SmLINC116523-IBu represent the **magenta** module (schitosomula); **(G)** and **(H)** SmLINC109028-IBu and **(I)** and **(J)** SmLINC116017-IBu represent the **pink** module (adult females). Bars represent standard deviation of the mean from two to thirty biological replicates for each stage. The ANOVA Tukey test was used to calculate the statistical significance of the expression differences among the parasite stage samples (ns: p-value ≥ 0.05; ****p-value ≤ 0.0001). For clarity purposes, we show only the highest p-value obtained in the ANOVA Tukey test for expression comparisons against one another among the stages.

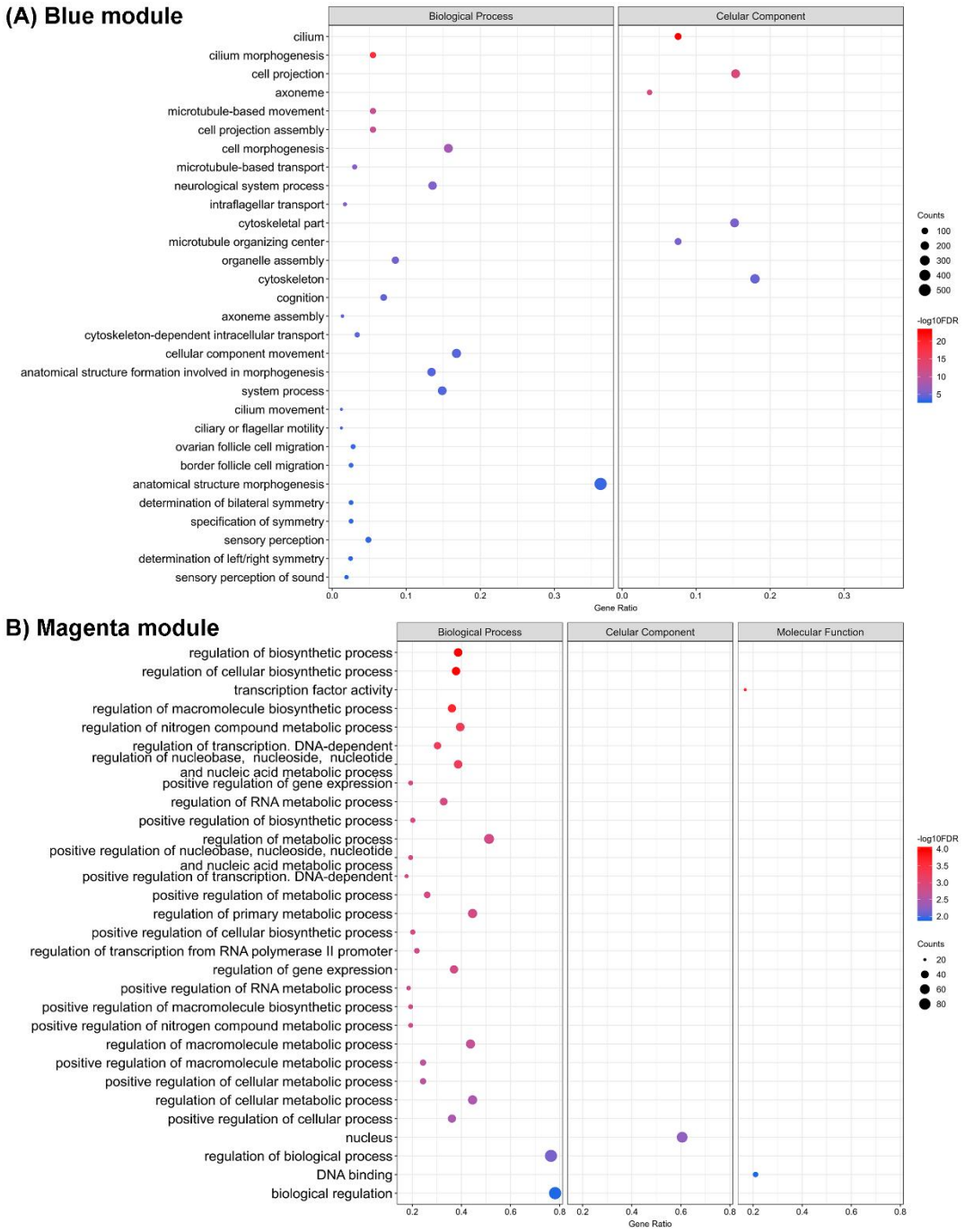
Supplementary Figure S7. Top 30 Gene Ontology most significantly enriched terms for protein-coding genes belonging to the greenyellow and yellow co-expression network modules. At left are the enriched GO term annotations. For the (**A**) greenyellow (gonads) and the (**B**) yellow (females) modules, the enriched GOs are separately represented into the three major GO term categories, namely Biological Process, Cellular Component and Molecular Function. No Molecular Function term was significantly enriched in the yellow module. The size of the circles is proportional to the number of genes (counts scale on the right) in each significantly enriched GO category, and the colors show the statistical significance of the enrichment, as indicated by the -log10 FDR values (color-coded scales at the right).



Supplementary Figure S8. Top 30 Gene Ontology most significantly enriched terms for protein-coding genes belonging to the brown and salmon co-expression network modules. At left are the enriched GO term annotations. For the (**A**) brown (gonads) and the (**B**) salmon (gonads) modules, the enriched GOs are separately represented into the three major GO term categories, namely Biological Process, Cellular Component and Molecular Function. The size of the circles is proportional to the number of genes (counts scale on the right) in each significantly enriched GO category, and the colors show the statistical significance of the enrichment, as indicated by the -log10 FDR values (color-coded scales at the right).



Supplementary Figure S9. Top 30 Gene Ontology most significantly enriched terms for protein-coding genes belonging to the pink and turquoise co-expression network modules. At left are the enriched GO term annotations. For the (**A**) pink (females) and the (**B**) turquoise (males) modules, the enriched GOs are separately represented into the three major GO term categories, namely Biological Process, Cellular Component and Molecular Function. The size of the circles is proportional to the number of genes (counts scale on the right) in each significantly enriched GO category, and the colors show the statistical significance of the enrichment, as indicated by the -log10 FDR values (color-coded scales at the right).



Supplementary Figure S10. Top 30 Gene Ontology most significantly enriched terms for protein-coding genes belonging to the blue and magenta co-expression network modules. At left are the enriched GO term annotations. For the (**A**) blue (females) and the (**B**) magenta (schistosomula) modules, the enriched GOs are separately represented into the three major GO term categories, namely Biological Process, Cellular Component and Molecular Function. No Molecular Function term was significantly enriched in the blue module. The size of the circles is proportional to the number of genes (counts scale on the right) in each significantly enriched GO category, and the colors show the statistical significance of the enrichment, as indicated by the -log10 FDR values (color-coded scales at the right).

3.2 Capítulo II - *S. japonicum*

Este capítulo é referente a identificação e anotação de lncRNAs e análises de co-expressão em *S. japonicum*. O manuscrito, intitulado *Dynamic expression of long non-coding RNAs throughout parasite sexual and neural maturation in Schistosoma japonicum,* foi submetido para o periódico *RNA Biology* (ISSN**:** 1555-8584) e atualmente ainda encontra-se em processo de revisão. O manuscrito, que está disposto na forma como foi submetido, foi escrito por mim, e revisado e aprovado pelos demais co-autores. Declaro que o trabalho aqui apresentado foi realizado por mim, exceto nas partes expressamente indicadas no texto abaixo.

Dynamic expression of long non-coding RNAs throughout parasite sexual and neural maturation in *Schistosoma japonicum*

Lucas F. Maciel[1,2], David A. Morales-Vicente[1,3], Sergio Verjovski-Almeida[1,3]*

[1] Laboratório de Expressão Gênica em Eucariotos, Instituto Butantan, São Paulo, SP 05503-900, Brazil

[2] Programa Interunidades em Bioinformática, Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, SP 05508-900, Brazil

[3] Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, São Paulo, SP 05508-900, Brazil

Abstract

Schistosoma japonicum is a flatworm which causes schistosomiasis, a neglected tropical disease. Due to the importance of sexual maturation for the parasite life-cycle and for host immunopathogenesis, *S. japonicum* RNA-seq analyses had been previously reported in the literature on females and males obtained during sexual maturation from 14 to 28 days post-infection in mouse, resulting in the identification of a number of protein-coding genes and pathways whose expression levels were related to sexual development; however, that work did not include an analysis of long non-coding RNAs (lncRNAs), transcripts that in mammals are known key regulators of vital processes. Here, we applied a pipeline to identify and annotate lncRNAs present in 66 *S. japonicum* RNA-seq libraries publicly available from different life-cycle stages, and performed a co-expression analysis to find stage-specific lncRNAs related to sexual maturation. We identified 12,291 *S. japonicum* expressed lncRNAs; of those, 6,593 were intergenic lncRNAs (lincRNAs), 4,694 antisense, and 1,004 sense lncRNAs. Sequence similarity search and synteny conservation indicated that some 14 % of *S. japonicum* lincRNAs have synteny conservation with *S. mansoni* lincRNAs. Co-expression analysis in males and females from 14 to 28 days post-infection showed that lncRNAs and protein-coding genes in *S. japonicum* have a dynamic co-expression throughout sexual maturation, having a differential regulation between males and females. These genes are related to nervous system development, lipid and drug metabolism and overall parasite survival. Co-expression pattern suggests that lncRNAs possibly regulate these processes or are regulated by the same activation program of the protein-coding genes.

Keywords: Parasitology, lncRNAs, WGCNA, gene co-expression network, synteny, nervous system

Introduction

Schistosomiasis is a disease caused by parasitic trematodes of the genus *Schistosoma*, classified by WHO as a neglected tropical disease [1]. Conservative estimates indicate that at least 230 million people worldwide are infected with *Schistosoma spp.* [2]. *Schistosoma japonicum* is one of the three main species which affects humans, prevalent in Asia, primarily China, Indonesia, and the Philippines [2].

This parasite has a very complex life-cycle comprised of several developmental stages, with a freshwater snail intermediate-host and a final mammalian host [3]. Once the parasites are inside the final host they migrate through blood circulation until they reach the mesenteric veins of the liver, where males and females pair through the gynecophoral canal in order to promote sexual development and female gonad maturation [4] [5]. The paired couples then migrate to the mesenteric veins of the gut, where each female of *S. japonicum* produces 1,000-3,000 eggs per day [6]. These eggs are released in the bloodstream, where they can actively pass through the intestinal wall and be excreted in the feces or be carried by the circulation and trapped in organs where they will cause immunopathologies [7].

Due to the clear importance of sexual maturation for both parasite reproduction and host immunopathogenesis, Wang et al. [8] sequenced the transcriptomes of female and male *S. japonicum* parasites at eight time points obtained from 14 days post-infection (dpi), before the pairing process, until 28 dpi when females and males are paired and sexually mature. In this study, the authors obtained valuable information about the reproductive biology of schistosomes, and were able to identify an insect-like hormonal regulatory pathway along the process of parasites sexual maturation [8].

Wang et al. [8] also identified novel transcripts that had no annotation because they did not match the known protein-coding genes in their reference dataset, and at the time the authors have not considered the possibility that some of them could actually be long non-coding RNAs (lncRNAs). LncRNAs are defined as transcripts longer than 200 nucleotides, without apparent protein-coding potential [9]. They were more recently annotated for the first time in *S. japonicum* by Liao et al. [10] using only two RNA-Seq libraries, one from males and one from females. In mammals, lncRNAs were already identified as important regulators of vital processes, including sexual maturation and reproduction [11]. In *S. mansoni,* our

group has recently shown that some lncRNAs have a tissue-specific expression in ovaries and testis [12], indicating a potential role of lncRNAs in schistosomes reproduction. Thus, the Wang et al. [8] RNA-Seq dataset represents a valuable and so far, unexplored source to identify lncRNAs with dynamic expression throughout sexual maturation in *S. japonicum* and to reveal possible regulatory players in the reproductive biology of these parasites.

The aim of the present work was to annotate a robust and more complete set of lncRNAs in *S. japonicum*, with the pipeline that we developed and previously applied to *S. mansoni* [12]. For this, we used the 66 publicly available RNA-Seq libraries from different life-cycle stages to define a reference set of *S. japonicum* lncRNAs, and we subsequently focused on the 48 libraries from the study of Wang et al. [8] to identify the expression patterns of lncRNAs throughout the sexual maturation process of males and females. Our results provide the basis for future studies on the mechanisms of action of lncRNAs in *S. japonicum* reproductive biology.

Results

Thousands of lncRNAs are expressed in *S. japonicum*

Our pipeline was able to reconstruct 61,298 *S. japonicum* expressed transcripts from the 66 RNA-Seq libraries that were used in this study. The steps to filter pre-mRNAs and mRNAs removed 49,007 transcripts. The remaining 12,291 transcripts, from 7,960 different genes, were classified as lncRNAs expressed in *S. japonicum*; on average 1.5 isoforms per lncRNA gene. Of those, 6,593 were intergenic lncRNAs (lincRNAs), 4,694 antisense lncRNAs, and 1,004 sense lncRNAs. For comparison, in *S. mansoni* a total of 633 RNA-Seq libraries were used and 16 thousand lncRNAs were identified [12]; the difference in the total number of lncRNAs is probably due to the smaller number of RNA-seq libraries from different stages and tissues available for transcript reconstruction in *S. japonicum*.

Liao et al. [10] had previously identified 3,247 and 3,033 potential lncRNAs in *S. mansoni* and *S. japonicum*, respectively. These authors used only two RNA-Seq libraries from *S. japonicum*, one from males and one from females [10]. Here we used a much higher number of RNA-Seq libraries from cercariae, sporocysts, schistosomula, early-development or adult males and females, including those two libraries used by Liao et al. [10], along with the newest version of the genome and transcriptome of *S. japonicum*, which were recently released with significant improvement in assembly contiguity [13]. Our group has recently developed an improved pipeline for identification of lncRNAs [12] and we showed that a

large set of the *S. mansoni* transcripts that were previously annotated as lncRNAs by us with a different pipeline [14] and also by Liao et al. [10], seem now to represent partially processed pre-mRNAs arising from novel protein-coding genes annotated in the newest version of the *S. mansoni* genome and transcriptome [12]. Considering the limited extent of the *S. japonicum* dataset used by Liao et al. [10], and the difficulties with the previous lncRNA identification pipelines [10] [14], we decided to use in further *S. japonicum* analyses only the 12,291 lncRNAs identified here by our present improved pipeline.

Synteny conservation is higher than gene sequence similarity in *Schistosoma* lncRNA genes

A blastn search was used to identify sequence similarity between the *S. japonicum* lncRNAs and lncRNAs from other species. Using a relaxed threshold (see Methods), only 1,578 *S. japonicum* transcripts (13 % of all 12,291 lncRNAs) presented at least one significant hit against *S. mansoni* lncRNAs **(Supplementary Data 1)**; on average there were 21.6 *S. mansoni* hits per *S. japonicum* transcript. Among these 1,578 *S. japonicum* transcripts with sequence similarity to *S. mansoni* lncRNAs, 503 were lincRNAs (7.6 % of all 6,593 *S. japonicum* lincRNAs). This is in accordance with the work of Vasconcelos et al. [14], which demonstrated that genomic regions containing lincRNAs in *S. mansoni* have some sequence conservation among *Schistosoma* species, but much less conservation than that of protein-coding genes.

When the same search was performed against human lncRNAs, the number of hits drastically dropped to 19 lncRNAs **(Supplementary Data 1);** on average there were 10.5 human hits per *S. japonicum* transcript. The low number of similar lncRNAs between human and *S. japonicum* is in accordance with the fast evolution identified in lncRNAs from mammalian species [15].

Because in other species it was demonstrated that some lncRNAs have synteny conservation, even when there is a lack of sequence conservation [16], and because this may indicate an orthologous relationship between these transcripts and a possible functional conservation [16], we searched for syntenic protein-coding blocks between *S. japonicum* and *S. mansoni* genomes. In our synteny analysis, we identified 1,990 protein-coding genes syntenic blocks covering a representative part of both genomes (**Fig. 1, green lines**). Next, we looked for the lincRNAs that were mapped inside these syntenic blocks and verified that there were 4,254 *S. japonicum* lincRNAs, a total of 64 % of all lincRNAs identified in *S. japonicum*.

Next, we extracted the orthologous groups of protein-coding genes identified by OrthoMCL (see Methods), and we looked for the closest protein-coding genes upstream and downstream from the 4,254 lincRNAs. It was possible to identify 934 distinct *S. japonicum* lincRNAs (14.1 % of all 6,593 *S. japonicum* lincRNAs) which were not only contained inside the orthologous syntenic blocks but also had at least one equivalent lincRNA in *S. mansoni* with the same pair of closest orthologous protein-coding genes both upstream and downstream **(Supplementary Data 2)**. Given that there were variable numbers of lincRNA isoforms per locus, the number of matching pairs was 3,144 **(Supplementary Data 2)**, an average of 3.4 matching pairs per each of the 934 distinct *S. japonicum* lincRNAs. The orthologous relationship of these transcripts suggests a possible functional conservation, and shows that there were twice as many lincRNAs with syntenic conservation than with sequence similarity between *S. japonicum* and *S. mansoni.*

Gender-specific lncRNAs transcriptome

Multidimensional scaling (MDS) analysis (**Fig. 2a**) and the heatmap with all protein-coding and lncRNA transcripts expressed in males and females from 14 to 28 days post-infection (**Fig. 2b**) showed that sexually mature females (22 to 28 days post-infection) had a very different expression profile from immature females (14 to 20 days post-infection) and from males. These results are very similar to the ones presented by Wang et al. [8], even though they had not detected the lncRNAs and they used a different genome and transcriptome version, with a different set of bioinformatic tools.

In fact, when we looked only at the lncRNAs (**Fig. 3a and 3b**) we could again separate the sexually mature females from the immature females and from males. This suggests that lncRNAs have an activation program similar to that of protein-coding genes throughout development in *S. japonicum* males and females and that there are sets of lncRNAs which are gender-specific.

It is also similar to the results of Lu et al. [4] which demonstrated in *S. mansoni* that non-paired sexually immature females have a similar expression profile of protein-coding genes to that of males, being the effect of pairing on gene expression more pronounced in the female worms [4].

Differential eigengene network analysis reveals different patterns of nervous system differentiation in males and females during sexual maturation

Function of the vast majority of lncRNAs in all species is still unknown, but one way to predict function on a genome-wide scale is through the guilt-by-association approach [17,18], with the construction of gene co-expression networks between protein-coding genes and lncRNAs, combined with gene ontology enrichment analyses. In *S. mansoni*, we showed that lncRNAs have a dynamic expression throughout life-cycle progression and are hub genes in the gene co-expression networks [12].

To identify the different *S. japonicum* lncRNAs expression profiles and to assess whether the relationship between consensus modules is preserved in males and females throughout sexual maturation, we performed a differential eigengene network analysis with the weighted gene co-expression network analysis (WGCNA) package [19]. In this approach, gene co-expression networks for males and females were first built separately and then the co-expression modules that were shared by both networks were detected, which were named consensus modules. Consensus modules may represent biological pathways that are shared among the compared conditions, in our case males and females, and a differential relationship between the modules may reveal important differences in pathway regulation [20].

In our set, we were able to identify 11 consensus modules (**Fig. 4**). These modules were composed of dozens to thousands lncRNAs, and the ratio between lncRNAs and mRNAs comprising the modules varied from 24 to 50 % (**Table 1** and **Supplementary Data 3**).

Each consensus module was then represented by one eigengene, which is the first principal component of the gene expression data of all transcripts comprising that module and is highly correlated with the expression profile of these transcripts [20]. To identify differences in pathway regulation between the sexes, we examined for both male and female co-expression networks the relationship between all consensus module eigengenes, comparing the modules in pairs (**Fig. 5**). **Fig. 5 panels a and b** show the clustering dendrograms of consensus module eigengenes in females and males, respectively, while **panels 5c and 5f** show the correlation between eigengene pairs represented as a heatmap for females and males, respectively. It is possible to see that in female samples (**Fig. 5a and 5c**) there were two meta-modules (clusters of highly correlated module eigengenes), one composed by magenta, green, yellow and blue module eigengenes and the other composed by the remaining module eigengenes. In males (**Fig. 5b and 5f**), the relationship between the eigengenes and consequently the two meta-modules were not totally preserved. **Fig. 5e** shows the heatmap indicating the preservation of these relationships between the eigengene pairs in females (**Fig. 5c**) and males (**Fig. 5f**). Finally, in **Fig. 5d** we have the preservation measure for each

consensus module eigengene; an overall preservation D = 0.62 was obtained, showing that some modules were preserved in both sexes throughout sexual maturation. Overall, the analyses pointed to biological pathways which had a differential regulation in males and females.

Besides the pairwise comparison between the consensus module eigengenes, we found that the eigengenes were correlated to an external trait, which in this case was the number of days post-infection. The correlation with the days is shown in **Fig. 5** (see "days" label among the modules) but can be better seen in **Fig. 6**, as detailed below.

We found that some module eigengenes had opposite patterns of change as a function of the days post-infection in males and females, such as in the brown module (**Fig. 6d**). In this module, the eigengene value increased in males and decreased in females as the days post-infection increased (**Fig. 6d**). This was reflected in a positive coefficient of correlation (0.94) with the days post-infection in males (**Fig. 6b**, brown module) and a negative coefficient of correlation (-0.81) in females (**Fig. 6a**, brown module). Thus, there was no consensus module relationship between males and females for the brown module, reflecting in the NA annotation in **Fig. 6c**.

Gene ontology enrichment analysis showed that hundreds of protein-coding genes related to neuron projection, differentiation, nervous system development and G-protein receptor signaling pathways were present in the brown module (**Fig. 7 and Supplementary Data 4**). This confirms that, both in *S. mansoni* and *S. japonicum*, sexually mature males have a more active neuronal regulation than sexually mature females, including those mediated by G-protein signaling, which may point to a higher importance of neuronal processes in males during reproduction [21].

Furthermore, sperm motility and axoneme assembly, important for reproduction in males, were enriched in the brown module. Notably, these pathways were also found as enriched in one co-expression module in *S. mansoni* that has lncRNAs as hub genes and is more expressed in the testes [12]. The brown module has 1,607 lncRNAs and 74 out of the 934 lincRNAs that our synteny analysis identified as possibly having orthologous lncRNAs in *S. mansoni*.

The brown module shares dozens of GO terms that were found as enriched by Wang et al. [8] when they performed the GO analysis with the top 901 transcripts in *S. japonicum* males with the highest positive correlation with the percentage of paired worms. Among the shared GO terms, it is included synaptic vesicle endocytosis, cardiac muscle contraction,

reproduction, and the amine transports activity that was further explored by Wang et al. [8]. It should be noted that in the original work [8], the authors did not identify that these transcripts had an opposite pattern of change as a function of sexual maturation in females (**Fig. 6d**).

Expression of genes associated with lipid metabolism and host survival increases simultaneously in males and females

Also, we identified module eigengenes with the same pattern of expression both in males and females as the days post-infection increased (**Fig. 6e-6f**); of note, the presence of patterns of simultaneous change in expression in both sexes has not been explored by Wang et al. [8], who performed GO analyses with sets of genes that exclusively changed either in males or in females. The blue module, whose pattern of expression was positively correlated with the days post-infection in both sexes (**Fig. 6e**), had a number of enriched GO terms associated with lipid and carbohydrates metabolism (**Fig. 8** and **Supplementary Data 5**). It is proposed that glycolysis provides energy sufficient for the survival of schistosomes while vitellocytes are highly dependent on oxidative phosphorylation of lipids [22]. Schistosomes infecting mice living on high-fat diets have a higher fecundity rate than worms infecting mice living with the regular diet [23]. Vitellogenesis was also enriched in the blue module.

Other important pathways associated with survival inside the host were enriched (**Supplementary Data 5**), including drug metabolic process, cell activation involved in immune response, neutrophil-mediated immunity and negative regulation of growth of symbiont in the host. The blue module had a total of 1,520 lncRNAs, including 158 lincRNAs with possible orthologs identified in *S. mansoni*.

The expression of a large number of genes related to neurogenesis is repressed simultaneously in males and females

Importantly, the turquoise eigengene represents genes whose pattern of expression decreased simultaneously in both sexes with the days post-infection (**Fig. 6f**). The turquoise module eigengene showed a negative correlation coefficient with the days post-infection both in females (-0.91) and in males (-0.96) (**Fig. 6a and 6b**), and thus displayed a negative consensus module relationship (-0.91) (**Fig. 6c**). Most interestingly, GO analysis indicated that the turquoise module was enriched with genes related to nervous system development and neurogenesis (**Fig. 9 and Supplementary Data 6**).

The number of protein-coding genes associated with nervous system development in the turquoise module (810 genes) (**Fig. 9 and Supplementary Data 6**) was much higher than the one present in the brown module (425 genes) (**Fig. 7 and Supplementary Data 4**). This indicates that in *S. japonicum* the genes which belong to the brown module are related to nervous system development processes specific to males, being associated with axon projection and male-female signaling, while the genes belonging to the turquoise module are associated with nervous system development that occurs in both sexes and must take place in the first weeks of infection; the latter processes seem to be mostly complete when all worms are paired and mature, as the genes related to these processes were massively turned off.

In this respect, it is known that the parasites pass through considerable changes in different tissues after entering the host, including in the nervous system [24], with some of these changes driven by stem cells in the schistosomula stage (first two weeks of infection) [25,26]. In *S. mansoni* it was recently demonstrated that protein-coding genes associated with embryogenesis, neuronal development, brain development and cell-fate differentiation, such as SOX, procadherin family, Wnt and frizzled receptors, have the highest expression in the sixth day post-infection followed by a steady decline towards the adult stage [27]. In *S. japonicum*, the ortholog protein-coding genes of these genes were all present in the turquoise module: SOX, EWB00_011227 (mRNA16657); procharedin family, EWB00_002205 (mRNA3141), EWB00_00339 (mRNA4912), EWB00_009292 (mRNA13839); Wnt, EWB00_009528 (mRNA14229); frizzled receptor, EWB00_011205 (mRNA16607 and mRNA16608). Besides, the turquoise module contains the neuroendocrine protein 7B2, EWB00_00462 (mRNA6788 and mRNA6789) that was identified as a marker gene for cells of the cephalic ganglia in schistosomula obtained two days post-infection in single-cell RNA-Seq experiments [26].

The turquoise was the largest module, with 8,161 transcripts and 40 % of them were lncRNAs, including 177 lincRNAs with possible orthologs identified in *S. mansoni*. Using the guilt-by-association approach we propose that some of these lncRNAs may regulate the nervous system development processes.

Different sets of genes related to cell replication and general nucleotide metabolism have their expression changed only in one sex throughout the sexual maturation process

We identified eigengene modules with correlation with the days post-infection only in one sex, such as the yellow and green modules, that had a positive correlation only in females

(Fig. 6a to 6c). The green module is more associated with general nucleotide metabolism (**Supplementary Data 7**) while the yellow module is mainly composed of genes associated with gene expression regulation and cell replication (**Supplementary Data 8**), pathways that are also enriched in *S. mansoni* gonads [12]. This pattern of expression was explored by Wang et al. [8], when they performed GO enrichment analysis with 645 transcripts in females with highest correlation with the percentage of females with developed vitellaria. Some of the GO terms identified by them were also present in the green and yellow modules enrichment.

The black module had no correlation with the days post-infection in females and a negative correlation in males and is enriched mainly with genes associated with meiosis, cell cycle regulation, DNA replication and repair **(Supplementary Data 9**). The number of protein-coding genes in the pink and greenyellow modules were too small to produce reliable GO analysis. The modules red, purple, magenta group had no GO enrichment that passed the statistical threshold. The grey group, comprised of genes that do not belong to any module, also had no GO enrichment.

Overall, the identification of lncRNAs co-expressed with protein-coding genes within all modules points to a dynamic expression of lncRNAs throughout sexual maturation in *S. japonicum*; these transcripts are correlated with protein-coding genes related to schistosome survival and reproduction, possibly regulating these processes.

Discussion

Here we have reported the identification of 12 thousand lncRNAs expressed in different life-cycle stages of *S. japonicum*, including cercariae, sporocysts, schistosomula, early-development or adult males and females. Sequence similarity search and synteny conservation indicated that the lncRNAs of *S. japonicum* have synteny conservation with *S. mansoni* lncRNAs even when there is a lack of sequence conservation. Pegueroles et al. [28] analyses of lncRNAs sequence or synteny conservation between the *Caenorhabditis* spp. also identified a much higher number of lncRNAs with synteny conservation than with sequence conservation [28].

We have shown that the lncRNAs in *S. japonicum* have a dynamic expression throughout sexual maturation in both males and females, being correlated to genes involved with nervous system development, lipid and carbohydrate metabolism, drug metabolism and parasite survival. This correlation can suggest that: (1) lncRNAs are regulated by the same activation program of the protein-coding genes; or (2) they are regulating the expression of

these protein-coding genes throughout the development in males and females. The role of lncRNAs regulating gene expression in sexual and tissue maturation in other species has been already reported in the literature [29-31]. Interestingly, it has been shown in amniotes that lncRNAs with dynamic expression patterns across developmental stages show signatures of enrichment for functionality, including a higher number of transcription factor binding sites in their promoters, in comparison with non-dynamic lncRNAs, thus suggesting a stronger and more complex transcriptional regulation [32].

Finally, this study is the initial step towards the functional characterization of the role of *S. japonicum* lncRNAs in sexual maturation. Through synteny and co-expression analyses good candidates can be prioritized for future functional studies. Once the roles of lncRNAs are identified, and given the clear importance of sexual maturation for both parasite reproduction and host immunopathogenesis, lncRNAs involved in these processes could be used as therapeutic targets. The lack of sequence similarity between *S. japonicum* and human lncRNAs, with only 19 lncRNAs having sequence similarity to human lncRNAs, is a very interesting feature for making them good candidates for therapies against schistosomes; as we have previously suggested, targeting the parasite lncRNA transcripts would have a reduced chance of unwanted off-target effects against the host [12]. Although the progression toward the clinic has been slow, lncRNAs represent potentially good therapeutic targets for human diseases [33-35].

Methods

Transcriptome assembly and lncRNAs classification

For this work, we used 66 *S. japonicum* RNA-Seq libraries publicly available at SRA from the following stages: cercariae, sporocysts, schistosomula, early-development or adult males and females (**Supplementary Data 10**). The most recent SKCS01000000 version of the *S. japonicum* genome and transcriptome [13] were used as references.

[Nota: O desenho e a construção do pipeline automatizado, que estão descritos no parágrafo abaixo, são o resultado de uma colaboração com um dos co-autores do artigo, o aluno David A. Morales-Vicente. A aplicação desse pipeline sobre os dados de RNA-Seq de *S. japonicum* e a geração dos resultados descritos na respectiva seção foram realizados por mim.]

The pipeline that was developed and applied by Maciel et al. [12] to identify lncRNAs in *S. mansoni*, was now applied to the above indicated *S. japonicum* RNA-Seq datasets; all parameters used in the pipeline outlined below were previously described in detail [12]. Briefly, adapters and low-quality reads were removed by fastp [36] v 0.19.4, then reads were mapped against the genome with STAR [37] v 2.6.1c in a two-pass mode. RNA-Seq library strandedness was inferred by RSeQC [38] v 2.6.5 and this information was used in transcripts reconstruction and expression levels quantification. Multi mapped reads were removed with Samtools [39] v 1.3 and uniquely mapped reads were used for transcript reconstruction for each library with Scallop [40] v 10.2. A consensus transcriptome from all libraries was built using TACO [41] v 0.7.3.

Transcripts shorter than 200 nt, monoexonic or with exon-exon overlap with protein-coding genes from the same genomic strand were removed from the set. The coding potential of the remaining transcripts was evaluated by means of the FEELnc tool [42] v 0.1.1 with shuffle mode and by CPC2 tool [43] v 0.1. Only transcripts classified as lncRNAs by both tools were kept. The open reading frames (ORFs) were extracted from each putative lncRNA using ORFfinder v 0.4.3 (https://www.ncbi.nlm.nih.gov/orffinder/) and submitted to annotation by eggNOG-mapper webtool [44]. Transcripts with no hits against the eukaryote eggNOG database were then considered as lncRNAs.

When any transcript isoform was classified as a protein-coding mRNA at any step, all transcripts mapping to the same genomic locus were removed to avoid eventual pre-mRNAs. After this final step, a new GTF file was created containing the lncRNAs identified here, plus the protein-coding genes previously annotated by Luo et al. [13].

LncRNAs conservation analysis

To evaluate sequence conservation of lncRNAs from *S. japonicum* across species, a search with the blastn tool [45] v 2.6.0 with a relaxed e-value cutoff of 1e-3 was performed against the sets of *S. mansoni* lncRNAs [12] and human lncRNAs [46] (GENCODE v 31).

Synteny analysis was performed to identify lncRNAs which were contained inside syntenic blocks. In order to do that, genome, proteome and CDS sequences in FASTA format and the GFF3 protein-coding genes annotation from *S. japonicum* and *S. mansoni* were provided to Synima pipeline [47]. The pipeline was run with default parameters, with OrthoMCL [48] v 1.4 as the method to identify orthologous protein-coding genes and DAGchainer [49] to identify the syntenic blocks. Bedtools [50] v 2.27.1 was used to compare

the coordinates from the syntenic blocks identified by the Synima pipeline with the coordinates from the lncRNAs of *S. japonicum*.

Bedtools was also used to identify the closest protein-coding genes upstream and downstream of the intergenic lncRNAs from *S. japonicum* and *S. mansoni*.

Weighted gene co-expression network analysis

The new GTF file was used as the reference along with the genome sequence for mapping the reads of each RNA-seq library from Wang et al. [8], again using the STAR [37] v 2.6.1c, now in the one-pass mode, followed by gene expression quantification with RSEM [51] v 1.3. To reduce noise, transcripts with low expression (sum of counts < 10) were removed and counts transformed by Variance stabilizing transformation using the vst function from DESeq2 package [52] v 1.24.0. From the vst counts we performed a Multidimensional scalling (MDS) using the cmdscale R function using Euclidian distance.

Differential eigengene network analysis was performed, using the Wang et al. [8] RNA-Seq dataset and the weighted gene co-expression network analysis (WGCNA) package [19] v 1.68. For consensus network construction, female and male counts were provided as two different sets in blockwiseConsensusModules function (parameters power = 10, minModuleSize = 100, deepSplit = 2, mergeCutHeight = 0.25, networkType = signed).

Module preservation was plotted using the function plotEigengeneNetworks with the days-post infection as the external information, as described in section II.5 from WGCNA tutorials webpage (https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/). The equations behind the function are detailed by Langfelder and Horvath [20].

Gene Ontology (GO) Enrichment

Protein-coding genes identified by Luo et al. [13] were submitted to GO annotation by eggNOG-mapper [53]. Based on this annotation **(Supplementary Data 11)**, we performed GO enrichment analyses with BINGO [54] for the consensus modules identified by WGCNA. We used a hypergeometric test, the whole annotation as the reference set, and FDR ≤ 0.05 was used as the significance threshold.

Data Availability

The data sets analyzed in this study can be found in the SRA repository (https://www.ncbi.nlm.nih.gov/sra). The specific accession numbers for each and all data sets

that were downloaded from these databases and used here are given in **Supplementary Data 10**. The GTF containing the lncRNAs annotated in this work is available at http://schistosoma.usp.br/.

Disclosure of interest

The authors report no conflict of interest.

Funding

This work was supported by the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) grant numbers 2014/03620-2 and 2018/23693-5 to SV-A. LFM received a FAPESP fellowship (grant number 2018/19591-2) and DAM-V received a fellowship from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). SV-A laboratory was also supported by institutional funds from Fundação Butantan and received an established investigator fellowship award from CNPq, Brasil.

References

1. WHO WHO (2015) Investing to overcome the global impact of neglected tropical diseases: third WHO report on neglected tropical diseases 2015. Geneva, Switzerland: World Health Organization.
2. Colley DG, Bustinduy AL, Secor WE, King CH (2014) Human schistosomiasis. The Lancet 383: 2253-2264.
3. Basch PF (1976) Intermediate host specificity in *Schistosoma mansoni*. Exp Parasitol 39: 150-169.
4. Lu Z, Sessler F, Holroyd N, Hahnel S, Quack T, et al. (2016) Schistosome sex matters: a deep view into gonad-specific and pairing-dependent transcriptomes reveals a complex gender interplay. Scientific Reports 6: 31150.
5. McManus DP, Dunne DW, Sacko M, Utzinger J, Vennervald BJ, et al. (2018) Schistosomiasis. Nature Reviews Disease Primers 4: 13.
6. Cheever AW, Macedonia JG, Mosimann JE, Cheever EA (1994) Kinetics of Egg Production and Egg Excretion by *Schistosoma mansoni* and *S. japonicum* in Mice Infected with a Single Pair of Worms. The American Journal of Tropical Medicine and Hygiene 50: 281-295.

7. Wilson MS, Mentink-Kane MM, Pesce JT, Ramalingam TR, Thompson R, et al. (2007) Immunopathology of schistosomiasis. Immunology & Cell Biology 85: 148-154.
8. Wang J, Yu Y, Shen H, Qing T, Zheng Y, et al. (2017) Dynamic transcriptomes identify biogenic amines and insect-like hormonal regulation for mediating reproduction in *Schistosoma japonicum*. Nature Communications 8: 14693.
9. Cao H, Wahlestedt C, Kapranov P (2018) Strategies to Annotate and Characterize Long Noncoding RNAs: Advantages and Pitfalls. Trends in Genetics 34: 704-721.
10. Liao Q, Zhang Y, Zhu Y, Chen J, Dong C, et al. (2018) Identification of long noncoding RNAs in *Schistosoma mansoni* and *Schistosoma japonicum*. Exp Parasitol 191: 82-87.
11. Taylor DH, Chu ET-J, Spektor R, Soloway PD (2015) Long non-coding RNA regulation of reproduction and development. Molecular Reproduction and Development 82: 932-956.
12. Maciel LF, Morales-Vicente DA, Silveira GO, Ribeiro RO, Olberg GGO, et al. (2019) Weighted Gene Co-Expression Analyses Point to Long Non-Coding RNA Hub Genes at Different *Schistosoma mansoni* Life-Cycle Stages. Frontiers in Genetics 10: 823.
13. Luo F, Yin M, Mo X, Sun C, Wu Q, et al. (2019) An improved genome assembly of the fluke *Schistosoma japonicum*. PLoS Negl Trop Dis 13: e0007612.
14. Vasconcelos EJR, daSilva LF, Pires DS, Lavezzo GM, Pereira ASA, et al. (2017) The *Schistosoma mansoni* genome encodes thousands of long non-coding RNAs predicted to be functional at different parasite life-cycle stages. Sci Rep 7: 10508.
15. Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, et al. (2014) The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature 505: 635-640.
16. Herrera-Ubeda C, Marin-Barba M, Navas-Perez E, Gravemeyer J, Albuixech-Crespo B, et al. (2019) Microsyntenic Clusters Reveal Conservation of lncRNAs in Chordates Despite Absence of Sequence Conservation. Biology (Basel) 8: 61.
17. Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, et al. (2012) Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome research 22: 577-591.
18. Lefever S, Anckaert J, Volders P-J, Luypaert M, Vandesompele J, et al. (2017) decodeRNA- predicting non-coding RNA functions using guilt-by-association. Database : the journal of biological databases and curation 2017: bax042.
19. Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9: 559.

20. Langfelder P, Horvath S (2007) Eigengene networks for studying the relationships between co-expression modules. BMC Systems Biology 1: 54.
21. Lu Z, Spänig S, Weth O, Grevelding CG (2019) Males, the Wrongly Neglected Partners of the Biologically Unprecedented Male–Female Interaction of Schistosomes. Frontiers in Genetics 10: 796.
22. Pearce EJ, Huang SC-C (2015) The metabolic control of schistosome egg production. Cellular Microbiology 17: 796-801.
23. Alencar ACMD, Neves RH, Aguila MB, Mandarim-de-Lacerda CA, Gomes DC, et al. (2009) High fat diet has a prominent effect upon the course of chronic schistosomiasis mansoni in mice. Memorias Do Instituto Oswaldo Cruz 104: 608-613.
24. Collins JJ, 3rd, King RS, Cogswell A, Williams DL, Newmark PA (2011) An atlas for *Schistosoma mansoni* organs and life-cycle stages using cell type-specific markers and confocal microscopy. PLoS Negl Trop Dis 5: e1009.
25. Wang B, Lee J, Li P, Saberi A, Yang H, et al. (2018) Stem cell heterogeneity drives the parasitic life cycle of *Schistosoma mansoni*. Elife 7: e35449.
26. Diaz Soria CL, Lee J, Chong T, Coghlan A, Tracey A, et al. (2019) Single-cell atlas of the first intra-mammalian developmental stage of the human parasite *Schistosoma mansoni*. bioRxiv: 754713.
27. Wangwiwatsin A, Protasio AV, Wilson S, Owusu C, Holroyd NE, et al. (2019) Transcriptome of the parasitic flatworm *Schistosoma mansoni* during intra-mammalian development. bioRxiv: 757633.
28. Pegueroles C, Iraola-Guzmán S, Chorostecki U, Ksiezopolska E, Saus E, et al. (2019) Transcriptomic analyses reveal groups of co-expressed, syntenic lncRNAs in four species of the genus *Caenorhabditis*. RNA biology 16: 320-329.
29. Anguera MC, Ma W, Clift D, Namekawa S, Kelleher RJ, 3rd, et al. (2011) Tsx produces a long noncoding RNA and has general functions in the germline, stem cells, and brain. PLoS Genet 7: e1002248.
30. Golicz AA, Bhalla PL, Singh MB (2018) lncRNAs in Plant and Animal Sexual Reproduction. Trends Plant Sci 23: 195-205.
31. Lawson H, Vuong E, Miller RM, Kiontke K, Fitch DH, et al. (2019) The Makorin lep-2 and the lncRNA lep-5 regulate lin-28 to schedule sexual maturation of the *C. elegans* nervous system. Elife 8: e43660.

32. Sarropoulos I, Marin R, Cardoso-Moreira M, Kaessmann H (2019) Developmental dynamics of lncRNAs across mammalian organs and species. Nature 571: 510-514.
33. Matsui M, Corey DR (2017) Non-coding RNAs as drug targets. Nat Rev Drug Discov 16: 167-179.
34. Blokhin I, Khorkova O, Hsiao J, Wahlestedt C (2018) Developments in lncRNA drug discovery: where are we heading? Expert Opin Drug Discov 13: 837-849.
35. Harries LW (2019) RNA Biology Provides New Therapeutic Targets for Human Disease. Front Genet 10: 205.
36. Chen S, Zhou Y, Chen Y, Gu J (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34: i884-i890.
37. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, et al. (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15-21.
38. Wang L, Wang S, Li W (2012) RSeQC: quality control of RNA-seq experiments. Bioinformatics 28: 2184-2185.
39. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.
40. Shao M, Kingsford C (2017) Accurate assembly of transcripts through phase-preserving graph decomposition. Nature Biotechnology 35: 1167.
41. Niknafs YS, Pandian B, Iyer HK, Chinnaiyan AM, Iyer MK (2016) TACO produces robust multisample transcriptome assemblies from RNA-seq. Nature Methods 14: 68.
42. Wucher V, Legeai F, Hedan B, Rizk G, Lagoutte L, et al. (2017) FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. Nucleic Acids Res 45: e57.
43. Kang YJ, Yang DC, Kong L, Hou M, Meng YQ, et al. (2017) CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res 45: W12-W16.
44. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, et al. (2017) Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. Mol Biol Evol 34: 2115-2122.
45. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.

46. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, et al. (2019) GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res 47: D766-D773.

47. Farrer RA (2017) Synima: a Synteny imaging tool for annotated genome assemblies. BMC Bioinformatics 18: 507.

48. Li L, Stoeckert CJ, Jr., Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Research 13: 2178-2189.

49. Haas BJ, Delcher AL, Wortman JR, Salzberg SL (2004) DAGchainer: a tool for mining segmental genome duplications and synteny. Bioinformatics 20: 3643-3646.

50. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841-842.

51. Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12: 323.

52. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15: 550.

53. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, et al. (2017) Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. Molecular Biology and Evolution 34: 2115-2122.

54. Maere S, Heymans K, Kuiper M (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. Bioinformatics 21: 3448-3449.

Tables and Figures

Table 1. Number of transcripts per module and percentage of lncRNAs in each module.

Module color	Total number of transcripts	Number of lncRNAs	% of lncRNAs
Black	336	82	24
Blue	4,465	1,520	34
Brown	4,400	1,607	37
Green	2,288	960	42
Greenyellow	128	53	41
Magenta	219	91	42
Pink	288	98	34
Purple	174	75	43
Red	521	260	50
Turquoise	8,161	3,300	40
Yellow	2,830	877	31
Total	23,810	8,923	37

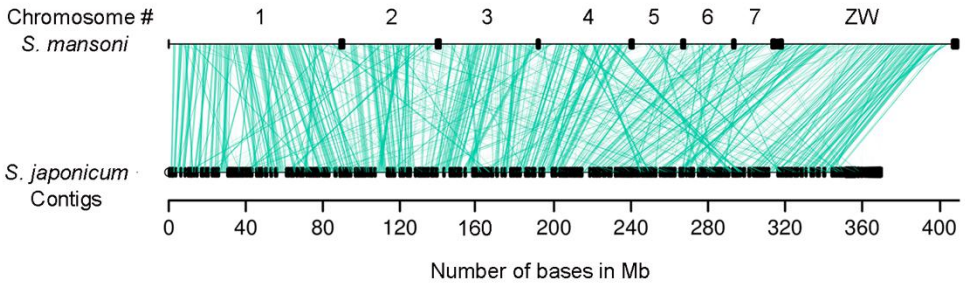


Figure 1 – Synteny of protein-coding genes between *S. mansoni* and *S. japonicum* genomes. Syntenic blocks are indicated by the green lines connecting the genomes. Species names are shown at left and the genomes are represented by horizontal black lines, with vertical black lines indicating scaffold/contig borders. For *S. mansoni* the chromosome number annotations of the longest scaffolds are indicated above the horizontal black line.

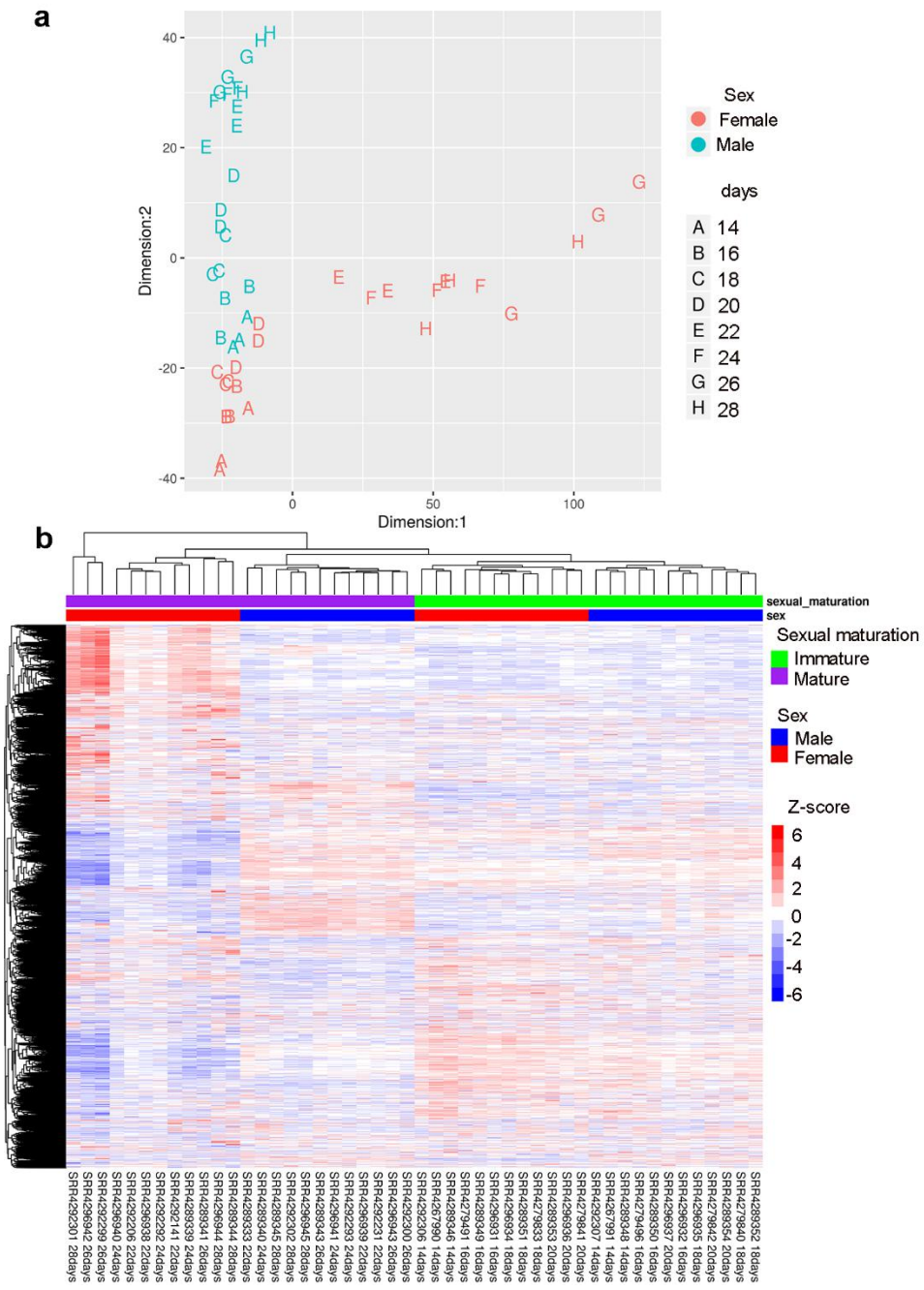


Figure 2 – Protein-coding and lncRNA expression profile of *S. japonicum* throughout sexual maturation. **(a)** Multidimensional scaling (MDS) analyses of protein-coding and lncRNA genes expression together, detected by RNA-seq in 48 samples of *S. japonicum* males (blue) and females (red) perfused from mice at 14 to 28 days post-infection. The number of days post-infection are indicated by the letters A to H according to the legend at right; three replicate samples for each day were analyzed. **(b)** Gene expression heatmap of protein-coding and lncRNA genes together (each in one line), across all 48 samples (each in one column) as indicated by the sample ID label at the bottom. Unsupervised clustering using the Euclidean distance was performed; expression of each gene is shown as the z-score (from −6 to 6), which is the number of standard deviations above (red) or below (blue) the mean expression value of that gene across all RNA-seq libraries; the z-score color scale is shown on the right. Samples from days 14 to 20 are labeled at the top as immature parasites (green), and from days 22 to 28 as mature parasites (purple). Males are labeled at the top in blue and females in red.

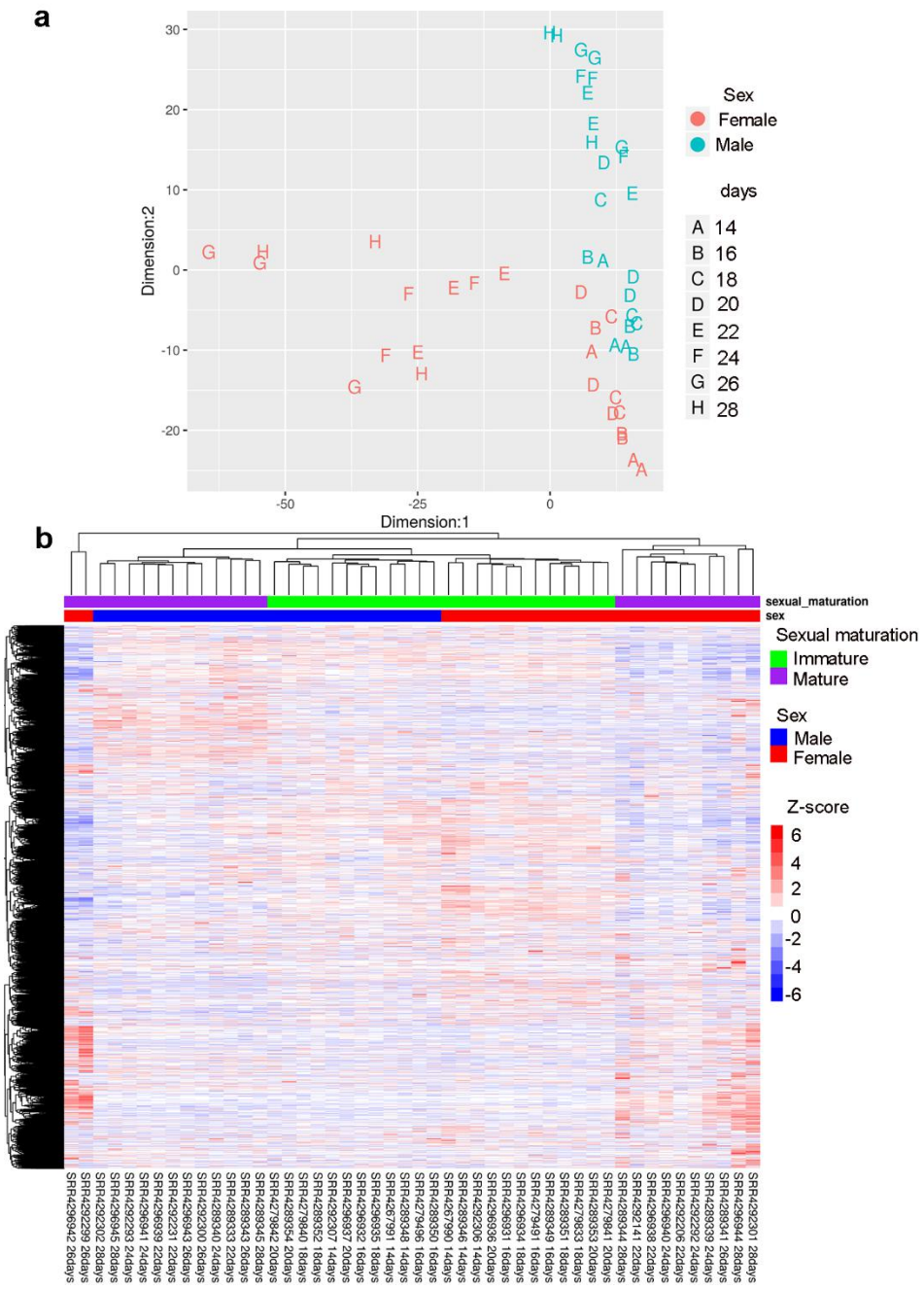


Figure 3 – LncRNA expression profile of *S. japonicum* throughout sexual maturation. (a) Multidimensional scaling (MDS) analyses of lncRNA genes expression alone, detected by RNA-seq in 48 samples of *S. japonicum* males (blue) and females (red) perfused from mice at 14 to 28 days post-infection. The number of days post-infection are indicated by the letters A to H according to the legend at right; three replicate samples for each day were analyzed. **(b)** Gene expression heatmap of lncRNA genes (each in one line), across all 48 samples (each in one column) as indicated by the sample ID label at the bottom. Unsupervised clustering using the Euclidean distance was performed; expression of each lncRNA gene is shown as the z-score (from −6 to 6), which is the number of standard deviations above (red) or below (blue) the mean expression value of that gene across all RNA-seq libraries; the z-score color scale is shown on the right. Samples from days 14 to 20 are labeled at the top as immature parasites (green), and from days 22 to 28 as mature parasites (purple). Males are labeled at the top in blue and females in red.

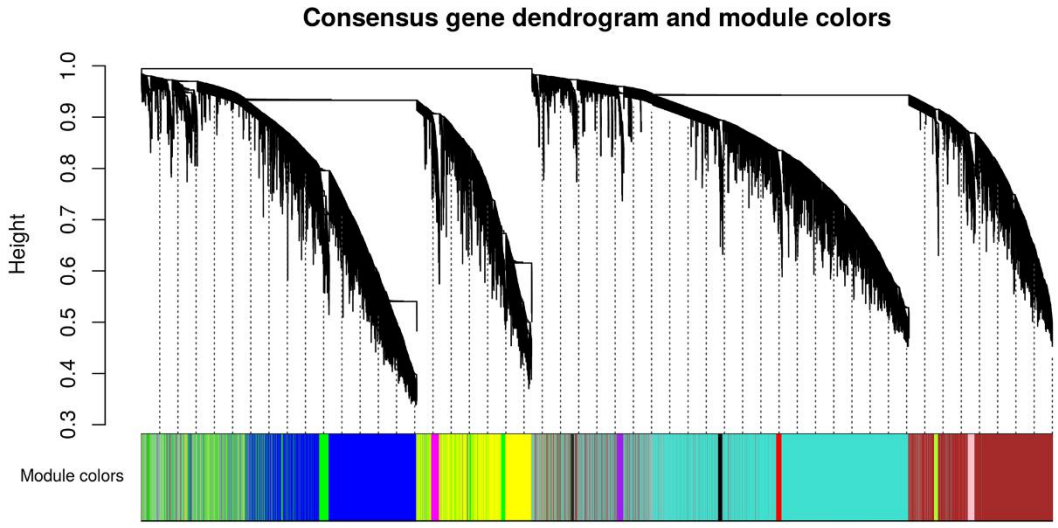


Figure 4 – Consensus modules identified by WGCNA. Gene hierarchical cluster dendrogram based on a dissimilarity measure of the Topological Overlap Matrix (1 – TOM) calculated by WGCNA and the color labels correspond to the different gene co-expression consensus modules identified between male and female datasets.

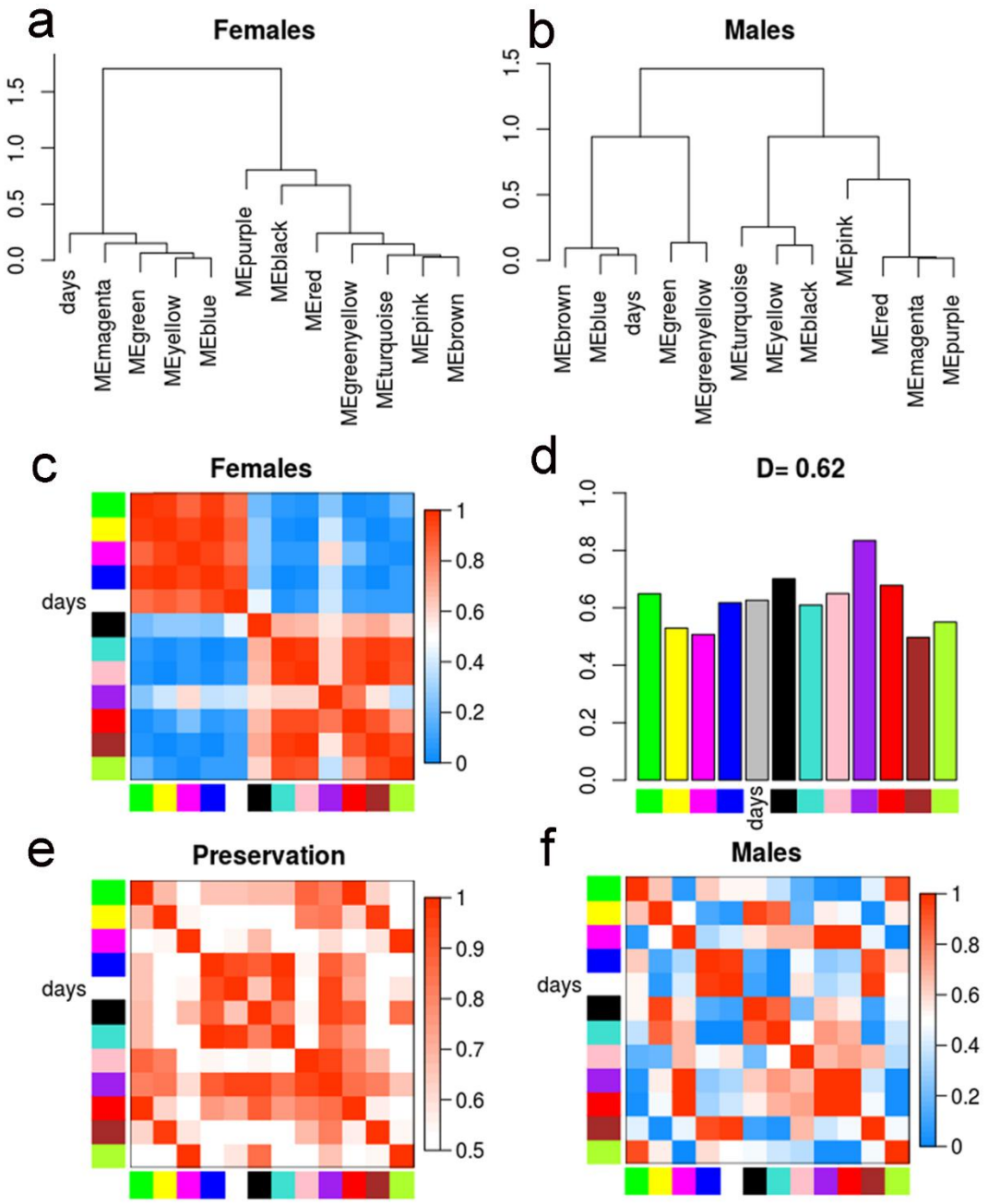


Figure 5 – Differential eigengene network analysis. **(a, b)** Clustering dendrograms of the consensus module eigengenes for females and males, respectively. **(c and f)** Females (c) and males (f) heatmaps of eigengene adjacencies (correlation matrix) in the consensus module eigengenes network. Each row and column correspond to one of the eleven module eigengenes (colors). The correlation between the eigengene and the number of days post-infection (days, white) is also indicated. The cells are colored according to the scale at right; red indicates high adjacency (positive correlation) and blue low adjacency (negative correlation). **(d)** Preservation measure for each consensus eigengene. Each colored bar corresponds to the eigengene of the corresponding module color. The y-axis gives the eigengene preservation measure. The D value denotes the overall preservation of the eigengene networks. **(e)** Heatmap of adjacencies in the preservation network between females and males. Each row and column correspond to a consensus module; saturation of the red color is proportional to preservation according to the color legend.

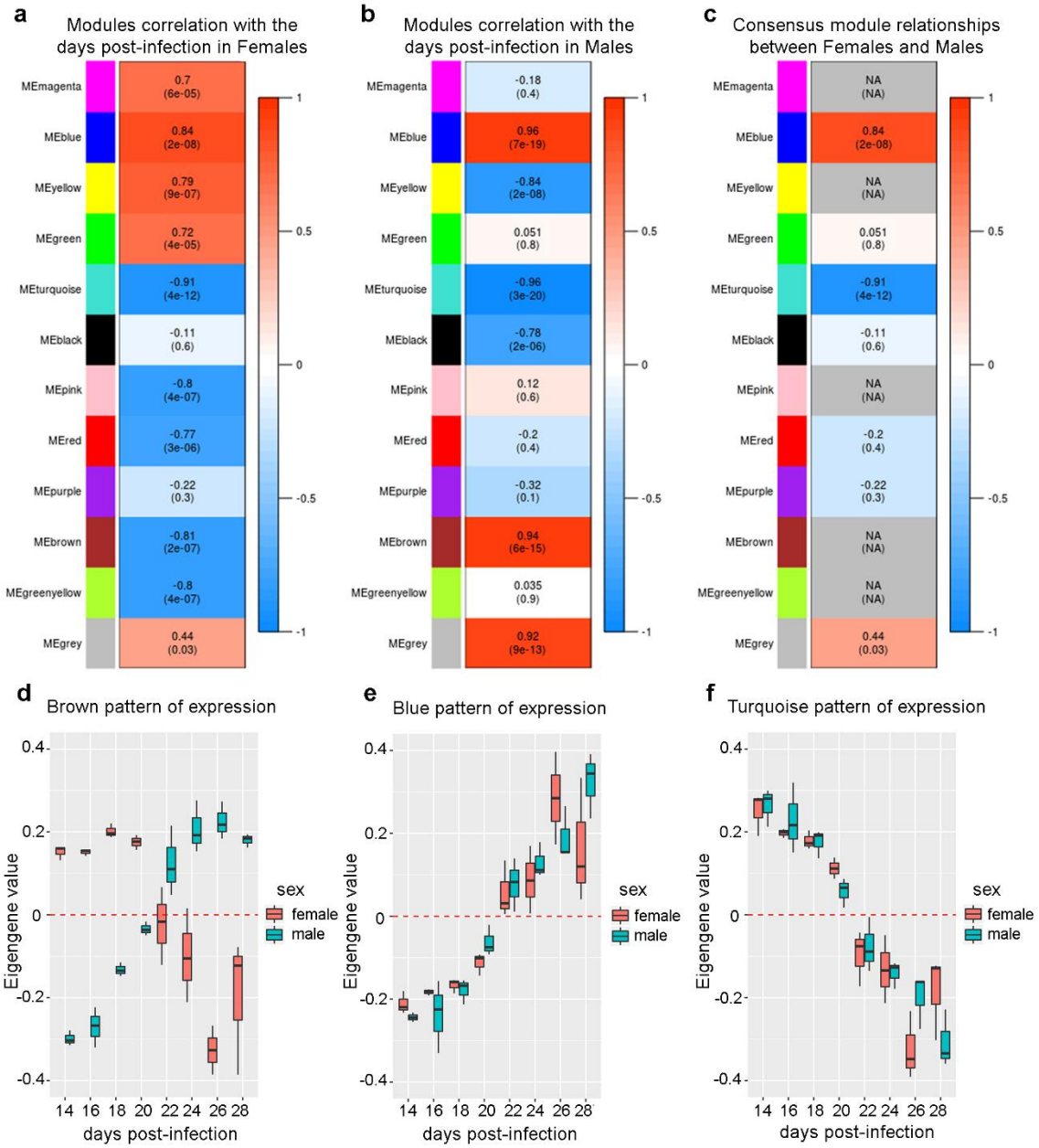


Figure 6 – Relationship between the module eigengenes (ME) and the days post-infection. (a, b) Module eigengenes (ME) correlation with the days post-infection in females (a) and males (b). Each row in the tables corresponds to a module; the number shows the correlation of the corresponding module eigengene with the days, with p-value given in parentheses. Tables are color-coded by correlation according to the color scales at right. **(c)** Consensus module relationships between the female and male module eigengenes and the days post-infection. Consensus relationship correlation is shown when both go in the same direction; missing (NA) entries indicate that the correlations in the male and female datasets have opposite signs. **(d to f)** The boxplots indicate the value of the first principal component, called eigengene, for the samples of males (blue) and females (red) from 14 to 28 days post-infection for the modules brown (d), blue (e) and turquoise (f).

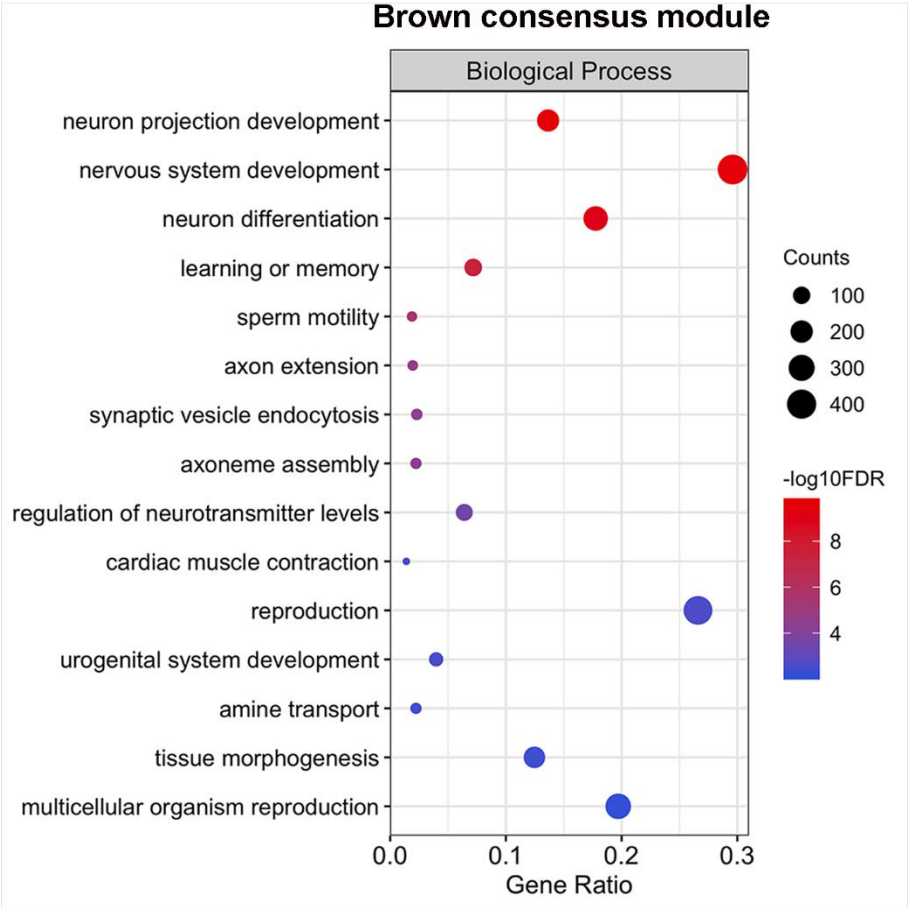


Figure 7 – Selected fifteen Gene Ontology enriched terms for protein-coding genes of the brown consensus module. Only GO terms from the Biological Process category are represented. The size of the circles is proportional to the number of genes (counts scale on the right) in each significantly enriched GO term, and the colors show the statistical significance of the enrichment, as indicated by the -log10 FDR values (color-coded scale at the right).

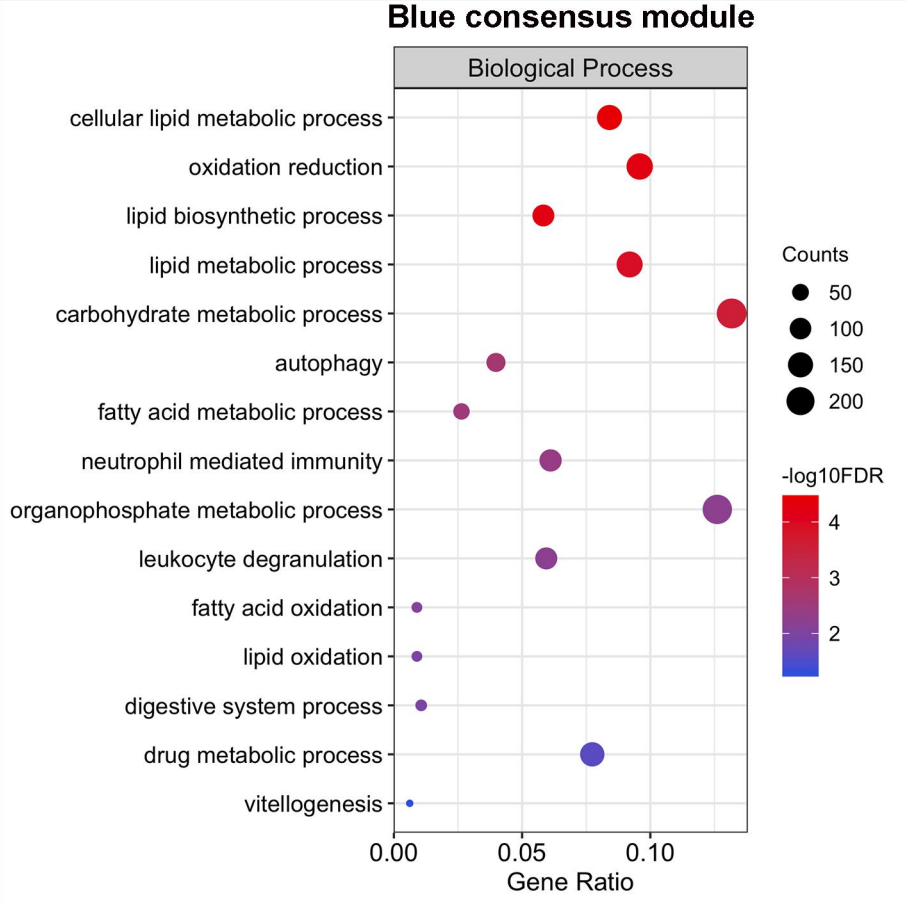


Figure 8 – Selected fifteen Gene Ontology enriched terms for protein-coding genes of the blue consensus module. Only GO terms from the Biological Process category are represented. The size of the circles is proportional to the number of genes (counts scale on the right) in each significantly enriched GO term, and the colors show the statistical significance of the enrichment, as indicated by the -log10 FDR values (color-coded scale at the right).

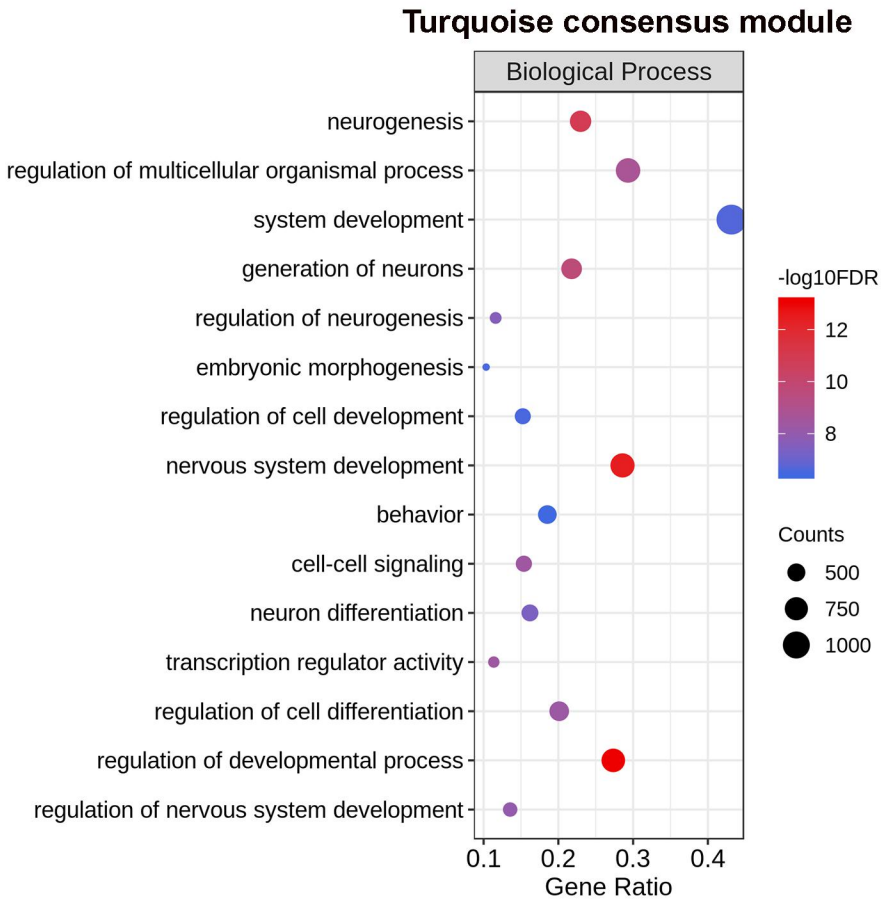


Figure 9 – Top fifteen most significantly enriched Gene Ontology terms from the Biological Process category for protein-coding genes of the turquoise consensus module. At left are the enriched GO term annotations. The size of the circles is proportional to the number of genes (counts scale on the right) in each significantly enriched GO term, and the colors show the statistical significance of the enrichment, as indicated by the -log10 FDR values (color-coded scale at the right).

4 CONCLUSÕES E PERSPECTIVAS

Diante dos resultados obtidos neste trabalho podemos concluir que:

1) Foi possível anotar um conjunto mais robusto e atualizado de milhares de lncRNAs expressos pelos parasitas *S. mansoni* e *S. japonicum*, 16 e 12 mil respectivamente.
2) LncRNAs de *S. mansoni* e *S. japonicum* possuem uma relação de sintenia mesmo quando não há conservação de sequência, sendo isso uma característica de possível manutenção de funcionalidade.
3) Em *S. mansoni*, lncRNAs apresentam expressão que é estágio, tecido e célula-específico; possuindo assim uma expressão dinâmica ao longo do ciclo de vida. Além disso, lncRNAs são genes *hub* de diversas redes de co-expressão que possuem genes codificadores de proteínas associados a replicação, reprodução e metabolismo.
4) Em *S. japonicum*, conseguimos identificar que lncRNAs tem dinâmica ao longo do processo de maturação sexual de machos e fêmeas, e está em co-expressão com genes codificadores de proteínas associados aos mais diversos processos biológicos, como desenvolvimento nervoso, desenvolvimento sexual, produção de ovos, metabolismo de lipídeos, entre outros.

As redes de co-expressão podem sugerir possíveis papeis para os lncRNAs, pelo princípio de *guilt-by-association*, quando existe correlação com proteínas de função conhecida. Além disso, a manutenção de sintenia destes lncRNAs entre *S. mansoni* e *S. japonicum* aponta para uma possível manutenção de funcionalidade.

Com base nessas informações, nosso grupo tem como perspectiva realizar estudos sobre funcionalidade desses lncRNAs, para que se possa entender melhor a biologia do parasita e possivelmente identificar potenciais alvos para terapias gênicas. Entre os estudos para caracterização do papel de lncRNAs, pode-se exemplificar o seu silenciamento usando-se RNAs dupla-fita que alvejem os lncRNAs candidatos, em parasitas mantidos em cultura nos diversos estágios de vida, seguido da verificação de possíveis efeitos fenotípicos, tais como diminuição da viabilidade dos parasitas, perda de pareamento entre machos e fêmeas ou diminuição da produção de ovos pelas fêmeas. A mudança no nível de expressão dos mRNAs dos genes codificadores de proteína co-expressos com um lncRNA candidato, em função do silenciamento deste lncRNA, é também um dado que poderá apontar para a possível relação entre esses genes. A eventual mudança em larga-escala no nível de expressão de mRNAs de genes codificadores de proteína, causada pelo silenciamento de um lncRNA candidato, poderá

apontar para um efeito regulador mais amplo deste lncRNA sobre o programa de ativação gênica do parasita; a identificação de alguma via metabólica enriquecida entre os genes codificadores de proteína afetados poderá sugerir o papel desse lncRNA sobre a referida via.

5 REFERÊNCIAS

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. ISSN 1091-6490 (Electronic)
0027-8424 (Linking). Disponível em: < https://www.ncbi.nlm.nih.gov/pubmed/25691743 >.

BLOKHIN, I. et al. Developments in lncRNA drug discovery: where are we heading? **Expert Opin Drug Discov,** v. 13, n. 9, p. 837-849, Sep 2018. ISSN 1746-045X (Electronic)
1746-0441 (Linking). Disponível em: < https://www.ncbi.nlm.nih.gov/pubmed/30078338 >.

BURKE, M. L. et al. Immunopathogenesis of human schistosomiasis. **Parasite Immunology,** v. 31, n. 4, p. 163-176, 2009. Disponível em: < https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-3024.2009.01098.x >.

CAO, H.; WAHLESTEDT, C.; KAPRANOV, P. Strategies to Annotate and Characterize Long Noncoding RNAs: Advantages and Pitfalls. **Trends in Genetics,** v. 34, n. 9, p. 704-721, 2018. ISSN 0168-9525. Disponível em: < https://doi.org/10.1016/j.tig.2018.06.002 >. Acesso em: 2018/08/24.

CDC. Centers for Disease Control and Prevention, Schistosomiasis Biology. 7/11/2012 2012. Disponível em: < https://www.cdc.gov/parasites/schistosomiasis/biology.html >. Acesso em: 20/07/2018.

_____. Centers for Disease Control and Prevention, Parasites - Schistosomiasis. 20/04/18 2018. Disponível em: < https://www.cdc.gov/parasites/schistosomiasis/ >. Acesso em: 20/07/2018.

CERASE, A. et al. Xist localization and function: new insights from multiple levels. **BMC Genome Biology,** v. 16, n. 1, p. 166, August 15 2015. ISSN 1474-760X. Disponível em: < https://doi.org/10.1186/s13059-015-0733-y >.

CHOI, S.-W.; KIM, H.-W.; NAM, J.-W. The small peptide world in long noncoding RNAs. **Briefings in Bioinformatics**, p. bby055-bby055, 2018. ISSN 1467-5463. Disponível em: < http://dx.doi.org/10.1093/bib/bby055 >.

COLLINS, J. J. et al. An Atlas for *Schistosoma mansoni* Organs and Life-Cycle Stages Using Cell Type-Specific Markers and Confocal Microscopy. **PLoS Neglected Tropical Diseases**, San Francisco, USA**,** v. 5, n. 3, p. e1009, 03/08
10/28/received
12/15/accepted 2011. ISSN 1935-2727
1935-2735. Disponível em: < http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3050934/ >.

CREDENDINO, S. C. et al. Tissue- and Cell Type-Specific Expression of the Long Noncoding RNA Klhl14-AS in Mouse **International Journal of Genomics,** v. 2017, p. 7, 2017. Disponível em: < https://doi.org/10.1155/2017/9769171 >.

CUPIT, P. M.; CUNNINGHAM, C. What is the mechanism of action of praziquantel and how might resistance strike? **FUTURE MEDICINAL CHEMISTRY,** v. 7, n. 6, p. 701-705, 2015. Disponível em: < https://www.future-science.com/doi/abs/10.4155/fmc.15.11 >.

ENGREITZ, J. M. et al. Local regulation of gene expression by lncRNA promoters, transcription and splicing. **Nature,** v. 539, p. 452, 10/26/online 2016. Disponível em: < http://dx.doi.org/10.1038/nature20149 >.

FANG, Y.; FULLWOOD, M. J. Roles, Functions, and Mechanisms of Long Non-coding RNAs in Cancer. **Genomics, Proteomics & Bioinformatics,** v. 14, n. 1, p. 42-54, 2016/02/01/ 2016. ISSN 1672-0229. Disponível em: < http://www.sciencedirect.com/science/article/pii/S1672022916000371 >.

FERNANDES, C. R. J. et al. Long Non-Coding RNAs in the Regulation of Gene Expression: Physiology and Disease. **Non-Coding RNA,** v. 5, n. 1, 2019. ISSN 2311-553X.

FITZPATRICK, J. M. et al. Anti-schistosomal Intervention Targets Identified by Lifecycle Transcriptomic Analyses. **PLoS Neglected Tropical Diseases**, San Francisco, USA**,** v. 3, n. 11, p. e543, ISSN 1935-2727 1935-2735. Disponível em: < http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2764848/ >.

GAWRONSKI, K. A. B.; KIM, J. Single cell transcriptomics of noncoding RNAs and their cell-specificity. **Wiley Interdisciplinary Reviews: RNA,** v. 8, n. 6, p. e1433, 2017/11/01 2017. ISSN 1757-7004. Disponível em: < https://doi.org/10.1002/wrna.1433 >. Acesso em: 2018/08/27.

GEYER, K. K. et al. The anti-fecundity effect of 5-azacytidine (5-AzaC) on *Schistosoma mansoni* is linked to dis-regulated transcription, translation and stem cell activities. **International Journal for Parasitology: Drugs and Drug Resistance,** v. 8, n. 2, p. 213-222, 2018/08/01/ 2018. ISSN 2211-3207. Disponível em: < http://www.sciencedirect.com/science/article/pii/S2211320718300150 >.

GEYER, K. K. et al. Cytosine methylation regulates oviposition in the pathogenic blood fluke *Schistosoma mansoni*. **Nature Communications,** v. 2, p. 424, 08/09/online 2011. Disponível em: < http://dx.doi.org/10.1038/ncomms1433 >.

GOLICZ, A. A.; BHALLA, P. L.; SINGH, M. B. lncRNAs in Plant and Animal Sexual Reproduction. **Trends in Plant Science,** v. 23, n. 3, p. 195-205, 2018/03/01/ 2018. ISSN 1360-1385. Disponível em: < http://www.sciencedirect.com/science/article/pii/S1360138517302856 >.

GOMES CASAVECHIA, M. T. et al. Systematic review and meta-analysis on *Schistosoma mansoni* infection prevalence, and associated risk factors in Brazil. **Parasitology,** v. 145, n. 8, p. 1000-1014, 2018. ISSN 0031-1820. Disponível em: < https://www.cambridge.org/core/article/systematic-review-and-metaanalysis-on-schistosoma-mansoni-infection-prevalence-and-associated-risk-factors-in-brazil/724D5B8D10BE5ECE9101FABD6C963BB9 >.

GRAY, D. J. et al. Diagnosis and management of schistosomiasis. **BMJ,** v. 342, 2011. Disponível em: < http://www.bmj.com/content/342/bmj.d2651.abstract >.

HARRIES, L. W. RNA Biology Provides New Therapeutic Targets for Human Disease. **Front Genet,** v. 10, n. 205, p. 205, 2019-March-08 2019. ISSN 1664-8021 (Print) 1664-8021 (Linking). Disponível em: < https://www.ncbi.nlm.nih.gov/pubmed/30906315 >.

HOFFMANN, K. F.; WYNN, T. A.; DUNNE, D. W. Cytokine-mediated host responses during schistosome infections; walking the fine line between immunological control and immunopathology. **Advances in parasitology,** v. 52, p. 265-307, 2002 2002. ISSN 0065-308X. Disponível em: < http://europepmc.org/abstract/MED/12521263 https://doi.org/10.1016/S0065-308X(02)52014-5 >.

JOHNSSON, P. et al. Evolutionary conservation of long noncoding RNAs; sequence, structure, function. **Biochimica et biophysica acta,** v. 1840, n. 3, p. 1063-1071, 10/27 2014. ISSN 0006-3002. Disponível em: < http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3909678/ >.

KIM, D. H. et al. Single cell transcriptome analysis reveals dynamic changes in lncRNA expression during reprogramming. **Cell stem cell,** v. 16, n. 1, p. 88-101, 2015. ISSN 1934-5909
1875-9777. Disponível em: < http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4291542/ >.

KITAGAWA, M. et al. Cell cycle regulation by long non-coding RNAs. **Cellular and Molecular Life Sciences,** v. 70, n. 24, p. 4785-4794, 2013/12/01 2013. ISSN 1420-9071. Disponível em: < https://doi.org/10.1007/s00018-013-1423-0 >.

KNOLL, M.; LODISH, H. F.; SUN, L. Long non-coding RNAs as regulators of the endocrine system. **Nature Reviews Endocrinology,** v. 11, p. 151, 01/06/online 2015. Disponível em: < https://doi.org/10.1038/nrendo.2014.229 >.

LEFEVER, S. et al. decodeRNA- predicting non-coding RNA functions using guilt-by-association. **Database : the journal of biological databases and curation,** v. 2017, p. bax042, 2017. ISSN 1758-0463. Disponível em: < https://www.ncbi.nlm.nih.gov/pubmed/29220434 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5502368/ >.

LEUCCI, E. et al. Melanoma addiction to the long non-coding RNA SAMMSON. **Nature,** v. 531, p. 518, 03/23/online 2016. Disponível em: < https://doi.org/10.1038/nature17161 >.

LIAO, Q. et al. Genome-wide identification and functional annotation of *Plasmodium falciparum* long noncoding RNAs from RNA-seq data. **Parasitology Research,** v. 113, n. 4, p. 1269-1281, 2014/04/01 2014. ISSN 1432-1955. Disponível em: < https://doi.org/10.1007/s00436-014-3765-4 >.

LIAO, Q. et al. Identification of long noncoding RNAs in *Schistosoma mansoni* and *Schistosoma japonicum*. **Experimental Parasitology,** v. 191, p. 82-87, 2018/08/01/ 2018. ISSN 0014-4894. Disponível em: < http://www.sciencedirect.com/science/article/pii/S0014489417305040 >.

LIAO, Q. et al. Identification of long noncoding RNAs in *Schistosoma mansoni* and *Schistosoma japonicum*. **Exp Parasitol,** v. 191, p. 82-87, Aug 2018. ISSN 1090-2449 (Electronic)
0014-4894 (Linking). Disponível em: < https://www.ncbi.nlm.nih.gov/pubmed/29981293 >.

LU, Z. et al. Schistosome sex matters: a deep view into gonad-specific and pairing-dependent transcriptomes reveals a complex gender interplay. **Scientific Reports,** v. 6, p. 31150, 08/08/online 2016. Disponível em: < http://dx.doi.org/10.1038/srep31150 >.

_____. A gene expression atlas of adult *Schistosoma mansoni* and their gonads. **Scientific Data,** v. 4, p. 170118, 08/22/online 2017. Disponível em: < http://dx.doi.org/10.1038/sdata.2017.118 >.

LUO, F. et al. An improved genome assembly of the fluke *Schistosoma japonicum*. **PLOS Neglected Tropical Diseases,** v. 13, n. 8, p. e0007612, 2019. Disponível em: < https://doi.org/10.1371/journal.pntd.0007612 >.

LUO, Y. et al. Single-Cell Transcriptome Analyses Reveal Signals to Activate Dormant Neural Stem Cells. **Cell,** v. 161, n. 5, p. 1175-1186, 2015/05/21/ 2015. ISSN 0092-8674. Disponível em: < http://www.sciencedirect.com/science/article/pii/S0092867415003748 >.

MATSUI, M.; COREY, D. R. Non-coding RNAs as drug targets. **Nat Rev Drug Discov,** v. 16, n. 3, p. 167-179, Mar 2017. ISSN 1474-1784 (Electronic)
1474-1776 (Linking). Disponível em: < https://www.ncbi.nlm.nih.gov/pubmed/27444227 >.

MENG, L. et al. Towards a therapy for Angelman syndrome by targeting a long non-coding RNA. **Nature,** v. 518, p. 409, 12/01/online 2014. Disponível em: < https://doi.org/10.1038/nature13975 >.

NAM, J.-W.; CHOI, S.-W.; YOU, B.-H. Incredible RNA: Dual Functions of Coding and Noncoding. **Molecules and Cells,** v. 39, n. 5, p. 367-374, ISSN 1016-8478
0219-1032. Disponível em: < http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4870183/ >.

OLIVEIRA, V. F. et al. Identification of 170 New Long Noncoding RNAs in *Schistosoma mansoni*. **BioMed Research International,** v. 2018, p. 9, 2018. Disponível em: < https://doi.org/10.1155/2018/1264697 >.

PAL, D.; RAO, M. R. S. Long Noncoding RNAs in Pluripotency of Stem Cells and Cell Fate Specification. In: RAO, M. R. S. (Ed.). **Long Non Coding RNA Biology**. Singapore: Springer Singapore, 2017. p.223-252. ISBN 978-981-10-5203-3.

PANG, K. C.; FRITH, M. C.; MATTICK, J. S. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. **Trends in Genetics,** v. 22, n. 1, p. 1-5, 2006/01/01/ 2006. ISSN 0168-9525. Disponível em: < http://www.sciencedirect.com/science/article/pii/S0168952505003227 >.

PICARD, M. A. L. et al. Sex-Biased Transcriptome of *Schistosoma mansoni*: Host-Parasite Interaction, Genetic Determinants and Epigenetic Regulators Are Associated with Sexual Differentiation. **PLoS Neglected Tropical Diseases**, San Francisco, CA USA**,** v. 10, n. 9, p.

e0004930, ISSN 1935-2727 1935-2735. Disponível em: < http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5038963/ >.

PINTACUDA, G.; YOUNG, A. N.; CERASE, A. Function by Structure: Spotlights on Xist Long Non-coding RNA. **Frontiers in Molecular Bioscience,** v. 4, n. 90, 2017-December-19 2017. ISSN 2296-889X. Disponível em: < https://www.frontiersin.org/article/10.3389/fmolb.2017.00090 >.

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. ISSN 1935-2735 (Electronic)
1935-2727 (Linking). Disponível em: < https://www.ncbi.nlm.nih.gov/pubmed/22253936 >.

ROSA, A.; BALLARINO, M. Long Noncoding RNA Regulation of Pluripotency. **Stem Cells International,** v. 2016, p. 1797692, 11/30
03/04/received
07/07/accepted 2016. ISSN 1687-966X
1687-9678. Disponível em: < http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4677244/ >.

SARROPOULOS, I. et al. Developmental dynamics of lncRNAs across mammalian organs and species. **Nature,** v. 571, n. 7766, p. 510-514, 2019/07/01 2019. ISSN 1476-4687. Disponível em: < https://doi.org/10.1038/s41586-019-1341-x >.

SIMONA, G. et al. Long Noncoding RNAs and Cardiac Disease. **Antioxidants & Redox Signaling,** v. 29, n. 9, p. 880-901, 2018. Disponível em: < https://www.liebertpub.com/doi/abs/10.1089/ars.2017.7126 >.

ST. LAURENT, G.; WAHLESTEDT, C.; KAPRANOV, P. The Landscape of long noncoding RNA classification. **Trends in Genetics,** v. 31, n. 5, p. 239-251, 2015. ISSN 0168-9525. Disponível em: < https://doi.org/10.1016/j.tig.2015.03.007 >. Acesso em: 2018/08/24.

TARASHANSKY, A. J. et al. Self-assembling Manifolds in Single-cell RNA Sequencing Data. **bioRxiv**, 2018. Disponível em: < http://biorxiv.org/content/early/2018/07/07/364166.abstract >.

UTZINGER, J. et al. From innovation to application: Social–ecological context, diagnostics, drugs and integrated control of schistosomiasis. **Acta Tropica,** v. 120, p. S121-S137, 2011/09/01/ 2011. ISSN 0001-706X. Disponível em: < http://www.sciencedirect.com/science/article/pii/S0001706X10002330 >.

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. **Scientific Reports,** v. 7, n. 1, p. 10508, 2017/09/05 2017. ISSN 2045-2322. Disponível em: < https://doi.org/10.1038/s41598-017-10853-6 >.

WANG, B. et al. Stem cell heterogeneity drives the parasitic life cycle of *Schistosoma mansoni*. **eLife,** v. 7, p. e35449, 07/10
01/26/received

06/08/accepted 2018. ISSN 2050-084X. Disponível em: < http://www.ncbi.nlm.nih.gov/pmc/articles/PMC6039179/ >.

WANG, J. et al. Dynamic transcriptomes identify biogenic amines and insect-like hormonal regulation for mediating reproduction in *Schistosoma japonicum*. **Nature Communications,** v. 8, p. 14693, 03/13/online 2017. Disponível em: < http://dx.doi.org/10.1038/ncomms14693 >.

WASHIETL, S.; KELLIS, M.; GARBER, M. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. **Genome Research,** v. 24, n. 4, p. 616-628, ISSN 1088-9051 1549-5469. Disponível em: < http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3975061/ >.

WHO, W. H. O. **Investing to overcome the global impact of neglected tropical diseases: third WHO report on neglected tropical diseases 2015**. Geneva, Switzerland: World Health Organization, 2015. ISBN 978 92 4 156486 1.

WILSON, R. A. The saga of schistosome migration and attrition. **Parasitology,** v. 136, n. 12, p. 1581-1592, 2009. ISSN 0031-1820. Disponível em: < https://www.cambridge.org/core/article/saga-of-schistosome-migration-and-attrition/70AC43F6399DC161E22443E0C2D9C545 >.

WU, W. et al. Tissue-specific Co-expression of Long Non-coding and Coding RNAs Associated with Breast Cancer. **Scientific Reports,** v. 6, p. 32731, 09/06/online 2016. Disponível em: < http://dx.doi.org/10.1038/srep32731 >.

ZIJIAN, Z. Long non-coding RNAs in Alzheimer's disease. **Current Topics in Medicinal Chemistry,** v. 16, n. 5, p. 511-519, 2016. ISSN 1568-0266/1873-4294. Disponível em: < http://www.eurekaselect.com/node/133966/article >.

ZONI, A. C.; CATALÁ, L.; AULT, S. K. Schistosomiasis Prevalence and Intensity of Infection in Latin America and the Caribbean Countries, 1942-2014: A Systematic Review in the Context of a Regional Elimination Goal. **PLoS Neglected Tropical Diseases**, San Francisco, CA USA**,** v. 10, n. 3, p. e0004493, ISSN 1935-2727 1935-2735. Disponível em: < http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4805296/ >.

ANEXOS

Anexo A – Artigo publicado na *Frontiers in Genetics* que foi apresentado na seção de manuscritos desta dissertação, Capítulo I.

frontiers in Genetics

ORIGINAL RESEARCH
published: 12 September 2019
doi: 10.3389/fgene.2019.00823

Weighted Gene Co-Expression Analyses Point to Long Non-Coding RNA Hub Genes at Different *Schistosoma mansoni* Life-Cycle Stages

*Lucas F. Maciel[1,2], David A. Morales-Vicente[1,3], Gilbert O. Silveira[1,3], Raphael O. Ribeiro[1,3], Giovanna G. O. Olberg[1], David S. Pires[1], Murilo S. Amaral[1] and Sergio Verjovski-Almeida[1,3]**

[1] Laboratório de Expressão Gênica em Eucariotos, Instituto Butantan, São Paulo, Brazil, [2] Programa Interunidades em Bioinformática, Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, Brazil, [3] Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, São Paulo, Brazil

OPEN ACCESS

Edited by:
Gabriel Rinaldi,
Wellcome Trust Sanger Institute (WT), United Kingdom

Reviewed by:
Thiago Motta Venancio,
Universidade Estadual do Norte Fluminense Darcy Ribeiro, Brazil
Zhigang Lu,
Wellcome Trust Sanger Institute (WT), United Kingdom

***Correspondence:**
Sergio Verjovski-Almeida
verjo@iq.usp.br

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Genetics

Received: 04 April 2019
Accepted: 09 August 2019
Published: 12 September 2019

Citation:
Maciel LF, Morales-Vicente DA, Silveira GO, Ribeiro RO, Olberg GGO, Pires DS, Amaral MS and Verjovski-Almeida S (2019) Weighted Gene Co-Expression Analyses Point to Long Non-Coding RNA Hub Genes at Different Schistosoma mansoni Life-Cycle Stages. Front. Genet. 10:823. doi: 10.3389/fgene.2019.00823

Long non-coding RNAs (lncRNAs) (>200 nt) are expressed at levels lower than those of the protein-coding mRNAs, and in all eukaryotic model species where they have been characterized, they are transcribed from thousands of different genomic *loci*. In humans, some four dozen lncRNAs have been studied in detail, and they have been shown to play important roles in transcriptional regulation, acting in conjunction with transcription factors and epigenetic marks to modulate the tissue-type specific programs of transcriptional gene activation and repression. In *Schistosoma mansoni*, around 10,000 lncRNAs have been identified in previous works. However, the limited number of RNA-sequencing (RNA-seq) libraries that had been previously assessed, together with the use of old and incomplete versions of the *S. mansoni* genome and protein-coding transcriptome annotations, have hampered the identification of all lncRNAs expressed in the parasite. Here we have used 633 publicly available *S. mansoni* RNA-seq libraries from whole worms at different stages (n = 121), from isolated tissues (n = 24), from cell-populations (n = 81), and from single-cells (n = 407). We have assembled a set of 16,583 lncRNA transcripts originated from 10,024 genes, of which 11,022 are novel *S. mansoni* lncRNA transcripts, whereas the remaining 5,561 transcripts comprise 120 lncRNAs that are identical to and 5,441 lncRNAs that have gene overlap with *S. mansoni* lncRNAs already reported in previous works. Most importantly, our more stringent assembly and filtering pipeline has identified and removed a set of 4,293 lncRNA transcripts from previous publications that were in fact derived from partially processed mRNAs with intron retention. We have used weighted gene co-expression network analyses and identified 15 different gene co-expression modules. Each parasite life-cycle stage has at least one highly correlated gene co-expression module, and each module is comprised of hundreds to thousands lncRNAs and mRNAs having correlated co-expression patterns at different stages. Inspection of the top most

highly connected genes within the modules' networks has shown that different lncRNAs are hub genes at different life-cycle stages, being among the most promising candidate lncRNAs to be further explored for functional characterization.

Keywords: parasitology, RNA-seq, single-cell sequencing data, *Schistosoma mansoni*, long non-coding RNAs, weighted genes co-expression network analysis

INTRODUCTION

Schistosomiasis is a neglected tropical disease, caused by flatworms from the genus *Schistosoma*, with estimates of more than 250 million infected people worldwide and responsible for 200 thousand deaths annually at the Sub-Saharan Africa (Who, 2015). *Schistosoma mansoni*, prevalent in Africa and Latin America, is one of the three main species related to human infections (Cdc, 2018). In America, it is estimated that 1 to 3 million people are infected by *S. mansoni* and over 25 million live in risk areas, being Brazil and Venezuela the most affected (Zoni et al., 2016). The prevalence of this disease is correlated to social–economic and environmental factors (Gomes Casavechia et al., 2018).

This parasite has a very complex life-cycle comprised of several developmental stages, with a freshwater snail intermediate-host and a final mammalian host (Basch, 1976). Recently, it has been shown that epigenetic changes are required for life-cycle progression (Roquis et al., 2018). However, little is known about the genes and molecules that drive this process through the life-cycle stages of *S. mansoni*. A better understanding of the gene expression regulation mechanisms and of their components may lead to new therapeutic targets (Batugedara et al., 2017), and one key element could be the long non-coding RNAs (lncRNAs) (Blokhin et al., 2018).

LncRNAs are defined as transcripts longer than 200 nucleotides, without apparent protein-coding potential (Cao et al., 2018). The term "apparent" is included because it is already known that some lncRNAs actually have dual function roles, being functional both as lncRNAs and through peptides shorter than 100 amino acids that they encode (Nam et al., 2016; Choi et al., 2018). In mammalians, lncRNAs regulate gene expression through different mechanisms (Bhat et al., 2016), including mediating epigenetic modifications (Hanly et al., 2018), and were shown to be important in vital processes, such as cell cycle regulation (Kitagawa et al., 2013), pluripotency maintenance (Rosa and Ballarino, 2016), and reproduction (Golicz et al., 2018).

In *S. mansoni*, the expression of lncRNAs at different life-cycle stages was first detected by our group in 2011 using microarrays (Oliveira et al., 2011). Subsequently, large-scale identification of *S. mansoni* lncRNAs has been reported in three studies from our group and from others that analyzed high-throughput RNA-sequencing (RNA-seq) data (Vasconcelos et al., 2017; Liao et al., 2018; Oliveira et al., 2018), but each of them has used a the limited number of data sets (from 4 to 88 RNA-seq libraries). Because each work used different mapping tools and parameters (Vasconcelos et al., 2017; Liao et al., 2018; Oliveira et al., 2018), and given that Liao et al. (2018) did not compare their lncRNAs with the previously published ones, part of the lncRNAs are redundant among the three reports. In addition, the lncRNAs were annotated against the old version 5.2 of the genome and protein-coding transcriptome (Protasio et al., 2012); as a result, a set of transcripts that were previously annotated as lncRNAs (Vasconcelos et al., 2017; Liao et al., 2018; Oliveira et al., 2018), seem now to represent partially processed pre-mRNAs arising from novel protein-coding genes annotated in the new version 7.1 of the transcriptome (https://parasite.wormbase.org/Schistosoma_mansoni_prjea36577/); these transcripts were previously annotated as having no coding potential due to intron retention, as exemplified in **Supplementary Figure S1**. Besides, these three works used expression data from whole parasites, while it is known from other species that lncRNAs have tissue- and cell-specific expression (Wu et al., 2016; Credendino et al., 2017).

The aim of the present work is to identify and annotate a robust and more complete set of lncRNAs that agrees with the most updated transcriptome annotation, and to analyze RNA-seq data sets still non-annotated for the presence of lncRNAs—e.g., gonads (Lu et al., 2016) and single-cell (Tarashansky et al., 2018; Wang et al., 2018) RNA-seq libraries. The goal is to provide a foundation that will enable future studies on the role of lncRNAs in *S. mansoni* biology, which could eventually identify potential new therapeutic targets.

MATERIALS AND METHODS

Transcripts Reconstruction

To identify new lncRNAs, 633 publicly available RNA-seq libraries from whole worms at different stages (miracidia, n = 1; sporocysts, n = 1; cercariae, n = 8; schistosomula, n = 11; juveniles, n = 9; adult males, n = 34; adult females, n = 37; and mixed adults, n = 20), from tissues (testes, n = 6; ovaries, n = 5; posterior somatic tissues, n = 3; heads, n = 5; and tails, n = 5), from cell populations (n = 81) and from single cells (from juveniles, n = 370 and mother sporocysts stem cells, n = 37) were downloaded from the SRA and ENA databases (**Supplementary Table S1**). The only whole-worm stage that was not included was eggs, because there is a single RNA-seq library available in the public domain (Anderson et al., 2016), which has only 252,000 egg reads, an amount that is fourfold lower than the minimum number of reads per library in the other whole-worm libraries that we used (namely 1 million good quality reads), being a too-low coverage for an unbiased detection of stage- or tissue-specific lncRNAs in complex

organisms (Sims et al., 2014). The new versions of the genome (v 7) and transcriptome (v 7.1), which were used as reference in this study, were downloaded from the WormBase ParaSite resource (Howe et al., 2017) at https://parasite.wormbase.org/Schistosoma_mansoni_prjea36577/.

Quality control was done with fastp v 0.19.4 (Chen et al., 2018) (default parameters), removing adapters and low-quality reads. The reads in each library were then mapped against the genome with STAR v 2.6.1c in a two-pass mode, with parameters indicated by STAR's manual as the best ones to identify new splicing sites and transcripts (Dobin et al., 2013). RSeQC v 2.6.5 (Wang et al., 2012) was used to identify RNA-Seq library strandedness to be used in transcripts reconstruction and expression levels quantification. For each library, multi mapped reads were removed with Samtools v 1.3 (Li et al., 2009) and uniquely mapped reads were used for transcript reconstruction with Scallop v 10.2 (–min_mapping_quality 255 -min_splice_boundary_hits 2) (Shao and Kingsford, 2017). A new splicing site should be confirmed at least by two reads to be considered. A consensus transcriptome from all libraries was built using TACO v 0.7.3 (–filter-min-length 200 -isoform-frac 0.05), an algorithm that reconstructs the consensus transcriptome from a collection of individual assemblies (Niknafs et al., 2016). As described by Niknafs et al. (2016), TACO employs change point detection to break apart complex loci and correctly delineate transcript start and end sites and a dynamic programming approach to assemble transcripts from a network of splicing patterns (Niknafs et al., 2016).

LncRNAs Classification

In the consensus transcriptome, transcripts shorter than 200 nt, monoexonic or with exon-exon overlap with protein-coding genes from the same genomic strand were removed from the set. The coding potential of the remaining transcripts was evaluated by means of the FEELnc tool v 0.1.1 (Wucher et al., 2017) with shuffle mode, which uses a random forest machine-learning algorithm and classifies these transcripts into lncRNAs or protein-coding genes, and also by CPC2 v 0.1 (Yang et al., 2017), which classifies through a support vector machine model using four intrinsic features. Only transcripts classified as lncRNAs by both tools were kept. ORFfinder v 0.4.3 (https://www.ncbi.nlm.nih.gov/orffinder/) was used to extract the putative longest open reading frames (ORFs); these putative peptides were then submitted to orthology-based annotation with eggNOG-mapper webtool (HMMER mapping mode) (Huerta-Cepas et al., 2017). Transcripts with no hits against the eukaryote eggNOG database were then considered as lncRNAs. If any transcript isoform was classified as a protein-coding mRNA at any step, all transcripts mapping to the same genomic *locus* were removed to avoid eventual pre-mRNAs. After this final step, a lncRNAs GTF file was created.

Histone Marks

To identify histone H3 lysine 4 trimethylation (H3K4me3) and H3 lysine 27 trimethylation (H3K27me3) marks near the transcription start site (TSS) of lncRNAs, we used 12 libraries of Chromatin Immunoprecipitation Sequencing (ChIP-Seq) data generated by Roquis et al. (2018) for cercariae, schistosomula, and adults (**Supplementary Table S1**), which had more than 90% overall mapping rate. The reads were downloaded from the SRA database and mapped against the genome v 7 with Bowtie2 v 2.3.4.3 (Langmead and Salzberg, 2012) (parameters end-to-end, -sensitive, -gbar 4). Because there are no input data sets publicly available in the SRA database for the Roquis et al. (2018) paper, we were not able to exactly reproduce the pipeline that was described in the Methods section of that paper, which used the input as a reference for peak calling. Instead, we used HOMER v 4.10 (Heinz et al., 2010) for removing multi-mapped and duplicated reads and for significant peak calling as described by Anderson et al. (2016), an approach also used by Vasconcelos et al. (2017) in the first large-scale annotation of lncRNAs in *S. mansoni*. The number of reads in the peak should be at least fourfold higher than in the peaks of the surrounding 10-kb area and the Poisson p-value threshold cutoff was 0.0001. The lncRNAs with significant histone mark peaks within 1-kb distance upstream and downstream from their TSS were annotated. The lncRNAs with overlapping marks are shown with an intersection diagram that was plotted using the UpSetR tool v 1.3.3 (Lex et al., 2014). The Venn diagram tool at http://bioinformatics.psb.ugent.be/beg/tools/venn-diagrams was used for generating the lists of lncRNA genes belonging to each intersection set.

Co-Expression Networks

The lncRNAs GTF file was then added to the *S. mansoni* public protein-coding transcriptome version 7.1 GTF file, and the resulting protein-coding + lncRNAs GTF was used as the reference together with the genome sequence v.7 for mapping the reads of each RNA-seq library under study, again using the STAR tool, now in the one pass mode, followed by gene expression quantification with RSEM v 1.3 (Li and Dewey, 2011). Weighted gene co-expression network analyses v 1.68 (WGCNA) (Langfelder and Horvath, 2008) were then performed to identify modules related to the life-cycle stages and tissues of the organism. For this purpose, only libraries from whole worms or from tissues with more than 50% of the reads uniquely mapped were used. To reduce noise, only transcripts with expression greater than 1 transcript per million (TPM) in at least half of the libraries in one or more stages/tissues were considered. Expression levels were measured in log space with a pseudocount of 1 (log2 (TPM+1)), and we set the transcript expression to zero when log2 (TPM+1) <1. For the construction of the adjacency matrix, the power adjacency function for signed networks was applied with the soft-thresholding beta parameter equal to 14, which resulted in a scale-free topology model fit index ($R^2 = 0.935$). The adjacency matrix was then converted to the Topological Overlap Matrix (TOM) and the dissimilarity TOM (1 − TOM) was calculated (Langfelder and Horvath, 2008).

Correlation between the modules and the stages was calculated based on the Pearson correlation coefficient between

the expression levels of the transcripts belonging to each module along the stages, as suggested in the WGCNA tutorial (https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/). As miracidia and sporocysts have only one library each, are closely related stages of development, and were clustered together as an outgroup based on their overall expression patterns (as shown in the Results), we decided to consider both stages together as one group (miracidia/sporocysts) to calculate the correlation and p-values between modules and stages.

The Gene Trait Significance (GS) was calculated based on the correlation of an individual transcript and the trait, which in our case was always the stage of higher absolute Pearson correlation coefficient with the module where the transcript belongs. For example, for a transcript that belongs to the red module (most highly correlated with testes, see Results), the correlation was calculated between the expression of the transcript in the testes libraries and the expression of the transcript in all other non-testes libraries.

Gene Ontology (GO) Enrichment

Protein-coding genes were submitted to eggNOG-mapper (Huerta-Cepas et al., 2017) for annotation of GO terms. Based on this annotation (available at **Supplementary Table S2**), we performed GO enrichment analyses with BINGO (Maere et al., 2005). For each module, we used a hypergeometric test, the whole annotation as reference set, and FDR ≤ 0.05 was used as the significance threshold.

Single-Cell Analyses

The expression levels were quantified in single-cell RNA-seq libraries from juveniles' stem cells (Tarashansky et al., 2018) and mother sporocysts stem cells (Wang et al., 2018) by RSEM. We used Scater v 1.10.1 (Mccarthy et al., 2017) to normalize and identify high-quality single-cell RNA-Seq libraries, i.e., those that have at least 100,000 total counts and at least 1,000 different expressed transcripts, as recommended by Mccarthy et al. (2017); all libraries were classified as high quality.

Next, we used the R package Single-Cell Consensus Clustering (SC3) tool v 1.10.1 (Kiselev et al., 2017), which performs an unsupervised clustering of scRNA-seq data. Based on the clusters identified, we used the plot SC3 markers function to find marker genes based on the mean cluster expression values. These markers are highly expressed in only one of the clusters and indicate the specific expression at the cell level. As described by Kiselev et al. (2017), the area under the receiver operating characteristic (ROC) curve is used to quantify the accuracy of the prediction. A p-value is assigned to each gene by using the Wilcoxon signed rank test. Genes with the area under the ROC curve (AUROC) > 0.85 and with p-value < 0.01 are defined as marker genes.

Parasite Materials

All parasite materials were from a BH isolate of *S. mansoni* maintained by passage through golden hamster (*Mesocricetus auratus*) and *Biomphalaria glabrata* snails. Eggs were purified from livers of hamsters previously infected with *S. mansoni*, according to Dalton et al. (1997). After purification, eggs were added to 10 ml of distilled water and exposed to a bright light. Supernatant containing hatched miracidia was removed every 30 min for 2 h and replaced by fresh water. The supernatants containing the miracidia were pooled and chilled on ice, and miracidia were then recovered by centrifugation at 15,000*g* for 20 s (Dalton et al., 1997). Supernatant was discarded and miracidia stored in RNAlater (Ambion) until RNA extraction.

Cercariae were collected from snails infected with 10 miracidia each. Thirty-five days after infection, the snails were placed in the dark in water and then illuminated for 2 h to induce shedding. The emerging cercariae were collected by centrifugation, washed with PBS once, and then stored in RNAlater (Ambion) until RNA extraction.

Schistosomula were obtained by mechanical transformation of cercariae and separation of their bodies as previously described (Basch, 1981), with some modifications. Briefly, cercariae were collected as described above and then suspended in 15 ml of M169 medium (Vitrocell, cat number 00464) containing penicillin/streptomycin, amphotericin (Vitrocell, cat number 00148). Mechanical transformation was performed by passing the cercariae 10 times through a 23G needle. To separate schistosomula from the tails, the tail-rich supernatant was decanted and the sedimented bodies resuspended in a further 7 ml of M169 medium. The procedure was repeated until less than 1% of the tails remained. The newly transformed schistosomula were maintained for 24 h in M169 medium (Vitrocell, cat number 00464) supplemented with penicillin/streptomycin, amphotericin, gentamicin (Vitrocell, cat number 00148), 2% fetal bovine serum, 1 μM serotonin, 0.5 μM hypoxanthine, 1 μM hydrocortisone, and 0.2 μM triiodothyronine at 37°C and 5% CO_2. Schistosomula cultivated for 24 h were collected, washed three times with PBS and stored in RNAlater (Ambion) until RNA extraction.

Adult *S. mansoni* worms were recovered by perfusion of golden hamsters that had been infected with 250 cercariae, 7 weeks previously. Approximately 200 *S. mansoni* (BH strain) adult worm pairs were freshly obtained through the periportal perfusion of hamster, as previously described (Anderson et al., 2016; Vasconcelos et al., 2017). After perfusion, the adult worm pairs were kept for 3 h at 37°C and 5% CO_2 in Advanced RPMI Medium 1640 (Gibco, 12633-012) supplemented with 10% fetal bovine serum, 12 mM HEPES (4-(2-hydroxyethyl) piperazine-1-ethanesulfonic acid) pH 7.4, and 1% penicillin/streptomycin, amphotericin (Vitrocell, cat number 00148). After 3 h of incubation, the adult worm pairs were collected, washed three times with PBS, and stored in RNAlater (Ambion) until RNA extraction. Before the extraction of RNA from males or females, adult worm pairs were manually separated in RNAlater (Ambion) using tweezers.

RNA Extraction, Quantification, and Quality Assessment

Total RNA from eggs (E), miracidia (Mi), cercariae (C), and schistosomula (S) was extracted according to

Vasconcelos et al. (2017). Briefly, 100,000 eggs, 15,000 miracidia, 25,000 cercariae, or 25,000 schistosomula were ground with glass beads in liquid nitrogen for 5 min. Then, the Qiagen RNeasy Micro Kit (Cat number 74004) was used for RNA extraction and purification according to the manufacturer's instructions, except for the DNase I treatment, the amount of DNase I was doubled and the time of treatment was increased to 45 min.

Male (M) or female (F) adult worms were first disrupted in Qiagen RLT buffer using glass potters and pestles. RNA from males or females was then extracted and purified using the Qiagen RNeasy Mini Kit (Cat number 74104), according to the manufacturer's instructions, except for the DNase I treatment, which was the same used for egg, miracidia, cercariae, and schistosomula RNA extraction.

All the RNA samples were quantified using the Qubit RNA HS Assay Kit (Q32852, Thermo Fisher Scientific), and the integrity of RNAs was verified using the Agilent RNA 6000 Pico Kit (5067-1513 Agilent Technologies) in a 2100 Bioanalyzer Instrument (Agilent Technologies). Four biological replicates were assessed for each life cycle stage, except for schistosomula, for which three biological replicates were assessed.

Reverse Transcription and Quantitative PCR (qPCR) Assays

The reverse transcription (RT) reaction was performed with 200 ng of each total RNA sample using the SuperScript IV First-Strand Synthesis System (18091050; Life Technologies) and random hexamer primers in a 20-μL final volume. The obtained complementary DNAs (cDNAs) were diluted four times in DEPC water, and quantitative PCR was performed using 2.5 μL of each diluted cDNA in a total volume of 10 μL containing 1X LightCycler 480 SYBR Green I Master Mix (04707516001, Roche Diagnostics) and 800 nM of each primer in a LightCycler 480 System (Roche Diagnostics). Primers for selected transcripts (**Supplementary Table S3**) were designed using the Primer 3 tool (http://biotools.umassmed.edu/bioapps/primer3_www.cgi), and each real-time qPCR was run in two technical replicates. The results were analyzed by comparative Ct method (Livak and Schmittgen, 2001). Real-time data were normalized in relation to the level of expression of Smp_090920 and Smp_062630 reference genes.

RESULTS

LncRNAs Identification and Annotation

Using 633 publicly available *S. mansoni* RNA-seq libraries from whole worms at different stages, from isolated tissues, from cell-populations, and from single-cells (see Methods), our pipeline assembled a consensus transcriptome comprised of 78,817 transcripts, of which 7,954 were classified as intergenic lncRNAs (lincRNAs), 7,438 as antisense lncRNAs, and 1,191 as sense lncRNAs, totalizing 16,583 lncRNA transcripts originated from 10,024 genes (on average, 1.65 lncRNA isoforms per lncRNA gene); the summary of all six filtering steps in the pipeline is presented in **Table 1**. With the FEELnc lncRNA classification tool (**Table 1, step 5**), the most important feature for transcripts classification was the ORF coverage (**Supplementary Figure S2A**), i.e., the fraction of the total length of the transcript that is occupied by the longest predicted ORF. In the FEELnc model training process, an optimal coding probability cutoff (0.348) was identified, which resulted in 0.962 sensitivity and specificity of mRNA classification (**Supplementary Figure S2B**). Analogous information is not provided in the output of the CPC2 classification tool (**Table 1, step 5**). Only the lncRNAs classified as such by both prediction tools were retained in the final set (**Table 1**).

From the total set of 16,583 lncRNAs obtained here, 11,022 are novel *S. mansoni* lncRNAs, whereas the remaining 5,561 transcripts comprise 120 lncRNAs that are identical to previously published ones, and 5,441 lncRNAs that have gene overlap with *S. mansoni* lncRNAs already reported in previous works (Vasconcelos et al., 2017; Liao et al., 2018; Oliveira et al., 2018) (**Supplementary Table S4**). In particular, among the 7,029 lincRNAs previously published ones reported by our group (Vasconcelos et al., 2017), a total of 4,368 transcripts have partial or complete sequence overlap with the lncRNAs obtained here, whereas the remaining 2,661 (37.8%) transcripts previously annotated by Vasconcelos et al. (2017) are no longer in the present updated *S. mansoni* lncRNAs data set.

Among the transcripts in the public data set that were previously classified as lncRNAs (Vasconcelos et al., 2017; Liao et al., 2018; Oliveira et al., 2018) and are now excluded, a total of 4,293 were reconstructed in our assembly; however, they were removed from our set of lncRNAs because they were partially processed pre-mRNA transcripts that have exon-exon overlap with new protein-coding genes of version 7.1. The remaining transcripts previously classified as lncRNAs were reconstructed here but were removed by the more stringent, presently used filtering steps. We have created a track on the *S. mansoni* UCSC-like genome browser (http://schistosoma.usp.br/), where the set

TABLE 1 | Summary of transcripts removed at each filtering step and the final set of *S. mansoni* lncRNAs.

Pipeline step number	Removed transcripts	Remaining transcripts (Genes)
1. Total assembled transcripts		78,817 (42,337)
2. Remove short transcripts (<200 nt)	11	78,806
3. Remove monoexonic transcripts	27,255	51,551
4. Remove transcripts that overlap exon-exon with known Sm protein-coding genes	31,183	20,368
5. Remove transcripts with coding potential (FEELnc and/or CPC2 tools)	3,618	16,750
6. Remove transcripts with hits on eggNOG-Mapper	167	16,583
7. Total lncRNAs identified		16,583 (10,024)
Long intergenic non-coding RNAs		7,954
Antisense long non-coding RNAs		7,438
Sense long non-coding RNAs		1,191

of 16,583 lncRNAs obtained here can be visualized and the GTF and BED files can be downloaded. In **Figure 1**, we show a selected protein-coding desert genomic *locus* on chromosome 2 covering 245 kilobases, which harbors only three protein-coding genes and where we identified seven lincRNAs, two sense lncRNAs, and one antisense lncRNA that were not previously described.

To identify the contribution from each type of RNA-Seq library to the final lncRNAs set, we used the TACO transcriptome assembler to obtain the transcriptomes of the four following groups: whole organisms, tissues, cell populations, and single cells. The result is presented in **Figure 2** and shows that each type of sample contributed with at least 1,000 unique lncRNAs, detected only in that group. It is worthy to mention that around 4% of the 16,583 lncRNAs are lost when the four transcriptomes are reconstructed separately.

Almost all lncRNAs encode short canonical ORFs within their sequences, however, as described by Verheggen et al. (2017), one can evaluate if these ORFs are originated only by random nucleotide progression by comparing the relative sizes of ORFs using the reverse-complement of the sequence as a control. As presented in **Figure 3**, it is very clear that the size distribution of *bona fide S. mansoni* mRNA ORFs (sense) from the annotated v 7.1 transcriptome is greatly shifted toward longer sizes, compared with the size distribution of random ORFs found in their reverse-complement sequences. It is also possible to observe that the size distribution of ORFs found both within the lncRNAs (sense) and within their reverse-complement sequences is very similar and is also similar to the size distribution of random ORFs in the reverse-complement sequence of mRNAs.

Histone Marks at the TSS of LncRNAs as Evidence of Regulation

As reported earlier, cross-matching of the lncRNAs genomic coordinates with the genomic coordinates of different publicly available histone mark profiles, obtained by ChIP-Seq at different life-cycle stages, adds another layer of functionality evidence for this class of RNAs (Vasconcelos et al., 2017; Cao et al., 2018). We used the data for two different histone marks obtained by Roquis et al. (2018) in cercariae, schistosomula, and adult parasites, namely, H3K4me3 that is generally associated with active transcription, and H3K27me3 associated to transcription repression (Barski et al., 2007). First, we analyzed the histone mark profiles of H3K4me3 and H3K27me3 around the TSS of protein-coding genes through the stages, and they were very similar to the ones presented by Roquis et al. (2018) (**Supplementary Figure S3**). **Figure 4** shows that these marks are also present around the TSS of *S. mansoni* lncRNAs at the three different life-cycle stages; a comparison with **Supplementary Figure S3** shows that these marks are less abundant in lncRNAs than that in the protein-coding genes loci and more spread away of the lncRNAs TSSs when compared with protein-coding genes. This profile is

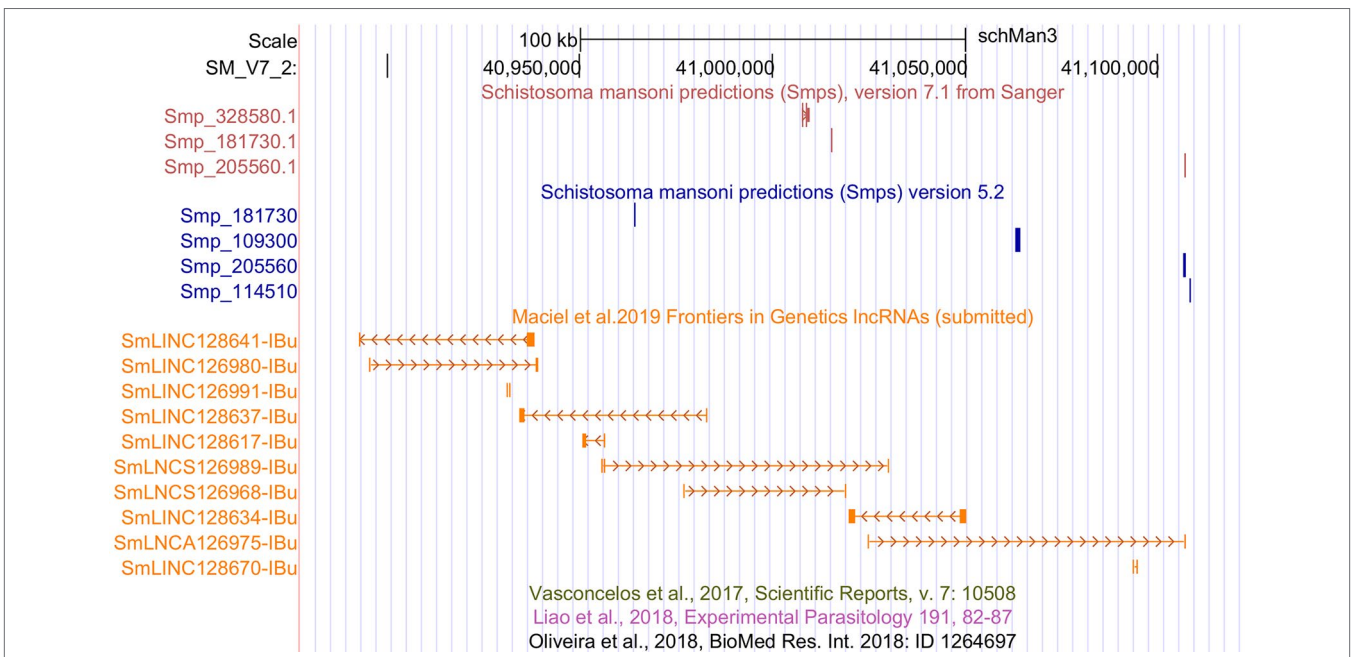


FIGURE 1 | Novel *S. mansoni* lncRNAs discovered in a protein-coding desert *locus*. Snapshot of a *S. mansoni* genome browser image, showing a region spanning 245 kb on chromosome 2 with coordinates SM_V7_2:40,877,676-41,122,371 (top black row). The red track (top) shows the three protein-coding genes from transcriptome version 7.1, whereas the blue track (middle) represents the protein-coding genes from version 5.2. The orange track (lower track) shows seven intergenic lncRNAs (SmLINCnnnnnn-IBu), two sense lncRNAs (SmLNCSnnnnnn-IBu), and one antisense lncRNA (SmLNCAnnnnnn-IBu) that were not annotated by the previously published works on lncRNAs, of which there are three empty tracks at the bottom, namely, Vasconcelos et al. (2017), Liao et al. (2018), and Oliveira et al. (2018).

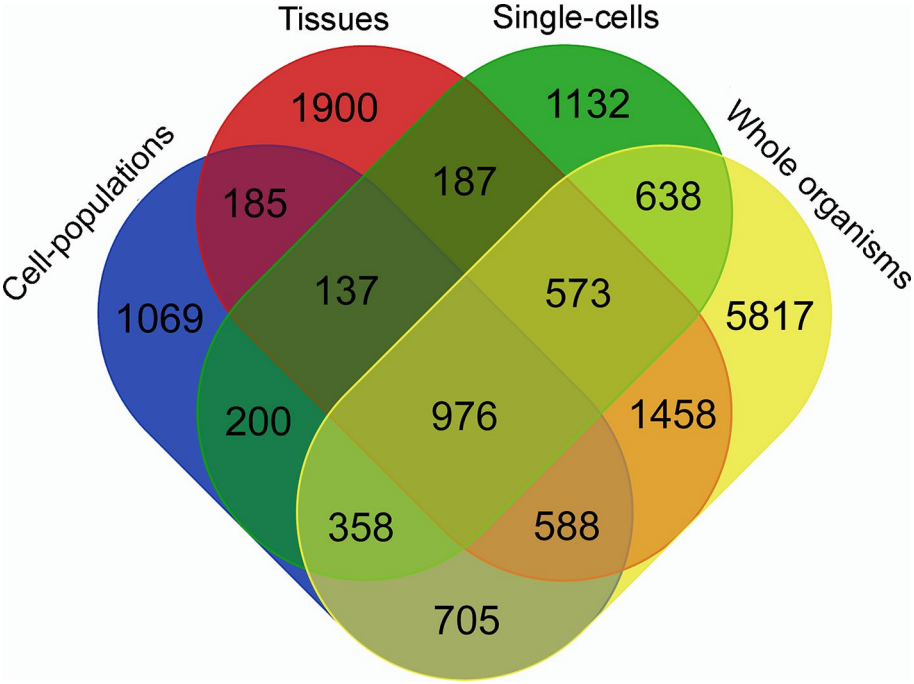


FIGURE 2 | Venn diagram representing the specific contribution from each type of RNA-Seq library to the *S. mansoni* lncRNAs set. TACO assembler was run separately for the RNA-Seq data from samples of four groups: whole organisms (yellow), tissues (red), cell-populations (blue) and single-cells (green), and each value indicates the number of transcripts that were reconstructed specifically with samples from groups indicated in each intersection.

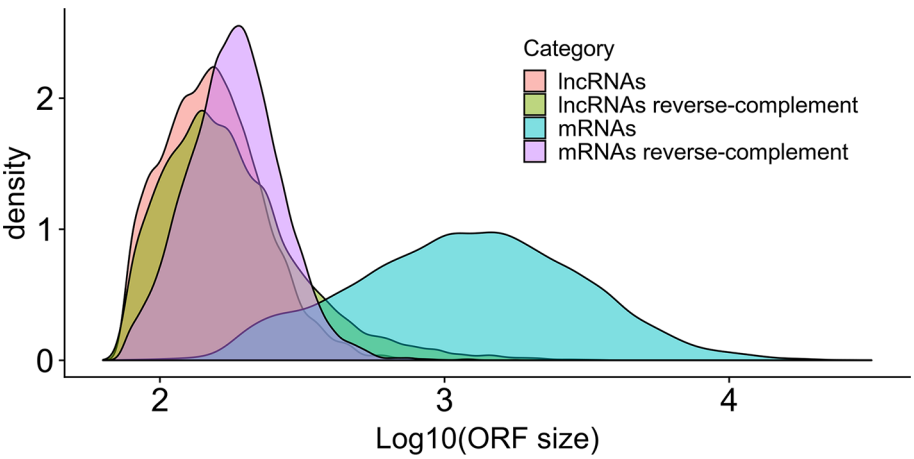


FIGURE 3 | Size distribution in *S. mansoni* of the longest canonical ORFs in lncRNAs and in mRNAs. The graph shows the density (y-axis) of the different sizes for the longest detected ORFs (in nucleotides, x-axis) of all lncRNAs (pink), of all mRNAs (blue) and of their reverse-complement sequences as controls (green and purple, respectively).

similar to that observed by Sati et al. (2012) when comparing histone marks around the TSS of human protein-coding genes and lncRNAs. A total of 8,599 lncRNA transcripts have at least one histone modification mark within 1 kb from their TSS (**Supplementary Table S5**), being 3,659 lincRNAs, 4,188 antisense lncRNAs, and 752 sense lncRNAs. A comparison of the lists of lncRNAs having a given histone mark at their TSS at either of the three different life-cycle stages (**Figure 5**) shows that the most abundant mark is the transcriptional repressive mark, H3K27me3. This mark is present at the TSS of different sets of lncRNAs at each of the three stages, with abundancies ranging from 1,334 lncRNAs with the H3K27me3 mark exclusively in schistosomula to 1,147 lncRNAs with the mark exclusively in adults and 1,024 lncRNAs with the mark exclusively in cercariae (**Figure 5, red**). In addition, the transcriptional activating mark H3K4me3 is present at the

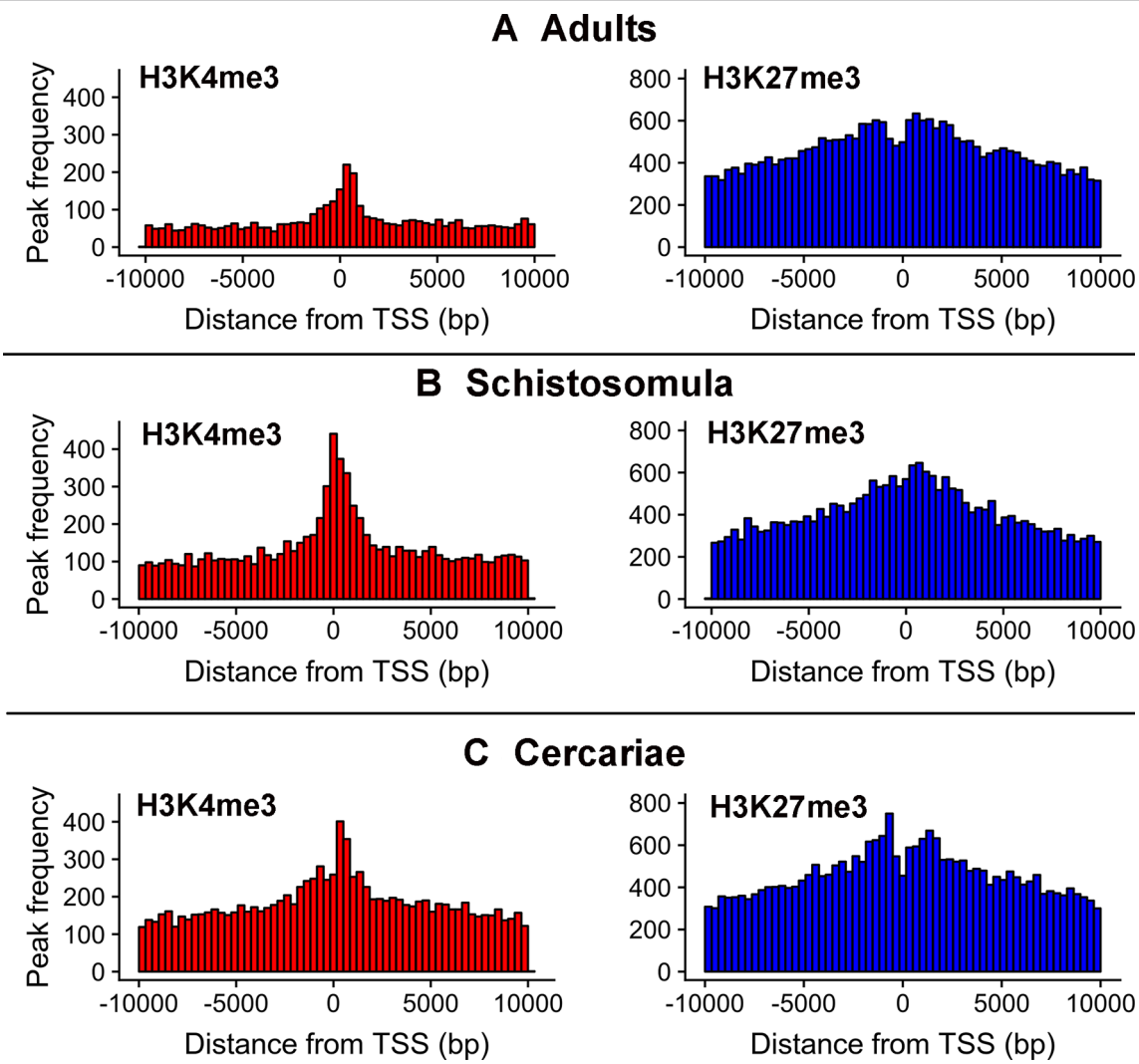


FIGURE 4 | Epigenetic histone marks H3K4me3 and H3K27me3 surrounding the TSS of *S. mansoni* lncRNAs. The frequency of the H3K4me3 marks (red) or of the H3K27me3 marks (blue) mapping within 10 kb around the TSS of all lncRNAs in **(A)** adults, **(B)** schistosomula and **(C)** cercariae was computed.

TSS of a different set of lncRNAs, with abundancies ranging from 740 lncRNAs with the H3K4me3 mark exclusively in schistosomula to 282 lncRNAs with the mark exclusively in cercariae, and 214 lncRNAs with the mark exclusively in adults (**Figure 5, green**). Interestingly, among the lncRNAs with the most abundant patterns of marks at their TSS, there are 316 lncRNAs in cercariae that have the characteristic marks of bivalent poised promoters (having both H3K4me3 and H3K27me3 marks at their TSS) (Voigt et al., 2013) (**Figure 5, blue**). This is analogous to the marks at the TSS of protein-coding genes in cercariae, where most genes have the bivalent mark (Roquis et al., 2018), indicating that lncRNAs are under a similar transcriptional regulatory program as the protein-coding genes in cercariae. **Supplementary Table S5** has a complete UpSet plot similar to that of **Figure 5**, showing the number of lncRNAs found in all different intersections, along with the lists of lncRNAs belonging to each intersection set.

Gene Co-Expression Analyses

Once we identified our final lncRNAs set, we applied weighted gene co-expression network analyses (WGCNA) to integrate the expression level differences observed for lncRNAs and mRNAs among all life-cycle stages and the gonads, using all RNA-seq libraries available. The file containing expression levels (in TPM) for all transcripts in all 633 RNA-Seq libraries is available at http://schistosoma.usp.br/. After normalization and gene filtering (see Methods), 90 libraries out of the 112 from the different stages (mixed-sex adults were not included) remained in the WGCNA analyses, and 19,258 transcripts were retained (12,693 protein-coding genes and 6,565 lncRNAs).

Samples from miracidia, sporocysts, schistosomula, cercariae, and gonads (testes and ovaries) were correctly clustered together by their expression correlation, based on Euclidian distance metrics (**Figure 6**). For samples from adult worms, in spite of the fact that we have one cluster branch mainly composed of

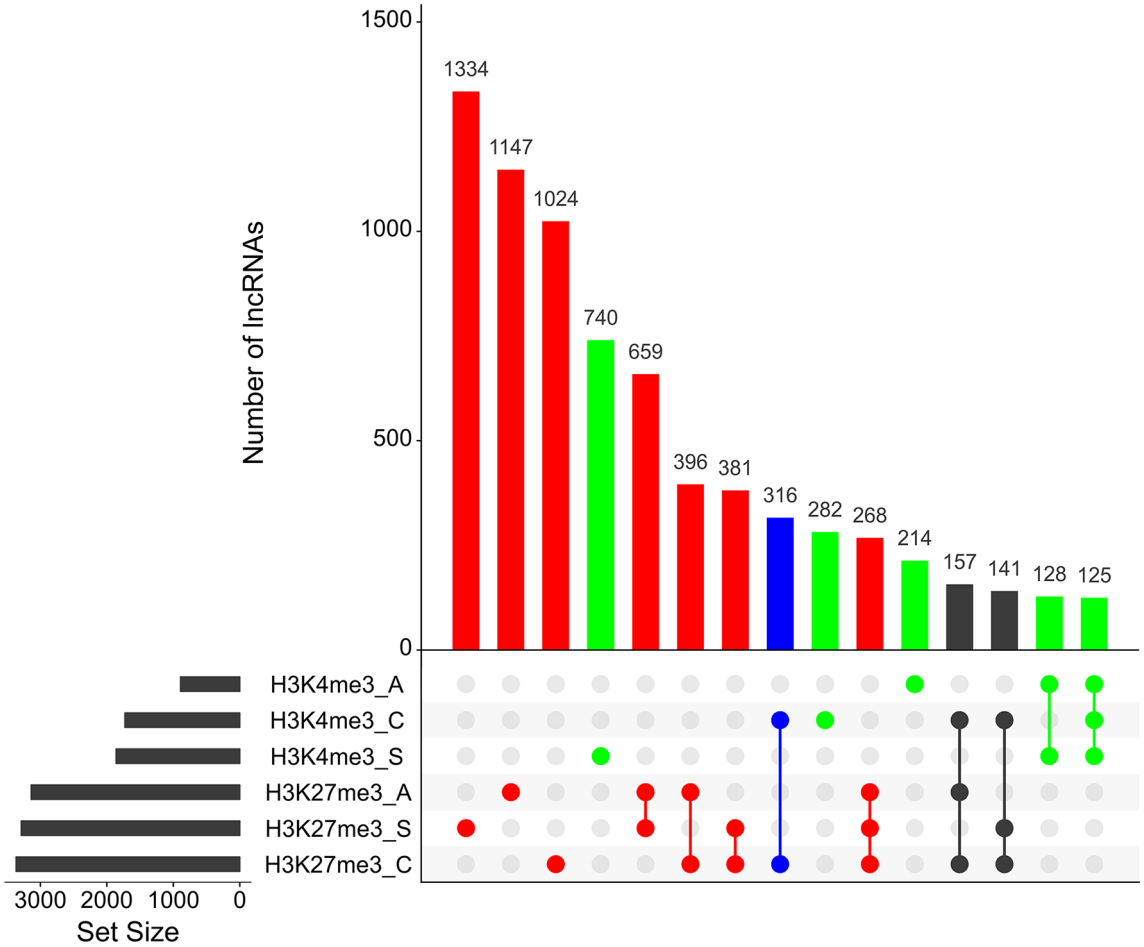


FIGURE 5 | Hundreds of *S. mansoni* lncRNAs have histone transcriptional activating or repressive marks at their TSS exclusively in one life-cycle stage. The UpSet intersection diagram shows the number of *S. mansoni* lncRNAs (y-axis) that have been detected in each of the intersection sets, indicated by the connected points in the lower part of the plot, as having the H3K4me3 transcriptional activating marks (green) and/or the H3K27me3 repressive marks (red) within 1 kb (upstream or downstream) from their TSS. Six histone mark data sets indicated at the bottom left were analyzed: H3K4me3_A in adults, H3K4me3_C in cercariae, H3K4me3_S in schistosomula, H3K27me3_A in adults, H3K27me3_S in schistosomula, and H3K27me3_C in cercariae, and each set size black bar represents the number of lncRNAs that contain the indicated histone mark at the indicated stage. The top 15 most enriched intersection sets are shown; all intersection sets and the lists of lncRNAs in each intersection set are shown in **Supplementary Table S5**. The intersection set in blue shows the number of lncRNAs with the simultaneous H3K4me3_C/H3K27me3_C marks at their TSS in cercariae, characteristic of poised promoters.

females, and another mainly composed of males, there are some male samples among the female ones, and vice versa. Besides, due to the known similarity between males and juveniles (Wang et al., 2017), their samples were not well separated. It is interesting to note that immature females, which were shown to have a similar expression profile as that of males (Lu et al., 2016), are clustered here in the male branch. As the WGCNA performs an unsupervised co-expression analysis, we decided to keep all male and female samples in the analysis, including those that are clustered apart from their main group, in order not to add a bias in the construction of modules.

We identified 15 different lncRNAs/mRNAs co-expression modules (**Figure 7**), the sizes ranging from 215 to 3,318 transcripts (**Table 2** and **Supplementary Table S6**). The ratio between the number of lncRNAs and mRNAs that comprise each module varies among the modules; thus, whereas lncRNAs comprise 86% of the transcripts in the cyan module, only 5% of the transcripts from the black module are lncRNAs (**Table 2**).

A Pearson correlation analysis indicates that each stage/tissue has at least one module whose gene expression has a statistically significant positive correlation with that stage or tissue (**Figure 8**). Some stages also have modules that have a statistically significant negative correlation, such as the black module that is negatively correlated with miracidia/sporocysts. For the black module, the transcripts that compose the module have an expression in miracidia/sporocysts that is lower when compared with the overall expression of those transcripts across the other stages. The gray color represents the group of transcripts with a highly heterogeneous co-expression pattern that could not cluster into any of the 15 modules. In fact, it can be seen in **Figure 8** that in

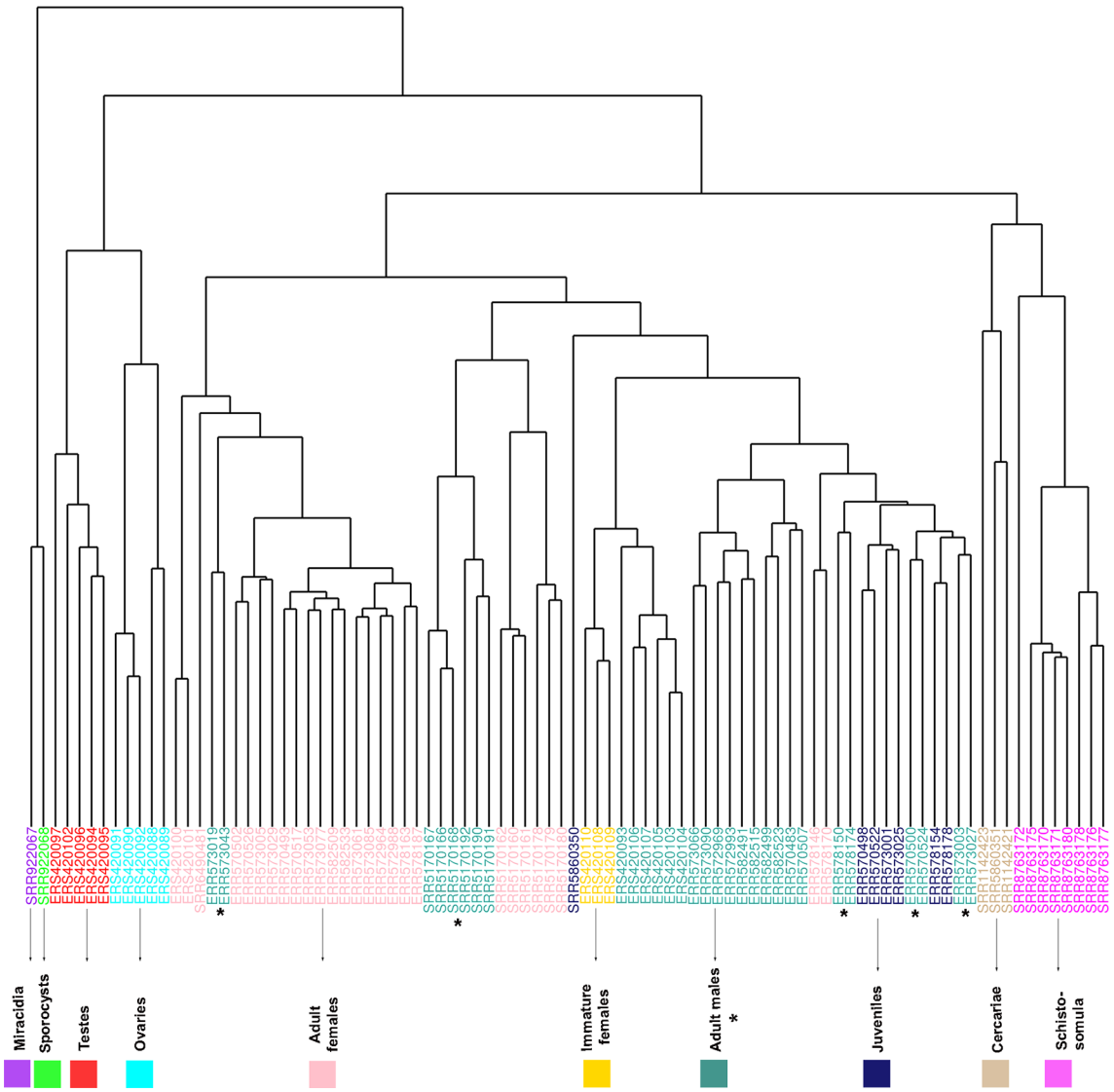


FIGURE 6 | RNA-seq samples clustering based on Euclidian distance according to the expression levels of all genes used for WGCNA. The expression levels of all lncRNA and mRNA genes from all RNA-seq data sets analyzed in the WGCNA were used for unsupervised clustering of the samples, including RNA-seq data sets from adult females (pink), adult males (turquoise), cercariae (tan), immature females (gold), juveniles (midnight blue), miracidia (purple), ovaries (cyan), schistosomula (magenta), sporocysts (green), and testes (red). The SRA or ENA accession number for each RNA-seq library is indicated at each leaf. The asterisks mark the adult male data sets, whose clustering pattern is the most spread one.

this group, the best correlation coefficient obtained in juveniles is lower ($|r| = 0.32$), and the p-value is much higher ($p = 0.002$) than the best parameters that were obtained in at least one stage for any module ($|r| \geq 0.51$ and $p \leq 3e\text{-}07$). Here, our choice of keeping in the WGCNA analysis, those male and female samples that cluster apart from their main group (**Figure 6**) have an impact, decreasing the correlation coefficient of the modules mostly correlated to males or females (pink or turquoise, respectively) when compared with correlation coefficients in the other stages/tissues, nevertheless, they still have a statistically significant high correlation.

We chose three RNA-seq library samples from each of the nine different stages/tissues (among all the libraries under analysis) to construct a representative expression heatmap (**Figure 9**). This heatmap shows the expression across all stages of the top 50 transcripts with the highest gene module membership (GMM) to the most correlated module of each stage (as seen in **Figure 8**) (for GMM definition see WGCNA background and glossary, available at https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/) (Langfelder and Horvath, 2008). The heatmap (**Figure 9**) confirms that the top transcripts belonging to one module are more expressed in one given stage/tissue, which is the stage/tissue with which the module has the highest correlation. It is noteworthy that female library SRR5170160, which clustered inside the male group (**Figure 6**) when all

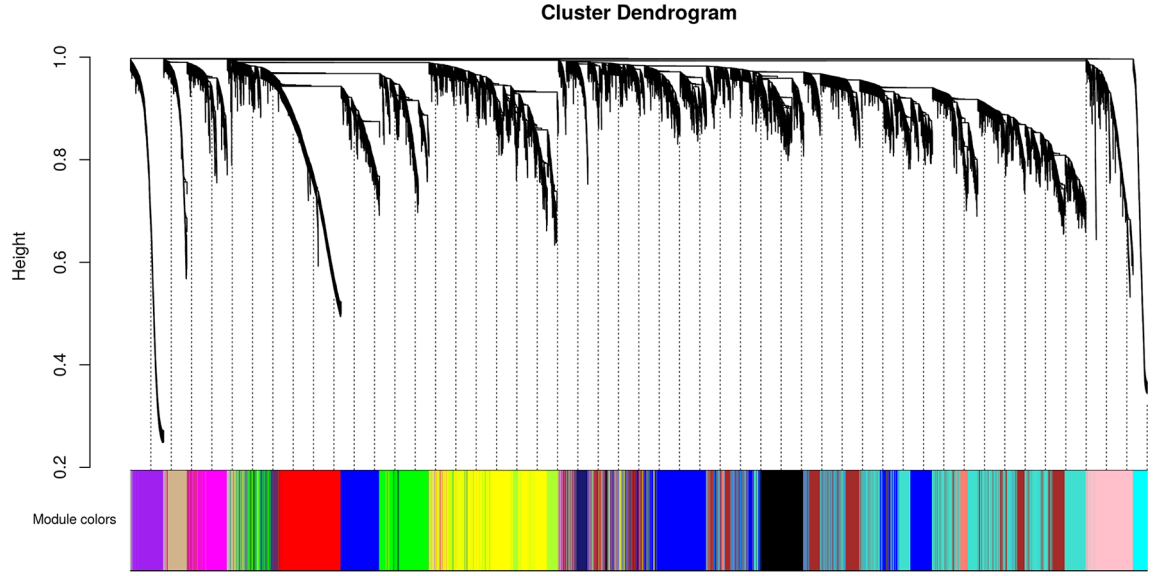


FIGURE 7 | Identification of gene co-expression modules among the different RNA-seq libraries analyzed with the WGCNA tool. Gene hierarchical cluster dendrogram based on a dissimilarity measure of the Topological Overlap Matrix (1 – TOM) calculated by WGCNA, together with the 15 assigned module colors.

TABLE 2 | Number of transcripts per module and percentage of lncRNAs in each module.

Module color	Total number of transcripts	mRNAs	lncRNAs	% of lncRNAs	Stage of higher absolute correlation value
Black	989	940	49	5	Miracidia/Sporocysts
Blue	3,211	2,688	523	16	Juveniles
Brown	2,466	1,761	705	29	Gonads
Cyan	253	36	217	86	Ovaries
Green	1,308	841	467	35	Gonads
Greenyellow	502	417	85	17	Gonads
Magenta	748	273	475	64	Schistosomula
Midnight blue	215	43	172	80	Juveniles
Pink	840	506	334	40	Adult Females
Purple	590	274	316	54	Miracidia/Sporocysts
Red	1,254	333	921	73	Testes
Salmon	267	230	37	14	Gonads
Tan	356	158	198	56	Cercariae
Turquoise	3,318	2,470	848	26	Adult Males
Yellow	2,067	1,398	669	32	Adult Males

filtered transcripts under analysis were used for clustering, now is correctly clustered with the other female samples (**Figure 9**) when only the top 50 transcripts with the highest GMM are considered. Also, juveniles share with adult males a similar expression pattern of the top 50 male genes, which is in line with the clustering of juveniles along with males in the analysis of **Figure 6**.

Validation of lncRNAs Expression by RT-qPCR

We designed PCR primer pairs for a selected set of eleven lincRNAs belonging to five different modules, as determined by WGCNA, to detect their expression along the different *S. mansoni* life-cycle stages and to eventually validate their different expression levels at the stages. Our selection was based on the Gene Trait Significance score (GS score) (**Supplementary Table S7**) of each lncRNA in the module where it belongs, which varies from −1 to 1, using the stages as external information (see Methods). The higher the absolute value of the GS score, the more biologically significant and correlated to the stage of interest is the transcript expression. For the RT-qPCR assays, we used samples from eggs (E), miracidia (Mi), cercariae (C), schistosomula (S), adult males (M), and females (F).

First, we measured the expression of five protein-coding genes that were used as stage markers (Parker-Manuel et al.,

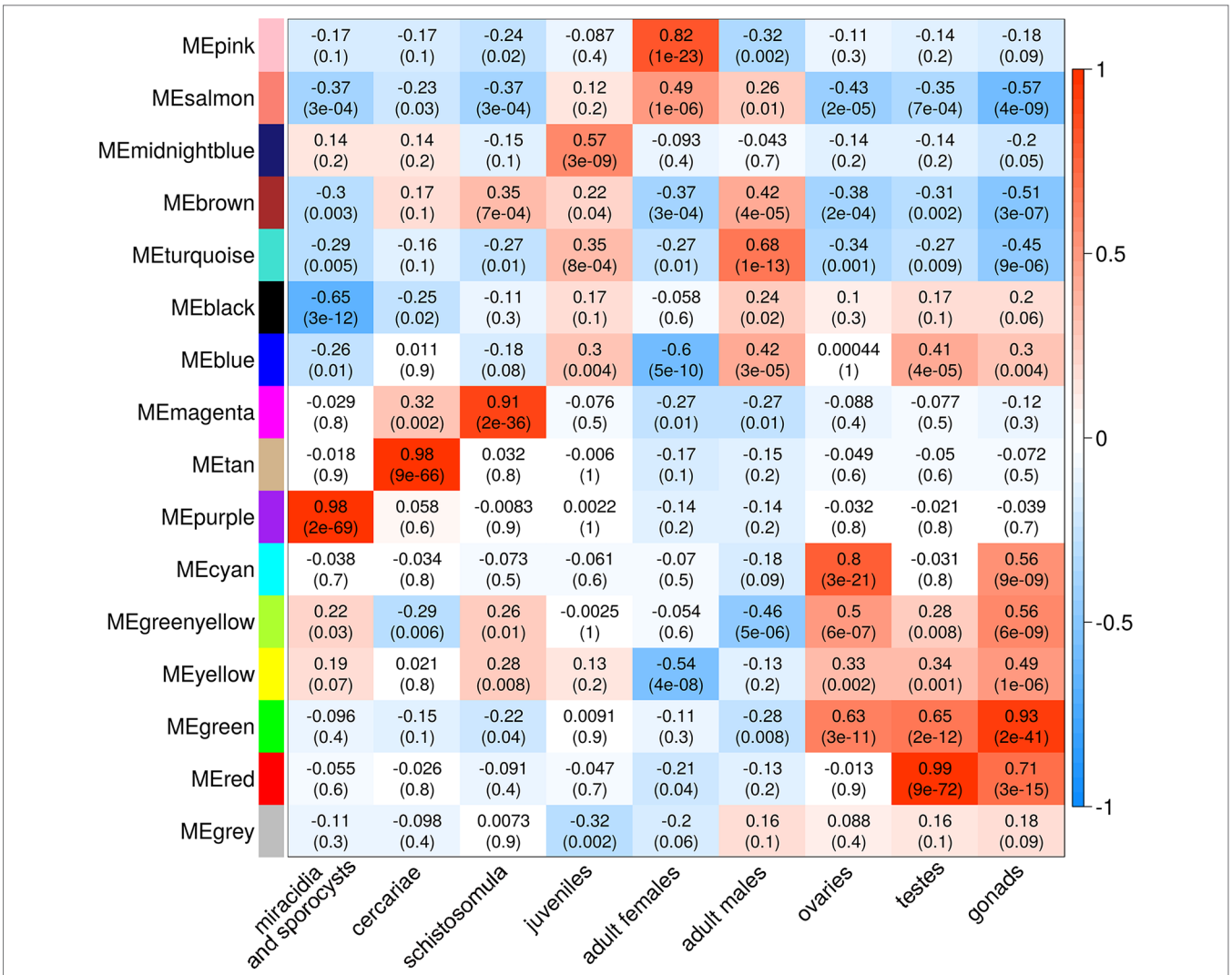


FIGURE 8 | Each parasite life-cycle stage (or tissue) has at least one highly correlated gene co-expression module. Each cell in the table shows the Pearson correlation (with the p-value in parenthesis) between each of the 15 co-expression modules determined by WGCNA (indicated at left) and the stages/tissues of *S. mansoni* (indicated at the bottom). The cells are colored according to the scale (at right), which is related to the Pearson correlation coefficient.

2011; Anderson et al., 2016), and we found that in our RNA samples, they were more highly expressed at the predicted stages (**Supplementary Figure S4**).

Then, we tested the selected eleven lincRNAs and detected that they were expressed in at least one of the six stages that were assayed; specifically, each of six lincRNAs were more highly expressed at the stage predicted by the correlation with the modules **(Figure 10)**, at four life-cycle stages: two more highly expressed in miracidia (SmLINC158013-IBu and SmLINC123205-IBu, purple module), two in cercariae (SmLINC123474-IBu and SmLINC134196-IBu, tan module), one in schistosomula (SmLINC105065-IBu, magenta module), and one in males (SmLINC100046-IBu, turquoise module) **(Figure 10)**. In **Supplementary Figure S5** we present the values in transcripts per million reads (TPM) from the RNA-seq libraries for each of these six validated lincRNAs. Additionally, the five other lincRNAs that were tested were detected as expressed across all stages; however, they were not differentially expressed as predicted by the RNA-seq (**Supplementary Figure S6**). This indicates that there is variability of lncRNAs expression between the experimental conditions and parasite strain used in our assays and those found among the dozens of samples that are publicly available.

Protein-Coding Genes Ontology Enrichment and lncRNA Hub Genes in the Modules

Gene ontology (GO) enrichment analyses show that the protein-coding genes belonging to the red module, which have a correlation of 0.99 with testes, are enriched with processes related to sperm motility such as cilium movement and the

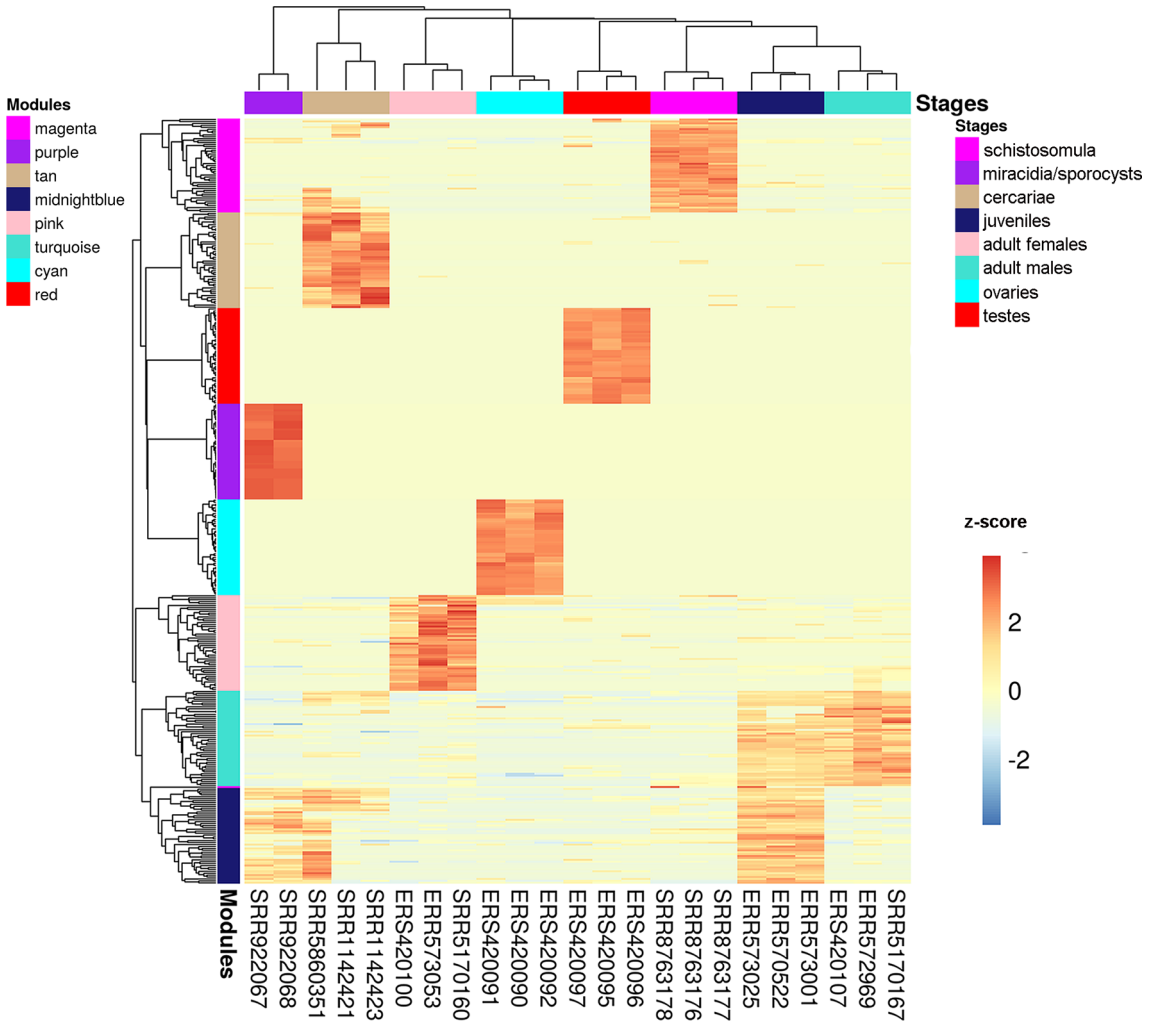


FIGURE 9 | Gene expression heatmap across the life-cycle stages/tissues of the parasite. Representative heatmap of gene expression levels for the top 50 genes (each in one line) with the highest GMM values from each of the eight modules (indicated at the left) with the highest positive correlation to each stage/tissue (indicated at the top). Expression data from three chosen RNA-seq libraries (one in each column) were picked as representative libraries for each stage/tissue, and their SRA or ENA accession numbers are given at the bottom; for miracidia/sporocysts only two RNA-seq libraries were available. Unsupervised clustering using the Euclidean distance was performed; expression of each gene (in one line) is shown as the z-score (from –3 to 3), which is the number of standard deviations above (red) or below (blue) the mean expression value of that gene across all RNA-seq libraries; the z-score color scale is shown on the right.

axoneme assembly **(Figure 11A)**. Besides, the green module, correlated with both ovaries and testes, is enriched with proteins associated with cellular replication **(Figure 11B)**. All other modules with GO enrichment, which in general are enriched with proteins associated to general metabolism, are presented in **Supplementary Figures S7–S10**. The black, cyan, midnight blue, purple, and tan modules have no significantly enriched GO terms due to the small number of protein-coding genes with GO annotation within each of these modules.

All transcripts that belong to the same module are connected; however, to better visualize this, gene co-expression networks were constructed only with the most connected genes (as determined by the adjacency threshold) (**Figure 12**), and they show, along with the correlation values presented in **Supplementary Table S7**, that some lncRNAs are hub genes from the network. **Figures 12A**, **B** show lncRNA hub genes in the co-expression networks from the purple and tan modules strongly correlated with miracidia/sporocysts and cercariae life-cycle stages, respectively. In both modules, the lncRNAs represent around half of the transcripts that comprise the modules (see **Table 2**). However, there are some cases, such as in the red module, where three quarters of the member transcripts are lncRNAs, and among the most connected genes in that co-expression network, almost all are lncRNAs (**Figure 12C**). Also, in the blue module only, 16% of the member transcripts are lncRNAs, and only one is among the most connected genes in the co-expression network (**Figure 12D**). All the gene networks for all modules in a format compatible with Cytoscape are available at **Supplementary Table S8**. An adjacency cutoff threshold of 0.1 was used.

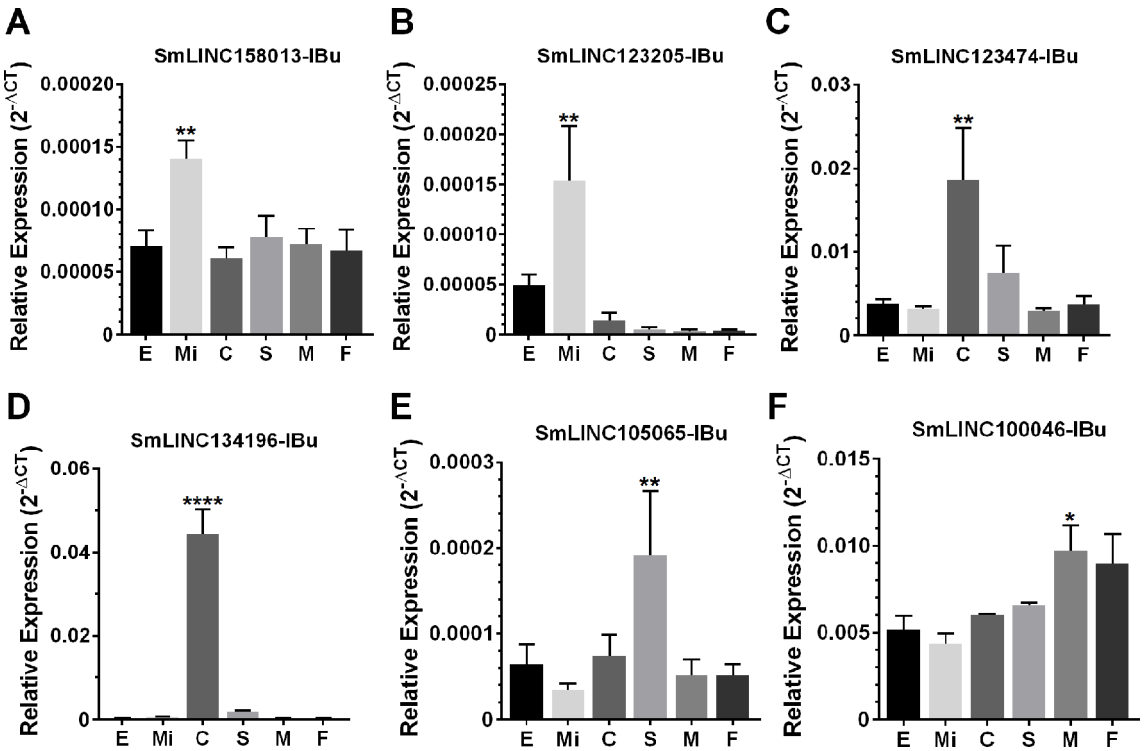


FIGURE 10 | Confirmation by RT-qPCR of the module-specific lincRNAs relative expression. Six lincRNAs were measured at different developmental stages of *S. mansoni*. From left to right in the x-axis, lincRNAs were measured in RNA samples from eggs (E), miracidia (Mi), cercariae (C), *in vitro* mechanically transformed schistosomula cultivated for 24 h (S), adult males (M) and females (F). The lincRNAs relative gene expression was calculated against the geometric mean of two housekeeping genes: Smp_090920 and Smp_062630. **(A)** and **(B)** show SmLINC158013-IBu and SmLINC123205-IBu representing the **purple** module, specific for miracidia/sporocysts. In **(C)** and **(D),** the SmLINC123474-IBu and SmLINC134196-IBu representing the cercariae-specific **tan** module. In **(E),** the schistosomula-specific lincRNA SmLINC105065-IBu from the **magenta** module and **(F)** the adult male-specific lincRNA SmLINC100046-IBu from the **turquoise** module. Bars represent standard deviation of the mean from four biological replicates for each stage. Two technical replicates were assayed for each of the four biological replicates per stage. The ANOVA Tukey test was used to calculate the statistical significance of the expression differences among the parasite stage samples (*p value≤0.05; **p value≤0.01; ****p value≤0.0001). For clarity purposes, we show only the highest p value obtained in the ANOVA Tukey test for expression comparisons against one another among the stages.

LncRNAs Expressed in Single Cells

Finally, analyses using single-cell data from two stages, mother sporocysts stem cells and juveniles' stem cells, identified three different clusters. Cluster 1 is composed of a subgroup of juvenile stem cells, cluster 2 is composed of all mother sporocysts stem cells, and cluster 3 is composed of a second and smaller subgroup of juvenile stem cells (**Figure 13A**). The marker gene analyses show, for the first time in *S. mansoni,* that lncRNAs have specific expression also at the single-cell level, where from the top 10 markers that allow us to differentiate mother sporocysts stem cells from juvenile stem cells, eight are lncRNAs (**Figure 13B**), confirming the stage specificity of lncRNAs also seen in whole worm analyses by WGCNA. Besides, another lncRNA was identified as a marker for cluster 3 when compared with the other two clusters (**Figure 13B**).

DISCUSSION

When the human genome was first sequenced, the vast genomic regions that lie between protein-coding genes (intergenic regions) were considered junk DNA; one decade later, the Encyclopedia of DNA Elements (ENCODE) project found that 80% of the human genome serves some biochemical purpose (Pennisi, 2012), including giving rise to the transcription of nearly 10,000 lncRNAs (Derrien et al., 2012). Although we are still at the beginning of the studies with lncRNAs, with the vast majority of their roles and mechanisms of action in human beings still unknown, it is now clear that most of the lncRNAs are transcribed from intergenic regions and are key regulators in vital processes (Kitagawa et al., 2013; Rosa and Ballarino, 2016; Golicz et al., 2018), being associated to several pathologies in humans, such as cancer (Fang and Fullwood, 2016), Alzheimer's (Zijian, 2016), and cardiac diseases (Simona et al., 2018).

In *S. mansoni*, with the release in 2012 of version 5.2 of the genome and annotated transcriptome (Protasio et al., 2012), and with the accumulation until 2017 of large amounts of information on gene expression obtained through 88 publicly available RNA-seq libraries, our group decided to map the RNA-seq data and identify the lncRNAs repertoire expressed in this parasite (Vasconcelos et al., 2017); this was followed by two other papers that provided an additional set of lncRNAs

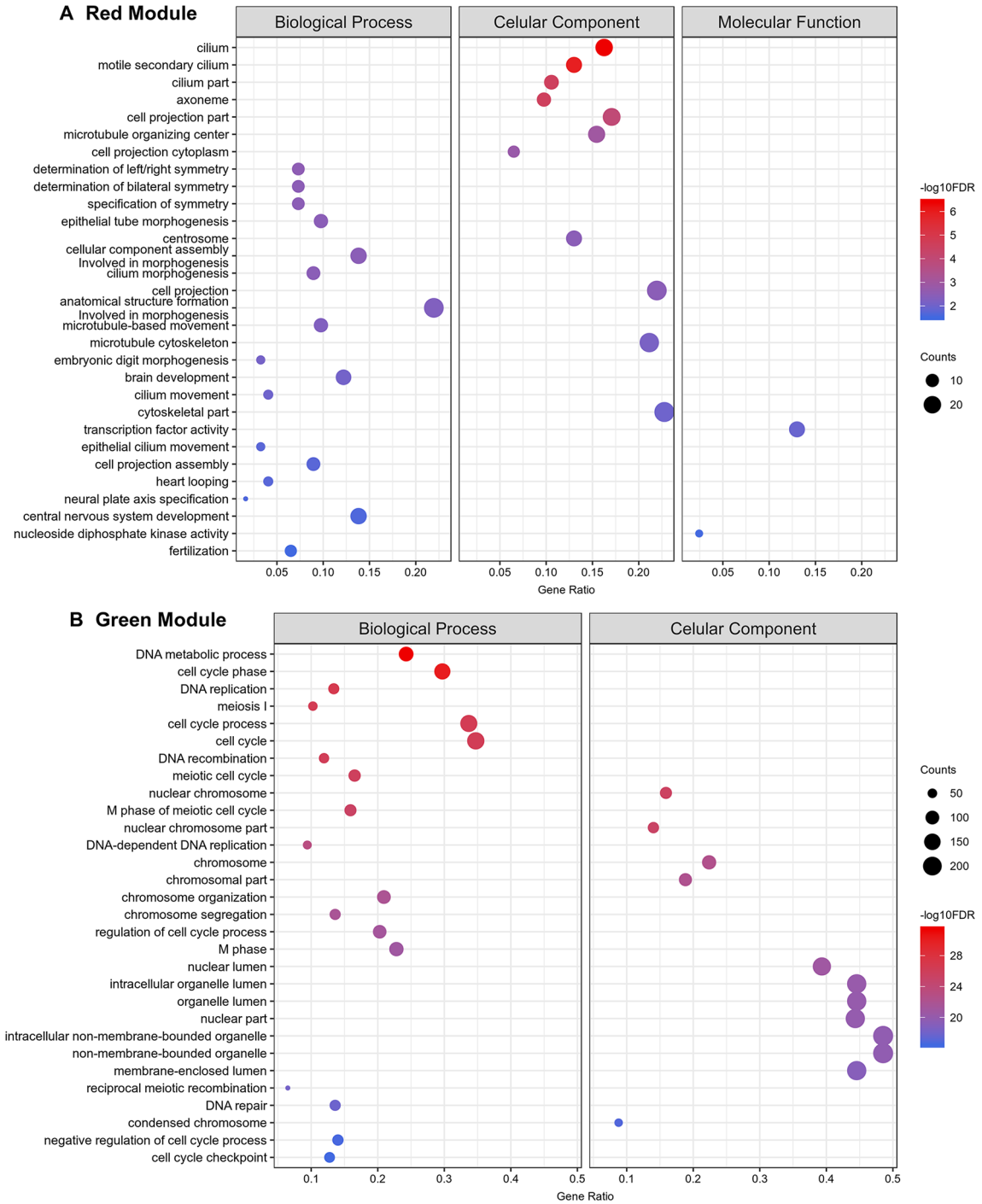


FIGURE 11 | Top 30 Gene Ontology most significantly enriched terms for protein-coding genes belonging to the red and green co-expression network modules. At left are the enriched GO term annotations. For the **(A)** red (testes) and the **(B)** green (gonads) modules, the enriched GOs are separately represented into the three major GO term categories, namely Biological Process, Cellular Component and Molecular Function. No Molecular Function term was significantly enriched in the green module. The size of the circles is proportional to the number of genes (counts scale on the right) in each significantly enriched GO category, and the colors show the statistical significance of the enrichment, as indicated by the -log10 FDR values (color-coded scales at the right).

(Liao et al., 2018; Oliveira et al., 2018). In the present work, by extending the analysis to 633 publicly available RNA-seq libraries, and by performing a detailed curation of the assembled transcripts, we observed that at the sequencing depth obtained with the current RNA-seq data sets, a considerable amount of partially processed pre-mRNAs is being sequenced. These pre-mRNAs give rise to assembled transcript units showing intron retention and frequent stop codons in the retained introns,

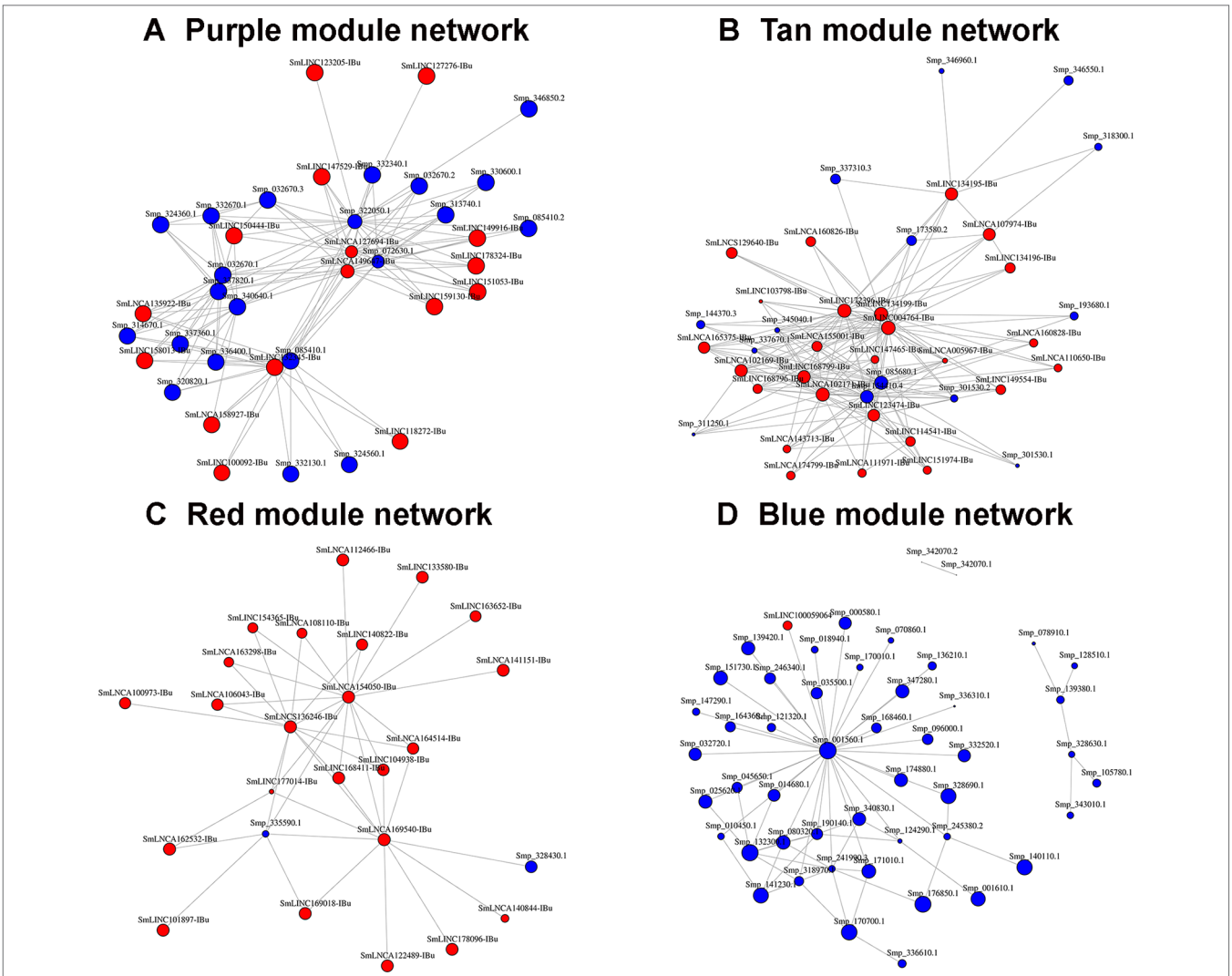


FIGURE 12 | Gene co-expression networks. The top most highly connected genes from the **(A)** purple (miracidia/sporocysts), **(B)** tan (cercariae), **(C)** red (testes) or **(D)** blue (females) co-expression network modules are shown. The adjacency thresholds are 0.77, 0.48, 0.58, and 0.26, respectively. Each red circle represents one lncRNA (SmLINC, SmLNCA, or SmLNCS), and each blue circle represents one protein-coding gene (Smp_). Circle sizes are related to the intramodular connectivity value for each transcript.

and therefore, these transcripts can be mistakenly annotated as lncRNAs.

In fact, the failure to identify partially processed pre-mRNA in previous publications (Vasconcelos et al., 2017; Liao et al., 2018) may explain the report of probable protein-coding genes as lncRNAs (**Supplementary Figure S1**). Our current pipeline has removed at step 4 a total of 31,183 assembled transcripts that had partial or total exon-exon overlap on the same genomic strand with known *S. mansoni* protein-coding genes, and this included around 14,000 assembled transcripts that represented fully processed mature protein-coding transcripts that exactly matched the annotated v 7.1 transcripts from the Wellcome Sanger Institute, as well as some 17,000 assembled transcripts that for the most part represent partially processed pre-mRNAs with intron retention; among the latter are 4,293 transcripts that were previously classified as lncRNAs (Vasconcelos et al., 2017; Liao et al., 2018; Oliveira et al., 2018) and are now excluded. With the six stringent filtering steps used in the present work, we are confident that our final set of 16,583 lncRNAs is a robust representation of the lncRNAs complement expressed in *S. mansoni*, of which 11,022 transcripts are novel lncRNAs, and 5,561 have gene overlap with lncRNAs already reported in previous works (Vasconcelos et al., 2017; Liao et al., 2018; Oliveira et al., 2018).

One question that has been raised about lncRNAs is the possibility that their function is executed through translation into short peptides, a concern that arises from the fact that almost all lncRNAs encode short canonical ORFs within their sequences (Verheggen et al., 2017); the fact that the size distribution of ORFs found within our set of lncRNAs (sense) is very similar to

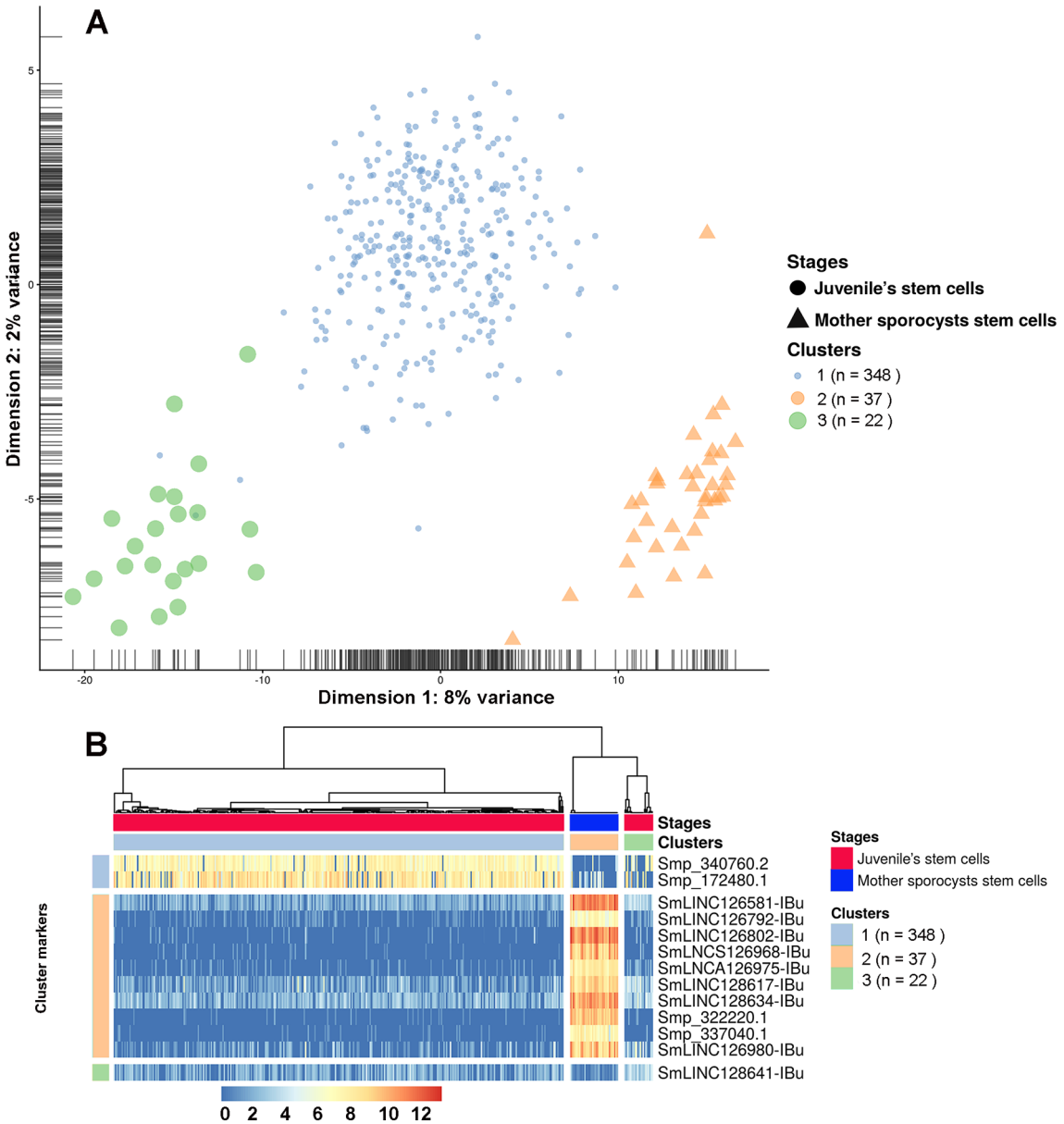


FIGURE 13 | Single-cell expression analysis identified three different cell population clusters when comparing *S. mansoni* juveniles' stem cells and mother sporocysts stem cells and lncRNAs as gene markers at the single-cell level. **(A)** Single-cell RNA-Seq data from two RNA-Seq libraries, one from juveniles' stem cells and another from mother sporocysts' stem cells, were analyzed with the SC3 tool that performed an unsupervised clustering of the cells based on the single-cell gene expression data. Principal component analysis plot, where the symbol colors and sizes indicate the three clusters identified by SC3, and the shapes indicate the two life-cycle stages from which the stem cells were isolated. The symbol size is inversely related to the number of cells that belong to the cluster. **(B)** In the marker-gene expression matrix (log-transformation, represented by the color scale), the statistically significant gene markers are the rows, and the cells are columns. The life-cycle stage from which each cell was isolated is indicated by the color bar at the top (stages). The clusters of cells are separated by white vertical lines and are indicated by the second color bar at the top (clusters). The cluster marker genes are separated by white horizontal lines, the markers groups are indicated at left, and the names of the marker genes at right. Only the top 10 most significant marker genes are shown for cluster 2.

the size distribution of random ORFs found within their reverse-complement sequences and within the reverse-complement sequence of mRNAs suggests that the putative short ORFs from the lncRNAs identified here are indeed random ORFs, most probably not translated into short functional peptides. Nevertheless, future functional characterization in *S. mansoni* of selected lncRNAs may eventually include a search for a possible dual function role (Nam et al., 2016; Choi et al., 2018) both as lncRNA and through a translated short peptide.

Histone marks were found here at the TSS of lncRNAs, and the identification of different sets of lncRNAs that have at their TSS the transcriptional activation H3K4me3 mark, or the repressive H3K27me3 mark, when the three life-cycle stages are compared, suggests that lncRNAs expression in *S. mansoni*

is regulated by an epigenetic program. This finding reinforces the hypothesis that different lncRNAs may play important roles along the parasite life-cycle, and the sets of lncRNAs identified in this analysis might be the first candidates to be explored for further functional characterization.

Gene co-expression networks correlated to the different *S. mansoni* life-cycle stages were identified by our analyses, and they pointed to sets of protein-coding genes and lncRNAs with expression most correlated to one given stage. This information provides an initial platform for prioritizing the lncRNAs to be selected for further direct functional characterization, which will include a search for altered *S. mansoni* phenotypes upon knockdown of lncRNA candidates. In *Plasmodium falciparum*, the knockdown of antisense lncRNAs has down-regulated the active var gene, a gene related to immune evasion, erasing the epigenetic memory and substantially changing the var gene expression pattern (Amit-Avraham et al., 2015). In analogy, it is expected that characterization of lncRNAs in *S. mansoni* will help to recognize the biochemical pathways where they play a functional role, will permit to identify their interacting protein partners, and will eventually point to relevant ways of intervention in the parasite physiology.

Due to the complex and diverse mechanisms displayed by lncRNAs in regulating protein-coding genes and miRNAs, the majority of studies have not progressed beyond cell or animal models, and progression toward the clinic has been slow (Harries, 2019). Nevertheless, lncRNAs represent potentially good therapeutic targets (Matsui and Corey, 2017; Blokhin et al., 2018; Harries, 2019). As reviewed by Matsui and Corey (2017), in Angelman syndrome model mouse, the administration of antisense oligonucleotides (ASOs), which target the Ube3a-ATS lncRNA for degradation, partially reversed some cognitive defects associated with the disease in the animals (Meng et al., 2014). Also, in xenograft melanoma models, the intravenous injection of ASOs targeting the lncRNA SAMMSON caused p53 activation, tumor growth suppression, decreased cell proliferation, and increased apoptosis (Leucci et al., 2016). In this respect, it is noteworthy that lncRNAs are considerably less conserved between species when compared with protein-coding genes (Pang et al., 2006; Blokhin et al., 2018), and that only a few dozen ancient lncRNAs have conserved orthologs between ancient non-amniote *Xenopus* and the closest amniote chicken model animals (Necsulea et al., 2014), which shows that lncRNAs have evolutionarily conserved gene regulatory functions but low-sequence conservation across distant species (Necsulea et al., 2014). This feature reduces the chances that targeting a lncRNA in *S. mansoni*, for example, with ASOs, will cause unwanted off-target effects against the mammalian host.

DATA AVAILABILITY

The data sets analyzed in this study can be found in the SRA repository (https://www.ncbi.nlm.nih.gov/sra) and in the ENA repository (https://www.ebi.ac.uk/ena). The specific accession numbers for each and all data sets that were downloaded from these databases and used here are given in **Supplementary Table S1**.

ETHICS STATEMENT

All protocols involving animals were conducted in accordance with the Ethical Principles in Animal Research adopted by the Brazilian College of Animal Experimentation (COBEA), and the protocol/experiments have been approved by the Ethics Committee for Animal Experimentation of Instituto Butantan (CEUAIB Protocol number 1777050816).

AUTHOR CONTRIBUTIONS

LM, MA and SV-A conceived the project. LM and SV-A designed the experiments and wrote the paper. LM and DM-V performed the *in silico* analyses. MA, GS, RR, and GO performed the wet lab experiments and analyses. LM and SV-A analyzed and interpreted the data. DP contributed with informatic resources.

FUNDING

This work was supported by the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) grant numbers 2014/03620-2 and 2018/23693-5 to SV-A. LM, GS and RR received FAPESP fellowships (grant numbers 2018/19591-2, 2018/24015-0 and 2017/22379-2, respectively) and DM-V received a fellowship from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). SV-A laboratory was also supported by institutional funds from Fundação Butantan and received an established investigator fellowship award from CNPq, Brasil.

ACKNOWLEDGMENTS

We thank Dr. J.C. Setubal for access to the computational facilities of the Bioinformatics Laboratory of Instituto de Química, Universidade de São Paulo (USP). We also acknowledge Patricia Aoki Miyasato and Dr. Eliana Nakano, Laboratorio de Malacologia, Instituto Butantan, for maintaining the *S. mansoni* life cycle.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00823/full#supplementary-material

REFERENCES

Amit-Avraham, I., Pozner, G., Eshar, S., Fastman, Y., Kolevzon, N., Yavin, E., et al. (2015). Antisense long noncoding RNAs regulate var gene activation in the malaria parasite *Plasmodium falciparum*. *Proc. Nat.l Acad. Sci. U. S. A.* 112, E982–E991. doi: 10.1073/pnas.1420855112

Anderson, L., Amaral, M. S., Beckedorff, F., Silva, L. F., Dazzani, B., Oliveira, K. C., et al. (2016). *Schistosoma mansoni* egg, adult male and female comparative gene expression analysis and identification of novel genes by RNA-Seq. *PLoS Negl. Trop. Dis.* 9, e0004334. doi: 10.1371/journal.pntd.0004334

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., et al. (2007). High-resolution profiling of histone methylations in the human genome. *Cell* 129, 823–837. doi: 10.1016/j.cell.2007.05.009

Basch, P. F. (1976). Intermediate host specificity in *Schistosoma mansoni*. *Exp. Parasitol.* 39, 150–169. doi: 10.1016/0014-4894(76)90022-9

Basch, P. F. (1981). Cultivation of *Schistosoma mansoni in vitro*. I. Establishment of cultures from cercariae and development until pairing. *J. Parasitol.* 67, 179–185. doi: 10.2307/3280632

Batugedara, G., Lu, X. M., Bunnik, E. M., and Le Roch, K. G. (2017). The role of chromatin structure in gene regulation of the human malaria parasite. *Trends Parasitol* 33, 364–377. doi: 10.1016/j.pt.2016.12.004

Bhat, S. A., Ahmad, S. M., Mumtaz, P. T., Malik, A. A., Dar, M. A., Urwat, U., et al. (2016). Long non-coding RNAs: mechanism of action and functional utility. *Non-coding RNA Research* 1, 43–50. doi: 10.1016/j.ncrna.2016.11.002

Blokhin, I., Khorkova, O., Hsiao, J., and Wahlestedt, C. (2018). Developments in lncRNA drug discovery: where are we heading? *Expert Opin. Drug Discov.* 13, 837–849. doi: 10.1080/17460441.2018.1501024

Cao, H., Wahlestedt, C., and Kapranov, P. (2018). Strategies to annotate and characterize long noncoding rnas: advantages and pitfalls. *Trends Genet.* 34, 704–721. doi: 10.1016/j.tig.2018.06.002

Cdc. (2018). *Centers for Disease Control and Prevention, Parasites - Schistosomiasis* [Online]. Available: https://www.cdc.gov/parasites/schistosomiasis/ [Accessed 20/07/2018 2018].

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890. doi: 10.1093/bioinformatics/bty560

Choi, S. W., Kim, H. W., and Nam, J. W. (2018). The small peptide world in long noncoding RNAs. *Brief. Bioinformatics*, bby055. doi: 10.1093/bib/bby055

Credendino, S. C., Lewin, N., De Oliveira, M., Basu, S., Andrea, B., Amendola, E., et al. (2017). Tissue- and cell type-specific expression of the long noncoding RNA Klhl14-AS in Mouse. *Int. J. Genomics* 2017, 7. doi: 10.1155/2017/9769171

Dalton, J. P., Day, S. R., Drew, A. C., and Brindley, P. J. (1997). A method for the isolation of schistosome eggs and miracidia free of contaminating host tissues. *Parasitology* 115 (Pt 1), 29–32. doi: 10.1017/S0031182097001091

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* 22, 1775–1789. doi: 10.1101/gr.132159.111

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Fang, Y., and Fullwood, M. J. (2016). Roles, functions, and mechanisms of long non-coding RNAs in cancer. *Genomics Proteomics Bioinformatics* 14, 42–54. doi: 10.1016/j.gpb.2015.09.006

Golicz, A. A., Bhalla, P. L., and Singh, M. B. (2018). lncRNAs in plant and animal sexual reproduction. *Trends Plant Sci.* 23, 195–205. doi: 10.1016/j.tplants.2017.12.009

Gomes Casavechia, M. T., De Melo, G.D.a.N., Da Silva Fernandes, A. C. B., De Castro, K. R., Pedroso, R. B., Da Silva Santos, T., et al. (2018). Systematic review and meta-analysis on *Schistosoma mansoni* infection prevalence, and associated risk factors in Brazil. *Parasitology* 145, 1000–1014. doi: 10.1017/S0031182017002268

Hanly, D. J., Esteller, M., and Berdasco, M. (2018). Interplay between long non-coding RNAs and epigenetic machinery: emerging targets in cancer? *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 373, 20170074. doi: 10.1098/rstb.2017.0074

Harries, L. W. (2019). RNA biology provides new therapeutic targets for human disease. *Front. Genet.* 10, 205. doi: 10.3389/fgene.2019.00205

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., et al. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* 38, 576–589. doi: 10.1016/j.molcel.2010.05.004

Howe, K. L., Bolt, B. J., Shafie, M., Kersey, P., and Berriman, M. (2017). WormBase ParaSite—a comprehensive resource for helminth genomics. *Mol. Biochem. Parasitol.* 215, 2–10. doi: 10.1016/j.molbiopara.2016.11.005

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., Von Mering, C., et al. (2017). Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.* 34, 2115–2122. doi: 10.1093/molbev/msx148

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., et al. (2017). SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* 14, 483. doi: 10.1038/nmeth.4236

Kitagawa, M., Kitagawa, K., Kotake, Y., Niida, H., and Ohhata, T. (2013). Cell cycle regulation by long non-coding RNAs. *Cell. Mol. Life Sci.* 70, 4785–4794. doi: 10.1007/s00018-013-1423-0

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9, 559. doi: 10.1186/1471-2105-9-559

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Leucci, E., Vendramin, R., Spinazzi, M., Laurette, P., Fiers, M., Wouters, J., et al. (2016). Melanoma addiction to the long non-coding RNA SAMMSON. *Nature* 531, 518. doi: 10.1038/nature17161

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). UpSet: Visualization of intersecting sets. *IEEE Trans. Vis. Comput. Graph.* 20, 1983–1992. doi: 10.1109/TVCG.2014.2346248

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12, 323. doi: 10.1186/1471-2105-12-323

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. and Genome Project Data Processing, S. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Liao, Q., Zhang, Y., Zhu, Y., Chen, J., Dong, C., Tao, Y., et al. (2018). Identification of long noncoding RNAs in *Schistosoma mansoni* and *Schistosoma japonicum*. *Exp. Parasitol.* 191, 82–87. doi: 10.1016/j.exppara.2018.07.001

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2–DDCT Method. *Methods* 25, 402–408. doi: 10.1006/meth.2001.1262

Lu, Z., Sessler, F., Holroyd, N., Hahnel, S., Quack, T., Berriman, M., et al. (2016). Schistosome sex matters: a deep view into gonad-specific and pairing-dependent transcriptomes reveals a complex gender interplay. *Sci Rep.* 6, 31150. doi: 10.1038/srep31150

Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 21, 3448–3449. doi: 10.1093/bioinformatics/bti551

Matsui, M., and Corey, D. R. (2017). Non-coding RNAs as drug targets. *Nat. Rev. Drug Discov.* 16, 167–179. doi: 10.1038/nrd.2016.117

Mccarthy, D. J., Campbell, K. R., Wills, Q. F., and Lun, A. T. L. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* 33, 1179–1186. doi: 10.1093/bioinformatics/btw777

Meng, L., Ward, A. J., Chun, S., Bennett, C. F., Beaudet, A. L., and Rigo, F. (2014). Towards a therapy for Angelman syndrome by targeting a long non-coding RNA. *Nature* 518, 409. doi: 10.1038/nature13975

Nam, J. W., Choi, S. W., and You, B. H. (2016). Incredible RNA: dual functions of coding and noncoding. *Mol. Cells* 39, 367–374. doi: 10.14348/molcells.2016.0039

Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., et al. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* 505, 635–640. doi: 10.1038/nature12943

Niknafs, Y. S., Pandian, B., Iyer, H. K., Chinnaiyan, A. M., and Iyer, M. K. (2016). TACO produces robust multisample transcriptome assemblies from RNA-seq. *Nat. Methods* 14, 68. doi: 10.1038/nmeth.4078

Oliveira, K. C., Carvalho, M. L., Maracaja-Coutinho, V., Kitajima, J. P., and Verjovski-Almeida, S. (2011). Non-coding RNAs in schistosomes: an unexplored world. *An. Acad. Bras. Cienc.* 83, 673–694. doi: 10.1590/S0001-37652011000200026

Oliveira, V. F., Mota, E. A., Jannotti-Passos, L. K., Coelho, P. M. Z., Mattos, A. C. A., Couto, F. F. B., et al. (2018). Identification of 170 new long noncoding RNAs in *Schistosoma mansoni*. *Biomed Res. Int.* 2018, 1264697. doi: 10.1155/2018/1264697

Pang, K. C., Frith, M. C., and Mattick, J. S. (2006). Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet.* 22, 1–5. doi: 10.1016/j.tig.2005.10.003

Parker-Manuel, S. J., Ivens, A. C., Dillon, G. P., and Wilson, R. A. (2011). Gene expression patterns in larval *Schistosoma mansoni* associated with infection of the mammalian host. *PLoS Negl. Trop. Dis.* 5, e1274. doi: 10.1371/journal.pntd.0001274

Pennisi, E. (2012). Genomics. ENCODE project writes eulogy for junk DNA. *Science* 337, 1159, 1161. doi: 10.1126/science.337.6099.1159

Protasio, A. V., Tsai, I. J., Babbage, A., Nichol, S., Hunt, M., Aslett, M. A., et al. (2012). A systematically improved high quality genome and transcriptome of the human blood fluke *Schistosoma mansoni*. *PLoS Negl. Trop. Dis.* 6, e1455. doi: 10.1371/journal.pntd.0001455

Roquis, D., Taudt, A., Geyer, K. K., Padalino, G., Hoffmann, K. F., Holroyd, N., et al. (2018). Histone methylation changes are required for life cycle progression in the human parasite *Schistosoma mansoni*. *PLOS Pathog.* 14, e1007066. doi: 10.1371/journal.ppat.1007066

Rosa, A., and Ballarino, M. (2016). Long noncoding RNA regulation of pluripotency. *Stem Cells Int.* 2016, 1797692. doi: 10.1155/2016/1797692

Sati, S., Ghosh, S., Jain, V., Scaria, V., and Sengupta, S. (2012). Genome-wide analysis reveals distinct patterns of epigenetic features in long non-coding RNA loci. *Nucleic Acids Res.* 40, 10018–10031. doi: 10.1093/nar/gks776

Shao, M., and Kingsford, C. (2017). Accurate assembly of transcripts through phase-preserving graph decomposition. *Nat. Biotechnol.* 35, 1167. doi: 10.1038/nbt.4020

Simona, G., Antonio, S. S., Yvan, D., and Fabio, M., (2018). Long noncoding RNAs and cardiac disease. *Antioxid. Redox. Signal.* 29, 880–901. doi: 10.1089/ars.2017.7126

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., and Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* 15, 121–132. doi: 10.1038/nrg3642

Tarashansky, A. J., Xue, Y., Quake, S. R., and Wang, B. (2018). Self-assembling manifolds in single-cell RNA sequencing data. *bioRxiv*. doi: 10.1101/364166

Vasconcelos, E. J. R., Dasilva, L. F., Pires, D. S., Lavezzo, G. M., Pereira, A. S. A., Amaral, M. S., et al. (2017). The *Schistosoma mansoni* genome encodes thousands of long non-coding RNAs predicted to be functional at different parasite life-cycle stages. *Sci Rep.* 7, 10508. doi: 10.1038/s41598-017-10853-6

Verheggen, K., Volders, P.-J., Mestdagh, P., Menschaert, G., Van Damme, P., Gevaert, K., et al. (2017). Noncoding after all: biases in proteomics data do not explain observed absence of lncRNA translation products. *J. Proteome Res.* 16, 2508–2515. doi: 10.1021/acs.jproteome.7b00085

Voigt, P., Tee, W.-W., and Reinberg, D. (2013). A double take on bivalent promoters. *Genes Dev.* 27, 1318–1338. doi: 10.1101/gad.219626.113

Wang, B., Lee, J., Li, P., Saberi, A., Yang, H., Liu, C., et al. (2018). Stem cell heterogeneity drives the parasitic life cycle of *Schistosoma mansoni*. *Elife* 7, e35449. doi: 10.7554/eLife.35449

Wang, J., Yu, Y., Shen, H., Qing, T., Zheng, Y., Li, Q., et al. (2017). Dynamic transcriptomes identify biogenic amines and insect-like hormonal regulation for mediating reproduction in *Schistosoma japonicum*. *Nat. Commun.* 8, 14693. doi: 10.1038/ncomms14693

Wang, L., Wang, S., and Li, W. (2012). RSeQC: quality control of RNA-seq experiments. *Bioinformatics* 28, 2184–2185. doi: 10.1093/bioinformatics/bts356

Who, W. H. O. (2015). *Investing to overcome the global impact of neglected tropical diseases: third WHO report on neglected tropical diseases 2015*. Geneva, Switzerland: World Health Organization.

Wu, W., Wagner, E. K., Hao, Y., Rao, X., Dai, H., Han, J., et al. (2016). Tissue-specific co-expression of long non-coding and coding RNAs associated with breast cancer. *Sci Rep.* 6, 32731. doi: 10.1038/srep32731

Wucher, V., Legeai, F., Hédan, B., Rizk, G., Lagoutte, L., Leeb, T., et al. (2017). FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Res.* 45, e57–e57. doi: 10.1093/nar/gkw1306

Yang, D.-C., Kong, L., Wei, L., Hou, M., Kang, Y.-J., Meng, Y.-Q., et al. (2017). CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.* 45, W12–W16. doi: 10.1093/nar/gkx428

Zijian, Z. (2016). Long non-coding RNAs in Alzheimer's disease. *Curr. Top. Med. Chem.* 16, 511–519. doi: 10.2174/1568026615666150813142956

Zoni, A. C., Catalá, L., and Ault, S. K. (2016). Schistosomiasis prevalence and intensity of infection in Latin America and the Caribbean Countries, 1942-2014: a systematic review in the context of a regional elimination goal. *PLoS Negl. Trop. Dis.* 10, e0004493. doi: 10.1371/journal.pntd.0004493

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Maciel, Morales-Vicente, Silveira, Ribeiro, Olberg, Pires, Amaral and Verjovski-Almeida. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.