



Universidade Federal de Minas Gerais
Programa de Pós-Graduação em Bioinformática



*O Spliced leader trans-splicing no parasito *Schistosoma mansoni**

Mariana Boroni

Orientadora: Prof^ª. Dr^ª. Glória Regina Franco
Co-orientadora: Dr^ª. Marina de Moraes Mourão

Belo Horizonte, agosto de 2014

Mariana Lima Boroni Martins

***O spliced leader trans-splicing no
parasito *Schistosoma mansoni****

Tese de Doutorado apresentada ao
Programa de Doutorado em Bioinformática,
Departamento de Bioquímica e Imunologia
da UFMG como requisito parcial para a
obtenção do título de Doutor em
Bioinformática

Orientadora: Prof^a. Dr^a. Glória Regina Franco

Co-orientadora: Dr^a. Marina de Moraes
Mourão

Universidade Federal de Minas Gerais

Programa de Pós-Graduação em Bioinformática

Belo Horizonte, agosto de 2014



ATA DA DEFESA DE TESE

Mariana Lima Boroni Martins

47/2014
 entrada
 2º/2010
 CPF:
 013.105.816-90

Às oito horas do dia **08 de agosto de 2014**, reuniu-se, no Instituto de Ciências Biológicas da UFMG, a Comissão Examinadora de Tese, indicada pelo Colegiado de Programa, para julgar, em exame final, o trabalho intitulado: "**O spliced leader trans-splicing no parasito Schistosoma mansoni**", requisito para obtenção do grau de Doutora em Bioinformática. Abrindo a sessão, a Presidente da Comissão, **Dra. Marina de Moraes Mourão (Co-orientadora)**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra à candidata, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa da candidata. Logo após, a Comissão se reuniu, sem a presença da candidata e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

| Prof./Pesq. | Instituição | Indicação |
|---|-----------------------------|-----------|
| Dra. Marina de Moraes Mourão (Co-orientadora) | Fiocruz | APROVADA |
| Dr. Benilton de Sa Carvalho | UNICAMP | Aprovada |
| Dr. Pedro A. F. Galante | Hospital Sírio Libanês - SP | APROVADA |
| Dr. Guilherme Correa de Oliveira | FIOCRUZ | APROVADA |
| Dr. Vasco Ariston de Carvalho Azevedo | UFMG | APROVADA |

Pelas indicações, a candidata foi considerada: APROVADA
 O resultado final foi comunicado publicamente à candidata pela Presidente da Comissão. Nada mais havendo a tratar, a Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.
Belo Horizonte, 08 de agosto de 2014.

Dra. Marina de Moraes Mourão (Co-orientadora) Marina Mourão
 Dr. Benilton de Sa Carvalho Benilton de Sa Carvalho
 Dr. Pedro A. F. Galante Pedro A. F. Galante
 Dr. Guilherme Correa de Oliveira Guilherme Correa de Oliveira
 Dr. Vasco Ariston de Carvalho Azevedo Vasco Ariston de Carvalho Azevedo

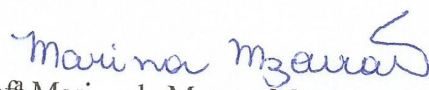
Dr. Vasco Ariston de C. Azevedo
 Prof. Titular e Coordenador do Programa de Pós Graduação em Bioinformática
 ICB/UFMG

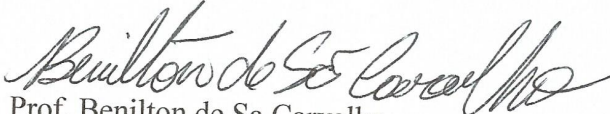



"O spliced leader trans-splicing no parasito Schistosoma mansoni"


Mariana Lima Boroni Martins

Tese aprovada pela banca examinadora constituída pelos Professores:


Profª Marina de Moraes Mourão (Co-orientadora)
Fiocruz


Prof. Benilton de Sa Carvalho
UNICAMP


Prof. Pedro A. F. Galante
Hospital Sírio Libanês - SP


Prof. Guilherme Correa de Oliveira
FIOCRUZ


Prof. Vasco Ariston de Carvalho Azevedo
UFMG

Belo Horizonte, 08 de agosto de 2014.


Dr. Vasco Ariston de C. Azevedo
Prof. Titular e Coordenador do Programa de
Pós Graduação em Bioinformática
ICB/UFMG

AGRADECIMENTOS

Ao apoio das pessoas que mais amo e prezo e que são fundamentais na minha vida: meu pai **Eduardo**, minha mãe **Denise** e meus irmãos **Natália** e **Matheus**.

À minha orientadora, Profa. **Glória Franco**, pela oportunidade, pela confiança e toda a orientação que muito contribuiu para o meu crescimento profissional e pessoal. Ainda quero agradecer pelo carinho e amizade ao longo desses quatro anos e por ser também um exemplo a ser seguido. Obrigada por todas valiosas lições.

À minha co-orientadora, Dra. **Marina Mourão** por toda ajuda e orientação durante a condução desse trabalho. Obrigada pela paciência e carinho.

Ao Prof. **José Marcos Ribeiro** e toda a sua equipe, por toda atenção, gentileza e contribuição nos experimentos de anotação funcional. Sem a sua ajuda, esse trabalho não teria existido.

À amiga **Neuzinha Antunes**, por toda preocupação em sempre nos proporcionar o melhor ambiente de trabalho e por nos acolher tão bem, além é claro do apoio técnico excepcional.

À Profa. **Andrea Macedo** e ao Prof. **Carlos Renato**, pelas valiosas contribuições científicas.

Ao **Juliano, Rennan e Flávio**, pela disponibilidade e ajuda no preparo das bibliotecas para sequenciamento.

À amiga **Sílvia Dias**, por todo auxílio no preparo do material biológico e aos valiosos ensinamentos de vida.

Aos estudantes de iniciação científica, **André Reis, Núbia Monteiro, Dani de Laett e Carol Castro** que me mostraram que para ser uma pessoa respeitada, não é necessário se impor, e sim, saber ser o exemplo.

Aos meus amigos **Carol Reis, Terciane, Mika, Pedrão, Mari Cota, Mari Kunrath, Elizângela, Carol Furtado, Ferdi, Mari Rocha e Livinha** pelos valiosos conselhos e por sempre estarem presentes na minha vida.

Aos meus queridos amigos e companheiros do **Laboratório de Genética Bioquímica**, pelo apoio, incentivo, ajuda nos experimentos e principalmente pela convivência diária tão feliz.

Ao **Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq** pela concessão da bolsa de estudo de doutorado, à **Fundação de Amparo à Pesquisa do Estado de Minas Gerais - FAPEMIG**, pelo financiamento do projeto e à agência de fomento internacional **Burroughs Wellcome Fund - BWF** (Schistosome Toolbox Collaborative Research Travel Award, que permitiu minha ida ao NIH).

SUMÁRIO

| | |
|---|------------|
| Agradecimentos | I |
| Resumo | V |
| Abstract | VI |
| Lista de Abreviaturas | VII |
| Índice de Ilustrações | IX |
| Índice de Tabelas | 1 |
| 1. Introdução | 2 |
| 1.1. A esquistossomose | 2 |
| 1.2. O ciclo de vida do <i>S. mansoni</i> | 3 |
| 1.3. O genoma do <i>S. mansoni</i> | 4 |
| 1.4. Processamento de transcritos por <i>cis -splicing</i> e <i>spliced leader trans-splicing</i> | 6 |
| 1.5. O <i>SLTS</i> em <i>S. mansoni</i> | 11 |
| 1.6. Sequenciamento de transcritos por RNA-Seq | 14 |
| 2. Objetivos | 16 |
| 3. Material e Métodos | 17 |
| 3.1. Conjuntos de dados utilizados nesse trabalho | 17 |
| 3.2. Obtenção do Material Biológico..... | 18 |
| 3.3. Extração dos RNAs | 18 |
| 3.4. Preparo das bibliotecas de cDNAs e sequenciamento no equipamento Illumina HiSeq 2000 | 19 |
| 3.5. Preparo das bibliotecas de cDNA e sequenciamento no equipamento Ion Torrent PGM..... | 20 |
| 3.6. Busca por dados de RNA-Seq de <i>S. mansoni</i> em bancos de dados..... | 22 |
| 3.6.1. Identificação da sequência do SL no conjunto de dados Total RNA-Seq | 22 |
| 3.7. <i>Pipeline</i> para análise dos dados de RNA-Seq gerados na plataforma Illumina HiSeq 2000 | 23 |
| 3.7.1. Análise de qualidade , mapeamento e contagem das reads..... | 23 |
| 3.8. <i>Pipeline</i> para análise dos dados de RNA-Seq gerados na plataforma Ion Torrent..... | 26 |
| 3.8.1. Identificação da sequência do SL no conjunto de dados SL Enriched.. | 26 |
| 3.8.2. Tratamento de qualidade das <i>reads</i> | 27 |

| | | |
|-----------|--|------------|
| 3.8.3. | Mapeamento e contagem das reads..... | 27 |
| 3.9. | processamento dos dados..... | 30 |
| 3.10. | Anotação e análise funcional das sequências..... | 30 |
| 3.11. | Identificação das vias metabólicas onde atuam as proteínas codificadas pelos transcritos processados SLTS..... | 33 |
| 3.12. | Identificação de transcritos policistrônicos em <i>S. mansoni</i> | 33 |
| 3.13. | Identificação no genoma de genes que contêm inserção da sequência do SL | 34 |
| 3.14. | Análise dos sítios aceptores de SL..... | 34 |
| 3.15. | Análise de sinais de <i>splicing</i> | 34 |
| 4. | Resultados e Discussão..... | 36 |
| 4.1. | Obtenção do RNA total..... | 36 |
| 4.2. | Obtenção das bibliotecas de cDNA enriquecidas em transcritos processados por SLTS..... | 37 |
| 4.3. | Obtenção das reads na plataforma Illumina HiSeq 2000..... | 38 |
| 4.4. | Obtenção das <i>reads</i> na plataforma Ion Torrent PGM™ System..... | 38 |
| 4.5. | Avaliação da qualidade das <i>reads</i> geradas..... | 38 |
| 4.6. | Identificação e remoção do SL e tratamento de qualidade nas <i>reads</i> geradas na plataforma de Ion Torrent PGM™ System..... | 51 |
| 4.7. | Obtenção das <i>reads</i> em bancos de dados públicos..... | 51 |
| 4.8. | Mapeamento das <i>reads</i> no genoma de referência..... | 55 |
| 4.9. | Comparação entre os três métodos utilizados para obtenção de sequências processadas por SLTS..... | 59 |
| 4.10. | O mecanismo de SLTS no parasito <i>S. mansoni</i> | 62 |
| 4.10.1. | Genes identificados: comparação entre as réplicas biológicas e os diferentes tipos de conjunto de dados..... | 62 |
| 4.10.2. | Identificação de sequências do SL no genoma de <i>S. mansoni</i> | 66 |
| 4.10.3. | Identificação de transcritos policistrônicos em <i>S. mansoni</i> | 68 |
| 4.10.4. | Análise funcional dos transcritos processados por SLTS..... | 79 |
| 4.10.1. | Características dos genes processados por SLTS..... | 82 |
| 4.10.2. | Competição entre SLTS e <i>cis-splicing</i> | 88 |
| 4.11. | O mecanismo de SLTS em diferentes fases do ciclo de vida do parasito <i>S. mansoni</i> | 98 |
| 4.11.1. | Genes identificados: comparação entre as réplicas biológicas e os diferentes tipos de conjunto de dados..... | 98 |
| 4.11.2. | Expressão diferencial de genes processados por SLTS entre diferentes fases do parasito..... | 102 |
| 4.11.1. | Classes gênicas funcionais diferencialmente representadas entre as fases | 103 |
| 4.11.2. | Categorias gênicas e análises de clusterização..... | 104 |
| 4.11.3. | SLTS alternativo entre os diferentes estágios de desenvolvimento do parasito. | 109 |
| 5. | Conclusões..... | 113 |

| | |
|--|------------|
| 6. Referências Bibliográficas..... | 116 |
| 7. Anexo 1 : Landscape of the spliced leader trans-splicing mechanism in <i>Schistosoma mansoni</i> | 123 |
| 8. Anexo 2 : The spliced leader trans-splicing mechanism in different organisms: molecular details and possible biological roles..... | 124 |
| 9. Anexo 3 : Evaluation of the <i>Schistosoma mansoni</i> Y-box-binding protein (SMYB1) potential as a vaccine candidate against schistosomiasis..... | 125 |

RESUMO

O *spliced leader trans-splicing* (SLTS) é um processo de maturação do pré-mRNA que ocorre em diversos organismos, incluindo o parasito trematódeo *Schistosoma mansoni*. Várias funções têm sido atribuídas ao SLTS, mas, no caso específico de *S. mansoni*, a sua importância como um mecanismo de regulação pós-transcricional ainda não foi determinada. Usando três diferentes estratégias para estudar transcritos que possuem a sequência do SL (*spliced leader*), geramos um conjunto extenso de dados compreendendo uma vasta gama de genes processados por SLTS. A partir dos resultados obtidos neste estudo, estimamos que 77% dos genes codificadores de proteínas anotados na 5ª versão do genoma de *S. mansoni* podem sofrer SLTS em alguma fase do ciclo de vida do parasito. Os níveis de expressão dos genes identificados variam amplamente e genes processados por SLTS foram relacionados a diversas classes funcionais. Nossos resultados indicam que o mecanismo de SLTS não parece estar particularmente enviesado para um conjunto específico de genes, caracterizando-o como um mecanismo ubíquo. Nossa análise revelou uma extensa heterogeneidade de sítios aceptores de SLTS ocorrendo em outrons e íntrons de vários genes, assim como características que distinguem íntrons alvos de *cis-splicing* daqueles alvos de *trans-splicing*. Vimos que os sítios alvos de SLTS podem ser diferencialmente utilizados durante os distintos estágios do ciclo de vida do parasito, influenciando o repertório e possivelmente os níveis de expressão de proteínas. Em conjunto, nossos dados mostram a importância do mecanismo de SLTS na modulação da expressão gênica do parasito associada ao seu desenvolvimento e adaptação a diferentes condições ambientais.

ABSTRACT

Spliced leader dependent *trans-splicing* (SLTS) is a pre-RNA maturation process that occurs in a diverse array of organisms, including the trematode parasite *Schistosoma mansoni*. SLTS has several functional roles assigned to it, but its importance as a post-transcriptional regulatory mechanism on *S. mansoni* is yet to be determined. Using three different strategies for studying transcripts harboring the SL (spliced leader) sequence we generated a broad dataset of SLTS genes. From the results presented herein we estimate that 77% of protein-coding genes annotated in the *S. mansoni* reference genome (5th version) may undergo SLTS at some stage of the parasite life cycle. The expression levels of the identified trans-spliced genes span several orders of magnitude and SL-containing reads mapped to genes over a wide spectrum of functional classes. Our results indicate that SLTS is not particularly biased to a specific set of genes, characterizing it as a ubiquitous mechanism. Our analysis revealed an extensive heterogeneity of SL acceptor sites occurring in outtrons and introns of several genes, in addition to attributes that can distinguish cis-spliced from trans-spliced introns. We showed that the SLTS target sites can be differentially regulated during the distinct life stages, influencing the protein repertoires and possibly the expression levels of different proteins. Together, our data show the importance of the SLTS mechanism in the gene expression modulation associated with the parasite development and its adaptation to different environmental conditions.

LISTA DE ABREVIATURAS

BH - *Benjamini-Hochberg*

BLAT - *BLAST-like Alignment Tool*

BLAST - *Basic Local Alignment Search Tool*

cDNA - *Ácido Desoxirribonucleico complementar*

DEG - *Genes Diferencialmente Expressos (Differentially Expressed Genes)*

ESTs - *Etiquetas de Sequências Expressas (Expressed Sequence Tags)*

FDR - *Fração de Falsos Positivos (False Discovery Rate)*

GED - *Expressão gênica dirigindo os eventos de SLTS*

GO - *Gene Ontology*

HMG-CoA - *3-hidróxi-3-metilglutaril-Coenzima A*

hnRNPs - *Partículas ribonucleoproteicas nucleares heterogeneas (Heterogeneous Nuclear Ribonucleoproteins)*

KEGG - *Kyoto Encyclopedia of Genes and Genomes*

KS - *Kolmogorov-Smirnov*

mRNA - *Ácido Ribonucleico mensageiro*

NCBI - *National Center for Biotechnology Information*

NGS - *Sequenciamento de Nova Geração (Next Generation Sequencing)*

nt - *Nucleotídeos*

ORFs - *Fases de leitura aberta (Open Reading Frames)*

PCA - *Análise de Componentes Principais (Principal Component Analysis)*

PCC - *Coeficiente de Correlação de Pearson (Pearson Correlation Coefficient)*

PCR - *Reação em Cadeia da Polimerase (Polymerase Chain Reaction)*

RNA-Seq - *Sequenciamento em massa de cDNA (RNA Sequencing)*

RT-PCR - *PCR quantitativo em tempo real*

rRNA - *Ácido Ribonucleico ribossômico*

SL - Sequência líder de *splicing* (*Spliced Leader*)

SLTS - *Spliced Leader Trans-Splicing*

snRNP - pequena partícula ribonucleoprotéica nuclear (*Small Nuclear Ribonucleic Particle*)

SRA - *Sequence Read Archive*

SVM - Máquina de Vetores de Suporte (*Support Vector Machine*)

tRNA - Ácido Ribonucleico transportador

TSD - *Trans-Splicing* dirigindo os eventos de SLTS

UbCRBP - proteína de ligação ao complexo ubiquinol-citocromo C redutase (*Ubiquinol-Cytochrome C Reductase Complex Binding Protein*)

uORF - *Upstream Open Reading Frame*

VST - Transformação da Estabilização da Variância (*Variance-Stabilizing Transformation*)

ÍNDICE DE ILUSTRAÇÕES

| | |
|--|----|
| Figura 1 - Distribuição global estimada da esquistosomose para o ano de 2010..... | 3 |
| Figura 2 - Precursores e produtos do <i>Spliced leader trans-splicing</i> | 9 |
| Figura 3 - Sequência e estrutura do SL RNA de <i>S. mansoni</i> | 13 |
| Figura 4 – Representação da estratégia experimental utilizada em cada uma das três abordagens distintas para identificação dos transcritos processados por SLTS em <i>S. mansoni</i> | 17 |
| Figura 5 – Detecção de transcritos policistrônicos..... | 34 |
| Figura 6 – Perfil de qualidade dos RNAs totais extraídos de diferentes fases do ciclo de vida do parasito <i>S. mansoni</i> | 36 |
| Figura 7 – Perfil eletroforético dos cdnas gerados à partir de transcritos que sofrem processamento por slts de diferentes fases de <i>S. mansoni</i> | 37 |
| Figura 8 - Valores de qualidade por base. | 40 |
| Figura 9 - Valor de qualidade média por sequência..... | 42 |
| Figura 10 - Conteúdo de base ao longo da sequência..... | 44 |
| Figura 11 - Conteúdo GC por base | 46 |
| Figura 12 - Conteúdo GC por sequência..... | 48 |
| Figura 13 - Sequências duplicadas. | 50 |
| Figura 14 - Diagrama de Venn representando os genes processados por SLTS identificados por métodos diferentes..... | 60 |
| Figura 15 – Comparação entre as duas réplicas biológicas das bibliotecas SL Trapping. | 62 |
| Figura 16 – Comparação entre os conjuntos de dados SL Trapping e RNA-Seq Filtered. | 64 |
| Figura 17 – Sítios de inserção da sequência do SL nos diferentes tipos de biblioteca... | 66 |
| Figura 18 - SL concatenado a um gene no genoma de <i>S. mansoni</i> | 68 |
| Figura 19 - Inserção do SL no dicistron enolase e UbCRBP | 69 |
| Figura 20 – Transcrito policistrônico em <i>S. mansoni</i> | 70 |

| | |
|---|-----|
| Figura 21 – Classificação funcional dos transcritos de <i>S. mansoni</i> | 81 |
| Figura 22 - Vias metabólicas cujas proteínas envolvidas são derivadas de transcritos processadas por SL <i>trans-splicing</i> | 84 |
| Figura 23 – Expressão gênica e frequência de SLTS. | 85 |
| Figura 24 – Origem cromossômica, isoformas originadas de <i>splicing</i> alternativo e número de exons de transcritos processados por SLTS. | 88 |
| Figura 25 - Sítios de SLTS alternativo. | 89 |
| Figura 26 – Comparação entre os sítios de <i>trans-splicing</i> ocorrendo em outrons e íntrons. | 90 |
| Figura 27 – Posição da inserção da sequência do SL nos transcritos. | 92 |
| Figura 28 – Motivos dos sítios de <i>splicing</i> em íntrons..... | 93 |
| Figura 29 – Tamanho dos íntrons e exons dos transcritos de acordo com sua classificação e sua posição relativa no transcrito..... | 94 |
| Figura 30 – Características dos Sítios de <i>splicing</i> subdivididas de acordo com a classificação do íntron e sua posição relativa no transcrito..... | 95 |
| Figura 31 - Características dos Sítios de ramificação subdividida de acordo com as características do íntron e sua posição relativa no transcrito..... | 97 |
| Figura 32 - Comparação entre as duas réplicas biológicas das bibliotecas SL Enriched de cada fase do ciclo de vida..... | 100 |
| Figura 33- Comparação entre as amostras. | 101 |
| Figura 34 - Diagrama de Venn representando os transcritos identificados nas bibliotecas das diferentes fases do ciclo de vida de <i>S. mansoni</i> | 102 |
| Figura 35 – Genes processados por SLTS diferencialmente expressos entre as diferentes fases do ciclo de vida de <i>S. mansoni</i> | 103 |
| Figura 36 - Classificação funcional dos dos genes diferencialmente expressos entre as fases de <i>S. mansoni</i> | 107 |
| Figura 37 – Heatmap mostrando a clusterização hierárquica entre as amostras de diferentes estágios de desenvolvimento e os genes..... | 108 |
| Figura 38 - Classificação funcional dos dos genes agrupados em diferentes clusteres de acordo com o seu padrão de expressão nos estágios de desenvolvimento do parasito. | 109 |

Figura 39 – Entrada alternativa da sequência do SL em genes processados por SLTS em diferentes estágios de desenvolvimento de *S. mansoni*.112

ÍNDICE DE TABELAS

| | |
|--|-----|
| Tabela 1 - Tipos de BLAST, bases de dados e parâmetros utilizados para anotação dos transcritos. | 31 |
| Tabela 2 – Dados gerados na plataforma Illumina Hiseq 2000. | 38 |
| Tabela 3 – Dados gerados na plataforma Ion Torrent PGM™ System. | 38 |
| Tabela 4 – <i>Reads</i> das bibliotecas SL Enriched sequenciadas na plataforma Ion Torrent PGM™ System após identificação e remoção da sequência do SL e filtro de qualidade. | 53 |
| Tabela 5 - Dados de RNA-Seq de estudos de transcriptômica do parasito <i>S. mansoni</i> depositados no repositório público SRA do NCBI. | 54 |
| Tabela 6 - Estatística da identificação da sequência do SL nas <i>reads</i> das bibliotecas RNA-Seq advindas do banco de dados público SRA. | 55 |
| Tabela 7 - Estatísticas do alinhamento das <i>reads</i> geradas a partir das diferentes bibliotecas no genoma de referência de <i>S. mansoni</i> | 56 |
| Tabela 8 – Genes concatenados com a sequência do SL..... | 67 |
| Tabela 9 – Transcritos Policistrônicos em <i>S. mansoni</i> Error! Bookmark not defined. | |
| Tabela 10- Classes enriquecidas entre os clusters gênicos. | 106 |

1. INTRODUÇÃO

1.1. A ESQUISTOSSOMOSE

A esquistossomose é uma doença parasitária causada por organismos do filo platelminto, classe trematodea, ordem digenea, família schistosomatidae e gênero *Schistosoma*, que compreendem várias espécies, dentre elas, o *Schistosoma mansoni*, endêmico em países da África e América Latina.

A esquistossomose é uma parasitose considerada a segunda doença tropical mais prevalente e mais devastadora socioeconomicamente, depois apenas da malária. Estimativas da Organização Mundial da Saúde (*World Health Organization*, WHO, 2013) indicam que existem 243 milhões de pessoas acometidas pela esquistossomose e que mais de 780 milhões de pessoas vivem em áreas com risco de infecção em 78 países, principalmente aqueles considerados subdesenvolvidos ou em desenvolvimento (ENGELS *et al.*, 2002), onde as condições de higiene e saneamento básico são precárias. A distribuição global (Figura 1) e a taxa de infecção da esquistossomose têm mudado pouco nos últimos 20 anos (WHO, 2013).

Atualmente, o controle da esquistossomose tem sido focado especialmente na quimioterapia utilizando o praziquantel e o uso em larga escala desse fármaco tem gerado grandes discussões acerca da resistência à droga (MANN *et al.*, 2010; STANDLEY *et al.*, 2010). Esses dados reforçam a importância do estudo científico das espécies causadoras dessa doença, como o *S. mansoni*, agente causador desta doença no Brasil.

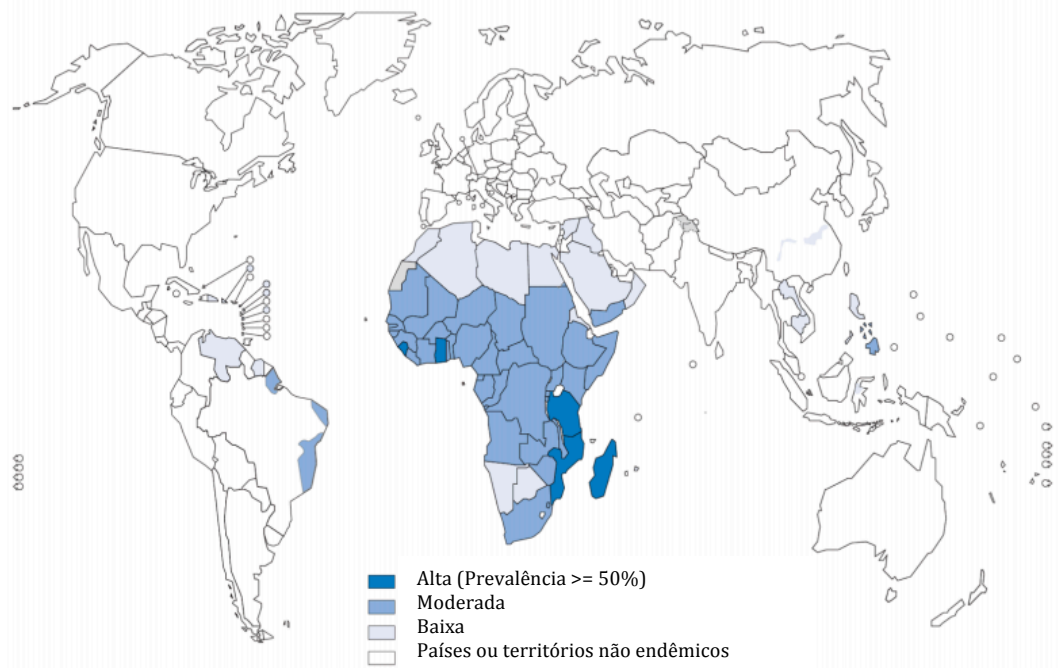


FIGURA 1 - DISTRIBUIÇÃO GLOBAL ESTIMADA DA ESQUISTOSOMOSE PARA O ANO DE 2010. [reproduzida de (WHO, 2013)]. Diferentes tonalidades em azul representam a prevalência da infecção em indivíduos na população.

1.2. O CICLO DE VIDA DO *S. MANSONI*

O *S. mansoni* apresenta um ciclo de vida complexo, envolvendo a sobrevivência do parasito em ambientes bastante distintos. Diferentemente de outros platelmintos, o *S. mansoni* apresenta dimorfismo sexual quando adulto. Os vermes adultos pareados (macho e fêmea) residem no sistema porta hepático (PESSOA; MARTINS, 1982). Para oviposição, os casais migram para as veias mesentéricas inferiores, onde cada fêmea deposita de 200 a 300 ovos por dia próximo à parede intestinal. Esses ovos atingem o lúmen do intestino, sendo eliminados conjuntamente com as fezes e continuando o ciclo de vida do parasito, mas também podem ganhar as veias mesentéricas, ou os capilares do sistema porta, nos quais se alojam podendo causar uma patologia severa, incluindo uma resposta inflamatória granulomatosa do hospedeiro e fibrose. Em contato com a água fresca, os ovos eclodem e liberam larvas de vida livre, os miracídios, que

infectam caramujos (*Biomphalaria spp.*) Nestes hospedeiros, os parasitos se reproduzem assexuadamente por duas gerações de esporocistos e posteriormente são transformados na forma infectiva, as cercárias, que são liberadas na água. As cercárias infectam o hospedeiro definitivo, comumente o homem, através da penetração ativa pela pele e transformam-se em esquistossômulos. Após vários dias, os parasitos saem do tecido cutâneo através dos vasos do sistema sanguíneo ou linfático e viajam primeiramente até os pulmões, onde se desenvolvem antes de atingir o sistema porta-hepático. Somente quando atingem este último é que os vermes começam a se alimentar de sangue e atingem a maturidade sexual (DAVIS, 2002; PESSOA; MARTINS, 1982).

Essa grande adaptabilidade do parasito a diferentes ambientes reflete drásticas mudanças morfológicas no curso do seu ciclo de vida e, para tais mudanças, espera-se que haja um controle da expressão gênica igualmente complexo.

1.3. O GENOMA DO *S. MANSONI*

O cariótipo de *S. mansoni* compreende 7 pares de autossomos e 1 par de cromossomos sexuais (fêmea=ZW, macho=ZZ) (BERRIMAN *et al.*, 2009). O início da análise de genes em larga escala de *S. mansoni* se deu em meados da década de 1990, com as primeiras publicações de Etiquetas de Sequências Espressas (*Expressed Sequence Tags*, ESTs) do parasito por Franco e colaboradores (FRANCO *et al.*, 1995, 1997). Mais tarde, outros trabalhos contribuíram para a descoberta de novos genes em *S. mansoni* (FRANCO *et al.*, 2000; SHABAAN *et al.*, 2003; VERJOVSK-ALMEIDA *et al.*, 2003). Entretanto, o sucesso no uso de ESTs para descobrir novos genes é limitado principalmente pela pouca representatividade de

genes de baixa expressão nas bibliotecas de cDNA e pela utilização de sequenciadores capilares (método de Sanger), que é, hoje em dia, um método caro, laborioso e ultrapassado, devido ao advento das técnicas de sequenciamento de última geração.

O primeiro rascunho do genoma de *S. mansoni* foi publicado em 2009 com mais de 360 milhões de bases. O genoma foi montado em 19.022 *scaffolds* e anotado com 11.809 genes correspondendo a 13.197 transcritos (BERRIMAN *et al.*, 2009). Quase metade do genoma (45%) é composto de regiões contendo elementos repetitivos. Mais recentemente, a montagem do genoma foi melhorada sistematicamente, utilizando os dados originais do rascunho do genoma, novos dados obtidos por sequenciamento de Sanger e também sequenciamento de nova geração. A nova montagem do genoma resultou em uma versão menos fragmentada com apenas 885 *scaffolds* e 364,5 milhões de bases, sendo 86% dessas mapeadas nos cromossomos. A estrutura de 45% dos genes até então descobertos foi refinada utilizando-se dados de sequenciamento de RNA em larga escala (RNA-Seq) e foram identificados novos genes e variantes de *splicing* (PROTASIO *et al.*, 2012).

Atualmente, devido ao enorme progresso na área genômica e a crescente quantidade de informação gerada com os novos sequenciadores de larga escala, dispomos de uma quantidade enorme de dados. Entretanto, ainda existe um déficit nas caracterizações das funções gênicas. O recente avanço na análise do genoma do *S. mansoni* tem possibilitado uma maior compreensão da complexa biologia do parasito e a identificação de vias metabólicas e de sinalização que podem representar pontos chave para intervenção e controle da doença.

1.4. PROCESSAMENTO DE TRANSCRITOS POR *CIS* –*SPLICING* E *SPLICED LEADER TRANS-SPLICING*

Dentre as numerosas e complexas formas de regulação gênica pós-transcricional, sabe-se que em *S. mansoni*, além do *cis-splicing* de transcritos recém sintetizados, ocorre também o processamento por *spliced leader trans-splicing* (SLTS). Ambos os mecanismos utilizam a maquinaria básica do spliceossomo (DENKER; ZUCKERMAN, 2002).

Logo durante a transcrição, o pré-mRNA é complexado com uma grande variedade de partículas ribonucleoproteicas nucleares heterogeneas formando uma grande estrutura conhecida como partículas ribonucleoproteicas mensageiras (mRNPs). Essas partículas complexadas ao RNA nascente, denominadas hnRNPs, atuam no destino do mRNA, desde a sua transcrição e processamento no núcleo até sua tradução e degradação no citoplasma (MÜLLER-MCNICOLL; NEUGEBAUER, 2013).

A montagem do spliceossomo ocorre durante a síntese do RNA. A snRNP U1 (um dos componentes do spliceossomo) liga-se aos sítios doadores 5' (GU) de *splicing* do transcrito nascente e com a ajuda de hnRNPs, que empacotam o RNA nascente aproximando regiões distantes, auxilia no direcionamento correto do *splicing* no pre-mRNA, juntamente com a U2 snRNP. A U2 snRNP é responsável pelo reconhecimento e ligação ao trato de polipirimidina e ao sítio de ramificação, ou *branch point* (A) assim como ao sítio 3' acceptor de *splicing* (AG) (revisado em (BENTLEY, 2014; MATERA; WANG, 2014; MÜLLER-MCNICOLL; NEUGEBAUER, 2013)). O dinucleotídeo AG no sítio acceptor está geralmente associado a um trato de polipirimidinas, sendo estas as características mais proeminentes e mais

altamente conservadas nos pré-mRNAs substratos de *splicing*. Em mamíferos, os tratos de polipirimidina presentes nos íntrons possuem uma função dupla: eles atuam na identificação do sítio de ramificação, assim como na seleção do sítio de *splicing* (REED, 1989).

A interação das snRNPs U1 e U2 com o transcrito nascente é mediada pelo domínio carboxi-terminal da RNA-polimerase II, e U1 e U2 interagem entre si para formar o pré-spliceossomo, num processo dependente de helicases. Em um passo subsequente, um complexo formado por três snRNPs (U4-U6•U5) é recrutado ao pré-spliceossomo, que é rearranjado com o auxílio de helicases de RNA para formar um complexo ativo cataliticamente. Como consequência, a partícula U4 e U1 são liberadas e o complexo atua na primeira etapa do *splicing*, gerando um exon 5' livre e uma estrutura intermediária de laço contendo o íntron ligado ao exon 3'. Esse complexo é novamente rearranjado de forma a catalisar a segunda etapa de trans-esterificação do *splicing*, resultando num complexo pós-spliceossômico e na união dos exons e remoção do íntron na forma de laço. Finalmente as snRNPs U2, U5 e U6 são liberadas da molécula de mRNP e recicladas em rodadas adicionais de *splicing* (revisado em (MATERA; WANG, 2014)).

A escolha do sítio de *splicing* é regulada por diversos fatores que se associam tanto aos exons quanto aos íntrons do pré-mRNA e que podem tanto promover, quanto inibir o reconhecimento de sítios de *splicing* próximos. A associação das hnRNPs ao pré-mRNA é muito importante para esconder sítios de *splicing* fortes ou expor sítios de *splicing* fracos, favorecendo a remoção ou inclusão de exons alternativos pelo spliceossomo, respectivamente (revisado em (MATERA; WANG, 2014; MÜLLER-MCNICOLL; NEUGEBAUER, 2013)).

Diferentemente do que ocorre no *cis-splicing*, onde exons de um mesmo transcrito são unidos para formar um mRNA maduro e funcional, no SLTS uma sequência identificada como *spliced leader* (SL) é doada para alguns pré-mRNAs receptores, formando a região 5' terminal de mRNAs maduros (Figura 2). As sequências de SL originam-se de pequenos RNAs não codificadores (SL RNAs) de 40 a 140 nt, não poliadenilados, que têm um sítio doador de *splicing* e um cap 5' modificado (2,2,7-trimetilguanosina - TMG) (revisado em (LASDA; BLUMENTHAL, 2011)). O mecanismo de processamento por SLTS é muito parecido com o *cis-splicing*, pois utiliza a maquinaria básica do spliceossomo, o mesmo motivo no RNA para indicar os sítios de *splicing* (AG no sítioceptor e GU no sítio doador) e uma extensão de polipirimidina necessária para atrair a maquinaria de *splicing* (DENKER; ZUCKERMAN, 2002). A principal diferença identificada é que, no *trans-splicing*, o SL RNA interage com proteínas formando uma snRNP estruturalmente e funcionalmente relacionada aos U snRNAs (U1, U2, U4, U5 e U6) (MATERA; WANG, 2014; MICHAELI, 2011). Entretanto, no SLTS a participação da U1 não é obrigatória e a snRNP, formada a partir da associação do SL RNA com os fatores de *splicing*, é consumida durante o processo de SLTS, uma vez que a sequência do SL é doada da partícula para formar o RNA maduro (BRUZIK *et al.*, 1988; THOMAS; CONRAD; BLUMENTHAL, 1988). Dessa forma, todos os transcritos processados por SLTS se iniciam com a mesma sequência. Adicionalmente é formado um produto em forma de Y contendo o restante do SL RNA (íntron) unido por uma ligação 2',5'-fosfodiéster à adenosina do outtron (parte que foi removida do mRNA no processo de *trans-splicing*), que é análogo ao produto em forma de laço formado durante o *cis-splicing* (revisado em (BITAR *et al.*, 2013; BLUMENTHAL, 2005; LASDA; BLUMENTHAL, 2011; STOVER; KAYE; CAVALCANTI, 2006)) (Figura 2).

No mecanismo de SLTS, a interação entre os fatores que se ligam ao trato de polipirimidinas e ao sítio acceptor de *splicing* se mostram muito importantes, uma vez que a eliminação de um sítio acceptor nativo implica a ocorrência do *trans-splicing* no próximo sítio AG e acúmulo do intermediário na forma de Y (HUMMEL; GILLESPIE; SWINDLE, 2000). Ainda, em tripanossomatídeos, o trato de polipirimidina foi identificado como o principal fator para a acurácia do SLTS e o seu tamanho foi relacionado a sua eficácia (MATTHEWS; TSCHUDI; ULLU, 1994; SCHÜRCH *et al.*, 1994).

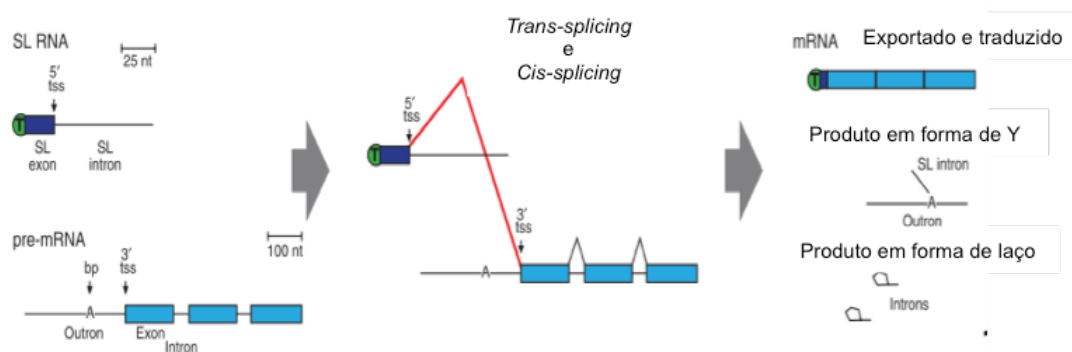


FIGURA 2 - PRECURSORES E PRODUTOS DO SPLICED LEADER TRANS-SPLICING. Painel à esquerda - Moléculas precursoras: SL RNA e pré-mRNA. A barra de escala indica o comprimento relativo das moléculas de RNA. Bp representa o sítio de de ramificação (*branch point*) e tss representa os sítios de SLTS (*trans-splicing site*). Painel do meio - O SL *trans-splicing* (linhas vermelhas) une o sítio de *trans-splicing* 5' localizado no SL RNA ao sítio 3' do *trans-splicing* localizado no pré-mRNA. Já o *cis-splicing* une os exons de um mesmo transcrito (linhas pretas). Painel à direita - Produtos resultantes de reações de *splicing*. O RNA maduro contém o exon SL e o TMG cap na extremidade 5' e pode ser exportado para o citoplasma para ser traduzido. O produto em forma de Y é o intron do SL RNA acoplado ao outtron do pré-mRNA. É análogo ao laço formado durante a remoção dos íntrons no *cis-splicing*. Estes produtos são rapidamente degradados. Caixas azuis representam os exons e as linhas pretas representam íntrons e outtrons. O círculo verde representa o TMG cap no exon do SL RNA. [reproduzida de Lasda e Blumenthal, 2011].

Diversas funções foram atribuídas ao SLTS, entre elas: prover o cap 5' aos transcritos da RNA polimerase I (em kinetoplastídeos), modificação necessária para a estabilidade, transporte e tradução do mRNA (revisado em (HASTINGS, 2005; LASDA; BLUMENTHAL, 2011; STOVER; KAYE; CAVALCANTI, 2006)); aumentar a flexibilidade da região 5' dos transcritos, podendo gerar vantagens adaptativas aos organismos (SAITO *et al.*, 2013); regular a expressão gênica em nível pós-transcricional ao substituir códons de iniciação dentro do primeiro exon

(que estejam ou não fora de fase de leitura) durante a remoção do outron. Finalmente, a função melhor caracterizada do SLTS, inicialmente descrita em tripanosomatídeos (HUANG; VAN DER PLOEG, 1991a), é a resolução de policistrons em moléculas monocistrônicas, cada uma atuando como um mRNA distinto que são, em seguida, traduzidas separadamente (DAVIS; HODGSON, 1997; EVANS; BLUMENTHAL, 2000; GIULIANO; BLAXTER, 2006; LEE; SOMMER, 2003). Estudos no nematóide *Caenorhabditis elegans* descrevem unidades policistrônicas de 2-8 genes transcritos a partir de um promotor comum, geralmente separados por espaçadores de ~ 100 pb, mas em casos raros, tão distantes quanto 1 - 2kb do outro (revisado em (BLUMENTHAL; GLEASON, 2003; BLUMENTHAL, 2004, 2005)). Embora as unidades policistrônicas existam em diversos organismos com SLTS, em muitos casos a maioria dos mRNAs processados por SLTS não resultam de transcritos policistrônicos (revisado em (LASDA; BLUMENTHAL, 2011)).

A perspectiva tradicional de que o SLTS normalmente ocorre no exon 5' terminal de um gene se baseia na idéia de que a presença de sinais de *cis-splicing*, tais como sequências doadoras *upstream*, perturbariam o processo de SLTS (CONRAD; LEA; BLUMENTHAL, 1995; CONRAD et al., 1991). No entanto, essa visão geral foi parcialmente revista por trabalhos recentes em *C. elegans* nos quais observou-se a ocorrência de SLTS em locais internos no transcrito, que não estão associados com a extremidade 5' destes. Adicionalmente, ensaios com um gene repórter em *T. brucei* identificaram atributos característicos do trato de polipirimidinas e ponto de ramificação associados com SLTS (SIEGEL; TAN; CROSS, 2005). Em *Ciona intestinalis* proporções substanciais de alvos de SLTS estão aparentemente em locais próximos ao respectivos sítio aceptores principais de SLTS (MATSUMOTO et al., 2010).

Até o momento, o SLTS foi identificado em um grupo relativamente pequeno e diverso de organismos, após ser descrito pela primeira vez em tripanosomatídeos (MURPHY; WATKINS; AGABIAN, 1986; SUTTON, 1986) e em seguida em *C. elegans* (KRAUSE; HIRSH, 1987). Em platelmintos, a primeira evidência de *trans-splicing* foi reportada em *S. mansoni* (RAJKOVIC *et al.*, 1990). O SLTS *trans-splicing* também já foi descrito em cnidários, cinetóforos, rotíferos, crustáceos, esponjas, cetognatas, tunicados (urocordados), dinoflagelados e euglenozóides. Por sua vez, Taxa nos quais o mecanismo de SLTS ainda não foi demonstrado incluem plantas, fungos, insetos e vertebrados (BITAR *et al.*, 2013; DERELLE *et al.*, 2010; DOURIS; TELFORD; AVEROF, 2010; LASDA; BLUMENTHAL, 2011).

A porcentagem dos genes que são sujeitos ao processamento por SLTS varia entre as diferentes espécies. Em tripanosomatídeos, por exemplo, 100% dos transcritos sofrem *trans-splicing* (NILSSON *et al.*, 2010). Já em *C. elegans* e *C. intestinalis* cerca de 70% (ALLEN *et al.*, 2011) e 58% (MATSUMOTO *et al.*, 2010) dos genes sofrem *trans-splicing*, respectivamente.

1.5. O SLTS EM *S. MANSONI*

Em *S. mansoni*, Rajkovic e colaboradores (1990) mostraram que o SL do parasito contém 36 nucleotídeos derivados de um RNA não poliadenilado de 90 nucleotídeos. O transcrito de 90 nucleotídeos não possui similaridade de sequência com outros organismos, sugerindo assim que o gene SL de *S. mansoni* seja espécie-específico. Foram identificadas 54 cópias do gene SL RNA, sendo 43 destas idênticas, seis variantes distintas e cinco pseudogenes (Figura 3) (COPELAND *et al.*, 2009).

Neste parasito, o SLTS aparentemente não está relacionado a genes específicos, localizações subcelulares e/ou fases do ciclo de vida (DAVIS; HARDWICK; TAVERNIER, 1995; DAVIS; HODGSON, 1997; MOURÃO et al., 2013). Alguns estudos de caso demonstraram locais de inserção alternativa para o SLTS no gene de 3-hidróxi-3-metilglutaril-Coenzima A (HMG-CoA) redutase e também no gene da proteína de ligação ao complexo ubiquinol-citocromo C redutase (UbcRBP) (MOURÃO *et al.*, 2013; RAJKOVIC *et al.*, 1990). Utilizando a técnica de RNA-Seq, Protasio e colaboradores (2012) identificaram 1.178 genes sujeitos ao processamento por SLTS em diferentes fases do ciclo de vida de *S. mansoni*, estimando-se que cerca de 10% dos genes sofrem esse tipo de processamento (DAVIS; HARDWICK; TAVERNIER, 1995; PROTASIO et al., 2012). Os autores observaram ainda, em alguns casos, um segundo sítio acceptor de SL, usualmente entre 20–50 nt de distância do primeiro sítio acceptor.

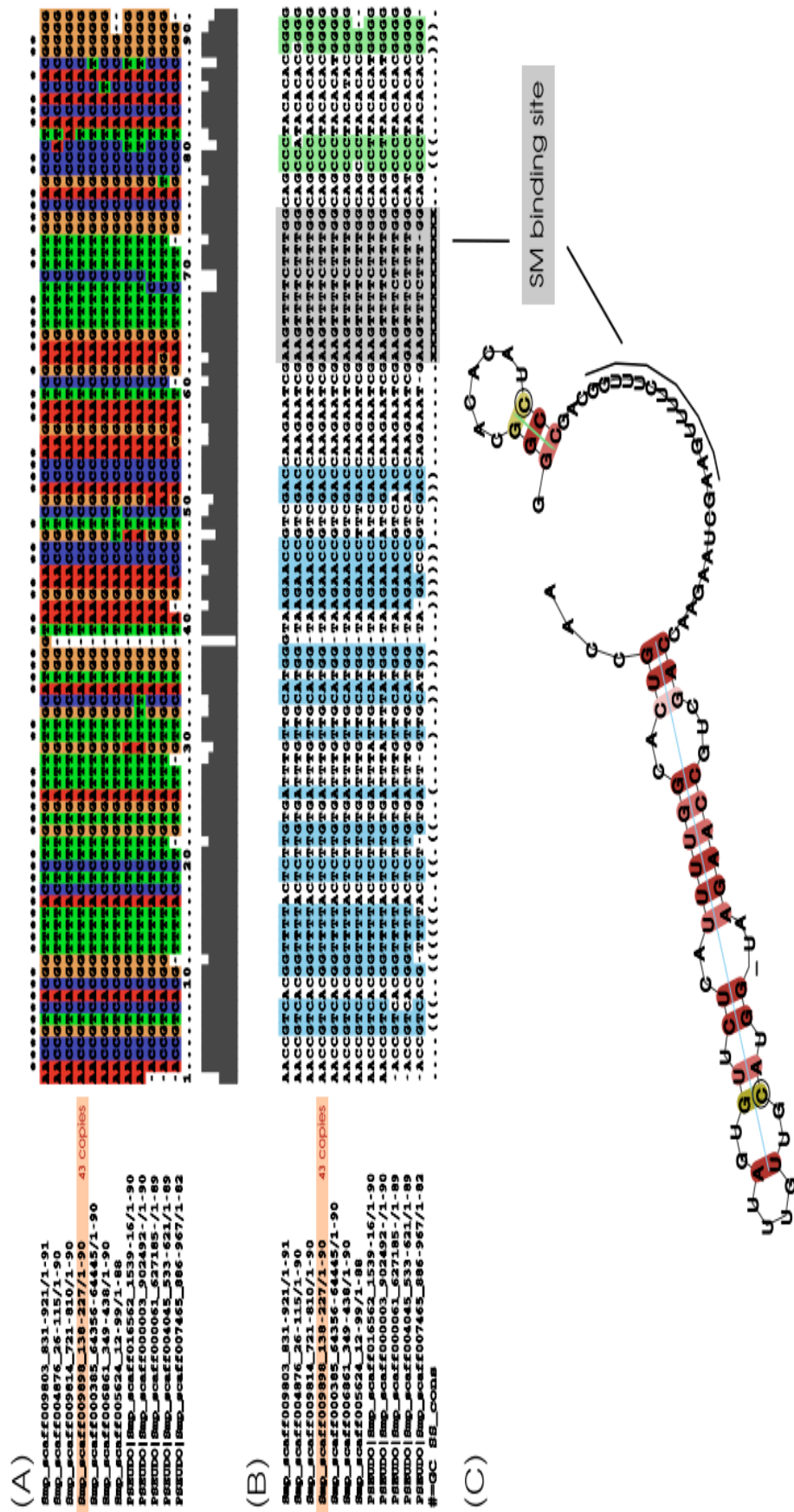


FIGURA 3 - SEQUÊNCIA E ESTRUTURA DO SL RNA DE *S. MANSONI*. A - alinhamentos gerados pelo Clustal mostrando os genes candidatos para SL RNAs e possíveis pseudogenes. B - alinhamentos no modo Emac Ralee, com os elementos estruturais assinalados em cores. A estrutura secundária consenso é representada na última linha. Regiões assinaladas em verde e em azul: regiões de nucleotídeos pareados. Região assinalada em cinza: sítio de ligação às *SM-binding proteins* (proteínas acessórias de *splicing*) C - estrutura secundária do SL RNA predita por RNAalifold [reproduzida de COPELAND *et al.*, 2009].

Davis e Hodgson (1997) introduziram o modelo UbCRBP / enolase como um possível transcrito policistônico resolvido por SLTS em *S. mansoni*. Protasio e colaboradores (2012) estenderam esse conjunto para 46 possíveis unidades policistrônicas, com distância intergênica de até 200 pb. Entretanto, todos estes estudos esboçam uma imagem limitada sobre o mecanismo de STLS neste parasito e, portanto, estão subestimando o verdadeiro impacto deste fenômeno.

1.6. SEQUENCIAMENTO DE TRANSCRITOS POR RNA-SEQ

Com o advento das novas tecnologias de sequenciamento, atualmente é possível avaliar a expressão gênica através do sequenciamento em massa de moléculas de cDNA (GARBER *et al.*, 2011; WANG; GERSTEIN; SNYDER, 2009). Este método, conhecido como RNA-Seq, possibilita o estudo abrangente de transcriptomas, com alta sensibilidade, economia de tempo e custo.

Em experimentos de RNA-Seq, os fragmentos de cDNA são sequenciados e mapeados nos genes correspondentes, idealmente, em posições únicas. Adequadamente normalizadas, as contagens de fragmentos podem ser usadas como uma medida da abundância relativa dos transcritos. Ainda, o RNA-Seq permite a detecção de mutações pontuais nos transcritos, identificação de fusão de transcritos, descoberta de novas classes de RNA e novos eventos de *splicing* alternativo, além da análise de expressão de alelos (COSTA *et al.*, 2010; FLINTOFT, 2008; MARIONI *et al.*, 2008; ROBERTS *et al.*, 2011; WANG; GERSTEIN; SNYDER, 2009; WILHELM; LANDRY, 2009). O grande sucesso das novas tecnologias de sequenciamento na transcriptômica se deve também ao fato de estas possibilitarem a superação de uma das maiores limitações dos projetos de sequenciamento cDNAs derivados de bibliotecas, que geram ESTs - a brusca

redução no número de sequências novas amostradas com o aumento na quantidade de informação sequenciada (GARBER *et al.*, 2011; WANG; GERSTEIN; SNYDER, 2009).

Assim, tendo em vista que o mecanismo de SLTS em *S. mansoni* ainda não foi bem caracterizado, estudos que visem a identificação de transcritos processados por SLTS em diferentes fases do ciclo de vida do parasito, especialmente utilizando o sequenciamento em massa do transcriptoma para uma análise em larga escala, podem auxiliar a determinar a importância desse processo na regulação pós-transcricional da expressão gênica.

2. OBJETIVOS

i. Objetivo geral:

Compreender as implicações biológicas do mecanismo de *SL trans-splicing* em *S. mansoni* através da identificação de transcritos que são processados por SLTS em diferentes fases do ciclo de vida do parasito utilizando-se a técnica de RNA-Seq.

ii. Objetivos específicos:

- Construção e sequenciamento de bibliotecas de RNA-Seq enriquecidas em transcritos que sofrem processamento por SLTS em diferentes fases do ciclo de vida do parasito *S. mansoni* (miracídio, esporocisto, cercária, esquistossômulo e vermes adultos);
- Identificação e classificação funcional dos genes processados por SLTS e busca por possíveis funções e processos biológicos diferencialmente representados em genes processados por SLTS;
- Avaliação dos sítios aceptores de *trans-splicing* e sua correlação com a ocorrência de *trans-splicing* alternativo;
- Comparação do perfil de expressão e processamento de genes sujeitos a SLTS entre diferentes fases do ciclo de vida do parasito.

3. MATERIAL E MÉTODOS

3.1. CONJUNTOS DE DADOS UTILIZADOS NESSE TRABALHO

No decorrer das seções seguintes descreveremos as três abordagens utilizadas neste trabalho para a análise de transcritos processados por SLTS em *S. mansoni* (Figura 4).

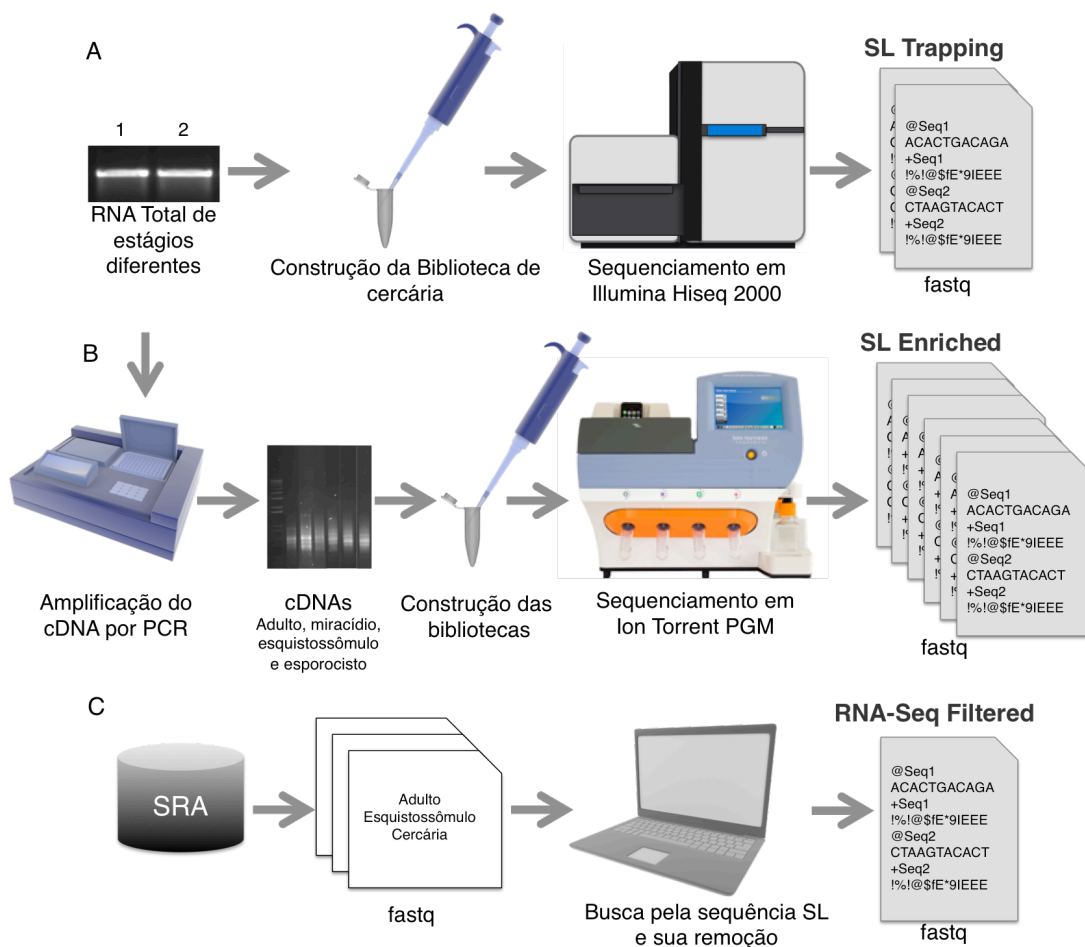


FIGURA 4 – REPRESENTAÇÃO DA ESTRATÉGIA EXPERIMENTAL UTILIZADA EM CADA UMA DAS TRÊS ABORDAGENS DISTINTAS PARA IDENTIFICAÇÃO DOS TRANSCRITOS PROCESSADOS POR SLTS EM *S. MANSONI*. A – RNA total de cercária foi utilizado para a produção de bibliotecas de cDNA para sequenciamento na plataforma Illumina Hiseq 2000 (RNA-Seq), utilizando uma metodologia modificada, na qual somente transcritos que se iniciam com a sequência do SL foram sequenciados, produzindo o conjunto de dados denominado SL Trapping. B – RNA total das fases esquistossômulo, esporocisto, miracídio e adulto foram utilizados para produzir cDNAs utilizando como iniciadores uma sequência complementar ao SL e outra ao adaptador presente na cauda de poli-A. Os cDNAs enriquecidos em transcritos processados por SLTS foram então utilizados para construção de bibliotecas de fragmentos para sequenciamento na plataforma Ion Torrent PGM, produzindo o conjunto de dados denominado SL Enriched. C – Reads contendo a sequência do SL foram obtidas a partir de dados de RNA-Seq depositados no repositório público SRA referentes a 3 fases do ciclo de vida de *S. mansoni* (vermes adultos, cercárias e esquistossômulos), para produção do conjunto de dados denominado RNA-Seq Filtered.

3.2. OBTENÇÃO DO MATERIAL BIOLÓGICO

O material biológico (cercárias e miracídios, cepa LE de *S. mansoni*) foi gentilmente cedido pela pesquisadora Liana K. J. Passos responsável pelo Moluscário Lobato Paraense no Centro de Pesquisa René Rachou - FIOCRUZ, Belo Horizonte – MG, onde o ciclo de vida do parasito *S. mansoni* é mantido. Parte das cercárias foi transformada *in vitro* em esquistossômulos por meio de estresse mecânico, utilizando-se uma seringa para estimular a perda da cauda (BASCH, 1981) e os esquistossômulos foram cultivados por 7 dias em meio MEM-Suplementado (2% de Soro Fetal Bovino e 1% de antibiótico penicilina-estreptomicina). Os vermes adultos foram obtidos através de perfusão de *hamsters* (Golden) infectados por 7-8 semanas (SMITHERS; TERRY, 1965), em parceria com a pos-doc Sílvia Dias. Para a obtenção dos esporocistos, parte dos miracídios foi cultivada em meio RPMI1640 suplementado (5% de soro fetal bovino e 100 µg/ml de antibiótico penicilina-estreptomicina) por 48 horas em estufa tipo BOD a 27°C para garantir a transformação destes em esporocistos primários através da perda das placas ciliares (KAWANAKA; SIDNER; CARTER, 1985).

3.3. EXTRAÇÃO DOS RNAs

Para a extração dos RNAs foram utilizados o reagente Trizol (Invitrogen) e o kit RNeasy (Invitrogen), segundo normas do fabricante. Foi feita uma eletroforese em gel desnaturante de agarose 1% para visualização da qualidade do RNA extraído. O RNA total foi tratado com DNase por 30 min a 37°C, utilizando o kit Ambion® TURBO DNA-free™ (Invitrogen), segundo manual do fabricante, com o objetivo de remover DNA genômico. Posteriormente, a densidade óptica (OD

260/280) do RNA tratado foi medida, utilizando-se o espectrofotômetro NanoDrop Spectrophotometer ND-1000 (Thermo Scientific).

3.4. PREPARO DAS BIBLIOTECAS DE cDNAs E SEQUENCIAMENTO NO EQUIPAMENTO

ILLUMINA HiSeq 2000

Duas alíquotas de RNA total de cercária (correspondentes a duas réplicas biológicas) foram precipitadas na presença de etanol absoluto e acetato de sódio 0,3 M e enviadas para a *facility* FASTERIS (Suíça, <https://www.fasteris.com>), onde a qualidade do RNA total foi avaliada com o equipamento Agilent 2100 Bioanalyzer (Agilent Technologies). O material foi processado para construção das bibliotecas e posterior sequenciamento, segundo o protocolo descrito por Nilsson e colaboradores (2010). Em suma, os mRNAs poliadenilados foram capturados usando Dynalbeads oligo-(dT) (Invitrogen) de acordo com o fabricante. A síntese da primeira fita de cDNA foi feita utilizando-se iniciadores randômicos de 6 nt e Superscript II reverse transcriptase (Invitrogen). A síntese da segunda fita foi produzida utilizando a enzima Taq polimerase (New England Biolabs), e um iniciador específico ([BIOT]5'-AATGATACGGCGACCACCGAGATCTACACTCTTGTGATTTGTTGCATG -3') que contém parte do adaptador Illumina HiSeq 2000 e parte da sequência do SL de *S. mansoni* marcada em negrito. O cDNA produzido foi purificado, utilizando-se o kit Qiagen MinElute column (Qiagen) e eluído em tampão TE (Tris-HCl 10 mM, pH 8,0; EDTA 1 mM). A biblioteca foi então preparada para posterior sequenciamento no equipamento Illumina HiSeq 2000, seguindo as recomendações do fabricante com a seguinte modificação: para o sequenciamento, um iniciador específico foi utilizado 5'GAGATCTACACTCTTGTGATTTGTTGCATG3', de forma que somente os

cDNAs que se iniciam com a sequência do SL foram sequenciados. Cada uma das réplicas biológicas foi sequenciada utilizando 1/8 da lâmina, produzindo sequências do tipo *single* de 100 pb. Essas bibliotecas foram denominada SL Trapping (Figura 4A).

3.5. PREPARO DAS BIBLIOTECAS DE cDNA E SEQUENCIAMENTO NO EQUIPAMENTO ION TORRENT PGM

Dez µg de RNA total das fases miracídio, esporocisto, esquistossômulo e vermes adultos foram purificados utilizando-se esferas magnéticas Dynalbeads C1 Streptavidin (Invitrogen) e oligos dT biotinilados para a captura dos mRNAs através da complementariedade com a cauda poli-A. Após a captura, a solução contendo o mRNA e as esferas serviu de substrato para a síntese de cDNA. Para isso foi utilizado o kit SuperScript III Reverse Transcriptase (Invitrogen), conforme instruções do fabricante, e um iniciador complementar à cauda poli-A com uma sequência estendida

5'CGGTATTTTCAGTCGGTGTTCAAACCTTTTTTTTTTTTTTTTTTTT3' V=A,G,C,

correspondente ao sítio para anelamento do iniciador reverso no passo seguinte de amplificação. Os cDNAs foram então amplificados utilizando o iniciador complementar ao SL de *S. mansoni*, o iniciador complementar à sequência estendida dado oligo dT(BREHM *et al.*, 2000) à concentração de 0.5 pM cada iniciador e o kit GoTaq DNA Polymerase (Promega), seguindo recomendações do fabricante. As reações de PCR foram realizadas no termociclador (Bio-Rad), seguindo o programa descrito a seguir:

- desnaturação a 95°C por 5 min
- 5 ciclos de:
 - desnaturação a 95°C por 1 min

- anelamento dos iniciadores a 60°C por 1 min
- extensão a 72°C por 1 min e 30 s
- 5 ciclos de:
 - desnaturação a 95°C por 1 min
 - anelamento dos iniciadores a 59°C por 1 min
 - extensão a 72°C por 1 min e 30 s
- 5 ciclos de:
 - desnaturação a 95°C por 1 min
 - anelamento dos iniciadores a 58°C por 1 min
 - extensão a 72°C por 1 min e 30 s
- 23 ciclos de:
 - desnaturação a 95°C por 1 min
 - anelamento dos iniciadores a 57°C por 1 min
 - extensão a 72°C por 1 min e 30 s

Os cDNAs produzidos foram analisados em gel de agarose 1% corados com brometo de etídio após corrida eletroforética e purificados utilizando-se o kit MinElute PCR Purification Kit (Qiagen), segundo recomendações do fabricante. Inicialmente, 100 ng de cDNA foi fragmentado aleatoriamente por 8 min utilizando o sistema Ion Shear™ (Life Technologies) e fragmentos em torno de 200 pb foram recuperados utilizando-se o sistema E-Gel® SizeSelect™ Gels (Life Technologies). Em seguida, as extremidades dos fragmentos de cDNA foram reparadas e ligadas a adaptadores específicos fornecidos pelo kit de preparo de biblioteca Ion Plus Fragment Library Kit (Life Technologies), conforme instrução do fabricante. Os fragmentos de cDNA ligados aos adaptadores foram equalizados quanto a sua concentração para ~100 pM, utilizando-se o Ion Library Equalizer™ Kit e então, 4 µl da biblioteca equalizada foram amplificados através de PCR em emulsão, ligados a esferas magnéticas, utilizando-se o kit Ion PGM™ Template OT2 200 Kit (Life Technologies). As esferas magnéticas positivas foram posteriormente enriquecidas e preparadas para deposição no Ion 316™ Chip Kit - Ion Torrent™ (100Mb) (Life Technologies) e sequenciadas em equipamento Ion Torrent PGM™ System (Life Technologies) utilizando-se o Ion PGM™ Sequencing 200 Kit (Life Technologies).

Para cada uma das quatro fases analisadas foram produzidos dois sequenciamentos, cada um correspondente a uma réplica biológica. Essas bibliotecas foram denominadas SL Enriched (Figura 4B).

3.6. BUSCA POR DADOS DE RNA-SEQ DE *S. MANSONI* EM BANCOS DE DADOS

Foi realizada uma busca na base de dados públicos SRA (*Sequence Read Archive* - NCBI, <http://www.ncbi.nlm.nih.gov/sra>) (LEINONEN; SUGAWARA; SHUMWAY, 2011) por sequências geradas na plataforma Illumina HiSeq 2000 durante estudos de transcriptômica em diferentes fases do ciclo de vida do parasito *S. mansoni*. As sequências foram convertidas para o formato fastq usando o parser fastq-dump (SRA Toolkit) e o arquivo resultante foi dividido utilizando-se um *in house script* em perl Split_paired_ends.pl para gerar os pares de *reads*. O conjunto de sequências totais encontrados nos bancos de dados foi denominado Total RNA-Seq.

3.6.1. IDENTIFICAÇÃO DA SEQUÊNCIA DO SL NO CONJUNTO DE DADOS TOTAL RNA-SEQ

A fim de se obter apenas as sequências dos transcritos que, de fato, foram processados por SLTS no conjunto de *reads* obtidas no SRA, empregou-se o algoritmo SortPairedEnds (desenvolvido por Haibao Tang, em <http://github.com/tanghaibao/trimReads>) com o parâmetro `-s 19` (score mínimo para considerar a sequência válida, sendo considerado +1 ponto para cada pareamento correto e -3 para pareamento incorreto, ou para a abertura e extensão de *gaps*) para identificar a sequência do SL de *S. mansoni*: 5'AACCGTCACGGTTTTACTCTTGTGATTTGTTGCATG3' (RAJKOVIC *et al.*, 1990), ou seu complemento reverso nas *reads* e separá-las das demais. Este programa

classifica os pares de *reads* em um conjunto que contém a sequência de interesse e em outro que não a contém. Os arquivos de saída contendo as *reads* com a sequência do SL foram então submetidos ao programa trimReads (desenvolvido por Haibao Tang, em <http://github.com/tanghaibao/trimReads>) para remoção das sequências SL nas *reads*, utilizando-se as flags `-s 19, -q 0` (para não remover sequências de bases pelas suas qualidades) e `-m 0` (para não descartar nenhuma sequência e manter as *reads* pareadas). A remoção da sequência do SL é necessária para que posteriormente ocorra o correto alinhamento das *reads* no genoma de referência de *S. mansoni*. Em ambos os programas a sequência de interesse é identificada utilizando-se o algoritmo Waterman-Eggert (WATERMAN; EGGERT; LANDER, 1992). Em seguida, foi utilizado um *in house script* em perl Split_paired_ends_SL.pl para gerar os arquivos *paired-ends* `<*_1.fastq>` e `<*_2.fastq>`. O conjunto de dados constituído pelas *reads* que continham a sequência do SL foi denominado RNA-Seq Filtered (Figura 4C).

3.7. PIPELINE PARA ANÁLISE DOS DADOS DE RNA-SEQ GERADOS NA PLATAFORMA

ILLUMINA HiSeq 2000

3.7.1. ANÁLISE DE QUALIDADE , MAPEAMENTO E CONTAGEM DAS READS

A qualidade de todas as *reads* foi avaliada utilizando-se o programa FastQC (WWW.dioinformatics.babraham.ac.uk/prpjects/fastqc), que possui diferentes módulos de análises para o controle de qualidade das sequências brutas produzidas pelos equipamentos de NGS. Os gráficos gerados pelo programa permitem avaliar a qualidade geral das bibliotecas e tomar decisões importantes que afetam as análises posteriores. Os arquivos de extensão .fastq dos conjuntos de dados SL Trapping, Total RNA-Seq e RNA-Seq Filtered foram separadamente

submetidos ao programa TopHat 2.0 (TRAPNELL; PACHTER; SALZBERG, 2009) a fim de se obter o alinhamento das sequências no genoma de *S. mansoni* (versão 5, obtido em <ftp://ftp.sanger.ac.uk/pub/pathogens/Schistosoma/mansoni/>), utilizando os seguintes parâmetros: -i 10 (tamanho mínimo do íntron), -I 30000 (tamanho máximo do íntron), --coverage-search (para a busca por junções exon-exon baseada em cobertura) -j combined.juncs (alimenta o programa com um conjunto de junções exon-exon conhecidas) e -G v5.07.08.12.chado.raw.gff. Esta última opção fornece ao programa um conjunto de anotações baseado no modelo gênico disponível, para que a primeira etapa do alinhamento ocorra no transcriptoma virtual do parasito (o arquivo utilizado foi obtido em ftp://ftp.sanger.ac.uk/pub/pathogens/Schistosoma/mansoni/genome/Gene_models/).

Métricas que descrevem os alinhamentos foram coletadas utilizando-se o programa CollectAlignmentMetrics, do pacote de ferramentas Picard (LI *et al.*, 2009). O alinhamento foi também verificado visualmente utilizando o visualizador IGV 2.1 (THORVALDSDÓTTIR; ROBINSON; MESIROV, 2013) (Broad Institute, obtido em <http://www.broadinstitute.org/software/igv/download/>). Através de uma filtragem realizada com a ferramenta Samtools versão 0.1.19 (LI *et al.*, 2009) (com os parâmetros `view -bhq 20 -F 0x100`), somente *reads* com alinhamento único e qualidade de alinhamento ≥ 20 na escala Phred foram aceitas em análises posteriores.

A contagem bruta do número de *reads* alinhadas para cada gene presente no modelo gênico de *S. mansoni* (v5.07.08.12.chado.raw.gff), disponível no repositório GeneDB (versão 5), foi obtida com o pacote HTSeq Python versão

0.5.3p3 (disponível em <http://www-huber.embl.de/users/anders/HTSeq/doc/index.html>). As contagens foram obtidas para todos os três conjuntos de dados (SL Trapping, Total RNA-Seq e RNA-Seq Filtered), utilizando as opções `--stranded=yes` e `--mode=intersection-strict`. Uma abordagem semelhante foi utilizada para obter contagens para os exons, porém utilizando um arquivo de anotação de genes modificado por um *script* em python chamado `dexseq_prepare_annotation.py` do pacote DEXSeq (ANDERS; REYES; HUBER, 2012) para o software R.

Para as bibliotecas do tipo SL Trapping, somente genes reproduzidos em ambas replicatas biológicas e com média de *reads* ≥ 10 contagens foram mantidos nas análises posteriores.

Durante os últimos anos, uma série de abordagens para a normalização dos dados de sequenciamento de RNA surgiram na literatura, diferindo tanto no tipo de estratégia estatística adotada, quanto no viés acarretado na análise. No entanto, como os dados continuam a acumular-se, não existe um consenso claro sobre o método de normalização apropriada a ser usado (DILLIES et al., 2013).

Neste trabalho nós focamos no uso do pacote estatístico DESeq2, que é baseado em modelos lineares que seguem uma distribuição binomial negativa para estimar dados de dispersão e o logaritmo do fold change. DESeq2 executa para cada gene um teste de hipótese para verificar se existem evidências suficientes para decidir contra a hipótese nula de que não existe qualquer efeito das condições sobre a expressão do gene e que a diferença observada entre duas condições é causada apenas por uma variabilidade experimental. Para calcular a significância de um gene, o pacote utiliza a relação entre a variância dos dados (ou

dispersão) e a sua média. O resultado deste teste é relatado como um valor de p que indica a probabilidade de que a diferença observada entre as duas condições, ou mesmo uma diferença ainda mais acentuada, seria encontrada na situação descrita pela hipótese nula. Este valor de p-value é ajustado utilizando o método de ajuste Benjamini-Hochberg (BH). Este método calcula para cada gene um valor p ajustado que responde a seguinte pergunta: se todos os genes com valor de p menor ou igual ao limiar de valor p deste gene foram considerados significativos, qual seria a fração de falsos positivos entre eles (a taxa de detecção falsa, FDR)? Estes valores, chamados de valores de pBH-ajustados, são utilizados para definir um *cutoff* de significância (ANDERS; HUBER, 2010).

Dessa forma, normalização da contagem dos genes e obtenção de seus valores de expressão foi realizada através do pacote DESeq2 versão 1.2.5 (ANDERS; HUBER, 2010) para o software R, e uma normalização adicional por tamanho de cada gene foi realizada no qual uma matriz contendo o comprimento de cada um dos genes foi incluída no grupo de dados Total RNA-Seq. Os valores de abundância dos genes estimados pelo programa no grupo de dados Total RNA-Seq, normalizados pelo seu tamanho, foram comparados com suas frequências de SLTS, encontradas nas bibliotecas do tipo SL Trapping e RNA-Seq Filtered.

3.8. PIPELINE PARA ANÁLISE DOS DADOS DE RNA-SEQ GERADOS NA PLATAFORMA ION TORRENT

3.8.1. IDENTIFICAÇÃO DA SEQUÊNCIA DO SL NO CONJUNTO DE DADOS SL ENRICHED

As sequências obtidas no sequenciador Ion Torrent PGM™ System (Life Technologies) também foram analisadas quanto à presença da sequência do SL. As

reads contendo a sequência em questão foram aparadas utilizando-se o algoritmo *fastq-mcf* do pacote *ea-utils* (ARONESTY, 2013) com os parâmetros: *-m 25* (número mínimo de pb a ser identificado na *read*) *-p 4* (porcentagem máxima de discrepância entre as sequências) *-l 20* (tamanho mínimo da sequência remanescente).

3.8.2. TRATAMENTO DE QUALIDADE DAS *READS*

Como o sequenciador Ion Torrent PGM™ System (Life Technologies) trabalha com *reads* com qualidade de sequenciamento médio em torno de 20 na escala Phred, as *reads* sofreram um tratamento criterioso de controle de qualidade para que as bases de qualidade ruim fossem removidas, melhorando a qualidade global dos dados. Nesta etapa utilizamos o algoritmo PRINSEQ (SCHMIEDER; EDWARDS, 2011) com os parâmetros: *-min_len 20* (tamanho mínimo da *read* remanescente) *-min_qual_mean 18* (qualidade média mínima da *read*) *-ns_max_p 60* (porcentagem máxima nas sequências de nucleotídeos desconhecidos -N) *-trim_ns_right 5* (apara cauda poli-N, com este comprimento mínimo na extremidade 3') *-trim_tail_right 5* (apara caudas poli-A/T com este tamanho mínimo na na extremidade 5') *-trim_tail_left 5* (apara caudas poli-A/T com este tamanho mínimo na extremidade 3') *-trim_qual_step 1* (tamanho do passo utilizado para mover a janela deslizante) *-trim_qual_right 20* (apara a sequência pelo índice de qualidade mínima a partir da extremidade 3') *-trim_qual_type mean* (tipo de cálculo a ser utilizado no índice de qualidade) *-trim_qual_rule lt* (regra de comparação para o índice de qualidade de valor calculado na janela deslizante) *-trim_qual_window 5* (tamanho da janela deslizante).

3.8.3. MAPEAMENTO E CONTAGEM DAS *READS*

Após criterioso controle de qualidade, as *reads* remanescentes foram utilizadas para alinhamento no genoma e transcriptoma de referência de *S. mansoni*. Os arquivos de extensão .fastq dos conjuntos de dados SL Enriched foram separadamente submetidos ao programa TopHat 2.0 (TRAPNELL; PACHTER; SALZBERG, 2009). Como as *reads* obtidas pelo sequenciamento na plataforma Ion Torrent PGM™ System tem qualidade média 20 na escala Phred, permitiu-se um maior número de *mismatches* durante a fase de alinhamento das *reads* no genoma de referência, comparando-se com o *pipeline* utilizado para as *reads* obtidas pelo sequenciamento na plataforma Illumina HiSeq 2000, que tem qualidade média em torno de 30 na escala Phred. Dessa forma, foram utilizados os seguintes parâmetros: -i 10 (tamanho mínimo do íntron), -I 30000 (tamanho máximo do íntron), --coverage-search (permite a busca por junções exon-exon baseado em cobertura) -G v5.07.08.12.chado.raw.gff (fornece ao programa um conjunto de anotações baseado no modelo gênico disponível, para que a primeira etapa do alinhamento ocorra no transcriptoma virtual do parasito (o arquivo utilizado foi obtido em ftp://ftp.sanger.ac.uk/pub/pathogens/Schistosoma/mansoni/genome/Gene_models/), -j combined.juncs (alimenta o programa com um conjunto de junções exon-exon conhecidas), --read-mismatches 4 (número máximo de *mismatches* permitido nas *reads* alinhadas) --read-gap-length 4 (número máximo de *gaps* permitido nas *reads* alinhadas) --read-edit-dist 4 (alinhamentos finais de *reads* com mais do que esta distância de edição são descartados) -a 5 (comprimento “âncora” para encontrar junções exon-exon) -m 1 (número máximo de *mismatches* que podem aparecer na região de “âncora” de um alinhamento que ocorreu numa junção exon-

exon) --max-insertion-length 5 (tamanho máximo de uma inserção) --max-deletion-length 5 (tamanho máximo de uma deleção).

Os arquivos unmmapped.bam contendo as *reads* não mapeadas nessa primeira etapa foram convertidos para o formato .fastq utilizando-se a ferramenta SamToFastq.jar do pacote Picard (LI *et al.*, 2009). As *reads* não mapeadas foram então submetidas a uma segunda rodada de alinhamento utilizando o alinhador Bowtie2 (LANGMEAD; SALZBERG, 2012), com os parâmetros --local (neste modo, Bowtie2 executa alinhamento local das *reads*, procurando por locais de semelhança entre as sequências sem ter de considerar todo o comprimento destas, podendo ignorar alguns caracteres de ambas as extremidades das *reads*, caso isso maximize a pontuação de alinhamento da mesma.) e --very-sensitive-local (modo projetado para que o programa seja executado de forma mais veloz, sensível e precisa). O arquivo de saída no formato .sam foi convertido para o formato .bam e fundido com o arquivo de mapeamento obtido na primeira etapa com o programa TopHat 2.0 (TRAPNELL; PACHTER; SALZBERG, 2009), utilizando-se o pacote Samtools (LI *et al.*, 2009).

As métricas dos alinhamentos foram coletadas utilizando-se o programa CollectAlignmentMetrics do pacote de ferramentas Picard (Li et al., 2009). O alinhamento foi também verificado visualmente utilizando o visualizador IGV 2.1 (Broad Institute, obtido em <http://www.broadinstitute.org/software/igv/download/>).

As contagens brutas do número de *reads* alinhadas para cada gene presente no modelo gênico de *S. mansoni* (v5.07.08.12.chado.raw.gff), disponível no repositório GeneDB (versão 5), foram obtidas com o pacote HTSeq Python versão

0.5.3p3 (disponível em <http://www-huber.embl.de/users/anders/HTSeq/doc/index.html>), utilizando as opções “--stranded=no” e “-mode=intersection-strict”. Para a contagem de *reads* em cada exon, utilizou-se a mesma abordagem descrita para as sequências de Illumina HiSeq 2000.

3.9. PROCESSAMENTO DOS DADOS

O arquivo de anotação foi utilizado para obter o valor de comprimento dos genes, exons e íntrons. O processamento de dados, assim como sua visualização foram realizados empregando a versão 3.0.2 do programa R (R Core Team, 2013). Alguns gráficos foram construídos usando o pacote ggplot2 para R, versão 0.9.3.1

3.10. ANOTAÇÃO E ANÁLISE FUNCIONAL DAS SEQUÊNCIAS

Como a maioria dos genes que sofrem processamento por SLTS estão anotadas como "proteínas hipotéticas", nós re-anotamos as sequências de mRNA previstos nos modelos gênicos do parasito. Para a anotação funcional dos transcritos foi feita uma colaboração com Dr. José Marcos C. Ribeiro, que possui ampla experiência na análise de dados e desenvolvimento de ferramentas para anotação de sequências geradas em experimentos de transcriptômica pelas novas plataformas de sequenciamento. O Dr. Ribeiro é chefe da seção de Biologia de Vetores no *Laboratory of Malaria and Vector Research*, no NIAID (*National Institute of Allergy and Infectious Diseases*) situado no NIH (National Institutes of Health) Bethesda, Maryland, USA e supervisionou a anotação das sequências de *S. mansoni* durante meu estágio no NIH.

Todas as sequências de nucleotídeos referentes aos genes contidos no modelo gênico de *S. mansoni* (versão 5) foram utilizadas para realizar buscas por similaridade em diversos bancos de dados utilizando os programas BLAST (ALTSCHUL *et al.*, 1997): BLASTX, BLASTN ou RPS-BLAST, conforme realizado por Karim e colaboradores (2011).

Os tipos de BLAST, as bases de dados e os parâmetros utilizados estão listados na **Tabela 1**. As traduções preditas das proteínas foram submetidas ao servidor SignalP (NIELSEN *et al.*, 1997) para ajudar a identificar os produtos de tradução que poderiam ser secretados, ao servidor TMHMM (SONNHAMMER; VON HEIJNE; KROGH, 1998) para detectar as hélices transmembrana, ao servidor NetOglyc para detectar possíveis O-glicosilações do tipo mucina (HANSEN *et al.*, 1998) e ao servidor ProP (DUCKERT; BRUNAK; BLOM, 2004) para identificar possíveis sítios de clivagem (contendo Arg e Lys) para a furinas, que são enzimas conhecidas por converter pré-proteínas em seus produtos biologicamente ativos.

Tabela 1 - Tipos de BLAST, bases de dados e parâmetros utilizados para anotação dos transcritos.

| Tipo | DB | Parâmetros |
|-------------|--|--------------------------------|
| BLASTN | Mit-pla (Genbank) | -IT -JT -v10 -b3 -e1e-10 -FF |
| | rRNA (GenBank) | -IT -JT -v10 -b3 -e1e-10 -FF |
| | Rfam (Sanger) (GARDNER <i>et al.</i> , 2011) | -IT -JT -v10 -b3 -e1e-10 -FF |
| BLASTX | NR (Genbank) | -IT -JT -v10 -b3 -e100 -FF -CF |
| | NR_Acelomata (Subset do GenBank) | -IT -JT -v10 -b3 -e100 -FF -CF |
| | Swiss-Prot (UniProtKB) | -IT -JT -v10 -b3 -e100 -FF -CF |
| | WormBase (HARRIS <i>et al.</i> , 2010) | -IT -JT -v10 -b3 -e100 -FF -CF |
| | Gene Ontology – GO (LEWIS; | -IT -JT -v10 -b3 -e1e-4 -FF - |

| | | |
|---------------|---------------------------------------|-----------------------------------|
| RPS- BLAST | ASHBURNER; REESE, 2000) | CF |
| | KEGG Orthology (KANEHISA; GOTO, 2000) | -v1 -b1 -e1e-5 -FF |
| | COG (TATUSOV <i>et al.</i> , 2003) | -IT -JT -v10 -b10 -e10 -FF -pF |
| | Pfam (PUNTA <i>et al.</i> , 2012) | -IT -JT -v10 -b10 -e10 -FF -pF |
| | CDD (MARCHLER-BAUER, 2002) | -IT -JT -v10 -b10 -e10 -FF -pF |
| | SMART (SCHULTZ <i>et al.</i> , 2000) | -IT -JT -v10 -b10 -e10 -FF -pF |
| | TIGRFAMS (J. Craig Venter Institute) | -IT -JT -v10 -b10 -e10 -FF -pF |
| | PRK | -IT -JT -v10 -b10 -e10 -FF -pF |
| | TE-CLASS | -IT -JT -v10 -b10 -e1e-15 -FF -pF |

Os vários resultados foram tabulados em uma planilha Excel com hyperlinks que permitiram a anotação das sequências de forma automatizada, assim como sua classificação funcional, utilizando o programa Classifier, escrito em Visual Basic pelo Dr. Ribeiro, que leva em consideração palavras-chave dos *matches* de todos os resultados de BLAST, assim como os *e-values*, os resultados para SignalP, domínios transmembranares e glicosilação para classificar os transcritos em aproximadamente 30 categorias funcionais. Em alguns casos foi feita uma correção manual da anotação nos resultados finais. O programa Class Table Maker, também escrito pelo Dr. Ribeiro, foi utilizado para calcular a frequência das classes funcionais dos transcritos nas bibliotecas, utilizando os valores de contagens obtidos. Para identificar o enriquecimento de determinada classe nos transcritos que sofrem SLTS, foi realizado um teste estatístico χ^2 ($p < 0,05$) para averiguar se a diferença de frequência encontrada para cada classe é estatisticamente significante entre o grupo de transcritos que sofre SLTS e os

genes expressos na fase cercária de *S. mansoni*. As classes funcionais dos genes presentes em unidades policistrônicas cujos transcritos são processados por SLTS também foram comparadas às classes funcionais dos genes processados por SLTS, em geral.

3.11. IDENTIFICAÇÃO DAS VIAS METABÓLICAS ONDE ATUAM AS PROTEÍNAS CODIFICADAS PELOS TRANSCRITOS PROCESSADOS SLTS

Os identificadores UniProtKB/TrEMBL associados às proteínas de *S. mansoni* cujos transcritos são processados por SLTS foram combinados aos seus respectivos valores de frequência de processamento (contagens do grupo SL Trapping). O programa iPath (YAMADA *et al.*, 2011) foi utilizado para representar as diferentes vias metabólicas de *S. mansoni* baseado no banco de dados KEGG (KANEHISA; GOTO, 2000; KANEHISA *et al.*, 2012), destacando as vias associadas a proteínas codificadas por transcritos processados por SLTS, de acordo com a frequência de processamento de cada transcrito.

3.12. IDENTIFICAÇÃO DE TRANSCRITOS POLICISTRÔNICOS EM *S. MANSONI*

Os grupos gênicos pertencentes a um mesmo cromossomo, contidos na mesma fita de DNA e com distância intergênica de até 200 pb (isto é, a distância entre a extremidade 3' de um gene *upstream* e a extremidade 5' de um gene *downstream*, Figura 5), foram detectados no modelo de anotação gênica de *S. mansoni* usando um *script* personalizado. Em seguida, foi avaliada a ocorrência de SLTS nos grupos de genes. Foram assinalados como possíveis transcritos policistrônicos aqueles grupos gênicos que apresentaram o processamento por SLTS pelo menos em todos os genes *downstream*.

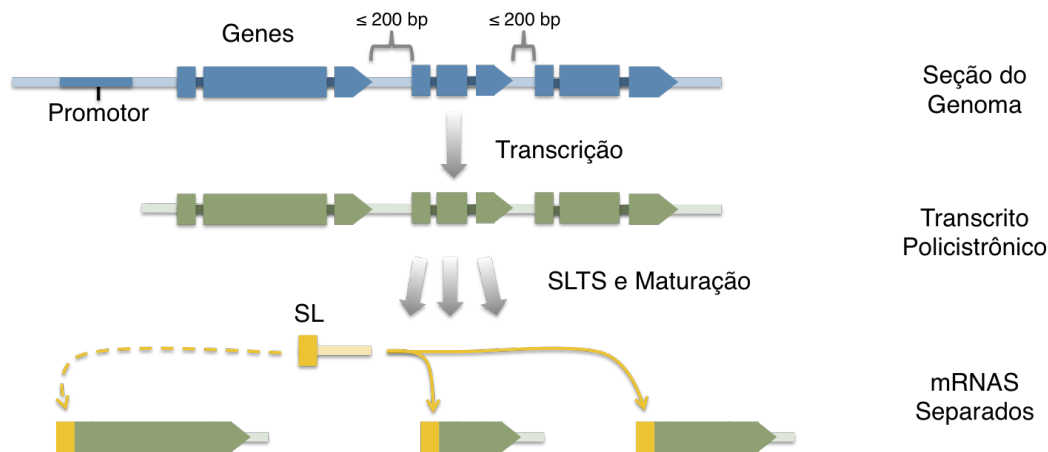


FIGURA 5 - DETECÇÃO DE TRANSCRITOS POLICISTRÔNICOS. Genes presentes em uma mesma fita de DNA e com distância intergênica de até 200 pb foram identificados. Em seguida averigou-se a ocorrência do processamento por SLTS nos genes *downstream* para identificação dos transcritos policistrônicos.

3.13. IDENTIFICAÇÃO NO GENOMA DE GENES QUE CONTÊM INSERÇÃO DA SEQUÊNCIA DO SL

Utilizando o algoritmo BLAT (KENT, 2002) (com os `-t=dna -q=dna -minIdentity=90 -out=blast8 -maxGap=1 -fine`), uma busca pela sequência do SL de *S. mansoni* foi realizada no genoma do parasito. Os genes de SL RNAs foram identificados e removidos do conjunto de resultados, que foram comparados com o modelo gênico do parasito para identificar os genes presentes nas coordenadas onde a sequência do SL foi encontrada.

3.14. ANÁLISE DOS SÍTIOS ACEPTORES DE SL

As sequências que flanqueiam os locais de inserção da sequência do SL correspondentes a 20 nt *upstream* e 4 nt *downstream* do sítio aceptor de *splicing* foram obtidas utilizando a ferramenta BedTools v2.17.0 (QUINLAN; HALL, 2010) e o software WebLogo 3.0 (CROOKS *et al.*, 2004) foi utilizado para gerar a representação gráfica dos padrões encontrados em um alinhamento múltiplo das sequências.

3.15. ANÁLISE DE SINAIS DE *SPLICING*

Com o intuito de investigar características dos íntrons que sofrem *cis-* ou *trans-splicing*, focamos em um grupo de genes que apresentam, pelo menos, um exon interno com média > 10 contagens para SLTS, e com valores positivos em ambas as bibliotecas. Nestes genes, o íntron anterior ao exon com os valores mais altos de SLTS foi identificado como sendo o alvo principal de SLTS e todos os outros íntrons do mesmo gene foram considerados como íntrons que sofreram *cis-splicing* em transcritos que sofrem *trans-splicing*. Um outro conjunto de tamanho aproximadamente igual de genes que não sofrem *trans-splicing* e apresentaram níveis semelhantes de expressão foi também utilizado para obter informações de íntrons que sofrem somente *cis-splicing*. Os conjuntos de íntrons foram então empregados para predição de potenciais sítios aceptores e doadores de *splicing* utilizando os modelos de sítios de *splicing* do programa GeneID (BLANCO; PARRA; GUIGÓ, 2007) (<http://genome.crg.es/software/geneid/geneid.html>), treinados a partir de sítios de *splicing* anotados utilizando-se cadeias ocultas de Markov de primeira ordem. Para avaliar as características dos pontos de ramificação (*branch point*), aplicamos às sequências intrônicas anotadas diferentes modelos integrados no programa Support Vector Machine SVM-BPfinder (CORVELO *et al.*, 2010) (http://regulatorygenomics.upf.edu/Software/SVM_BP/).

4. RESULTADOS E DISCUSSÃO

4.1. OBTENÇÃO DO RNA TOTAL

A qualidade do RNA total extraído para as diferentes fases do ciclo de vida do parasito foi verificada utilizando o equipamento Agilent 2100 Bioanalyzer (Figura 6A) e por eletroforese em gel de agarose 1% desnaturante (Figura 6B).

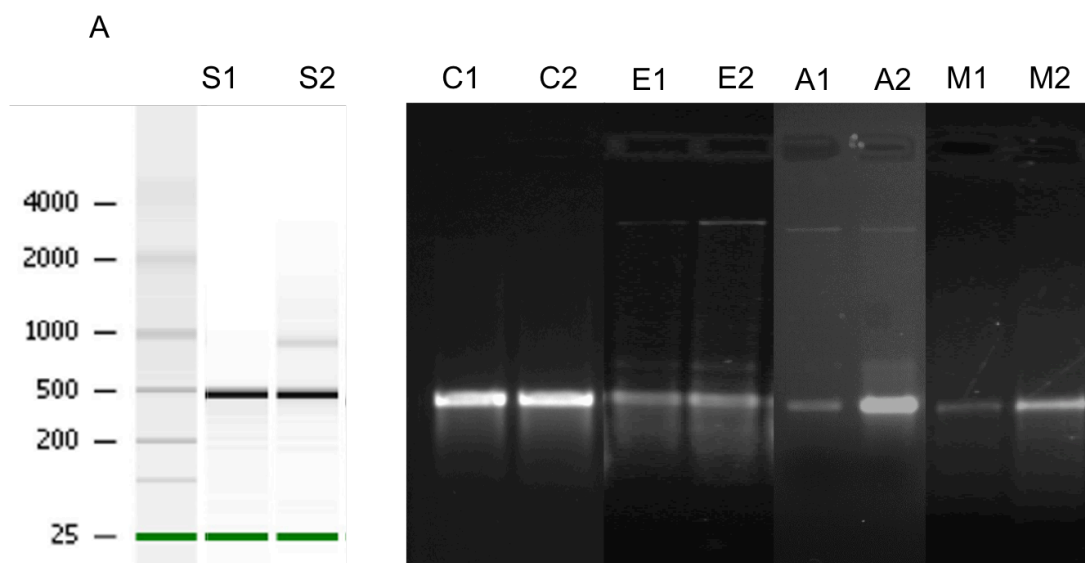


FIGURA 6 - PERFIL DE QUALIDADE DOS RNAs TOTAIS EXTRAÍDOS DE DIFERENTES FASES DO CICLO DE VIDA DO PARASITO *S. mansoni*. A - Gel virtual dos RNAs da fase esporocisto (S1 e S2) gerado pelo Agilent 2100 Bioanalyzer. B - Gel de agarose 1% desnaturante corado com brometo de etídio dos RNAs totais extraídos das fases cercária (C1 e C2), esquistossômulo (E1 e E2), vermes adultos (A1 e A2) e miracídio (M1 e M2) do parasito *S. mansoni*. Apenas uma banda é observada em todos os géis e esta representa a comigração dos fragmentos do rRNA 28S (28S alpha e 28S beta) e rRNA18S.

O perfil eletroforético do rRNA de *S. mansoni* apresenta apenas uma banda, uma vez que o rRNA 28S sofre uma clivagem gerando dois fragmentos (28S alpha e 28S beta) de aproximadamente o mesmo tamanho do rRNA 18S. Dessa forma, não é possível visualizar um banda maior e outra menor no gel de agarose 1% (TENNISWOOD; SIMPSON, 1982). O RNA obtido, como pode ser verificado na Figura 6, é de ótima qualidade pois apresenta uma banda íntegra e pouco sinal de degradação (representado por uma alta quantidade de RNA de baixo peso molecular).

4.2. OBTENÇÃO DAS BIBLIOTECAS DE cDNA ENRIQUECIDAS EM TRANSCRITOS PROCESSADOS POR SLTS

Os RNAs totais das fases vermes adultos, esquistossômulo, miracídio e esporocisto foram tratados com DNase e amplificados, utilizando-se como iniciadores uma sequência complementar ao SL de *S. mansoni* e outra complementar a uma extensão da cauda de poli-A. Após a amplificação, parte dos cDNAs foram utilizados para visualização por eletroforese em gel de agarose 1% desnaturante (Figura 7).

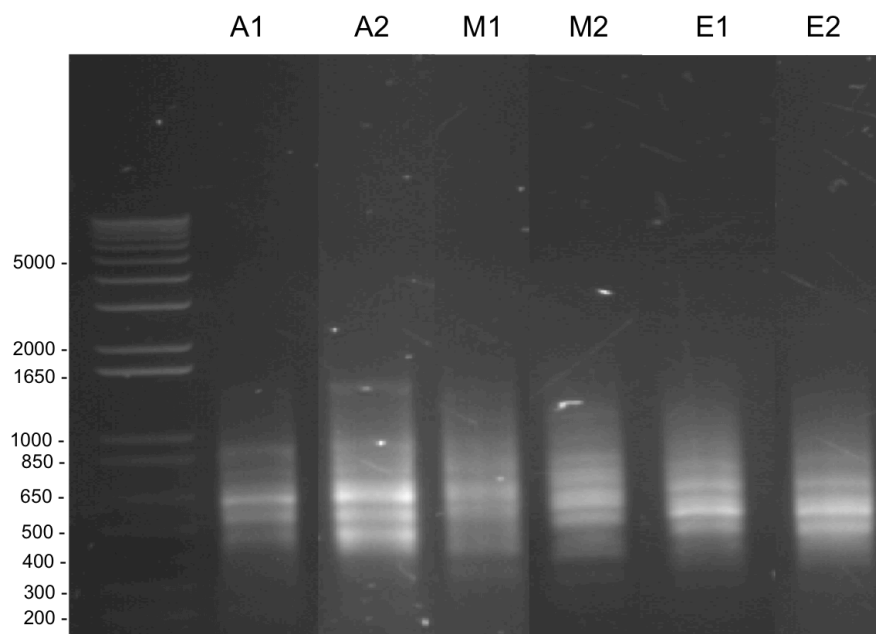


FIGURA 7 - PERFIL ELETROFORÉTICO DOS CDNAS GERADOS À PARTIR DE TRANSCRITOS QUE SOFREM PROCESSAMENTO POR SLTS DE DIFERENTES FASES DE *S. MANSONI*. Observa-se no gel de agarose 1% desnaturante um rastro ao fundo e algumas bandas mais proeminentes em cada amostra, uma vez que os transcritos amplificados apresentam tamanhos variados, que compreendem de 1650 a 100 pb.

Para cada fase, foram produzidas duas réplicas biológicas contendo um conjunto de cDNAs enriquecidos em transcritos processados por SLTS, produzindo um padrão característico no gel (Figura 7). Estes cDNAs foram utilizados para produção das bibliotecas do tipo SL Enriched e posterior sequenciamento em Plataforma Ion Torrent PGM™ System (Life Technologies).

4.3. OBTENÇÃO DAS READS NA PLATAFORMA ILLUMINA HISEQ 2000

Foram sequenciadas na plataforma Illumina HiSeq 2000 duas réplicas biológicas da fase cercária em dias diferentes, utilizando-se um protocolo modificado para produção dos conjuntos de dados SL Trapping, que contém transcritos que se iniciam com a sequência do SL. O resultado das corridas está detalhado na Tabela 2 .

TABELA 2 - DADOS GERADOS NA PLATAFORMA ILLUMINA HISEQ 2000.

| Nome da Biblioteca | Número de Acesso ¹ | Fase | Plataforma | Tamanho das reads | %GC | Número de reads |
|--------------------|-------------------------------|----------|---------------------|-------------------|-----|-----------------|
| SL Trapping 1 | SRR1134198 | Cercária | Illumina Hiseq 2000 | 100 pb | 37 | 11.520.178 |
| SL Trapping 2 | SRR1134204 | Cercária | Illumina Hiseq 2000 | 100 pb | 37 | 30.332.894 |

¹ Número de acesso do depósito no banco de dados SRA(NCBI).

4.4. OBTENÇÃO DAS READS NA PLATAFORMA ION TORRENT PGM™ SYSTEM

Foram sequenciadas na plataforma Ion Torrent PGM™ System duas réplicas biológicas das fases adulto, esquistossômulo, miracídio e esporocisto, utilizando-se cDNAs enriquecidos em transcritos que sofrem processamento por SLTS, resultando no conjunto de dados denominado SL Enriched. O resultado das corridas nesta plataforma está detalhado na Tabela 3 :

TABELA 3 - DADOS GERADOS NA PLATAFORMA ION TORRENT PGM™ SYSTEM.

| Nome da Biblioteca | Fase | Tipo de Biblioteca | Tamanho das reads | %GC | Número de reads |
|--------------------|-----------------|--------------------|-------------------|-----|-----------------|
| SL Enriched A1 | Adulto | Single | 12-353 pb | 36 | 2.586.306 |
| SL Enriched A2 | Adulto | Single | 8-368 pb | 37 | 2.802.032 |
| SL Enriched E1 | Esquistossômulo | Single | 8-330 pb | 35 | 2.562.467 |
| SL Enriched E2 | Esquistossômulo | Single | 8-350 pb | 35 | 3.372.875 |
| SL Enriched M1 | Miracídio | Single | 8-366 pb | 36 | 3.402.016 |
| SL Enriched M2 | Miracídio | Single | 8-371 pb | 35 | 2.792.890 |
| SL Enriched S1 | Esporocisto | Single | 8-372 pb | 36 | 3.479.656 |
| SL Enriched S2 | Esporocisto | Single | 8-376 pb | 36 | 3.127.173 |

4.5. AVALIAÇÃO DA QUALIDADE DAS READS GERADAS

A qualidade das *reads* utilizadas nas análises foi avaliada segundo os quesitos descritos a seguir:

a. Valores de qualidade por base

A Figura 8 mostra uma visão geral dos valores de qualidade das bases em cada posição nos arquivos fastq gerados após o sequenciamento das bibliotecas. Em geral, um *score* de qualidade é considerado aceitável quando é superior ao valor 20 na escala Phred, o que indica a probabilidade de ocorrência de 1 erro a cada 100 pb. Para as bibliotecas sequenciadas na plataforma Illumina HiSeq 2000 (Figura 8A), a média da qualidade por base foi em torno de 30 na escala Phred. Já para as demais bibliotecas sequenciadas no equipamento Ion Torrent PGM™ System (Figura 8B-E), a média foi em torno de 20. Essa diferença era esperada, uma vez que a taxa de erro descrita para a plataforma Illumina HiSeq 2000 é menor do que para a plataforma Ion Torrent PGM™ System (QUAIL *et al.*, 2012). Apesar de os valores de qualidade das *reads* sequenciadas na plataforma Illumina HiSeq 2000 serem em geral maiores do que para a outra plataforma, é descrito na literatura que para Ion Torrent PGM™ System, os valores Phred de qualidade estimados para as bases são muitas vezes subestimados, uma vez que ao alinhar as *reads* geradas no genoma de referência, observa-se muitas vezes uma maior acurácia das bases (BRAGG *et al.*, 2013; LOMAN *et al.*, 2012; ROTHBERG *et al.*, 2011). Além disso, nota-se que há uma queda nos valores de qualidade na porção 3' terminal das *reads*, o que é um artefato dos sequenciadores utilizados (DOHM *et al.*, 2008).

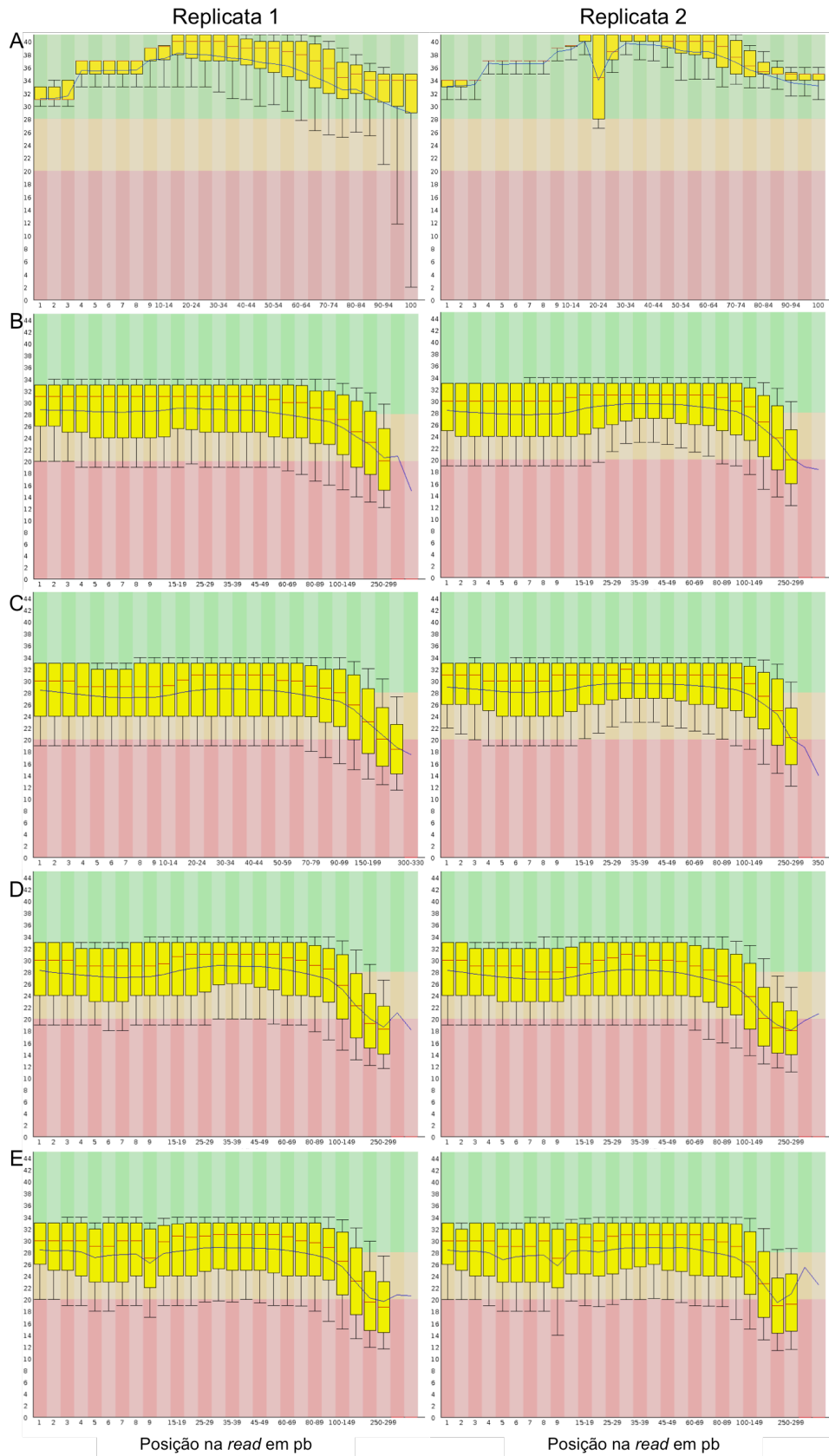


FIGURA 8 - VALORES DE QUALIDADE POR BASE. A linha vermelha central corresponde ao valor da mediana; as caixas amarelas representam o intervalo inter-quartil (25-75%); as barras superior e inferior representam os 10% e 90% dos

pontos; a linha azul representa a qualidade média das bases. **A** – Sequências geradas na plataforma Illumina HiSeq 2000, fase cercária, replicatas 1 e 2; **B** – Sequências geradas na plataforma Ion Torrent PGM™ System, fase adulto, replicatas 1 e 2; **C** – Sequências geradas na plataforma Ion Torrent PGM™ System, fase esquistossômulo, replicatas 1 e 2; **D** – Sequências geradas na plataforma Ion Torrent PGM™ System, fase miracídio, replicatas 1 e 2; **E** – Sequências geradas na plataforma Ion Torrent PGM™ System, fase esporocisto, replicatas 1 e 2.

É possível notar que a qualidade das reads obtidas pelo sequenciador Ion Torrent PGM™ System é mais estável, enquanto que na plataforma Illumina HiSeq 2000 a qualidade diminui consideravelmente após 50 ciclos, o que pode ser causado pelo decaimento do sinal de fluorescência com o aumento do tamanho da sequência (LIU *et al.*, 2012). Ainda é possível observar pela Figura 8 que as leituras obtidas no sequenciador Illumina HiSeq 2000 possuem tamanho único de 100 pb, enquanto as leituras obtidas pelo equipamento Ion Torrent PGM™ System possuem tamanho variado de 8 a 350 pb.

b. Valor de qualidade média por sequência

Esta análise permite verificar se um subconjunto das sequências tem valores de qualidade muito baixos, levando-se em consideração a média de qualidade por base de cada sequência. Caso seja essa a realidade, ao invés de realizar uma corte das sequências, o melhor a fazer seria descartar as sequências de qualidade baixa. Na Figura 9, é possível observar uma distribuição unimodal do valor de qualidade para todas as bibliotecas. No entanto, bibliotecas obtidas pelo sequenciamento na plataforma Illumina HiSeq 2000 (Figura 9A) apresentam menor variância e uma média centrada em um valor de Phred 37, o que mostra que a grande maioria das sequências apresenta alta qualidade. Por outro lado, sequências obtidas no equipamento Ion Torrent PGM™ System (Figura 9B-E) se mostraram menos homogêneas, além de ter média centrada em torno de Phred 27, o que mostra que essas bibliotecas apresentam leituras com menor qualidade.

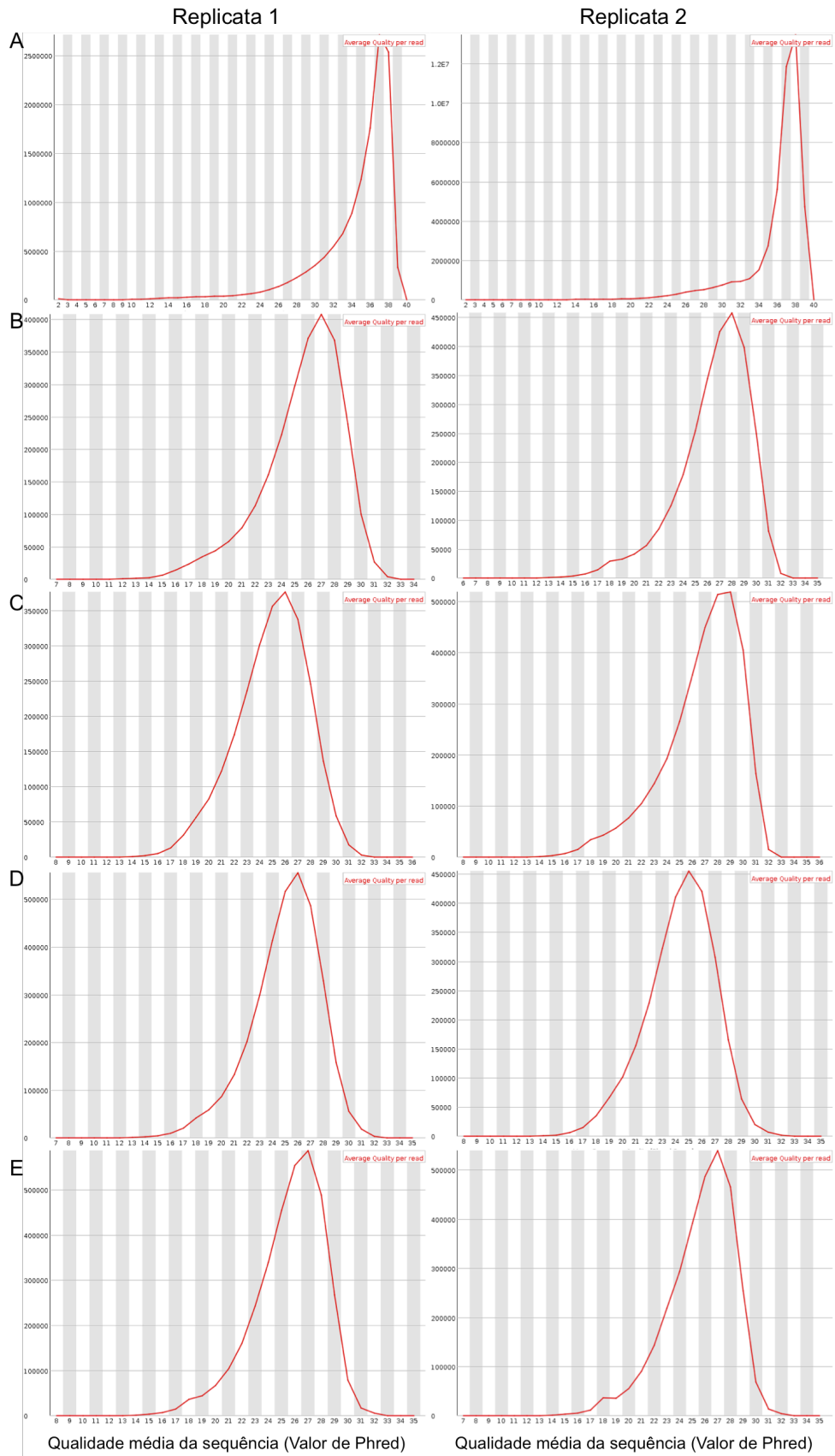


FIGURA 9 - VALOR DE QUALIDADE MÉDIA POR SEQUÊNCIA. A - Sequências geradas na plataforma Illumina HiSeq 2000, fase cercária, replicatas 1 e 2; B - Sequências geradas na plataforma Ion Torrent PGM™ System, fase adulto, replicatas 1 e 2;

C – Sequências geradas na plataforma Ion Torrent PGM™ System, fase esquistossômulo, replicatas 1 e 2; D – Sequências geradas na plataforma Ion Torrent PGM™ System, fase miracídio, replicatas 1 e 2; E – Sequências geradas na plataforma Ion Torrent PGM™ System, fase esporocisto, replicatas 1 e 2.

c. Conteúdo de bases nas sequências

Em uma biblioteca aleatória seria esperado pouca ou nenhuma diferença na proporção (conteúdo) de bases por posição na sequência, de modo que as linhas nos gráficos a seguir (Figura 10) devem ficar paralelas ao eixo das abscissas. É possível observar um pequeno grau de oscilação para cada uma das bases em todas as bibliotecas, mas nada que indique um grande viés para uma posição em específico das leituras, com exceção das bibliotecas geradas para esporocisto, nas quais a oscilação do conteúdo GC por posição foi maior (Figura 10E). Além disso, nos dois tipos de biblioteca avaliados, as sequências analisadas apresentam um conteúdo de A e T maior do que G e C, o que é esperado considerando que a medida do conteúdo das bases costuma ser espécie-específico, e em *S. mansoni* o conteúdo GC é de 36% (PROTASIO *et al.*, 2012). Ainda é possível observar um aumento de conteúdo de algumas bases nas últimas posições das leituras geradas pelo equipamento Ion Torrent PGM™ System, um artefato gerado pela tecnologia usada.

Vieses no conteúdo GC podem ser introduzidos em vários passos durante o preparo da biblioteca e posterior sequenciamento, como por exemplo, durante a amplificação da biblioteca por PCR (fragmentos com conteúdo GC médio são preferencialmente amplificados), ou amplificação do cluster, durante a seleção do tamanho dos fragmentos, utilizando técnicas aquecem o gel para recuperação do DNA, que por consequência reduzem a quantidade de fragmentos de baixo conteúdo GC, e ainda durante o sequenciamento em si devido à preferências da enzima utilizada (AIRD *et al.*, 2011).

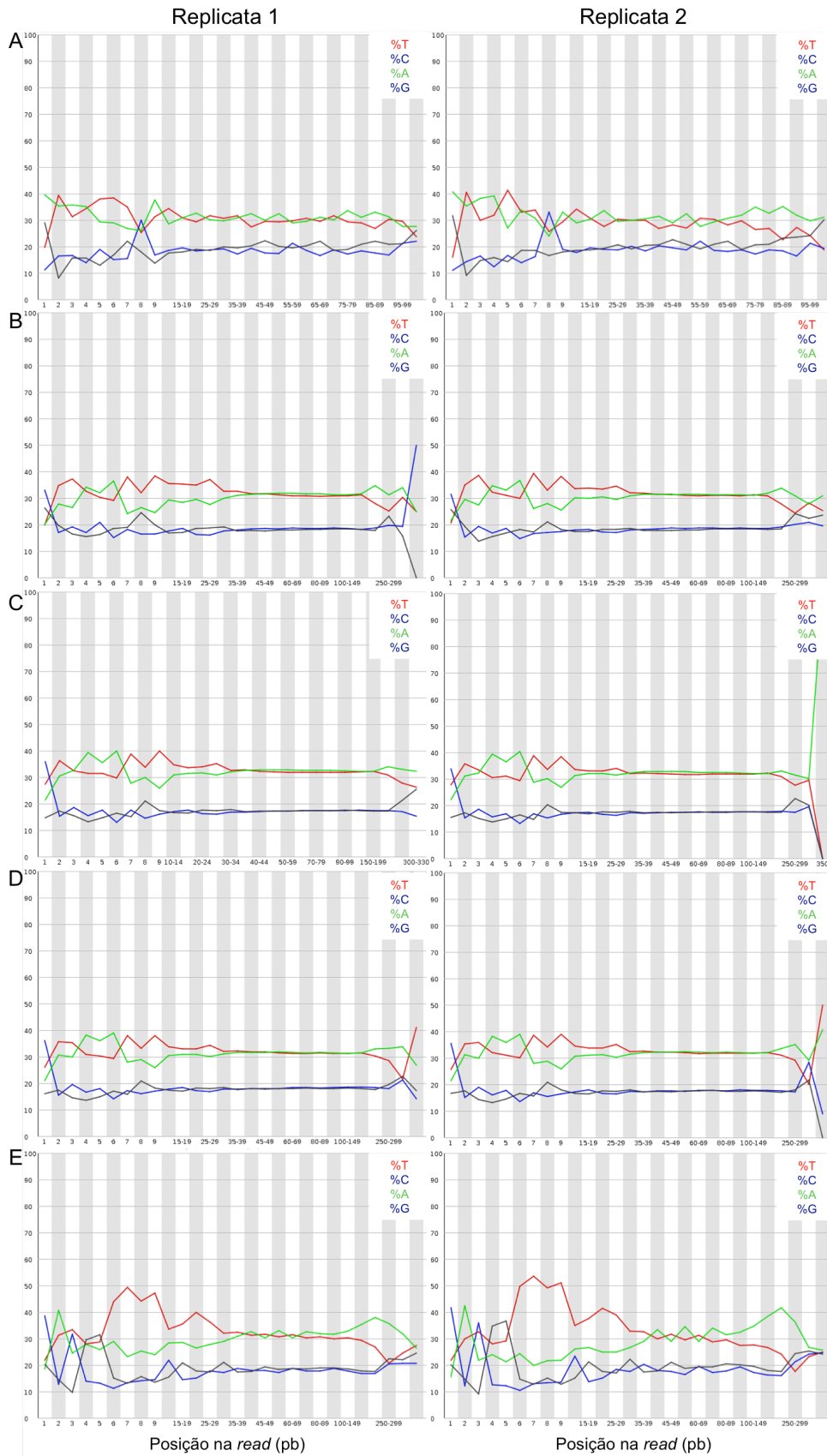


FIGURA 10 - CONTEÚDO DE BASE AO LONGO DA SEQUÊNCIA. A figura mostra a proporção de cada base por posição nas sequências geradas. **A** - Sequências geradas na plataforma Illumina HiSeq 2000, fase cercária, replicatas 1 e 2; **B** -

Sequências geradas na plataforma Ion Torrent PGM™ System, da fase adulto, replicatas 1 e 2; **C** – Sequências geradas na plataforma Ion Torrent PGM™ System, fase esquistossômulo, replicatas 1 e 2; **D** – Sequências geradas na plataforma Ion Torrent PGM™ System, fase miracídio, replicatas 1 e 2; **E** – Sequências geradas na plataforma Ion Torrent PGM™ System, fase esporocisto, replicatas 1 e 2.

d. Conteúdo de GC em cada posição ao longo da sequência

Em uma biblioteca randômica, espera-se que o conteúdo de GC não varie consideravelmente ao longo das posições de uma sequência. A Figura 11 mostra a proporção de G e C em cada posição para todas as sequências geradas e é possível perceber que há certa variação, principalmente para as primeiras e últimas bases nos dois tipos de bibliotecas, o que não é incomum tratando-se da natureza dos iniciadores utilizados na construção das bibliotecas nas plataformas Illumina HiSeq 2000 (Figura 11A) e Ion Torrent PGM™ System (Figura 11B-E). No entanto, no geral, observa-se uma tendência a manter o conteúdo GC constante ao longo da *read*. Além disso, assim como esperado para o conteúdo de bases nas sequências, o conteúdo GC oscilou em torno do mesmo valor em todas as bibliotecas.

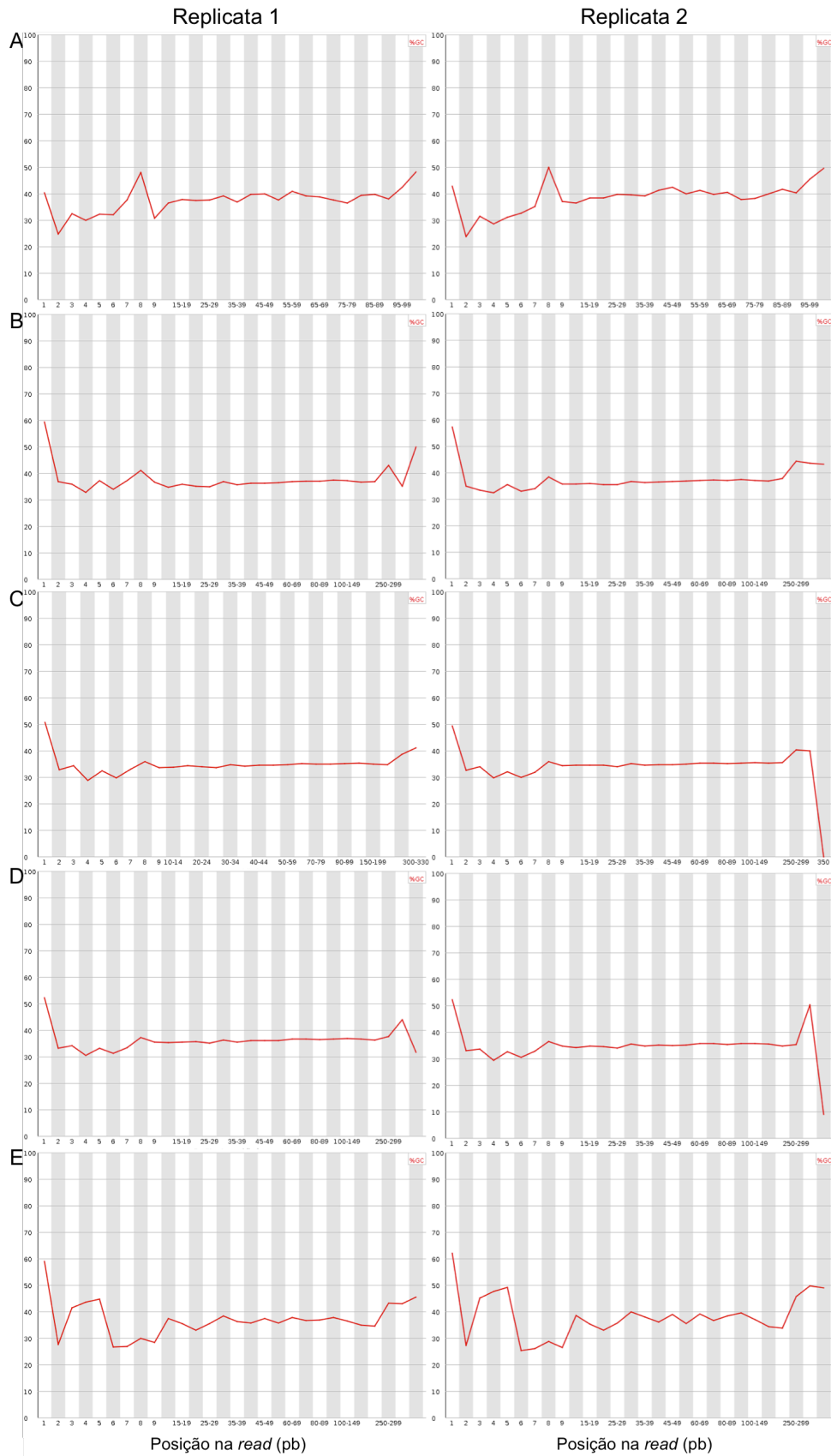


FIGURA 11 - CONTEÚDO GC POR BASE. A – Sequências geradas na plataforma Illumina, fase cercária, replicatas 1 e 2; **B** – Sequências geradas na plataforma Ion Torrent PGM, fase adulto, replicatas 1 e 2; **C** – Sequências geradas na plataforma Ion

Torrent PGM, fase esquistossômulo, replicatas 1 e 2; **D** – Sequências geradas na plataforma Ion Torrent PGM, fase miracídio, replicatas 1 e 2; **E** – Sequências geradas na plataforma Ion Torrent PGM, fase esporocisto, replicatas 1 e 2.

e. Distribuição do conteúdo de GC por sequência

O modelo de distribuição do conteúdo GC por sequência é uma curva normal onde o valor central correspondente ao teor global médio de GC do genoma. Nas bibliotecas de cercária, do tipo SL Trapping, (Figura 12A) observa-se que a distribuição real do conteúdo GC por leituras possui alguns picos de maior conteúdo GC, indicando a presença de sequências super-representadas nesta biblioteca, o que é de se esperar dado que a biblioteca gerada é enriquecida com transcritos que sofrem processamento por SLTS. Já nas bibliotecas do tipo SL Enriched, geradas no equipamento Ion Torrent PGM™ System (Figura 12B-E), observa-se que o pico de conteúdo GC também foi maior do que o esperado a partir da distribuição teórica, novamente reforçando que essa pode ser uma característica das bibliotecas enriquecidas em transcritos contendo a sequência do SL, entretanto, estas apresentam-se com uma distribuição mais uniforme.

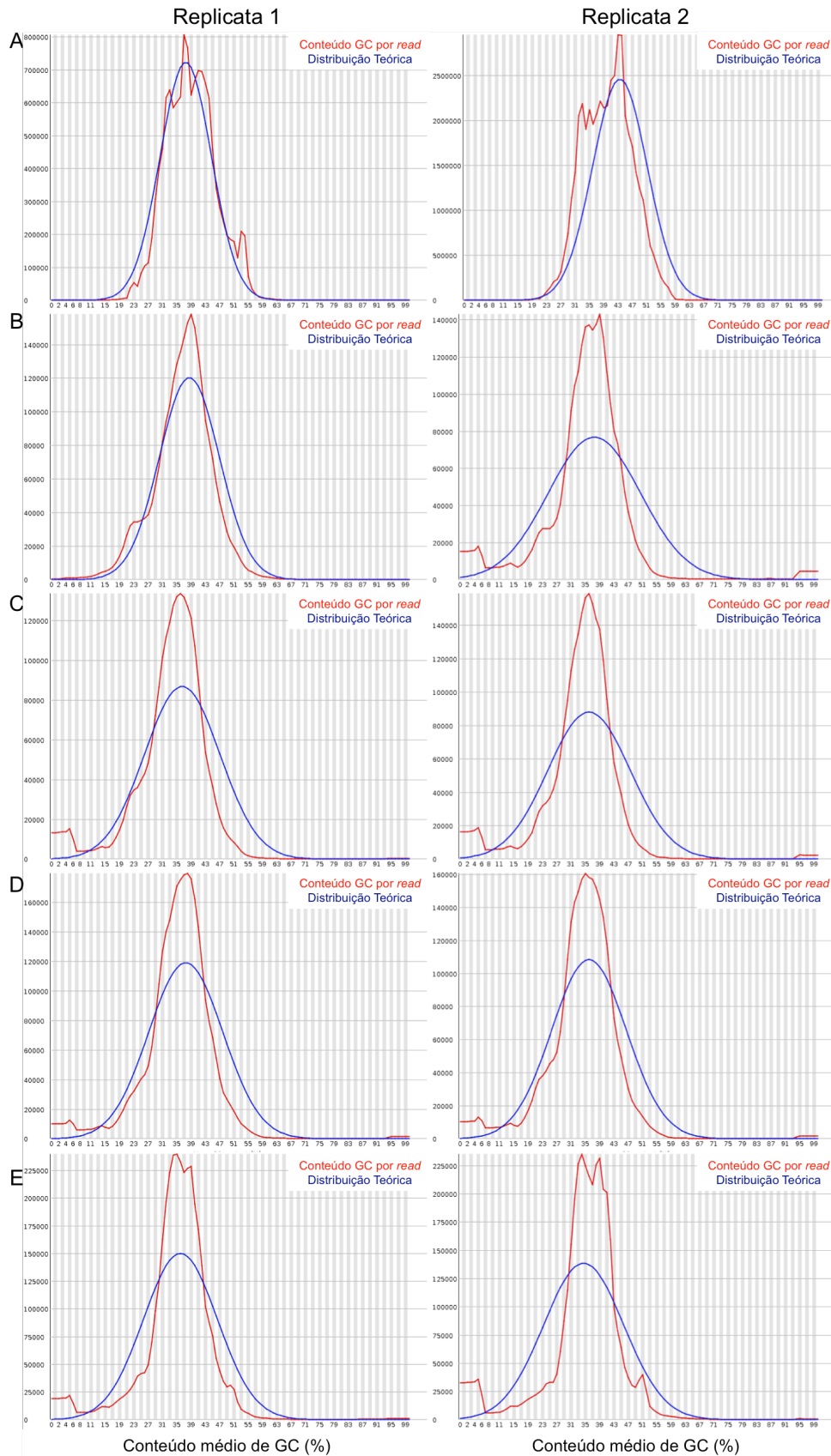


FIGURA 12 - CONTEÚDO GC POR SEQUÊNCIA. O gráfico representa a média do conteúdo de GC por *read* gerada comparado com uma distribuição normal modelada a partir do conteúdo de GC do genoma; **A** - Sequências geradas na plataforma

Illumina, fase cercária, replicatas 1 e 2; **B** – Sequências geradas na plataforma Ion Torrent PGM, fase adulto, replicatas 1 e 2; **C** – Sequências geradas na plataforma Ion Torrent PGM, fase esquistossômulo, replicatas 1 e 2; **D** – Sequências geradas na plataforma Ion Torrent PGM, fase miracídio, replicatas 1 e 2; **E** – Sequências geradas na plataforma Ion Torrent PGM, fase esporocisto, replicatas 1 e 2.

f. Sequências duplicadas

Esta análise é feita com um pequeno subconjunto dos dados, representativo do conjunto de todos os dados gerados. Espera-se que nesse subconjunto apenas uma pequena porção das sequências apresentem sequências duplicadas (redundantes), advindas de amplificação por PCR. O aumento do valor observado na porção final dos dois gráficos abaixo (Figura 13) nos permite inferir o grau de redundância nas duas bibliotecas. É importante notar que a biblioteca de cercária SL Trapping replicata 2 (Figura 13A) possui um alto grau de redundância, em torno de 85%, mostrando que muitas das sequências geradas vieram do sequenciamento de produtos de PCR. Este fato se deve a maior profundidade da biblioteca de cercária SL Trapping, na qual cerca de 30 milhões de leituras foram geradas na replicata 2, enquanto na replicata 1 cerca de 10 milhões de leituras foram geradas.

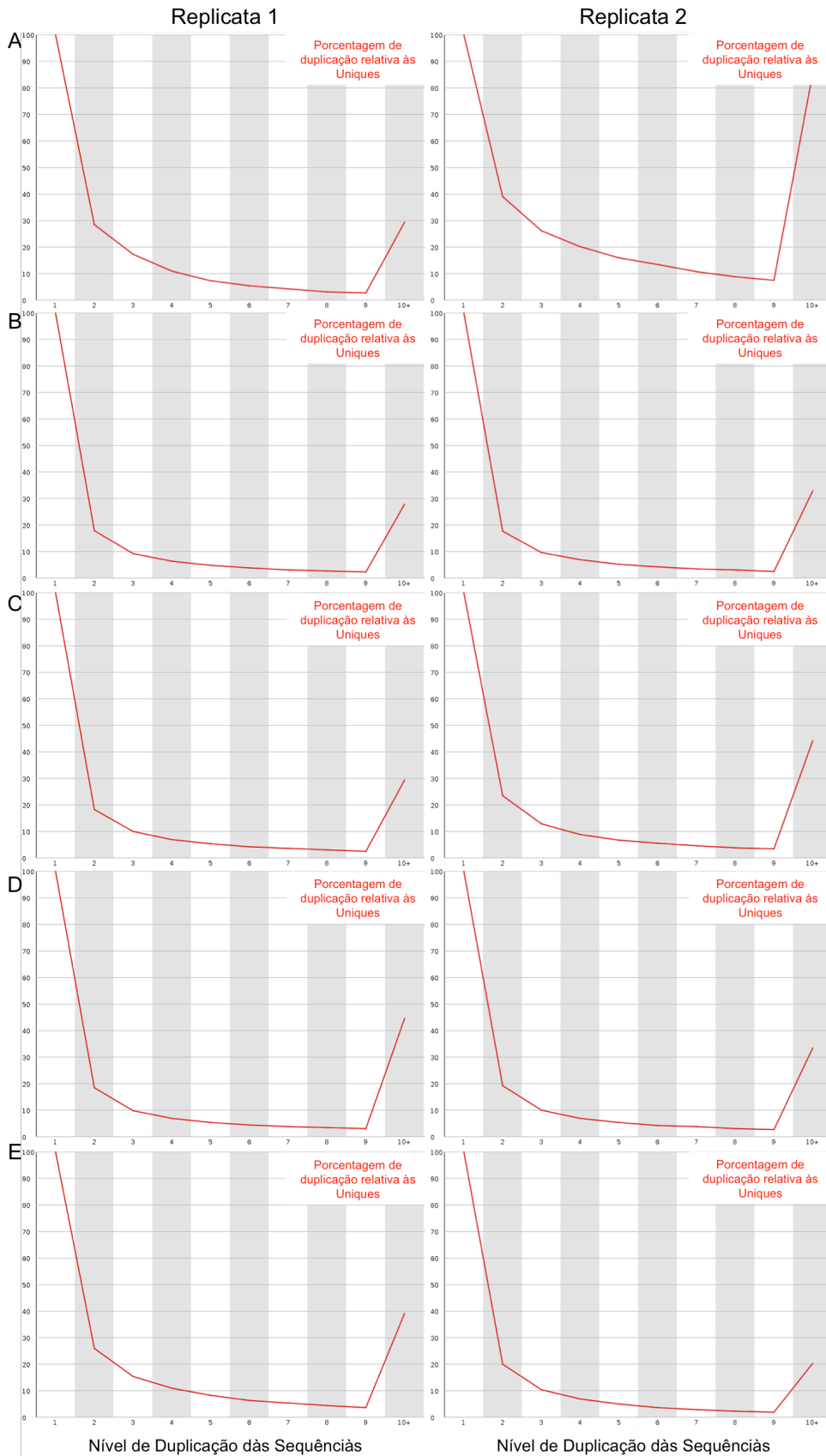


FIGURA 13 - SEQUÊNCIAS DUPLICADAS. O gráfico reporta o grau de duplicação (redundância) em um pequeno subconjunto de dados, mostrando o número relativo de sequências com diferentes graus de duplicação. **A** – Sequências

geradas na plataforma Illumina, fase cercária, replicatas 1 e 2; **B** – Sequências geradas na plataforma Ion Torrent PGM, fase adulto, replicatas 1 e 2; **C** – Sequências geradas na plataforma Ion Torrent PGM, fase esquistossômulo, replicatas 1 e 2; **D** – Sequências geradas na plataforma Ion Torrent PGM, fase miracídio, replicatas 1 e 2; **E** – Sequências geradas na plataforma Ion Torrent PGM, fase esporocisto, replicatas 1 e 2.

4.6. IDENTIFICAÇÃO E REMOÇÃO DO SL E TRATAMENTO DE QUALIDADE NAS *READS*

GERADAS NA PLATAFORMA DE ION TORRENT PGM™ SYSTEM

Buscou-se pela sequência do SL nas *reads* produzidas pelo equipamento Ion Torrent PGM, para posterior remoção desta. Ainda, após a avaliação da qualidade das *reads* geradas na plataforma Ion Torrent PGM™ System, foi realizado um tratamento de qualidade para remoção das bases de baixa qualidade e de *reads* menores do que 20 pb. As *reads* que permaneceram após o corte foram utilizadas nos passos seguintes e corresponderam a 97,36% do total de *reads* geradas. O resumo deste processo está descrito na Tabela 4.

4.7. OBTENÇÃO DAS *READS* EM BANCOS DE DADOS PÚBLICOS

Além das *reads* obtidas pelo sequenciamento em larga escala (plataformas Illumina HiSeq 2000 e Ion Torrent PGM™ System), outras 12 bibliotecas de RNA-Seq de diferentes fases do ciclo de vida do parasito *S. mansoni* foram encontradas no repositório SRA e incluídas no estudo. As características das bibliotecas estão listadas a seguir na Tabela 5.

A presença da sequência do SL foi verificada nas *reads* e utilizada como filtro no caso das sequências obtidas do SRA. Apenas uma pequena porcentagem das *reads* obtidas (0,04% em média) nas bases de dados públicos apresentaram a sequência do SL. Nas *reads* sequenciadas a partir da biblioteca SL Trapping, a porcentagem foi ainda menor, 0,00002%, o que é esperado, dada a metodologia utilizada para construção e sequenciamento das *reads*. A quantidade total de *reads*

e a quantidade total de *reads* com a sequência do SL por fase do ciclo de vida do parasito estão apresentadas na Tabela 6.

As *reads* filtradas contendo a sequência do SL, correspondendo a um total de 91.188 *reads*, foram aparadas para remoção da sequência do SL e constituíram o conjunto de dados RNA-Seq Filtered (Figura 4C).

TABELA 4 - READS DAS BIBLIOTECAS SL ENRICHED SEQUENCIADAS NA PLATAFORMA ION TORRENT PGM™ SYSTEM APÓS IDENTIFICAÇÃO E REMOÇÃO DA SEQUÊNCIA DO SL E FILTRO DE QUALIDADE.

| Nome da Biblioteca | Fase | Número inicial de reads | Número de reads com SL | % de reads com SL | Número final de reads após filtro de qualidade | % de reads de boa qualidade | Tamanho médio da read |
|---------------------------|-----------------|--------------------------------|-------------------------------|--------------------------|---|------------------------------------|------------------------------|
| SL Enriched A1 | Adulto | 2.586.306 | 113.404 | 4,38% | 2.498.801 | 96,62% | 169,22 pb |
| SL Enriched A2 | Adulto | 2.802.032 | 173.102 | 6,17% | 2.712.395 | 96,80% | 188,38 pb |
| SL Enriched E1 | Esquistossômulo | 2.562.467 | 44.496 | 1,73% | 2.501.894 | 97,64% | 178,64 pb |
| SL Enriched E2 | Esquistossômulo | 3.372.875 | 127.373 | 3,77% | 3.302.391 | 97,91% | 207,99 pb |
| SL Enriched M1 | Miracídio | 3.402.016 | 37.966 | 1,11% | 3.325.680 | 97,76% | 177,09 pb |
| SL Enriched M2 | Miracídio | 2.792.890 | 17.865 | 0,64% | 2.722.168 | 97,47% | 158,63 pb |
| SL Enriched S1 | Esporocisto | 3.479.656 | 42.872 | 1,23% | 3.317.906 | 95,35% | 163,04 pb |
| SL Enriched S2 | Esporocisto | 3.127.173 | 44.064 | 1,41% | 2.929.283 | 93,67% | 156,70 pb |

TABELA 5 - DADOS DE RNA-SEQ DE ESTUDOS DE TRANSCRIPTÔMICA DO PARASITO *S. MANSONI* DEPOSITADOS NO REPOSITÓRIO PÚBLICO SRA DO NCBI.

| Número de Acesso ¹ | Identificador | Espécie | Fase | Plataforma | Tipo de Biblioteca | Tamanho da read | Número de reads |
|-------------------------------|---------------|-------------------|-----------------|-----------------------------|--------------------|-----------------|-----------------|
| ERX009276 | ERR022873 | <i>S. mansoni</i> | Adulto_mix | Illumina Genome Analyzer II | Paired ends | 100 pb | 21.042.510 |
| ERX009279 | ERR022872 | <i>S. mansoni</i> | Cercária | Illumina Genome Analyzer II | Paired ends | 100 pb | 33.692.780 |
| ERX009282 | ERR022875 | <i>S. mansoni</i> | Cercária | Illumina Genome Analyzer II | Paired ends | 100 pb | 25.995.126 |
| ERX009278 | ERR022877 | <i>S. mansoni</i> | Cercária | Illumina Genome Analyzer II | Paired ends | 100 pb | 61.554.460 |
| ERX009277 | ERR022878 | <i>S. mansoni</i> | Cercária | Illumina Genome Analyzer II | Paired ends | 100 pb | 43.748.766 |
| ERX009284 | ERR022874 | <i>S. mansoni</i> | Esquistossômulo | Illumina Genome Analyzer II | Paired ends | 100 pb | 14.096.330 |
| ERX009281 | ERR022876 | <i>S. mansoni</i> | Esquistossômulo | Illumina Genome Analyzer II | Paired ends | 100 pb | 47.454.768 |
| ERX009274 | ERR022879 | <i>S. mansoni</i> | Esquistossômulo | Illumina Genome Analyzer II | Paired ends | 100 pb | 58.628.978 |
| ERX009285 | ERR022880 | <i>S. mansoni</i> | Esquistossômulo | Illumina Genome Analyzer II | Paired ends | 100 pb | 50.616.612 |
| ERX009280 | ERR022881 | <i>S. mansoni</i> | Esquistossômulo | Illumina Genome Analyzer II | Paired ends | 100 pb | 50.441.286 |
| ERX009283 | ERR022882 | <i>S. mansoni</i> | Esquistossômulo | Illumina Genome Analyzer II | Paired ends | 100 pb | 44.496.358 |
| ERX009275 | ERR022883 | <i>S. mansoni</i> | Esquistossômulo | Illumina Genome Analyzer II | Paired ends | 100 pb | 48.970.182 |

¹ Número de acesso do depósito no banco de dados SRA(NCBI).

TABELA 6 - ESTATÍSTICA DA IDENTIFICAÇÃO DA SEQUÊNCIA DO SL NAS *READS* DAS BIBLIOTECAS RNA-SEQ ADVINDAS DO BANCO DE DADOS PÚBLICO SRA.

| Fase do ciclo de vida do <i>S. mansoni</i> | Número total de <i>reads</i> | Número de <i>reads</i> com SL Match | % de <i>reads</i> com SL Match |
|---|-------------------------------------|--|---------------------------------------|
| Adulto | 21.042.510 | 15.856 | 0,07% |
| Cercária | 164.991.132 | 41.108 | 0,02% |
| Esquistossômulo | 314.704.514 | 125.412 | 0,04% |

4.8. MAPEAMENTO DAS *READS* NO GENOMA DE REFERÊNCIA

As *reads* dos conjuntos de dados SL Trapping, SL Enriched, RNA-Seq Filtered e RNA-Seq foram mapeadas no genoma de referência de *S. mansoni*. Na Tabela 7 são apresentadas todas as estatísticas dos alinhamentos com o programa TopHat2 (sequências geradas nas plataformas Illumina e Ion Torrent PGM) e Bowtie2 (sequências geradas na plataforma Ion Torrent PGM).

As *reads* geradas no equipamento Illumina HiSeq 2000 tiveram um mapeamento em torno de 80% usando somente o programa TopHat2, sendo 6% a porcentagem média de alinhamentos múltiplos, enquanto que as *reads* geradas na plataforma Ion Torrent PGM™ System tiveram uma porcentagem bem menor, em torno de 60%. As *reads* geradas na plataforma Ion Torrent PGM™ System não mapeadas na primeira rodada com o programa TopHat2 foram re-mapeadas utilizando-se o programa Bowtie2, alcançando-se dessa forma um mapeamento geral em torno de 97% das sequências, com 4,6% de alinhamentos múltiplos.

TABELA 7 - ESTATÍSTICAS DO ALINHAMENTO DAS READS GERADAS A PARTIR DAS DIFERENTES BIBLIOTECAS NO GENOMA DE REFERÊNCIA DE S. MANSONI.

| Biblioteca | SL Trapping | SL Trapping | RNA-Seq Filtered | RNA-Seq | RNA-Seq | RNA-Seq | RNA-Seq | RNA-Seq |
|--------------------------|---------------------|---------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| Número de Acesso (SRA) | SRR1134198 | SRR1134204 | - | ERR022873 | ERR022872 | ERR022875 | ERR022877 | ERR022878 |
| Fase | Cercária | Cercária | Mix | Adult | Cercária | Cercária | Cercária | Cercária |
| Replicata | R1 | R2 | R1 | R1 | R1 | R2 | R3 | R4 |
| Sequenciador | Illumina HiSeq 2000 | Illumina HiSeq 2000 | Illumina Genome Analyzer II | Illumina Genome Analyzer II | Illumina Genome Analyzer II | Illumina Genome Analyzer II | Illumina Genome Analyzer II | Illumina Genome Analyzer II |
| Alinhador | TopHat2 | TopHat2 | TopHat2 | TopHat2 | TopHat2 | TopHat2 | TopHat2 | TopHat2 |
| Categoria das reads | single | single | paired-ends | paired-ends | paired-ends | paired-ends | paired-ends | paired-ends |
| Total de reads | 11.520.178 | 30.332.894 | 182.376 | 21.042.510 | 33.692.780 | 25.995.126 | 61.554.460 | 43.748.766 |
| Reads alinhadas | 10.337.019 | 26.464.540 | 169.974 | 15.939.109 | 24.021.480 | 20.242.179 | 55.551.095 | 39.954.524 |
| % Reads Alinhadas | 90% | 87% | 93% | 76% | 71% | 78% | 90% | 91% |
| Alinhamento aos pares | - | - | 161.000 | 14.795.216 | 22.182.224 | 18.926.778 | 53.894.070 | 38.925.894 |
| % Alinhamentos aos pares | - | - | 95% | 93% | 92% | 94% | 97% | 97% |
| Múltiplos Alinhamentos | 477.703 | 1.961.952 | 18.391 | 3.026.082 | 6.351.948 | 3.558.948 | 11.742.720 | 8.400.609 |
| % Múltiplos Alinhamentos | 5% | 7% | 11% | 19% | 26% | 18% | 21% | 21% |

TABELA 7 - ESTATÍSTICAS DO ALINHAMENTO DAS READS GERADAS A PARTIR DAS DIFERENTES BIBLIOTECAS NO GENOMA DE REFERÊNCIA DE S. MANSONI.

| Biblioteca | RNA-Seq | RNA-Seq | RNA-Seq | RNA-Seq | RNA-Seq | RNA-Seq | RNA-Seq |
|--------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|
| Número de Acesso (SRA) | ERR022874 | ERR022876 | ERR022879 | ERR022880 | ERR022881 | ERR022882 | ERR022883 |
| Fase | Esquistossômulo | Esquistossômulo | Esquistossômulo | Esquistossômulo | Esquistossômulo | Esquistossômulo | Esquistossômulo |
| Replicata | R1 | R2 | R3 | R4 | R5 | R6 | R7 |
| Sequenciador | Illumina Genome Analyzer II TopHat2 | Illumina Genome Analyzer II TopHat2 | Illumina Genome Analyzer II TopHat2 | Illumina Genome Analyzer II TopHat2 | Illumina Genome Analyzer II TopHat2 | Illumina Genome Analyzer II TopHat2 | Illumina Genome Analyzer II TopHat2 |
| Alinhador | paired-ends | paired-ends | paired-ends | paired-ends | paired-ends | paired-ends | paired-ends |
| Categoria das reads | paired-ends | paired-ends | paired-ends | paired-ends | paired-ends | paired-ends | paired-ends |
| Total de reads | 14.096.330 | 47.454.768 | 58.628.978 | 50.616.612 | 50.441.286 | 44.496.358 | 48.970.182 |
| Reads alinhadas | 9.330.473 | 35.370.176 | 48.604.040 | 39.831.980 | 39.692.697 | 35.118.941 | 38.585.052 |
| % Reads Alinhadas | 66% | 75% | 83% | 79% | 79% | 79% | 79% |
| Alinhamento aos pares | 8.856.942 | 33.919.810 | 44.635.282 | 35.829.120 | 35.253.026 | 30.297.874 | 34.148.994 |
| % Alinhamentos aos pares | 95% | 96% | 92% | 90% | 89% | 86% | 89% |
| Múltiplos Alinhamentos | 1.920.134 | 7.019.591 | 6.525.086 | 6.118.816 | 7.577.651 | 6.090.464 | 6.117.843 |
| % Múltiplos Alinhamentos | 21% | 20% | 13% | 15% | 19% | 17% | 16% |

TABELA 7 - ESTATÍSTICAS DO ALINHAMENTO DAS READS GERADAS A PARTIR DAS DIFERENTES BIBLIOTECAS NO GENOMA DE REFERÊNCIA DE S. MANSONI.

| Biblioteca | SL Enriched | | SL Enriched | | SL Enriched | | SL Enriched | | SL Enriched | | SL Enriched | |
|--------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|
| | Adulto | Adulto | Esquistossômulo | Esquistossômulo | Miracídio | Miracídio | Miracídio | Miracídio | Esporocisto | Esporocisto | Esporocisto | Esporocisto |
| Fase | R1 | R2 | R1 | R2 | R1 | R2 | R1 | R2 | R1 | R2 | R1 | R2 |
| Replicata | Ion Torrent PGM™ System | Ion Torrent PGM™ System | Ion Torrent PGM™ System | Ion Torrent PGM™ System | Ion Torrent PGM™ System | Ion Torrent PGM™ System | Ion Torrent PGM™ System | Ion Torrent PGM™ System | Ion Torrent PGM™ System | Ion Torrent PGM™ System | Ion Torrent PGM™ System | Ion Torrent PGM™ System |
| Sequenciador | TopHat2 & Bowtie2 | TopHat2 & Bowtie2 | TopHat2 & Bowtie2 | TopHat2 & Bowtie2 | TopHat2 & Bowtie2 | TopHat2 & Bowtie2 | TopHat2 & Bowtie2 | TopHat2 & Bowtie2 | TopHat2 & Bowtie2 | TopHat2 & Bowtie2 | TopHat2 & Bowtie2 | TopHat2 & Bowtie2 |
| Alinhador | single | single | single | single | single | single | single | single | single | single | single | single |
| Categoria das reads | 2.498.801 | 2.712.395 | 2.501.894 | 3.302.391 | 3.325.680 | 2.722.168 | 3.325.680 | 2.722.168 | 3.325.680 | 3.325.680 | 2.722.168 | 2.722.168 |
| Número inicial de reads | 2.400.027 | 2.595.976 | 2.435.828 | 3.223.396 | 3.243.958 | 2.640.768 | 3.243.957 | 2.640.768 | 3.243.957 | 3.243.957 | 2.640.772 | 2.640.772 |
| Reads alinhadas | 96% | 96% | 97% | 98% | 98% | 97% | 98% | 97% | 98% | 98% | 97% | 97% |
| % Reads Alinhadas | 129.328 | 110.561 | 90.437 | 122.512 | 125.913 | 89.370 | 125.913 | 89.370 | 125.781 | 125.781 | 89.374 | 89.374 |
| Múltiplos Alinhamentos | 5% | 4% | 4% | 4% | 4% | 3% | 4% | 3% | 7% | 7% | 6% | 6% |
| % Múltiplos Alinhamentos | | | | | | | | | | | | |

4.9. COMPARAÇÃO ENTRE OS TRÊS MÉTODOS UTILIZADOS PARA OBTENÇÃO DE SEQUÊNCIAS PROCESSADAS POR SLTS

As três metodologias utilizadas para obtenção das sequências foram comparadas com relação aos genes identificados que sofrem processamento por SLTS.

A profundidade da biblioteca SL Trapping foi muito maior do que a biblioteca final resultante da filtragem por sequências que continham o SL nas *reads* advindas do SRA (RNA-Seq Filtered), mesmo utilizando uma quantidade muito grande de *reads* durante a fase de triagem (das 503.979.020 *reads* iniciais, somente 182.376 *reads* continham a sequência do SL e foram consideradas na análise posterior, contra 10.220.382 *reads* produzidas no sequenciamento da biblioteca SL Trapping da fase cercária). Já para as bibliotecas do tipo SL Enriched, foi produzido um número total de 24.125.415 *reads* construídas a partir das fases adulto, esquistosômulo, miracídio e esporocisto.

Um diagrama de Venn (Figura 14) ilustra a relação entre os transcritos identificados em cada biblioteca. A biblioteca SL Trapping apresentou um número maior de transcritos exclusivos (3.280) em relação às demais, SL Enriched (375) e RNA-Seq Filtered (41), como esperado, dadas as características das bibliotecas do tipo SL Trapping e a sua alta profundidade, uma vez que a capacidade de detectar e quantificar uma maior diversidade de transcritos, assim como transcritos raros, aumenta com o aumento da profundidade do sequenciamento (TARAZONA *et al.*, 2011). Dessa forma, fomos capazes de identificar um número maior de genes processados por SLTS utilizando a metodologia SL Trapping na plataforma Illumina HiSeq 2000, representado pelo

círculo verde. As três diferentes bibliotecas compartilharam um grande número de transcritos (1.363), mostrando a alta reprodutibilidade das metodologias utilizadas.

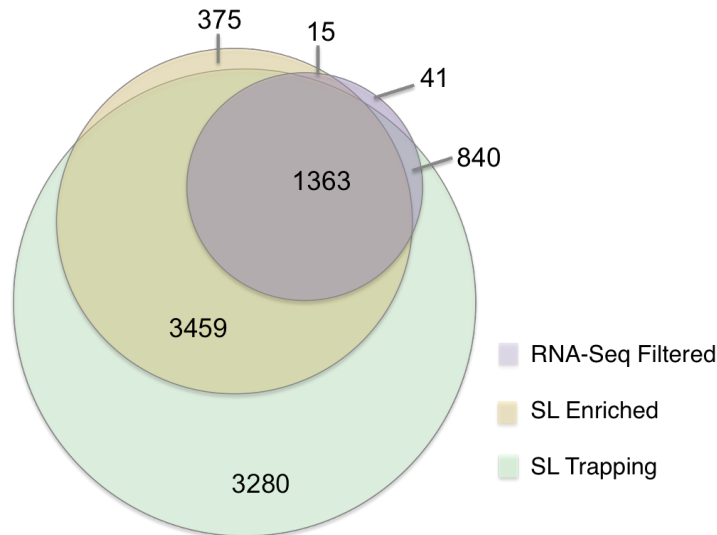


FIGURA 14 - DIAGRAMA DE VENN REPRESENTANDO OS GENES PROCESSADOS POR SLTS IDENTIFICADOS POR MÉTODOS DIFERENTES. O círculo em verde representa o conjunto de genes identificados na biblioteca SL Trapping. O círculo em amarelo representa os genes que foram identificados utilizando-se as bibliotecas do tipo SL Enriched. O círculo em lilás representa o conjunto de genes identificados na biblioteca resultante da filtragem das *reads* por sequências contendo o SL (RNA-Seq Filtered).

Considerando todos os genes que sofrem SLTS observados, utilizando os três diferentes tipos de bibliotecas, nós identificamos 8.457 genes processados por SLTS, representando 63% de todos os 13.322 genes anotados na 5ª versão do genoma de referência de *S. mansoni*. Essa proporção aumenta para 77% se considerarmos apenas nos genes codificadores de proteínas (8.344 genes processados por SLTS em 10.787 genes codificadores de proteínas). Esse valor é substancialmente maior que os valores previamente estimados na literatura. Em um estudo focando em um pequeno conjunto de genes, foi estimado que aproximadamente 10% dos genes sofriam processamento por SLTS em *Schistosoma spp.* (DAVIS; HARDWICK; TAVERNIER, 1995). Uma estimativa

semelhante foi obtida por Protasio e colaboradores (2012), que empregaram a técnica padrão de RNA-Seq.

Nossas observações elevam o mecanismo de SLTS em *S. mansoni* a outro nível de prevalência, o mesmo observado em estudos realizados no nematoide *C. elegans* e no urocordado *C. intestinalis*, nos quais 70% (ALLEN *et al.*, 2011) e 58% (MATSUMOTO *et al.*, 2010) de todos os transcritos são processados por SLTS, respectivamente. O conjunto de eventos de processamento por SLTS apresentado aqui, apesar de extenso (compreende eventos nas fases cercária, adulto, esquistossômulo, miracídio e esporocisto), é incompleto, uma vez que a baixa profundidade dos sequenciamentos gerados na plataforma Ion Torrent PGM™ System não permitiu a detecção de transcritos de baixa expressão. Dessa forma, esperamos que este conjunto seja estendido no futuro, através do estudo de transcritos processados por SLTS em outras fases/condições experimentais, utilizando tecnologias que forneça uma alta profundidade de sequenciamento, uma vez que é sabido que a expressão gênica varia com as condições ambientais e o processamento por SLTS varia entre as diferentes fases do ciclo de vida de *S. mansoni* (MOURÃO *et al.*, 2013).

Para fins didáticos, dividirei o restante da seção de resultados em duas partes. Em um primeiro momento tratarei do estudo do mecanismo de SLTS no parasito *S. mansoni*, utilizando as bibliotecas SL Trapping e RNA-Seq Filtered, focando principalmente na fase de cercária. Já na segunda parte me concentrarei na identificação da regulação diferencial de genes processados por SLTS entre diferentes estágios do ciclo de vida do parasito utilizando as bibliotecas SL Enriched.

4.10. O MECANISMO DE SLTS NO PARASITO *S. MANSONI*

4.10.1. GENES IDENTIFICADOS: COMPARAÇÃO ENTRE AS RÉPLICAS BIOLÓGICAS E OS DIFERENTES TIPOS DE CONJUNTO DE DADOS

Foi observada uma forte correlação qualitativa entre as replicatas biológicas SL Trapping. Na replicata SL Trapping 1 foram identificados 8.506 *loci*, enquanto que na replicata SL Trapping 2, identificamos 7.495 *loci*, sendo que 70% desses estão presentes em ambas as bibliotecas (Figura 15A). Uma correlação quantitativa também foi identificada entre as replicatas (coeficiente de correlação de Pearson igual a 0,8 , Figura 15B), mostrando que os alvos de SLTS são altamente reproduzíveis, o que enfatiza a qualidade dos dados gerados.

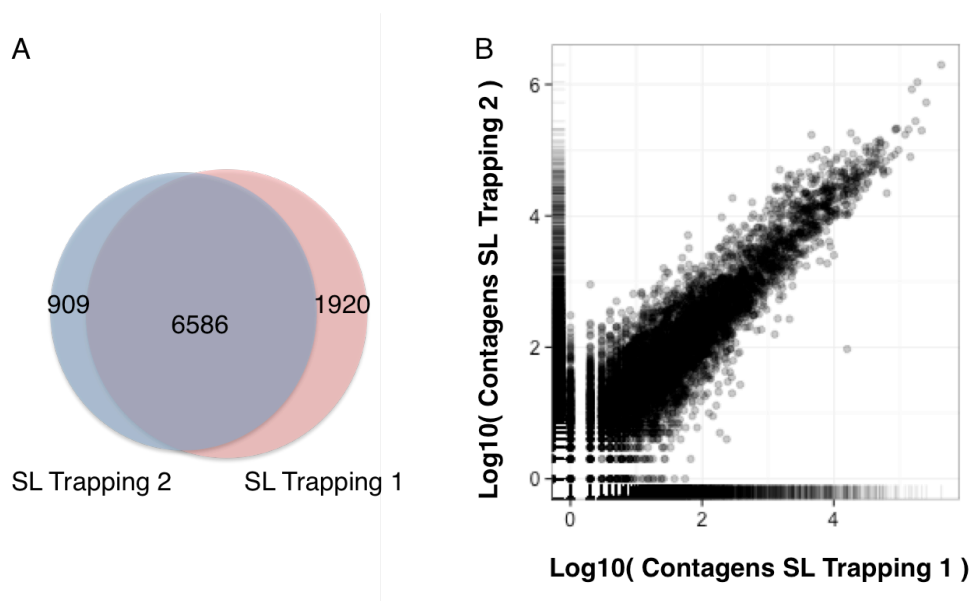


FIGURA 15 - COMPARAÇÃO ENTRE AS DUAS RÉPLICAS BIOLÓGICAS DAS BIBLIOTECAS SL TRAPPING. A- Genes detectados nas duas réplicas biológicas. **B -** Correlação entre as réplicas biológicas (PCC = 0,86).

Já para a biblioteca gerada pela compilação de *reads* filtradas de experimentos de RNA-Seq (RNA-Seq Filtered), obtivemos um número muito pequeno de transcritos (2.459, círculo rosa na Figura 16A) processados por

SLTS, uma vez que o número de *reads* recuperadas contendo o SL foi baixo (<0,04% do total de *reads* de RNA-Seq).

Entretanto, apesar da origem heterogênea de ambos os tipos de biblioteca, as bibliotecas SL Trapping abrangeram 95% dos transcritos processados por SLTS contidos na biblioteca RNA-Seq Filtered (Figura 16A). Os 123 transcritos remanescentes, encontrados somente no conjunto de dados RNA-Seq Filtered, podem ser explicados pela presença de RNA das fases esquistossômulo e adulto, além da fase cercária, nesse conjunto de dados, ao passo que o conjunto de dados SL Trapping contém somente transcritos expressos na fase cercária. Ainda, desses 123 transcritos, apenas 22 são expressos na fase cercária (círculo lilás na Figura 16B).

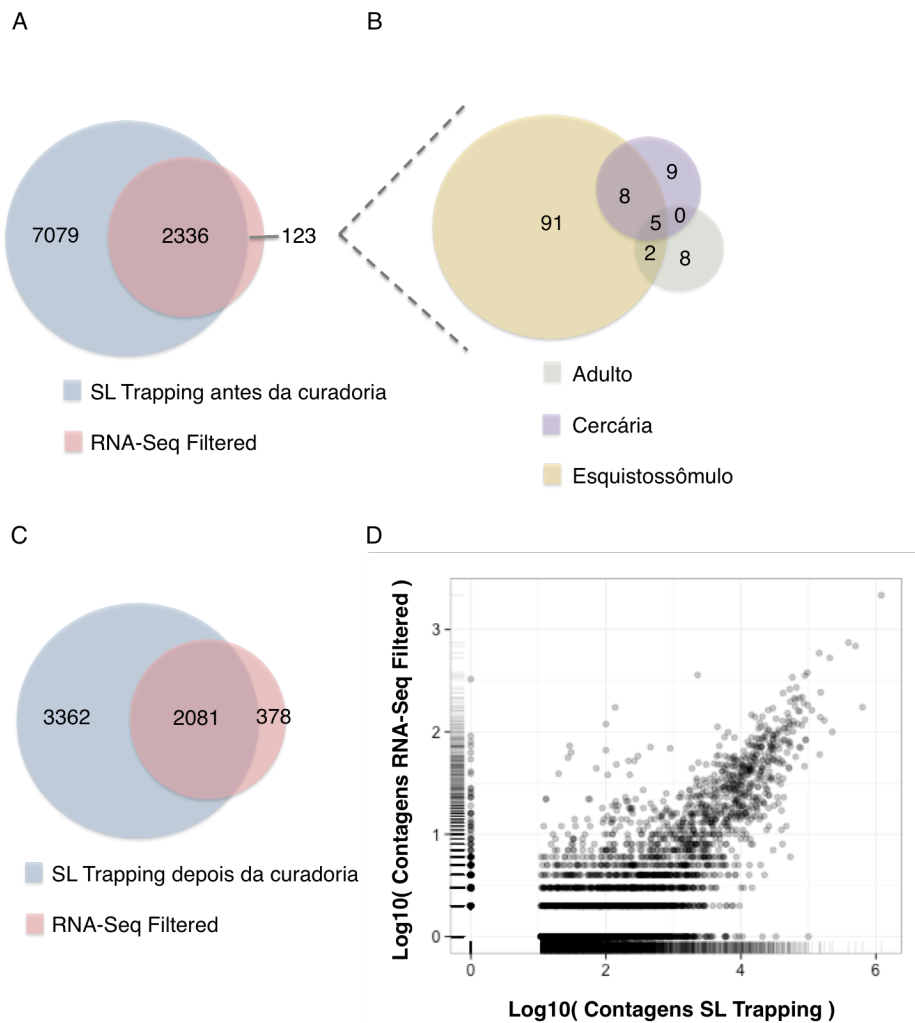


FIGURA 16 - COMPARAÇÃO ENTRE OS CONJUNTOS DE DADOS SL TRAPPING E RNA-SEQ FILTERED. **A** - Genes detectados nos dois tipos de bibliotecas antes da curadoria da biblioteca SL Trapping; **B** - Genes exclusivos da biblioteca RNA-Seq Filtered divididos por fases do ciclo de vida do *S. mansoni*; **C** - Comparação entre os dois conjuntos de dados após curadoria da SL Trapping; **D** - Correlação entre os dois conjuntos de dados (PCC = 0,5).

Para aumentar a confiabilidade dos nossos dados, fizemos uma curadoria dos mesmos (loci devem ser reprodutíveis em ambas as réplicas e possuírem média de contagens maior ou igual a 10 *reads*). De um total de 9.415 *loci* identificados nas duas réplicas biológicas de SL Trapping (Figura 15A), restringimos para 5.443 genes processados por SLTS após curadoria, círculo azul na Figura 16C, para serem considerados nas análises posteriores. Dessa forma, após a curadoria, o conjunto de dados SL Trapping abrangeu 85% dos genes processados por SLTS identificados no conjunto de dados RNA-Seq Filtered

(Figura 16C). Os 2.081 eventos de SLTS (Figura 16C) em comum entre SL Trapping and RNA-Seq Filtered representam genes que sofrem SLTS mais frequentemente (mediana de contagens igual a 507 para eventos compartilhados por ambos os conjuntos de dados *versus* mediana 44 para eventos detectados exclusivamente no conjunto de dados SL Trapping). Ainda, quantitativamente, os dois conjuntos exibem correlação moderada (PCC= 0,5 , Figura 16D).

Mesmo com diferenças consideráveis na forma como as duas bibliotecas foram originadas, a análise visual do mapeamento das sequências de ambas no genoma confirma a confiabilidade e reprodutibilidade dos dados. Para um dado gene, os picos de cobertura com *reads* advindas de ambos os tipos de bibliotecas foram encontrados exatamente na mesma posição do gene, indicando corretamente onde ocorre a inserção da sequência do SL. Como exemplo, nós analisamos os sinais de SLTS no gene Smp_034190, que mostra sinais de SLTS comuns aos dois conjuntos de dados nos exons terminais 5' e 3' do transcrito. Apesar de o conjunto de dados contidos na biblioteca RNA-Seq Filtered mostrar um sinal forte nos exons 2 e 3 e este sinal não ser reproduzido no conjunto SL Trapping, essa situação pode ser explicada pela diversidade de transcritos expressos nos estágios do ciclo de vida acessados em cada tipo de biblioteca (cercária *versus* cercária, adulto e esquistossômulo) ou pelo tipo de biblioteca (*single versus paired-ends*). O conjunto SL Trapping, em contraste, apresenta um sinal fraco no exon 5, que não é reproduzido pelo conjunto RNA-Seq Filtered, possivelmente por causa de sua baixa sensibilidade.

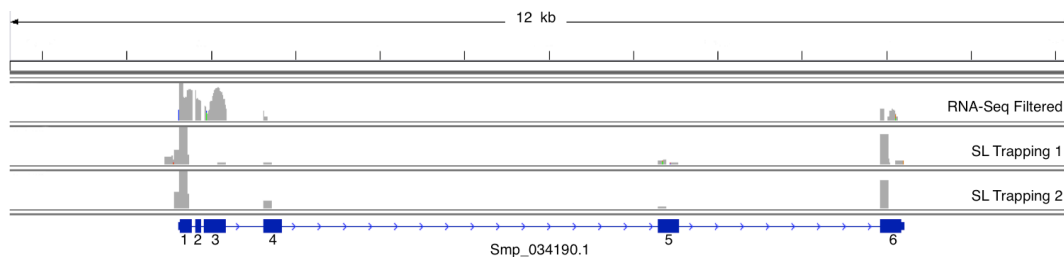


FIGURA 17 - SÍTIOS DE INSERÇÃO DA SEQUÊNCIA DO SL NOS DIFERENTES TIPOS DE BIBLIOTECA. Esquema produzido no IGV da estrutura exon-ínton do transcrito Smp_034190.1 com as contagens de *reads* superpostas dos conjuntos de dados RNA-Seq Filtered (faixa de cima), SL Trapping 1 (faixa do meio), e SL Trapping 2 (faixa de baixo).

Em conjunto, nossos resultados demonstram que SL Trapping é uma técnica poderosa e reprodutível capaz de detectar sinais também identificados pelo método de RNA-Seq convencional, entretanto com uma resolução incomparavelmente maior, capaz de capturar até mesmo eventos de SLTS que ocorrem em baixa frequência.

4.10.2. IDENTIFICAÇÃO DE SEQUÊNCIAS DO SL NO GENOMA DE *S. MANSONI*

Em 25% dos transcritos de dinoflagelados foi observada a presença de sequências SL adicionais em *tandem* nos transcritos, provavelmente resultado de uma retro-transcrição eventual de um transcrito processado por SLTS que em seguida foi incorporado ao genoma (SLAMOVITS; KEELING, 2008). Procurando no genoma de *S. mansonii* pela sequência do SL, além dos 5 genes de SL RNA anotados na 5ª versão do genoma de referência do parasito e 7 outros genes de SL RNA não anotados, nós detectamos 32 *loci* no genoma contendo a sequência do SL. Destes, 11 sítios foram encontrados em genes previamente anotados, sendo 6 deles identificados como elementos transponíveis (Tabela 8), os quais são provavelmente cDNAs que sofreram processamento por SLTS e foram retrotranscritos, sendo os retrotransposons os elementos mais comumente repetidos no genoma (BERRIMAN *et al.*, 2009; DEMARCO *et al.*, 2004).

TABELA 8 – GENES CONCATENADOS COM A SEQUÊNCIA DO SL.

| | Cromossoma | | Spliced leader | | Gene | Comentário |
|---------|------------|----------|----------------|-----|--------------|-----------------------|
| | Início | Fim | Início | Fim | | |
| Chr_1 | 8917800 | 8917767 | 1 | 34 | Smp_123830.1 | - |
| Chr_1 | 48421666 | 48421692 | 10 | 36 | - | - |
| Chr_1 | 57525190 | 57525163 | 8 | 35 | - | - |
| Chr_2 | 42605 | 42638 | 3 | 36 | - | Elemento Transponível |
| Chr_2 | 48216 | 48189 | 3 | 30 | - | - |
| Chr_2 | 2838790 | 2838761 | 7 | 36 | Smp_170980.1 | - |
| Chr_2 | 3230664 | 3230638 | 1 | 27 | Smp_133980.1 | Elemento Transponível |
| Chr_2 | 7626236 | 7626222 | 22 | 36 | - | - |
| Chr_2 | 15182842 | 15182810 | 4 | 36 | Smp_105420.2 | - |
| Chr_2 | 22148229 | 22148248 | 1 | 20 | - | - |
| Chr_3 | 8035976 | 8036000 | 1 | 25 | - | - |
| Chr_5 | 4122475 | 4122510 | 1 | 36 | - | Elemento Transponível |
| Chr_5 | 6425379 | 6425348 | 5 | 36 | - | - |
| Chr_5 | 7448107 | 7448142 | 1 | 36 | - | - |
| Chr_6 | 378665 | 378700 | 1 | 36 | - | - |
| Chr_6 | 2654089 | 2654115 | 6 | 32 | Smp_123190.1 | - |
| Chr_6 | 7123848 | 7123883 | 1 | 36 | - | - |
| Chr_6 | 7149339 | 7149374 | 1 | 36 | - | - |
| Chr_6 | 9818358 | 9818390 | 4 | 36 | - | - |
| Chr_7 | 1011639 | 1011669 | 6 | 36 | Smp_127680.1 | - |
| Chr_W | 2532171 | 2532139 | 4 | 36 | Smp_143010.1 | - |
| Chr_W | 16217864 | 16217833 | 5 | 36 | - | - |
| Chr_W | 18163122 | 18163153 | 2 | 33 | - | - |
| Chr_W | 18199959 | 18199990 | 2 | 33 | Smp_095000.2 | Elemento Transponível |
| Chr_W | 31525431 | 31525400 | 5 | 36 | - | - |
| Chr_W | 36429422 | 36429446 | 7 | 31 | Smp_025150.1 | - |
| Chr_W | 54185929 | 54185905 | 6 | 30 | - | - |
| SC_0097 | 343513 | 343544 | 5 | 36 | - | - |
| SC_0125 | 495991 | 495967 | 5 | 29 | Smp_032150.1 | - |
| SC_0129 | 880047 | 880016 | 5 | 36 | Smp_212530.1 | - |
| SC_0147 | 222428 | 222459 | 5 | 36 | - | Elemento Transponível |
| SC_0201 | 292767 | 292740 | 5 | 32 | - | Elemento Transponível |

A figura a seguir (Figura 18) mostra uma região no genoma onde é possível ver a sequência do SL concatenada a um elemento transponível:

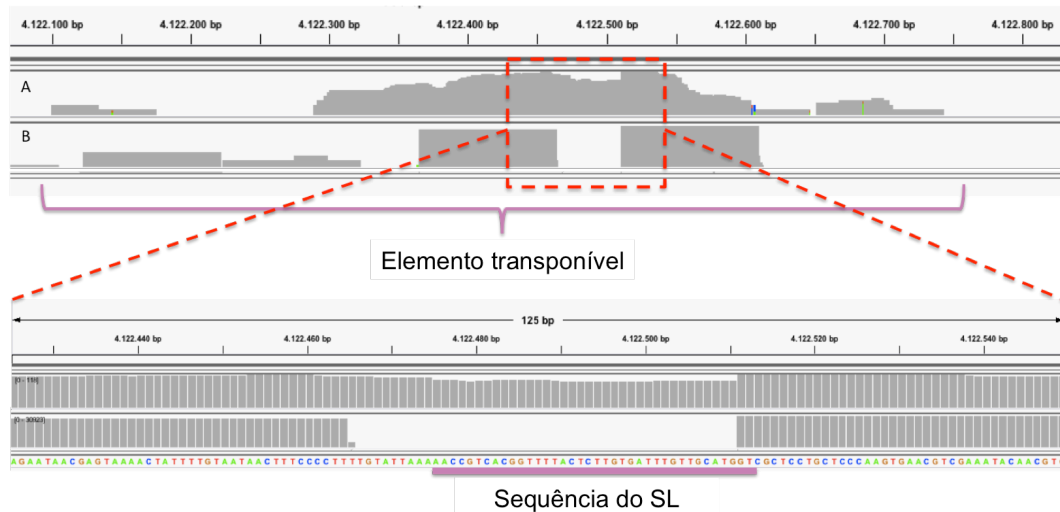


FIGURA 18 - SL CONCATENADO A UM GENE NO GENOMA DE *S. MANSONI*. Foram encontradas sequências SL concatenadas a diferentes genes no genoma do parasito, dentre eles, o elemento transponível Jockey_Ele1_ORF2, em evidência na figura.

Os 17 sítios restantes estão localizados em regiões onde não existem genes anotados, indicando que *loci* adicionais contendo a sequência do SL podem existir. Para todas as nossas análises subsequentes, os genes identificados contendo a sequência do SL concatenada não foram considerados.

4.10.3. IDENTIFICAÇÃO DE TRANSCRITOS POLICISTRÔNICOS EM *S. MANSONI*

A adição da sequência do SL em cinetoplastídeos e em um subconjunto de genes em outros organismos como nematóides, cordados e platelmintos, tem como função resolver mRNAs policistrônicos em mRNAs monocistrônicos capeados. No caso do nematóide *C. elegans*, primeiro animal onde foi descoberta a presença dos genes organizados em *operons*, uma sequência do SL diferente é utilizada para resolver os transcritos policistrônicos, denominada SL2 (SPIETH *et al.*, 1993).

Em *S. mansoni* foi reportada a presença de um possível transcrito dicistrônico: foi encontrado um gene intimamente ligado *upstream* ao gene da enolase (que sofre *trans-splicing*). Este gene produz um mRNA que codifica uma

proteína de ligação a ubiquinol, a UbCRBP, que é um componente do complexo de ubiquinol-citocromo-C redutase. A distância entre o sítio de poliadenilação da UbCRBP e o sítio acceptor de *trans-splicing* da enolase é excepcionalmente curta, de apenas 54 nucleotídeos (Davis e Hodgson 1997). Dessa forma esses dois genes devem ser transcritos como um dicistron que é resolvido em 2 monocistrons por *trans-splicing*. Observamos a partir dos nossos dados que o transcrito de enolase recebe o SL nos exons 1, 2, 3 e 4. Também observamos que o próprio transcrito de UbCRBP é alvo de *trans-splicing* (Figura 19).

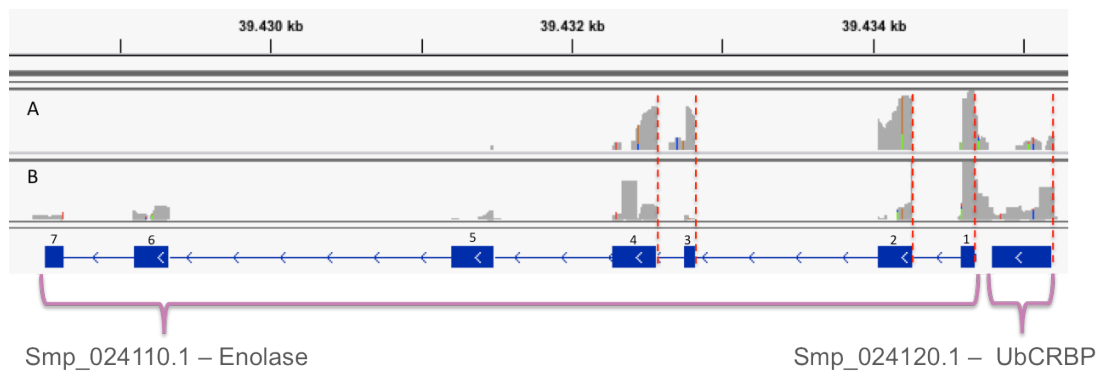


FIGURA 19 - INSERÇÃO DO SL NO DICISTRON ENOLASE E UBCRBP . A – cobertura da biblioteca RNA-Seq Filtered. **B** - cobertura da biblioteca SL Trapping. As linhas e as caixas em azul indicam os modelos gênicos e as setas o sentido da transcrição – versão 5 do genoma de *S. mansoni*.

Pelos nossos resultados e diferentemente do que ocorre em *C. elegans*, parece que em *S. mansoni* a mesma sequência do SL é utilizada tanto no *trans-splicing* de transcritos monocistrônicos quanto de policistrônicos.

Protasio e colaboradores (2012), utilizando resultados do sequenciamento em massa de transcritos do parasito também identificaram outros 46 possíveis *operons* (com distâncias intergênicas de até 200 pb) e conseguiram validar a existência de 4 pela técnica de RT-PCR e sequenciamento de Sanger.

Na tentativa de identificarmos mais transcritos policistrônicos, foi avaliada a ocorrência do *trans-splicing* nos conjuntos de genes com distância

intergênica de até 200 pb. Ao examinar a distribuição genômica dos sítios aceptores de *trans-splicing*, descobrimos muitos pares de genes vizinhos na mesma fita, inesperadamente próximos: dentre os 139 grupos gênicos identificados com distância intergênica de até 200 pb, nós identificamos 65 unidades dicistrônicas e uma tricistrônica (**Error! Reference source not found.**) expressas em cercária, incluindo 34 dos 46 policistrons identificados por Protasio e colaboradores (2012).

Um exemplo de um transcrito dicistrônico, resolvido por SL *trans-splicing*, é mostrado na Figura 20 abaixo:

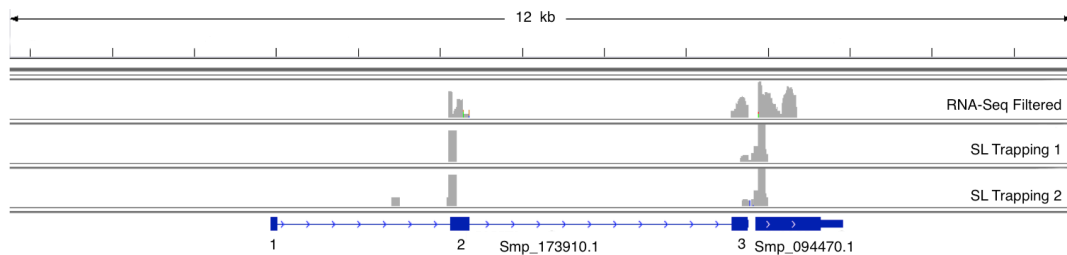


FIGURA 20 - TRANSCRITO POLICISTRÔNICO EM *S. MANSONI*. Esquema produzido no IGV da estrutura exon-ítron de um transcrito dicistrônico contendo os genes Smp_173910 e Smp_094470. As linhas e as caixas em azul indicam os modelos gênicos e as setas o sentido da transcrição - versão 5 do genoma de *S. mansoni*. As contagens de *reads* dos conjuntos de dados estão superpostas: RNA-Seq Filtered (faixa de cima), SL Trapping 1 (faixa do meio), e SL Trapping 2 (faixa de baixo).

O tamanho das distâncias intergênicas em transcritos policistrônicos foi analisada e foi encontrada média igual a 99 pb e mediana de 92 pb.

TABELA 9 - TRANSCRITOS POLICISTRÔNICOS EM S. MANSONI.

| Gene ID | Nome do Cromossoma/ <i>scaffold</i> | Gene Início (bp) | Gene Final (bp) | Fita | Contagens SL Trapping | Contagens RNA-Seq Filtered |
|------------|-------------------------------------|------------------|-----------------|------|-----------------------|----------------------------|
| Smp_035380 | 1 | 14187931 | 14193281 | -1 | 54 | 0 |
| Smp_035370 | 1 | 14193421 | 14204012 | -1 | 71 | 0 |
| Smp_165930 | 1 | 18236352 | 18262772 | -1 | 37 | 0 |
| Smp_165920 | 1 | 18262788 | 18290955 | -1 | 14397 | 57 |
| Smp_196960 | 1 | 25888899 | 25908258 | -1 | 1979 | 2 |
| Smp_062830 | 1 | 25908264 | 25911401 | -1 | 3073 | 0 |
| Smp_025950 | 1 | 28521922 | 28522896 | -1 | 44130 | 109 |
| Smp_136210 | 1 | 28523083 | 28542376 | -1 | 707 | 2 |
| Smp_147930 | 1 | 48611905 | 48631493 | -1 | 52976 | 117 |
| Smp_147940 | 1 | 48631555 | 48634343 | -1 | 0 | 0 |
| Smp_034530 | 1 | 20078234 | 20080251 | 1 | 5375 | 9 |
| Smp_140830 | 1 | 20080385 | 20093784 | 1 | 3951 | 9 |
| Smp_071410 | 1 | 23188766 | 23190268 | 1 | 0 | 0 |
| Smp_160940 | 1 | 23190281 | 23214263 | 1 | 42 | 0 |
| Smp_073100 | 1 | 25307262 | 25324241 | 1 | 60 | 0 |
| Smp_073110 | 1 | 25324272 | 25361527 | 1 | 46 | 0 |

| | | | | | | |
|------------|---|----------|----------|----|-------|-----|
| Smp_025680 | 1 | 29106744 | 29107406 | 1 | 13 | 0 |
| Smp_025670 | 1 | 29107468 | 29111085 | 1 | 9231 | 17 |
| Smp_057140 | 1 | 31874979 | 31888523 | 1 | 0 | 0 |
| Smp_210460 | 1 | 31888675 | 31905711 | 1 | 64648 | 205 |
| Smp_004980 | 1 | 32465706 | 32466400 | 1 | 0 | 0 |
| Smp_124830 | 1 | 32466532 | 32482444 | 1 | 15349 | 46 |
| Smp_114590 | 1 | 40397522 | 40397935 | 1 | 0 | 0 |
| Smp_126670 | 1 | 40397973 | 40403918 | 1 | 18 | 0 |
| Smp_054740 | 1 | 42616219 | 42617239 | 1 | 0 | 0 |
| Smp_054750 | 1 | 42617264 | 42618041 | 1 | 12 | 0 |
| Smp_021280 | 1 | 46291396 | 46292742 | 1 | 1311 | 24 |
| Smp_133700 | 1 | 46292801 | 46333910 | 1 | 3675 | 23 |
| Smp_147850 | 1 | 48191969 | 48204847 | 1 | 13 | 0 |
| Smp_048080 | 1 | 48204848 | 48234070 | 1 | 105 | 14 |
| Smp_049400 | 1 | 49538630 | 49539037 | 1 | 0 | 0 |
| Smp_148620 | 1 | 49539229 | 49601334 | 1 | 249 | 15 |
| Smp_023170 | 2 | 17043396 | 17046962 | -1 | 26047 | 56 |
| Smp_023160 | 2 | 17047103 | 17047518 | -1 | 0 | 0 |

| | | | | | | |
|------------|---|----------|----------|----|-------|----|
| Smp_181000 | 2 | 32630474 | 32641351 | -1 | 27890 | 33 |
| Smp_180990 | 2 | 32641402 | 32642612 | -1 | 127 | 0 |
| Smp_047570 | 2 | 8620772 | 8621493 | 1 | 0 | 0 |
| Smp_211090 | 2 | 8621593 | 8657951 | 1 | 750 | 8 |
| Smp_176820 | 2 | 10086396 | 10106309 | 1 | 0 | 0 |
| Smp_176830 | 2 | 10106420 | 10107211 | 1 | 62 | 1 |
| Smp_048870 | 2 | 13653727 | 13654119 | 1 | 0 | 0 |
| Smp_048880 | 2 | 13654224 | 13660782 | 1 | 30224 | 52 |
| Smp_030020 | 3 | 19427918 | 19432296 | -1 | 1633 | 0 |
| Smp_030030 | 3 | 19432396 | 19433051 | -1 | 0 | 0 |
| Smp_015070 | 3 | 27099424 | 27104411 | -1 | 0 | 1 |
| Smp_117810 | 3 | 27104540 | 27104980 | -1 | 0 | 0 |
| Smp_129400 | 3 | 8542187 | 8588954 | 1 | 1641 | 0 |
| Smp_129390 | 3 | 8589128 | 8621563 | 1 | 203 | 0 |
| Smp_195160 | 3 | 14603068 | 14605791 | 1 | 0 | 0 |
| Smp_076650 | 3 | 14605977 | 14608129 | 1 | 336 | 7 |
| Smp_079750 | 3 | 17159107 | 17165294 | 1 | 12 | 0 |
| Smp_079760 | 3 | 17165365 | 17195326 | 1 | 16634 | 70 |

| | | | | | | |
|------------|---|----------|----------|----|-------|----|
| Smp_084900 | 3 | 18259801 | 18263095 | 1 | 0 | 0 |
| Smp_084890 | 3 | 18263172 | 18285856 | 1 | 15568 | 55 |
| Smp_158860 | 4 | 1090333 | 1098520 | -1 | 81 | 0 |
| Smp_158850 | 4 | 1098678 | 1099490 | -1 | 13631 | 8 |
| Smp_038640 | 4 | 5355206 | 5376392 | -1 | 10197 | 18 |
| Smp_038630 | 4 | 5376531 | 5377308 | -1 | 1429 | 4 |
| Smp_090820 | 4 | 16513428 | 16531471 | -1 | 3963 | 21 |
| Smp_090830 | 4 | 16531654 | 16532553 | -1 | 151 | 0 |
| Smp_080150 | 4 | 20844962 | 20848086 | -1 | 29956 | 70 |
| Smp_165320 | 4 | 20848206 | 20849564 | -1 | 194 | 22 |
| Smp_146660 | 4 | 4661545 | 4665148 | 1 | 515 | 3 |
| Smp_211050 | 4 | 4665320 | 4685228 | 1 | 11543 | 17 |
| Smp_182770 | 4 | 5005996 | 5009886 | 1 | 0 | 0 |
| Smp_038730 | 4 | 5009937 | 5018397 | 1 | 966 | 1 |
| Smp_143150 | 4 | 5489737 | 5491450 | 1 | 75 | 0 |
| Smp_143140 | 4 | 5491465 | 5492480 | 1 | 0 | 1 |
| Smp_058970 | 4 | 14980460 | 14980774 | 1 | 124 | 0 |
| Smp_153860 | 4 | 14980940 | 14989331 | 1 | 22010 | 19 |

| | | | | | | |
|--------------|------------------------|----------|----------|----|-------|-----|
| Smp_090790 | 4 | 16484716 | 16485516 | 1 | 69 | 0 |
| Smp_090800 | 4 | 16485568 | 16508629 | 1 | 25312 | 103 |
| Smp_035730 | 4 | 29430290 | 29432952 | 1 | 0 | 0 |
| Smp_141410 | 4 | 29432985 | 29502926 | 1 | 20180 | 71 |
| Smp_087810 | 6 | 4984893 | 5003037 | -1 | 25 | 0 |
| Smp_087820 | 6 | 5003221 | 5003631 | -1 | 0 | 0 |
| Smp_149350 | 6 | 13918650 | 13922461 | -1 | 8599 | 38 |
| Smp_149360 | 6 | 13922543 | 13935119 | -1 | 69 | 0 |
| sma.28s-25.1 | Chr_1.unplaced.SC_0076 | 1210131 | 1211825 | -1 | 2316 | 1 |
| sma.58s-24.1 | Chr_1.unplaced.SC_0076 | 1211937 | 1212090 | -1 | 0 | 0 |
| Smp_210170 | Chr_2.unplaced.SC_0108 | 176500 | 182759 | -1 | 36241 | 54 |
| Smp_210160 | Chr_2.unplaced.SC_0108 | 182903 | 183328 | -1 | 53942 | 196 |
| Smp_033590 | Chr_2.unplaced.SC_0108 | 183420 | 184862 | -1 | 108 | 0 |
| Smp_210190 | Chr_2.unplaced.SC_0120 | 994194 | 994415 | -1 | 13 | 0 |
| Smp_210180 | Chr_2.unplaced.SC_0120 | 994520 | 994993 | -1 | 0 | 0 |
| Smp_006980 | Chr_3.unplaced.SC_0192 | 119052 | 119399 | 1 | 0 | 0 |
| Smp_006970 | Chr_3.unplaced.SC_0192 | 119468 | 154747 | 1 | 29854 | 90 |
| Smp_058720 | SC_0013 | 562455 | 564360 | -1 | 95 | 0 |

| | | | | | | |
|------------|---------|---------|---------|----|-------|-----|
| Smp_153730 | SC_0013 | 564379 | 565639 | -1 | 0 | 0 |
| Smp_167410 | SC_0037 | 832808 | 848588 | -1 | 597 | 0 |
| Smp_167420 | SC_0037 | 848738 | 849016 | -1 | 0 | 0 |
| Smp_170170 | SC_0049 | 241782 | 269078 | -1 | 3656 | 0 |
| Smp_170180 | SC_0049 | 269235 | 269312 | -1 | 0 | 0 |
| Smp_103330 | SC_0134 | 10170 | 10785 | 1 | 873 | 0 |
| Smp_103320 | SC_0134 | 10836 | 17720 | 1 | 41520 | 127 |
| Smp_124980 | SC_0151 | 181315 | 187287 | -1 | 173 | 1 |
| Smp_005390 | SC_0151 | 187472 | 211416 | -1 | 8194 | 23 |
| Smp_041680 | SC_0180 | 369836 | 370274 | 1 | 206 | 0 |
| Smp_144870 | SC_0180 | 370376 | 383034 | 1 | 9100 | 36 |
| Smp_176350 | SC_0186 | 219108 | 241634 | -1 | 11 | 0 |
| Smp_098610 | SC_0186 | 241693 | 241921 | -1 | 0 | 0 |
| Smp_174010 | SC_0221 | 122246 | 122636 | 1 | 53 | 0 |
| Smp_174000 | SC_0221 | 122786 | 128361 | 1 | 49 | 0 |
| Smp_142990 | W | 2480399 | 2510166 | -1 | 26379 | 40 |
| Smp_143000 | W | 2510220 | 2511677 | -1 | 33661 | 85 |
| Smp_038420 | W | 2586348 | 2599957 | -1 | 6941 | 14 |

| | | | | | | |
|------------|---|----------|----------|----|---------|------|
| Smp_038430 | W | 2600104 | 2601891 | -1 | 0 | 0 |
| Smp_080550 | W | 17151880 | 17161368 | -1 | 496 | 1 |
| Smp_080560 | W | 17161564 | 17176237 | -1 | 0 | 0 |
| Smp_012050 | W | 18787660 | 18787965 | -1 | 27890 | 72 |
| Smp_128610 | W | 18788117 | 18788748 | -1 | 0 | 1 |
| Smp_059880 | W | 24216002 | 24227750 | -1 | 17249 | 69 |
| Smp_059870 | W | 24227886 | 24234769 | -1 | 284 | 2 |
| Smp_094470 | W | 27274100 | 27275167 | -1 | 69851 | 119 |
| Smp_173910 | W | 27275255 | 27281082 | -1 | 5092 | 34 |
| Smp_024110 | W | 39428503 | 39434676 | -1 | 1204777 | 2159 |
| Smp_024120 | W | 39434794 | 39435186 | -1 | 1691 | 1 |
| Smp_180360 | W | 45208849 | 45214329 | -1 | 11207 | 15 |
| Smp_180370 | W | 45214425 | 45215987 | -1 | 0 | 0 |
| Smp_130610 | W | 47852787 | 47856198 | -1 | 6167 | 22 |
| Smp_015440 | W | 47856235 | 47858949 | -1 | 323 | 0 |
| Smp_160270 | W | 56711445 | 56725483 | -1 | 11811 | 31 |
| Smp_160260 | W | 56725546 | 56727998 | -1 | 0 | 0 |
| Smp_033510 | W | 13438993 | 13439214 | 1 | 0 | 0 |

| | | | | | | |
|------------|---|----------|----------|---|-------|----|
| Smp_140260 | W | 13439223 | 13476729 | 1 | 31 | 0 |
| Smp_165540 | W | 17054298 | 17063040 | 1 | 224 | 1 |
| Smp_080520 | W | 17063162 | 17079509 | 1 | 16537 | 47 |
| Smp_088390 | W | 22140971 | 22141853 | 1 | 1612 | 6 |
| Smp_088380 | W | 22141940 | 22142540 | 1 | 28370 | 35 |
| Smp_151690 | W | 29778749 | 29779204 | 1 | 166 | 0 |
| Smp_210900 | W | 29779260 | 29791257 | 1 | 54074 | 5 |
| Smp_163740 | W | 58948422 | 59051479 | 1 | 25 | 0 |
| Smp_163730 | W | 59051484 | 59133867 | 1 | 909 | 1 |

Os policistrons putativos identificados estão localizados principalmente no chr1 (25,75% de todos os policistrons), chrW (22,72%) e chr4 (15,15%), em contraste com os demais genes que são processados por SLTS os quais estão distribuídos ao longo dos cromossomos seguindo a mesma distribuição dos genes expressos em cercária, uma vez que em *S. mansoni* o número de genes encontrados em cada cromossomo é proporcional ao tamanho de cada cromossomo. Os 133 genes constituintes das unidades policistrônicas em conjunto constituem apenas ~2% de todos os 5.821 genes processados por SLTS identificados em cercária, entretanto, nossas observações subestimam o verdadeiro número de unidades policistrônicas, uma vez que anotações incompletas de transcritos, policistrons com distância intergênica maiores que 200 pb, ou não expressos no nosso conjunto de dados podem estar subestimando o número de policistrons.

A maior parte dos genes componentes dos policistrons é conservada e 23,33% são transcritos monoexônicos, 131 são genes codificadores de proteínas e dois são rRNAs.

4.10.4. ANÁLISE FUNCIONAL DOS TRANSCRITOS PROCESSADOS POR SLTS

O pequeno número de genes previamente descritos como processados por SLTS (DAVIS; HARDWICK; TAVERNIER, 1995; PROTASIO et al., 2012) impossibilitaram pesquisadores de identificar funções gênicas relacionadas a este mecanismo em *S. mansoni*. Com a grande profundidade da biblioteca gerada por nós (SL Trapping), foi possível investigar funções biológicas comuns entre os produtos gênicos derivados de transcritos processados por SLTS.

Para uma melhor compreensão do processo, as sequências codificadoras CDS de *S. mansoni* foram tabuladas em planilha Excel com hyperlinks para vários resultados de BLAST e valores de contagens. Das 4.793 sequências anotadas anteriormente como “Hypothetical protein” na versão 5 do genoma, cerca de 42% foram re-anotadas nesse trabalho. Os resultados de BLAST foram utilizados para atribuir uma classe a cada gene conhecido do parasito, incluindo aqueles que são processados por SLTS.

Para atribuir funções biológicas aos genes processados por SLTS foi feita uma análise do enriquecimento de categorias gênicas (KARIM; SINGH; RIBEIRO, 2011). Comparou-se a frequência das classes gênicas encontradas nos 5.057 genes codificadores de proteínas processados por SLTS, nos 131 genes que fazem parte de transcritos policistrônicos, assim como nos 7.987 genes codificadores de proteínas expressos na fase cercária. Em seguida, um teste χ^2 foi realizado para averiguar se a diferença de frequência entre o grupo *trans-splicing* e o transcriptoma total expresso na fase cercária, para cada classe, é estatisticamente significativa.

Quando comparamos a distribuição funcional dos genes que sofrem processamento por SLTS com o transcriptoma expresso na fase cercária, nós não encontramos muitas classes funcionais enriquecidas estatisticamente significantes (p-value < 0.05, teste χ^2 ; Figura 21). As duas únicas exceções foram as classes “Desconhecido” e “Fator de transcrição” (p-value < 0,024 e p-value < 0,0003 respectivamente, teste χ^2 ; Figura 21). As mesmas classes não foram encontradas enriquecidas quando os transcritos da biblioteca RNA-Seq Filtered foram analisados. Utilizando a metodologia de análise de enriquecimento de ontologias gênicas (GO enrichment) em transcritos processados por SLTS em *C.*

intestinalis, Matsumoto e colaboradores (2010) encontraram classes associadas com membrana plasmática / sistema de endomembranas, homeostase de Ca²⁺ e componentes do citoesqueleto/actina. Já em *C. elegans*, genes envolvidos no desenvolvimento e regulação biológica estão super-representados no grupo que sofre processamento por SLTS (ALLEN *et al.*, 2011), mas essas análises foram feitas comparando-se os genes expressos que sofrem SLTS com todos os genes do parasito como *background*.

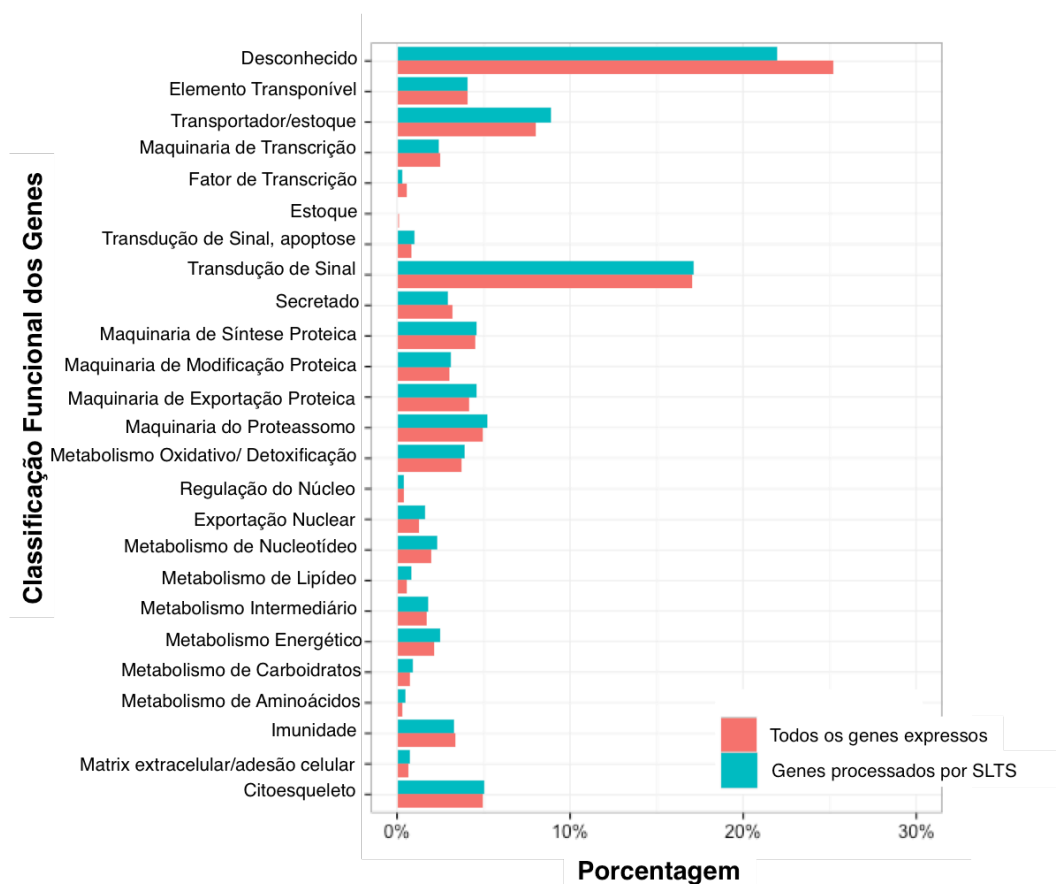


FIGURA 21 - CLASSIFICAÇÃO FUNCIONAL DOS TRANSCRITOS DE *S. MANSONI*. Em azul a frequência dos transcritos que sofrem processamento por SLTS e em vermelho a frequência de todos os transcritos expressos na fase cercária.

Adicionalmente, as proteínas codificadas por transcritos processados por SLTS foram associadas às vias metabólicas do parasito *S. mansoni* disponíveis no banco de dados KEGG (Kyoto Encyclopedia of Genes and Genomes) (KANEHISA; GOTO, 2000; YAMADA *et al.*, 2011), sendo também associadas à frequência de

processamento por SLTS (representado pela largura das linhas vermelhas na Figura 22). Foi também visto que genes que sofrem SLTS não apresentam enriquecimento significativo nas vias presentes no KEGG, demonstrando que as proteínas derivadas de transcritos processados por SLTS podem atuar em diversas vias metabólicas centrais de *S. mansoni*.

Esta observação está de acordo com o esperado, uma vez que o mecanismo de SLTS em *S. mansoni* não parece estar associado com nenhum tecido específico, fase do desenvolvimento ou gênero, tão pouco com genes específicos ou famílias gênicas (DAVIS; HARDWICK; TAVERNIER, 1995; DAVIS, 1996; MOURÃO et al., 2013). Nossos resultados também estão em concordância com um estudo recente realizado em *Eutreptiella sp.*, no qual foi reportado que os genes que sofrem SLTS são funcionalmente muito diversos, caracterizando portanto este como um mecanismo ubíquo na alga euglenóide (KUO et al., 2013). Entretanto, os genes processados por SLTS componentes de policistrons são enriquecidos na categoria “Maquinaria de exportação proteica” (p-value < 0,002, teste χ^2).

4.10.1. CARACTERÍSTICAS DOS GENES PROCESSADOS POR SLTS

Nós estimamos os níveis de expressão de todos os genes expressos na fase cercária em *S. mansoni* a partir de um experimento de RNA-Seq (obtido no banco de dados do SRA) e investigamos a correlação entre a expressão gênica e a eficiência do mecanismo de SLTS. Em geral, nós observamos que genes processados por SLTS são em média mais expressos do que os genes não processados (p-value < 1.304e-08, em um teste t para duas amostras - Welch), uma vez que a média dos valores de expressão normalizados dos genes que

sofrem SLTS é ~3 vezes maior comparada com aqueles que não são processados (760.9108 versus 2480.4370, Figura 23A)

Entretanto, na Figura 23B é possível observar uma fraca correlação entre as frequências de SLTS e os níveis expressão gênica ($fit= 0,204$ coeficiente de correlação de Pearson, $rank= 0,208$ coeficiente de correlação de Spearman) e que existem genes que são altamente expressos e pouco processados por SLTS e vice-versa (quadrante abaixo à direita e quadrante acima à esquerda, respectivamente na , Figura 23B). Assim como já bem descrito na literatura (BENGTSSON *et al.*, 2005; ISLAM *et al.*, 2011), nossa análise não permite rejeitar a hipótese de que o logaritmo dos níveis de expressão gênica seguem uma distribuição diferente da normal (p-valor 0,62 , teste Kolmogorov-Smirnov, histograma em rosa na Figura 23C). Entretanto, as frequências do SLTS (histograma em azul na Figura 23C) desviam do comportamento normal (p-valor $< 0,11$, teste Kolmogorov-Smirnov,), acumulando em genes com valores altos e baixos de contagem, confirmando nossas observações prévias da existência de alguns genes em particular com frequências de SLTS muita alta (ou seja, com log frequências > 2) ou muito baixa (ou seja, com o registro de frequências em torno de 0). Apesar da sensibilidade para detectar eventos de *trans-splicing* aumentar com o aumento da expressão gênica, nossos resultados mostram que a eficiência do processamento por SLTS não é somente determinada pela taxa de transcrição de um gene, sugerindo a existência de mecanismos particulares que facilitam ou impedem o SLTS.

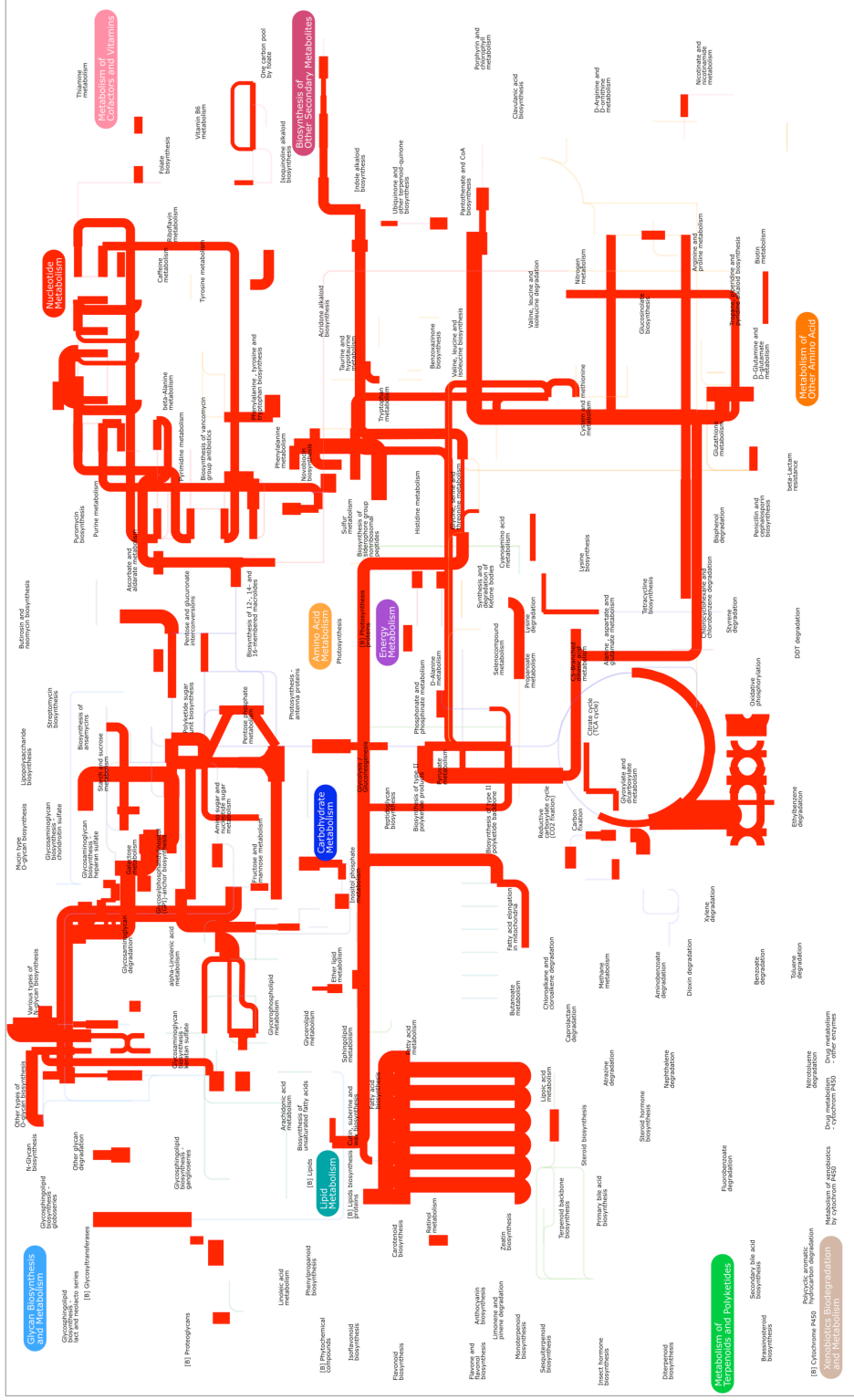


FIGURA 22 - VIAS METABÓLICAS CUJAS PROTEÍNAS ENVOLVIDAS SÃO DERIVADAS DE TRANSCRITOS PROCESSADAS POR SL. TRANS-SPLICING. A figura foi feita utilizando-se o programa iPath. Os valores de frequência de SLTs foram associados para cada KO representado. As linhas em vermelho representam vias metabólicas cujas proteínas participantes são derivadas de transcritos que sofrem *trans-splicing* frequentemente e a largura da linha está relacionada à frequência com que o transcrito é processado por SLTs.

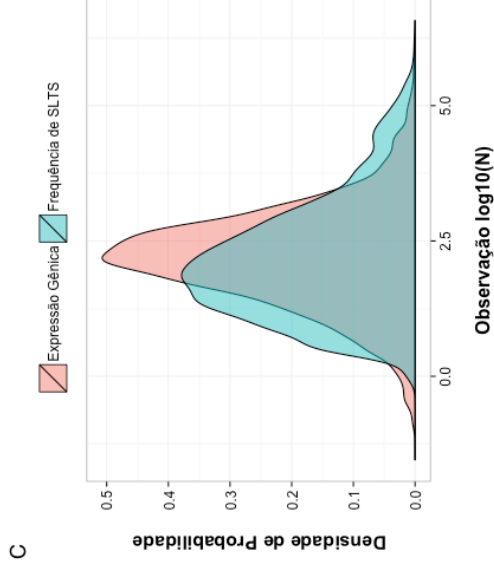
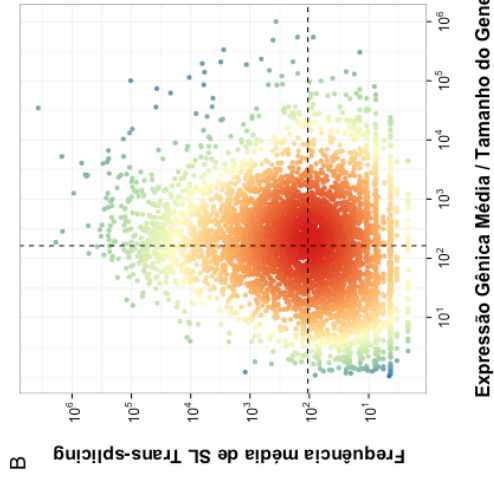
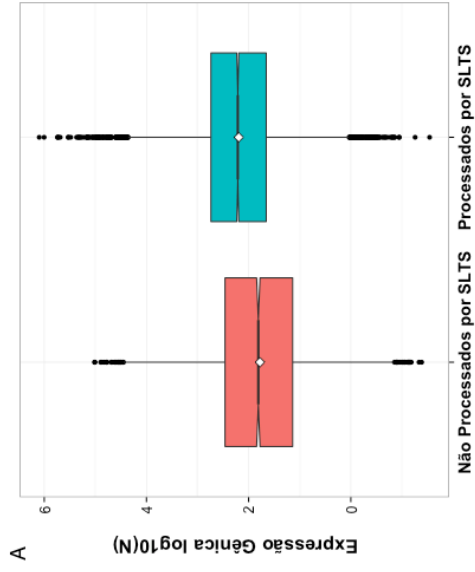


FIGURA 23 – EXPRESSÃO GÊNICA E FREQUÊNCIA DE SLTS. **A** – Logaritmo dos níveis de expressão normalizados de 3.928 genes que não sofrem processamento por SLTS e de 7.165 genes que sofrem processamento por SLTS (p-value < 2.2×10^{-16} , em um teste Kolmogorov-Smirnov para duas amostras). **B** – Correlação entre o Log dos níveis de expressão normalizados (eixo x, estimado a partir das reads de RNA-Seq) e Log da frequência normalizada de SLTS (eixo y, estimado a partir do conjunto de dados SL Trapping) dos genes processados por SLTS (fit= 0,204 coeficiente de correlação de Pearson, rank= 0,208 coeficiente de correlação de Spearman). Linhas pontilhadas em cinza marcam as medianas de ambos os indicadores (163 para expressão gênica e 106 para frequência de SLTS), separando genes processados com baixa (quadrante inferior esquerdo) e alta (quadrante superior esquerdo) frequência de SLTS relativo aos seus níveis de expressão. **C** – Distribuição dos níveis de expressão gênica (curva em rosa) e frequência de SLTS (curva em azul).

Em seguida, buscamos avaliar a conexão entre expressão gênica e eficiência de SLTS ao investigar as características que discriminam genes que tem baixa expressão e elevado número de eventos de SLTS (o *trans-splicing* dirige os eventos de SLTS – TSD) de genes que tem alta expressão com um número baixo de eventos de SLTS (a expressão gênica dirige os eventos de SLTS – GED). Nós usamos 2.400 *outliers* (1.200 *outliers* pertencentes ao quadrante inferior direito e 1.200 *outliers* pertencentes ao quadrante superior esquerdo na Figura 23B), que correspondem a 33,5% do total de genes analisados processados por SLTS. Como esperado, não existem preferências significativas para uma certa localização cromossômica, nem por genes GED nem por genes TSD quando estes são comparados com a distribuição cromossômica dos genes expressos em cercária com nível de significância de 0,01. (Figura 24A).

Ao avaliar o número de transcritos produzidos por *splicing* alternativo nos genes GED e TSD, vimos que o número de isoformas diferentes produzidas por *cis-splicing* alternativo decresce exponencialmente para os genes GED (barras azuis na figura Figura 24B), assim como ocorre para o transcriptoma de *S. mansoni* (barras vermelhas na Figura 24B). O conjunto GED apresenta mais genes com maior número de transcritos alternativos por gene, apesar desta diferença não ter significância estatística (p-value= 0,96, barras azuis). Entretanto, para os genes TSD, observamos um número de *cis-splicing* alternativo significativamente menor (barras rosas na Figura 24B, p-value= 0,08, teste Kolmogorov-Smirnov para duas amostras), sendo que a maior parte deles apresenta não mais que 3 transcritos alternativos por gene.

De forma similar, avaliamos o número de exons existentes nos genes pertencentes aos grupos GED e TSD, e vimos que o número de exons nos genes

de *S. mansoni* diminui exponencialmente (barras vermelhas na Figura 24C). Tanto o conjunto de genes GED quanto o de genes TSD contém menos elementos monoexônicos do que o observado no transcriptoma total do parasito, e um número significativamente maior de genes com número de exons variando entre valores moderados a altos (p-value= 0,0005 para ambos os grupos de genes GED e TSD, teste Kolmogorov-Smirnov para duas amostras).

Nossos resultados mostram que ambos os conjuntos de genes GED e TSD apresentam maior grau de complexidade de arquitetura gênica, considerando o número de exons compreendidos em um locus. Entretanto, enquanto no conjunto de genes GED a diversidade de exons também traduz-se em um aumento do número de *splicing* alternativos, nos genes TSD existe um acúmulo preferencial de genes com mais exons que exibem um número restrito de transcritos alternativos, o que pode sugerir uma maior eficiência do processamento por SLTS na ausência de estruturas complexas de *cis-splicing*.

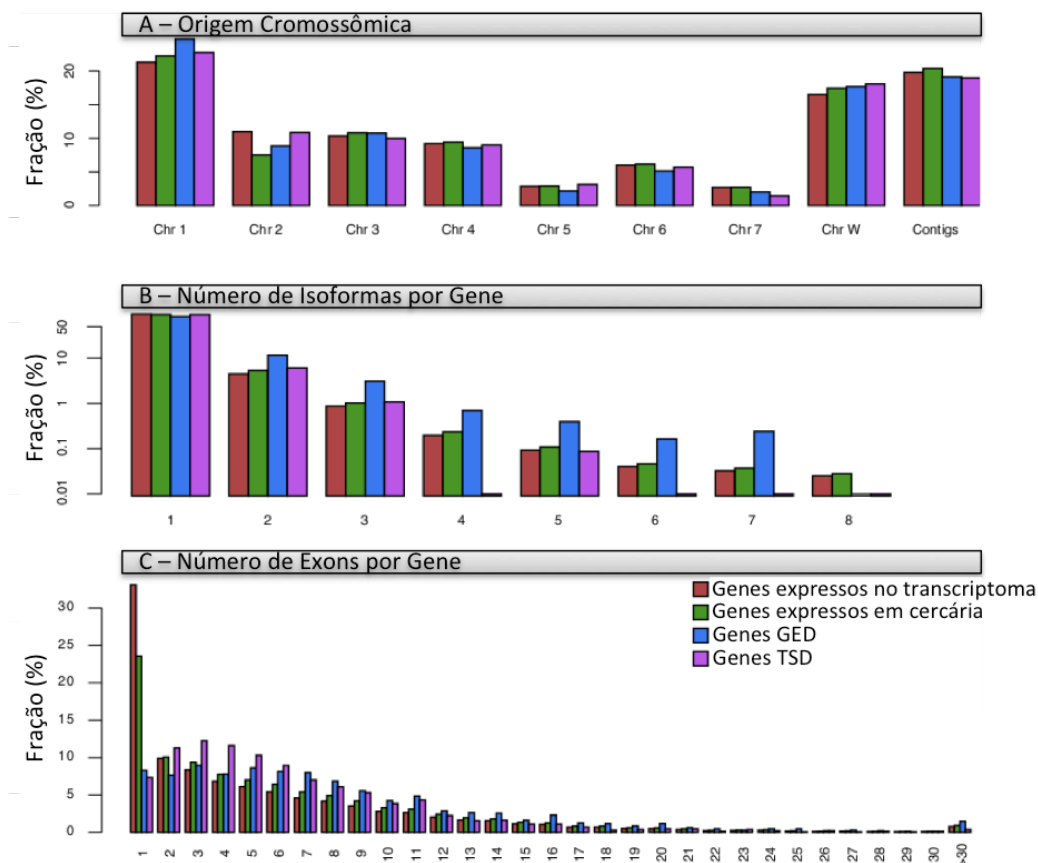


FIGURA 24 – ORIGEM CROMOSSÔMICA, ISOFORMAS ORIGINADAS DE *SPLICING* ALTERNATIVO E NÚMERO DE EXONS DE TRANSCRITOS PROCESSADOS POR SLTS. **A** – Frequência relativa dos genes localizados em cada cromossomo é comparada entre o conjunto de genes do transcriptoma inteiro de *S. mansoni* (barras em vermelho), genes expressos em cercária (barras verdes), assim como genes que tem alta expressão com uma baixa frequência de eventos de SLTS- GED (barras azuis) e genes que tem baixa expressão e uma alta frequência de eventos de SLTS – TSD (barras roxas). **B** – Distribuição de transcritos alternativos anotados por gene no transcriptoma total de *S. mansoni* (barras vermelhas), no conjunto de genes expressos em cercária (barras verdes), no conjunto de genes GED (barras azuis) e no conjunto de genes TSD (barras rosas). **C** – Distribuição do número de exons por genes no transcriptoma total de *S. mansoni* (barras vermelhas), nos genes expressos em cercária (barras verdes), nos genes GED (barras azuis) e nos genes TSD (barras rosas).

4.10.2. COMPETIÇÃO ENTRE SLTS E *CIS-SPLICING*

Em contraste com a visão clássica de que o primeiro exon é o substrato exclusivo para a inserção do SL (CONRAD et al., 1991), nós observamos muitos eventos de processamento por SLTS ocorrendo nos demais exons, tanto no conjunto de dados representado pela biblioteca SL Trapping, quanto pela biblioteca RNA-Seq Filtered (Figura 25).

Em concordância com resultados prévios, uma análise quantitativa confirmou que os eventos de SLTS ocorrem de forma significativamente mais

eficiente no primeiro exon, com uma média de contagens de eventos de SLTS igual a 2.295 no primeiro exon *versus* uma média de contagens igual a 1.035 nos demais exons que não o primeiro. Entretanto, apesar de a entrada da sequência do SL ocorrer principalmente no primeiro exon, nós observamos que este evento representa apenas 18% de todos os eventos, sugerindo que uma grande variedade de eventos de inserção da sequência do SL ocorrem em sítios distantes da porção 5' do transcrito (Figura 25).

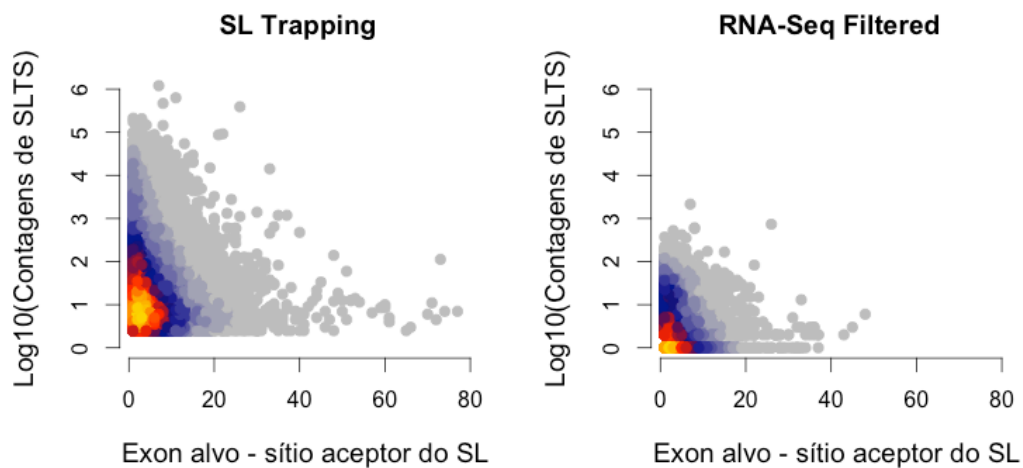


FIGURA 25 - SÍTIOS DE SLTS ALTERNATIVO. Frequência de SLTS (eixo y) como função da posição no transcrito, onde os exons são ordenados no sentido 5'->3' (eixo x). A frequência foi calculada utilizando-se as contagens obtidas na biblioteca SL Trapping (painel a esquerda) e RNA-Seq Filtered (painel a direita).

Interessantemente, os eventos de SLTS que ocorrem internamente nos transcritos (em íntrons) exibem um maior grau de anotação do sítio de *splicing* comparados com os sítios de *trans-splicing* que ocorrem no outron na porção 5' do transcrito (63% *versus* 7%, Figura 26A), e de forma coerente, exibem também maiores sinais de sítios canônicos de *splicing* (Figura 26A). Sítios aceptores de *trans-splicing* em íntrons (barras vermelhas) mostram uma fração consideravelmente maior do dinucleotídeo canônico AG no sítio de *splicing* do que sítios aceptores de *trans-splicing* em outrons (barras azuis). A conservação

do dinucleotídeo no sítio acceptor reflete o grau de anotação dos sítios mapeados como alvos de SLTS. Trabalhos anteriores realizados em *T. brucei* (NILSSON *et al.*, 2010) mostraram que cerca de 20% dos sítios aceptores de *trans-splicing* alternativo contém dinucleotídeos diferentes de AG, com o motivo GG ocorrendo em 7% dos casos enquanto TG, AA, GA e AC foram encontrados em 2% dos sítios.

Nossas observações concordam com trabalhos prévios (CONRAD *et al.*, 1991) que mostram o SLTS como um mecanismo mais eficiente, e aparentemente mais facilitado, em sítios aceptores localizados próximos ao início do transcrito, provavelmente devido à ausência de competição com outros sítios doadores de *splice* em *cis*.

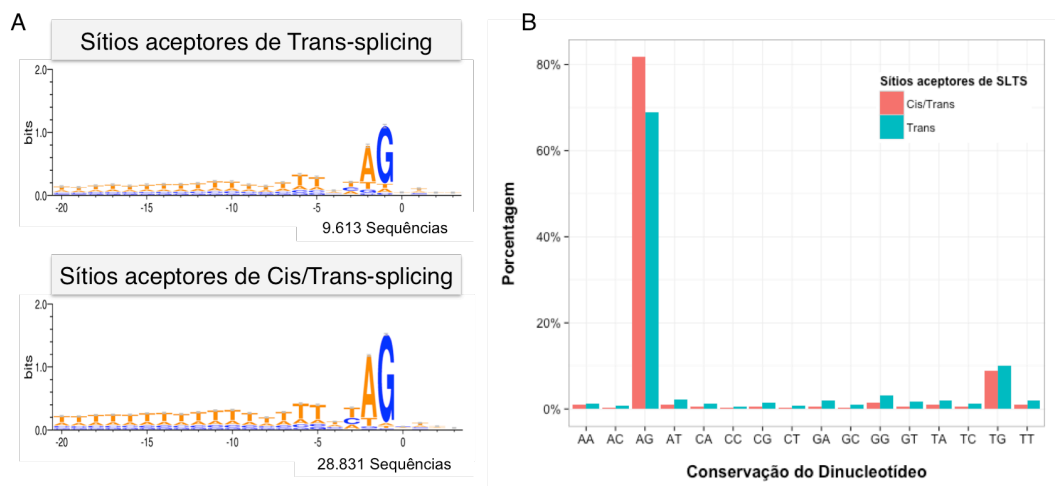


FIGURA 26 - COMPARAÇÃO ENTRE OS SÍTIOS DE TRANS-SPLICING OCORRENDO EM OUTRONS E ÍNTRONS. **A** - Logo das sequências previstas e anotadas do sítio acceptor de *trans-splicing*. As posições resultantes do mapeamento da biblioteca SL Trapping foram recuperadas e mapeamentos ocorrendo no primeiro exon de um transcrito ou além do primeiro exon foram agrupados e respectivamente denominados como sítios aceptores de *trans-splicing* (painel superior, N=9.613 com 7% dos sítios anotados) e sítios aceptores de *cis/trans-splicing* (painel inferior, N=28.831 com 63% dos sítios anotados). Os 20 nt *upstream* e os 4 nt *downstream* aos sítios onde foram encontrados a inserção da sequência do SL foram analisados; **B** - Conservação do sítio acceptor. Sítios que sofrem *trans-splicing* em íntrons são representados por barras vermelhas e sítios aceptores de *trans-splicing* em outrons por barras azuis.

Ao agrupar os eventos de SLTS por posição de entrada da sequência do SL relativa ao exon de cada transcrito, observamos que existe uma preferência de inserção da sequência do SL nos exons das extremidades 5' e 3' do transcrito, independentemente do tamanho do transcrito (painel a esquerda na Figura

27A). Como esperado, não observamos uma cobertura aumentada de *reads* de RNA-Seq nestes exons da extremidade dos transcritos (painel a direita na Figura 27A), confirmando que as observações de frequência de SLTS não são impactadas por artefatos de expressão desigual dos exons. Foi também observado nos transcritos que sofrem SLTS que o tamanho dos exons terminais são geralmente maiores (Figura 27B), e os íntrons alvos de SLTS são geralmente maiores que os íntrons processados somente por *cis-splicing* (Figura 27C). Em conjunto, estas observações mostram a existência de características particulares que governam o mecanismo de SLTS decorrente de uma geometria especial entre exon-íntron nas extremidades dos transcritos.

Para investigar as propriedades que discriminam os íntrons que sofrem SLTS dos que não sofrem, nós empregamos a classificação proposta por Davis e colaboradores (1995) de *cis-spliced* íntrons, *cis-spliced* íntrons em transcritos processados por SLTS e *trans-spliced* íntrons.

Ao analisar a sequência de um subconjunto de regiões intrônicas, foi possível confirmar registros prévios sobre uma maior frequência de polipirimidinas em sítios aceptores de *cis-splicing* que podem atuar como substratos para *trans-splicing* (sequência mais à direita de cada triplete na Figura 28), ao comparar com a intensidade do sinal observado nos sítios aceptores exclusivos de *cis-splicing* (sequência mais à esquerda de cada triplete na Figura 28). O conteúdo de uridinas é ainda mais forte próximo ao sítio aceptor de *splicing* (AG).

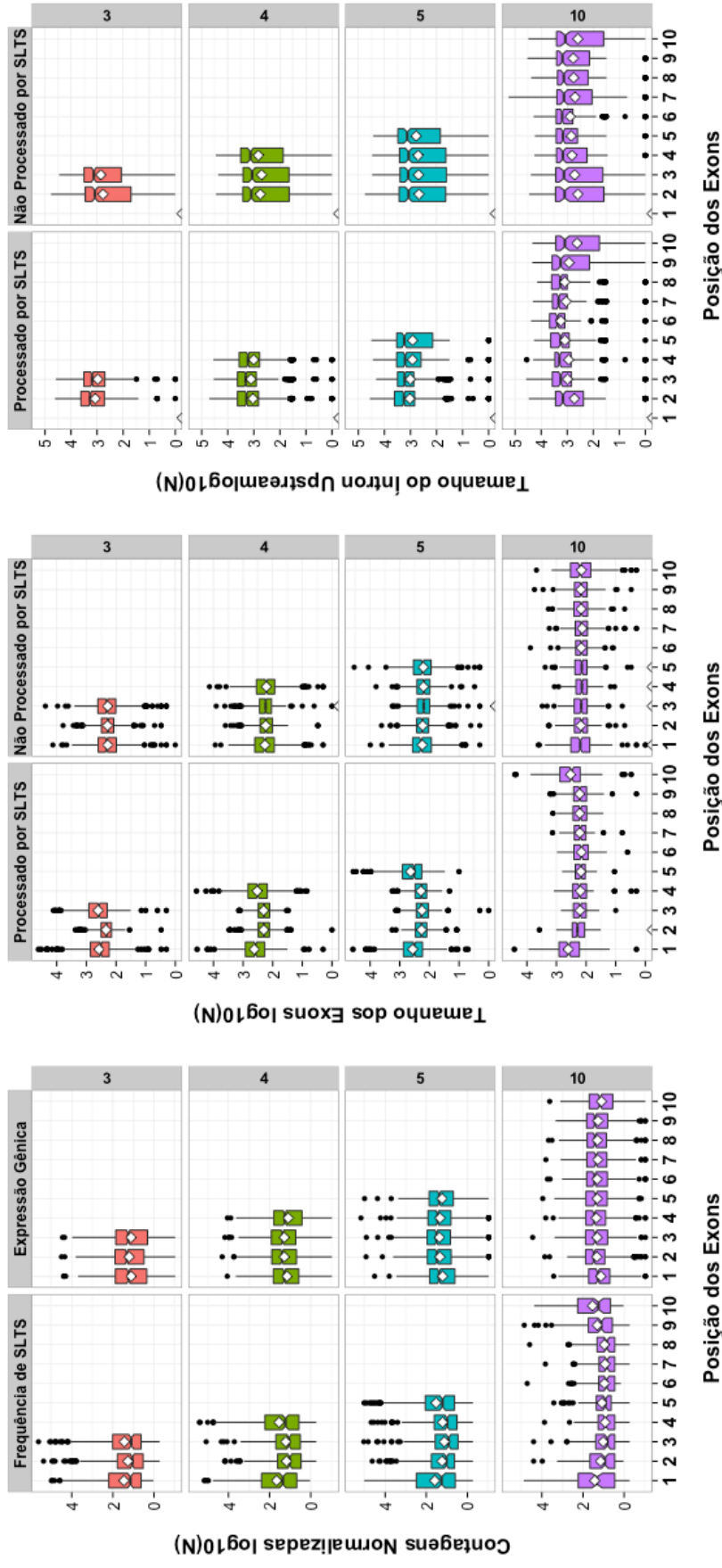


FIGURA 27 – POSIÇÃO DA INSERÇÃO DA SEQUÊNCIA DO SL NOS TRANSCRITOS. A - Distribuição das frequências de SLTS (painel à esquerda) e níveis de expressão (painel à direita) observado em cada exon de acordo com a sua posição no transcrito (eixo x) e agrupado por número de exons anotados em cada gene (escala no eixo y, de cima para baixo). B - Distribuição das frequências do tamanho dos exons estratificados pela posição do exon dentro de cada transcrito (eixo x) nos genes que sofrem SLTS (painel à esquerda) e nos genes onde não foram observados sinais de SLTS (painel à direita). C - Distribuição das frequências do tamanho dos introns prévios aos exons estratificados pela posição do exon dentro de cada transcrito (eixo x) nos genes que sofrem SLTS (painel à esquerda) e nos genes onde não foram observados sinais de SLTS (painel à direita).

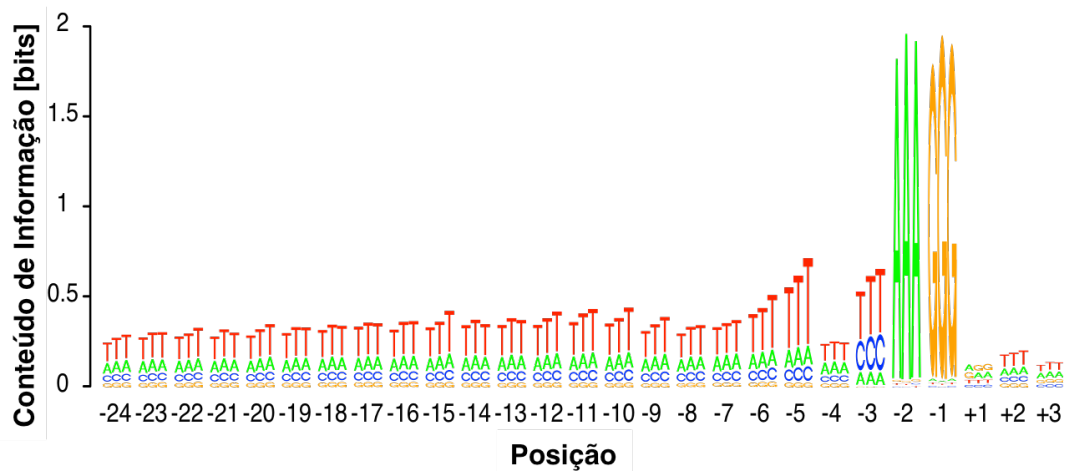


FIGURA 28 – MOTIVOS DOS SÍTIOS DE SPLICING EM ÍNTRONS. As sequências à esquerda de cada triplete representam íntrons que são exclusivamente processados por *cis-splicing*. As sequências do meio representam íntrons processados por *cis-splicing* em transcritos que são processados por SLTS. Já as sequências à direita representam íntrons que atuam como substrato para o mecanismo de SLTS.

O trato de polipirimidinas tem sido relacionado como um dos elementos mais importantes no *splicing* (COOLIDGE; SEELY; PATTON, 1997) e, em tripanosomatídeos, este tem sido relacionado à eficiência de *splicing* (HUANG; VAN DER PLOEG, 1991b). Em um estudo sobre o pre-mRNA policistrônico de tubulina (MATTHEWS; TSCHUDI; ULLU, 1994), a substituição de aminoácidos em sua sequência gerou um bloqueio do correto *splicing* da molécula, mostrando que o trato de polipirimidina no sítio acceptor 3' é crucial para o correto *trans-splicing*.

Ainda, para melhor entender a competição que ocorre entre *cis-splicing* e *trans-splicing*, fizemos uma distinção adicional entre os tipos de íntrons em: íntron único, íntrons iniciais, intermediários, e íntrons finais e avaliamos o tamanho desses íntrons, e dos exons que os flanqueiam. Como resultado, observamos que os íntrons que atuam como substrato para SLTS (caixas roxas na Figura 29A) são geralmente maiores que os íntrons que atuam como substrato para *cis-splicing* (caixas verdes na Figura 29A), especialmente em

íntrons de transcritos monointrônicos e íntrons internos. Adicionalmente, o comprimento dos íntrons decresce conforme percorremos o gene da posição 5' para 3'. Os íntrons que são substratos para *cis-splicing* em transcritos processados por SLTS se comportam da mesma forma, com a exceção dos monointrônicos, nos quais o SLTS ocorre na porção 5' do pre-mRNA, no outron. Esses resultados corroboram achados em *C. elegans* e a teoria de que a maquinaria de *splicing* pode interpretar íntrons longos como outrons, uma vez que a probabilidade de ocorrer SLTS dentro de um íntron aumenta com o seu tamanho (ALLEN *et al.*, 2011; CHOI; NEWMAN, 2006). No caso de transcritos monointrônicos, vimos também que o exon *upstream* ao íntron usado como sítio de SLTS tende a ser maior (caixas roxas na Figura 29B), assim como os exons *downstream* ao íntron usado como sítio de SLTS (caixas roxas na Figura 29C), principalmente se estes representam o último exon do transcrito.

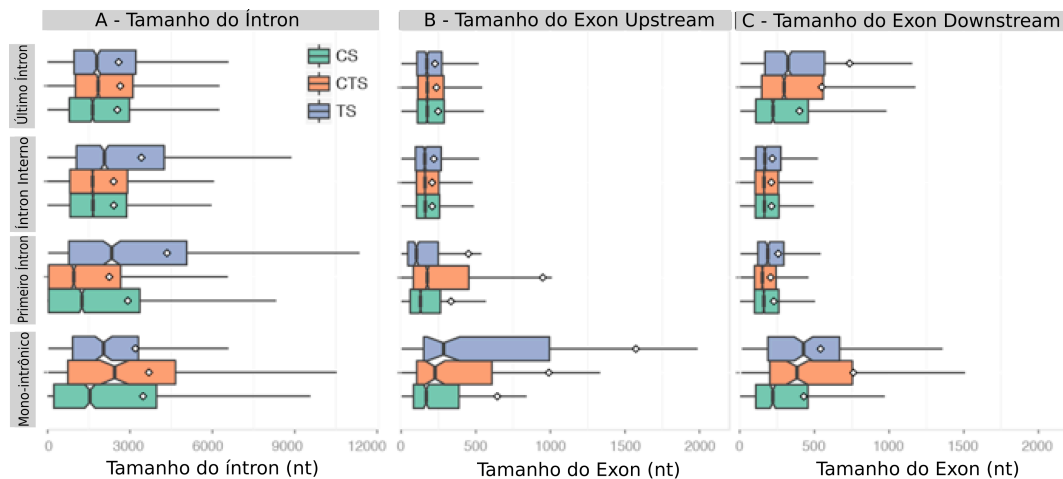


FIGURA 29 - TAMANHO DOS ÍNTRONS E EXONS DOS TRANSCRITOS DE ACORDO COM SUA CLASSIFICAÇÃO E SUA POSIÇÃO RELATIVA NO TRANSCRITO. O Tamanho dos íntrons (A), assim como dos exons *downstream* (B) e *upstream* (C) em relação ao tipo de íntron foram avaliados (ou seja, íntrons que são substratos para *cis-splicing* – caixas verdes, íntrons que são substratos para *cis-splicing* em transcritos processados por SLTS – caixas laranjas e íntrons que são substratos para SLTS – caixas roxas) de acordo com a posição do íntron no transcrito (ou seja, íntrons em transcritos monointrônicos, primeiro íntron, íntron interno, e último íntron).

Ao conjunto de sequências intrônicas previamente descritas aplicamos um algoritmo de predição de potenciais sítios aceptores e doadores de *splicing* e

vimos que a força dos sítios doadores de *splicing* aumenta ligeiramente de 5' para 3', e esta força é particularmente fraca no caso dos transcritos monointrônicos onde o íntron sofre *cis-splicing* (caixa verde na Figura 30A). Já para os sítios aceptores, verificamos que estes são geralmente mais fortes nos introns utilizados como substrato para SLTS do que nos introns utilizados no *cis-splicing*, e essa diferença diminui de 5' para 3'. Essa força é especialmente reduzida nos transcritos monointrônicos, onde o SLTS ocorre no outtron (caixa verde na Figura 30B).

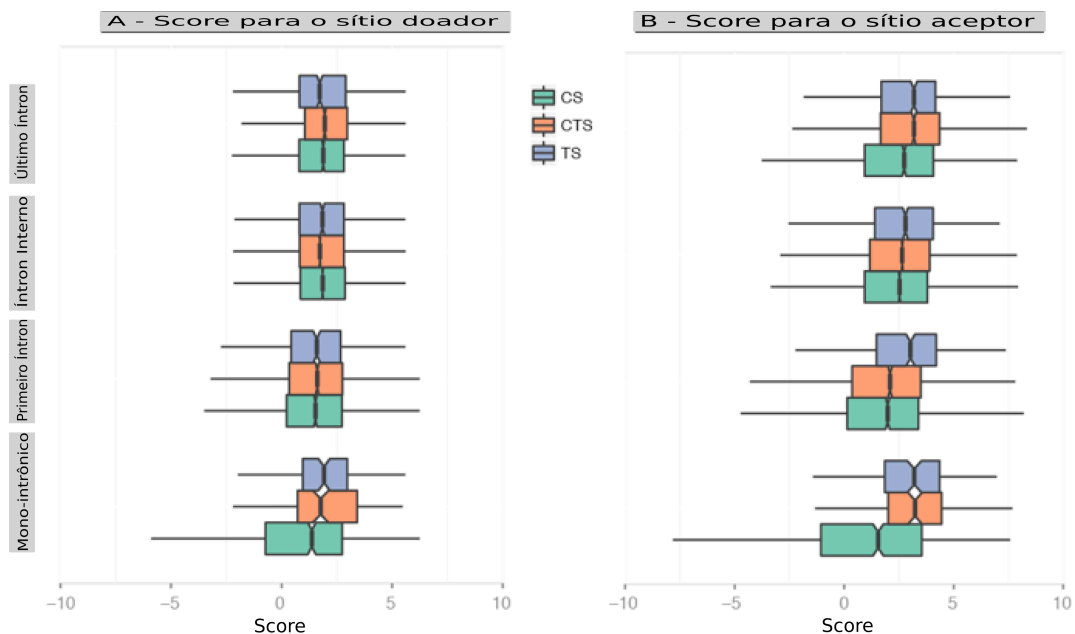


FIGURA 30 - CARACTERÍSTICAS DOS SÍTIOS DE *SPLICING* SUBDIVIDIDAS DE ACORDO COM A CLASSIFICAÇÃO DO ÍNTRON E SUA POSIÇÃO RELATIVA NO TRANSCRITO. Distribuição dos scores de sítios doadores (A) e aceptores (B) de *splicing* em relação ao tipo de íntron (ou seja, introns que são substratos para *cis-splicing* – caixas verdes, introns que são substratos para *cis-splicing* em transcritos processados por SLTS – caixas laranjas e introns que são substratos para SLTS – caixas roxas) de acordo com a posição do íntron no transcrito (ou seja, introns em monointrônicos, primeiro íntron, íntron interno, e último íntron). Os *boxplots* mostram a distribuição dos *log-odds scores*, ou seja, a força do sítio de *splice*, computado por um modelo de Markov treinado para sequências de sítios de *splice*.

Para avaliar as características dos pontos de ramificação (*branch point*), aplicamos às sequências intrônicas anotadas diferentes modelos integrados no programa SVM-BPfinder e vimos que o *score* para os *branch points* não variam substancialmente entre as diferentes classes de introns (Figura 31A). Entretanto, vimos que a distância entre os *branch points* e o sítio de *splice* 3' tendem a ser

maiores nos íntrons processados por SLTS, especialmente se considerarmos transcritos monointrônicos e os primeiros íntrons (Figura 31B). Ainda, vimos que íntrons que sofrem *cis-splicing* mostram uma zona de exclusão AG reduzida (caixa verde na Figura 31C), assim como o tamanho do trato de polipirimidina (caixa verde na Figura 31D). Em íntrons processados por SLTS, o trato de polipirimidina é mais próximo ao *branch point* (Caixa roxa na Figura 31E) e mais rico em polipirimidinas (Caixa roxa na Figura 31F). Trabalhos prévios mostraram que o tamanho do trato de polipirimidina é importante para a ocorrência de reações de *splicing in vitro* (BINDEREIF; GREEN, 1986) e Coolidge e colaboradores (1997), utilizando experimentos com mutações sítio-dirigidas, propuseram que tratos de polipirimidinas compreendendo 11 uridinas contínuas são tratos muito fortes, e que nestes casos, a posição do trato de polipirimidinas entre o sítio de ramificação e o sítio acceptor 3' não é importante. Em contraste, a diminuição do número de uridinas para cinco ou seis resíduos faz com que haja a necessidade de o trato de polipirimidinas estar localizado imediatamente adjacente ao AG para que haja uma maior eficiência. De forma análoga, vimos uma alta frequência de uridinas em regiões intrônicas próximas aos sítios aceptores de SLTS comparadas aos sítios aceptores de *cis-splicing* em transcritos processados por SLTS ou aos sítios aceptores de *cis-splicing* em transcritos processados por *cis-splicing* (Figura 28 e Figura 31).

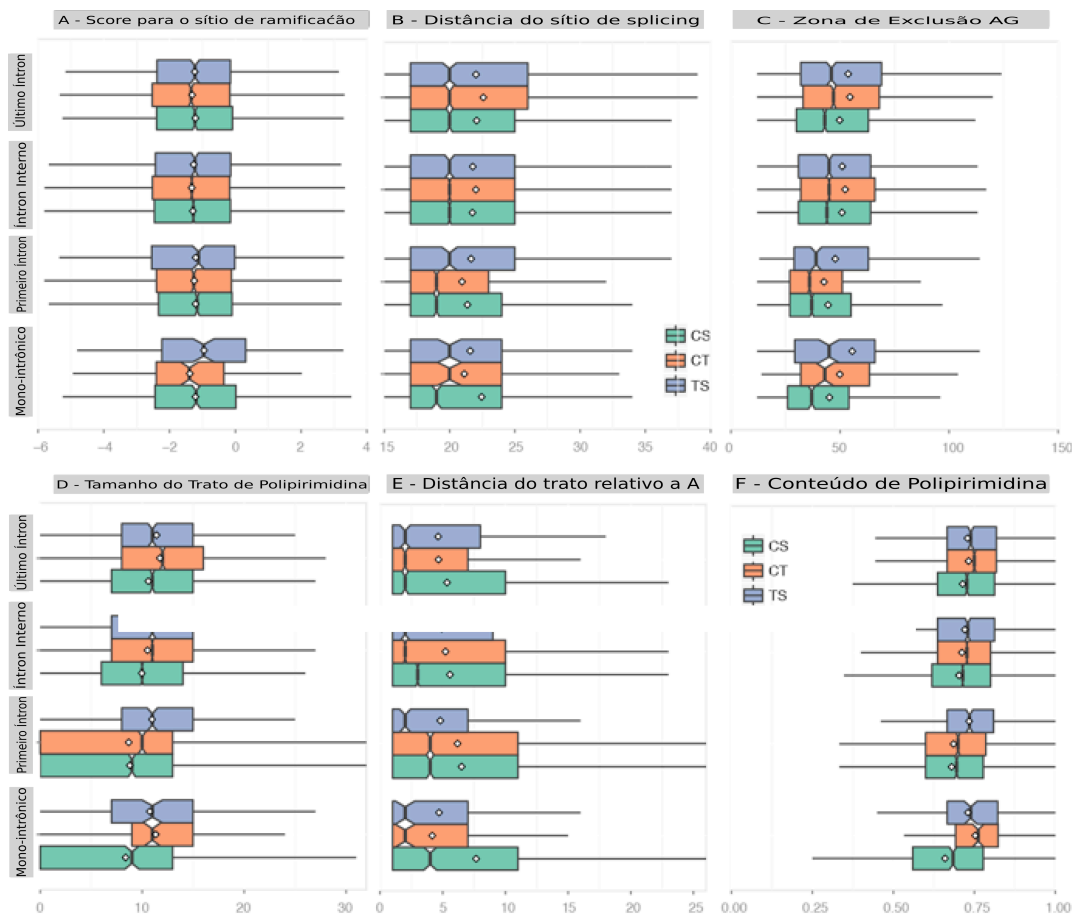


FIGURA 31 - CARACTERÍSTICAS DOS SÍTIOS DE RAMIFICAÇÃO SUBDIVIDIDA DE ACORDO COM AS CARACTERÍSTICAS DO ÍNTRON E SUA POSIÇÃO RELATIVA NO TRANSCRITO. Considerando o primeiro branch point à partir do sítio de splice 3', características diferentes relacionadas ao branch point são mostradas separadamente de acordo com tipo de íntron (ou seja, íntrons que são substratos para *cis-splicing* – caixas verdes, íntrons que são substratos para *cis-splicing* em transcritos processados por SLTS – caixas laranjas e íntrons que são substratos para SLTS – caixas roxas) de acordo com a posição do íntron no transcrito (ou seja, íntrons em transcritos monoíntrônicos, primeiro~, interno~, e último íntrons). **A** – Branch point score; **B** – Distância do sítio de splice; **C** - Zona de exclusão AG; **D** – Tamanho do trato de polipirimidina; **E** – Distância do trato de polipirimidina relativo à adenina; **F** – Conteúdo de polipirimidina.

Essas informações mostram que íntrons que são substrato para SLTS são geralmente mais longos (Figura 29), exibem um índice aumentado na identificação de sítios de *splicing* (Figura 30), e que os tratos de polipirimidina estão mais próximos ao ponto de ramificação, assim como com um conteúdo mais rico de polipirimidinas do que os íntrons que são substratos para o *cis-splicing* (Figura 31). As nossas observações confirmam estudos semelhantes realizados em *T. brucei* (SIEGEL; TAN; CROSS, 2005) que mostram que íntrons que sofrem *cis-splicing* são mais curtos e com tratos de polipirimidinas menos desenvolvidos. Os nossos resultados mostram que eventos de SLTS em sítios de

splice além da porção 5' do transcrito (SLTS em íntrons) não são uma observação casual, mas apresentam características particulares com implicações sobre o mecanismo de *trans-splicing*.

Em organismos onde *cis-splicing* e *trans-splicing* co-existem, não está claro como o sítio 5' doador do SL RNA se associa de forma eficiente com o sítio 3' acceptor no pre-mRNA e como é decidido quando um sítio de *splicing* será substrato para *cis-splicing* ou *trans-splicing*. A decisão pode depender da presença de sítios 5' canônicos de *splicing* e pode ser modulada por fatores como a U1 snRNP e certas proteínas SR em tripanosomatídeos (LIANG *et al.*, 2003). Em *Ascaris lumbricoides*, estudos recentes identificaram duas proteínas SL RNP específicas, sendo que uma delas interage com a SF1/BBP (proteína ligadora do *branch point*) e a associação entre essas proteínas e o *branch point*, juntamente com U2AF65, explica como a ponte entre o SL RNA e os pre-mRNAs é feita em nematoides (DENKER; ZUCKERMAN, 2002). Dessa forma, o spliceossomo poderia ser visto como uma maquinaria única carregando tanto U1 quanto SL RNP, com a ocorrência da competição de ligação entre U1 e SL RNP como principal determinante para selecionar *cis* ou *trans-splicing* (LIANG *et al.*, 2003). Esses dados corroboram nossos achados em *S. mansoni*, onde características que favorecem um determinado sítio 3' acceptor de *splicing* como substrato para o mecanismo de SL *trans-splicing* foram destacadas.

4.11. O MECANISMO DE SLTS EM DIFERENTES FASES DO CICLO DE VIDA DO PARASITO *S. MANSONI*

4.11.1. GENES IDENTIFICADOS: COMPARAÇÃO ENTRE AS RÉPLICAS BIOLÓGICAS E OS DIFERENTES TIPOS DE CONJUNTO DE DADOS

Para analisarmos se o mecanismo de SLTS pode ser regulado de forma diferencial entre as fases do ciclo de vida de *S. mansoni*, nós construímos bibliotecas de duas réplicas biológicas de 4 diferentes estágios do ciclo de vida do parasito (adulto, esquistossômulo, miracídio e esporocisto) enriquecidas em transcritos processados por SLTS. Foi observada uma forte correlação quantitativa entre as replicatas biológicas de adulto (PCC = 0,91, Figura 32A), esquistossômulo (PCC = 0,98, , Figura 32B) e miracídio (PCC = 0,94, Figura 32C), assim como qualitativa, pois tiveram respectivamente cerca de 59%, 53% e 48% dos *loci* identificados compartilhados pelas réplicas biológicas. Já as bibliotecas de esporocisto (PCC = 0,89, Figura 32D) apresentaram somente 39% de compartilhamento dos *loci* identificados. Esses dados enfatizam a qualidade dos dados gerados.

Para avaliar a semelhança global entre as amostras e averiguar a qualidade das réplicas biológicas, utilizou-se a função plotPCA, disponível no pacote DESeq2 que emprega o algoritmo Análise de Componentes Principais (PCA). Para evitar que a medida de distância seja dominada por genes altamente variáveis, e que exista uma contribuição equilibrada de todos os genes, utilizamos uma função implementado pelo pacote DESeq2, denominada transformação de estabilização da variância (VST).

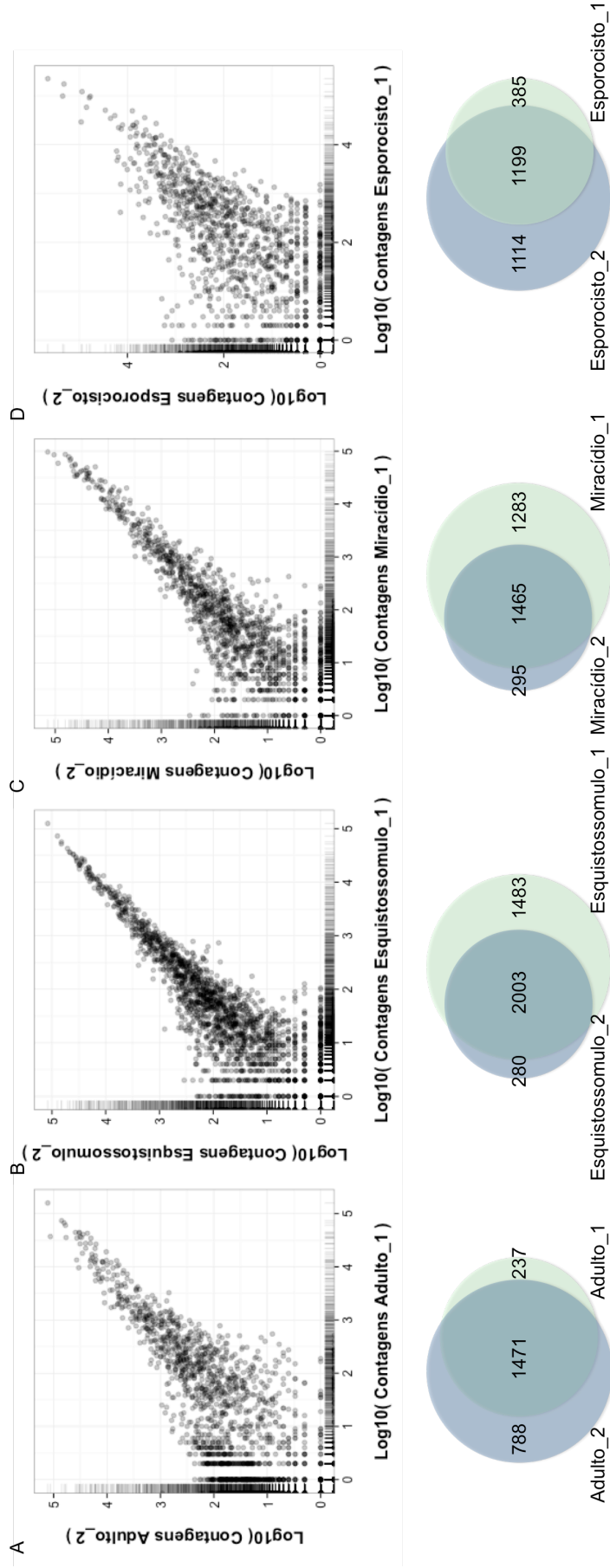


FIGURA 32 - COMPARAÇÃO ENTRE AS DUAS RÉPLICAS BIOLÓGICAS DAS BIBLIOTECAS SL ENRICHED DE CADA FASE DO CICLO DE VIDA. **A** - Painel superior: correlação entre as réplicas biológicas da fase adulto (PCC = 0,91); Painel inferior: Genes detectados nas duas réplicas biológicas da fase adulto. **B** - Painel superior: correlação entre as réplicas biológicas da fase esquistossomulo (PCC = 0,98); Painel inferior: Genes detectados nas duas réplicas biológicas da fase esquistossomulo. **C** - Painel superior: correlação entre as réplicas biológicas da fase miracídio (PCC = 0,94); Painel inferior: Genes detectados nas duas réplicas biológicas da fase miracídio. **D** - Painel superior: correlação entre as réplicas biológicas da fase esporocisto (PCC = 0,89); Painel inferior: Genes detectados nas duas réplicas biológicas da fase esporocisto.

Apesar das duas réplicas biológicas de cada fase agruparem-se em conjunto, como pode ser visto na Figura 33, as réplicas biológicas da fase esporocisto são bastante diferentes entre si e se mostram heterogêneas das demais, o que pode enviesar a análise.

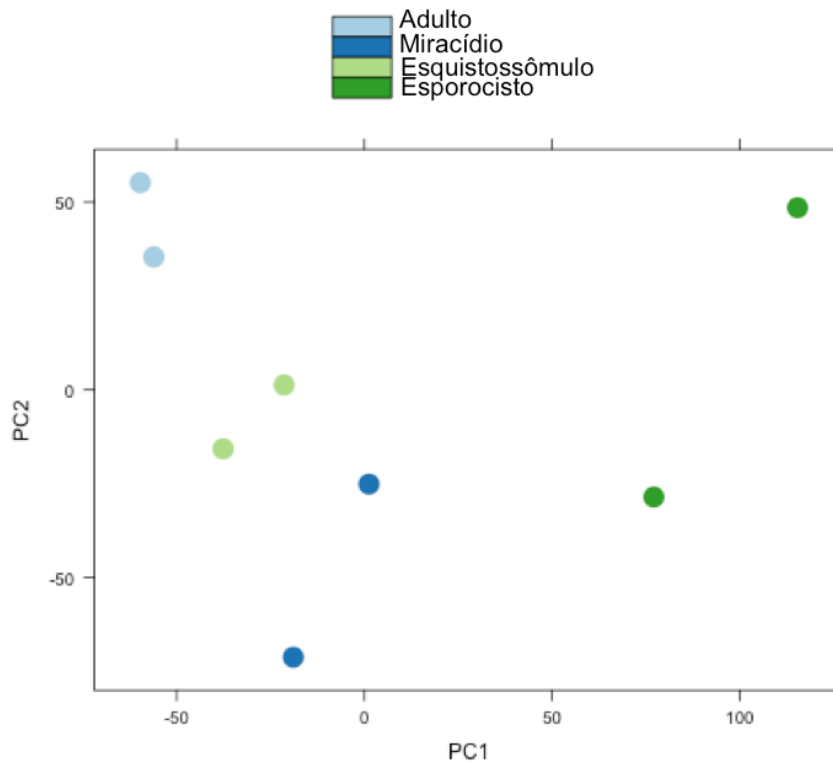


FIGURA 33- COMPARAÇÃO ENTRE AS AMOSTRAS. -Análise de componentes principais (PCA) entre as amostras após transformação da estabilização da variância.

O diagrama de Venn a seguir na Figura 34 ilustra a relação entre o número de genes identificados nas diferentes fases do ciclo de vida do parasito. O círculo em verde representa os transcritos identificados na biblioteca correspondente à fase esquistossômulo (maior número de *reads* e, portanto, maior número de genes identificados), enquanto os círculos amarelo, vermelho e azul representam respectivamente as fases esporocisto, miracídio e adulto. O número de genes em comum identificados entre as diferentes fases foi 1.384.

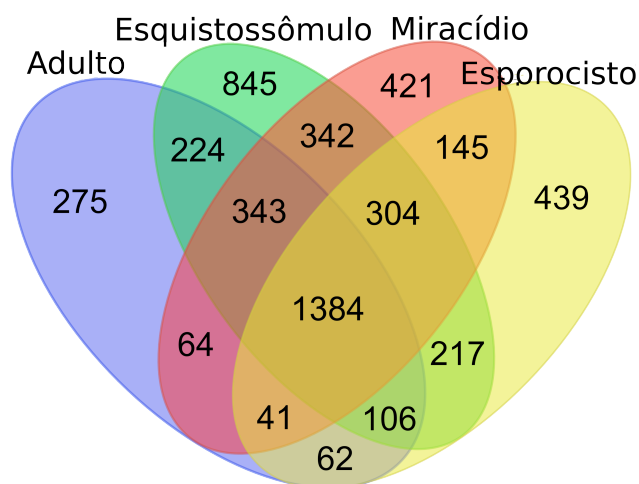


FIGURA 34 - DIAGRAMA DE VENN REPRESENTANDO OS TRANSCRITOS IDENTIFICADOS NAS BIBLIOTECAS DAS DIFERENTES FASES DO CICLO DE VIDA DE *S. MANSONI*. O círculo em azul representa o conjunto de transcritos identificados na fase adulto. O círculo em verde, na fase esquistossômulo em vermelho na fase miracídio e o círculo em amarelo representa o conjunto de transcritos identificados na biblioteca esporocisto.

4.11.2. EXPRESSÃO DIFERENCIAL DE GENES PROCESSADOS POR SLTS ENTRE DIFERENTES FASES DO PARASITO

Para testar a expressão diferencial dos genes que sofrem SLTS entre as quatro diferentes fases, utilizamos o pacote estatístico DESeq2 (ANDERS; HUBER, 2010). Utilizando um valor de cutoff de p_{BH} -ajustado $\leq 0,01$ e de $|\log_2\text{FoldChange}| \geq 1$ foram observados vários genes fase-específicos, assim como genes que tiveram expressão aumentada ou diminuída entre as fases, indicando a importância desse tipo de processamento pós-transcricional na regulação da expressão gênica do parasito em suas diferentes fases do ciclo de vida. Foram encontrados 164 genes processados por SLTS diferencialmente expressos (DEG) entre as fases adulto e miracídio, 72 entre as fases adulto e esquistossômulo, 639 entre adulto e esporocisto, 40 entre esquistossômulo e miracídio, 645 entre esquistossômulo e esporocisto e 418 entre esporocisto e miracídio. A Figura 35 ilustra a relação entre o número de DEG entre as diversas

condições avaliadas, representados por pontos azuis e os demais, representados por pontos vermelhos.

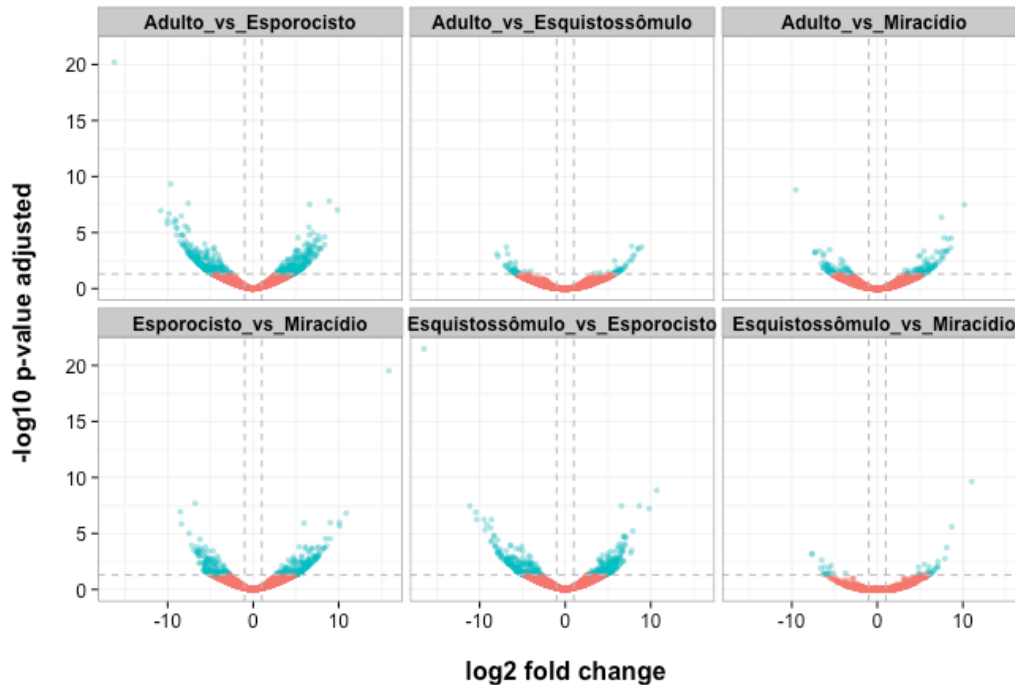


FIGURA 35 – GENES PROCESSADOS POR SLTS DIFERENCIALMENTE EXPRESSOS ENTRE AS DIFERENTES FASES DO CICLO DE VIDA DE *S. MANSONI*. A expressão dos genes processados por SLTS foi avaliada em cada uma das quatro fases do ciclo de vida do parasito e comparadas para determinar os DEG entre as fases, representados por pontos azuis. Eixo y representa a significância do teste estatístico (-log 10 p-value) enquanto o eixo x representa a razão entre os valores de expressão estimados em cada condição testada (log 2 fold change).

4.11.1. CLASSES GÊNICAS FUNCIONAIS DIFERENCIALMENTE REPRESENTADAS ENTRE AS FASES

Os genes diferencialmente expressos (DEG) identificados em cada comparação entre as fases foram utilizados para verificar a presença de funções biológicas enriquecidas: Na comparação entre as fases Adulto x Esquistossômulo e Esquistossômulo x Miracídio, nenhuma classe foi encontrada enriquecida (Figura 36D, respectivamente). Já para as fases Adulto x Miracídio, a classe gênica “Transdução de sinal” apresentou-se enriquecida (p-value < 0,003, (Figura 36A). Na comparação entre Adulto x Esporocisto, vimos que as classes “Metabolismo Intermediário” (p-value < 0,0006), “Metabolismo de Nucleotídeos” (p-value <

0,04), “Exportação Nuclear” (p-value < 1,05086E-05), “Maquinaria de Síntese Proteica” (p-value < 0,009), “Maquinaria de Transcrição” (p-value < 0,016), “Elementos Tramponíveis” (p-value < 0,011) e “Desconhecidos” (p-value < 0,04) encontraram-se enriquecidas (Figura 36C). As classes “Exportação Nuclear” (p-value < 0,0002), “Maquinaria de Síntese Proteica” (p-value < 5,98E-05), “Secretados” (p-value < 0,015), “Maquinaria de Transcrição” (p-value < 0,001), “Transportadores/Estoque” (p-value < 0,003) e “Desconhecidos” (p-value < 0,03) encontraram-se enriquecidas na comparação entre os genes Esquistossômulo e Miracídio (Figura 36E). Para as fases Esporocisto x Miracídio (Figura 36F) vimos que as categorias “Metabolismo Energético” (p-value < 0,02), “Metabolismo intermediário” (p-value < 0,04), “Exportação Nuclear” (p-value < 1,099E-09), “Maquinaria de Síntese Proteica” (p-value < 0,005), “Transportadores/Estoque” (p-value < 0,014) e “Elementos Transponíveis” (p-value < 0,041) encontram-se diferencialmente representadas entre essas fases.

4.11.2. CATEGORIAS GÊNICAS E ANÁLISES DE AGRUPAMENTO

Com o objetivo de identificar genes que são processados por SLTS e co-expressos, fizemos uma análise de agrupamento com os padrões de expressão dos 1.000 genes com a maior média de contagem entre as fases. Para tanto, os valores de contagens foram transformados utilizando o algoritmo VST implementado no pacote DESeq2. Na Figura 37 é possível observar um dendograma na parte superior, mostrando o agrupamento das amostras. Este procedimento também reflete a grande diferença entre as amostras de esporocistos com relação às demais. No canto esquerdo da figura é possível observar o agrupamento de genes que possuem padrões de expressão

semelhantes. O dendograma foi aparado em três grandes grupos gênicos, coloridos em vermelho, azul e verde (Figura 37). Como previsto, com base nas alterações morfológicas e ambientais encontradas nos diferentes estágios do desenvolvimento de *S. mansoni*, existem diferenças claras no padrão de processamento de genes por SLTS entre os estágios (Figura 37).

As categorias significativamente enriquecidas nos diferentes grupos são descritas na Tabela 10:

O Grupo 1, representado em vermelho nas figuras 39 e 40, apresenta a categoria “Maquinaria de síntese proteica” enriquecida. Esta classe gênica está mais expressas em esporocistos em relação às outras fases. O estágio esporocisto reside no caramujo (hospedeiro intermediário) e a função biológica chave desta fase do ciclo de vida é apoiar a diferenciação e desenvolvimento de um grande número de cercárias, que é a fase larval aquática infectante do hospedeiro mamífero. Durante a fase de esporocisto muitos dos transcritos mais expressos estão relacionados com produtos gênicos que atuam na síntese de proteínas em geral (subunidades ribossomais 40S e 60S, fator de alongamento), e no correto dobramento das proteínas (JOLLY *et al.*, 2007).

O Grupo 2 apresentam genes enriquecidos nas classes “Maquinaria de síntese Proteica”, “Maquinaria de Transcrição” e “Metabolismo de Nucleotídeos” (representado em azul nas figuras 39 e 40). Estes genes encontram-se menos expressos em esporocisto e expressos em maior intensidade na fase adulta.

Já o Grupo 3, representado em verde nas figuras 39 e 40, encontra-se enriquecido em genes relacionados com a “Maquinaria de modificação proteica” e “Metabolismo oxidantes/detoxificação”, apresentando muitos genes de proteínas relacionadas ao proteassomo, uma vez que na fase adulta o parasito

passa por grandes estresses oxidativos causados pelo sistema imune humano e pelo seu próprio metabolismo, e o sistema proteassomal é responsável pela homeostase proteica durante o estresse oxidativo (DE PAULA *et al.*, 2014).

TABELA 10- CLASSES ENRIQUECIDAS ENTRE OS CLUSTERS GÊNICOS. (p-value < 0.05, teste χ^2)

| | Grupo 1 | Grupo 2 | Grupo 3 |
|-------------------------------------|---------|---------|---------|
| Citoesqueleto | | | |
| Matrix extracelular/ adesão celular | | | |
| Imunidade | X | X | |
| Metabolismo de aminoácido | | | |
| Metabolismo de carboidrato | | | |
| Metabolismo energético | X | | |
| Metabolismo intermediário | | | |
| Metabolismo de lipídeos | | | |
| Metabolismo de Nucleotídeos | X | | |
| Exportação do núcleo | | | |
| Regulação do núcleo | | | |
| Metabolismo oxidantes/detoxificação | | | X |
| Maquinaria do Proteasomo | | | X |
| Maquinaria de exportação proteica | X | | |
| Maquinaria de modificação proteica | | | X |
| Maquinaria de síntese proteica | X | X | X |
| Secretado | X | | X |
| Transdução de sinal | | X | X |
| Transdução de sinal, apoptose | X | | |
| Estoque | | | |
| Fator de transcrição | | | |
| Maquinaria de transcrição | | X | X |
| Transportadores/ estoque | X | X | X |
| Elementos transponíveis | X | X | X |
| Desconhecido | X | | |
| Desconhecido, conservado | | | |
| Viral | | | |



FIGURA 36 - CLASSIFICAÇÃO FUNCIONAL DOS GENES DIFERENCIALMENTE EXPRESSOS ENTRE AS FASES DE S. MANSONI. Em vermelho está representado a frequência das classes funcionais relacionadas aos genes do parasito e em azul aos genes que sofrem SLTS diferencialmente expressos (DEG) entre as fases: **A** - Adulto x Miracido; **B** - Adulto x Esporocisto; **C** - Adulto x Esporocisto; **D** - Esquistossômo x Miracido; **E** - Esquistossômo x Esporocisto; **F** - Esquistossômo x Miracido.

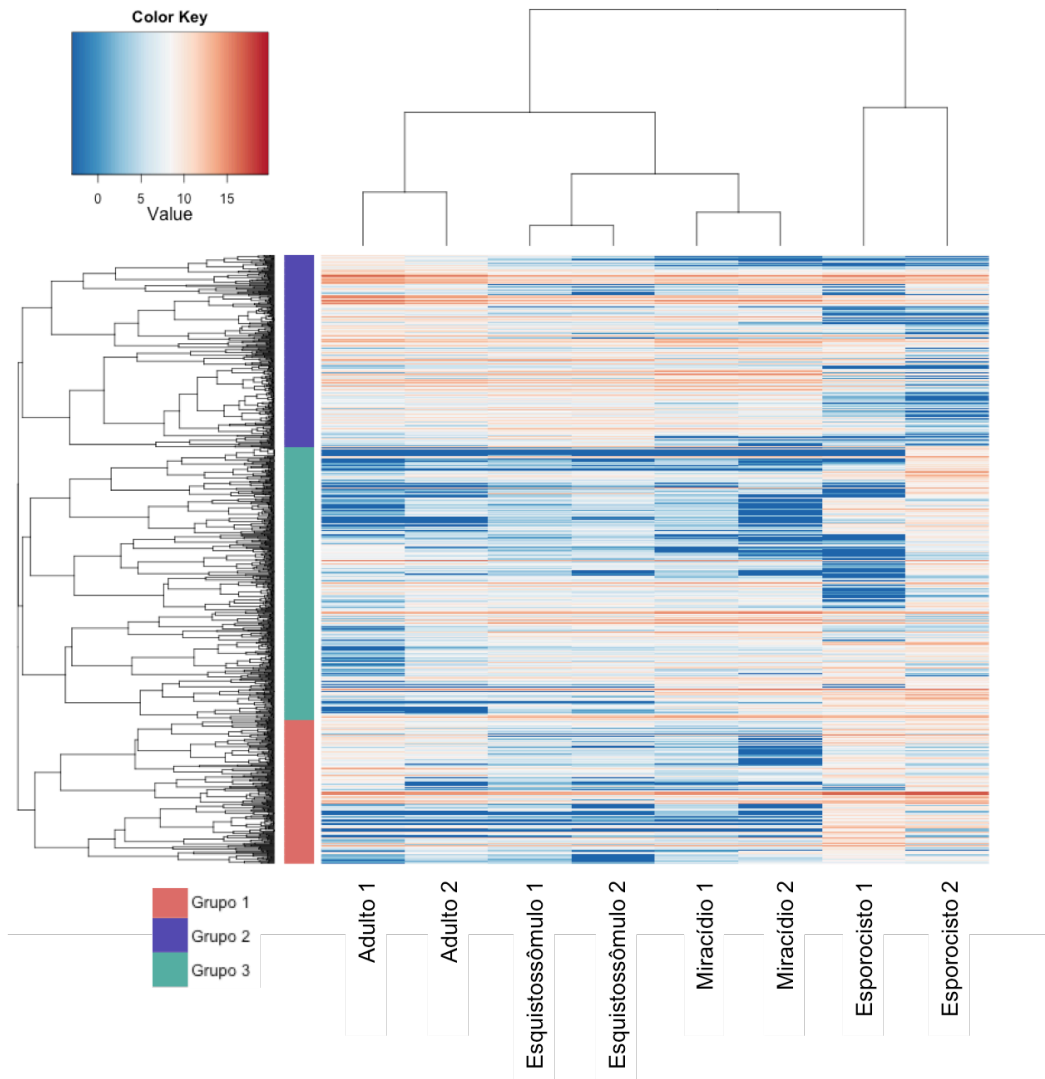


FIGURA 37 - HEATMAP MOSTRANDO A CLUSTERIZAÇÃO HIERÁRQUICA ENTRE AS AMOSTRAS DE DIFERENTES ESTÁGIOS DE DESENVOLVIMENTO E OS GENES. Foram utilizados 1000 dados dados de expressão após transformação por estabilização de variância. Utilizando-se a ferramenta cutree separamos os dados de expressão em quatro grandes clusters gênicos de co-expressão, identificados pelas cores amarelo, verde, rosa e roxo. Genes super-expressos são mostrados em vermelho e genes de baixa expressão em azul. As linhas em branco representam ausência de uma regulação de ativação ou inibição.

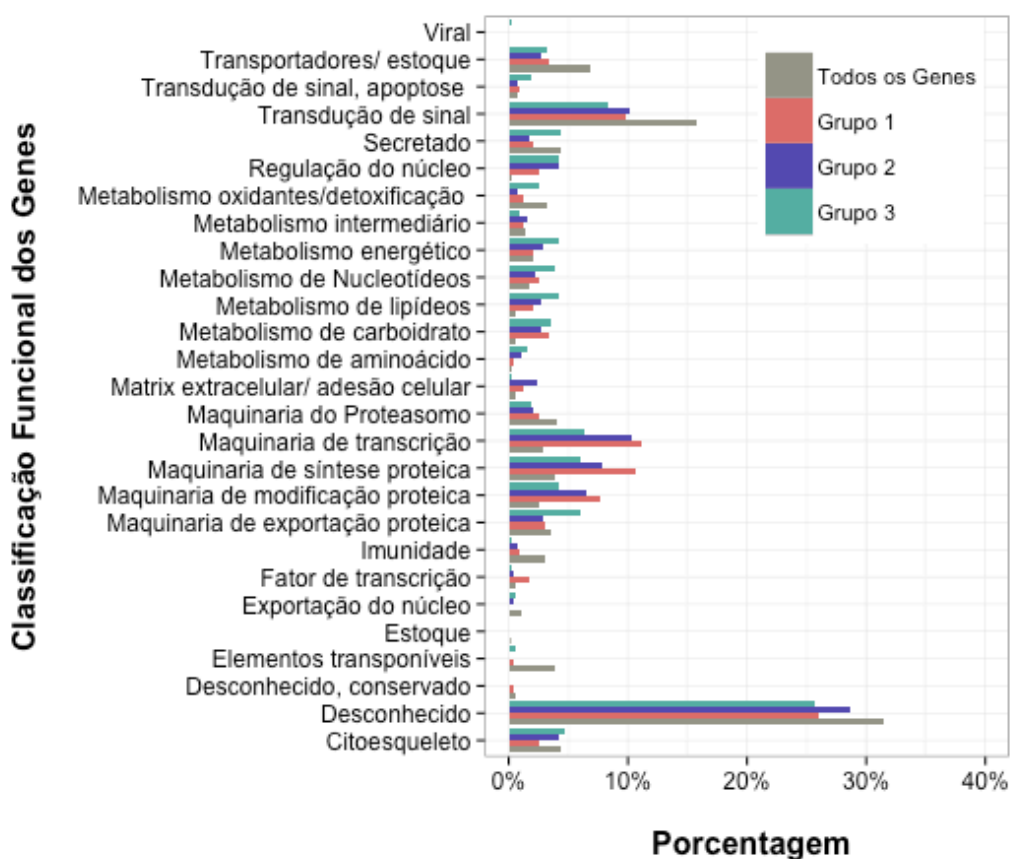


FIGURA 38 - CLASSIFICAÇÃO FUNCIONAL DOS DOS GENES AGRUPADOS EM DIFERENTES GRUPOS DE ACORDO COM O SEU PADRÃO DE EXPRESSÃO NOS ESTÁGIOS DE DESENVOLVIMENTO DO PARASITO.

4.11.3. SLTS ALTERNATIVO ENTRE OS DIFERENTES ESTÁGIOS DE DESENVOLVIMENTO DO PARASITO.

Com o intuito de ESTUDAR a entrada alternativa da sequência do SL nos transcritos processados por SLTS, nós avaliamos a expressão de cada exon nos genes processados por SLTS, utilizando o pacote do bioconductor DEXSeq v1.10.6. De forma surpreendente, vimos que o SLTS alternativo é regulado diferentemente entre os estágios distintos do parasito. Na Figura 39A é possível observar a entrada alternativa do SL em relação às diferentes fases usando como exemplo os genes Smp_006680 e Smp_008660. No gene Smp_006680 a entrada

da sequência do SL ocorre no terceiro exon tanto para a fase adulta (linha em verde) quanto para esquistossômulo (linha em laranja). Entretanto, na fase esquistossômulo é possível também observar a entrada da sequência do SL no último exon (exon 6), logo após um íntron longo (características do mecanismo de SLTS em *S. mansoni* já discutidas anteriormente). Entretanto, na Figura 39B podemos observar que o gene Smp_006680 é mais expresso na fase adulto do que na fase esquistossômulo (PROTASIO *et al.*, 2012), mostrando que não existe uma correlação direta entre os níveis de expressão e a frequência de SLTS. Já no segundo exemplo, é possível observar o processamento de SLTS ocorrendo no exon 6, tanto na fase esquistossômulo (linha em laranja) quanto em esporocisto (linha em azul), com entrada alternativa no último exon para esquistossômulo. Correlacionando com dados de expressão gênica (PROTASIO *et al.*, 2012), é notório que este gene encontra-se mais expresso em adulto do que em esquistossômulo. Entretanto, o mesmo gene, apesar de muito expresso em adultos, não é processado por SLTS nesta fase. O SLTS alternativo tem sido observado em diversos organismos: um estudo realizado em *Leishmania major* (RASTROJO *et al.*, 2013) mostrou que 50% dos genes no parasito possuem sítios de inserção de SL adicionais e observações similares já foram reportadas em estudos de RNA-Seq para *T. brucei* (NILSSON *et al.*, 2010), *C. elegans* (ALLEN *et al.*, 2011) e *C. intestinalis* (MATSUMOTO *et al.*, 2010). Entretanto, as funções do SLTS alternativo ainda não são muito bem conhecidas. Existe um caso documentado de SLTS alternativo do gene LYT1 em *T. cruzi*, cuja maturação alternativa pela entrada da sequência do SL permite a expressão de uma isoforma proteica que apresenta propriedades funcionais e localização celular diferentes (BENABDELLAH; GONZÁLEZ-REY; GONZÁLEZ, 2007). Outras funções

propostas para o SLTS alternativo são: inibição da tradução, devido à falta da AUG, mudança de alvo devido à perda de peptídeo sinal, sinal de localização celular ou sinal de ancoramento de membrana, exclusão ou inclusão de elementos regulatórios como uORFs, que poderiam atuar na eficiência da tradução e estabilidade do transcrito e ainda uso alternativo de ORFs (NILSSON *et al.*, 2010; SAITO *et al.*, 2013).

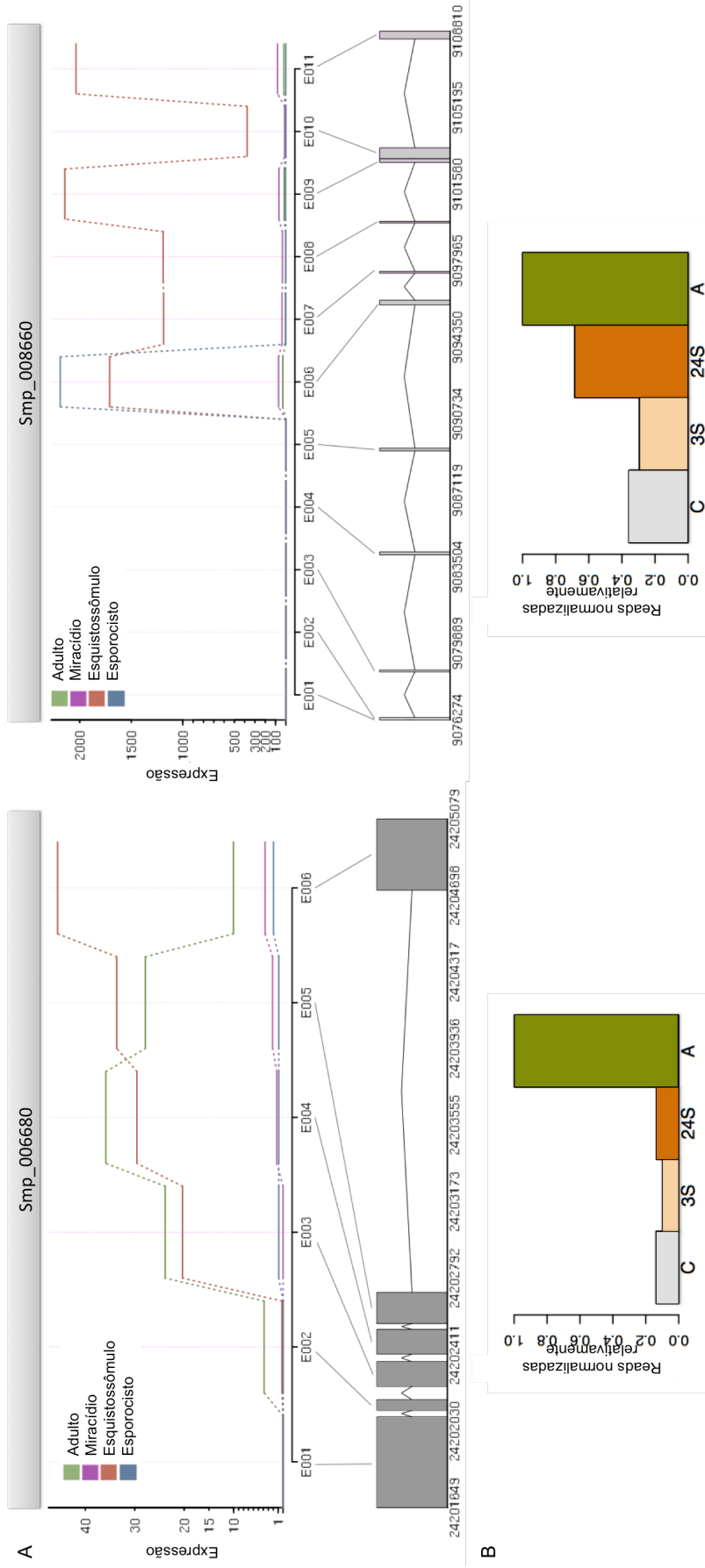


FIGURA 39 – ENTRADA ALTERNATIVA DA SEQUÊNCIA DO SI EM GENES PROCESSADOS POR SLTS EM DIFERENTES ESTÁGIOS DE DESENVOLVIMENTO DE *S. MANSONI*. A – Gráfico mostrando a expressão identificada em cada exon do transcrito para as fases adulto (verde), miracídio (rosa), esquistossômulo (laranja) e esporocisto (azul). B – Valores de expressão gênica nas fases cercária (cinza), esquistossômulo após 3 horas de transformação (laranja claro), esquistossômulo após 24 horas de transformação (laranja escuro) e vermes adultos (verde). Esses valores foram obtidos pelos dados de Protasio *et al.*, 2012 .

5. CONCLUSÕES

Neste trabalho, nós apresentamos o primeiro estudo aprofundado do mecanismo de SLTS no *S. mansoni*, no qual foram empregadas três diferentes metodologias para geração de dados. Isso permitiu um estudo minucioso dos eventos de SLTS em cinco diferentes fases do ciclo de vida do parasito. Nossas investigações demonstram que a técnica de SL Trapping permite a detecção de uma ampla gama de transcritos e, principalmente, permite estudar os sítios de inserção da sequência SL. Ainda, foi observado que os dados aqui gerados puderam ser reproduzidos empregando as outras metodologias utilizadas nesse trabalho, RNA-Seq Filtered e SL Enriched, pois houve uma grande sobreposição dos transcritos identificados nesses conjuntos de dados, além da confirmação de resultados prévios obtidos utilizando-se a técnica de RNA-Seq. Adicionalmente, esta técnica estendeu o limite de resolução do transcriptoma processado por SLTS em uma profundidade sem precedentes, permitindo-nos explorar uma pluralidade de eventos raros de SLTS, que não foram contemplados em estudos prévios. Dessa forma, nossas análises indicam que cerca de 77% dos genes codificadores de proteínas em *S. mansoni* podem sofrer processamento por SLTS em alguma fase de seu ciclo de vida. A geração das bibliotecas de SL Trapping também permitiram a detecção de novos genes candidatos, destacando a relevância desta abordagem como uma fonte de informação para melhorar a anotação do genoma do parasito.

Em nosso estudo identificamos 32 loci no genoma contendo a sequência do SL, sendo que 11 desses sítios puderam ser associados a genes que possivelmente sofreram uma retro-transcrição eventual após processamento por

SLTS e posterior inserção no genoma. Ainda, extendemos o número de 46 policistrons (PROTASIO *et al.*, 2012) para um total de 65 transcritos dicistrônicos e uma unidade tricistrônica conhecidos atualmente no parasito. Foi também observado que a mesma sequência do SL é utilizada no processamento tanto dos transcritos monocistrônicos quanto dos policistrônicos. Como esperado, não observamos preferências para uma categoria funcional ou origem cromossômica para os genes processados por SLTS, sugerindo que o SLTS é um mecanismo ubíquo, capaz de atuar em genes envolvidos em diferentes vias metabólicas em *S. mansoni*.

Os resultados apresentados aqui mostram que a detecção de SLTS é tecnicamente facilitada em genes de altos valores de expressão, entretanto observamos que genes de baixa expressão podem ser processados por SLTS com uma frequência relativamente alta, não sendo possível estabelecer uma correlação direta entre a expressão gênica e a frequência de SLTS em *S. mansoni*. Mostramos também que genes que exibem maiores números de isoformas decorrentes de *splicing* alternativo apresentam frequências menores de SLTS, enquanto que genes com alto número de exons, mas com baixo número de isoformas de *cis-splicing* comparativamente são mais sujeitos ao processamento por SLTS, sugerindo uma maior eficiência do spliceossomo na ausência de estruturas complexas de *splicing* competindo em *cis*. Apesar de a entrada da sequência SL ocorrer principalmente no primeiro exon, foi apontado nesse estudo que uma proporção substancial dos eventos de SLTS em *S. mansoni* ocorrem em íntrons e nossas investigações revelaram que existe uma preferência de inserção da sequência SL nos íntrons que flanqueiam os exons das extremidades 5' e 3' do transcrito, independentemente do tamanho do

transcrito. Ainda foi possível observar que o aumento do tamanho dos íntrons pode favorecer a predisposição para o *trans-splicing* e que íntrons sujeitos ao processamento por SLTS apresentam sítios aceptores mais fortes assim como um trato de polipirimidina mais longo e com um conteúdo mais rico em pirimidinas, mostrando que vários fatores desempenham um papel fundamental na decisão de quando um sítio de *splicing* será substrato para *cis-splicing* ou *trans-splicing*.

Foi também ressaltada a existência de uma regulação diferencial do mecanismo de SLTS entre os diferentes estágios de desenvolvimento de *S. mansoni*. Esta regulação envolve a entrada alternativa da sequência do SL nos transcritos gerando isoformas diferentes em diferentes fases, suportando a idéia de que o *trans-splicing* alternativo tem uma função importante na regulação do desenvolvimento do parasito. Entretanto, o impacto do SLTS alternativo nos produtos proteicos gerados em *S. mansoni* permanece desconhecida.

Por fim, esse trabalho mostra a importância do mecanismo de SLTS na regulação gênica de *S. mansoni*, um parasita humano causador de uma doença importante e negligenciada. Compreender o mecanismo de SLTS pode auxiliar no desenvolvimento futuro de uma ferramenta de intervenção terapêutica direcionada para os componentes específicos desse mecanismo em *S. mansoni*.

6. REFERÊNCIAS BIBLIOGRÁFICAS

- AIRD, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. **Genome biology**, v. 12, n. 2, p. R18, 2011.
- ALLEN, M. A. *et al.* A global analysis of *C. elegans* trans-splicing. **Genome Research**, v. 21, p. 255–264, 2011.
- ALTSCHUL, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Research**, v. 25, n. 17, p. 3389–3402, 1997.
- ANDERS, S.; HUBER, W. Differential expression analysis for sequence count data. **Genome Biology**, v. 11, n. 10, p. R106, 2010.
- ANDERS, S.; REYES, A.; HUBER, W. Detecting differential usage of exons from RNA-seq data. **Genome Research**, v. 22, n. 10, p. 2008–2017, 2012.
- ARONESTY, E. Comparison of Sequencing Utility Programs. **The Open Bioinformatics Journal**, v. 7, p. 1–8, 2013.
- BASCH, P. F. Cultivation of *Schistosoma mansoni* in vitro. I. Establishment of cultures from cercariae and development until pairing. **The Journal of Parasitology**, v. 67, n. 2, p. 179–185, 1981.
- BENABDELLAH, K.; GONZÁLEZ-REY, E.; GONZÁLEZ, A. Alternative trans-splicing of the *Trypanosoma cruzi* LYT1 gene transcript results in compartmental and functional switch for the encoded protein. **Molecular Microbiology**, v. 65, n. 6, p. 1559–1567, 2007.
- BENGTSSON, M. *et al.* Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. **Genome Research**, v. 15, n. 10, p. 1388–1392, 2005.
- BENTLEY, D. L. Coupling mRNA processing with transcription in time and space. **Nature reviews. Genetics**, v. 15, n. 3, p. 163–175, 2014.
- BERRIMAN, M. *et al.* The genome of the blood fluke *Schistosoma mansoni*. **Nature**, v. 460, n. 7253, p. 352–358, 2009.
- BINDEREIF, A.; GREEN, M. R. Ribonucleoprotein complex formation during pre-mRNA splicing in vitro. **Molecular and Cellular Biology**, v. 6, n. 7, p. 2582–2592, 1986.
- BITAR, M. *et al.* The spliced leader trans-splicing mechanism in different organisms: molecular details and possible biological roles. **Frontiers in Genetics**, v. 4, p. 199, 2013.
- BLANCO, E.; PARRA, G.; GUIGÓ, R. Using geneid to identify genes. **Current protocols in bioinformatics / editorial board, Andreas D. Baxevanis ... [et al.]**, v. Chapter 4, p. Unit 4.3, jun. 2007.
- BLUMENTHAL, T. Operons in eukaryotes. **Briefings in Functional Genomics & Proteomics**, v. 3, n. 3, p. 199–211, 2004.
- BLUMENTHAL, T. Trans-splicing and operons. **WormBook: the online review of C. elegans biology**, p. 1–9, 2005.

- BLUMENTHAL, T.; GLEASON, K. S. *Caenorhabditis elegans* operons: form and function. **Nature reviews. Genetics**, v. 4, n. 2, p. 112–120, 2003.
- BRAGG, L. M. *et al.* Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. **PLoS Computational Biology**, v. 9, n. 4, p. e1003031, 2013.
- BRUZIK, J. P. *et al.* Trans splicing involves a novel form of small nuclear ribonucleoprotein particles. **Nature**, v. 335, n. 6190, p. 559–562, 1988.
- CHOI, J.; NEWMAN, A. P. A two-promoter system of gene expression in *C. elegans*. **Developmental Biology**, v. 296, n. 2, p. 537–544, 2006.
- CONRAD, R. *et al.* Insertion of part of an intron into the 5' untranslated region of a *Caenorhabditis elegans* gene converts it into a trans-spliced gene. **Molecular and Cellular Biology**, v. 11, n. 4, p. 1921–1926, 1991.
- CONRAD, R. C.; LEA, K.; BLUMENTHAL, T. SL1 trans-splicing specified by AU-rich synthetic RNA inserted at the 5' end of *Caenorhabditis elegans* pre-mRNA. **RNA**, v. 1, n. 2, p. 164–170, 1995.
- COOLIDGE, C. J.; SEELY, R. J.; PATTON, J. G. Functional analysis of the polypyrimidine tract in pre-mRNA splicing. **Nucleic Acids Research**, v. 25, n. 4, p. 888–896, 1997.
- COPELAND, C. S. *et al.* Homology-based annotation of non-coding RNAs in the genomes of *Schistosoma mansoni* and *Schistosoma japonicum*. **BMC Genomics**, v. 10, p. 464–477, 2009.
- CORVELO, A. *et al.* Genome-wide association between branch point properties and alternative splicing. **PLoS Computational Biology**, v. 6, n. 11, p. e1001016, 2010.
- COSTA, V. *et al.* Uncovering the Complexity of Transcriptomes with RNA-Seq. **Journal of Biomedicine and Biotechnology**, v. 2010, p. 1–19, 2010.
- CROOKS, G. E. *et al.* WebLogo: a sequence logo generator. **Genome research**, v. 14, n. 6, p. 1188–90, 2004.
- DAVIS, A. Schistosomiasis. In: MANSON, P. S.; COOK, G. C.; ZUMLA, A. (Eds.). **Manson's tropical diseases**. Edinburgh: Saunders, 2002. p. 1434–1469.
- DAVIS, R. E. Spliced leader RNA trans-splicing in metazoa. **Parasitology Today**, v. 12, n. 1, p. 33–40, 1996.
- DAVIS, R. E.; HARDWICK, C.; TAVERNIER, P. RNA Trans-splicing in Flatworms. Analysis of trans-spliced mRNAs and genes in the human parasite, *Schistosoma mansoni*. **The Journal of Biological Chemistry**, v. 270, n. 37, p. 21813–21819, 1995.
- DAVIS, R. E.; HODGSON, S. Gene linkage and steady state RNAs suggest trans-splicing may be associated with a polycistronic transcript in *Schistosoma mansoni*. **Molecular and Biochemical Parasitology**, v. 89, n. 1, p. 25–39, out. 1997.
- DE PAULA, R. G. *et al.* Biochemical characterization and role of the proteasome in the oxidative stress response of adult *Schistosoma mansoni* worms. **Parasitology Research**, 2014.
- DEMARCO, R. *et al.* Saci-1 , -2 , and -3 and Perere , Four Novel Retrotransposons with High Transcriptional Activities from the Human Parasite *Schistosoma mansoni*. **Journal of Virology**, v. 78, n. 6, p. 2967–2978, 2004.

- DENKER, J.; ZUCKERMAN, D. New components of the spliced leader RNP required for nematode trans-splicing. **Nature**, v. 417, p. 667–670, 2002.
- DERELLE, R. *et al.* Convergent origins and rapid evolution of spliced leader trans-splicing in metazoa: Insights from the Ctenophora and Hydrozoa. **RNA**, v. 16, n. 4, p. 696–707, 2010.
- DILLIES, M.-A. *et al.* A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. **Briefings in Bioinformatics**, v. 14, n. 6, p. 671–683, 2013.
- DOHM, J. C. *et al.* Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. **Nucleic Acids Research**, v. 36, n. 16, p. e105, 2008.
- DOURIS, V.; TELFORD, M. J.; AVEROF, M. Evidence for multiple independent origins of trans-splicing in Metazoa. **Molecular Biology and Evolution**, v. 27, n. 3, p. 684–693, 2010.
- DUCKERT, P.; BRUNAK, S.; BLOM, N. Prediction of proprotein convertase cleavage sites. **Protein engineering, design & selection : PEDS**, v. 17, n. 1, p. 107–112, 2004.
- ENGELS, D. *et al.* The global epidemiological situation of schistosomiasis and new approaches to control and research. **Acta Tropica**, v. 82, n. 2, p. 139–146, 2002.
- EVANS, D.; BLUMENTHAL, T. Trans splicing of polycistronic *Caenorhabditis elegans* pre-mRNAs: analysis of the SL2 RNA. **Molecular and Cellular Biology**, v. 20, n. 18, p. 6659–6667, 2000.
- FLINTOFT, L. Transcriptomics: Digging deep with RNA-Seq. **Nature Reviews Genetics**, v. 9, n. 8, p. 568–568, 2008.
- FRANCO, G. R. *et al.* Identification of new *Schistosoma mansoni* genes by the EST strategy using a directional cDNA library. **Gene**, v. 152, n. 2, p. 141–147, 1995.
- FRANCO, G. R. *et al.* Evaluation of cDNA Libraries from Different Developmental Stages of *Schistosoma mansoni* for Production of Expressed Sequence Tags (ESTs). **DNA Research**, v. 4, n. 3, p. 231–240, 1997.
- FRANCO, G. R. *et al.* The Schistosoma gene discovery program: state of the art. **International Journal for Parasitology**, v. 30, p. 453–463, 2000.
- GARBER, M. *et al.* Computational methods for transcriptome annotation and quantification using RNA-seq. **Nature Methods**, v. 8, n. 6, p. 469–477, 2011.
- GARDNER, P. P. *et al.* Rfam: Wikipedia, clans and the “decimal” release. **Nucleic Acids Research**, v. 39, p. D141–145, 2011.
- GUILIANO, D. B.; BLAXTER, M. L. Operon Conservation and the Evolution of trans-Splicing in the Phylum Nematoda. **PLoS Genetics**, v. 2, n. 11, p. 12, 2006.
- HANSEN, J. E. *et al.* NetOglyc: Prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. **Glycoconjugate Journal**, v. 15, n. 2, p. 115–130, 1998.
- HARRIS, T. W. *et al.* WormBase: a comprehensive resource for nematode research. **Nucleic Acids Research**, v. 38, p. D463–467, 2010.
- HASTINGS, K. E. M. SL trans-splicing : easy come or easy go ? **Trends in genetics**, v. 21, n. 4, p. 240–247, 2005.

- HUANG, J.; VAN DER PLOEG, L. H. Maturation of polycistronic pre-mRNA in *Trypanosoma brucei*: analysis of trans splicing and poly (A) addition at nascent RNA transcripts from the hsp70 locus. **Molecular and cellular biology**, v. 11, n. 6, p. 3180–3190, 1991a.
- HUANG, J.; VAN DER PLOEG, L. H. Requirement of a polypyrimidine tract for trans-splicing in trypanosomes: discriminating the PARP promoter from the immediately adjacent 3' splice acceptor site. **The EMBO journal**, v. 10, n. 12, p. 3877–85, dez. 1991b.
- HUMMEL, H. S.; GILLESPIE, R. D.; SWINDLE, J. Mutational analysis of 3' splice site selection during trans-splicing. **The Journal of biological chemistry**, v. 275, n. 45, p. 35522–31, 10 nov. 2000.
- ISLAM, S. et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. **Genome research**, v. 21, n. 7, p. 1160–7, jul. 2011.
- JOLLY, E. R. et al. Gene expression patterns during adaptation of a helminth parasite to different environmental niches. **Genome biology**, v. 8, n. 4, p. R65, jan. 2007.
- KANEHISA, M. et al. KEGG for integration and interpretation of large-scale molecular data sets. **Nucleic acids research**, v. 40, n. Database issue, p. D109–14, jan. 2012.
- KANEHISA, M.; GOTO, S. KEGG: kyoto encyclopedia of genes and genomes. **Nucleic acids research**, v. 28, n. 1, p. 27–30, 1 jan. 2000.
- KARIM, S.; SINGH, P.; RIBEIRO, J. M. C. A Deep Insight into the Sialotranscriptome of the Gulf Coast Tick, *Amblyomma maculatum*. **PLoS ONE**, v. 6, n. 12, p. e28525, 21 dez. 2011.
- KAWANAKA, M.; SIDNER, R. A.; CARTER, C. E. In vitro transformation of *Schistosoma japonicum* miracidia to young sporocysts in a culture system for egg maturation. **The Journal of parasitology**, v. 71, n. 3, p. 368–70, jun. 1985.
- KENT, W. J. BLAT--the BLAST-like alignment tool. **Genome research**, v. 12, n. 4, p. 656–64, abr. 2002.
- KRAUSE, M.; HIRSH, D. A trans-spliced leader sequence on actin mRNA in *C. elegans*. **Cell**, v. 49, n. 6, p. 753–761, jun. 1987.
- KUO, R. C. et al. Transcriptomic study reveals widespread spliced leader trans-splicing, short 5'-UTRs and potential complex carbon fixation mechanisms in the euglenoid Alga *Eutreptiella* sp. **PloS one**, v. 8, n. 4, p. e60826, jan. 2013.
- LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. **Nature methods**, v. 9, n. 4, p. 357–9, abr. 2012.
- LASDA, E. L.; BLUMENTHAL, T. Trans-splicing. **Developmental Biology**, v. 2, p. 417–434, 2011.
- LEE, K.-Z.; SOMMER, R. J. Operon structure and trans-splicing in the nematode *Pristionchus pacificus*. **Molecular biology and evolution**, v. 20, n. 12, p. 2097–103, dez. 2003.
- LEINONEN, R.; SUGAWARA, H.; SHUMWAY, M. The sequence read archive. **Nucleic acids research**, v. 39, n. Database issue, p. D19–21, jan. 2011.
- LEWIS, S.; ASHBURNER, M.; REESE, M. G. Annotating eukaryote genomes. **Current Opinion in Structural Biology**, v. 10, n. 3, p. 349–354, jun. 2000.
- LI, H. et al. The Sequence Alignment/Map format and SAMtools. **Bioinformatics**, v. 25, n. 16, p. 2078–2079 DO – 10.1093/bioinformatics/btp352, ago. 2009.

- LIANG, X. -H. *et al.* trans and cis Splicing in Trypanosomatids: Mechanism, Factors, and Regulation. **Eukaryotic Cell**, v. 2, n. 5, p. 830–840, 2003.
- LIU, L. *et al.* Comparison of next-generation sequencing systems. **Journal of Biomedicine & Biotechnology**, v. 2012, p. 251364, 2012.
- LOMAN, N. J. *et al.* Performance comparison of benchtop high-throughput sequencing platforms. **Nature Biotechnology**, v. 30, n. 5, p. 434–439, 2012.
- MANN, V. H. *et al.* Culture for genetic manipulation of developmental stages of *Schistosoma mansoni*. **Parasitology**, v. 137, p. 451–462, 2010.
- MARCHLER-BAUER, A. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. **Nucleic Acids Research**, v. 30, n. 1, p. 281–283, 2002.
- MARIONI, J. C. *et al.* RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. **Genome Research**, v. 18, n. 9, p. 1509–1517, 2008.
- MATERA, A G.; WANG, Z. A day in the life of the spliceosome. **Nature Reviews. Molecular Cell Biology**, v. 15, n. 2, p. 108–121, 2014.
- MATSUMOTO, J. *et al.* High-throughput sequence analysis of *Ciona intestinalis* SL trans-spliced mRNAs: alternative expression modes and gene function correlates. **Genome Research**, v. 20, n. 5, p. 636–645, 2010.
- MATTHEWS, K. R.; TSCHUDI, C.; ULLU, E. A common pyrimidine-rich motif governs trans-splicing and polyadenylation of tubulin polycistronic pre-mRNA in trypanosomes. **Genes & Development**, v. 8, n. 4, p. 491–501, 1994.
- MICHAELI, S. Trans-splicing in trypanosomes: machinery and its impact on the parasite transcriptome. **Future Microbiology**, v. 6, n. 4, p. 459–474, 2011.
- MOURÃO, M. DE M. *et al.* A directed approach for the identification of transcripts harbouring the spliced leader sequence and the effect of trans-splicing knockdown in *Schistosoma mansoni*. **Mem Inst Oswaldo Cruz**, v. 108, n. 6, p. 707–717, 2013.
- MÜLLER-MCNICOLL, M.; NEUGEBAUER, K. M. How cells get the message: dynamic assembly and function of mRNA-protein complexes. **Nature Reviews. Genetics**, v. 14, n. 4, p. 275–87, 2013.
- MURPHY, W. J.; WATKINS, K. P.; AGABIAN, N. Identification of a novel Y branch structure as an intermediate in trypanosome mRNA processing: Evidence for Trans splicing. **Cell**, v. 47, n. 4, p. 517–525, 1986.
- NIELSEN, H. *et al.* Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. **Protein Engineering**, v. 10, n. 1, p. 1–6, 1997.
- NILSSON, D. *et al.* Spliced Leader Trapping Reveals Widespread Alternative Splicing Patterns in the Highly Dynamic Transcriptome of *Trypanosoma brucei*. **PLoS Pathogens**, v. 6, n. 8, p. e1001037, 2010.
- PESSOA, S. B.; MARTINS, A. V. **Parasitologia Médica**. 11^a. ed. Rio de Janeiro: Guanabara Koogan, 1982.
- PROTASIO, A. V *et al.* A systematically improved high quality genome and transcriptome of the human blood fluke *Schistosoma mansoni*. **PLoS neglected tropical diseases**, v. 6, n. 1, p. e1455, 2012.

- PUNTA, M. *et al.* The Pfam protein families database. **Nucleic Acids Research**, v. 40, p. D290–301, 2012.
- QUAIL, M. A. *et al.* A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. **BMC Genomics**, v. 13, n. 1, p. 341, 2012.
- QUINLAN, A. R.; HALL, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. **Bioinformatics**, v. 26, n. 6, p. 841–842, 2010.
- RAJKOVIC, A. *et al.* A spliced leader is present on a subset of mRNAs from the human parasite *Schistosoma mansoni*. **Biochemistry**, v. 87, p. 8879–8883, 1990.
- RASTROJO, A. *et al.* The transcriptome of *Leishmania major* in the axenic promastigote stage: transcript annotation and relative expression levels by RNA-seq. **BMC Genomics**, v. 14, n. 1, p. 223, 2013.
- REED, R. The organization of 3' splice-site sequences in mammalian introns. **Genes & Development**, v. 3, n. 12B, p. 2113–2123, 1989.
- ROBERTS, A. *et al.* Identification of novel transcripts in annotated genomes using RNA-Seq. **Bioinformatics**, v. 27, n. 17, p. 2325–2329, 2011.
- ROTHBERG, J. M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. **Nature**, v. 475, n. 7356, p. 348–352, 2011.
- SAITO, T. L. *et al.* The transcription start site landscape of *C. elegans*. **Genome Research**, v. 23, n. 8, p. 1348–1361, 2013.
- SCHMIEDER, R.; EDWARDS, R. Quality control and preprocessing of metagenomic datasets. **Bioinformatics**, v. 27, n. 6, p. 863–4, 2011.
- SCHULTZ, J. *et al.* SMART: a web-based tool for the study of genetically mobile domains. **Nucleic Acids Research**, v. 28, n. 1, p. 231–234, 2000.
- SCHÜRCH, N. *et al.* Accurate polyadenylation of procyclin mRNAs in *Trypanosoma brucei* is determined by pyrimidine-rich elements in the intergenic regions. **Molecular and Cellular Biology**, v. 14, n. 6, p. 3668–75, jun. 1994.
- SHABAAN, A. M. *et al.* Analysis of *Schistosoma mansoni* genes using the expressed sequence Tag approach. **Techniques**, v. 50, n. 1, p. 259–268, 2003.
- SIEGEL, T. N.; TAN, K. S. W.; CROSS, G. A. M. Systematic study of sequence motifs for RNA trans splicing in *Trypanosoma brucei*. **Molecular and Cellular Biology**, v. 25, n. 21, p. 9586–9594, 2005.
- SLAMOVITS, C. H.; KEELING, P. J. Widespread recycling of processed cDNAs in dinoflagellates. **Current Biology**, v. 18, n. 13, p. R550–552, 2008.
- SMITHERS, S. R.; TERRY, R. J. The infection of laboratory hosts with cercariae of *Schistosoma mansoni* and the recovery of the adult worms. **Parasitology**, v. 55, n. 4, p. 695–700, 1965.
- SONNHAMMER, E. L.; VON HEIJNE, G.; KROGH, A. A hidden Markov model for predicting transmembrane helices in protein sequences. **Proceedings / ... International Conference on Intelligent Systems for Molecular Biology; ISMB. International Conference on Intelligent Systems for Molecular Biology**, v. 6, p. 175–182, 1998.

- SPIETH, J. *et al.* Operons in *C. elegans*: Polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions. **Cell**, v. 73, n. 3, p. 521–532, 1993.
- STANDLEY, C. J. *et al.* Molecular epidemiology and phylogeography of *Schistosoma mansoni* around Lake Victoria. **Parasitology**, v. 137, p. 1937–1949, 2010.
- STOVER, N. A.; KAYE, M. S.; CAVALCANTI, A. R. O. Spliced leader trans-splicing. **Current Biology**, v. 16, n. 1, p. 8–9, 2006.
- SUTTON, R. Evidence for Trans splicing in trypanosomes. **Cell**, v. 47, n. 4, p. 527–535, 1986.
- TARAZONA, S. *et al.* Differential expression in RNA-seq: A matter of depth. **Genome Research**, v. 21, p. 2213–2223, 2011.
- TATUSOV, R. L. *et al.* The COG database: an updated version includes eukaryotes. **BMC Bioinformatics**, v. 4, p. 41, 2003.
- TENNISWOOD, M. P. R.; SIMPSON, A. J. G. The extraction, characterization and in vitro translation of RNA from adult *Schistosoma mansoni*. **Parasitology**, v. 84, n. 02, p. 253–261, 1982.
- THOMAS, J. D.; CONRAD, R. C.; BLUMENTHAL, T. The *C. elegans* Trans-spliced leader RNA is bound to Sm and has a trimethylguanosine cap. **Cell**, v. 54, n. 4, p. 533–539, 1988.
- THORVALDSDÓTTIR, H.; ROBINSON, J. T.; MESIROV, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. **Briefings in Bioinformatics**, v. 14, n. 2, p. 178–192, 2013.
- TRAPNELL, C.; PACTER, L.; SALZBERG, S. L. TopHat: discovering splice junctions with RNA-Seq. **Bioinformatics**, v. 25, n. 9, p. 1105–1111, 2009.
- VERJOVSK-ALMEIDA, S. R. *et al.* Transcriptome analysis of the acoelomate human parasite *Schistosoma mansoni*. **Nature Genetics**, v. 35, n. 2, p. 148–157, 2003.
- WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics. **Nature Reviews Genetics**, v. 10, n. 1, p. 57–63, 2009.
- WATERMAN, M. S.; EGGERT, M.; LANDER, E. Parametric sequence comparisons. **Proceedings of the National Academy of Sciences of the United States of America**, v. 89, n. 13, p. 6090–6093, 1992.
- WHO - WORLD HEALTH ORGANIZATION. **Schistosomiasis. Fact sheet 115**. Disponível em: <<http://www.who.int/mediacentre/factsheets/fs115/en/>>.
- WILHELM, B. T.; LANDRY, J. RNA-Seq — quantitative measurement of expression through massively parallel. **Methods**, v. 48, n. 3, p. 249–257, 2009.
- YAMADA, T. *et al.* iPath2.0: interactive pathway explorer. **Nucleic Acids Research**, v. 39, p. W412–415, 2011.

7. ANEXO 1 : LANDSCAPE OF THE SPLICED LEADER TRANS-SPLICING
MECHANISM IN SCHISTOSOMA MANSONI

De: Genome Biology Editorial editorial@genomebiology.com 
Assunto: 1224997222134377 Landscape of the spliced leader trans-splicing mechanism in Schistosoma mansoni
Data: 1 de julho de 2014 04:10
Para: Marina M Mourão mourao.marina@gmail.com
Cc: Mariana Boroni marianaboroni@gmail.com, Michael Sammeth micha@sammeth.net, Mainá Bitar duonlumo@gmail.co,
Andréa M Macedo andrea@icb.ufmg.br, Carlos R Machado crmachad@icb.ufmg.br, Marina M Mourão
mourao.marina@gmail.com, Glória R Franco gfrancoufmg@gmail.com

Article title: Landscape of the spliced leader trans-splicing mechanism in Schistosoma mansoni
MS ID : 1224997222134377
Authors : Mariana Boroni, Michael Sammeth, Mainá Bitar, Andréa M Macedo, Carlos R Machado, Marina M Mourão and Glória R Franco
Journal : Genome Biology

Dear Dr Mourão

Thank you for submitting your article. This acknowledgement and any queries below are for the submitting author. This e-mail has also been copied to each author on the paper. Please bear in mind that all queries regarding the paper should be made through the submitting author.

A pdf file has been generated from your submitted manuscript and figures. We would be most grateful if you could check this file and let us know if any aspect is missing or incorrect. Any additional files you uploaded will also be sent in their original format for review.

http://genomebiology.com/imedia/1224997222134377_article.pdf (2834K)

For your records, please find below link(s) to the correspondence you uploaded with this submission. Please note there may be a short delay in creating this file.

http://genomebiology.com/imedia/9027270321343786_comment.pdf

If deemed suitable, we will assign peer reviewers as soon as possible, and will aim to contact you with an initial decision on the manuscript within four weeks.

In the meantime, if you have any queries about the manuscript you may contact us on editorial@genomebiology.com. We would also welcome feedback about the online submission process.

Best wishes,

The Genome Biology Editorial Team

Tel: +44 (0) 20 3192 2000
Facsimile: +44 (0) 20 3192 2011
e-mail: editorial@genomebiology.com
Web: <http://genomebiology.com/>

1 **Landscape of the spliced leader trans-splicing mechanism in *Schistosoma***

2 ***mansoni***

3

4 **Mariana Boroni¹, Michael Sammeth², Mainá Bitar¹, Andréa M. Macedo¹, Carlos**

5 **R. Machado¹, Marina M. Mourão³ and Glória R. Franco^{1§}**

6

7 ¹ Laboratório de Genética Bioquímica, Departamento de Bioquímica e

8 Imunologia, Universidade Federal de Minas Gerais, Belo Horizonte, Minas

9 Gerais, Brazil;

10 ² Laboratório Nacional de Computação Científica, Petrópolis, Rio de Janeiro,

11 Brazil;

12 ³ Grupo de Genômica e Biologia Computacional, Centro de Pesquisas René

13 Rachou, Fundação Oswaldo Cruz, Belo Horizonte, Minas Gerais, Brazil;

14

15 [§]Corresponding author

16

17 Email addresses:

18 MBo: marianaboroni@gmail.com

19 MS: micha@sammeth.net

20 MBi: duonlumo@gmail.com

21 MMM: mourao.marina@gmail.com

22 AMM: andrea@icb.ufmg.br

23 CRM: crmachad@icb.ufmg.br

24 GRF: gfrancoufmg@gmail.com

25 **Abstract**

26 **Background**

27 Spliced leader dependent trans-splicing (SLTS) is a pre-RNA maturation
28 process that occurs in diverse organisms, including the trematode parasite
29 *Schistosoma mansoni*. Several functional roles have been assigned to SLTS, but for *S.*
30 *mansoni* specifically, its importance as a post-transcriptional regulatory mechanism
31 has not been determined.

32 **Results**

33 Using a focused strategy to capture transcripts that harbor the SL sequence and
34 retrieving SL-containing reads from RNA-Seq experiments, we generated a broad
35 dataset of SLTS genes. From the results presented herein, we estimate that 56% of the
36 protein-coding genes expressed in the *S. mansoni* cercariae stage undergo SLTS. The
37 expression levels of the identified trans-spliced genes span several orders of
38 magnitude, which indicates that the mechanism is not particularly biased to a specific
39 set of genes based on their expression patterns. In addition, SL-containing reads
40 mapped to genes over a wide spectrum of functional classes emphasizes the notion
41 that the SLTS does not act on a particular gene category. Our analysis shows extensive
42 heterogeneity of SL acceptor sites along transcripts and intron attributes that can
43 distinguish them from cis-spliced introns.

44 **Conclusions**

45 Our results show that, in this parasite, SLTS does not selectively affect specific
46 gene categories and that gene expression is not the only determinant of SLTS
47 efficiency. Particularly for introns where SLTS competes with splicing in cis for the
48 acceptor substrate, we identified discriminating features that might impact the type of

49 splicing used and serve as a basis for future studies on the trans-splicing mechanism.

50

51 **Keywords**

52 RNA-Seq, Spliced Leader Trans-Splicing, Alternative Trans-Splicing,

53 *Schistosoma mansoni*

54

55 **Background**

56 In contrast to conventional splicing in cis, trans-splicing joins exons of two
57 different primary transcripts from *a priori* unrelated genomic origins. One prominent
58 variant of trans-splicing, spliced leader trans-splicing (SLTS), has been reported in
59 several eukaryotic organisms of the rotifera, chordata, cnidaria, dinoflagellata,
60 euglenozoa, nematoda and platyhelminthes phyla [1]. Both cis- and trans-splicing
61 processes are catalyzed by the spliceosome, which is an aggregate of molecules
62 composed of small nuclear RNAs (snRNAs) that are associated with proteins [2].
63 Compared with cis-splicing, in SLTS, the spliced leader (SL) RNA substitutes for the
64 snRNA U1 function in the spliceosome [3, 4], which, in cis-splicing, recognizes a
65 special RNA signature, the splice donor site. SLTS is oriented and joins the SL exon
66 (as the 5' splice donor counterpart) to the 3' splice acceptor sequence in the target
67 RNA molecule. Therefore, mRNA produced by SLTS begins with the SL sequence
68 and lacks the original 5' UTR, which is referred to as the outtron (reviewed in [1, 5,
69 6]). To form the intermediate Y-branch and lariat products, which are necessary for
70 SLTS and cis-splicing, respectively, both mechanisms require a branch point that is
71 located upstream of the acceptor site and is accompanied by an accessory
72 polypyrimidine tract to attract the splicing machinery.

73 In recent years, several functions have been described for SLTS, such as
74 increased transcript stability by adding a specialized cap to mRNA, which is required
75 for transport and translation (reviewed in [5–7]); increased flexibility of the 5' end
76 transcript structures [8]; and regulation of gene expression by substituting start codons
77 during outtron removal, which takes advantage of the alternative trans-splice sites [9].
78 Finally, the first and best characterized function of the SLTS, as described in
79 trypanosomatids [10], is the resolution of polycistronic transcripts into monocistronic
80 molecules for independent translation [11–14]. The same function was reported in
81 different organisms of other taxa; the best documented case is the nematode
82 *Caenorhabditis elegans* (reviewed in [15, 16]). Although polycistronic units may exist
83 in most organisms in which the SLTS mechanism is present, in many cases, most
84 trans-spliced mRNAs are not from polycistronic transcripts (reviewed in [5]).

85 Traditional reports on the SLTS mechanism suggest that it always occurs at the
86 5'-terminal exon of a gene based on the premise that the presence of cis-splicing
87 signals, such as upstream donor sequences, can disturb the SLTS process. However,
88 recently published articles have reported SLTS at sites distal to the 5'-end of
89 transcripts [16, 17], which suggests that the splicing machinery can interpret long
90 introns as outtrons, because the likelihood of SLTS in an intron increases with its size
91 [16]. Reporter gene assays in *T. brucei* indicate specific attributes of the
92 polypyrimidine tract and branch point associated with SLTS [18], and in *Ciona*
93 *intestinalis*, substantial proportions of the SLTS targets are described as cryptic sites
94 near the respective main trans-spliced acceptor [19].

95 The flatworm *S. mansoni* is the predominant etiologic agent of the neglected
96 tropical disease schistosomiasis in Africa and Brazil. The disease is classified as the
97 second most socioeconomically devastating in the world (after malaria), and it affects

198 over 243 million people in 78 countries [20, 21]. SLTS in *S. mansoni* uses a 36 nt SL
199 exon derived from non-polyadenylated 90-nt-long SL RNA [22] and is not clearly
100 associated with a specific gene category, subcellular localization or life-cycle stage
101 [17, 23]. Particular case studies demonstrate alternative trans-splice sites in the 3-
102 hydroxy-3-methylglutaryl CoA reductase gene transcript and, more recently, in the
103 ubiquinol-cytochrome C reductase complex ubiquinol binding protein (UbcCRBP)
104 gene transcript [17, 23]. Davis and Hodgson [24] introduced the UbcCRBP/enolase
105 gene model as a proof-of-principle for the presence of polycistrons that undergo trans-
106 splicing in *S. mansoni*, which Protasio and colleagues [25] extended to 46 potential
107 polycistronic units with intergenic distances up to 200 bp based on analyses of whole-
108 transcriptome RNA-Seq experiments. These studies suggest a limited scope for the
109 SLTS mechanism in *S. mansoni* and seem to underestimate the true impact of this
110 phenomenon on schistosome biology considering its complex life cycle, which
111 requires intricate transcription and post-transcription gene regulation mechanisms
112 [26].

113 To better characterize the SLTS mechanism in *S. mansoni* and account for its
114 importance as a regulatory mechanism in this parasite, we performed a comprehensive
115 study by combining two different strategies, mainly focusing on transcripts from
116 cercariae, adult worms and schistosomulae. To the best of our knowledge, our SLTS
117 data are the most representative for *S. mansoni* and were extensively analyzed, which
118 supports the conclusions on the role of the SLTS mechanism in this organism. Such
119 observations may be further extended to organisms from other taxa for which the
120 SLTS mechanism has been characterized and may contribute to constructing a more
121 thorough understanding of SLTS evolution and function in eukaryotes.

122

123 **Results and Discussion**

124 **Directed capture of SL-containing transcripts was used to provide a**
125 **comprehensive landscape of SLTS in *S. mansoni***

126 To explore the functions of the SLTS mechanism in *S. mansoni*, we used two
127 approaches. (i) From the publicly available RNA-Seq experiments, which include
128 approximately 250 million reads from different *S. mansoni* life-cycle stages, we
129 retrieved 91,188 reads containing the 36-nucleotide SL sequence, herein referred to as
130 the RNA-Seq Filtered dataset (Additional file 1: Table S1). (ii) We produced two
131 biological replicates of a SL Trapping experiment performed using RNA from the
132 cercariae stage and targeting transcripts containing the SL sequence (herein referred to
133 as “Trapping 1” and “Trapping 2” datasets with 11,520,178 and 30,332,894 reads,
134 respectively). From these data, 34,929,746 reads (83%) were uniquely mapped to the
135 *S. mansoni* reference genome (5th version).

136 We observed a marked reproducibility between the SL Trapping datasets: 70%
137 (6,586) of the 8,506 genes detected in Trapping 1 and 7,495 genes from Trapping 2
138 overlap (Additional file 1: Fig. S1A). Additionally, the SLTS genes in the two datasets
139 exhibit a strong quantitative correlation (0.86 Pearson correlation coefficient - PCC,
140 Additional file 1: Fig. S1B). We expected the SLTS signals from filtering the RNA-
141 Seq data to be much weaker (<0.04% of the total RNA-Seq reads contain the SL
142 sequence), and accordingly, we only retrieved 2,459 SL-containing transcripts from
143 the whole-transcriptome RNA-Seq datasets. However, despite their distinct origins,
144 the Trapping datasets combined include 95% of the trans-spliced transcripts contained
145 in the RNA-Seq Filtered dataset (Fig. 1A). The absence of the remaining 5% (123) of
146 transcripts (Fig. 1B) in the SL Trapping datasets can be explained by the presence of
147 RNA from 2 other stages (schistosomulae and adult worms) in the RNA-Seq Filtered

148 dataset, whereas the Trapping datasets only contain cercariae RNA. Merely 22
149 transcripts exclusively from the RNA-Seq Filtered dataset are expressed in cercariae
150 (Fig. 1B).

151 For further analyses, we curated the data and considered only the loci
152 identified in both Trapping replicates with at least 10 counts. From the total number of
153 9,415 loci, 5,443 remained after the curation (Fig. 1C), which represents
154 approximately 58% of the loci. Despite the heterogeneous origins of the experimental
155 material (cercariae versus cercariae, schistosomulae and adult mixed genders), the
156 final Trapping datasets cover approximately 85% of the RNA-Seq Filtered dataset
157 (Fig. 1C). An interesting observation is that the 2,081 genes (Fig. 1C) shared by both
158 the Trapping and RNA-Seq Filtered datasets are frequently trans-spliced genes (507
159 read counts on average among the genes shared by the Trapping and Filtered datasets
160 versus a 44 read counts on average for the genes exclusively detected in the Trapping
161 datasets), which reinforces the importance of the SL Trapping strategy as a method for
162 capturing rare trans-spliced transcripts. Quantitatively, the Trapping and Filtered
163 datasets exhibit a moderate correlation ($PCC=0.5$, Fig. 1D), and in general, the genes
164 identified in both datasets share the same SL insertion sites (Fig. 2A and B). For
165 example, we analyzed the SLTS signals through the [GeneDB:Smp_034190.1] gene
166 (Fig 2A), which shows consistently strong trans-spliced signals at the 5'- and 3'-
167 terminal exons in all datasets. While the RNA-Seq Filtered dataset exhibits a strong
168 signal for exons 2 and 3, the signal was not reproduced in the SL Trapping dataset,
169 most likely due to the limited types of life-cycle stages assessed. In contrast, the SL
170 Trapping datasets exhibit a weak trans-splicing signal in exon 5, which was not
171 observed in the RNA-Seq Filtered dataset, most likely due to its limited sensitivity.
172 Additionally, all of the datasets exhibit a weak trans-splicing signal in exon 4. Taken

173 together, our results demonstrate that SL Trapping is a powerful and reproducible
174 technique that detects signals similar to the signals produced using classical RNA-Seq
175 approaches, but at an incomparably higher resolution. This approach can also detect
176 additional, important, and low-frequency SLTS events.

177 In assessing the frequency of the SLTS mechanism, we identified 5,443 SL-
178 containing transcripts in cercariae, representing 43% of the 12,659 genes expressed in
179 cercariae. The proportion increases to 56% when protein-coding genes are considered
180 (4,467 trans-spliced genes from a total of 7,987 protein-coding genes expressed in
181 cercariae). These numbers are substantially higher than previous estimates that
182 indicate approximately 10% trans-spliced genes in schistosomes by Davis and
183 colleagues [23], who focused on a small subset of genes, and by Protasio and
184 colleagues [25], who exclusively employed a standard RNA-Seq approach. Our
185 observations are consistent with studies in the nematode *C. elegans* and in the
186 urochordata *C. intestinalis*, wherein 70% [16] and 58% [19] of all transcripts undergo
187 SLTS, respectively. The catalogue of *S. mansoni* SLTS events presented herein is
188 expected to be further extended as we focused predominantly on the cercariae stage
189 and the expression of trans-spliced genes is known to vary according to changes in
190 environmental conditions [17] and along the parasite life-cycle [26].

191 Among the trans-spliced genes identified, 749 are putatively novel genes (20%
192 of which are shared by both datasets) because they do not match the currently
193 annotated gene models. Additionally, 17 of the trans-spliced genes are ncRNAs,
194 primarily rRNAs. Taken together, these results demonstrate that in addition to
195 providing a comprehensive understanding of the SLTS mechanism in the parasite, the
196 SL Trapping approach serves as a complementary tool for transcriptome annotation.

197 **Features of the SLTS mechanism in *S. mansoni***

198 In dinoflagellates, it has been shown that certain genes acquired multiple in
199 tandem SL sequences that may have arisen from the reverse transcription of
200 previously trans-spliced transcripts and the subsequent reinsertion of the
201 corresponding cDNA in the genome [27]. By scanning the *S. mansoni* genome for the
202 SL sequence, in addition to the five SL RNA genes annotated in the 5th version of the
203 *S. mansoni* genome and seven non-annotated putative SL RNA genes, we detected 32
204 additional loci in the genome that include either full or partial SL sequences. Of these
205 loci, 11 matched previously annotated genes, and after we re-annotated the *S. mansoni*
206 transcriptome, we associated 6 genome-encoded SL sequences to transposable
207 elements (Additional file 1: Table S2). Presumably, these transposons correspond to
208 trans-spliced retro-transcribed elements because 40% of the *S. mansoni* genome is
209 composed of repetitive regions, such as retrotransposons [28], and these elements are
210 particularly active at the cercariae and schistosomulae stages [29]. Seventeen genes
211 that contain the spliced-leader portion on their genomic sequences are not currently
212 annotated as *S. mansoni* genes, which indicates the existence of additional gene loci
213 that contain the SL sequence. Through subsequent analyses of the SLTS mechanism,
214 we discarded the aforementioned genes that contain genome-encoded SL sequences.

215 Next, to investigate whether certain *S. mansoni* trans-spliced transcripts
216 originated from the processing of polycistronic units, we assessed the transcriptome
217 annotation to search for putative polycistron components that may be resolved
218 through SLTS. Of the 139 proposed polycistrons (gene clusters with intergenic
219 distances up to 200 bp in the *S. mansoni* transcriptome), we identified 65 *bona fide*
220 dicistronic and one tricistronic units (including 34 of the 46 previously identified
221 polycistrons, Additional file 2) with the SL sequence as evidence that they were trans-
222 spliced [25]. Putative polycistronic units are predominantly located at chr1 (25.75%

223 of all polycistrons), chrW (22.72%) and chr4 (15.15%), in contrast to the remaining
224 trans-spliced genes, which reflect the overall chromosomal distribution of all genes
225 expressed in cercariae and the chromosome sizes. Although the 133 polycistronic unit
226 gene components account for approximately 2% of the 5,821 trans-spliced genes, our
227 observations should be considered a lower boundary on the actual number of
228 polycistronic clusters because polycistron gene components with incomplete 5' or 3'
229 transcript annotations that are separated by unusually large spacers or not expressed in
230 our dataset would have escaped our observations. Moreover, our threshold is
231 supported by the average distance between the genes, which is 99 bp (min= 1 bp,
232 max= 200 bp). Most polycistron gene components are conserved; 23% are
233 monoexonic transcripts, 131 are protein-coding genes; and two are rRNAs. Figure 2B
234 shows an example of a putative polycistron [GeneDB:Smp_173910.1 and
235 GeneDB:Smp_094470.1], which was identified as a dicistron based on the intensity
236 and distribution of trans-spliced signals at the 5'-end of the downstream gene
237 [GeneDB:Smp_094470.1].

238 To further analyze the characteristics that discriminate genes subject to SLTS
239 from all protein-coding genes expressed in the cercariae stage, we analyzed the gene
240 function of 5,057 protein-coding, trans-spliced genes and compared the functions to
241 the 7,987 protein-coding genes expressed in cercariae. Employing a clustering
242 algorithm to a customized set of protein functional categories, we found that in
243 general, the different functional categories for the trans-spliced genes do not differ
244 significantly from all protein-coding genes expressed in cercariae, except in the
245 “Unknown” and the “Transcription factor” functional categories (p-value < 0.024 and
246 p-value < 0.0003, respectively, Chi-square test; Fig. 3). Additionally, the trans-spliced
247 genes did not exhibit significant enrichment in a specific KEGG (Kyoto Encyclopedia

248 of Genes and Genomes) pathway [30], which demonstrates that SLTS-derived gene
249 products act on essential pathways (Additional file 1: Fig. S2). This result is expected
250 because SLTS in *S. mansoni* is not likely associated with a particular tissue,
251 developmental phase or gender or with specific genes or gene families [17, 23, 31].
252 Our findings are consistent with recent transcriptomic studies in *Eutreptiella* sp.,
253 wherein the authors reported that the trans-spliced genes are functionally diverse and
254 thus characterized the SLTS as a ubiquitous mechanism in this euglenoid algae [32].

255 **Attributes of trans-spliced genes**

256 In *C. elegans*, gene expression positively correlates with trans-splicing
257 frequency [16]. Accordingly, we assessed the correlation between gene expression and
258 SLTS mechanism efficacy and estimated the expression levels of all trans-spliced
259 genes identified in the *S. mansoni* cercariae RNA-Seq experiment. We observed a
260 trend of higher expression of trans-spliced genes since the median normalized
261 expression value of the trans-spliced genes is approximately 5-fold greater than the
262 expression of non-trans-spliced genes (32.3 versus 161.2, p-value < 2.2×10^{-16} in a
263 two-sample Kolmogorov-Smirnov Test, Fig. 4A). However, Figure 4B shows highly
264 expressed genes that are rarely trans-spliced, and vice-versa (lower-right and upper-
265 left quadrants, respectively). Our analysis reveals log-normally distributed gene
266 expression levels for the overall transcriptome (pink histogram in Fig. 4C), but the
267 frequencies of transcripts undergoing trans-splicing deviate from a log-normal
268 distribution (blue histogram in Fig. 4C), which confirms our observations that certain
269 genes present particularly high/low SLTS frequencies. In summary, although the
270 frequency for trans-splicing events increases as expected with a rise in gene
271 expression, our results show that the efficiency of the SLTS process is not solely
272 determined by the transcription rate of a gene, which suggests that particular

273 mechanisms facilitate or hinder trans-splicing.

274 To further investigate the connection between gene expression and SLTS
275 frequency, we investigated the features that discriminate genes with low expression
276 and an associated high number of SLTS events (trans-splicing-driven – TSD) from
277 genes with high expression and a low number of SLTS events (gene-expression-
278 driven – GED). Thus, we selected 2,400 outliers (1,200 outliers from the upper-left
279 quadrant and 1,200 outliers from the lower-right quadrant of Fig. 4B). As observed,
280 there are no significant preferences for a certain chromosomal location in either GED
281 or TSD genes compared with the overall gene expression by chromosomes (Fig. 5A).
282 Based on an analysis of the entire *S. mansoni* transcriptome, most genes are expressed
283 as a single isoform or a few isoforms with alternative cis-splicing, and only a few
284 genes originate multiple transcript isoforms. The same phenomenon was observed for
285 GED genes (Fig. 5B). In contrast, we observed significantly less alternative cis-
286 splicing in TSD genes (p-value= 0.08, two-sample Kolmogorov-Smirnov Test), most
287 of which exhibited no more than three alternative isoforms per gene. Similarly, most
288 *S. mansoni* transcripts throughout the transcriptome were derived from genes with one
289 or few exons (Fig. 5C). GED and TSD include fewer single-exon genes than the entire
290 transcriptome, and these two categories exhibit more genes with a moderate to high
291 number of exons (p-value= 0.0005 for both, GED and TSD genes). Therefore, our
292 results show that both GED and TSD genes are highly complex considering the
293 number of exons that comprise a gene. However, while the higher number of exons in
294 GED genes correlates to more alternative splice isoforms, TSD genes include fewer
295 alternative transcripts compared to the number of exons, which suggests a higher
296 SLTS frequency in the absence of strong signals to drive alternative cis-splicing.

297 **Comparisons between SLTS and cis-splicing**

298 In contrast to the classical notion that the first exon is the only substrate for
299 SLTS [15], we observed many SLTS sites located far downstream of the annotated
300 transcription start sites. We observed that merely 18% of all SLTS events occur at the
301 first exon acceptor site and that the other 82% of events occur at other exons
302 (Additional file 1: Fig.S3). A quantitative analysis confirmed that SLTS events in the
303 first exons are significantly more frequent, with a mean reads count of 2,295 SLTS
304 events in the first exon versus a mean count of 1,035 events for other exons in the SL
305 Trapping dataset. Interestingly, the SLTS acceptor sites in introns exhibit greater
306 sequence conservation compared with SLTS acceptor sites in the 5'-end transcript
307 region (outtrons) (Additional file 1: Fig.S4A). Remarkably, the frequency of canonical
308 AG dinucleotides in the outtron/exon border was lower in SLTS sites, and another
309 dinucleotide (TG) was also observed at this position (Additional file 1: Fig.S4B). Our
310 observations are consistent with previous reports [15], which suggest that trans-
311 splicing is less constrained than cis-splicing in internal exons and more permissive at
312 acceptor sites around the transcription start sites, most likely due to the absence of
313 competition with splice donor substrates in cis. Previous reports on *T. brucei* [9] show
314 that 20% of the minor trans-splice acceptor sites contained a dinucleotide other than
315 AG. The dinucleotide GG occurred in 7% of these sites, while TG, AA, GA and AC
316 were observed in 2% of the minor splice sites.

317 Upon further analysis, we also observed an abundance of SLTS events at
318 different introns of transcripts, which is highly consistent with the annotated acceptor
319 sites and reflects canonical splice site signals reasonably well. Grouping SLTS events
320 by exon position within the transcripts revealed a preference for SLTS at the terminal
321 exons located either on the 5' or 3' end of the transcript (left panel in Fig. 6A). As
322 expected, we did not observe enhanced RNA-Seq coverage at the 5'- or 3'-end exons

323 (right panel in Fig. 6A), reinforcing that these SLTS observations are not particularly
324 affected by artifacts with unequal expression. We further found that the terminal
325 exons are longer (Fig. 6B) and the trans-spliced introns tend to be to some extent
326 longer (Fig. 6C) than their non-trans-spliced counterparts. These observations support
327 the existence of particular SLTS mechanistic features resulting from a special exon-
328 intron architecture in the trans-spliced transcripts.

329 The polypyrimidine tract is one of the most important sequence elements in
330 splicing [33], and in trypanosomes, it is related to splicing efficiency [34]. In a study
331 on tubulin polycistronic pre-mRNA in trypanosomes [35], block substitutions
332 indicated that the polypyrimidine tract at the 3' acceptor site is crucial for correct
333 trans-splicing. Furthermore, the length of the polypyrimidine tract is important for *in*
334 *vitro* splicing reactions [36], and Coolidge and collaborators [33] proposed that
335 polypyrimidine tracts comprising 11 continuous uridines are stronger, regardless of
336 their position between the branch point and the AG dinucleotide at the splice site.
337 Conversely, the authors also showed that when the number of continuous uridines is
338 decreased to values as low as five, they must be located immediately adjacent to the
339 AG for higher efficiency. Thus, we observed uridines at a higher frequency in intronic
340 regions proximal to the trans-splice acceptor sites compared with cis-splice acceptor
341 sites in trans-spliced transcripts or cis-splice acceptor sites in cis-spliced transcripts.
342 The uridine content is even greater proximal to the AG splice site (Fig. 7).

343 To investigate properties that discriminate trans-spliced from non-trans-spliced
344 introns, we employed the classification system generated by Davis and colleagues
345 [23] for cis-spliced introns (CS), cis-spliced introns in trans-spliced transcripts (CTS)
346 and trans-spliced introns (TS). By further distinguishing introns based on their
347 annotated transcript structure into single, first, intermediate, and last introns, we

348 derived the following rules for competition between splicing types. TS introns are
349 typically longer (Fig. 8A), exhibit increased acceptor splice site scores (Fig. 8B) and
350 include polypyrimidine tracts that are longer (Fig. 8C). Our observations support the
351 hypothesis that long introns can be interpreted as outtrons by the splicing machinery
352 [15, 37] and confirm similar studies on *T. brucei* [18], which report that cis-spliced
353 introns are shorter with less pronounced polypyrimidine tracts. Taken together, our
354 results show that SLTS events at acceptor sites other than outtrons are not a random
355 observation but exhibit particular characteristics with implications for the splicing
356 mechanism.

357

358 **Conclusions**

359 Here, we present the first thorough survey that employs the SL Trapping
360 protocol [9] to study SLTS events in *S. mansoni*. Our investigation demonstrates that
361 the SL Trapping technique is reproducible and confirms previous RNA-Seq results in
362 two biological replicates. Additionally, this approach extends the transcriptome
363 resolution limit to an unprecedented depth and facilitates the exploration of rare SLTS
364 events that were previously undetected in standard RNA-Seq experiments. Generating
365 the SL Trapping dataset also facilitated the detection of putative new genes, which
366 highlights the relevance of this approach for enhancing genome annotations. We also
367 identified 32 genes with the spliced leader sequence as a constituent of the gene body
368 and extended the 46 known polycistrons [25] for a total of 65 dicistronic units and one
369 tricistronic unit. As expected, we did not observe a preference for a functional
370 category or chromosomal origin in the trans-spliced genes, which suggests that the
371 SLTS is a ubiquitous mechanism in the parasite.

372 The results herein show that SLTS detection is easier in genes with higher

373 expression rates, but we also observed certain low-expressed genes that are frequently
374 trans-spliced. We also report that genes exhibiting comparatively higher levels of
375 alternative cis-splice isoforms are reduced in their trans-splicing activity, whereas
376 long genes with comparatively low alternative cis-splicing activity are more subject to
377 SLTS, which suggests the higher efficiency of the trans-splicing machinery in the
378 absence of complex splicing structures that compete in cis. We showed that a
379 substantial proportion of SLTS events occur in introns, and our investigation revealed
380 that increased intron length may enhance their predisposition for trans-splicing.
381 Furthermore, our data demonstrate that trans-spliced introns have stronger acceptor
382 sites as well as longer and richer polypyrimidine tracts. Overall, our results
383 demonstrate that several splicing features play a key role in committing an intron to
384 cis- or trans-splicing.

385

386 **Methods**

387 **Dataset compilation of RNA-Seq reads from SRA**

388 Publicly available RNA-Seq datasets from different stages of the *S. mansoni*
389 life cycle were included in this study. The dataset characteristics are listed in
390 Additional file 1: Table S1. Twelve raw files that were generated using Illumina
391 platform were downloaded from the SRA repository
392 (<http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi>) [38] and converted into the fastq
393 format. The trans-spliced sequences were filtered according to whether the 36 nt
394 sequence of the *S. mansoni* spliced leader was present using the program Illumina
395 reads adapter screening utilities / sortPairedReads (with -Q 33 -f sl_smansoni.fasta -s
396 19 parameters). Next, the SL sequence was trimmed from the reads using the program
397 Illumina reads adapter screening utilities / trimReads (with -f sl_smansoni.fasta -s 19 -

398 m 0 -q 0 -Q 33 parameters); these reads were compiled in a file referred to as the
399 RNA-Seq Filtered dataset.

400 **RNA isolation, library preparation and sequencing**

401 The *S. mansoni* (LE strain) life cycle was maintained at the Centro de
402 Pesquisas René Rachou (CPqRR), Fundação Oswaldo Cruz, Brazil. The total RNA
403 from *S. mansoni* (cercariae stage harvested from the intermediate host *Biomphalaria*
404 *glabrata* snails- Barreiro de Cima strain) was isolated using the TRIzol® Reagent
405 (Invitrogen) and RNeasy Kit (Qiagen) and then treated with Ambion® RNase-free
406 DNase I (Invitrogen). The RNA samples were quantified using the Nanodrop ND-
407 1000 (Thermo Scientific), and all samples showed an A260/A280 ratio higher than
408 1.8. In addition, the RNA integrity was verified using the Agilent 2100 Bioanalyzer.
409 The libraries were constructed and sequenced by the FASTERIS facility
410 (<https://www.fasteris.com>) following the protocol described by Nilsson and
411 colleagues [9] with the following modifications. The primer specific to the *S. mansoni*
412 SL sequence ([BIOT]5'-
413 AATGATACGGCGACCACCGAGATCTACACTCTTGTGATTTGTTGCATG -3')
414 was used to produce the second strand cDNA, and sequencing using the Illumina
415 HiSeq 2000 platform was performed with a specific sequencing primer (5'-
416 GAGATCTACACTCTTGTGATTTGTTGCATG -3'). Two libraries from independent
417 biological replicates were each sequenced using a 1x100 bp run and are referred to as
418 the SL Trapping datasets.

419 **Data processing and mapping of the reads to the *S. mansoni* genome**

420 The statistics and quality analyses of the reads were generated using the
421 FastQC software version 0.10.1

422 (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) by Barbhaham
423 Bioinformatics. To identify the SLTS targets, RNA-Seq Filtered reads trimmed by the
424 SL sequence and SL Trapping reads were aligned to the genomic sequence of *S.*
425 *mansoni* (5th version), which was downloaded from the GeneDB site
426 (<ftp://ftp.sanger.ac.uk/pub/pathogens/Schistosoma/mansoni/>), using the TopHat2
427 mapper [39] version 2.0.9 (with `-i 10 -I 30000 -p 8 --library-type fr-unstranded`
428 parameters). Reads that were mapped to the 5'-end of the annotated exons and novel
429 exons or to 5'-exon flanking regions were considered. Subsequently, all SRA RNA-
430 Seq reads from the cercariae stage were aligned to the genome using the same
431 approach before filtering for the SL sequence to estimate the expression levels of the
432 cercariae genes based on the GeneDB transcriptome annotation (5th version). The
433 mapped reads were filtered by quality with the threshold of Phred [40, 41] score 20
434 and unique mapping using Samtools version 0.1.19 [42] (with `-bhq 20 -F 0x100`
435 parameters).

436 **Gene expression and SLTS counts normalization**

437 The raw counts for each gene in the *S. mansoni* GeneDB gene annotation (5th
438 version) were obtained for all datasets using the HTSeq Python package version
439 0.5.3p3 [43] and the “`--stranded=yes`” option, as well as the “`--mode=intersection-`
440 `strict`” option for exon-intersection counting. The exon counts were produced using a
441 similar approach but with a modified gene annotation file using the Python script
442 `dexseq_prepare_annotation.py` from the DEXSeq R package version 1.10.4 [44]. As a
443 cutoff, only genes/exons with ≥ 10 counts and represented in both Trapping datasets
444 were considered for further analyses. Normalization of the gene/exon counts and their
445 expression values was performed using the DESeq2 R package version 1.2.5 [43],
446 including a matrix with the gene lengths for normalization.

447 **Functional annotation**

448 As most of the trans-spliced genes were annotated as “hypothetical proteins”,
449 we re-annotated the predicted mRNA sequences of *S. mansoni* gene models by
450 employing an established pipeline described in [45]. In brief, we performed blastx,
451 blastn, or rpsblast [46] searches for the coding DNA sequences against several
452 databases, including Swissprot [47], Gene Ontology [48], KOG [49], pfam [50], and
453 SMART [51], and a subset of the non-redundant protein database from NCBI that
454 contained vertebrate proteins. Further manual annotation was performed as required.
455 The results were used to assign a functional classification to the protein sequences
456 through the Classifier program (developed by Dr. José Marcos Chaves Ribeiro,
457 NIAID/NIH), which is based on a vocabulary of nearly 250 keywords found in
458 matches to all of the databases used, as well as their e-values, to produce nearly 30
459 functional categories. We performed a Chi-squared test (p-value < 0.05) to identify
460 significant enrichment of a given transcript class that undergoes trans-splicing
461 compared with the total *S. mansoni* transcriptome. The UniProtKB/TrEMBL
462 identifiers of *S. mansoni* protein gene models were associated with their respective
463 counts in the SL Trapping and RNA-Seq Filtered datasets. The program iPath [52]
464 was used to generate the *S. mansoni* metabolic pathways using the KEGG database
465 [30], and the pathways containing trans-spliced genes were represented by red lines;
466 with thickness reflecting the number of normalized read counts per gene.

467 **Identification of polycistronic units**

468 Gene groups located in the same chromosome and DNA strand with intergenic
469 distances of up to 200 bp (i.e., the distance from the 3' end of the upstream to the 5'
470 end of the next downstream annotated gene) were detected in the *S. mansoni* gene
471 model annotation using our customized script. We suggest the existence of *bona fide*

472 polycistronic units through assessing the occurrence of SLTS in their upstream genes.
473 Alleged polycistronic clusters were required to exhibit SLTS in at least one gene
474 downstream of the first gene in the polycistronic unit. Certain predictions were
475 visually verified using IGV 2.1 (Broad Institute) [53].

476 **Identification of gene loci containing the SL sequence**

477 We performed a BLAT [54] search (with `-t=dna -q=dna -minIdentity=90 -`
478 `out=blast8 -maxGap=1 -fine` parameters) on the *S. mansoni* genome for the 36 nt SL
479 sequence. SL genes encoding the SL RNA were identified and removed from the
480 results. After removing the SL genes, the genome was intersected with the gene model
481 annotation to identify genes containing embedded SL sequences.

482 **Identification of SLTS signals**

483 Sequences surrounding the splice sites where the SL sequence was inserted
484 were retrieved from genes using BedTools version 2.17.0 [55] and visualized as
485 sequence logos using the WebLogo 3.0 software [56]. To investigate feature changes
486 between the cis- and trans-spliced sites, we used a conservative approach that focused
487 on a set of trans-spliced genes with at least one trans-spliced exon at >10 counts to an
488 internal exon and positive values in both SL Trapping libraries. In these genes, the
489 intron with the highest count at its 3'-end was identified as the main target for trans-
490 splicing; all other introns for the same gene were considered cis-spliced introns in the
491 trans-spliced transcripts. A balanced set of genes not subject to trans-splicing but with
492 similar expression levels yielded a comparable number of introns spliced in cis. Using
493 these intron sets, we then employed the GeneID [57] splice site models to score
494 potential donor and acceptor sites, which represent the first order Markov chains
495 trained on annotated splice sites. To assess the branch point features, we applied

496 different models integrated into the Support Vector Machine SVM-BPfinder [58] to
497 the annotated intronic sequences.

498 **Additional analyses**

499 The annotation file generated was used to obtain the biotype of each gene and
500 the corresponding gene, exon and intron lengths. All data processing and visualization
501 was performed using R version 3.0.2 [55]. Certain plots were constructed using the
502 ggplot2 package for R, version 0.9.3.1.

503

504 **Competing interests**

505 The authors declare that they have no competing financial interests.

506

507 **Authors' contributions**

508 MBo prepared the RNA-Seq libraries, designed and conducted the *in silico*
509 experiments and analysis, contributed to discussions and wrote the manuscript. MS
510 participated in the *in silico* experiments/analysis, discussions and manuscript
511 preparation. MBi provided insights and contributed to the *in silico* experiments,
512 discussions, and manuscript preparation. MMM was involved in obtaining the
513 biological material/reagents for library construction, in the discussions and in revising
514 the manuscript. AMM and CRM were involved in the discussions, contributed expert
515 insights and reviewed the manuscript. GRF conceived the study, participated in its
516 design and coordination, contributed to discussions and prepared the manuscript. All
517 authors read and approved the final manuscript.

518

519 **Additional data files**

520 The following additional data are available with the online version of this
521 paper. Additional file 1 (PDF format) contains Supplementary Figures S1 to S4 as
522 well as Supplementary Tables S1 and S2. Additional file 2 (XLSX format) is a table
523 that lists the polycistronic genes identified through analyzing the trans-spliced *S.*
524 *mansoni* transcriptome.

525

526 **Data availability**

527 The RNA-Seq data for this study were obtained from the SRA archive; the
528 accession numbers are described in Additional file 1: Table S1. The SLTS trapping
529 experiments are available in the SRA archive, accession numbers [SRA:SRR1134198
530 and SRA:SRR1134204].

531

532 **Acknowledgements**

533 The authors would like to acknowledge the valuable contribution of Dr. José
534 Marcos Chaves Ribeiro, who assisted in functional annotation of the *S. mansoni*
535 transcripts, aided in the *in silico* experiments and provided the Classifier program for
536 functional annotation of transcripts. We also thank Dr. Andreza Chagas, Dr. Eric
537 Calvo and Neuza Antunes, who helped with the experimental procedures, Dr. Priscila
538 Grynberg, who helped with the *in silico* experiments and Dr. Liana Janotti-Passos for
539 maintaining the *S. mansoni* life cycle at Centro de Pesquisas René Rachou. This
540 research was supported by the funding agencies CNPq (140544/2011-9 fellowship for
541 MBo, grant id 480576/2010-6 for MMM, grant id 306232/2009-0 for GRF),
542 FAPEMIG (CBB- APQ-01715-11 for MMM and CBB - APQ-00529-13 for GRF),

543 CAPES (fellowship for MBI), and Burroughs Wellcome Fund - BWF (Schistosome
544 Toolbox Collaborative Research Travel Award for MBo).

545

546 **References**

- 547 1. Bitar M, Boroni M, Macedo AM, Machado CR, Franco GR: **The spliced leader**
548 **trans-splicing mechanism in different organisms: molecular details and**
549 **possible biological roles.** *Front Genet* 2013, **4**(199)1-14.
- 550 2. Denker J, Zuckerman D: **New components of the spliced leader RNP required**
551 **for nematode trans-splicing.** *Nature* 2002, **417**:667-670.
- 552 3. Thomas JD, Conrad RC, Blumenthal T: **The *C. elegans* Trans-spliced leader**
553 **RNA is bound to Sm and has a trimethylguanosine cap.** *Cell* 1988, **54**:533-
554 539.
- 555 4. Bruzik JP, Van Doren K, Hirsh D, Steitz JA: **Trans splicing involves a novel**
556 **form of small nuclear ribonucleoprotein particles.** *Nature* 1988, **335**:559-
557 562.
- 558 5. Lasda EL, Blumenthal T: **Trans-splicing.** *Dev Biol* 2011, **2**:417-434.
- 559 6. Stover NA, Kaye MS, Cavalcanti ARO: **Spliced leader trans-splicing.** *Curr Biol*
560 2006, **16**:8-9.
- 561 7. Hastings KEM: **SL trans-splicing: easy come or easy go?** *Trends Genet*
562 2005, **21**:240-247.

- 563 8. Saito TL, Hashimoto S, Gu SG, Morton JJ, Stadler M, Blumenthal T, Fire A,
564 Morishita S: **The transcription start site landscape of *C. elegans***. *Genome*
565 *Res* 2013, **23**:1348–1361.
- 566 9. Nilsson D, Gunasekera K, Mani J, Osteras M, Farinelli L, Baerlocher L, Roditi I,
567 Ochsenreiter T: **Spliced Leader Trapping Reveals Widespread Alternative**
568 **Splicing Patterns in the Highly Dynamic Transcriptome of *Trypanosoma***
569 ***brucei***. *PLoS Pathog* 2010, **6**:e1001037.
- 570 10. Lee K-Z, Sommer RJ: **Operon structure and trans-splicing in the nematode**
571 ***Pristionchus pacificus***. *Mol Biol Evol* 2003, **20**:2097–2103.
- 572 11. Evans D, Blumenthal T: **trans splicing of polycistronic *Caenorhabditis elegans***
573 **pre-mRNAs: analysis of the SL2 RNA**. *Mol Cell Biol* 2000, **20**:6659–6667.
- 574 12. Guiliano DB, Blaxter ML: **Operon Conservation and the Evolution of trans-**
575 **Splicing in the Phylum Nematoda**. *PLoS Genet* 2006, **2**:12.
- 576 13. Blumenthal T: **Operons in eukaryotes**. *Brief Funct Genomic Proteomic* 2004,
577 **3**:199–211.
- 578 14. Blumenthal T: **Trans-splicing and operons**. *WormBook* 2005:1–9.
- 579 15. Conrad R, Thomas J, Spieth J, Blumenthal T: **Insertion of part of an intron**
580 **into the 5' untranslated region of a *Caenorhabditis elegans* gene converts**
581 **it into a trans-spliced gene**. *Mol Cell Biol* 1991, **11**:1921–1926.
- 582 16. Allen MA, Hillier LW, Waterston RH, Blumenthal T: **A global analysis of *C.***
583 ***elegans* trans -splicing**. *Genome Res* 2011, **21**:255–264.

- 584 17. Mourão M de M, Bitar M, Lobo FP, Peconick AP, Grynberg P, Prosdocimi F,
585 Waisberg M, Cerqueira GC, Macedo AM, Machado CR, Yoshino T, Franco
586 **GR: A directed approach for the identification of transcripts harbouring**
587 **the spliced leader sequence and the effect of trans-splicing knockdown in**
588 ***Schistosoma mansoni*. Mem Inst Oswaldo Cruz 2013, 108:707–717.**
- 589 18. Siegel TN, Tan KSW, Cross GAM: **Systematic study of sequence motifs for**
590 **RNA trans splicing in *Trypanosoma brucei*. Mol Cell Biol 2005, 25:9586–**
591 **9594.**
- 592 19. Matsumoto J, Dewar K, Wasserscheid J, Wiley GB, Macmil SL, Roe BA, Zeller
593 RW, Satou Y, Hastings KEM: **High-throughput sequence analysis of *Ciona***
594 ***intestinalis* SL trans-spliced mRNAs: alternative expression modes and**
595 **gene function correlates. Genome Res 2010, 20:636–645.**
- 596 20. **The Carter Center. Schistosomiasis Control Program**
597 [<http://www.cartercenter.org/health/schistosomiasis/index.html>]
- 598 21. WHO - World Health Organization: *Schistosomiasis. Fact Sheet 115*. 2013.
- 599 22. Rajkovic A, Davis RE, Simonsen JN, Rottman FM: **A spliced leader is present**
600 **on a subset of mRNAs from the human parasite *Schistosoma mansoni*.**
601 *Biochemistry* 1990, **87:8879–8883.**
- 602 23. Davis RE, Hardwick C, Tavernier P: **RNA Trans-splicing in Flatworms.**
603 **Analysis of trans-spliced mRNAs and genes in the human parasite,**
604 ***Schistosoma mansoni*. J Biol Chem 1995, 270:21813–21819.**

- 605 24. Davis RE, Hodgson S: **Gene linkage and steady state RNAs suggest trans-**
606 **splicing may be associated with a polycistronic transcript in *Schistosoma***
607 ***mansoni*. *Mol Biochem Parasitol* 1997, **89**:25–39.**
- 608 25. Protasio A V, Tsai IJ, Babbage A, Nichol S, Hunt M, Aslett MA, Silva N De,
609 Velarde GS, Anderson TJC, Clark RC, Davidson C, P G, Holroyd NE,
610 Loverde PT, Lloyd C, Mcquillan J, Otto TD, Parker-manuel SJ, Quail MA,
611 Wilson RA, Dunne DW, Berriman M: **A Systematically Improved High**
612 **Quality Genome and Transcriptome of the Human Blood Fluke**
613 ***Schistosoma mansoni*. *PLoS Negl Trop Dis* 2012, **6**:e1455.**
- 614 26. Gobert GN, Moertel L, Brindley PJ, McManus DP: **Developmental gene**
615 **expression profiles of the human pathogen *Schistosoma japonicum*. *BMC***
616 ***Genomics* 2009, **10**:128.**
- 617 27. Slamovits CH, Keeling PJ: **Widespread recycling of processed cDNAs in**
618 **dinoflagellates. *Curr Biol* 2008, **18**:R550–2.**
- 619 28. Berriman M, Haas BJ, LoVerde PT, Wilson RA, Dillon GP, Cerqueira GC,
620 Mashiyama ST, Al-Lazikani B, Andrade LF, Ashton PD, Aslett MA,
621 Bartholomeu DC, Blandin G, Caffrey CR, Coghlan A, Coulson R, Day TA,
622 Delcher A, DeMarco R, Djikeng A, Eyre T, Gamble JA, Ghedin E, Gu Y,
623 Hertz-Fowler C, Hirai H, Hirai Y, Houston R, Ivens A, Johnston DA, et al.:
624 **The genome of the blood fluke *Schistosoma mansoni*. *Nature* 2009,**
625 **460:352–358.**
- 626 29. Demarco R, Kowaltowski AT, Machado AA, Soares MB, Gargioni C, Kawano
627 T, Rodrigues V, Madeira AMBN, Wilson RA, Menck CFM, Setubal C, Dias-

- 628 neto E, Leite LCC, Verjovski-almeida S: **Saci-1, -2, and -3 and Perere, Four**
629 **Novel Retrotransposons with High Transcriptional Activities from the**
630 **Human Parasite *Schistosoma mansoni*. *J Virol* 2004, 78:2967–2978.**
- 631 30. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.**
632 *Nucleic Acids Res* 2000, **28**:27–30.
- 633 31. Davis RE: **Spliced leader RNA trans-splicing in metazoa. *Parasitol Today***
634 1996, **12**:33–40.
- 635 32. Kuo RC, Zhang H, Zhuang Y, Hannick L, Lin S: **Transcriptomic study reveals**
636 **widespread spliced leader trans-splicing, short 5'-UTRs and potential**
637 **complex carbon fixation mechanisms in the euglenoid Alga *Eutreptiella***
638 **sp. *PLoS One* 2013, 8:e60826.**
- 639 33. Coolidge CJ, Seely RJ, Patton JG: **Functional analysis of the polypyrimidine**
640 **tract in pre-mRNA splicing. *Nucleic Acids Res* 1997, 25:888–896.**
- 641 34. Huang J, Van der Ploeg LH: **Requirement of a polypyrimidine tract for**
642 **trans-splicing in trypanosomes: discriminating the PARP promoter from**
643 **the immediately adjacent 3' splice acceptor site. *EMBO J* 1991, 10:3877–**
644 3885.
- 645 35. Matthews KR, Tschudi C, Ullu E: **A common pyrimidine-rich motif governs**
646 **trans-splicing and polyadenylation of tubulin polycistronic pre-mRNA in**
647 **trypanosomes. *Genes Dev* 1994, 8:491–501.**
- 648 36. Bindereif A, Green MR: **Ribonucleoprotein complex formation during pre-**
649 **mRNA splicing in vitro. *Mol Cell Biol* 1986, 6:2582–2592.**

- 650 37. Conrad RC, Lea K, Blumenthal T: **SL1 trans-splicing specified by AU-rich**
651 **synthetic RNA inserted at the 5' end of *Caenorhabditis elegans* pre-**
652 **mRNA.** *RNA* 1995, **1**:164–170.
- 653 38. Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, Cheng Y,
654 Cleland I, Faruque N, Goodgame N, Gibson R, Hoad G, Jang M, Pakseresht
655 N, Plaister S, Radhakrishnan R, Reddy K, Sobhany S, Ten Hoopen P,
656 Vaughan R, Zalunin V, Cochrane G: **The European Nucleotide Archive.**
657 *Nucleic Acids Res* 2011, **39**(Database issue):D28–31.
- 658 39. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with**
659 **RNA-Seq.** *Bioinformatics* 2009, **25**:1105–1111.
- 660 40. Ewing B, Green P: **Base-calling of automated sequencer traces using phred.**
661 **II. Error probabilities.** *Genome Res* 1998, **8**:186–194.
- 662 41. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer**
663 **traces using phred. I. Accuracy assessment.** *Genome Res* 1998, **8**:175–185.
- 664 42. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G,
665 Abecasis G, Durbin R, Subgroup 1000 Genome Project Data Processing: **The**
666 **Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009,
667 **25**:2078–2079.
- 668 43. Anders S, Huber W: **Differential expression analysis for sequence count data.**
669 *Genome Biol* 2010, **11**:R106.
- 670 44. Anders S, Reyes A, Huber W: **Detecting differential usage of exons from**
671 **RNA-seq data.** *Genome Res* 2012, **22**:2008–2017.

- 672 45. Karim S, Singh P, Ribeiro JMC: **A Deep Insight into the Sialotranscriptome**
673 **of the Gulf Coast Tick, *Amblyomma maculatum***. *PLoS One* 2011, **6**:e28525.
- 674 46. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment**
675 **search tool**. *J Mol Biol* 1990, **215**:403–410.
- 676 47. Bairoch A, Boeckmann B, Ferro S, Gasteiger E: **Swiss-Prot: juggling between**
677 **evolution and stability**. *Brief Bioinform* 2004, **5**:39–55.
- 678 48. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP,
679 Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L,
680 Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM,
681 Sherlock G: **Gene ontology: tool for the unification of biology. The Gene**
682 **Ontology Consortium**. *Nat Genet* 2000, **25**:25–9.
- 683 49. Koonin E V, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS,
684 Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Rogozin IB, Smirnov
685 S, Sorokin A V, Sverdlov A V, Vasudevan S, Wolf YI, Yin JJ, Natale DA: **A**
686 **comprehensive evolutionary classification of proteins encoded in complete**
687 **eukaryotic genomes**. *Genome Biol* 2004, **5**:R7.
- 688 50. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N,
689 Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer ELL, Eddy
690 SR, Bateman A, Finn RD: **The Pfam protein families database**. *Nucleic*
691 *Acids Res* 2012, **40**(Database issue):D290–301.
- 692 51. Schultz J, Copley RR, Doerks T, Ponting CP, Bork P: **SMART: a web-based**
693 **tool for the study of genetically mobile domains**. *Nucleic Acids Res* 2000,
694 **28**:231–234.

- 695 52. Yamada T, Letunic I, Okuda S, Kanehisa M, Bork P: **iPath2.0: interactive**
696 **pathway explorer**. *Nucleic Acids Res* 2011, **39**(Web Server issue):W412–415.
- 697 53. Thorvaldsdóttir H, Robinson JT, Mesirov JP: **Integrative Genomics Viewer**
698 **(IGV): high-performance genomics data visualization and exploration**.
699 *Brief Bioinform* 2013, **14**:178–192.
- 700 54. Kent WJ: **BLAT--the BLAST-like alignment tool**. *Genome Res* 2002, **12**:656–
701 664.
- 702 55. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing**
703 **genomic features**. *Bioinformatics* 2010, **26**:841–842.
- 704 56. Crooks GE, Hon G, Chandonia J-M, Brenner SE: **WebLogo: a sequence logo**
705 **generator**. *Genome Res* 2004, **14**:1188–1190.
- 706 57. Blanco E, Parra G, Guigó R: **Using geneid to identify genes**. *Curr Protoc*
707 *Bioinformatics* 2007, **Chapter 4**:Unit 4.3.
- 708 58. Corvelo A, Hallegger M, Smith CWJ, Eyras E: **Genome-wide association**
709 **between branch point properties and alternative splicing**. *PLoS Comput*
710 *Biol* 2010, **6**:e1001016.

711

712 **Figure Legends**

713 **Figure 1 - Comparison between the SL Trapping and the RNA-Seq**

714 **Filtered datasets**. A – A Venn Diagram representation of the SLTS events in the
715 Trapping and the RNA-Seq Filtered datasets before the Trapping dataset curation; B –
716 Genes exclusively detected in the RNA-Seq Filtered dataset among three compiled
717 stages; C – SLTS events in the Trapping and the RNA-Seq Filtered datasets after the

718 Trapping dataset curation; and D - Expression correlation between the trans-spliced
719 genes in the SL Trapping and RNA-Seq Filtered datasets (Pearson correlation
720 coefficient = 0.5).

721 **Figure 2 – Visualization of the SL insertion sites in genes.** A genome
722 browser view of the gene structure for two example genes with the read counts from
723 the RNA-Seq Filtered dataset (top track), the SL Trapping 1 dataset (middle track),
724 and the SL Trapping 2 dataset (bottom track) superimposed. A – Gene structure of the
725 trans-spliced transcript [GeneDB:Smp_034190.1]. B – Structure of the genes
726 [GeneDB:Smp-173910.1 and GeneDB:Smp_094470.1] presenting a genomic distance
727 of 87 bp from each other.

728 **Figure 3 - Functional classification of the trans-spliced genes.** Functional
729 annotations for the trans-spliced genes (blue bars) are quantitatively compared with
730 the background of all expressed genes (red bars). The bars represent the percentage of
731 transcripts classified into each functional category.

732 **Figure 4 – Comparison between gene expression and trans-splicing**
733 **frequency.** A – Normalized log-expression levels of non-trans-spliced genes
734 (estimated from the RNA-Seq dataset obtained from the SRA) are compared with the
735 log-expression levels of trans-spliced genes (estimated from the SL Trapping
736 datasets). B – Normalized expression for the trans-spliced genes (estimated from the
737 RNA-Seq dataset obtained from the SRA) is compared with the normalized SLTS
738 frequency (estimated from the SL Trapping datasets). Dashed black lines denote the
739 average levels of both indicators (163 for gene expression and 106 for SLTS
740 frequency), segregating the genes with relatively low SLTS frequency (lower-right
741 quadrant) from genes with relatively high SLTS frequency (upper-left quadrant) and
742 comparing their expression levels. C – The probability distribution of the gene

743 expression levels (pink curve) and SLTS frequencies (blue curve).

744 **Figure 5 – Chromosomal origin, alternative splicing isoforms and number**
745 **of exons in the trans-spliced genes.** A – The relative distribution of trans-spliced
746 genes with high expression levels (GED, blue bars) and trans-spliced genes with low
747 expression levels (TSD, purple bars) in each chromosome is compared with the
748 chromosome distribution of genes expressed on the entire *S. mansoni* transcriptome
749 (red bars) and genes expressed only in the cercariae transcriptome (green bars). B –
750 The relative distribution of alternative splicing transcript isoforms per gene in the
751 entire *S. mansoni* transcriptome, in the cercariae transcriptome, in the subset of GED
752 genes and in the subset of TSD genes. C – The relative distribution of transcripts from
753 genes with single or multiple exons in the entire *S. mansoni* transcriptome, in the
754 cercariae transcriptome, in the subset of GED genes and in the subset of TSD genes.

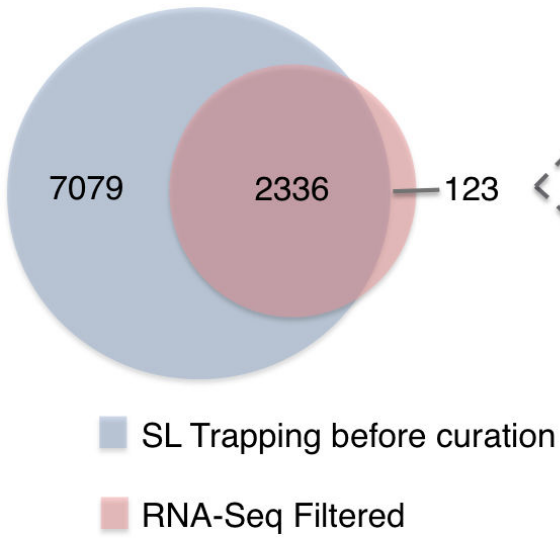
755 **Figure 6 – Correlations between trans-splicing occurrences, exons**
756 **position in gene bodies and both exon and intron lengths.** A – The distribution of
757 trans-splicing frequencies (estimated from SL Trapping) (left panel) is compared with
758 the expression levels of exons according to their position (estimated from the RNA-
759 Seq experiment obtained from the SRA) (right panel). Groups of genes were divided
760 according to their number of annotated exons. Only genes with three, four, five and
761 ten exons are shown. B – The distribution of exon lengths in trans-spliced and non-
762 trans-spliced genes that present 3, 4, 5 and 10 exons. C – The distribution of introns
763 lengths immediately upstream of the splicing event in trans-spliced and non-trans-
764 spliced genes presenting 3, 4, 5 and 10 exons.

765 **Figure 7 – Comparisons among acceptor splice site motifs in introns.**
766 Intron/exons boundaries represented by the region from -24 to +3 nucleotides and
767 containing acceptor splice sites in genes were analyzed based on their sequence

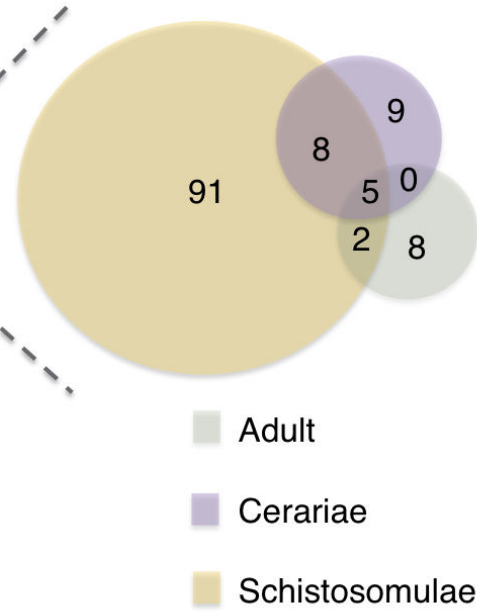
768 composition. Introns of genes that are exclusively cis-spliced correspond to sequences
769 on the left side of the triplets, cis-spliced introns in transcripts that undergo trans-
770 splicing correspond to sequences in the middle of the triplets, and the trans-spliced
771 introns that correspond to sequences on the right side of the triplets.

772 **Figure 8 – Analyses of intron lengths, strength of acceptor sites and**
773 **polypyrimidine tracts based on the intron positions.** A – The intron lengths are
774 plotted based on the intron position in genes (single, first, internal, and last introns) of
775 three distinct groups: CS - cis-spliced introns - green boxes, CTS - cis-spliced introns
776 in trans-spliced transcripts where the intron is not the primary SLTS target - red
777 boxes, and TS trans-spliced introns - blue boxes. B – The strengths of acceptor splice
778 sites (log-odds scores, as computed by a Markov model trained for the splice site
779 sequence) are plotted according to the intron position in genes of the same three
780 distinct groups. C – Length of the polypyrimidine tract in nt relative to the acceptor
781 sites (computed using different models integrated in the SVM-BPfinder) are plotted
782 according to intron position in genes of the same three distinct groups.

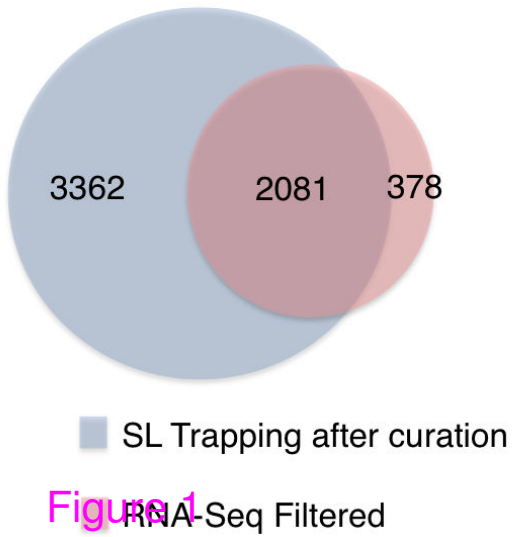
A



B



C



D

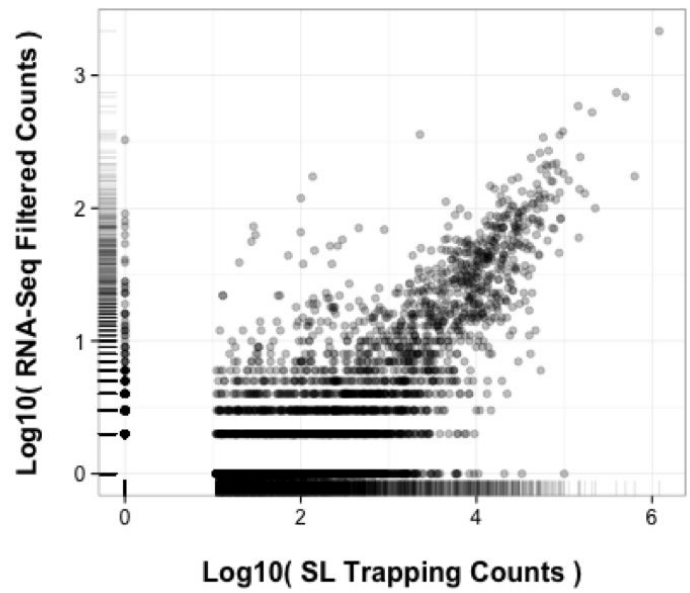


Figure 1

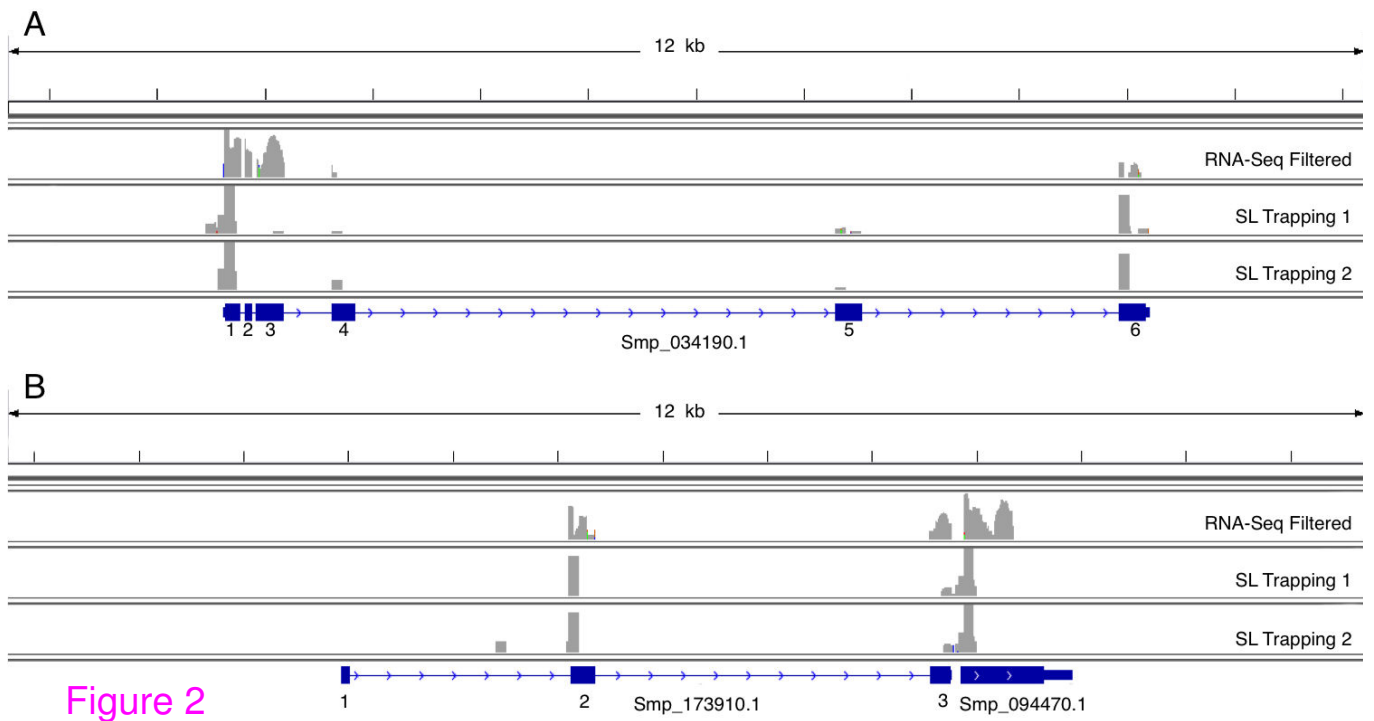


Figure 2

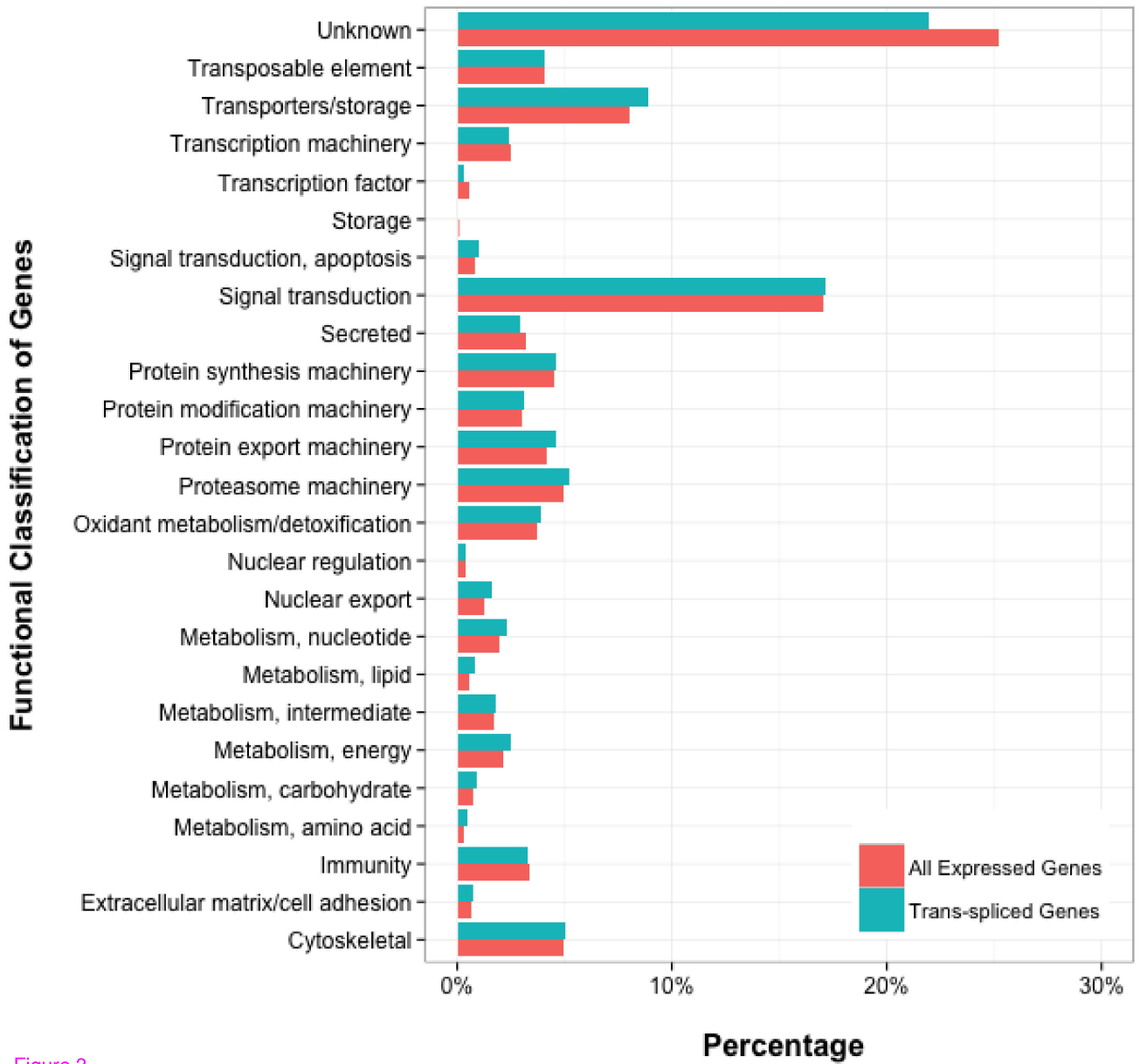


Figure 3

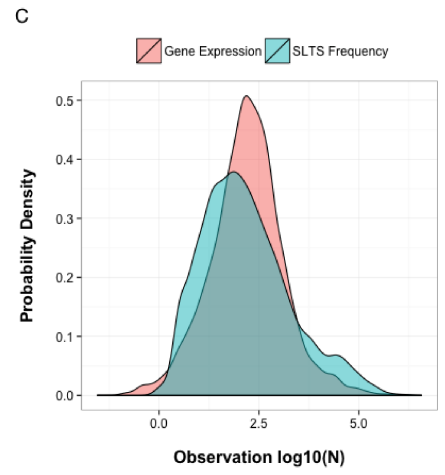
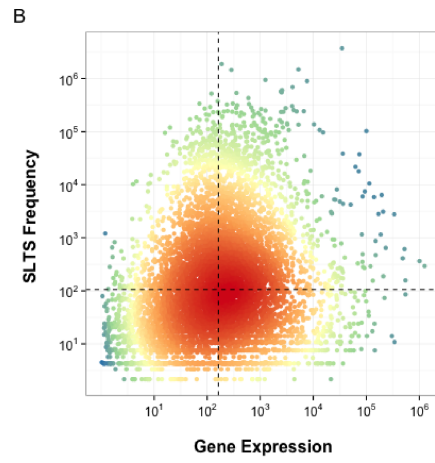
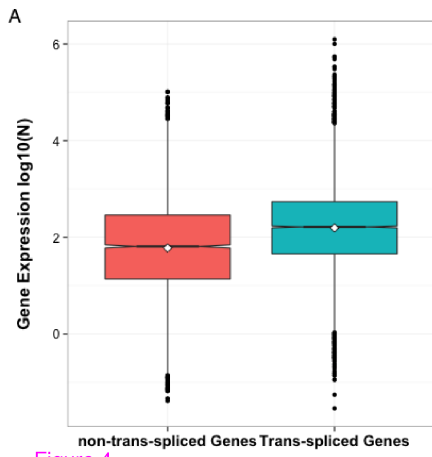


Figure 4

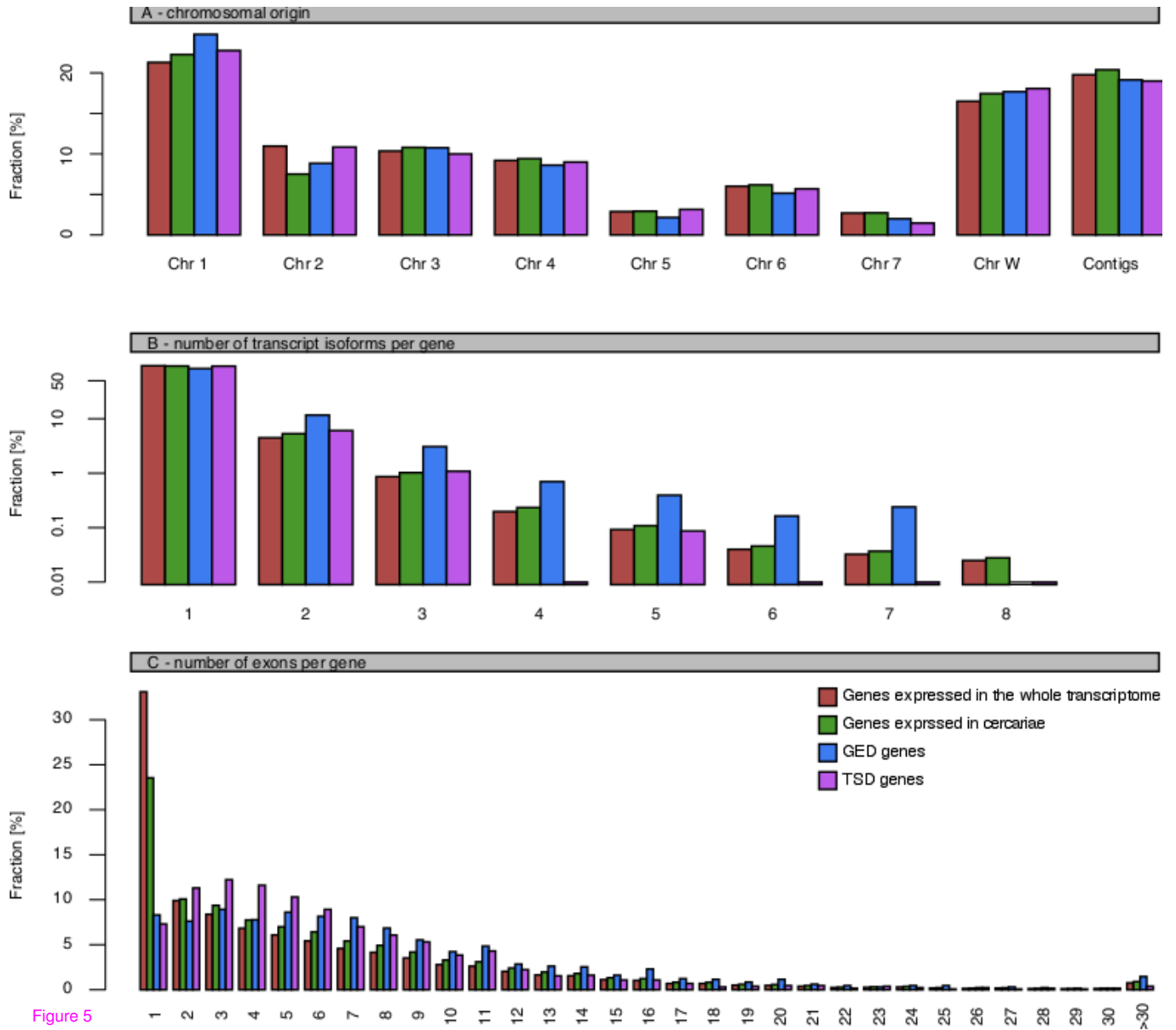


Figure 5

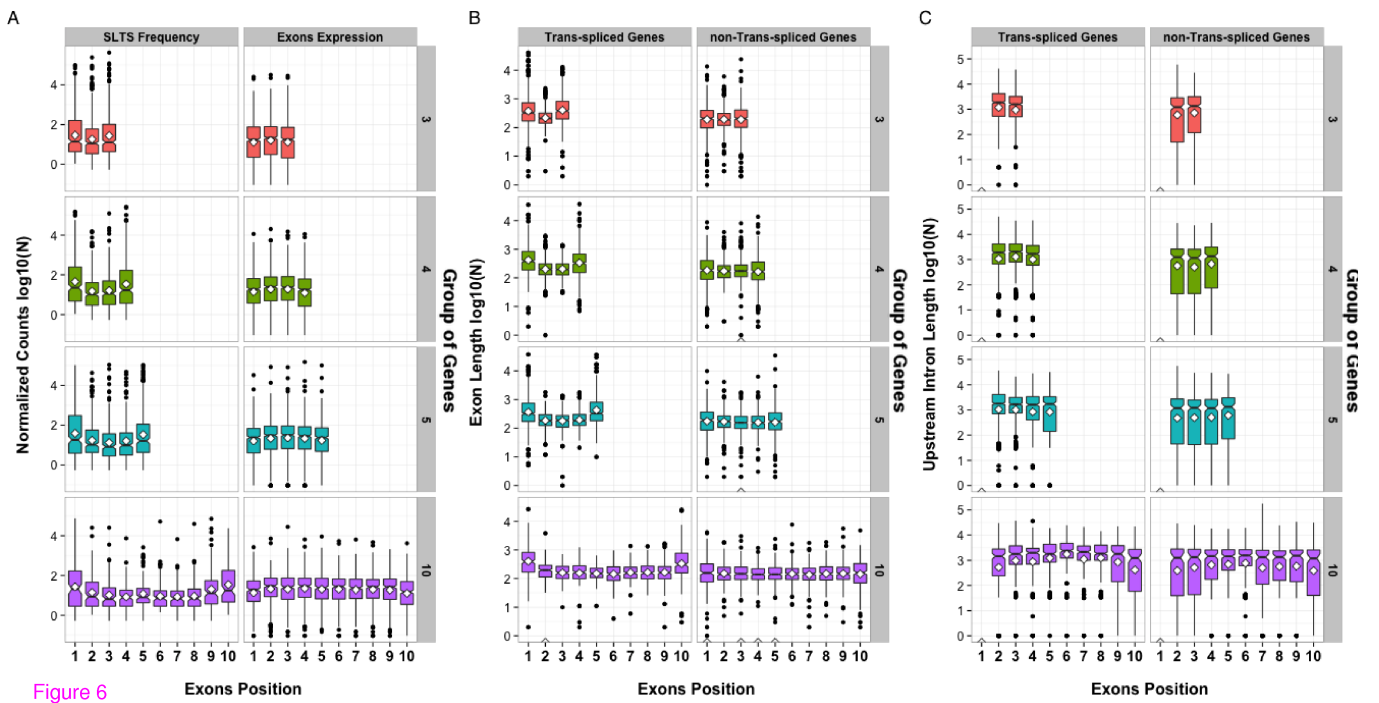


Figure 6

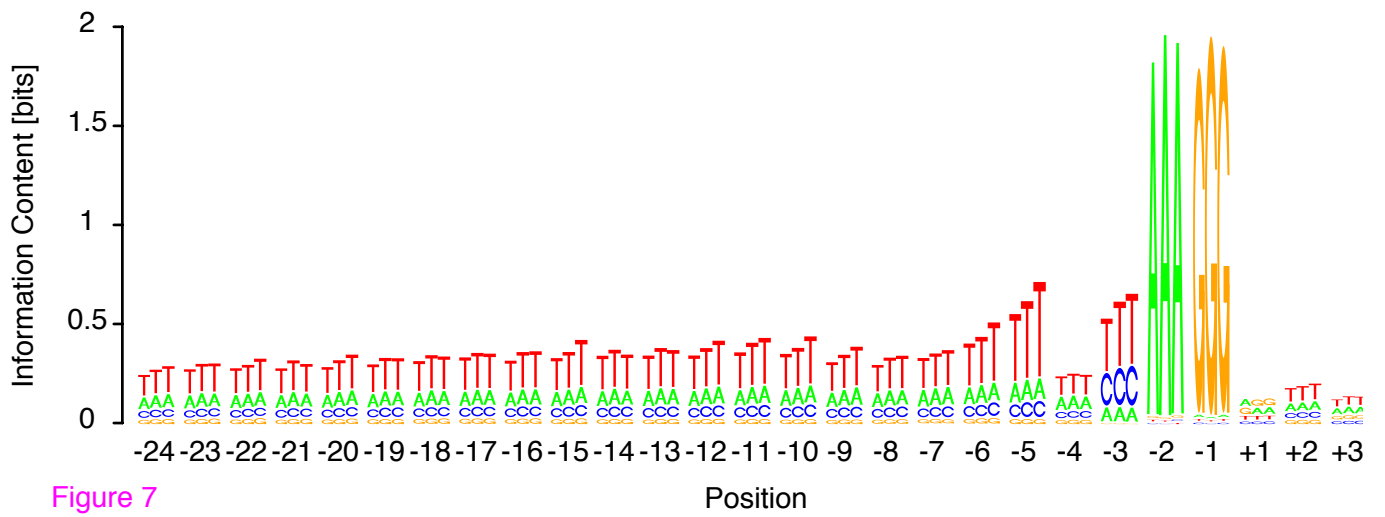
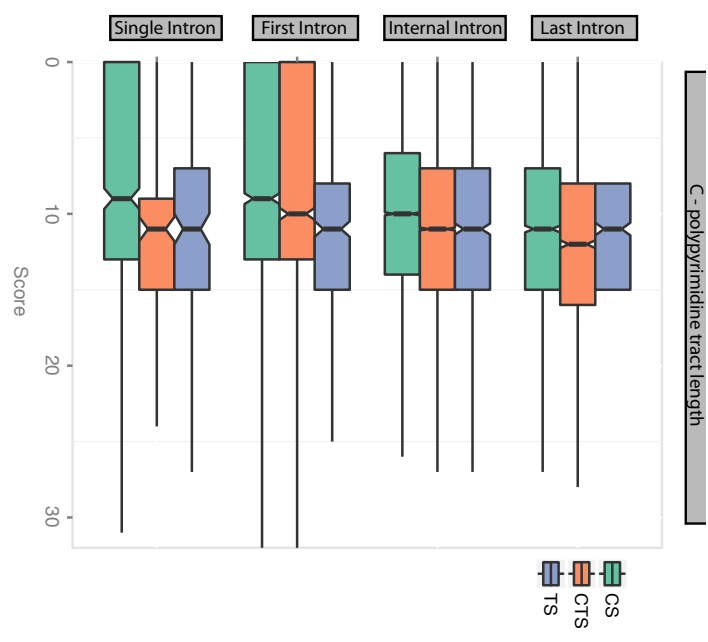
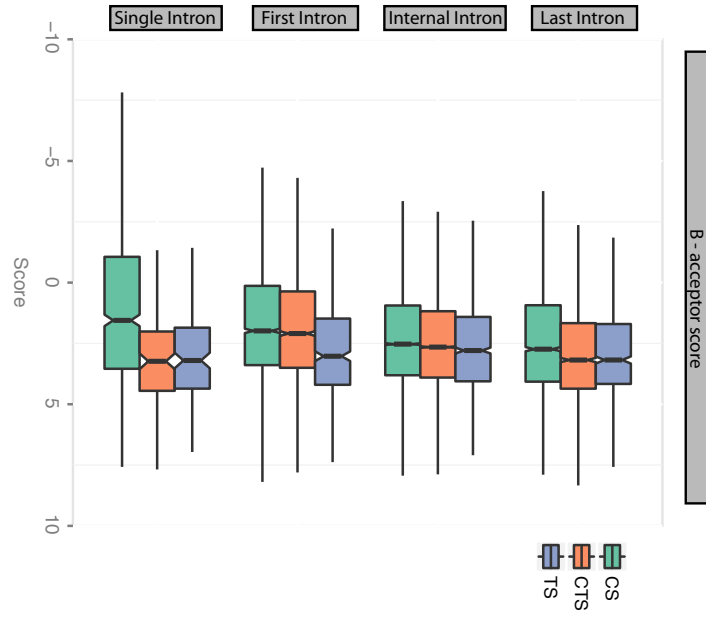
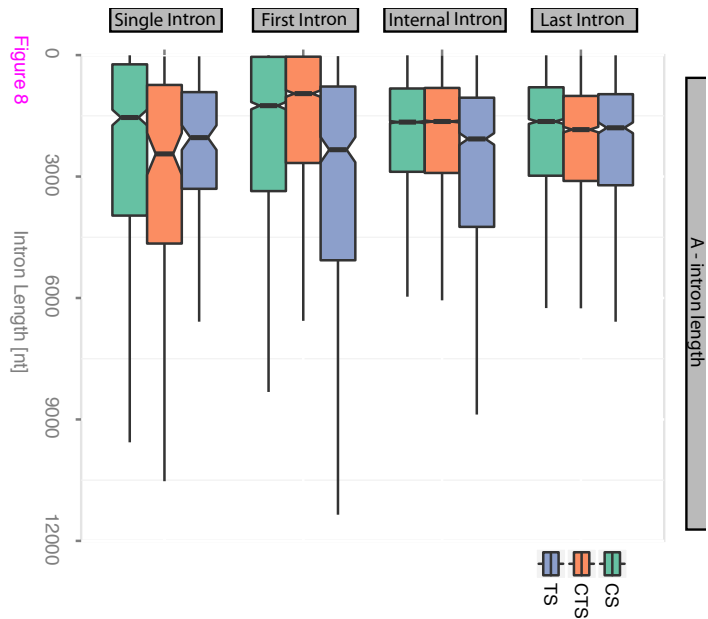


Figure 7



Additional files provided with this submission:

Additional file 1: Additional file 1.docx, 4492K

<http://genomebiology.com/imedia/1542275001134398/supp1.docx>

Additional file 2: Additional file 2.xlsx, 19K

<http://genomebiology.com/imedia/5515221213439830/supp2.xlsx>

8. ANEXO 2 : THE SPLICED LEADER TRANS-SPLICING MECHANISM IN DIFFERENT ORGANISMS: MOLECULAR DETAILS AND POSSIBLE BIOLOGICAL ROLES.



The spliced leader trans-splicing mechanism in different organisms: molecular details and possible biological roles

Mainá Bitar, Mariana Boroni, Andréa M. Macedo, Carlos R. Machado and Glória R. Franco*

Laboratório de Genética Bioquímica, Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Edited by:

Thiago M. Venancio, Universidade Estadual do Norte Fluminense, Brazil

Reviewed by:

Lakshminarayan M. Iyer, National Institutes of Health, USA
Vincius Maracaja-Coutinho, Universidad Mayor, Chile

*Correspondence:

Glória R. Franco, Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Av. Antônio Carlos 6627, Pampulha, Belo Horizonte, 31270-901, Brazil
e-mail: gfrancoufmg@gmail.com

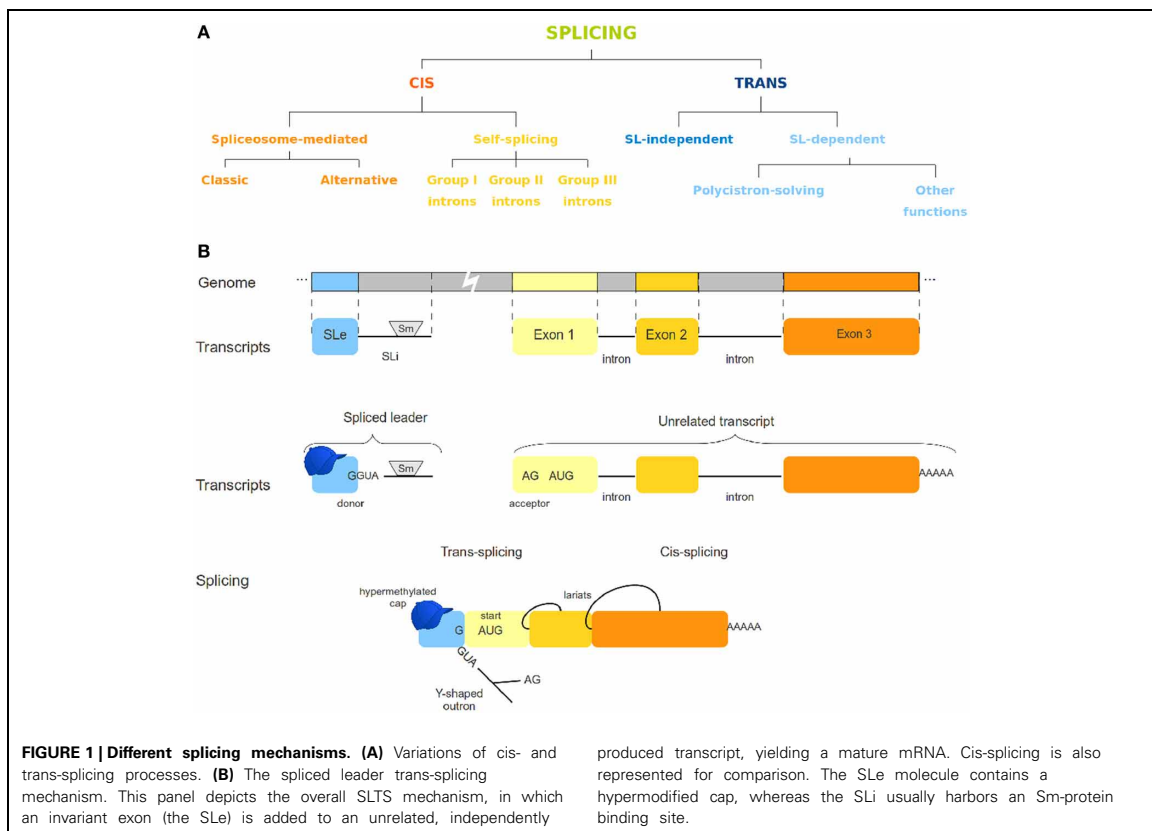
The spliced leader (SL) is a gene that generates a functional ncRNA that is composed of two regions: an intronic region of unknown function (SLi) and an exonic region (SLe), which is transferred to the 5' end of independent transcripts yielding mature mRNAs, in a process known as spliced leader trans-splicing (SLTS). The best described function for SLTS is to solve polycistronic transcripts into monocistronic units, specifically in Trypanosomatids. In other metazoans, it is speculated that the SLe addition could lead to increased mRNA stability, differential recruitment of the translational machinery, modification of the 5' region or a combination of these effects. Although important aspects of this mechanism have been revealed, several features remain to be elucidated. We have analyzed 157 SLe sequences from 148 species from seven phyla and found a high degree of conservation among the sequences of species from the same phylum, although no considerable similarity seems to exist between sequences of species from different phyla. When analyzing case studies, we found evidence that a given SLe will always be related to a given set of transcripts in different species from the same phylum, and therefore, different SLe sequences from the same species would regulate different sets of transcripts. In addition, we have observed distinct transcript categories to be preferential targets for the SLe addition in different phyla. This work sheds light into crucial and controversial aspects of the SLTS mechanism. It represents a comprehensive study concerning various species and different characteristics of this important post-transcriptional regulatory mechanism.

Keywords: spliced-leader, trans-splicing, non-coding RNAs, RNA sequence analysis, RNA secondary structure

INTRODUCTION

Splicing has been known for a long time as a post-transcriptional process for regulation of gene expression [reviewed by Gilbert (1978); Padgett et al. (1986) and further by several other authors]. To date, there are different known variants of both cis and trans-splicing mechanisms (Figure 1A). Cis-splicing differs from trans-splicing on the genomic origin of the transcripts involved. While cis-splicing is concerned with transcripts from a single gene, trans-splicing mechanisms act by connecting transcripts of otherwise unrelated genes (Figure 1). In the spliced leader trans-splicing (SLTS) mechanism, the exonic portion of the spliced leader (SLe) transcript is transferred to the 5' end of unrelated transcripts to yield a mature mRNA (Boothroyd and Cross, 1982) and reviewed by Liang et al. (2003) (Figure 1B). As first observed in trypanosomatids, its best described function is to resolve polycistronic transcripts into monocistronic units (Sather and Agabian, 1985) and reviewed by Preußner et al. (2012). Subsequently, SLTS was demonstrated to occur in other euglenozoans (Tessier et al., 1991) and several organisms, such as rotifers (Pouchkina-Stantcheva and Tunnacliffe, 2005), cnidarians (Stover and Steele, 2001), chordata (Vandenbergh et al., 2001), nematoda (Krause and Hirsh, 1987), platyhelminthes (Rajkovic et al., 1990), and dinoflagellates (Lidie and Van Dolah, 2007). Different biological roles have been proposed for this

mechanism, such as: (i) enhancing translation of trans-spliced transcripts by providing a specialized (hypermethylated) 5' cap structure for trans-spliced transcripts, (ii) stabilizing the mRNAs, and (iii) removing regulatory elements from the outtron, what has been called 5' UTR sanitization (Hastings, 2005; Matsumoto et al., 2010). In a recent work Nilsson and collaborators (Nilsson et al., 2010) discuss possible functions for the SLTS mechanism in *Trypanosoma brucei*. According to the authors, the differential insertion of the SLe sequence in alternative acceptor sites of genes could lead to: (i) translation blockage when the SLe insertion is upstream the protein start codon; (ii) alteration of protein subcellular location, when signaling sequences are eliminated by SLe insertion; (iii) inclusion or exclusion of uORFs or other regulatory elements from 5' end of transcripts; and (iv) translation of alternative ORFs. It has also been speculated that SLe addition could lead to increased mRNA stability, differential recruitment of the translational machinery, modification of the 5' UTR or a combination of those effects [reviewed by Hastings (2005); Stover et al. (2006); Lasda and Blumenthal (2011)]. From a recent study concerning the SLTS mechanism in the flatworm *Schistosoma mansoni* (Mourão et al., 2013), our group identified transcripts under trans-splicing regulation in different life-cycle stages, suggesting that the SLTS could account for differential protein levels and protein repertoires in different stages and/or



environmental conditions. Although important aspects of the SLTS mechanism in this parasite were elucidated, several other questions regarding this mechanism in *S. mansoni* and other organisms remained unanswered. These questions surround the existence of conserved motifs within SL sequences from different species, the possible emergence of SLTS in more complex taxa as plants and mammals, the role of the SLTS mechanism in different organisms, the peculiarities of the set of transcripts under the control of a given SLe and the structural aspects of the SL molecule.

We performed analyses on a great number of SL sequences, searching for answers to such questions. All those answers lead to one final question, which has been debated in the literature for a long time: “what is the origin of the SLTS mechanism and how has it evolved?” This work is devoted to the proposition of hypotheses based on observed features of this mechanism that may guide the composition of a future final answer to this question. As for now, although there is no conclusive statement, there are important observations that can help the characterization of the mechanism, its biological role, phylogenetic features and molecular details. In the course of this study, we have compiled a comprehensive dataset of SL sequences from several species of various phyla. This was the starting point for several computational analyses that explored different aspects of the SLTS

mechanism. To the best of our knowledge, no studies of this magnitude have been previously performed, regarding so many distinct features of this poorly understood mechanism in so many different species. Therefore, this work can largely contribute to a general overview of the SLTS in different biological contexts.

MATERIALS AND METHODS

SEQUENCE RETRIEVAL AND DATABASE GENERATION

We performed manual searches in the National Center for Biotechnology Information (NCBI <http://www.ncbi.nlm.nih.gov>) database to identify previously annotated SLe sequences. The searches were guided by sequence features, specifically considering entries under the “miscellaneous RNA” sobriquet. The exact expression informed in the search field was “misc_RNA[Feature key]spliced leader.” Once sequences were retrieved, a manual curation was carefully performed to exclude false positives and reduce redundancy. A consistent preliminary database (hereafter named SEED database) was then compiled containing manually curated and annotated sequences. Notably, only sequences from species in which the SLTS mechanism was previously described were included.

Using the sequences from the former mentioned SEED database as queries, we performed searches using BLAST

(Altschul et al., 1990) to expand the set of SLe sequences, and thereby generate a secondary dataset (hereafter named EXTENDED database). Searches were performed in the nucleotide collections (nt) database from NCBI using the blastn program (local version) with parameters automatically adjusted to address short sequences and allow for the retrieval of up to 500 matches for each query sequence. The results were then analyzed to identify SLe sequences based on information from the SEED database.

Several criteria were used to characterize a sequence as SLe to generate the EXTENDED dataset. As objective criteria, we only considered annotated sequences that displayed 90% or higher nucleotide identity and a query coverage exceeding 90% when compared to the respective query. More subjectively, for matches meeting the objective criteria, we have analyzed the presence of such sequences in the 5' end of transcripts from these species. As an additional step, all uncharacterized sequences from organisms in which the presence of the SLTS mechanism was not previously demonstrated were not included in the datasets at first. Uncharacterized sequences were further analyzed to confirm or disprove those as SLe based on literature and sequence annotation.

Along with the retrieval of SLe sequences from the NCBI database, SL gene candidates were also retrieved to compose a separate database. SL gene candidates were identified based on sequence annotation and further analyzed for exclusion of false positives and redundancy reduction. Whenever duplicated sequences from a given species were identified, only one copy was kept to eliminate redundancy. After manual curation, the sequences were compared to the SLe from the same species present in the EXTENDED database and only candidates containing the entire SLe sequence followed by an intronic region were further considered to be SL genes.

SEQUENCE SIMILARITY ASSESSMENT

During the construction of the previously mentioned databases, when more than one SLe sequence was retrieved for a given species, Clustal (Larkin et al., 2007) alignments were performed to guide redundancy reduction and false positive discovery. Subsequent manual alignments were performed to cluster sequences from the EXTENDED dataset. Sequence alignments were visually analyzed for the identification of duplicated sequences, such as completely identical sequences, partially identical sequences with missing residues and similar sequences presenting up to 5% (1 in every 20 positions, which is around the average size of the SLe sequences of most phyla) substitutions, deletions, or insertions. Notably, missing information at the 5' end of SLe sequences is most likely to occur due to incomplete sequencing. For this reason, whenever data were missing for multiple 5' end positions (more than 5%) in a sequence for a given species, conservation of such positions in the phylum was measured considering only the remaining sequences. To better observe consensus regions within SLe sequences, alignments were also used as input for the generation of sequence logos using the WebLogo suite (Crooks et al., 2004).

TRANS-SPLICED TRANSCRIPTS RETRIEVAL

Once the SLe EXTENDED database was built, its sequences were used as queries to search for transcripts from the respective species bearing the SLe in the 5' end. BLAST searches were performed using the same parameters as described above, the same program (blastn) and reference database (nt). The results were manually curated to yield lists of transcripts that undergo SLTS in a given species. Those lists were joined together in a comprehensive set to allow further inspection of transcript conservation among species. All transcripts annotated as “hypothetical protein” or “unknown protein” were excluded from further analysis. Transcripts coding for identical proteins in different species were clustered together in a database to allow the assessment of conservation between species and across phyla of transcripts under SLTS regulation.

In parallel, data from the recently published work of Protasio et al. (2012) were used to yield a preliminary database of *S. mansoni* SLe-containing sequences. Fasta-format sequences for proteins coded by SLe-containing transcripts were retrieved from GeneDB (Logan-Klumpler et al., 2012), in which transcripts that undergo trans-splicing were associated with the gene ontology (GO) ID number 0000870. In both cases described above, after retrieval, GO terms were assigned to each transcript sequence using the GoAnna and GoSlim (McCarthy et al., 2006), tools from the AgBase web portal (McCarthy et al., 2010). Subsequently, manual annotation and classification according to the main biological function was performed for all transcripts (except those retrieved from the *S. mansoni* database) based on literature data.

SECONDARY STRUCTURE GENERATION

To analyze structural conservation and topological features of the entire SL RNA molecule, all formerly mentioned SL gene sequences were submitted to RNAfold (Hofacker, 2003), a program from the Vienna package for RNA structure generation and energy assessment. Structures were then visually analyzed to search for conserved features in sequences from different species.

GENOMIC LOCATION OF SL SEQUENCES

The genomic location and copy number of the SLe sequences from two *Caenorhabditis* species were retrieved using BLAT (Kent, 2002) results and further analyzed. One hundred nucleotides downstream of each sequence were also retrieved to putatively represent the entire SL gene and allow for sequence comparison.

ONE NOTE REGARDING AUTOMATIZATION

All steps were aided by simple *shell* and/or *perl* scripts to automatically manipulate files, organize and categorize data, recognize specific patterns within the text and perform searches in a file. A methodological workflow is presented in **Figure 2** to illustrate each step of this study.

RESULTS

SEQUENCE RETRIEVAL AND DATABASE GENERATION

In the course of this work, we have assembled a comprehensive database of spliced leader sequences from several phyla. When a simple pattern-matching search was performed in NCBI's database to retrieve SLe sequences, 1161 matches were found.

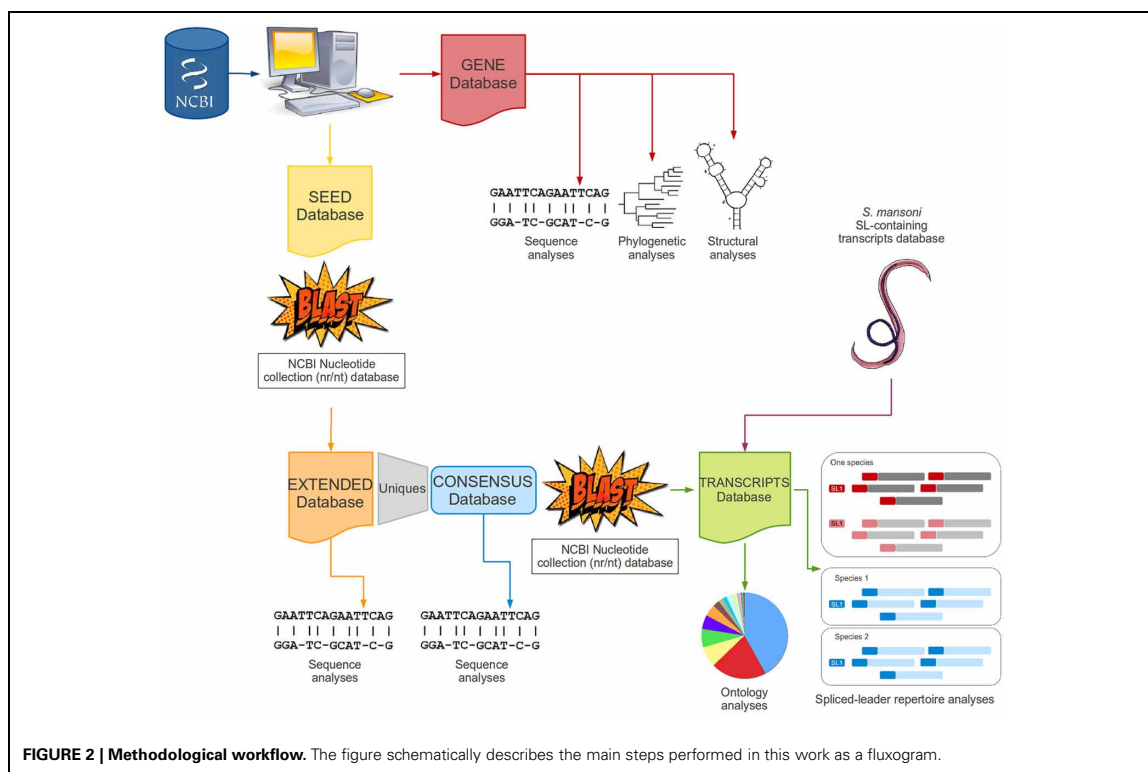


FIGURE 2 | Methodological workflow. The figure schematically describes the main steps performed in this work as a fluxogram.

Among those, the majority (757) were from kinetoplastids and almost half (544) were specifically from *Trypanosoma spp.* The rest included sequences from nematodes (182), flatworms (98), dinoflagellates (95), cnidarians (4), rotifers (2), and chordates (1) phyla, in which the SLTS mechanism was previously described. Sequences from organisms of other phyla were also retrieved, but not included in the datasets because of the lack of previous consistent evidence supporting the presence of SLTS or because no transcripts were found bearing the sequence in the 5' end. After redundancy reduction, false positive exclusion and validation by manual curation, a SEED database was defined containing only sequences with consistent evidence to be considered as SLE. This initial database was comprised 69 sequences from the seven different eukaryotic phyla (Supplementary Table 1): rotifera (2), chordata (1), cnidaria (2), dinoflagellate (8), euglenozoa (33), nematoda (18), and platyhelminthe (5).

These sequences were subsequently used as queries to extend the database based on BLAST similarity searches. The retrieved sequences were analyzed according to previously described criteria and a final EXTENDED database was generated that was comprised of 157 sequences from 148 different species (representing 81 genera) from the same seven phyla (Supplementary Table 2). Notably, all 157 sequences are, in fact, replicates of only 48 unique sequences, which were further clustered into 30 groups of highly similar sequences. This result indicates the high degree of sequence conservation, particularly between species

from the same phylum (as will be further discussed). These 30 SLE sequences originated a third database which we named the CONSENSUS database (Table 1).

In addition to the phyla previously represented within the SEED database, sequences from other phyla have also met the requirements to be considered SLE but were not included in the EXTENDED database. Species of arthropods, ciliophora, echinodermata, mollusca, mycetozoa, apicomplexa, plants and even a bacterial species presented putative SLE sequences (Supplementary Table 3), although most of these phyla have never been proven to harbor the SLTS mechanism. No obvious decision on whether they are real SLE sequences could be reached, and those were therefore excluded from our dataset.

DESCRIPTION OF THE EXTENDED DATABASE

In this section, we present a short description of sequences in the EXTENDED database according to phyla, which is fully available as a supplementary file (Supplementary Table 2) and graphically summarized in Figure 3. There are three species of rotifera in the dataset that share an identical 23 nucleotide SLE sequence that is enriched in adenines. All seven chordata species share an identical SLE sequence, apart from missing residues. The consensus is a 25 nucleotide sequence enriched in thymine nucleotides. Cnidarians are also represented by seven sequences from five species of the same genus (*Hydra*). All species share the exact same SLE sequence and, among those, two of the species present

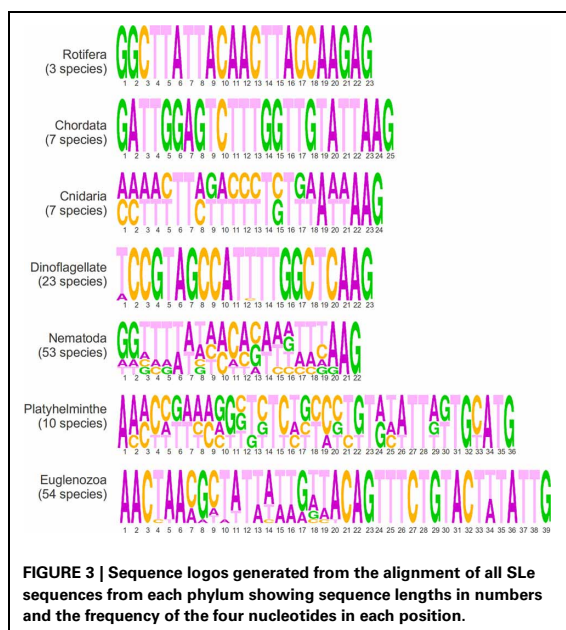
Table 1 | CONSENSUS Database.

| Consensus name | Consensus sequence | Length (composition) | Plenty |
|------------------|--|-----------------------|--------|
| Rotifera 1 | GGCTTATTACAACCTACCAAGAG | 23 (8A/5C/4G/6T) | 3/3 |
| Chordata 1 | GATTGGAGTATTTGGTTGTATTAAG | 25 (6A/8G/11T) | 7/7 |
| Cnidaria 1 | ACTTTTAGTCCCTGTGTAATAAG | 24 (6A/4C/4G/10T) | 5/7 |
| Cnidaria 2 | CAAACCTCTATTTCTTAATAAAG | 24 (9A/4C/1G/10T) | 2/7 |
| Dinoflagellate 1 | WCCGTAGCCATTTTGGCTCAAG | 22 (4A/6C/5G/6T/1W) | 23/23 |
| Nematoda 1 | GGTTTAATTACCCAAGTTTGAGGG | 22 (6A/3C/5G/8T) | 45/53 |
| Nematoda 2 | GGTTTAATTACCCAAGTTTAAG | 22 (7A/3C/4G/8T) | 2/53 |
| Nematoda 3 | GGTTTAACCAGTTAACCAAG | 22 (7A/5C/4G/6T) | 2/53 |
| Nematoda 4 | AGGTATTACCAGATCTAAAAG | 22 (9A/3C/4G/6T) | 1/53 |
| Nematoda 5 | TACCGTCAATTAATTTGAAG | 22 (7A/3C/3G/9T) | 1/53 |
| Nematoda 6 | GTAATAAGAAAACCTCAATAAAG | 22 (13A/2C/3G/4T) | 1/53 |
| Nematoda 7 | GGTTTTACCAGTATCTCAAG | 22 (5A/5C/4G/8T) | 1/53 |
| Platyhelminthe 1 | AACCGTCACGGTTTACTCTTGATTTGTTCATG | 36 (6A/7C/8G/15T) | 3/10 |
| Platyhelminthe 2 | AACCTTAACGGTTCTCTGCCCTGTATATTAGTGCATG | 37 (8A/9C/7G/13T) | 2/10 |
| Platyhelminthe 3 | AACTATAACGGYTCTCTGCCGTATATTAGTGCATG | 37 (9A/7C/8G/12T/1Y) | 2/10 |
| Platyhelminthe 4 | CACCGTTAATCGGTCCTTACCCTTGCACTTTGTATG | 36 (6A/9C/6G/14T/1R) | 3/10 |
| Euglenozoa 1 | AACTAACGCTATATAAGTATCAGTTTCTGTACTTTATTTG | 39 (12A/6C/5G/16T) | 21/54 |
| Euglenozoa 2 | AACTAACGCTATTTATTTGATACAGTTTCTGTACTATATTG | 39 (12A/6C/5G/16T) | 12/54 |
| Euglenozoa 3 | AACTAACGCTATTTATTTAGAACAGTTTCTGTACTATATTG | 39 (13A/6C/5G/15T) | 4/54 |
| Euglenozoa 4 | AACTAAAGTTATTTATTTGATACAGTTTCTGTACTATATTG | 39 (13A/4C/5G/17T) | 2/54 |
| Euglenozoa 5 | AACTAAAGCTTWTATTTAGAACAGTTTCTGTACTATATTG | 39 (13A/5C/5G/15T/1W) | 2/54 |
| Euglenozoa 6 | AACTAAAATTATTTATAATACAGTTTCTGTACTATATTG | 39 (15A/4C/3G/17T) | 1/54 |
| Euglenozoa 7 | AACTAAAGATTTTATTTGTTACAGTTTCTGTACTATATTG | 39 (12A/4C/5G/18T) | 1/54 |
| Euglenozoa 8 | AACTTACGCTATAAAAGTACAGTTTCTGTACTTTATTTG | 39 (12A/7C/5G/15T) | 2/54 |
| Euglenozoa 9 | AACTAACGCTATTTATTTGTTACAGTTTCTGTACTTTATTTG | 39 (10A/6C/5G/18T) | 3/54 |
| Euglenozoa 10 | AACTAACGCTAWAAAAGWTACAGTTTCTGTACTTTATTTG | 39 (13A/6C/5G/13T/2W) | 2/54 |
| Euglenozoa 11 | AACTAACGCATTTTGTGTACAGTTTCTGTACTTTATTTG | 39 (9A/6C/5G/19T) | 1/54 |
| Euglenozoa 12 | AACTAACGCTATATTTGTTACAGTTTCTGTACTTWTATTG | 39 (10A/6C/5G/17T/1W) | 1/54 |
| Euglenozoa 13 | AACTAACGCTATTTCTAGATACAGTTTCTGTACTTTATTTG | 39 (11A/7C/5G/16T) | 1/54 |
| Euglenozoa 14 | AACCAACGATTTAAAAGCTACAGTTTCTGTACTTTATTTG | 39 (13A/7C/5G/14T) | 1/54 |

*Total number of sequences in the phylum/Number of sequences matching consensus.

an additional SLe sequence, which is identical in both species. The two consensus sequences are composed of 24 nucleotides, of which, 10 are thymines. All 23 species of dinoflagellates in the database share an identical 22 nucleotide SLe sequence that is only degenerated in the first position (either A or T), with a balanced nucleotide composition. Notably, this is the only phylum in which, the SLe sequence itself carries the Sm-protein binding site (a T-rich element, which like in *C. elegans* and many other species is a AT₄₋₆G motif). There are 53 sequences from the nematoda phylum within the database. Among these, 45 are identical apart from a repetition of the nucleotide G of variable lengths at the 3' end. Among the remaining eight sequences, two are identical pairs and four are unique, one of which harbors an Sm binding site (the other two sequences present a putative inverted Sm binding motif). Most sequences have 22 nucleotides (apart from the variable repetition of guanines), and the majority are slightly richer in A and T nucleotides. Among the platyhelminthes, there are 10 species from seven different genera represented in the dataset. All SLe have a high percentage of thymines and are 36 or 37 nucleotides long. The sequences can be divided into four

different groups of two or three virtually identical sequences each. Notably, each group contains species of a unique order. A classical Sm-protein binding site was found in one of such groups, and another group presents a putative inverted binding site. Euglenozoans are unique organisms in this context because in these species all transcripts are processed by SLTS. This phylum is represented by 54 sequences of 39 nucleotides from 14 different genera. The sequences can be clustered together in 8 groups of identical sequences plus 6 isolated sequences that are not identical to any other sequence. In summary, the terminal regions (first 6 and last 20 nucleotides) are conserved in all sequences, whereas the central region is variable. All 16 *Leishmania* species share the exact same sequence, whereas the 22 *Trypanosoma* species were divided into 4 groups with identical sequences and 2 isolated sequences. Alternatively, with less stringency, SLe from this phyla could be clustered into three groups according to specific signatures within the last 20 nucleotides: (i) *Trypanosoma* species; (ii) *Leishmania*, *Leptomonas*, *Wallaceina*, and *Chritidia* species (with identical sequences); and (iii) the remaining genus, which presents more diverse sequences (the data presented in



this section are shown in **Figure 3**, **Table 1** and Supplementary Tables).

SLe SEQUENCE COMPARISON AMONG PHyla

When analyzing the EXTENDED dataset of SLe sequences, a high sequence conservation within each phyla was revealed, but a very low conservation among different phyla was found. One interesting feature we highlight as a general tendency is that sequence length is more conserved than sequence composition itself in any given phyla. There are some clear differences between sequences from euglenozoa and platyhelminthes and those from other phyla. Such differences include the sequence length and terminal residue identity. Whereas sequences from all other phyla range from 22 to 25 nucleotides in length, sequences from euglenozoa and platyhelminthes are 39 and 36–37 nucleotides long, respectively. As for the nucleotides at the 3' end, in which SL exon-intron cleavage occurs and the SLe is incorporated into mRNAs, euglenozoa and platyhelminthes have TTG and ATG patterns, respectively, whereas sequences from other phyla have an AAG pattern (except for one nematoda consensus sequence and the sequences from rotifera that end in GAG). Regarding the 5' end, 17 of the 18 consensus SLe sequences from euglenozoa and platyhelminthe present a conserved AAC pattern as the first nucleotide triad. For all other phyla, there is no conservation of nucleotides in the 5' end. Notably, all but five sequences from the CONSENSUS dataset present a TTT triplet, which in all dinoflagellates and three species from the other phyla, are part of the Sm binding site (in five other species it is part of a hypothetical inverted Sm binding motif) (**Table 1**). When considering all sequences in the CONSENSUS database, there is an evident enrichment in adenine (~30% of all nucleotides) and in thymine

(~40% of all nucleotides) in comparison to guanine and cytosine (which together comprise only ~30% of all nucleotides).

SL GENE SEQUENCE COMPARISON AMONG ALL PHyla

Several candidate SL gene sequences were identified through manually analyzing the retrieved sequences from the NCBI database. Within the preliminary dataset, we performed a manual curation to reduce redundancy and exclude false positives, and we also mapped the available SLe sequence to the initial portion of the putative genes of the same species. As a result, 30 sequences remained and were further separated according to phyla and analyzed. Gene length was relatively variable, from 75 to 123 nucleotides; although most sequences were approximately 100 nucleotides long (mean length is 107). The most evident patterns within sequences are the cleavage site in the exon-intron boarder and the presence of the Sm-protein binding site, which is usually in the intronic portion of the gene. Nevertheless, there is some degree of SLe sequence conservation among species from a given phylum, although it is much lower than for the SLe sequence alone. Differing from the exonic portion, the intronic nucleotide composition has an almost equal distribution of nucleotides with 24% adenine, 24% cytosine, 26% guanine, and 26% thymine.

SL GENOMIC LOCATION AND COMPOSITION IN TWO

CAENORHABDITIS SPECIES

Regarding the genomic position, it is already known that most SL genes are located near the 5S ribosomal RNA gene and comprises multiple copies in tandem (apud Hastings, 2005). We have used BLAT to map SLe sequences in the genomes of *C. elegans* and *C. remanei*. Both species have two different SLe sequences in the database, and these sequences are identical between species. One sequence [hereafter named as SLeI and identical to the SL1 sequence from Ross et al. (1995)] is shared among the majority of the nematoda species (45 of the 53 sequences from this phylum), and the other [hereafter named SLeII, closely related to the SL2 first described by Huang and Hirsh (1989) and identical to SLF as described by Ross et al. (1995)] is only common to these two *Caenorhabditis* species in our database.

In *C. elegans*, the SLeI sequence is repeated 13 times in chromosome V, twice in chromosome I and once in chromosome III (thus totaling 16 copies). The SLeII sequence is repeated twice in chromosome I and has only one copy in chromosomes III and IV (thus totaling four copies). The genome of *C. remanei* is more highly populated by SLeI sequences, which are represented by 135 copies. The SLeII sequence on the other hand seems to have only nine copies. The karyotype is not available in BLAT for the later species and therefore, it was not possible to map sequences onto chromosomes. Notably, the SLeI sequence, which is widespread among nematoda species, is always more represented in the genomes of these two species (16 vs. 4 copies in *C. elegans* and 135 vs. 9 copies in *C. remanei*).

Regarding the intronic portion of the putative SL gene sequences, we have analyzed all occurrences of the *C. elegans* SLeI in chromosomes V, III, and I. We observed that 10 repetitions have identical introns (all in chromosome V) and that the other putative SLeI gene sequences are very divergent from one another. Notably, these 10 identical sequences are the only

ones to present classical Sm-protein binding site. When evaluating the intronic portion of the four putative genes with the SLeII sequence in *C. elegans*, a high similarity is observed, although the sequences are not identical. These sequences also present Sm binding sites, although three of these have five thymine within the repetition, whereas the additional sequence has four. We have noticed a conserved GTTAG pattern in the four putative SLeII gene sequences that is also present at a different position in the 10 identical sequences and is absent in all other six putative SLeI gene sequences (two other short patterns—ACAA and GGAA—are also present in the 14 sequences, but are not exclusive to these sequences). Among the nine sequences of the putative SLeII genes in *C. remanei*, eight are identical (apart from one substitution in two sequences) and the other contains approximately 10% substitutions. All nine sequences contain classical Sm-protein binding sites. In addition, these sequences are closely related to the putative SLeII gene sequences of *C. elegans*, although not identical. When analyzing the set of 135 putative SLeII gene sequences in *C. remanei*, several clusters of highly similar or identical sequences are observed. One important observation is the lack of Sm binding sites in most of the sequences. Only 27 sequences bear Sm binding site motifs. These can be divided into groups of identical or nearly identical sequence; one of which is also closely related to the main group of *C. elegans* putative SLeI gene sequences. We have reconstructed a phylogenetic tree with all the SL sequences from both species (data not shown) and observed a more randomized distribution of SL genes not bearing Sm binding sites in comparison to sequences that contain this motif. We then narrowed our analysis to consider only Sm binding-containing sequences, and the corresponding tree is shown in the supplementary material (Supplementary Figure 1). The tree is divided into three main groups: (i) a group of SLeI gene sequences containing all *C. elegans* and approximately half the *C. remanei* SLeI gene sequences, (ii) a group of all SLeII gene sequences and (iii) a more distant group of *C. remanei* SLeI gene sequences. Each group can be further clustered into smaller groups with closely related sequences.

SL GENE STRUCTURAL COMPARISON

We have assigned secondary structures to all 30 SL gene sequences with RNAfold from the Vienna package. As a result of this structural analysis, a tendency for SL RNAs to form a Y shaped molecule was observed (Figure 4). The topology is the result of three stem-loops and a branch point and seems to be conserved in nearly all analyzed species, although the branch point position and stem-loop length are variable. The average free energy (ΔG) value for gene structures in fixed secondary structures was -30.6 Kcal/mol (ranging from -14.4 to -51.80 Kcal/mol).

ANALYSIS OF TRANSCRIPTS TRANS-SPLICED TO SPECIFIC SLe

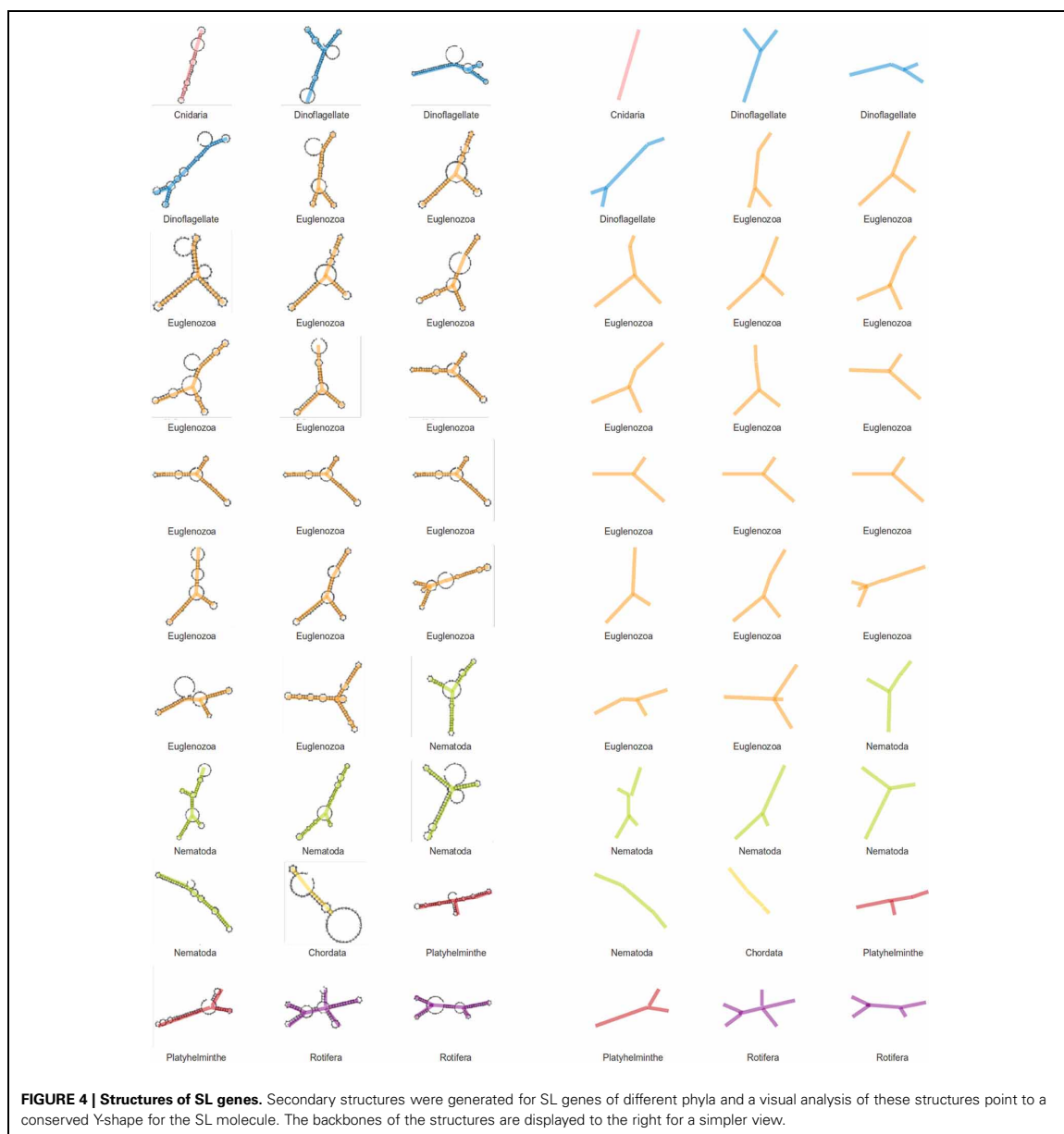
To assess whether the sets of trans-spliced transcripts from different organisms harboring the same SLe sequence are similar, we conducted BLAST searches in the nr/nt database using two different sequences as queries: (i) a sequence shared by different dinoflagellate species and (ii) a sequence shared by different nematoda species (other than *C. elegans*). For each SLe, we have allowed retrieval of up to 500 transcripts from different species

of each phylum (dinoflagellate and nematoda). We have excluded *C. elegans* from our survey because in this organism the addition of different SLe sequences in different transcripts has been investigated by high throughput sequencing (Allen et al., 2011) and could not be included here without introducing a substantial bias to the analyzed data. Among all retrieved dinoflagellate sequences, 133 remained after false positive exclusion and redundancy reduction. From these, 63 (over 47%) transcripts were shared by more than one species (totaling 30 different transcripts, from which many code for ribosomal proteins). The remaining 70 are specific to only one species. From the nematoda transcripts, after manual curation, we analyzed 158 transcripts from which 54 (over 34%) are shared by different species. We then decided to analyze whether in one given species two different SLe sequences would regulate different sets of transcripts. To this end, we used BLAST and searched the nr/nt database for transcripts from *H. vulgaris* bearing any of the two different SLe sequences. As a result we have identified 30 transcripts trans-spliced with one sequence and 13 with the other, none of which were related to both SLe sequences.

ANALYSES OF TRANS-SPLICED TRANSCRIPTS FROM ALL SPECIES

In a similar context, a more overall analysis was performed with the unique sequences of the EXTENDED database of SLe sequences as queries for BLAST searches (with the blastn program) within the nr/nt database. Search results and manual curation resulted in a final set of 455 transcripts (Figure 5 and Supplementary Table 4), among which five were annotated as “alternatively spliced.” From this set, we have previously excluded transcripts from euglenozoa species because in these organisms, SLe addition is ubiquitously used to solve polycistronic transcripts. Among these 455 transcripts, 237 are present in only one species and 218 are shared by more than one species (totaling 60 unique sequences). Additionally, 138 are common to species from different phyla (totaling 32 unique sequences). Enolase, calmodulin, glutathione S-transferase, ATP synthase subunits, cyclins, eukaryotic translation initiation factors, superoxide dismutase, ras-related proteins, and ribosomal proteins are among the most ubiquitous trans-spliced proteins, as these are found in species of at least three different phyla.

We have automatically assigned GoSlim terms to all transcripts and the results revealed no clear bias to any specific gene category. This was true for the entire set of transcripts and also for each individual phylum (data not shown). Because this protocol was not curated and has classified transcripts into generic and less informative classes, we have decided to manually annotate and classify all transcripts aiming to achieve a more specific, reliable and informative result. We have therefore separated transcripts according to their main biological function, generating a total of 30 different functional classes with different representations. The result was unexpected and revealed unique classes composed of transcripts that seem to more frequently undergo trans-spliced in each phylum, although this was not true for the entire set of transcripts, in which no specific category seemed to be dominant (Figure 5 and Supplementary Tables 4, 5).

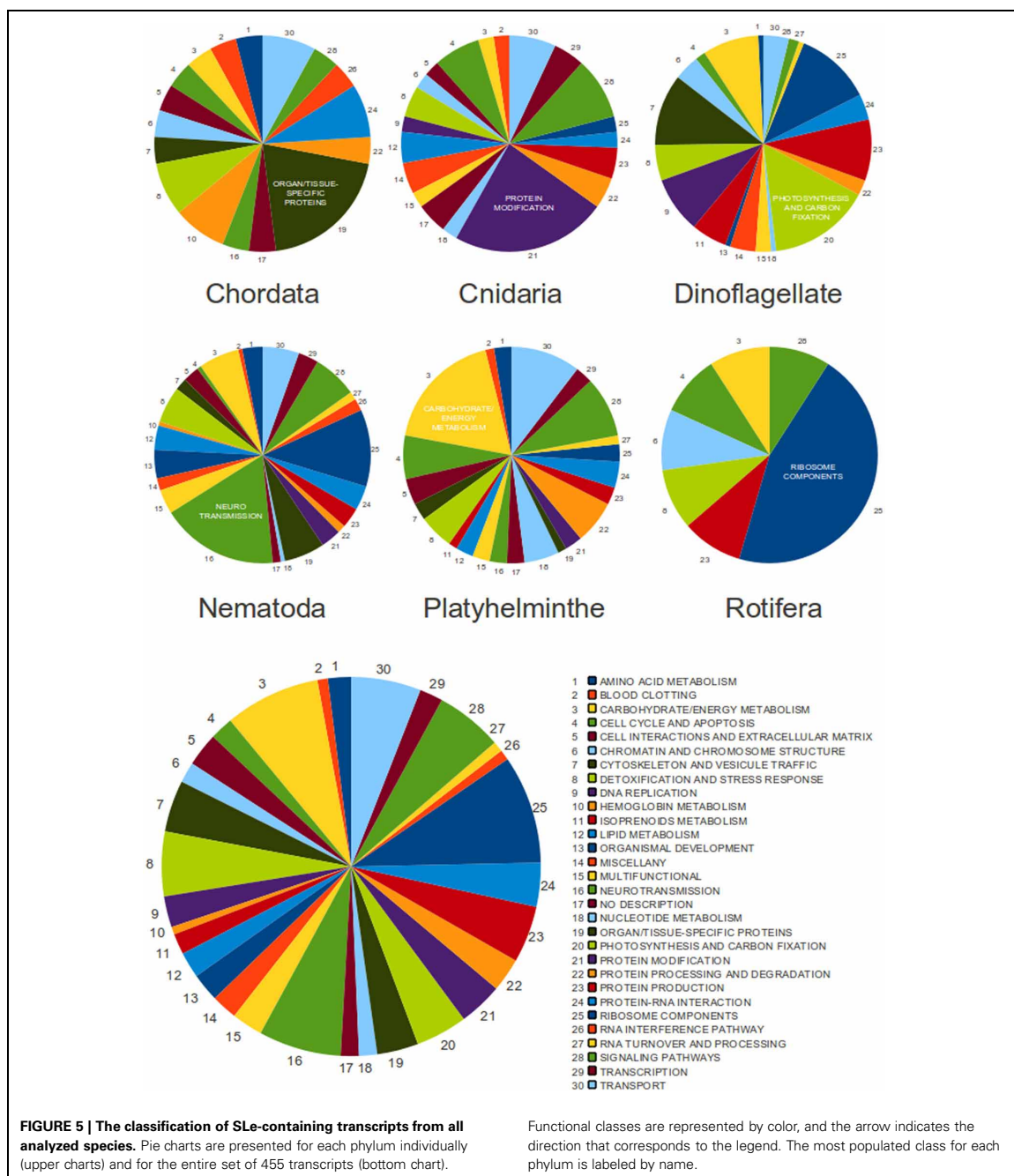


In parallel, using data available from the work of Protasio et al. (2012), we have generated a set of 1411 SLe-containing transcripts from *S. mansoni*. Preliminary analyses after GO assignment and GoSlim retrieval revealed no clear bias regarding the biological processes under the control of the SLTS mechanism, although the number of transcripts in a few biological processes was higher, such as whole-cell functions, metabolic processes, and organismal development (data not shown).

DISCUSSION

ARE SLe SEQUENCES CONSERVED AMONG DIFFERENT SPECIES?

One of the most discussed topics related to the SLTS mechanism is its evolutionary origin. There is previous evidence for a unique origin in a common ancestor but there is also evidence for multiple unrelated origins (Nilsen, 2001; Stover and Steele, 2001; Hastings, 2005). From the 157 sequences comprising the described EXTENDED database, it is possible to observe few



features that could universally define a SLe. Almost all sequences have a conserved WWG at the 3' end and a TTT pattern. As for the latter, we suggest it may be the vestige of a Sm binding site motif that was once present within the SLe and was further

lost or transferred to the intronic portion. Notably, there is a high sequence conservation rate within phyla, specifically to the level of subfamilies, although only the two aforementioned features are conserved among unrelated phyla. In addition to the

sequence itself, the SLe length is even more conserved within phyla, indicating this as a crucial characteristic of the molecule.

There are important characteristics that differentiate euglenozoa and platyhelminth consensus SLe sequences from other sequences, such as the length and composition of the 5' and 3' ends. This could reflect a distinct role of the SLTS mechanism in such species. Regarding trypanosomatids, the SLTS mechanism has a crucial role because in these organisms large regions of the genome are transcribed as polycistronic units and SLe incorporation is crucial for their resolution into monocistronic units. In platyhelminthes, the mechanism is supposed to be acting as an expression regulator for approximately 10% of all genes. A unique feature of the SLe sequence in this phylum is the presence of an ATG as the last nucleotide triad. This could account for an intrinsic start codon, that could generate an alternative open reading frame (ORF) in which the SLe insertion occurs within the transcript sequence, thus giving rise to alternative forms of the resulting protein. Nevertheless, the relationship between sequences from platyhelminthes and euglenozoans is not clear nor it is phylogenetically expected because these two phyla come from different eukaryotic kingdoms (animalia and excavata, respectively).

ARE SL GENE SEQUENCES CONSERVED AMONG DIFFERENT PHYLA?

Similarity between SL gene sequences from different phyla is less evident, but some features are conserved. Specifically, the last nucleotide of all SLe sequences is a guanine (G). This is because the cleavage site between the exon and intron is a highly conserved GGTA motif, from which the first nucleotide (G) is the last nucleotide of the exon and the other three (GTA) are in the beginning of the intron. Another conserved motif is the binding site for the Sm protein (usually a AT₄₋₆G motif). This conserved binding site is usually present in the intron, with the exception of the dinoflagellate sequences (and a few sequences from other phyla), in which the Sm binding site is located in the exon. The consequences of this exonic localization of the Sm binding site are not completely understood. Notably, intronic nucleotide composition is homogeneous (not biased for a specific nucleotide type), indicating a tendency for a lack of selective pressure in this region, except for the presence of the Sm binding site. From all analyzed phyla, euglenozoa species present the longest SLe sequences, with 1.5 times the size as compared to SLe from other species (which average 23 nucleotides). This discrepancy may indicate an independent origin of SLTS in euglenozoa, which may in turn be related to the unique transcription strategy adopted by such organisms and the role of SLe insertion in the post-transcriptional processing of polycistronic transcripts.

HOW CONSERVED ARE PUTATIVE SL GENES IN SPECIES OF CAENORHABDITIS?

There are two different SLe sequences in each of the two *Caenorhabditis* species in the EXTENDED dataset (we have internally named these as SLeI and SLeII). These sequences were mapped with BLAT on the genomes of the respective organisms, and the retrieved sequences of the putative SL genes (the SLe sequence plus 100 nucleotides downstream) were analyzed. The

results show a differential abundance of each SLe in the genomes, with the SLeI sequence being the most abundant in both species. Notably, this SLe is shared with the majority of the nematode species represented in the database. In each species, regarding sequence similarity, putative genes of identical SLeS seem to be more related to one another than to putative genes of other SLe. Although not all sequences are identical for a given SLe, sequence alignments present groups of closely related introns.

In their 1987 article, Krause and Hirsh (1987) reported the existence of more than 100 SLeI genes in *C. elegans*, which is much higher than we have observed. This discrepancy can be explained by the methodology used for SL gene identification, which considered sequences that were a 90% match to the SLe (20 out of 22 nucleotides) and was performed by Southern blot analysis. By contrast, in this study, we considered only 100% matches in the genome, thus restricting the number of retrieved sequences. Unfortunately, because the *C. elegans* genome was not available in 1987, we cannot compare our results to the previously published results.

When analyzing putative intronic sequences of different SLe, sequence conservation is more clear among species. There are only 10 SLeI gene sequences in *C. elegans* bearing the Sm binding site, and these are closely related to one group of SLeI sequences of *C. remanei*, in which only 27 sequences contain Sm binding sites. All of the remaining sequences (not bearing the Sm binding site) do not seem to be related to one another. The scenario is simpler for the SLeII sequences because there are fewer in both species. All four *C. elegans* SLeII sequences contain Sm binding sites and are similar. The nine *C. remanei* SLeII sequences are nearly identical, contain Sm binding sites and are closely related to the *C. elegans* SLeII sequences.

When we analyze the phylogenetic tree in the supplementary material (Supplementary Figure 1), we see two main groups: one comprising approximately half *C. remanei* SLeI gene sequences and another containing all other sequences. The latter is further divided into two groups, one with the SLeII gene sequences and another with SLeI gene sequences. This seems to indicate the existence of a more ancient SLe from which the SLeI and SLeII genes have arisen by duplication, most likely before speciation between *C. elegans* and *C. remanei* (given the similarity among intronic sequences of both species). The second group of SLeI genes from the later species may have arisen by duplication after speciation or (most likely given the divergence among sequences in this group) may be a result of duplication from a common ancestor prior to speciation and was then lost by the former species. Taken together, these results show that the orthologous SLe gene sequences (identical SLeS in different species) are more related to one another than to paralogous genes (other SLe in identical species). This most likely indicates that divergence between SLeI and SLeII took place before speciation.

ARE DIFFERENT SLe SEQUENCES FROM ONE UNIQUE SPECIES INCORPORATED INTO DIFFERENT TRANSCRIPTS?

This study observed that a given species may have more than one SLe sequence, supporting the literature [as in the classic case of *C. elegans* first reported in Huang and Hirsh (1989)].

This could account for the differential expression of transcripts or the production of different protein repertoires under certain environmental conditions or developmental stages. We have identified two different SLe sequences from *Hydra vulgaris* and analyzed the set of transcripts related to each in public databases. Although this cannot be considered a fully conclusive analysis, it gives an indication of the possible roles for different SLe sequences. Considering the two distinct SLe from *H. vulgaris*, whereas one was found in 30 annotated transcripts, the other was found in less than half of these (only 13 transcripts). There was no superposition of the two sub-groups, indicating that each SLe sequence could be trans-spliced to a distinct set of transcripts. Notably, the SLe that was added to a higher number of transcripts is the most conserved when compared to other species from identical genera. This result is in agreement with the recent work of Allen et al. (2011), where the authors conclude that in *C. elegans* the trans-splicing to SLeI or SLeII are mechanistically separate and distinct phenomena. On the other hand, the fact that in the consulted database both SLe sequences were related to the same number of transcripts is not expected given the higher prevalence (of >80%) of SLeI trans-splicing in *C. elegans* as reported on Allen et al. (2011).

ARE IDENTICAL SLe SEQUENCES IN DIFFERENT SPECIES INCORPORATED INTO DIFFERENT TRANSCRIPTS?

The analysis of transcripts bearing SLe sequences that are shared by different organisms from a same phylum indicated that these SLe sequences may control similar transcript repertoires. The first SLe sequence that was analyzed is the one conserved in all dinoflagellate species in the database. As a result, from the retrieved matches, 133 transcripts remained after curation with almost half (63) shared by more than one species. The other analyzed SLe is common to several nematode species. In this case, from the retrieved matches, 158 transcripts remained after manual curation, 54 of which are shared by different species from the same phylum. In the latter case, we have excluded *C. elegans* from our search, since in this organism it is already known that SLeI and SLeII have different roles and are therefore added to different sets of transcripts (Allen et al., 2011). Taken together, these results may indicate that identical SLe sequences in different species from the same phylum are incorporated into the same transcripts. The observation that not all transcripts are common to all species harboring identical SLe sequences could be a result of the restricted number of annotated transcripts deposited in publicly available databases. Sets of transcripts from different species that are regulated by the same SLe sequences, have between 34 and 47% common elements, potentially indicating an overall tendency for each SLe to regulate specific transcripts, regardless of the species. If this is true, then a given SLe will always be related to a given set of transcripts in all species (except maybe for species-specific genes), and therefore, different SLe from identical species would regulate different sets of transcripts. Indeed, Allen and collaborators have shown that, in *C. elegans*, spliced leader sequence SLeI is added either to monocistronic genes or to the first gene of a polycistron, while the SLeII sequence is added to internal genes of a polycistron. These two trans-splicing events are therefore mechanically unrelated

and may be functionally different. This observation alone could account for the lack of overlap between the sets of transcripts regulated by each SLe sequence in this species and maybe other species that carry genes in an operon structure. Additionally, the authors show that some genes can actually be trans-spliced to both SLeI and SLeII and these are internal genes in the polycistron for which there are independent promoters. Results show that, for genes preferably trans-spliced to SLeII sequences, some level of SLeI addition may still be observed. According to the authors, this may be due to the 10-fold excess of SLeI in the cells in comparison to SLeII or it may be that SLeI is added to transcripts in a constitutive frequency, while SLeII addition could be more specific. Unfortunately, there are limited data on trans-spliced transcripts of other species with multiple SLe sequences and therefore it was not possible here to reproduce the same observation in other species, apart from a limited analysis of *H. vulgaris* transcripts, in which we have found 30 SLeI-containing transcripts and 13 SLeII-containing transcripts, with no overlap among these sets.

ARE TRANSCRIPTS REGULATED BY THE TRANS-SPICING MECHANISM COMMON TO DIFFERENT SPECIES AND PHYLA?

We have analyzed over 450 transcripts bearing SLe sequences from all species in our EXTENDED database except euglenozoans (because this species add SLe to all transcripts). Of those species, almost half (48%) contain transcripts shared by at least two different species and almost one third (30%) contain transcripts of species from different phyla (136 transcripts, representing 32 unique sequences). Among the 32 transcripts represented in multiple phyla, most are conserved in all eukaryotes, and the related proteins perform basic functions in the cell, including involvement in ribosomal activity, cell structure, ATP synthesis, glucose metabolism, protein folding, antioxidant defense, DNA replication and translation. This may reflect a tendency of the SLTS mechanism to regulate ancient and conserved functions.

To thoroughly investigate the relationship between trans-spliced transcripts from different species and phyla, we have manually annotated and classified all 455 transcripts according to their biological function. In a surprising result, we have identified a dominant class of transcripts in each phylum, with no overlap among phyla. This could indicate that each phylum may have a preference to add the SLe to a specific gene category. We did not perform analyses to investigate if this is true at the species level, although previous data from studies in *S. mansoni* suggests it is not true (Mourão et al., 2013 and our previously described GO annotation of SLe-containing *S. mansoni* transcripts). There is no overall tendency when we observe the set of transcripts from all phyla. This is expected because each phylum has a bias to a different class. Notably, we have not identified genes related to host-parasite interactions undergoing SLTS in the parasitic organisms. We consider this to be an expected observation if one considers that the SLTS mechanism was derived early in the eukaryotic lineage. On the other hand, photosynthesis is the major class of trans-spliced transcripts for the dinoflagellate species (from which only two *Perkinsus* species are not photosynthetic) and this is a phylum-specific class in our study

because no species out of this phylum perform photosynthesis.

We cannot confirm that the observed bias to specific functional categories is not expected for a given phylum because we did not perform the same annotation and classification protocol for the entire set of transcripts for each species. For example, it is reasonable to expect that energy metabolism would be a broadly represented class in the transcriptome of most organisms and, accordingly, it is the major class among platyhelminth transcripts that undergo trans-splicing. However, in nematodes there is a bias for transcripts involved with neurotransmission to undergo trans-splicing. This is not an expected result. Despite the evidence above, there remains a limited number of SLe-containing transcripts in public databases, therefore, no final conclusions can be reached.

ARE THE TRANSCRIPTS UNDERGOING TRANS-SPICING IN A GIVEN ORGANISM RELATED TO ONE ANOTHER?

As we have observed in a previous study regarding the SLTS mechanism in *S. mansoni* (Mourão et al., 2013), we found no bias for any specific functional gene category to have transcripts undergoing SLTS. In a preliminary survey using a recently published *S. mansoni* SLe-containing cDNA dataset (Protasio et al., 2012), we have used GO annotations to assess whether specific gene categories could be found among the 1411 SLe-containing transcripts. The most abundant biological processes are whole-cell processes (as cell differentiation, cell cycle, cell death, cell communication, cell proliferation, cell growth, cell recognition, and cellular homeostasis), cellular component organization, transport, response to stimulus, signaling, protein function, protein expression, metabolic processes and organismal development, notably multicellular organismal development (which is the class with the highest number of related transcripts).

Taken these observations together, we can conclude that the SLTS mechanism does not regulate any specific biological process category. Nevertheless, some categories are slightly more represented than others and those categories are crucial processes for the metabolism of the cell and the organism. We can then speculate that the SLTS mechanism is of fundamental importance and that the disruption the mechanism should lead to serious consequences for the organism. This represents a notable result that places the SLTS mechanism in an important place for organismal development and survival.

ARE SL RNA STRUCTURES CONSERVED AMONG DIFFERENT SEQUENCES?

A structural RNA analysis was performed to identify possible structural conservation despite the lack of sequence similarity across phyla. Structures were generated for the 30 SL gene sequences mentioned previously, and the results do not show a clear conservation, although some features may be observed. Because SL gene sequences vary in length, structure complexity is also diverse. An almost universal topological identity can be observed among the different SL structures, with few exceptions. SL RNA structures have three stem loops and a bifurcation point. This topology can be defined as a Y shaped structure. Although

the bifurcation point location and stem-loop length may vary, the overall topology may be the only common aspect providing the SL RNA structure, which is a feature that is shared by all species.

CONCLUSIONS

In the course of this study, we have investigated several characteristics of the SLTS mechanism in different species from various phyla. Analyses of the conserved features revealed a close relationship between SLe sequences from the same phylum. However, sequences from different phyla do not share many common features. SL structures show a certain level of topological conservation across phyla with an overall Y-shaped structure.

Transcripts controlled by the SLTS mechanism are, to a certain extent, shared among different species and organisms from different phyla, although different biological functions are the focus of SLTS in different phyla. Keeling et al. (2005) have published a consistent classification for eukaryotes that includes five different “supergroups,” namely excavates, rhizaria, unikonts, chromalveolates, and plantae. Those phyla in which the SLTS mechanism was previously characterized are located in three of such supergroups: unikonts (rotifera, chordata, cnidaria, nematode, and platyhelminth), excavates (euglenozoa), and chromalveolates (dinoflagellates). The widespread presence of the mechanism could place a unique origin in an ancestor common to all eukaryotes (or at least to these groups). The outcome of this hypothesis, if supported, is the previous existence of the SLTS mechanism in all eukaryotes, although the progressive loss of relative importance reduces the chance of identifying SLe-containing transcripts in more complex organisms. It may even be true that some species have completely lost the SLTS mechanism when more robust regulatory mechanisms emerged. Despite speculation, no final answer can be reached and this is partly due to the restricted data currently available. Although we have covered here all species in which SLe sequences are annotated and deposited at NCBI, more data (specifically of species from other phyla) is needed to more precisely place the origin of this mechanism among eukaryotes.

Notably, there is also a clear bias in this study regarding the demonstration of the SLTS mechanism in new species. Each SLe sequence in the extended database has to be related to at least one sequence in the seed database. This indicates that we will most likely not incorporate new phyla into this database unless a new methodological approach is used. In this study, we have initiated an important characterization of the SLe identity, which can be further used to unravel the presence of the SLTS mechanism in new phyla. We reported that (i) all SLe sequences have a WWG pattern at the 3' end; (ii) a Sm binding site is always present, either in the exon or in the intron of the SL gene; (iii) the sequences are 22–25 or 36–39 nucleotides long; (iv) a SLe TTT pattern is present in almost all sequences; and (v) the RNA structure may be Y-shaped, bearing three stem-loops and a bifurcation point.

Another important contribution of this study is the observation that different SLe sequences in a given species control different transcripts and identical SLe sequences in different

species control identical transcripts. We hypothesize that each SLe sequence is always related to a given set of transcripts and, therefore, the expression of the respective protein repertoires could be switched on and off according to the presence of a SLe sequence. If this is correct, then the SLTS mechanism could be controlled by environmental changes and lead to translation of several processed transcripts, giving rise to a specific response. This hypothesis can be tested in further studies by activating different SLe sequences at different times and observing the phenotypic effects. Finally, we have shown that after annotation and classification of SLe-containing transcripts from all phyla that the SLTS seems to be directed to specific gene categories in each phylum.

AUTHOR CONTRIBUTIONS

Mainá Bitar retrieved the data, designed, and conducted all of the *in silico* experiments and analyses and wrote the manuscript. Mariana Boroni participated in the data retrieval, the *in silico*

experiments and in manuscript preparation. Andréa M. Macedo and Carlos R. Machado were involved in discussions, contributed with expert insights and guidance and reviewed the manuscript. Glória R. Franco designed all of the *in silico* experiments and analyses, contributed expert insights and guidance and reviewed the manuscript.

ACKNOWLEDGMENTS

The authors would like to acknowledge the valuable contributions of Dr. Priscila Grynberg to this work. Additionally, the financial support of the Brazilian funding agencies CAPES, CNPq, and FAPEMIG was of utmost importance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://www.frontiersin.org/journal/10.3389/fgene.2013.00199/abstract>

REFERENCES

- Allen, M. A., Hillier, L. W., Waterston, R. H., and Blumenthal, T. (2011). A global analysis of *C. elegans* trans-splicing. *Genome Res.* 21, 255–264. doi: 10.1101/gr.113811.110
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2
- Boothroyd, J. C., and Cross, G. A. (1982). Transcripts coding for variant surface glycoproteins of *Trypanosoma brucei* have a short, identical exon at their 5' end. *Gene* 20, 281–289. doi: 10.1016/0378-1119(82)90046-4
- Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res.* 14, 1188–1190. doi: 10.1101/gr.849004
- Gilbert, W. (1978). Why genes in pieces. *Nature* 271, 501. doi: 10.1038/271501a0
- Hastings, K. E. M. (2005). SL trans-splicing: easy come or easy go. *Trends Genet.* 21, 240–247. doi: 10.1016/j.tig.2005.02.005
- Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic Acids Res.* 31, 3429–3431. doi: 10.1093/nar/gkg599
- Huang, X. Y., and Hirsh, D. (1989). A second trans-spliced RNA leader sequence in the nematode *Caenorhabditis elegans*. *Proc Natl. Acad. Sci. U.S.A.* 86, 8640–8644. doi: 10.1073/pnas.86.22.8640
- Keeling, P. J., Burger, G., Durnford, D. G., Lang, B. F., Lee, R. W., Pearlman, R. E., et al. (2005). The tree of eukaryotes. *Trends Ecol. Evol.* 20, 670–676. doi: 10.1016/j.tree.2005.09.005
- Kent, W. J. (2002). BLAT—the BLAST-like alignment tool. *Genome Res.* 12, 656–664. doi: 10.1101/gr.229202
- Krause, M., and Hirsh, D. (1987). A trans-spliced leader sequence on actin mRNA in *C. elegans*. *Cell* 49, 753–761. doi: 10.1016/0092-8674(87)90613-1
- Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., et al. (2007). Clustal W and clustal X version 2.0. *Bioinformatics* 23, 2947–2948. doi: 10.1093/bioinformatics/btm404
- Lasda, E. L., and Blumenthal, T. (2011). Trans-splicing. *Wiley Interdiscip. Rev. RNA* 2, 417–434. doi: 10.1002/wrna.71
- Liang, X., Haritan, A., Uliel, S., and Michaeli, S. (2003). Trans and cis splicing in trypanosomatids: mechanism, factors, and regulation. *Eukaryot. Cell* 2, 830–840. doi: 10.1128/EC.2.5.830-840.2003
- Lidie, K. B., and Van Dolah, F. M. (2007). Spliced leader RNA-mediated trans-splicing in a Dinoflagellate, *Karenia brevis*. *J. Eukaryot. Microbiol.* 54, 427–435. doi: 10.1111/j.1550-7408.2007.00282.x
- Logan-Klumpler, F. J., De Silva, N., Boehme, U., Rogers, M. B., Velarde, G., McQuillan, J. A., et al. (2012). GeneDB—an annotation database for pathogens. *Nucleic Acids Res.* 40, D98–D108. doi: 10.1093/nar/gkr1032
- Matsumoto, J., Dewar, K., Wasserscheid, J., Wiley, G. B., Macmil, S. L., Roe, B. A., et al. (2010). High-throughput sequence analysis of *Ciona intestinalis* SL trans-spliced mRNAs: Alternative expression modes and gene function correlates. *Genome Res.* 20, 636–645. doi: 10.1101/gr.100271.109
- McCarthy, F. M., Gresham, C. R., Buza, T. J., Chouvarine, P., Pillai, L. R., Kumar, R., et al. (2010). AgBase: supporting functional modeling in agricultural organisms. *Nucleic Acids Res.* 39 (Suppl. 1), D497–D506. doi: 10.1093/nar/gkq1115
- McCarthy, F. M., Wang, N., Magee, G. B., Nanduri, B., Lawrence, M. L., Camon, E. B., et al. (2006). AgBase: a functional genomics resource for agriculture. *BMC Genomics* 7:229. doi: 10.1186/1471-2164-7-229
- Mourão, M. M., Bitar, M., Lobo, F. P., Peconick, A. P., Grynberg, P., Prodócimi, F., et al. (2013). A directed approach for the identification of transcripts harbouring the spliced leader sequence and the effect of trans-splicing knockdown in *Schistosoma mansoni*. *Mem. Inst. Oswaldo Cruz.* 108, 707–717. doi: 10.1590/0074-0276108062013006
- Nilsen, T. W. (2001). Evolutionary origin of SL-addition trans-splicing: still an enigma. *Trends Genet.* 17, 678–680. doi: 10.1016/S0168-9525(01)02499-4
- Nilsson, D., Gunasekera, K., Mani, J., Osteras, M., Farinelli, L., Baerlocher, L., et al. (2010). Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of *Trypanosoma brucei*. *PLoS Pathog.* 6:e1001037. doi: 10.1371/journal.ppat.1001037
- Padgett, R. A., Grabowski, P. J., Konarska, M. M., Seiler, S., and Sharp, P. A. (1986). Splicing of messenger RNA precursors. *Annu. Rev. Biochem.* 55, 1119–1150. doi: 10.1146/annurev.bi.55.070186.005351
- Pouchkina-Stantcheva, N. N., and Tunnacliffe, A. (2005). Spliced leader RNA-mediated trans-splicing in PHYLUM ROTIFERA. *Mol. Biol. Evol.* 22, 1482–1489. doi: 10.1093/molbev/msi139
- Preußer, C., Jaé, N., and Bindereif, A. (2012). mRNA splicing in trypanosomes. *Int. J. Med. Microbiol.* 302, 221–224. doi: 10.1016/j.ijmm.2012.07.004
- 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
- Rajkovic, A., Davis, R. E., Simonsen, J. N., and Rottman, F. M. (1990). A spliced leader is present on a subset of mRNAs from the human parasite *Schistosoma mansoni*. *Proc. Natl. Acad. Sci. U.S.A.* 87, 8879–8883. doi: 10.1073/pnas.87.22.8879
- Ross, L. H., Freedman, J. H., and Rubin, C. S. (1995). Structure and expression of novel spliced leader RNA genes in *Caenorhabditis elegans*. *J. Biol. Chem.* 270, 22066–22075. doi: 10.1074/jbc.270.37.22066
- Sather, S., and Agabian, N. (1985). A 5' spliced leader is added in trans to both alpha- and beta-tubulin transcripts in *Trypanosoma brucei*. *Proc. Natl. Acad. Sci. U.S.A.* 82, 5695–5699.
- Stover, N. A., Kaye, M. S., and Cavalcanti, A. R. (2006). Spliced leader trans-splicing.

- Curr. Biol.* 16, R8–R9. doi: 10.1016/j.cub.2005.12.019
- Stover, N. A., and Steele, R. E. (2001). Trans-spliced leader addition to mRNAs in a cnidarian. *Proc. Natl. Acad. Sci. U.S.A.* 98, 5693–5698. doi: 10.1073/pnas.101049998
- Tessier, L. H., Keller, M., Chan, R. L., Fournier, R., Weil, J. H., and Imbault, P. (1991). Short leader sequences may be transferred from small RNAs to pre-mature mRNAs by trans-splicing in *Euglena*. *EMBO J.* 10, 2621–2625.
- Vandenbergh, A. E., Meedel, T. H., and Hastings, K. E. M. (2001). mRNA 5'-leader trans-splicing in the chordates. *Genes Dev.* 15, 294–303. doi: 10.1101/gad.865401
- 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.
- Received: 14 July 2013; accepted: 19 September 2013; published online: 11 October 2013.
- Citation: Bitar M, Boroni M, Macedo AM, Machado CR and Franco GR (2013) The spliced leader trans-splicing mechanism in different organisms: molecular details and possible biological roles. *Front. Genet.* 4:199. doi: 10.3389/fgene.2013.00199
- This article was submitted to *Bioinformatics and Computational Biology*, a section of the journal *Frontiers in Genetics*.
- Copyright © 2013 Bitar, Boroni, Macedo, Machado and Franco. 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) or licensor 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.

9. ANEXO 3 : EVALUATION OF THE *SCHISTOSOMA MANSONI* Y-BOX-BINDING
PROTEIN (SMYB1) POTENTIAL AS A VACCINE CANDIDATE AGAINST
SCHISTOSOMIASIS



Evaluation of the *Schistosoma mansoni* Y-box-binding protein (SMYB1) potential as a vaccine candidate against schistosomiasis

Sílvia R. C. Dias¹, Mariana Boroni¹, Elizângela A. Rocha¹, Thomaz L. Dias¹, Daniela de Laet Souza¹, Fabrício M. S. Oliveira², Mainá Bitar¹, Andrea M. Macedo¹, Carlos R. Machado¹, Marcelo V. Caliar² and Glória R. Franco^{1*}

¹ Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

² Departamento de Patologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Edited by:

Arnon Dias Jurberg, Oswaldo Cruz Institute (IOC)/Oswaldo Cruz Foundation (Fiocruz), Brazil

Reviewed by:

Sheila Donnelly, University of Technology Sydney, Australia
Jose Tort, Universidad de la Republica, Uruguay

*Correspondence:

Glória R. Franco, Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Avenida Antônio Carlos 6627, Pampulha, Belo Horizonte, MG 31270-010, Brazil
e-mail: gfrancoufmg@gmail.com

Schistosomiasis is a neglected tropical disease, and after malaria, is the second most important tropical disease in public health. A vaccine that reduces parasitemia is desirable to achieve mass treatment with a low cost. Although potential antigens have been identified and tested in clinical trials, no effective vaccine against schistosomiasis is available. Y-box-binding proteins (YBPs) regulate gene expression and participate in a variety of cellular processes, including transcriptional and translational regulation, DNA repair, cellular proliferation, drug resistance, and stress responses. The *Schistosoma mansoni* ortholog of the human YB-1, SMYB1, is expressed in all stages of the parasite life cycle. Although SMYB1 binds to DNA or RNA oligonucleotides, immunohistochemistry assays demonstrated that it is primarily localized in the cytoplasm of parasite cells. In addition, SMYB1 interacts with a protein involved in mRNA processing, suggesting that SMYB1 functions in the turnover, transport, and/or stabilization of RNA molecules during post-transcriptional gene regulation. Here we report the potential of SMYB1 as a vaccine candidate. We demonstrate that recombinant SMYB1 stimulates the production of high levels of specific IgG1 antibodies in a mouse model. The observed levels of specific IgG1 and IgG2a antibodies indicate an actual protection against cercariae challenge. Animals immunized with rSMYB1 exhibited a 26% reduction in adult worm burden and a 28% reduction in eggs retained in the liver. Although proteins from the worm tegument are considered optimal targets for vaccine development, this study demonstrates that unexposed cytoplasmic proteins can reduce the load of intestinal worms and the number of eggs retained in the liver.

Keywords: *Schistosoma mansoni*, Y-box-binding protein 1, SMYB1, cytoplasmic antigen, vaccine candidates

INTRODUCTION

Schistosomiasis is the second most important neglected tropical disease causing approximately 280,000 deaths annually (King et al., 2006; Steinmann et al., 2006; Hotez et al., 2008). The disease remains endemic in several developing countries, including Brazil, where *Schistosoma mansoni* is the etiologic agent. The advent of praziquantel was essential to reduce morbidity and mortality due to schistosomiasis. However, the emergence of parasite resistant strains has been reported, raising concerns about the long-term effectiveness of this worldwide available drug (Doenhoff et al., 2002; Hotez et al., 2010). Therefore, the development of new drugs and additional control measures are essential to halt schistosomiasis dissemination. The development of a vaccine that significantly reduces parasitemia is desirable in order to allow a mass treatment with high level of protection and low costs (Chan, 1997; Katz, 1999; McManus, 1999).

Irradiated cercariae used for immunization in experimental animal models regularly induce >80% of protection (Souza et al., 1987; Lin et al., 2011; Tian et al., 2013). However, although some

promising antigens have been identified and tested in clinical trials, no effective vaccine against schistosomiasis is currently available. Indeed, most of the studied anti-schistosome targets are tegumental proteins, which directly interact with the host, but consistently do not show satisfactory protection levels (McManus and Loukas, 2008). Consequently, WHO has encouraged tests with new vaccine candidates such as cytoskeletal or cytoplasmic proteins which may be used as part of a multivalent vaccine (Wilson and Coulson, 2006). Additionally, vaccines based on nuclear/cytoplasmic proteins exhibit less chance to trigger an allergic response in the vaccinated individuals (Bethony et al., 2011), as they are not directly exposed to the host immune system.

In this context, the YBPs comprise a family of proteins that are found in most living organisms (Evdokimova et al., 2006) and contain a highly conserved nucleic acid-binding domain, the cold-shock domain (CSD), which possesses great similarity to bacterial cold-shock proteins (Wistow, 1990). In addition to the CSD, proteins from this family have a variable C-terminal TAIL domain predominantly composed of basic or

acid amino acids, which are responsible for either nucleic acid binding or protein-protein interactions (reviewed by Matsumoto and Bay, 2005). YBPs were originally identified as proteins that bind to DNA, RNA, and other proteins (Sommerville and Ladomery, 1996; Matsumoto and Wolffe, 1998; Valadão et al., 2002; Evdokimova et al., 2006; Dong et al., 2009; Mihailovich et al., 2010; Eliseeva et al., 2011). Subsequent studies demonstrated that YB-1, a member of this family, is a major component of ribonucleoprotein particles (mRNPs), working on pre-mRNA splicing, mRNA stability, and translation (Mihailovich et al., 2010; Brandt et al., 2012). Thus, these proteins regulate gene expression and participate in a variety of cellular processes, including transcriptional and translational regulation, induction of DNA repair, cellular proliferation, drug resistance, and stress responses to extracellular signals (Kohno et al., 2003; Mihailovich et al., 2010; Brandt et al., 2012).

In response to stress signals, including low temperatures, drugs that act on DNA, reactive oxygen species, and UV irradiation, the YB-1 protein can translocate from the cytoplasm to the nucleus and participate in gene regulation (Koike et al., 1997; Matsumoto and Wolffe, 1998; Kohno et al., 2003). One of the Y-box protein functions has been elucidated by studies of genes that are repressed in response to YB-1 overexpression in somatic cells. For example, an increase in cellular levels of the human YB-1 protein transcriptionally represses interferon-mediated activation of MHC class II genes (Ting et al., 1994). Subsequent analysis established that YB-1 stimulates the formation of single-stranded regions at the Y-box element (an inverted CCAAT motif) in a MHC class II gene promoter, preventing the loading and/or function of other transacting factors (MacDonald et al., 1995). In addition, it was reported that a synthetic protein can interact with YB-1, stimulating its translocation from the cytoplasm to the nucleus, where YB-1 binds to the promoters of collagen genes and suppresses their transcription, preventing the progression of systemic and hepatic fibrosis (Higashi et al., 2003a,b, 2011; Hasegawa et al., 2009). Currently, a number of genes involved in innate immune response processes and inflammation have been reported to be down- or up-regulated by the YB-1 protein (see the review by Raffetseder et al., 2012).

SMYB1 is a *S. mansoni* protein that belongs to the YBP family and was described by Franco et al. (1997). Due to the similarity between SMYB1 and Y-box proteins from other organisms, and the importance of these proteins in the control of gene expression, our group conducted several studies to characterize the SMYB1 protein. We reported that (i) the protein binds to double- or single-stranded DNA oligonucleotides, with a preference for sequences containing the CCAATT motif, (ii) the protein is expressed in all stages of the parasite life cycle, (iii) SMYB1 interacts with proteins involved in mRNA processing, and (iv) SMYB1 has a cytoplasmic localization (Franco et al., 1997; Valadão et al., 2002; de Oliveira et al., 2004; Rocha et al., 2013). Although the exact function of the SMYB1 protein in this parasite has not been determined, results presented by Valadão et al. (2002) and Rocha et al. (2013) suggested that, while SMYB1 may not act directly as a transcription factor, this protein may be necessary for the regulation of *S. mansoni* gene expression. These studies suggest that SMYB1 can function in the turnover,

transport, and stabilization of RNA molecules, acting as RNA chaperones (Valadão et al., 2002; de Oliveira et al., 2004; Rocha et al., 2013). Although intracellular proteins are not usually the first choice of immunogens for vaccination, several extracellular *S. mansoni* proteins have been previously tested with moderate success. We have therefore decided to test SMYB1 as a vaccine candidate against this parasite. To address this matter, we have used Bioinformatics tools to investigate SMYB1 sequence composition and structural features. We have further evaluated the protective efficacy of vaccination with recombinant SMYB1 (rSMYB1) against the *S. mansoni* infection in the murine model.

MATERIALS AND METHODS

ETHICS STATEMENT

Animal experiments were conducted in accordance with Brazilian Federal Law number 11,794, which regulates the scientific use of animals, and United States Institutional Animal Care and Use Committee (IACUC) guidelines. All protocols were approved by the Ethics Committee for Animal Experimentation (CETEA) at Universidade Federal de Minas Gerais under the protocol number 203/2011.

IN SILICO SEQUENCE ANALYSIS

National Center for Biotechnology Information (NCBI) BLAST (Altschul et al., 1990) searches using blastp and PSI-BLAST algorithms were performed against the UniProtKb database (The UniProt Consortium, 2013) using SMYB1 as query to identify possible SMYB1 paralogs with 90% minimal similarity. All subsequent analyses were performed for each of the three identified SMYB isoforms.

Online programs were used to assess functional characteristics of SMYB1. The InterProScan (Zdobnov and Apweiler, 2001) tool was used to recognize different protein signatures (representing protein domains, families, and functional sites) with default parameters. In addition, each SMYB protein isoform was subjected to a conserved domain search (CDS tool) (Marchler-Bauer and Bryant, 2004) from NCBI. Searches were performed against the conserved domain database (CDD v3.10; Marchler-Bauer et al., 2011) with e-values of either 0.01 or 0.001 and with or without applying the low complexity filter. The CDS analysis also points out the known DNA and RNA binding sites present within the predicted domain, by comparing to other proteins that bear the same domain.

The PredictProtein website (Rost et al., 2004) was used to generate information about the protein sequence. Several protein features can be assessed through this webserver, including amino acid composition, predicted protein binding sites and the effect of amino acid substitution. We have submitted all SMYB sequences to the PredictProtein server and retrieved specifically these three results. Protein binding sites are predicted by a machine-learning algorithm indirectly based on 3D structures to identify interacting residues using only the protein sequence as input. The effect of amino acid substitutions for each position is analyzed by exchanging the residue in each position by all other possibilities and investigating the structural/functional effect upon the protein as a whole. The impact of each point mutation is measured by a trained classifier algorithm that takes into account

several features, most importantly from evolutionary information retrieved from sequence alignments. The final output of this method is presented as a heatmap, in which each column represents one position in the protein sequence and each row represents one amino acid. The neutral substitutions are colored from white to dark green, while non-neutral are colored from white to dark red. The original amino acid is marked in black.

Intrinsically disordered regions of the three SMYB isoforms were identified using Disopred (Ward et al., 2004), a trained algorithm that accurately predicts disordered regions by comparison to a dataset of protein regions that could not be solved by X-ray crystallography and, therefore, are largely flexible. False positive rate (FPR) threshold was kept in its default value of 2%.

The secretory or non-secretory nature of the protein was predicted using SignalP 4.1 (Peterson et al., 2011), which identifies signal peptides, and the SecretomeP 2.0 server (Bendtsen et al., 2004), which predicts non-classical protein secretion pathways. Both types of prediction were performed using a default setting score of 0.5. The Euk-mPLoc 2.0 (Chou and Shen, 2010) and TargetP 1.1 Servers (Emanuelsson et al., 2000) were subsequently applied to predict the subcellular locations of SMYB1. GPI-modification sites, mucin type O-glycosylation sites, and N-glycosylation sites were analyzed using the GPI Prediction Server version 3 (Eisenhaber et al., 1999), NetOGlyc 4.0 Server (Steentoft et al., 2013), and NetNGlyc 1.0 (<http://www.cbs.dtu.dk/services/NetNGlyc/>), respectively. Predicted serine, threonine, and tyrosine phosphorylation sites were obtained using the NetPhos 2.0 Server (Blom et al., 1999).

T and B cell epitopes were predicted based on the amino acid sequences of SMYB1, using prediction tools located at the Immune Epitope Database and Analysis Resource (IEDB-AR), which is a database of experimentally characterized immune epitopes (i.e., B and T cell epitopes) in humans, non-human primates, rodents, and other animal species (<http://tools.immuneepitope.org/main/index.html>). Linear B cell epitopes were predicted using programs that incorporate solvent-accessible surface area calculations and contact distances into the prediction of B cell epitope potential along the length of the protein sequence. These programs consist of the Emini Surface Accessibility Prediction (Emini et al., 1985), Kolaskar and Tongaonkar Antigenicity (Kolaskar and Tongaonkar, 1990) and the BepiPred 1.0 server (Larsen et al., 2006). To predict T cell epitopes, neural network-based prediction of proteasomal cleavage sites (NetChop) (Nielsen et al., 2005) and T cell epitopes (NetCTL and NetCTLpan) (Larsen et al., 2005; Stranzl et al., 2010) were employed.

CLONING, EXPRESSION, AND PURIFICATION OF RECOMBINANT SMYB1

Initially, the SMYB1 cDNA (Accession no. U39883) was cloned into the pGEM-T Easy vector (Promega). The YB1fwNdeI (5'-CATATGGCGGACTAGACC-3') and YB1revHindIII (5'-AAGCTTGATCAGAGAATTTAAAGCGTC-3') primers were used for SMYB1 amplification from adult worm cDNA, generating an amplification product of 675 bp. The parameters for the PCR reaction were as follows: 1 cycle at 95°C for 6 min followed by 25 cycles of 1 min at 95°C, 1 min at 58°C, 1 min

at 72°C and a final cycle of 5 min at 72°C. The recombinant pGEM-SMYB1 vector was then digested with the enzymes *NdeI* and *HindIII* and the recovered insert was subcloned into the pET28aTEV vector, in-frame with the six histidine N-terminal (6xHis) tag. DNA sequencing was performed to confirm the presence and the correct orientation of the SMYB1 cDNA. *Escherichia coli* BL21 was transformed with the recombinant plasmid (pET28a-SMYB1) and grown in CIRCLEGROW medium (MP Biomedicals) supplemented with kanamycin (100 µg/ml), at 37°C, 180 rpm. Bacterial growth was monitored at OD600 nm until reach 0.4–0.6 and the expression of rSMYB1 was induced by the addition of 0.5 mM IPTG. After 4 h of induction, the bacterial cells were harvested by centrifugation at 7690 g for 20 min. The pellet was resuspended in 50 mL of column buffer (20 mM sodium phosphate; 300 mM NaCl; 20 mM imidazole, pH 7.4; 10% glycerol). Lysozyme (100 µg/mL) was subsequently added, and the cells were incubated for 15 min. The cells were then subjected to 3 cycles of heat shock (−80°C/37°C), followed by three 15 s cycles of sonication (Fisher Scientific) and three rounds of centrifugation at 5940 g for 20 min. The protein was purified from the supernatant by affinity chromatography on a HisTrap HP 5 mL Ni-Sepharose column (GE Healthcare) under denaturing conditions using the ÄKTA Prime Plus Liquid Chromatography System (GE Healthcare), according to the manufacturer's instructions. Fractions containing rSMYB1 were dialyzed against Tris-NaCl buffer (50 mM Tris; 20 mM NaCl, pH 7.4), which was changed every 12 h. The dialysis was performed for 36 h at 4°C using a >12 kDa dialysis tubing cellulose membrane (Sigma Aldrich). The protein was aliquoted and stored at −80°C until use. Protein concentration was determined using Bradford's method (Bradford, 1976). The recombinant protein was used as an antigen for immunization and in immunological experiments.

SDS-PAGE AND IMMUNOBLOTTING

SDS-PAGE of purified rSMYB1 was performed using 12% gels, and the gels were electroblotted onto nitrocellulose membranes for 30 min at 20 V using a semi-dry system (Bio-Rad). The membranes were blocked with phosphate-buffered saline (PBS) (130 mM NaCl, 2 mM KCl, 8 mM Na₂HPO₄, 1 mM KH₂PO₄) plus 0.05% Tween 20 (PBS-T) containing 5% dry milk (p/v) for 16 h at room temperature. The membrane was subsequently incubated in 1:2000 dilutions of an anti-His antibody (GE Healthcare) and peroxidase-conjugated anti-mouse IgG (Sigma Aldrich) in PBS-T for 1 h at room temperature. After washes using PBS-T, the membrane was developed using 3,3'-diaminobenzidine (Sigma Aldrich), according to the manufacturer's protocol. After developing, the membrane was washed using distilled water and dried on filter paper.

IMMUNIZATION OF MICE AND MEASUREMENT OF SPECIFIC ANTI-rSMYB1 ANTIBODIES

Female C57BL/6 mice ($n = 10$, per group) between 6 and 8 weeks of age were obtained from the Universidade Federal de Minas Gerais (UFMG) animal facility and supplied with commercial food and water *ad libitum*. Mice were subcutaneously injected in the nape of the neck with 25 µg of rSMYB1 on days

0, 15, and 30. The vaccine was formulated with the recombinant protein emulsified in complete Freund's adjuvant (CFA) (Sigma Aldrich) for the first immunization and incomplete Freund's adjuvant (IFA) (Sigma Aldrich) for subsequent immunizations. In the control group, Tris-NaCl buffer with Freund's adjuvant was administered using the same immunization protocol.

On the tenth day after each immunization, blood was collected from each experimental group by retro-orbital bleeding. The levels of specific anti-rSMYB1 antibodies were measured by indirect ELISA. Briefly, Maxisorp 96-well microtiter plates (Nunc) were coated with 5 µg/mL rSMYB1 in carbonate-bicarbonate buffer, pH 9.6, for 16 h at 4°C. The plates were then blocked for 2 h at room temperature with 200 µl of PBS-T plus 10% fetal bovine serum (FBS) (Life Technologies) per well. The serum from each mouse was diluted 1:100 in PBS-T, and a 100-µl sample was added to each well and incubated for 1 h at room temperature. Plate-bound antibody was detected using peroxidase-conjugated anti-mouse IgG, IgG1, and IgG2a (Sigma Aldrich) diluted to concentrations of 1:5000, 1:10000, and 1:2000 in PBS-T, respectively. Color reactions were developed by the addition of 100 µL per well of 200 pmol o-phenylenediamine (OPD) (Sigma Aldrich) in citrate buffer, pH 5.0, plus 0.04% H₂O₂ for 10 min. The reactions were stopped with 50 µL of 5% sulfuric acid per well. The plates were read at 492 nm using an ELISA plate reader (Bio-Rad).

CHALLENGE INFECTION WITH *S. MANSONI* AND WORM BURDEN RECOVERY

Cercariae of *S. mansoni* (LE strain) were maintained routinely in *Biomphalaria glabrata* snails at the Centro de Pesquisas René Rachou - Fiocruz (CPqRR) and prepared by exposing infected snails to light for 2 h to induce shedding. Cercariae numbers and viability were determined using a light microscope prior to infection. Challenge infection was performed 10 days after the final immunization. Mice were anaesthetized with 90 mg/kg of ketamine and 10 mg/kg of xylazine. The mice abdomens were shaved and they were exposed percutaneously to 100 cercariae of *S. mansoni* in water for 1 h using the ring method (Smithers and Terry, 1965). Forty-five days after challenge (DAC), the mice were sacrificed and the adult worms were perfused from the portal veins (Fonseca et al., 2004). Two independent experiments were performed to determine protection levels and 10 mice per group were used.

Protection was calculated by comparing the number of worms recovered from each vaccinated group with its respective control group, using the following formula: $PL = (WRCG - WREG) \times 100/WRCG$, where PL, protection level; WRCG, worms recovered from control group; and WREG, worms recovered from experimental group.

QUANTIFICATION OF *S. MANSONI* EGGS RETAINED IN THE LIVER

Quantification of *S. mansoni* eggs retained in the liver was performed according to the protocol described by Cheever (1968). To count the number of eggs in the liver, the organ was recovered from each experimental mouse, weighted and placed into 20 mL of a 5% KOH solution (p/v) in a 50 mL tube. Digestion occurred at room temperature for 48 h, and the samples were

subsequently mixed thoroughly. The solutions were centrifuged for 3 min at 200 g, and the pellets were resuspended in 20 mL PBS and vortexed. This step was repeated three times. After the last wash, eggs were resuspended in 5 mL of 10% buffered formaldehyde in PBS and maintained at room temperature until counting. An average of three counts was obtained per 50 µL solution to estimate the number of eggs per gram of tissue. Protection was calculated by comparing the number of eggs recovered from the vaccinated group to the number of eggs recovered from its respective control group, using the same formula used for adult worms.

HEPATIC GRANULOMA ANALYSIS

Liver sections from mice of control and vaccinated groups and infected with 100 cercariae were collected 45 days post-infection to evaluate the effect of immunization in granuloma formation. The liver sections removed from the central part of the left lateral lobe were fixed with 10% buffered formaldehyde in PBS. Histological sections were performed using microtome (4 µm) and stained in a slide with Gomory's trichromic. The granulomas were counted in Axiolab Carl Zeiss microscope using 10× objective lens. All slides were digitized by the Canon Lide 110 scanner, in 300 dpi resolution. The pixels of each histological section were fully screened, with subsequent creation of a binary image and the total area of the cut was calculated. The area of the lower cutoff was used as a minimum standard of tissue to be statistically analyzed. The results were expressed by the number of granulomas per area of liver (mm²). The area of granulomas was obtained through the KS300 software contained in Carl Zeiss image analyzer. Fifteen granulomas from each mouse with a single well-defined egg were randomly chosen at a microscope with 20× objective lens and scanned through a Q-Color3 microcamera (Olympus). Using a digital pad, the total area of granulomas was measured and the results were expressed in square micrometers (µm²).

HUMORAL RESPONSE AGAINST rSMYB1 AND *S. MANSONI* ANTIGENS AFTER CHALLENGE

Following immunization, blood was collected using the previously described protocol (see section Immunization of Mice and Measurement of Specific anti-rSMYB1 Antibodies) at day 0 (i.e., challenge) and day 45 of infection (i.e., sacrifice). Measurements of specific anti-SMYB1, anti-*Schistosoma* worm antigen protein (SWAP), and anti-soluble egg antigen (SEA) IgG, IgG1, and IgG2a antibodies in the sera were performed using indirect ELISA, as previously described.

STATISTICAL ANALYSIS

Statistical analysis was performed using Student's *t*-test in the GraphPad Prism 5.0 software package (La Jolla, CA, USA).

RESULTS

IN SILICO ANALYSES OF SMYB1 SEQUENCE

In *S. mansoni*, the SMYB1 protein (predicted molecular weight: 23805.20 Da, theoretical pI: 10.21) is encoded by the Smp_097800 gene, which produces three transcript isoforms: Smp_097800.1 (SMYB1), Smp_097800.2 (SMYB2), and

Smp_097800.3 (SMYB3) derived from alternative splicing (Figure 1A). BLAST searches using blastp and PSI-BLAST algorithms against the UniProtKb database revealed a paralog protein in *S. mansoni* (SMYBX_putative), encoded by the Smp_097750 gene, which produces a single transcript isoform (Smp_097750.1) (Figure 1B). Global alignment shows that the SMYB proteins are much conserved (more than 90% identity). The N-terminal region (CSD) is more conserved among all sequences, consistent with the fact that all Smp_097800 derived

isoforms share the first 156 amino acids, and only diverge in their C-terminal domain. Interestingly, the Smp_097750 derived isoform has an almost perfectly conserved CSD region (Figure 1C).

The InterProScan tool identified an N-terminal nucleic acid-binding OB-fold domain (IPR012340) in the SMYB isoforms (Figure 1C and Figure S1), which is found in the Y-box binding protein subfamily (PTHR11544:SF6). The presence of this domain was also confirmed by the CDS tool with high confidence (*e*-vaule of 0.001). The CDS tool has also identified a C-terminal

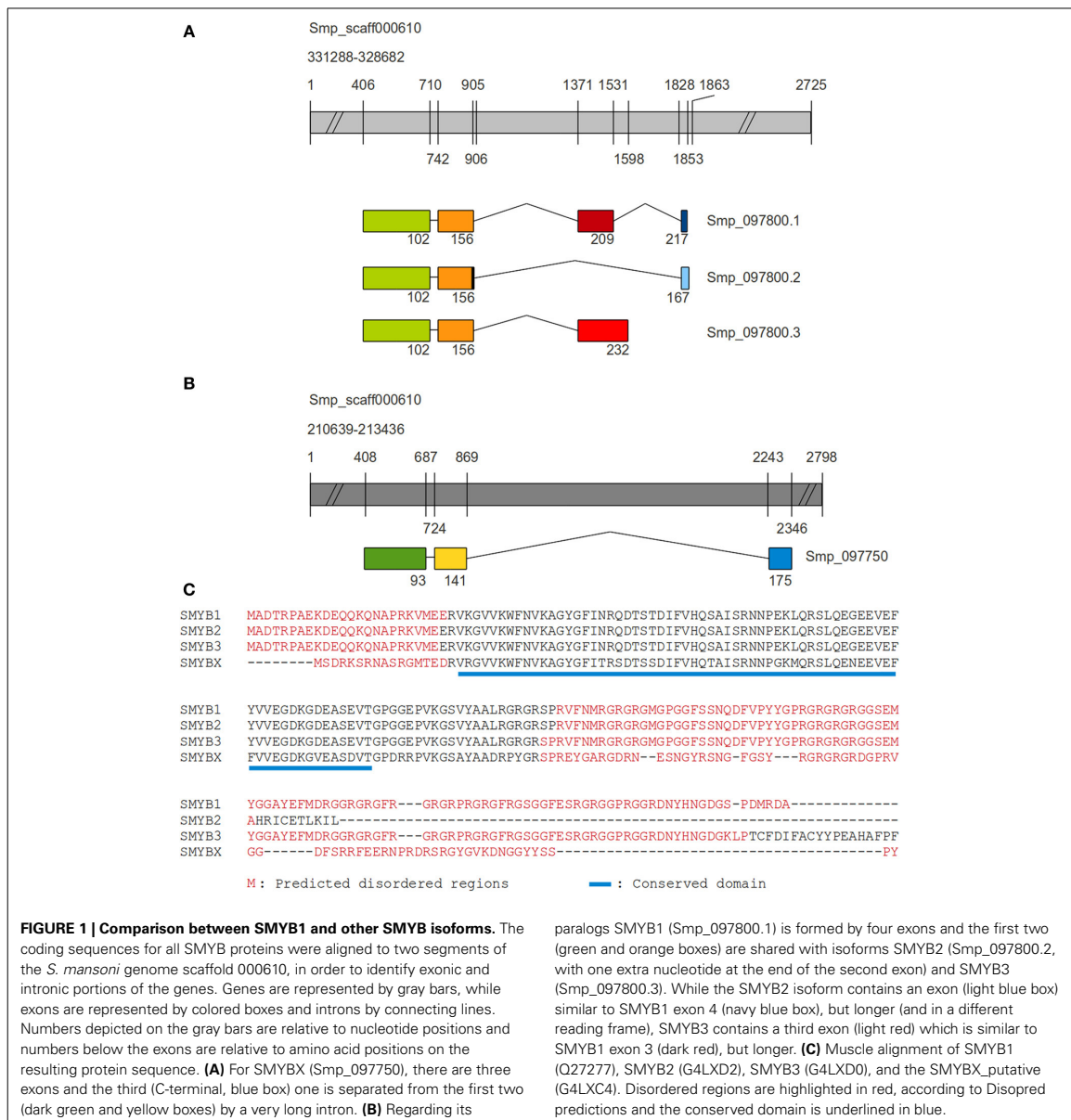


FIGURE 1 | Comparison between SMYB1 and other SMYB isoforms. The coding sequences for all SMYB proteins were aligned to two segments of the *S. mansoni* genome scaffold 000610, in order to identify exonic and intronic portions of the genes. Genes are represented by gray bars, while exons are represented by colored boxes and introns by connecting lines. Numbers depicted on the gray bars are relative to nucleotide positions and numbers below the exons are relative to amino acid positions on the resulting protein sequence. **(A)** For SMYB1 (Smp_097800.1), there are three exons and the third (C-terminal, blue box) one is separated from the first two (dark green and yellow boxes) by a very long intron. **(B)** Regarding its

paralogs SMYB1 (Smp_097800.1) is formed by four exons and the first two (green and orange boxes) are shared with isoforms SMYB2 (Smp_097800.2), with one extra nucleotide at the end of the second exon and SMYB3 (Smp_097800.3). While the SMYB2 isoform contains an exon (light blue box) similar to SMYB1 exon 4 (navy blue box), but longer (and in a different reading frame), SMYB3 contains a third exon (light red) which is similar to SMYB1 exon 3 (dark red), but longer. **(C)** Muscle alignment of SMYB1 (Q27277), SMYB2 (G4LXD2), SMYB3 (G4LXD0), and the SMYBX_putative (G4LXC4). Disordered regions are highlighted in red, according to Dispred predictions and the conserved domain is underlined in blue.

API5 domain (apoptosis inhibitor domain 5) approximately localized between residues 140 and 200 on the longer isoforms (SMYB1 and SMYB3) although with low confidence (e -value of 0.01). Further analysis may confirm this as an actual conserved domain or just an artifact (Figure S1).

The prediction of intrinsically disordered regions has characterized SMYB isoforms as mostly disordered proteins. It is interesting to observe that the conserved CSD is located away from the disordered regions (Figure S1). An additional region where the disorder probability suddenly drops (flanking the residue 180) is an interesting feature to be further investigated (Figure S1). Another interesting finding regarding the disorder is its relation to protein-binding residues. For all isoforms, predicted protein binding sites range from residues 1 to ~25, ~110 to the end of the sequence and position 65, which is the only predicted binding site out of the disordered region (Figure S1 and Supplementary Material).

When observing the SNAP results presented in Figure S1, one can easily identify the first ~20 N-terminal residues as predicted to contribute very little to the structure and function of SMYB isoforms, since all simulated mutation in such positions seem to have no effect to the proteins. On the other hand, the region where the CDS domain is located is the most important and mutations in this region can easily have a negative effect to protein structure and function. This is expected, since this is the only structured region of the proteins. Accordingly, the nucleic acid binding site regions are the most conserved within this domain, since the heatmap is dark red around these sites.

SMYB1 was predicted to be located in the cytoplasm and nucleus of *S. mansoni* cells, using the Euk-mPloc program. No cleavage sites or N-terminal presequences consistent with a mitochondrial targeting peptide or secretory pathway signal peptide were identified using the TargetP Server. In addition, the SecretomeP server revealed SecP scores below the cutoff

score (0.50), indicating a low possibility of secretion by the non-classical pathway.

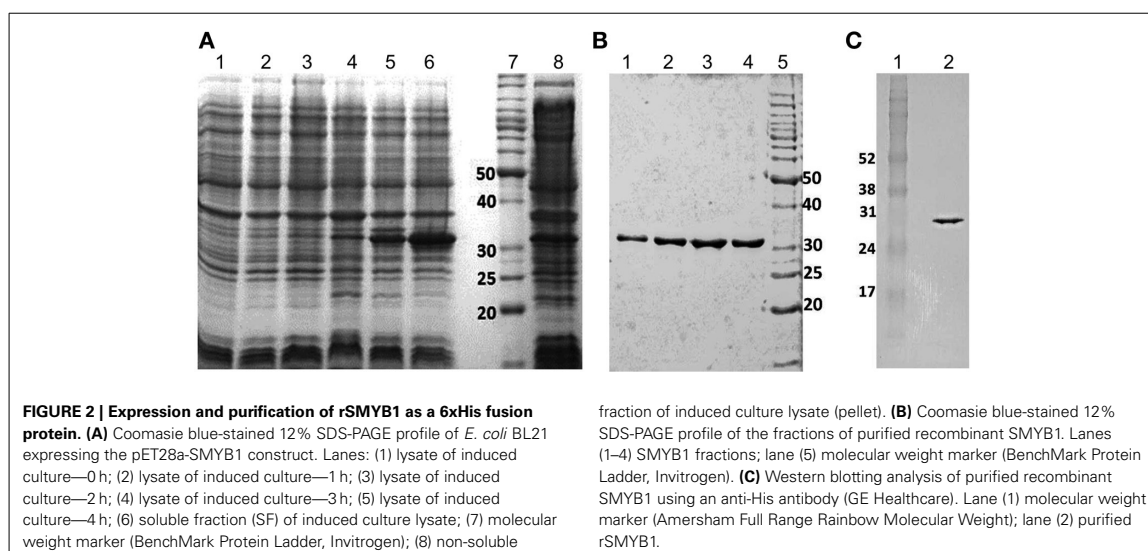
Additional Bioinformatics analyses of domain prediction, protein disorder, protein structure, and molecular interactions, as well as putative post-translational modifications (GPI modification, glycosylation, and phosphorylation sites) and B-cell and T-cell type epitope predictions for SMYB1 are presented and briefly discussed in the Supplementary Material (Table S1 and Figures S1, S2).

EXPRESSION AND PURIFICATION OF RECOMBINANT SMYB1

The SMYB1 gene was cloned into the pET28a expression vector, and the recombinant protein was successfully expressed as a 6xHis tag fusion protein. The transformed bacterial cells were treated with lysozyme, submitted to heat shock and sonication treatments, and the lysates were separated into soluble and insoluble fractions (Figure 2A). The protein was purified from the soluble fraction by affinity chromatography using His-binding columns under denaturing conditions (Figure 2B). The protein was then refolded by dialysis against Tris-NaCl buffer, with an approximate yield of 11 mg of protein/liter. The purity of the recombinant SMYB1-6xHis tag fusion protein was assessed using SDS-PAGE and Western blotting analysis with an anti-His antibody (Figure 2C), which revealed a protein of approximately 30 kDa.

HUMORAL RESPONSES TO rSMYB1

C57Bl/6 mice were immunized with three doses of rSMYB1 formulated with Freund's adjuvant, and the level of specific anti-rSMYB1 antibodies in the sera from the immune and placebo groups was evaluated using ELISA (Figure 3). Significant levels ($p < 0.01$) of specific anti-rSMYB1 IgG antibodies were detected after the first immunization, and these antibodies remained at a high levels after the second and third immunizations. To



determine the isotype of the antibody produced after immunization, IgG1 and IgG2a antibodies specific to rSMYB1 were also analyzed. The results revealed that rSMYB1 stimulates an IgG1 antibody response ($p < 0.05$) after the second dose (Figure 3). In the placebo group, no significant differences in specific IgG, IgG1, or IgG2a antibody levels were observed after immunization (data not shown).

S. MANSONI ADULT WORM RECOVERY

To determine the protective potential of rSMYB1, immunized mice were challenged with 100 *S. mansoni* cercariae. The worms were recovered by perfusion 6 weeks after challenge, and the results were expressed as the mean worm burden (mean \pm SD) as presented in Table 1. The animals immunized with rSMYB1 in Freund's adjuvant exhibited a 26% reduction in adult worm burden recovered from the mesenteric veins when compared to the control group ($p > 0.05$). No differences in male/female proportion were observed between the placebo and immune groups (data not shown). Similar results were observed in two independent experiments.

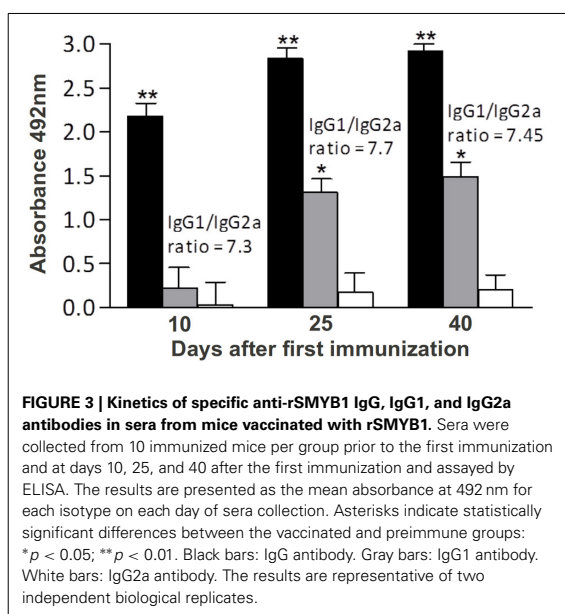


FIGURE 3 | Kinetics of specific anti-rSMYB1 IgG, IgG1, and IgG2a antibodies in sera from mice vaccinated with rSMYB1. Sera were collected from 10 immunized mice per group prior to the first immunization and at days 10, 25, and 40 after the first immunization and assayed by ELISA. The results are presented as the mean absorbance at 492 nm for each isotype on each day of sera collection. Asterisks indicate statistically significant differences between the vaccinated and preimmune groups: * $p < 0.05$; ** $p < 0.01$. Black bars: IgG antibody. Gray bars: IgG1 antibody. White bars: IgG2a antibody. The results are representative of two independent biological replicates.

Table 1 | Worm burden and protection level in mice vaccinated with the rSMYB1 protein.

| Group | Worm burden (mean \pm SD) | Protection |
|-------------------------------|-----------------------------|------------|
| Tris-NaCl + CFA/IFA (placebo) | 51.50 \pm 26.64 | – |
| rSMYB1 + CFA/IFA (immune) | 38.11 \pm 10.78 | 26% |

CFA, complete Freund's adjuvant; IFA, incomplete Freund's adjuvant. No statistically significant differences were observed between groups ($p > 0.05$). The data are representative of two independent biological assays.

QUANTIFICATION OF S. MANSONI EGGS RETAINED IN THE LIVER

In addition to worm counting, we evaluated the number of *S. mansoni* eggs retained in each gram of liver. The immunized group retained 28% less eggs in the liver than the placebo group ($p > 0.05$) (Figure 4). We have also measured the number of eggs laid by female adult worm recovered before and after immunization and found a 5.5% decrease in the number of eggs per female on the immunized group (the average was of 769.20 eggs/female on the placebo group against 726.69 eggs/female on the immunized group). Therefore, these results point to a combination between diminished egg production per female and decreased number of adult parasites in the host after immunization.

HISTOPATHOLOGICAL ANALYSIS

Histopathological analysis showed significantly fewer granulomas in the liver of animals immunized with rSMYB1 ($p < 0.05$) (Figure 5A). An associated decrease in the area of granulomas in the immunized mice group compared to the placebo group ($p < 0.05$) was also observed (Figures 5B, 6). However, no significant decrease in the area of fibrosis was detected when the two groups were compared ($p > 0.05$) (Figure 5C).

HUMORAL RESPONSE AGAINST rSMYB1 AND S. MANSONI ANTIGENS AFTER CHALLENGE

Levels of specific antibodies produced in response to the purified rSMYB1 protein in each group of mice after challenge were determined using ELISA. Surprisingly, the levels of rSMYB1-specific IgG, IgG1, and IgG2a antibodies in the immune group decreased after the third dose of the vaccine ($p > 0.05$) (Figure 7). In contrast, the placebo group exhibited increased levels of all antibodies against the protein. No statistically significant differences were observed between the immune and placebo groups at 45

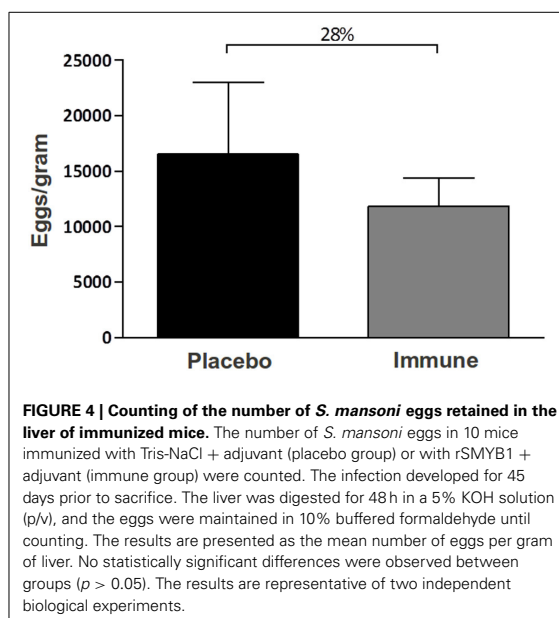
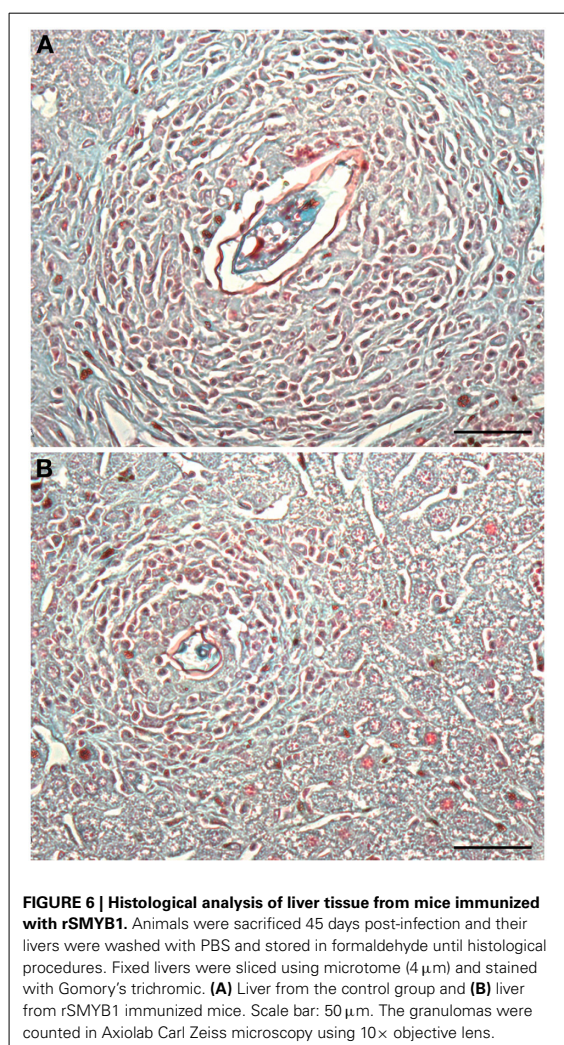
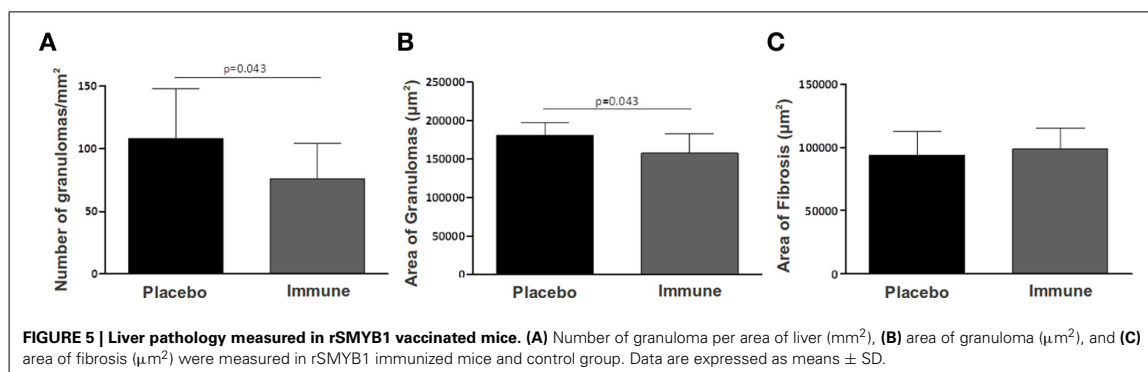


FIGURE 4 | Counting of the number of *S. mansoni* eggs retained in the liver of immunized mice. The number of *S. mansoni* eggs in 10 mice immunized with Tris-NaCl + adjuvant (placebo group) or with rSMYB1 + adjuvant (immune group) were counted. The infection developed for 45 days prior to sacrifice. The liver was digested for 48 h in a 5% KOH solution (p/v), and the eggs were maintained in 10% buffered formaldehyde until counting. The results are presented as the mean number of eggs per gram of liver. No statistically significant differences were observed between groups ($p > 0.05$). The results are representative of two independent biological experiments.



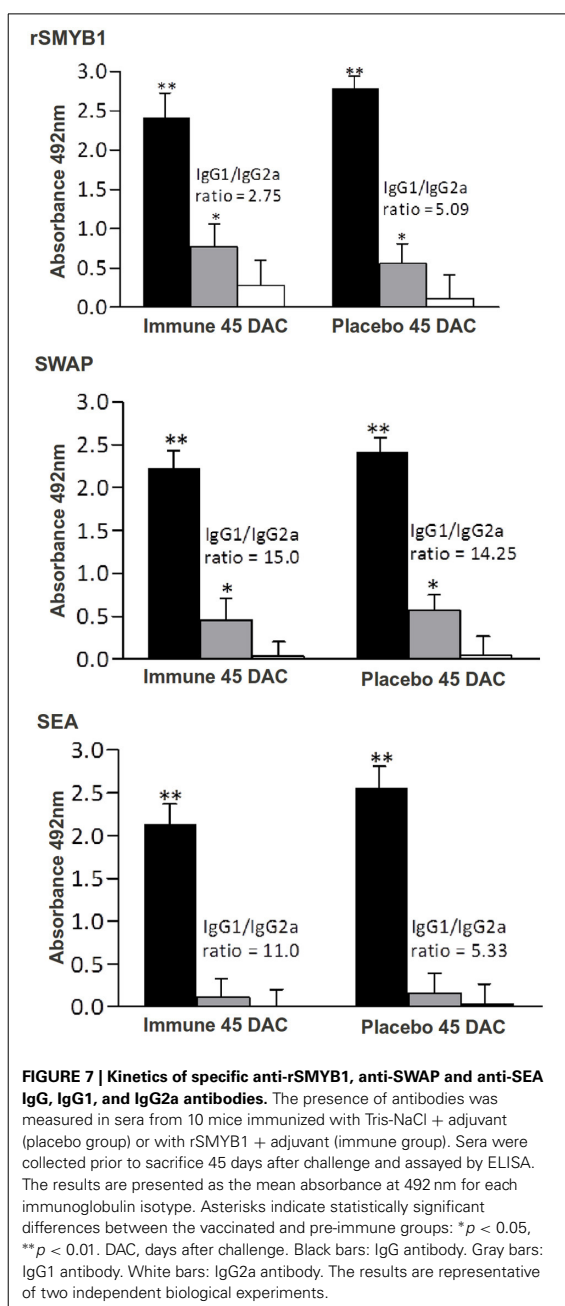
DAC ($p > 0.05$). The observed anti-rSMYB1 IgG1/IgG2a ratio decreased in the immunized mice.

To measure IgG, IgG1, and IgG2a antibodies against the specific *S. mansoni* antigens SWAP and SEA, the sera from vaccinated mice in each group were tested using ELISA (Figure 7). No specific anti-SWAP and anti-SEA IgG, IgG1, and IgG2a antibodies were detected before or during the immunizations (data not shown). After challenge, both groups developed significant levels of specific anti-SWAP and anti-SEA (Figure 7) IgG antibodies ($p < 0.01$). With respect to isotype, both groups developed a predominantly IgG1 antibody response ($p < 0.05$) against the SWAP antigen but not against the SEA antigen. No differences in IgG2a antibody response against these antigens were observed between the immune and placebo groups.

DISCUSSION

The long-term effective control of schistosomiasis will only occur as a result of combined vaccination and chemotherapy strategies with sanitation and public health control measures. Although evidences indicate that chemotherapy using praziquantel is effective in reducing the intensity of infection, as reinfection has been observed after chemotherapy, the use of this control strategy alone has been questioned (Wilson and Coulson, 2006). The irradiated cercarial vaccine elicits >80% protection in rodents and primates and other antigens identified in analyses of the *Schistosoma* proteome, transcriptome, glycome, and immunome, also exhibit protective potential (Oliveira et al., 2008). Nevertheless, the effectiveness of recombinant vaccines rarely exceeds 40% (McManus and Loukas, 2008; McWilliam et al., 2012), although new findings from El Ridi and collaborators (El Ridi and Tallima, 2013; El Ridi et al., 2014) are promising and depict a decrease of ~70–80% in worm burden using papain as adjuvant and focusing on *S. mansoni* cysteine peptidases as antigens. To date, vaccine candidates have been assessed using omics-derived high throughput approaches, such as proteomics, immunomics, and vaccinomics with promising results (DeMarco and Verjovski-Almeida, 2009; Loukas et al., 2011; McWilliam et al., 2012).

Many studies focus on tegument proteins as potential drug/vaccine targets because the tegument is a dynamic layer that represents the primary host-parasite interface and has close



proximity to the host blood and immune system (Jones et al., 2004; Pearce and Freitas, 2008; DeMarco and Verjovski-Almeida, 2009; Han et al., 2009; Loukas et al., 2011). Other studies focus on excretory/secretory (ES) proteins, molecules known to be released from live worms in the tissue culture and that may be

secreted into host tissues as the parasites move along the host body, feed, and produce eggs (Loukas et al., 2011). According to McManus and Loukas (2008), the apical membrane proteins expressed on the surfaces of the schistosomulum and the adult worm are the preferred vaccine targets. Therefore, the use of extracellular antigens for vaccine production is accompanied by inherent problems, for instance, the difficulty to produce recombinant proteins, since the majority of these antigens is processed through the classical secretory pathway and is subject of complex post-translational modifications, (e.g., glycosylation, specific processing, and disulfide bonds formation). Additionally, most antigens tested in WHO trials and by other groups are cytosolic or cytoskeletal components (e.g., paramyosin, Sm14, and GST) (Wilson and Coulson, 2006; McManus and Loukas, 2008; Oliveira et al., 2008). To our knowledge no study exploring the potential of a nucleic acid-binding protein as a *S. mansoni* vaccine candidate has been published. This is the first attempt to characterize such a protein as an antigen and to evaluate its protective efficacy as a vaccine against *S. mansoni* infection in the murine model.

In 2005, Carl and collaborators have stated that most nuclear systemic autoantigens contain long regions of structural disorder. They have studied properties of intrinsically disordered proteins in order to make connections linking disorder to antigenicity. The authors state that the amino acid composition of disordered regions (usually rich in Arg, Gly, Ser, Pro, Glu, Lys, Gln, and Ala residues) leads to a highly charged and low complexity molecule, typical properties of autoantigens (Plotz, 2003). Another property of autoantigens is their capacity to bind nucleic acids, as described by Plotz and cited by Carl and collaborators. Additionally, Plotz listed phosphorylation as a strong feature of autoantigens. All of these factors, namely enrichment in six of the listed amino acids (Arg, Gly, Ser, Pro, Glu, Lys), low complexity regions (such as repetitive sequence patterns), nucleic acid binding capacity and the abundance of phosphorylation sites (10 predicted) can be observed in the SMYB1 protein, thus corroborating its putative antigenic potential.

Although the Euk-mPLOC program predicted the SMYB1 localization in the cytoplasm and nucleus of cells, our group has previously demonstrated that this protein is predominantly located in the cytoplasm of cells from different life cycle stages of *S. mansoni*, suggesting that SMYB1 is probably acting in RNA metabolism in the cytoplasm. We also showed the presence of SMYB1 near the tegument in adult worms proposing an action on the translational regulation of tegument proteins (Rocha et al., 2013). Intrinsically disordered proteins have recently been characterized as the prevalent type of RNA and protein chaperones (Tompa and Csermely, 2004). Accordingly, it has been shown that YBPs and other cold-shock proteins typically act as chaperones that maintain mRNA in a single-stranded conformation to sustain the expression of genes that are necessary for cell growth, proliferation, and transformation (Jiang et al., 1997; Matsumoto and Wolffe, 1998; Salvetti et al., 1998; Tanaka et al., 2004; Evdokimova et al., 2006). YBPs are thought to play roles in a wide variety of responses to environmental stresses (Kohno et al., 2003). As such, SMYB1 localization in the cytoplasm of tegumental cells reinforces its

importance as a protein that acts responding to the stressing host environment.

Molecules that contain signal peptides or signal anchors are predicted to be excreted, secreted or membrane-anchored, directly interacting with the host immune system and, as stated above, constitute relevant targets for schistosome vaccines. The combined Bioinformatics results obtained in this study suggest that the SMYB1 protein is not secreted. However, Frye et al. (2009) reported that human YB-1 is secreted from cells during inflammatory stress after treatment with lipopolysaccharide, hydrogen peroxide or TGF β . In these cases, YB-1 is secreted not via the classical mechanism of protein secretion (i.e., via the Golgi apparatus and endoplasmic reticulum) but by a non-classical mechanism inside endolysosomal vesicles (Frye et al., 2009; Eliseeva et al., 2011). The question of whether SMYB1 is secreted or not needs further experimental investigation.

We reported here the successful cloning of SMYB1 cDNA into the pET28a vector and the expression of rSMYB1 in the soluble fraction of bacterial lysates. The discrepancy between the ~30 kDa protein molecular mass value calculated from SDS-PAGE and the ~24 kDa protein molecular mass value predicted from the cDNA is typical of Y-box proteins and related to the anomalous electrophoretic properties of these proteins (Deschamps et al., 1992) or to post-translational modification, such as phosphorylation (Salveti et al., 1998). We subsequently evaluated the antigenicity of the protein by investigating the murine humoral immune response to rSMYB1 and the impact of its immunization on adult worm and egg burden in mice challenged with 100 cercariae of *S. mansoni*. Recent data suggested that the establishment of a robust humoral response is likely the key for generating maximal immunity against schistosomes (Wynn and Hoffmann, 2000). A primary obstacle to the development of a schistosome vaccine is the lack of available knowledge concerning the type of immune response that should be induced. In the irradiated cercariae vaccination model, above 80% protection can be granted by a Th1, a Th2, or a mixed Th1/Th2 immune response (Wynn and Hoffmann, 2000). However, with respect to recombinant proteins, Th1-inducing antigens have been reported to confer protection against *Schistosoma* infection in the mouse model (Jankovic et al., 1996; Mountford et al., 1996; Zhang et al., 2001; Fonseca et al., 2004; Valardo et al., 2004; Li et al., 2005; Cardoso et al., 2008; Garcia et al., 2008).

In this study, C57Bl/6 mice immunized with rSMYB1 exhibited high levels of specific anti-SMYB1 IgG antibodies that emerged after the first immunization. Specific anti-SMYB1 IgG1 antibodies predominated over IgG2a antibodies, particularly after the second immunization. However, the IgG1/IgG2a ratio was reduced after the last immunization (i.e., during the challenge period). Antibody levels correlated with protective efficacy in our study. The antibody levels developed by mice immunized with rSMYB1 reduced in 26% the number of adult worm burden and in 28% the eggs/granuloma trapped in the liver. A critical issue in vaccine design is the use of an appropriate adjuvant to induce the suitable immune response. Although the CFA adjuvant, which triggers a Th1 response, cannot be used in humans (Heegaard et al., 2011), it is widely utilized in initial immunization trials. Further experiments combining rSMYB1 with suitable adjuvant

formulations for use in humans should be performed. In this sense, an interesting strategy would be to use papain as adjuvant, given that recently published articles have described very high protection rates related to the use of such molecule in vaccine candidates (El Ridi and Tallima, 2013; El Ridi et al., 2014).

S. mansoni adult worms live in the blood essentially unrecognized for many years, whereas schistosome eggs are a prominent target of the host immune response. In the first weeks of murine *S. mansoni* infection, a Th1 immune response is observed and the eggs deposited in the blood vessels by females that pass to the endothelial barrier and become trapped in the liver are immediately targeted by recruited immune cells that consist primarily of T-cells, eosinophils, and macrophages (Pearce and MacDonald, 2002; Wynn et al., 2004). Histopathology results show that in the initial phase of infection vaccination with SMYB1 seems to interfere with cell recruitment and migration in the liver. Consequently, the resulting granulomas, although presenting the same area of fibrosis, were fewer when compared to unvaccinated animals, showing the protective potential of the protein in the initial liver pathology.

Although tegument proteins are considered the main targets for vaccine development (Bergquist et al., 2002; McManus and Loukas, 2008), this study demonstrates that a cytoplasmic protein has the potential to be used as an immunogen, as we showed that SMYB1 could reduce the load of intestinal worms and eggs retained in the liver when it was used in vaccination trials and also that the protection levels achieved by SMYB1 are comparable to those obtained with other tegument and cytoskeleton proteins.

AUTHOR'S CONTRIBUTIONS

Sílvia R. C. Dias, Mariana Boroni, Thomaz L. Dias, Daniela de Laet Souza, Fabrício M. S. Oliveira, Elizângela A. Rocha, Mainá Bitar, Andrea M. Macedo, Carlos R. Machado, Marcelo V. Caliari, and Glória R. Franco contributed to the conception and design of the experiments; Sílvia R. C. Dias, Mariana Boroni, Thomaz L. Dias, Daniela de Laet Souza, Fabrício M. S. Oliveira, Elizângela A. Rocha, Mainá Bitar, Andrea M. Macedo, Carlos R. Machado, Marcelo V. Caliari, and Glória R. Franco performed data acquisition, analysis, and interpretation of the results; Sílvia R. C. Dias, Mariana Boroni, Mainá Bitar, Fabrício M. S. Oliveira, Marcelo V. Caliari, and Glória R. Franco contributed to drafting the manuscript; Sílvia R. C. Dias, Mariana Boroni, Mainá Bitar, Fabrício M. S. Oliveira, Andrea M. Macedo, Carlos R. Machado, Marcelo V. Caliari, and Glória R. Franco critically revised intellectual content of the work.

ACKNOWLEDGMENTS

This work was supported by CNPq, CAPES/PNPD, and FAPEMIG. The authors would like to thank Dr. Ronaldo Nagem and Dr. Daniela Castanheira Bartholomeu from Universidade Federal de Minas Gerais (UFMG) for giving permission to use the AKTA Prime System and Dr. Ronaldo Nagem for kindly providing the pET28aTEV vector; Dr. Éliada Mara Leite Rabelo from UFMG for kindly providing the *S. mansoni* extracts; Moluscário Lobato Paraense (CPqRR) for cercariae donation, Carlos Manoel Afonso for animal care; and Neuza Rodrigues Antunes for laboratory technical support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://www.frontiersin.org/journal/10.3389/fgene.2014.00174/abstract>

REFERENCES

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2
- Bendtsen, J. D., Jensen, L. J., Blom, N., von Heijne, G., and Brunak, S. (2004). Feature based prediction of non-classical and leaderless protein secretion. *Protein Eng. Des. Sel.* 17, 349–356. doi: 10.1093/protein/gzh037
- Bergquist, R., Al-Sherbiny, M., Barakat, R., and Olds, R. (2002). Blueprint for schistosomiasis vaccine development. *Acta Trop.* 82, 183–192. doi: 10.1016/S0001-706X(02)00048-7
- Bethony, J. M., Cole, R. N., Guo, X., Kamhawi, S., Lightowlers, M. W., Loukas, A., et al. (2011). Vaccines to combat the neglected tropical diseases. *Immunol. Rev.* 239, 237–270. doi: 10.1111/j.1600-065X.2010.00976.x
- Blom, N., Gammeltoft, S., and Brunak, S. (1999). Sequence- and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.* 294, 1351–1362. doi: 10.1006/jmbi.1999.3310
- Bradford, M. M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* 72, 248–254. doi: 10.1016/0003-2697(76)90527-3
- Brandt, S., Raffetseder, U., Djudjaj, S., Schreiter, A., Kadereit, B., Michele, M., et al. (2012). Cold shock Y-box protein-1 participates in signaling circuits with auto-regulatory activities. *Eur. J. Cell Biol.* 91, 464–471. doi: 10.1016/j.ejcb.2011.07.002
- Cardoso, F. C., Macedo, G. C., Gava, E., Kitten, G. T., Mati, V. L., de Melo, A. L., et al. (2008). *Schistosoma mansoni* tegument protein Sm29 is able to induce a Th1-type of immune response and protection against parasite infection. *PLoS Negl. Trop. Dis.* 2:e308. doi: 10.1371/journal.pntd.0000308
- Carl, P. L., Temple, B. R. S., and Cohen, P. L. (2005). Most nuclear systemic autoantigens are extremely disordered proteins: implications for the etiology of systemic autoimmunity. *Arthritis Res. Ther.* 7, R1360–R1374. doi: 10.1186/ar1832
- Chan, M. S. (1997). The global burden of intestinal nematode infections—fifty years on. *Parasitol. Today* 13, 438–443. doi: 10.1016/S0169-4758(97)01144-7
- Cheever, A. W. (1968). Conditions affecting the accuracy of potassium hydroxide digestion techniques for counting *Schistosoma mansoni* eggs in tissues. *Am. J. Trop. Med. Hyg.* 17, 38–64.
- Chou, K.-C., and Shen, H.-B. (2010). A new method for predicting the sub-cellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. *PLoS ONE* 5:e9931. doi: 10.1371/journal.pone.0009931
- DeMarco, R., and Verjovski-Almeida, S. (2009). Schistosomes - proteomics studies for potential novel vaccines and drug targets. *Drug Discov. Today* 14, 472–478. doi: 10.1016/j.drudis.2009.01.011
- de Oliveira, F. M., da Silva, I. C., Rumjanek, F. D., Valadao, A. F., Franco, G. R., Mesquita, R. D., et al. (2004). Functional properties of *Schistosoma mansoni* single-stranded DNA-binding protein SmpUR-alpha. Description of the interaction between SmpUR-alpha and SMYB1. *Mol. Biochem. Parasitol.* 135, 21–30. doi: 10.1016/S0166-6851(04)00003-9
- Deschamps, S., Viel, A., Garrigos, M., Denis, H., and le Maire, M. (1992). mRNP4, a major mRNA-binding protein from *Xenopus* oocytes is identical to transcription factor FRG Y2. *J. Biol. Chem.* 267, 13799–13802.
- Doenhoff, M. J., Kusel, J. R., Coles, G. C., and Cioli, D. (2002). Resistance of *Schistosoma mansoni* to praziquantel: is there a problem? *Trans. R. Soc. Trop. Med. Hyg.* 96, 465–469. doi: 10.1016/S0035-9203(02)90405-0
- Dong, J., Akcakanat, A., Stivers, D. N., Zhang, J., Kim, D., and Meric-Bernstam, F. (2009). RNA-binding specificity of Y-box protein 1. *RNA Biol.* 6, 59–64. doi: 10.4161/rna.6.1.7458
- Eisenhaber, B., Bork, P., and Eisenhaber, F. (1999). Prediction of potential GPI-modification sites in proprotein sequence. *J. Mol. Biol.* 292, 741–758. doi: 10.1006/jmbi.1999.3069
- Eliseeva, I. A., Kim, E. R., Guryanov, S. G., Ovchinnikov, L. P., and Lyabin, D. N. (2011). Y-box-binding protein 1 (YB-1) and its functions. *Biochemistry (Mosc.)* 76, 1402–1433. doi: 10.1134/S0006297911130049
- El Ridi, R., and Tallima, H. (2013). Vaccine-induced protection against murine schistosomiasis mansoni with larval excretory-secretory antigens and papain or type-2 cytokines. *J. Parasitol.* 99, 194–202. doi: 10.1645/GE-3186.1
- El Ridi, R., Tallima, H., Selim, S., Donnelly, S., Cotton, S., Santana, B. G., et al. (2014). Cysteine peptidases as schistosomiasis vaccines with inbuilt adjuvanticity. *PLoS ONE* 9:e85401. doi: 10.1371/journal.pone.0085401
- Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* 300, 1005–1016. doi: 10.1006/jmbi.2000.3903
- Emini, E. A., Hughes, J. V., Perlow, D. S., and Boger, J. (1985). Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. *J. Virol.* 55, 836–839.
- Evdokimova, V., Ovchinnikov, L. P., and Sorensen, P. H. (2006). Y-box binding protein 1: providing a new angle on translational regulation. *Cell Cycle* 5, 1143–1147. doi: 10.4161/cc.5.11.2784
- Fonseca, C. T., Brito, C. F., Alves, J. B., and Oliveira, S. C. (2004). IL-12 enhances protective immunity in mice engendered by immunization with recombinant 14 kDa *Schistosoma mansoni* fatty acid-binding protein through an IFN-gamma and TNF-alpha dependent pathway. *Vaccine* 22, 503–510. doi: 10.1016/j.vaccine.2003.07.010
- Franco, G. R., Garratt, R. C., Tanaka, M., Simpson, A. J., and Pena, S. D. (1997). Characterization of a *Schistosoma mansoni* gene encoding a homologue of the Y-box binding protein. *Gene* 198, 5–16. doi: 10.1016/S0378-1119(97)00261-8
- Frye, B. C., Halfter, S., Djudjaj, S., Muehlenberg, P., Weber, S., Raffetseder, U., et al. (2009). Y-box protein-1 is actively secreted through a non-classical pathway and acts as an extracellular mitogen. *EMBO Rep.* 10, 783–789. doi: 10.1038/embor.2009.81
- Garcia, T. C., Fonseca, C. T., Pacifico, L. G., Durães Fdo, V., Marinho, F. A., Penido, M. L., et al. (2008). Peptides containing T cell epitopes, derived from Sm14, but not from paramyosin, induce a Th1 type of immune response, reduction in liver pathology and partial protection against *Schistosoma mansoni* infection in mice. *Acta Trop.* 106, 162–167. doi: 10.1016/j.actatropica.2008.03.003
- Han, Z. G., Brindley, P. J., Wang, S. Y., and Chen, Z. (2009). Schistosoma genomics: new perspectives on schistosome biology and host-parasite interaction. *Annu. Rev. Genomics Hum. Genet.* 10, 211–240. doi: 10.1146/annurev-genom-082908-150036
- Hasegawa, M., Matsushita, Y., Horikawa, M., Higashi, K., Tomigahara, Y., Kaneko, H., et al. (2009). A novel inhibitor of Smad-dependent transcriptional activation suppresses tissue fibrosis in mouse models of systemic sclerosis. *Arthritis Rheum.* 60, 3465–3475. doi: 10.1002/art.24934
- Heegaard, P. M., Dedieu, L., Johnson, N., Le Potier, M. F., Mockey, M., Mutinelli, F., et al. (2011). Adjuvants and delivery systems in veterinary vaccinology: current state and future developments. *Arch Virol.* 156, 183–202. doi: 10.1007/s00705-010-0863-1
- Higashi, K., Inagaki, Y., Fujimori, K., Nakao, A., Kaneko, H., and Nakatsuka, I. (2003a). Interferon-gamma interferes with transforming growth factor-beta signaling through direct interaction of YB-1 with Smad3. *J. Biol. Chem.* 278, 43470–43479. doi: 10.1074/jbc.M302339200
- Higashi, K., Inagaki, Y., Suzuki, N., Mitsui, S., Mauviel, A., Kaneko, H., et al. (2003b). Y-box-binding protein YB-1 mediates transcriptional repression of human alpha 2(I) collagen gene expression by interferon-gamma. *J. Biol. Chem.* 278, 5156–5162. Erratum in: *J. Biol. Chem.* (2003). 278(14), 12598. doi: 10.1074/jbc.M208724200
- Higashi, K., Tomigahara, Y., Shiraki, H., Miyata, K., Mikami, T., Kimura, T., et al. (2011). A novel small compound that promotes nuclear translocation of YB-1 ameliorates experimental hepatic fibrosis in mice. *J. Biol. Chem.* 286, 4485–4492. doi: 10.1074/jbc.M110.151936
- Hotez, P. J., Bethony, J. M., Diemert, D. J., Pearson, M., and Loukas, A. (2010). Developing vaccines to combat hookworm infection and intestinal schistosomiasis. *Nat. Rev. Microbiol.* 8, 814–826. doi: 10.1038/nrmicro2438
- Hotez, P. J., Brindley, P. J., Bethony, J. M., King, C. H., Pearce, E. J., and Jacobson, J. (2008). Helminth infections: the great neglected tropical diseases. *J. Clin. Invest.* 118, 1311–1321. doi: 10.1172/JCI34261
- Jankovic, D., Aslund, L., Oswald, I. P., Caspar, P., Champion, C., Pearce, E., et al. (1996). Calpain is the target antigen of a Th1 clone that transfers protective immunity against *Schistosoma mansoni*. *J. Immunol.* 157, 806–814.
- Jiang, W., Hou, Y., and Inouye, M. (1997). CspA, the major cold-shock protein of *Escherichia coli*, is an RNA chaperone. *J. Biol. Chem.* 272, 196–202. doi: 10.1074/jbc.272.1.196

- Jones, M. K., Gobert, G. N., Zhang, L., Sunderland, P., and McManus, D. P. (2004). The cytoskeleton and motor proteins of human schistosomes and their roles in surface maintenance and host-parasite interactions. *Bioessays* 26, 752–765. doi: 10.1002/bies.20058
- Katz, N. (1999). Problems in the development of a vaccine against schistosomiasis mansoni. *Rev. Soc. Bras. Med. Trop.* 32, 705–711. doi: 10.1590/S0037-86821999000600014
- King, C. H., Sturrock, R. F., Kariuki, H. C., and Hamburger, J. (2006). Transmission control for schistosomiasis - why it matters now. *Trends Parasitol.* 22, 575–582. doi: 10.1016/j.pt.2006.09.006
- Kohno, K., Izumi, H., Uchiumi, T., Ashizuka, M., and Kuwano, M. (2003). The pleiotropic functions of the Y-box-binding protein, YB-1. *Bioessays* 25, 691–698. doi: 10.1002/bies.10300
- Koike, K., Uchiumi, T., Ohga, T., Toh, S., Wada, M., Kohno, K., et al. (1997). Nuclear translocation of the Y-box binding protein by ultraviolet irradiation. *FEBS Lett.* 417, 390–394. doi: 10.1016/S0014-5793(97)01296-9
- Kolaskar, A. S., and Tongaonkar, P. C. (1990). A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS Lett.* 276, 172–174. doi: 10.1016/0014-5793(90)80535-Q
- Larsen, J. E. P., Lund, O., and Nielsen, M. (2006). Improved method for predicting linear B-cell epitopes. *Immunome Res.* 2, 2. doi: 10.1186/1745-7580-2-2
- Larsen, M. V., Lundegaard, C., Lamberth, K., Buus S., Brunak, S., Lund, O., et al. (2005). An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC-I binding, TAP transport efficiency, and proteasomal cleavage predictions. *Eur. J. Immunol.* 35, 2295–2303. doi: 10.1002/eji.200425811
- Li, G. F., Wang, Y., Zhang, Z. S., Wang, X. J., Ji, M. J., Zhu, X., et al. (2005). Identification of immunodominant Th1-type T cell epitopes from *Schistosoma japonicum* 28 kDa glutathione-S-transferase, a vaccine candidate. *Acta Biochim. Biophys. Sin. (Shanghai)* 37, 751–758. doi: 10.1111/j.1745-7270.2005.00111.x
- Lin, D., Tian, F., Wu, H., Gao, Y., Wu, J., Zhang, D., et al. (2011). Multiple vaccinations with UV-attenuated cercariae in pig enhance protective immunity against *Schistosoma japonicum* infection as compared to single vaccination. *Parasit. Vectors* 4, 103. doi: 10.1186/1756-3305-4-103
- Loukas, A., Gaze, S., Mulvenna, J. P., Gasser, R. B., Brindley, P. J., Doolan, D. L., et al. (2011). Vaccinomics for the major blood feeding helminths of humans. *OMICS* 15, 567–577. doi: 10.1089/omi.2010.0150
- MacDonald, G. H., Itoh-Lindstrom, Y., and Ting, J. P. (1995). The transcriptional regulatory protein, YB-1, promotes single-stranded regions in the DNA promoter. *J. Biol. Chem.* 270, 3527–3533. doi: 10.1074/jbc.270.8.3527
- Marchler-Bauer, A., and Bryant, S. H. (2004). CD-Search: protein domain annotations on the fly. *Nucleic Acids Res.* 32, 327–331. doi: 10.1093/nar/gkh454
- Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., et al. (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* 39, D225–D229. doi: 10.1093/nar/gkq1189
- Matsumoto, K., and Bay, B. H. (2005). Significance of the Y-box proteins in human cancers. *J. Mol. Genet. Med.* 1, 11–17. doi: 10.4172/1747-0862.1000005
- Matsumoto, K., and Wolfe, A. P. (1998). Gene regulation by Y-box proteins: coupling control of transcription and translation. *Trends Cell Biol.* 8, 318–323. doi: 10.1016/S0962-8924(98)01300-2
- McManus, D. P. (1999). The search for a vaccine against schistosomiasis—a difficult path but an achievable goal. *Immunol. Rev.* 171, 149–161. doi: 10.1111/j.1600-065X.1999.tb01346.x
- McManus, D. P., and Loukas, A. (2008). Current status of vaccines for schistosomiasis. *Clin. Microbiol. Rev.* 21, 225–242. doi: 10.1128/CMR.00046-07
- McWilliam, H. E., Driguez, P., Piedrafita, D., McManus, D. P., and Meeusen, E. N. (2012). Novel immunomic technologies for schistosome vaccine development. *Parasite Immunol.* 34, 276–284. doi: 10.1111/j.1365-3024.2011.01330.x
- Mihailovich, M., Millitti, C., Gabaldón, T., and Gebauer, F. (2010). Eukaryotic cold shock domain proteins: highly versatile regulators of gene expression. *Bioessays* 32, 109–118. doi: 10.1002/bies.200900122
- Mountford, A. P., Anderson, S., and Wilson, R. A. (1996). Induction of Th1 cell-mediated protective immunity to *Schistosoma mansoni* by co-administration of larval antigens and IL-12 as an adjuvant. *J. Immunol.* 156, 4739–4745.
- Nielsen, M., Lundegaard, C., Lund, O., and Kesmir, C. (2005). The role of the proteasome in generating cytotoxic T cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics* 57, 33–41. doi: 10.1007/s00251-005-0781-7
- Oliveira, S. C., Fonseca, C. T., Cardoso, F. C., Farias, L. P., and Leite, L. C. (2008). Recent advances in vaccine research against schistosomiasis in Brazil. *Acta Trop.* 108, 256–262. doi: 10.1016/j.actatropica.2008.05.023
- Pearce, E. J., and Freitas, T. C. (2008). Reverse genetics and the study of the immune response to schistosomes. *Parasite Immunol.* 30, 215–221. doi: 10.1111/j.1365-3024.2007.01005.x
- Pearce, E. J., and MacDonald, A. S. (2002). The immunobiology of schistosomiasis. *Nat. Rev. Immunol.* 2, 499–511. doi: 10.1038/nri843
- Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* 8, 785–786. doi: 10.1038/nmeth.1701
- Plotz, P. H. (2003). The autoantibody repertoire: searching for order. *Nat. Rev. Immunol.* 3, 73–78. doi: 10.1038/nri976
- Raffetseder, U., Liehn, E. A., Weber, C., and Mertens, P. R. (2012). Role of cold shock Y-box protein-1 in inflammation, atherosclerosis and organ transplant rejection. *Eur. J. Cell Biol.* 91, 567–575. doi: 10.1016/j.ejcb.2011.07.001
- Rocha, E. A., Valadao, A. F., Rezende, C. M., Dias, S. R., Macedo, A. M., Machado, C. R., et al. (2013). Identification of a new *Schistosoma mansoni* SMYB1 partner: putative roles in RNA metabolism. *Parasitology* 140, 1085–1095. doi: 10.1017/S0031182013000413
- Rost, B., Yachdav, G., and Liu, J. (2004). The PredictProtein server. *Nucleic Acids Res.* 32, W321–W326. doi: 10.1093/nar/gkh377
- Salveti, A., Batistoni, R., Deri, P., Rossi, L., and Sommerville, J. (1998). Expression of DJY1, a protein containing a cold shock domain and RG repeat motifs, is targeted to sites of regeneration in planarians. *Dev. Biol.* 201, 217–229. doi: 10.1006/dbio.1998.8996
- Smithers, S. R., and Terry, R. J. (1965). The infection of laboratory hosts with cercariae of *S. mansoni* and the recovery of adult worms. *Parasitology* 55, 695–700.
- Sommerville, J., and Ladomery, M. (1996). Masking of mRNA by Y-box proteins. *FASEB J.* 10, 435–443.
- Souza, C. P., Araújo, N., Jannotti, L. K., and Gazzinelli, G. (1987). Fatores que podem afetar a criação e manutenção de caramujos infectados e a produção de cercárias de *Schistosoma mansoni* [Factors that might affect the creation and maintenance of infected snails and the production of *Schistosoma mansoni* cercariae]. *Mem. Inst. Oswaldo Cruz.* 82, 73–79. doi: 10.1590/S0074-02761987000100013
- Steenfot, C., Vakhrushev, S. Y., Joshi, H. J., Kong, Y., Vester-Christensen, M. B., Schjoldager, K. T., et al. (2013). Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. *EMBO J.* 32, 1478–1488. doi: 10.1038/emboj.2013.79
- Steinmann, P., Keiser, J., Bos, R., Tanner, M., and Utzinger, J. (2006). Schistosomiasis and water resources development: systematic review, meta-analysis, and estimates of people at risk. *Lancet Infect. Dis.* 6, 411–425. doi: 10.1016/S1473-3099(06)70521-7
- Stranzl, T., Larsen, M. V., Lundegaard, C., and Nielsen, M. (2010). NetCTLpan: pan-specific MHC class I pathway epitope predictions. *Immunogenetics* 62, 357–368. doi: 10.1007/s00251-010-0441-4
- Tanaka, K. J., Matsumoto, K., Tsujimoto, M., and Nishikata, T. (2004). CiYB1 is a major component of storage mRNPs in ascidian oocytes: implications in translational regulation of localized mRNAs. *Dev. Biol.* 272, 217–230. doi: 10.1016/j.ydbio.2004.04.032
- The UniProt Consortium. (2013). The UniProt Consortium Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.* 41, D43–D47. doi: 10.1093/nar/gks1068
- Tian, F., Hou, M., Chen, L., Gao, Y., Zhang, X., Ji, M., et al. (2013). Proteomic analysis of schistosomiasis japonica vaccine candidate antigens recognized by UV-attenuated cercariae-immunized porcine serum IgG2. *Parasitol. Res.* 112, 2791–2803. doi: 10.1007/s00436-013-3447-7
- Ting, J. P., Painter, A., Zeleznik-Le, N. J., MacDonald, G., Moore, T. M., Brown, A., et al. (1994). YB-1 DNA-binding protein represses interferon gamma activation of class II major histocompatibility complex genes. *J. Exp. Med.* 179, 1605–1611. doi: 10.1084/jem.179.5.1605
- Tompa, P., and Csermely, P. (2004). The role of structural disorder in the function of RNA and protein chaperones. *FASEB J.* 18, 1169–1175. doi: 10.1096/fj.04-1584rev
- Valadao, A. F., Fantappie, M. R., LoVerde, P. T., Pena, S. D., Rumjanek, F. D., and Franco, G. R. (2002). Y-box binding protein from *Schistosoma mansoni*:

- interaction with DNA and RNA. *Mol. Biochem. Parasitol.* 125, 47–57. doi: 10.1016/S0166-6851(02)00210-4
- Varaldo, P. B., Leite, L. C., Dias, W. O., Miyaji, E. N., Torres, F. I., Gebara, V. C., et al. (2004). Recombinant *Mycobacterium bovis* BCG expressing the Sm14 antigen of *Schistosoma mansoni* protects mice from cercarial challenge. *Infect. Immun.* 72, 3336–3343. doi: 10.1128/IAI.72.6.3336-3343.2004
- Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., and Jones, D. T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* 337, 635–645. doi: 10.1016/j.jmb.2004.02.002
- Wilson, R. A., and Coulson, P. S. (2006). Schistosome vaccines: a critical appraisal. *Mem. Inst. Oswaldo Cruz.* 101(Suppl. 1), 13–20. doi: 10.1590/S0074-02762006000900004
- Wistow, G. (1990). Cold shock and DNA binding. *Nature* 344, 823–824. doi: 10.1038/344823c0
- Wynn, T. A., and Hoffmann, K. F. (2000). Defining a schistosomiasis vaccination strategy - is it really Th1 versus Th2? *Parasitol. Today* 16, 497–501. doi: 10.1016/S0169-4758(00)01788-9
- Wynn, T. A., Thompson, R. W., Cheever, A. W., and Mentink-Kane, M. M. (2004). Immunopathogenesis of schistosomiasis. *J. Immunol.* 201, 156–167. doi: 10.1111/j.0105-2896.2004.00176.x
- Zdobnov, E. M., and Apweiler, R. (2001). InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17, 847–848. doi: 10.1093/bioinformatics/17.9.847
- Zhang, Y., Taylor, M. G., Johansen, M. V., and Bickle, Q. D. (2001). Vaccination of mice with a cocktail DNA vaccine induces a Th1-type immune response and partial protection against *Schistosoma japonicum* infection. *Vaccine* 20, 724–730. doi: 10.1016/S0264-410X(01)00420-0

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.

Received: 12 February 2014; accepted: 22 May 2014; published online: 11 June 2014.
Citation: Dias SRC, Boroni M, Rocha EA, Dias TL, de Laet Souza D, Oliveira FMS, Bitar M, Macedo AM, Machado CR, Caliani MV and Franco GR (2014) Evaluation of the *Schistosoma mansoni* Y-box-binding protein (SMYB1) potential as a vaccine candidate against schistosomiasis. *Front. Genet.* 5:174. doi: 10.3389/fgene.2014.00174
This article was submitted to *Evolutionary and Genomic Microbiology*, a section of the journal *Frontiers in Genetics*.
Copyright © 2014 Dias, Boroni, Rocha, Dias, de Laet Souza, Oliveira, Bitar, Macedo, Machado, Caliani and Franco. 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) or licensor 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.