Jéssica Silqueira Hickson Rios

Universidade Federal de Minas Gerais

IMPACTO DO PROCESSAMENTO DE TRANSCRITOS POR *SPLICED LEADER TRANS-SPLICING* NO REPERTÓRIO PROTEICO DE *Schistosoma mansoni*

Belo Horizonte
2017

Jéssica Silqueira Hickson Rios

Universidade Federal de Minas Gerais

IMPACTO DO PROCESSAMENTO DE TRANSCRITOS POR *SPLICED LEADER TRANS-SPLICING* NO REPERTÓRIO PROTEICO DE *Schistosoma mansoni*

Dissertação apresentada a Universidade Federal de Minas Gerais como requisito parcial para obtenção de título de Mestre pelo Programa de Pós-Graduação em Bioinformática.

Orientadora: Profa. D[ra] Glória Regina Franco

Co-orientadora: D[ra] Mariana Lima Boroni Martins

Belo Horizonte
2017

SUMÁRIO

RESUMO

O *Spliced Leader trans-splicing* (SLTS) é um mecanismo de inserção do éxon 5' ou *Spliced Leader* (SL) de RNAs específicos (SL RNAs) em moléculas de mRNAs receptoras. A relevância do SLTS foi atribuída a mecanismos de regulação pós-transcricional, como alterações na estabilidade das moléculas de mRNA, o melhoramento da eficiência do processo de tradução e o aumento da diversidade do repertório proteico. A participação do SLTS nestes processos foi sugerida em diversos organismos, inclusive em *Schistosoma mansoni*. Todavia, o impacto do processamento de RNAs por SLTS nas proteínas do parasito ainda não é conhecido. A partir do desenvolvimento de programas nas linguagens Perl e Python para analisar dados de sequenciamento de transcritos processados por SLTS, este trabalho permitiu a descrição de diferentes formas de atuação do SLTS em transcritos codificadores de proteínas em *S. mansoni* e gerou uma concepção das situações frequentes e eventualidades deste processamento. Os resultados apontam que existe uma maior ocorrência de SLTS nos sítios de *trans-splicing* em regiões não traduzíveis (5' UTR). Qualitativamente, foi possível observar que os transcritos processados por SLTS podem produzir proteínas diferentes das produzidas por transcritos não processados, uma vez que a inserção do SL pode alterar a fase de leitura dos transcritos pelos ribossomos. A maioria dos transcritos que possuem mais de um sítio de entrada alternativa do SL produzem peptídeos menores, sem que seja observado um tamanho padrão dos peptídeos gerados; entretanto, a maioria deles tem a fase de leitura alterada e não portam metionina alternativa para início da tradução. Assim, peptídeos pequenos podem estar sendo produzidos provavelmente sem função e a síntese de algumas proteínas pode ser perdida. A maior parte das proteínas geradas a partir dos transcritos sem mudança de fase de leitura perdem grandes porções ou sequências completas de domínios funcionais, o que reforça a ideia de que o SLTS pode prejudicar ou impedir a atividade das proteínas processadas. Os dados gerados mostram a diversidade de funções do mecanismo de SLTS que impactam diretamente na regulação da expressão gênica e consequentemente em proteínas do *S. mansoni*.

ABSTRACT

The Spliced Leader trans-splicing (SLTS) is an insertion mechanism of 5' exon (Spliced Leader or SL) of specific RNAs (SL RNA) in mRNAs receptor molecules. The relevance of SLTS was attributed to mechanisms of post-transcriptional regulation, for changes in stability of mRNA molecules, to improve translation and to increase the diversity of the protein repertoire. The participation of SLTS in these processes has been suggested in various organisms, including *Schistosoma mansoni*. However, the impact of processing of the RNAs by SLTS on the whole parasite protein set is not known. By developing programs in Perl and Python languages, to parse data from sequencing of transcripts processed for SLTS, this work allowed the description of different forms of SLTS performance in protein-coding transcripts in *S. mansoni* and produced the conception of frequent situations and contingencies in this process. The results show that there is a higher incidence of SLTS in trans-splicing sites in non-translatable regions (5' UTR). Qualitatively, it was possible to observe that transcripts processed by SLTS can produce different proteins from those produced by native transcripts since the insertion of the SL can change the reading frame of the transcripts by ribosomes. Most transcripts that have more than one point of SL alternative entrance are liable to produce smaller peptides, without observed a standard size of peptides generated, however, most of them have changed the reading frame and do not carry alternative methionine for initiation of translation. Thus, small peptides can be produced without being functional and the synthesis of some proteins may be being lost. The majority of the proteins generated from transcripts without changes in reading frame lose large portions or complete sequences of functional domains, which reinforces the idea that the SLTS may hinder or prevent the activity of the processed protein. The data show the diversity of roles in the mechanism of SLTS that impact directly in the regulation of gene expression and consequently in proteins of the *S. mansoni*.

LISTA DE ABREVIATURAS

WHO – World Health Organization

EM – Esquistossomose mansônica

PZQ – Praziquantel

MEGs – *Micro-exon genes*

NGS – *Next Generation Sequencing*

RNA-seq – *RNA-sequencing*

UFMG – Universidade Federal de Minas Gerais

CPqRR - Centro de Pesquisas René Rachou

snRNAs – *Small nuclear ribonucleic acids*

snRNPs – *Small nuclear ribonucleoproteins*

NTC – *NineTeen Complex*

SLTS – *Spliced Leader trans-splicing*

SL – *Spliced Leader*

cDNA – DNA complementar

CDS – *Coding sequence*

Sle – Porção exônica do SL

TMG – Trimetilguanosina

ORF – *Open reading frame*

UTR – *Untranslated regions*

NCBI – National Center for Biotechnology Information

GO – *Gene ontology*

uORFs – *Upstream open reading frame*

RNAi - *RNA interference*

ÍNDICE DE ILUSTRAÇÕES

ÍNDICE DE TABELAS

1 INTRODUÇÃO

1.1 Aspectos gerais da esquistossomose

A esquistossomose é uma doença parasitária que atinge principalmente a América do Sul, a Ásia e toda África subsaariana. Os agentes etiológicos mais comúns no hospedeiro humano são as espécies *Schistosoma mansoni* (*S. mansoni*), *Schistosoma haematobium* e *Schistosoma japonicum* (GROSSE, 1993). A figura 1 apresenta a distribuição mundial da doença.

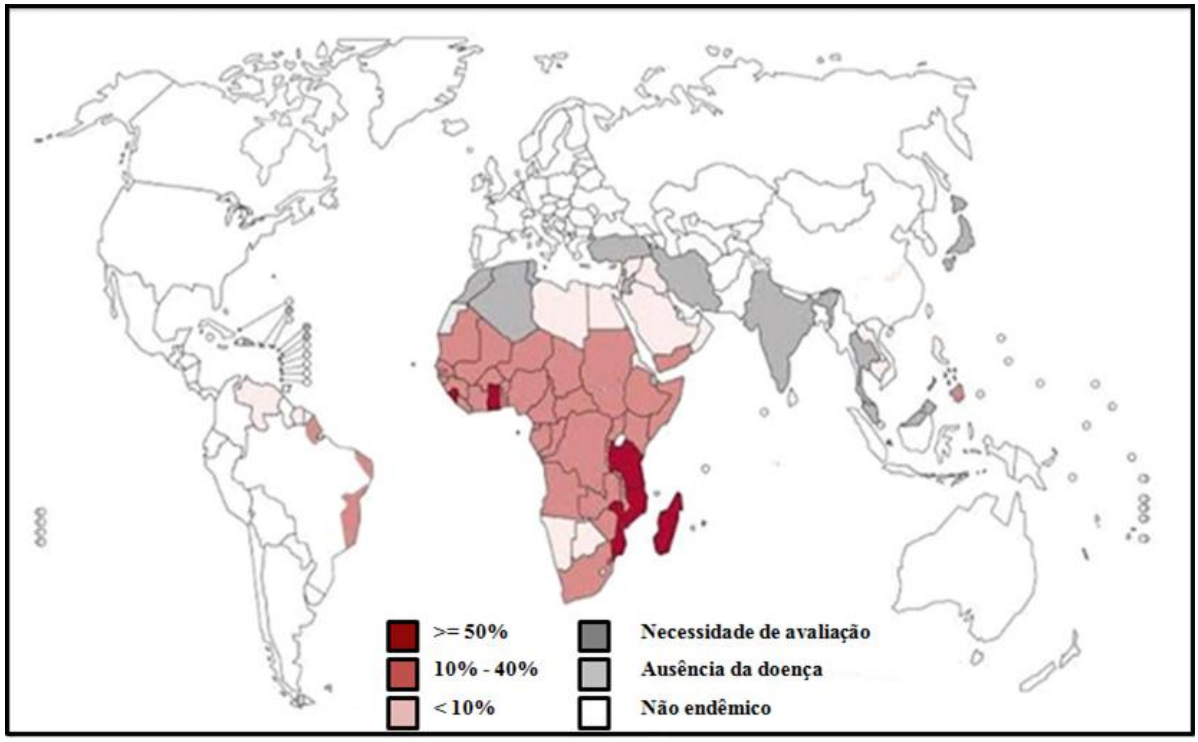


Figura 1: Distribuição mundial da esquistossomose. As cores representam regiões de prevalência maior ou igual a 50% (vermelho), de 10% a 40% (rosa escuro), menor que 10% (rosa claro), países que necessitam de avaliação para comprovar se foi alcançada a interrupção da transmissão (cinza escuro), sem a doença (cinza claro) e países não endêmicos (branco) (GROSSE, 1993).

Fonte: World Health Organization. Control of Neglected Tropical Diseases [reproduzida de (WHO, 2013)].

Detectada pela primeira vez em 1907 no estado da Bahia, a esquistossomose no Brasil ainda é considerada como grave problema de saúde pública, uma vez que prevalece em diversas regiões como causa de mortes. A propagação e persistência da doença no país são fatores diretamente relacionados com os movimentos migratórios combinados com as condições precárias de saúde, assim como com a disseminação dos hospedeiros intermediários (AMARAL; TAUIL; LIMA, 2006).

No território brasileiro, existem cerca de 2,5 milhões de indivíduos infectados pela esquistossomose mansônica (EM). As espécies de hospedeiros intermediários que se infectam naturalmente no estado de Minas Gerais, são os caramujos *Biomphalaria glabrata, B. straminea e B. tenagophila* (GUIMARÃES; FREITAS; DUTRA *et al*., 2013).

A patogenia da EM é caracterizada por uma reação inflamatória granulomatosa que pode ocasionar fibrose tecidual, hepatoesplenomegalia, hipertensão portal, ascite e varizes esofágicas (CARVALHO; COELHO; LENZI, 2008).

O ciclo de vida do parasito acontece quando fêmeas adultas depositam ovos nas veias mesentéricas do homem, seu hospedeiro definitivo. Ao defecar perto de ambientes aquáticos, este hospedeiro contamina as águas ao eliminar ovos juntamente com as fezes. Após serem depositados na água, os ovos liberam miracídios. Caso uma das espécies do hospedeiro intermediário habite este ambiente, os miracídios infectam estes caramujos, onde se transformam em cercárias. Estas são as formas que abandonam o organismo dos caramujos e ao encontrarem com o hospedeiro definitivo, infectam o mesmo, penetrando a pele e adquirindo a forma de esquistossômulos. Na fase adulta, os vermes migram para o sistema porta hepático e depois para o intestino, normalmente no mesentério e em todo intestino grosso. Em suas diferentes fases de desenvolvimento, o parasito não só sofre alterações morfológicas, como também fisiológicas e bioquímicas. Estas características denotam sua necessidade por uma regulação minuciosa da expressão gênica, o que evidencia a importância de estudos direcionados ao assunto (CARVALHO; COELHO; LENZI, 2008). A figura 2 mostra o ciclo de vida do parasito.

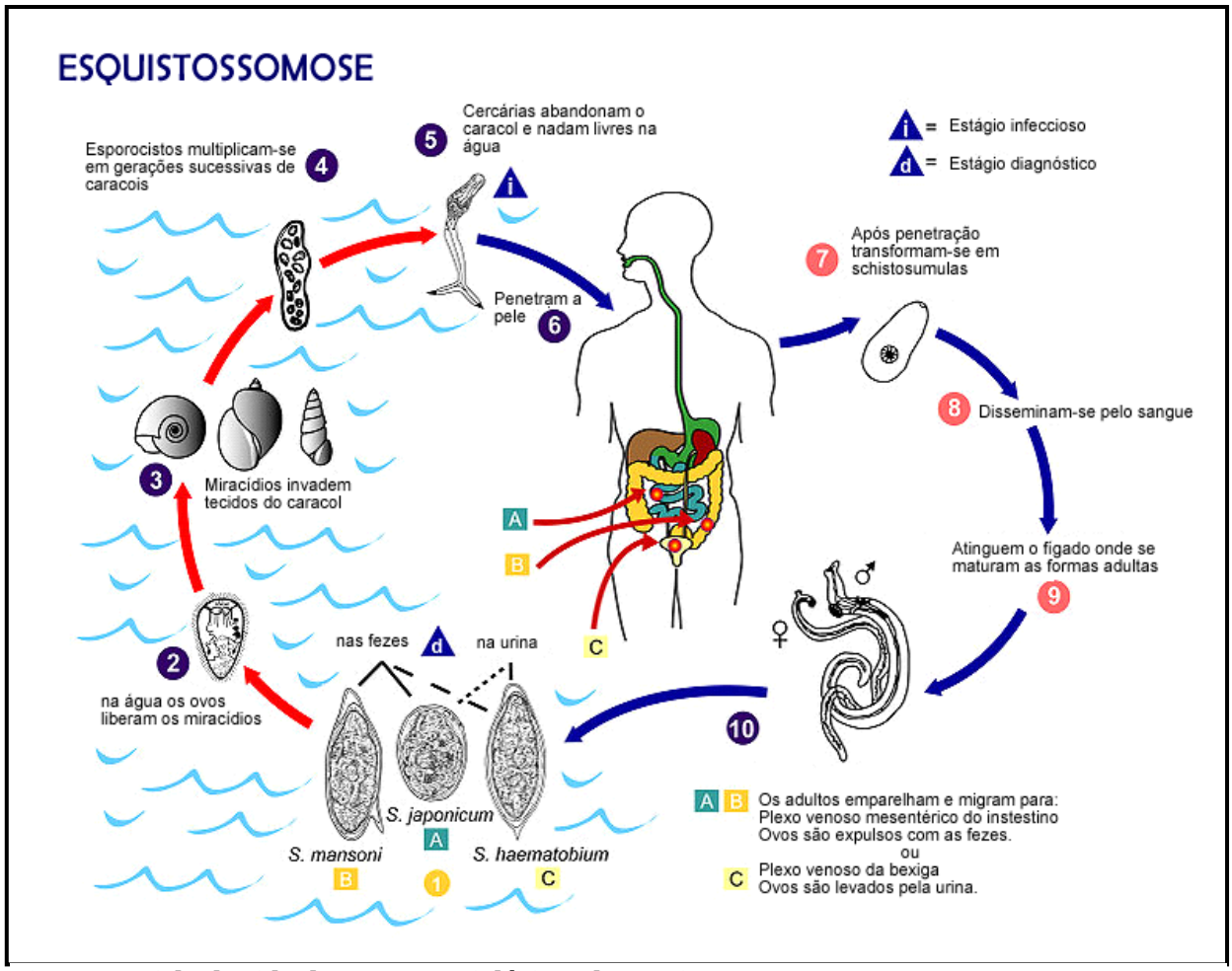


Figura 2: Ciclo de vida dos agentes etiológicos da esquistossomose.

Adaptado de: CDC http://www.cdc.gov/parasites/schistosomiasis/biology.html

O fármaco de escolha para tratamento da esquistossomose é o Praziquantel (PZQ), que atua na paralisação do agente e aumento da permeabilidade da membrana de suas células (COURA; CONCEIÇÃO, 2010). Esse fármaco foi descrito como inibidor dos canais de cálcio (PICCA-MATTOCIA *et al*., 2007) e como causador de danos no sistema excretor do parasito (OLIVEIRA *et al*., 2006).

O PZQ tem pouca efetividade contra as formas mais jovens do parasito e não contribui para prevenção de re-infecção, o que justifica uma necessidade de repetição do tratamento de tempos em tempos (WILSON; COULSON, 1999). Além disso, o PZQ está associado ao desenvolvimento de resistência medicamentosa dos parasitos frente ao seu mecanismo de ação (BOTROS; BENNETT, 2007). Estes também são aspectos que justificam a importância de estudos genéticos do parasito.

A maior parte dos infectados reside em regiões que não são submetidas a medidas de controle epidemiológico e pesquisas científicas (COURA; CONCEIÇÃO *et al*., 2010). Estas estratégias incluem o saneamento básico, aplicação de moluscicida e educação de saúde (KATZ; COELHO, 2008).

1.2 O Genoma de *S. mansoni*

O genoma do *S. mansoni* é constituído por sete pares de cromossomos autossomos e um par de cromossomos sexuais que, por sua vez, são representados como ZW nas fêmeas e ZZ nos machos. Em 2009, Berriman e colaboradores publicaram as primeiras informações do genoma de *S. mansoni*. Este foi descrito contendo 363 mega pares de bases, com pelo menos 11.809 genes codificadores de aproximadamente 13.197 transcritos e sendo caracterizados por longos *íntrons* (~1692 pb) e éxons menos extensos (~217 pb) (BERRIMAN *et al*., 2009).

A região codificante do genoma apresenta uma grande quantidade (~75% de sua sequência codificadora) de *micro-exon genes (*MEGs), em cujas sequencias foram identificadas diversas ocorrências de variantes de *splicing* alternativo derivadas de exclusão de éxons. É de extrema importância a compreensão da origem e funcionalidade desses eventos (BERRIMAN *et al*., 2009).

Com o auxilio das técnicas de sequenciamento de Sanger e *Next Generation Sequencing* (NGS), este genoma teve sua montagem bastante melhorada. A versão mais recente do genoma corresponde a 364,5 milhões de bases, além de ter permitido a descoberta de novos genes e variantes de *splicing* (PROTASIO *et al*., 2012).

1.3 O mecanismo de *splicing*

O *splicing* é o mecanismo de remoção de íntrons e união de éxons de uma mesma molécula de RNA, sendo o *cis-splicing* sua forma mais comum. O processamento foi decoberto em 1977 em adenovírus (BERGET; MOORE; SHARP, 1977; CHOW *et al*., 1977) e permite a maturação de pré-mRNAs pelo recrutamento de pelo menos cinco diferentes *small nuclear ribonucleic acids* (snRNAs) e proteínas que constituem o spliceossomo (WILL; LUHRMANN, 1997).

O processo é realizado pelo spliceossomo, um complexo ribonucleoprotéico composto por cinco pequenos RNAs nucleares (nRNAs), são eles: U1, U2, U4, U5 e U6. Os snRNAs encontram-se associados às proteínas, formando partículas ribonucleoproteicas (snRNPs). A construção e remodelamento destes snRNPs com o pré-mRNA, envolve duas etapas que caracterizam o processo de *cis-splicing*. Antes da iniciação do *cis-splicing*, o U1 snRNP é o primeiro componente que entra em contato com o RNA, sendo ligado a sítios doadores 5' (GU). Em seguida, o U2 snRNP associa-se a sítios receptores 3' (AG) normalmente associados a regiões de polipirimidinas e ao sítio de ramificação. A união de U1 + U2 + pré-mRNA é chamada de complexo A. Juntamente com o U1, o U2 permite que o *cis-splicing* ocorra no sentido correto. Subsequentemente, o snRNP triplo U4/U6·U5 é aderido. O complexo B é a união de U1 + U2 + U4/U6·U5 + pré-mRNA. O complexo B é associado ao *NineTeen Complex* (NTC) ou Prp19 para formação do spliceossomo (revisado em MATERA; WANG, 2014).

Posteriomente, o spliceossomo sofre dissociação dos snRNPs U1 e U4, passando de complexo B para complexo B*, que é a forma pronta para o início da primeira etapa do *cis-splicing*. Como consequência, o spliceossomo é ativado por catálise e ocorre a formação de um éxon 5' livre e um laço intermediário com o *íntron* ligado ao éxon 3' por transferência da ligação fosfodiéster entre o fosfato do sítio doador e uma hidroxila do ponto de ramificação. Desta forma, o complexo B* é rearranjado para formação do complexo C por remoção e adição de proteínas específicas. A estrutura sofre rearranjos consecutivos e o mecanismo prossegue com a remoção dos íntrons em laço e junção dos éxons. O complexo resultante é o *post-spliceosomal*, que em seguida é desmontado. Os snRNPs são reutilizados para montagem de novos spliceossomos por enzimas responsáveis pela hidrólise de ATP. A figura 3 ilustra cada etapa da montagem do spliceossomo (revisado em MATERA; WANG, 2014).

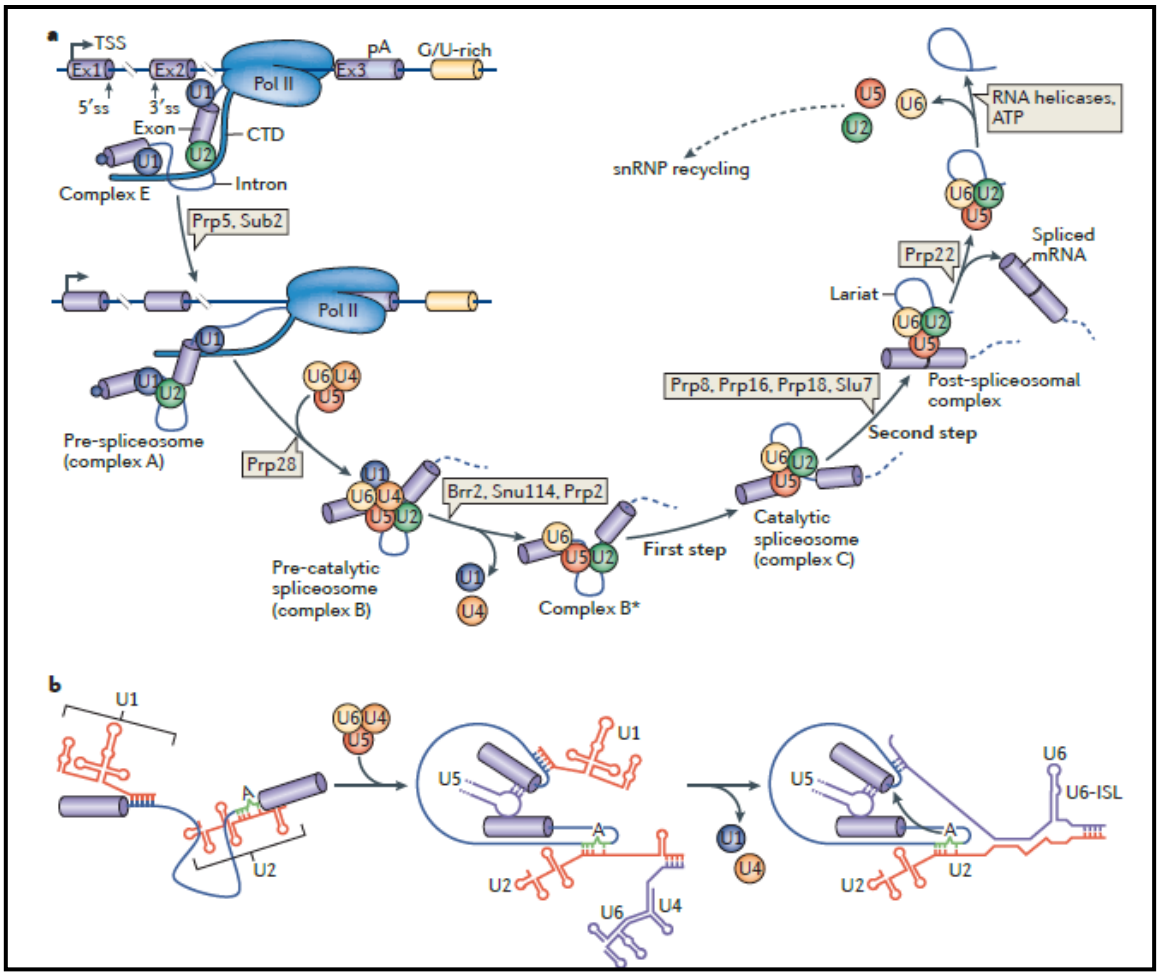


Figura 3: Montagem do spliceossomo. Recrutamento e rearranjo dos U snRNPs elucidando o gasto de energia em forma de ATP e participação de proteínas: (A) U1 e U2 se associam por reconhecimento dos sítios de *splicing* por mediação do domínio carboxi terminal (CTD) da polimerase II. U1 e U2 interagem para a formação de um pré-spliceossomo (complexo A). O snRNP triplo U4/U6·U5 é recrutado para formação do complexo B, que sofre rearranjos para tornar-se ativo (complexo B*). O U1 e o U4 são liberados. O complexo B* conclui a primeira etapa do *splicing* gerando o complexo C, que contém o éxon 1 livre (Ex1) e um laço intermediário entre o íntron e o éxon 2. O complexo C sofre rearranjos e finaliza o segundo passo do *splicing* com a geração de um complexo *post-spliceosomal* com os íntrons em laço e os éxons. O U4/U6·U5 é dissociado e os mRNPs reciclados para os próximos eventos de *splicing*. (B) Rearranjos no processo de *splicing*: o U1 e o U2 unem com a sequência de adenosina (A) do ponto de ramificação. Depois, o U4/U6·U5 é associado gerando novos pares de base entre U2 e U6 e U5 e sequências exônicas. O U4 é dissociado do U6 para expor a extremidade 5’ do U6 e o U1 é afastado. Uma rede de interações de pares de bases é feita entre o U6 e o U2 justapondo o sítio 5’ e o A. O U6 forma um *stem-loop* (U6 – ISL) necessário como sinal para a catálise do *splicing*.

Fonte: MATERA; WANG, 2014.

Cada nucleotídeo da região intrônica precisa ser removido com precisão, caso contrário, a fase leitura do mRNA poderá ser alterada podendo causar variações a produção de proteínas aberrantes (STALEY; GUTHRIE, 1998).

Para que o processo de *splicing* seja realizado corretamente, os elementos que devem ser reconhecidos são os sítios aceptores e doadores e o trato de polipirimidina (revisado em MATERA; WANG, 2014).

1.4 Variações de *splicing*

Diferentes variações de *splicing* foram descritas em eucariotos, dentre estas estão os eventos de: (l) *cis-splicing*; (ll) *trans-splicing* intragênico, que envolve a junção de éxons de moléculas de mRNAs idênticas; (lll) trans-*splicing* intergênico, em que ocorre a junção de éxons de mRNAs derivados de genes diferentes, podendo pertencer a cromossomos distintos (*trans-splicing* intercromossomal); (lV) *exo-endo trans-splicing*, quando um transcrito derivado de um plasmídeo exógeno é unido em *trans* com um mRNA endógeno; (V) *Spliced Leader trans-splicing* (SLTS), em que uma molécula específica de SL RNA doa um mini éxon para mRNAs (PREUßER; BINDEREIF, 2013; PREUßER; JAÉ; BINDEREIF, 2012).

Os processamentos de *trans-splicing* intragênico e intergênico são pouco conhecidos e especializados na formação de mRNAs funcionalmente importantes que parecem ser raros. Muitos destes eventos foram associados com o desenvolvimento de câncer e translocações cromossômicas (LI *et al*., 2008). *No exo-endo trans-splicing*, o mRNA produzido é quimérico e codificador de uma proteína funcional (PREUßER; BINDEREIF, 2013). No *Spliced Leader trans-splicing* (SLTS), uma sequência denominada *spliced leader* (SL), presente na extremidade 5’ de um RNA pequeno e especializado (SL RNA), é doada para alguns pré-mRNAs receptores, formando o éxon 5’ terminal de mRNAs maduros. Além do produto de *splicing* composto pelo SL e o pré-mRNA, ocorre a formação de uma estrutura em y proveniente da ligação do sítio de *splicing* 5' do SL RNA em uma adenina do *outron* (sequência de nucleotídeos no pré-mRNA que antecede sítio de *splicing* no qual o SL é inserido) (BLUMENTHAL, 2005). A figura 4 apresenta um esquema dos cinco tipos de processamentos de *splicing* descritos.

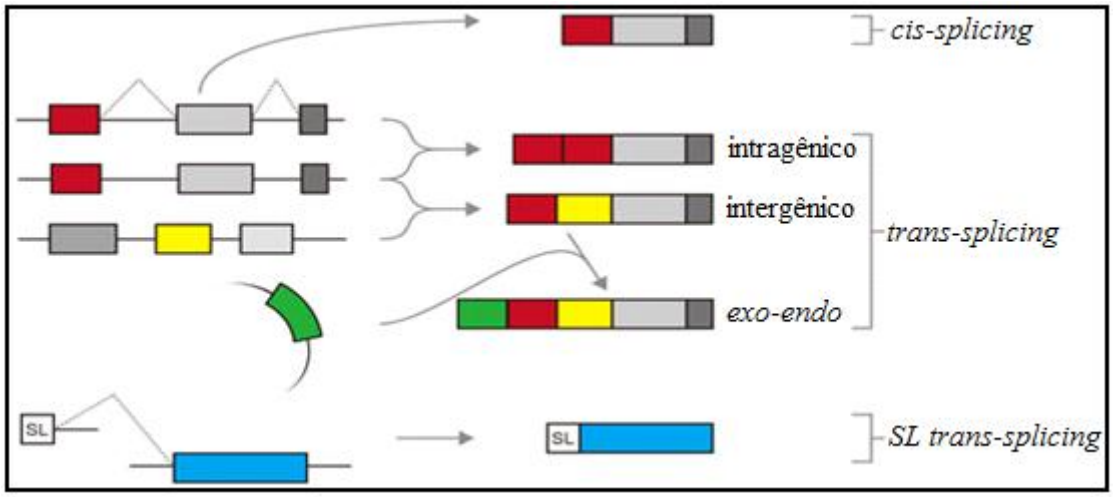


Figura 4: Formas de processamento de *splicing* em eucariotos. *Cis-splicing*, *trans-splicing* intragênico (exibido com uma resultante duplicação de éxons), *trans-splicing* intergênico, *exo-endo trans-splicing* e *SL trans-splicing*. Os íntrons estão representa pelas linhas, e *éxons* pelas caixas (*éxons* de mesma cor são idênticos e de cores diferentes são distintos).
Fonte: PREUßER; BINDEREIF, 2013.

O spliceossomo tem preferência pelo doador de *cis-splicing* que estiver disponível, porém, o mecanismo básico para esta característica é desconhecido (HASTINGS, 2005). Comparado ao *cis-splicing*, o *trans-splicing* utiliza muitos dos mesmos componentes spliceossômicos e sinalizações (LASDA; BLUMENTHAL, 2011). Além disso, alguns organismos usam o mesmo sítio de *splicing* 3' alternativamente para *trans* e *cis-splicing*. Eventos de *trans-splicing* foram relatados em rotíferos, (POUCHKINA-STANTCHEVA; TUNNACLIFFE, 2005), tunicatos (MATSUMOTO *et al*., 2010), tripanosomatídeos (HELM; WILSON; DONELSON *et al*., 2008), crustáceos (LASDA; BLUMENTHAL, 2011) e nematódeos (PETTITT *et al*., 2008), assim como em *S. mansoni* (MOURÃO *et al*., 2013).

Em tripanosomatídeos, a utilização de um sítio de *splicing* para processamentos de *splicing* em *cis* ou *trans* é uma ocorrência que pode depender da presença de sítios canônicos ou pode ser modulada pela U1 snRNP ou por proteínas SR. Foi demonstrado que o spliceossomo pode conter tanto o U1, quanto o SL RNP, havendo competição de ligação entre ambos, o que pode ser um mecanismo determinante para a ocorrência de *cis* ou *trans-splicing* (LIANG *et al*., 2003).

Em *S. mansoni*, a presença de sítios aceptores canônicos mais fortes e de um trato de polipirimidina maior, são determinantes da ocorrência de *splicing* em *cis* ou *trans* (BORONI, 2014).

1.5 *Spliced Leader trans-splicing*

1.5.1 Descoberta e distribuição filogenética

O processamento de SLTS foi descoberto em 1982, em transcritos de *Tripanosoma brucei* codificadores de glicoproteínas variáveis de superfície (VSGs). A sequência de 39 nucleotídeos, comum entre estes organismos, foi denominada *Spliced Leader* (SL) e estava contida no nomeado SL RNA (BOOTHROYD; CROSS, 1982). A descoberta ocorreu durante a comparação de sequências genômicas e de DNA complementar (cDNA), quando uma sequência curta de um gene, que não pode ser codificado, foi encontrada nas extremidades 5' de vários RNAs (revisado em LASDA; BLUMENTHAL, 2011).

Mais tarde, o SLTS foi descrito no nematódeo *Caenorhabditis elegans* (ABUBUCKER *et al*., 2014). A porção exônica do SL (SLe) possui tamanhos variados nos diferentes organismos, podendo ser constituída por 22 a 25 ou 36 a 39 nucleotídeos (BITAR *et al*., 2013). Alguns nematódeos possuem uma única classe de SL RNA para todos SLTS. No entanto, *C. elegans* e organismos relacionados, possuem duas classes distintas de SL RNAs (SL1 e SL2) que são utilizados de formas diferentes (ALLEN *et al*., 2011). O SL1 é inserido em *trans* em *outrons* de transcritos monocistrônicos e o SL2 é utilizado na resolução de transcritos policistrônicos (BLUMENTHAL, 2005).

Em organismos que possuem mais de uma sequência SL, diferentes sequências de SLe são inseridas em um conjunto de transcritos específicos. Isso sugere que cada sequência SLe está relacionada com um determinado conjunto de transcritos, ou seja, a expressão de seus respectivos repertórios de proteínas, pode ser controlada de acordo com a presença de uma determinada SLe e o SLTS pode ser regulado por mudanças ambientais e gerar uma resposta específica de acordo com os transcritos que são processados (BITAR *et al*., 2013).

Anos depois, o mecanismo de SLTS foi também encontrado em *S. mansoni* (MOURÃO *et al*., 2013) e no protista *Euglena gracilis* (FRANTZ *et al.,* 2000). Na década seguinte, o SLTS foi reportado em filos adicionais. O evento foi descrito na classe *Appendicularia* (GANOT *et al*., 2004), em crustáceos anfípodes, crustáceos copépodes, cordados primitivos (tunicados), *chaetognaths*, rotíferos *bdelloid*, em todas as espécies de platelmintos onde foi estudado em, cnidários hidrozoários, *ctenophores* e em diferentes espécies de dinoflagelados (LASDA; BLUMENTHAL, 2011). O SLTS parece ser empregado em genes de diferentes funções biológicas em cada filo (BITAR *et al*., 2013). Interessantemente, um número significativo de genes que geram moléculas de mRNA que não sofrem SLTS, também geram um número detectável de moléculas de mRNA que sofrem o processamento. Esta observação foi demonstrada em um estudo realizado em 2010 por Matsumoto e colaboradores por caracterização de genes de *Ciona intestinales* que sofreram *trans-splicing* (MATSUMOTO *et al*, 2010).

1.5.2 Componentes e mecanismo

Os SL RNAs existem como pequenos snRNPs complexados à proteínas Sm. São similares aos U snRNAs, sendo portanto, constituídos de motivos U-RNA, que por sua vez são regiões de ligação das proteínas Sm (AGABIAN, 1990). Algumas sequências líderes possuem uma hipermetilação m2, 2,7 GpppN trimetilguanosina (TMG) 5’ cap (cap 4 em tripanossomas) (LASDA; BLUMENTHAL, 2011).

No *cis-splicing*, os U snRNPs (U1, U2, U4, U5 e U6) atuam no reconhecimento dos sítios de *splicing* sobre a molécula de RNA sendo, portanto, fundamentais na remoção dos íntrons (LASDA; BLUMENTHAL, 2011). O SLTS usa a mesma maquinaria básica do spliceossomo utilizada no *cis-splicing* e os mesmos sítios doadores (GU) e aceptores (AG) (DENKER; ZUCKERMAN, 2002). No entanto o SLTS não necessita da participação do U1 (MATERA; WANG, 2014). A principal distinção entre o U1 e o SL é que diferente do U1, o SL é consumido durante a reação (LASDA; BLUMENTHAL, 2011). Um esquema do mecanismo de SLTS convencional é apresentado na figura 5.

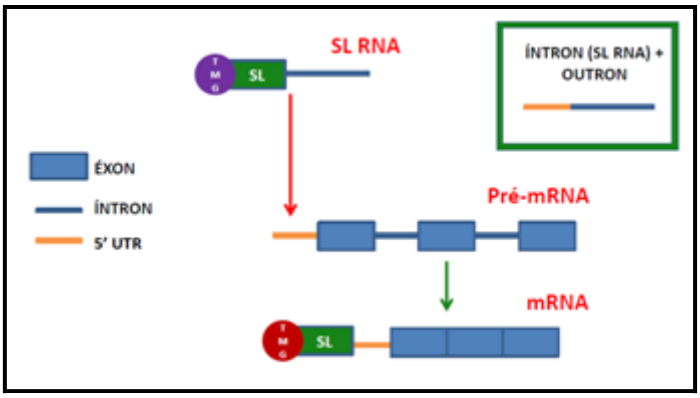


Figura 5: Esquema da forma convencional de SLTS (sítio de *splicing* na região 5' UTR).

1.5.3 Funções descritas

Estudos do SLTS em cinetoplastídeos (PREUßER; JAÉ; BINDEREIF, 2012), nematódeos (ABUBUCKER *et al*., 2014) e platelmintos (MOURÃO *et al*., 2013) permitiram avanços nos conhecimentos sobre a funcionalidade desse mecanismo. Nos diferentes filos, uma fração variante de populações de mRNAs sofre *trans-splicing*. Nos protistas cinetoplastídeos, aparentemente todas as moléculas de RNA sofrem SLTS. Isso ocorre porque muitos ou todos os seus genes são expressos como SL-*resolved operons* (CAMPBELL; THOMAS; STURM, 2003). *Operons* são unidades de transcrição constituídas por *clusters* em *tandem* de poucos ou muitos genes que caracterizam um transcrito policistrônico. Estes genes normalmente possuem funções relacionadas, sendo co-expressos e co-regulados por um único promotor (revisado em BLUMENTHAL, 1998). Genes *housekeeping*, por exemplo, são especificamente compatíveis com a organização de *operons*, pois eles presumivelmente não requerem controle transcricional independente (BLUMENTHAL; GLEASON, 2003).

O processamento de transcritos policistrônicos é a função mais bem conhecida do SLTS, atuando na separação de transcritos que podem ou não ter funções relacionadas. Independente da quantidade, tamanho ou distância entre os genes de transcritos policistrônicos, o SL *trans-splicing* é utilizado para transformar RNAs policistrônicos em RNAs maduros monocistrônicos por sítios aceptores *upstream* de cada *open reading frame* (ORF). RNAs mensageiros individuais produzidos por precursores policistrônicos podem então, ser exportados para fora do núcleo e

traduzidos (LASDA; BLUMENTHAL, 2011). Estes transcritos policistrônicos diferem dos *operons* de bactérias, que podem também ser encontrados ocasionalmente em plantas e moscas como transcritos dicistrônicos ou policistrônicos uma vez que passam por um processo de maturação do RNA e posterior tradução (BLUMENTHAL, 2004).

Outra função descrita do SLTS é a remoção da porção 5’ do pré-mRNA monocistrônico *upstream* ao sítio aceptor de *trans-splicing*, o *outron*, que pode conter elementos que possam comprometer o transporte, tradução ou estabilidade da molécula de mRNA. Por promover a remoção da região da extremidade 5’, o SLTS provavelmente facilita a evolução de novos sítios de iniciação de tradução (BLUMENTHAL, 1995) e pode permitir que os genes contenham uma grande variedade de componentes de sequência funcionais potencialmente úteis perto da extremidade 5’ do transcrito primário (por exemplo, a transcrição de diferentes elementos regulatórios do processo de tradução). O evento foi nomeado “sanitização da região 5’UTR e pode ser requerido em casos de existência de AUG *upstream* adicional que interfere na tradução da ORF correta (HASTINGS, 2005).

Em cinetoplastídeos, o SLTS foi descrito com a funcionalidade de fornecer a extremidade 5’ TMG cap para transcritos de RNA polimerase I não traduzíveis para permitir que eles possam ser traduzidos (HASTINGS, 2005; LASDA; BLUMENTHAL, 2011). A adição da sequência SL contendo o TMG cap modificado em cinetoplastídeos e nematódeos pode permitir um aumento da estabilidade do mRNA através de um recrutamento diferencial da maquinaria de tradução, e/ou alteração da região 5' UTR (LASDA; BLUMENTHAL, 2011). A figura 6 apresenta um esquema dos papéis descritos do SLTS.

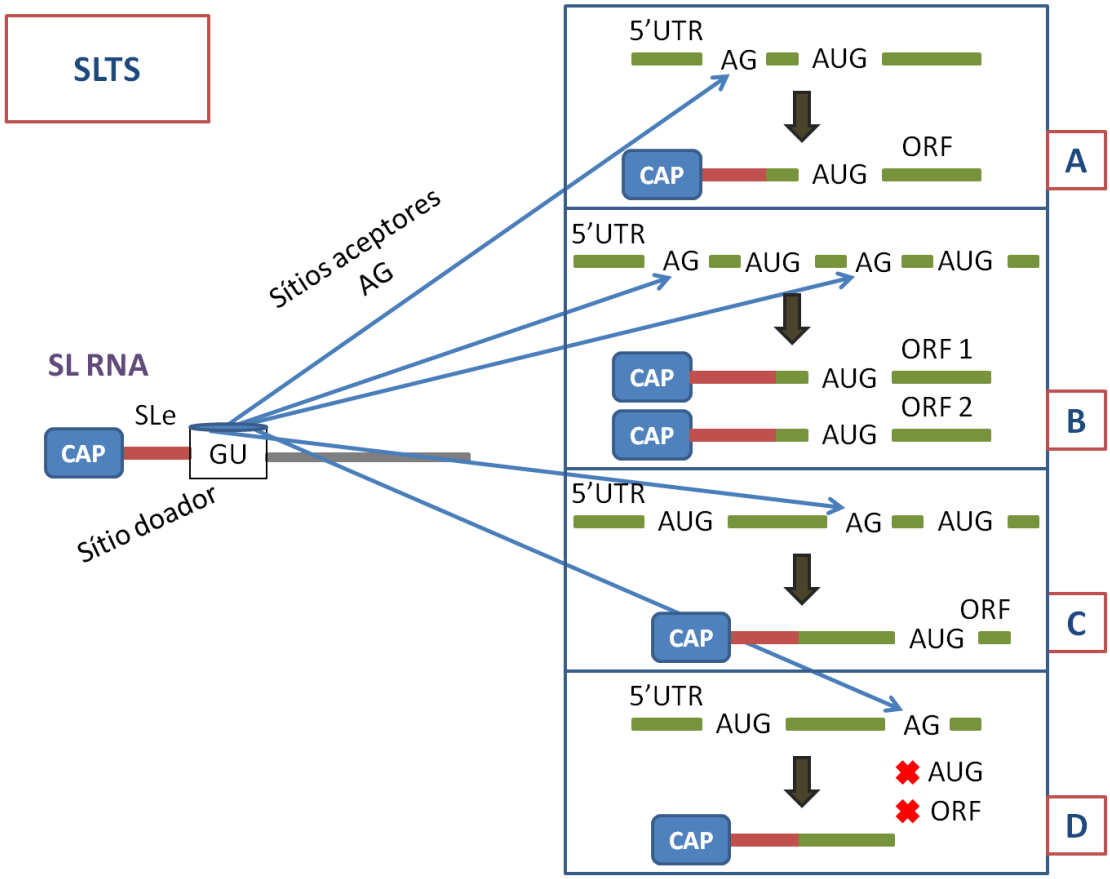


Figura 6: Visão geral das funções biológicas do SL *trans-splicing*. Ao lado esquerdo, o SL RNA é ilustrado com sua estrutura hiper modificada (TMG cap), sítio doador de *splicing* (GU), porção exônica (Sle) e porção intrônica. As setas azuis indicam eventos de *trans-splicing* de transferência do segmento de SL para sítios aceptores de *splicing* (AG) em mRNAs alvos (com representação de seus *codons* de iniciação). Estes eventos de *trans-splicing* tem os seguintes efeitos: (A) transformação de transcritos não traduzíveis por RNA polimerase l em traduzíveis através do ganho de uma estrutura cap; (B) resolução de *cistrons* intraduzíveis de pré-mRNAs policistrônicos através da separação e capeamento destes, tornando-os traduzíveis; (C) sanitização da região 5' UTR. Neste exemplo, o mecanismo está direcionado para a eliminação de um códon AUG *upstream* adicional juntamente com o segmento 5' descartado do mRNA (*outron*). Esta porção descartada pode interferir na eficiência da tradução da ORF e pode estar provocando instabilidade da molécula de RNA; (D) Regulação da expressão gênica via inativação da síntese proteica por ausência do códon de iniciação na sequência permanente após a inserção do SL.

Um estudo focado nas possíveis funções do SLTS, desenvolvido por Bitar e colaboradores em 2013, permitiu a análise de 450 transcritos com SLe em sequências de espécies de diversos filos. Os transcritos contendo o SL foram organizados em uma base de dados bastante completa, originada de uma busca por similaridade em outra base construída pelos autores a partir da identificação de anotações de sequências SLe no National Center for Biotechnology Information (NCBI). Todas as sequências recuperadas passaram por uma curadoria manual para exclusão de falsos positivos e redução de redundâncias. Cerca de metade dos transcritos contendo o SL foram encontrados em pelo menos dois organismos. A maior parte dos transcritos estava presente em todos eucariotos e suas proteínas atuam na síntese de ATP, metabolismo da glicose, dobramento de proteínas, defesa contra estresse oxidativo, replicação do DNA e tradução. Isso sugere que transcritos processados por SLTS atuam em funções antigas e conservadas (BITAR *et al*., 2013).

Em *C. intestinales*, foram descritos transcritos processados por SLTS associados a homeostase de Ca^{2+}, endomembranas e componentes do citoesqueleto/actina através de análise de *gene ontology* (GO) (MATSUMOTO *et al*., 2010).

1.5.4 *SL trans-splicing* em *S. mansoni*

A sequência SLe de *S. mansoni* consiste de 36 nucleotídeos (5'AACCGTCACGGTTTTACTCTTGTGATTTGTTGCATG3') proveniente de um SL RNA não poliadenilado composto por um total de 90 nucleotídeos. O mecanismo de SLTS não ocorre em todo conjunto de mRNAs do parasito (RAJKOVIC *et al*., 1990). Em 2013, Mourão e colaboradores identificaram transcritos que sofrem o processamento por SLTS em diferentes estágios do ciclo de vida do parasito, sugerindo o papel importante desse mecanismo na geração dos repertórios proteicos nos diferentes estágios de *S. mansoni*. Dentre os transcritos processados identificados, estão inclusos muitos codificadores de proteínas ribossomais e de algumas enzimas glicolíticas. (MOURÃO *et al*., 2013).

O mecanismo de SLTS não parece estar relacionado a uma determinada fase de desenvolvimento do parasito, sexo, tecidos e/ou genes específicos (MOURÃO *et al*., 2013). Esta é uma questão discutida desde 1995 por Davis e colaboradores, que indicaram que o SLTS não está associado a uma via específica, ao identificar quatro enzimas glicolíticas que não sofrem SLTS, enquanto que o transcrito derivado do gene

da enzima glicolítica enolase, que é derivada de um transcrito de SLTS (DAVIS *et al*., 1995).

Além disso, foi detectado que existe uma regulação diferencial por SLTS em todos os estágios de desenvolvimento do parasito. Esta regulação está associada com a entrada alternativa de SL em vários transcritos e consequente geração de isoformas distintas de transcritos (BORONI, 2014). O primeiro evento de SLTS alternativo descrito em *S. mansoni,* foi descoberto em 1990, após a identificação de um sítio de SLTS no terceiro éxon do transcrito codificador da proteína 3-hidroxi-3 metil-glutaril CoA redutase (RAJKOVIC *et al*., 1990). Mais recentemente, em 2013, o transcrito da ubiquinol-citocromo C-redutase, também foi identificado contendo um sítio de SLTS anterior ao segundo *exon*, além de conter um sítio aceptor *upstream* (MOURÃO *et al.*, 2013).

Entretanto, o impacto do SLTS nas proteínas derivadas destas diferentes isoformas de *trans-splicing* ainda permanece desconhecido. A figura 7 mostra um esquema das diferentes formas de SLTS alternativo que podem ocorrer. Como parâmetro de comparação, a figura do SLTS convencional (figura 5) também foi representada.

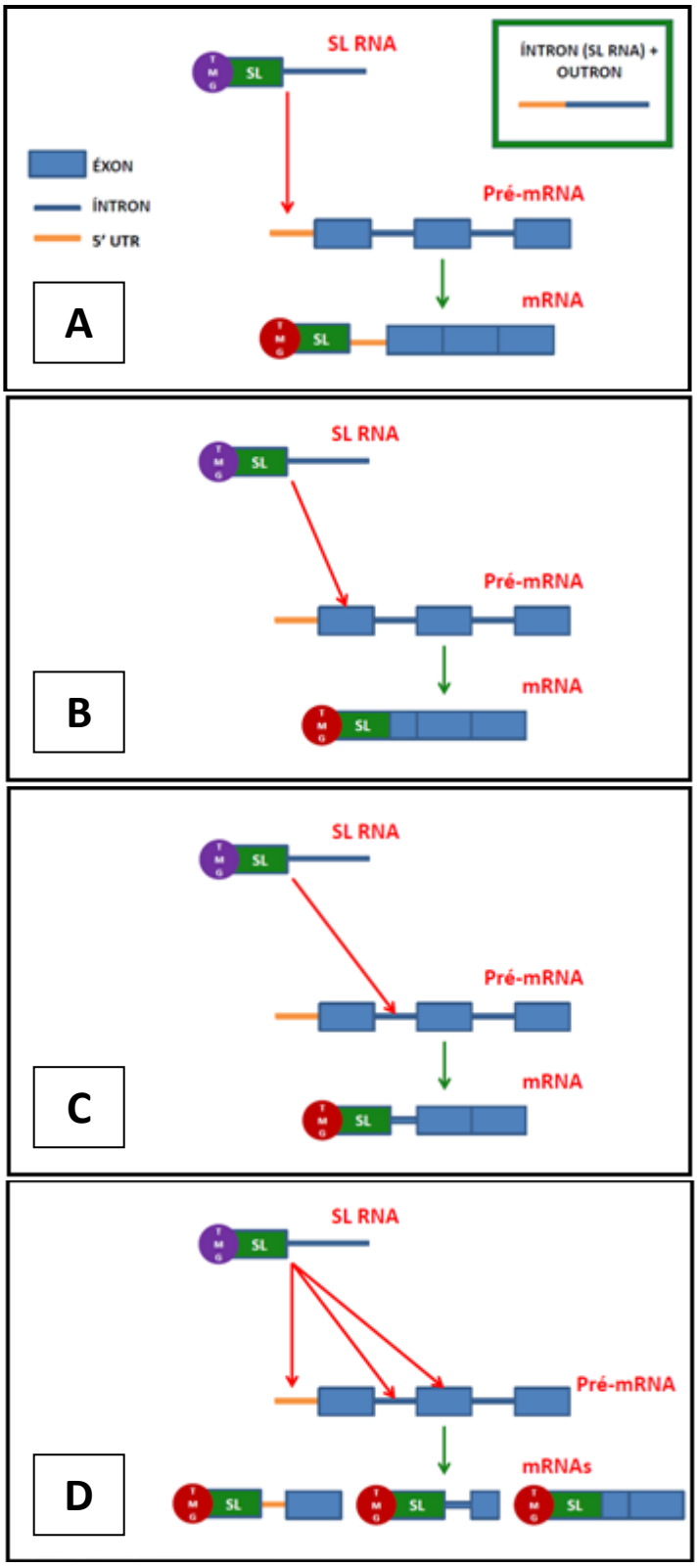


Figura 7: Formas alternativas e convencional de SLTS. O esquema em A (sítio de *splicing* na região 5' UTR) representa a forma convencional de SLTS. Em B (sítio de *splicing* no interior de uma *coding sequence* - CDS), C (sítio de *splicing* em região intrônica) e D (sítios na região 5' UTR, região intrônica e CDS).

1.6 Estudos de RNA-seq

O RNA-seq (*RNA Sequencing*) é uma tecnologia de *deep sequencing*, ou seja, sequenciamento profundo, capaz de detectar até mesmo RNAs transcritos com pouca frequência (WANG; GERSTEIN; SNYDER, 2009). O RNA-seq consiste na transformação de RNAs totais ou fracionados em bibliotecas de DNA complementar (cDNA), que são sequenciados necessitando ou não de amplificação, dependendo do tipo de tecnologia adotada (GARBER *et al.*, 2011).

A técnica foi desenvolvida para análise do perfil de *transcriptomas* e tem permitido a aplicação de novos métodos de mapeamento e quantificação gênica (NAGALAKSHMI *et al*., 2008).

A construção do *transcriptoma* de um determinado organismo através do RNA-seq, também possibilita a identificação de isoformas de RNA mais escassas provenientes de processamento por *splicing* alternativo (em *cis* e *trans*), podendo assim, auxiliar em estudos de expressão diferencial e características destes RNAs (WANG; GERSTEIN; SNYDER, 2009).

Em 2014 Boroni e colaboradores desenvolveram uma técnica denominada SL Trapping, capaz de detectar uma ampla gama de transcritos que sofrem SLTS em *S. mansoni*. A nova metodologia foi empregada utilizando um protocolo descrito por Nilsson e colaboradores em 2010. Foram capturados os mRNAs poliadenilados e o cDNA produzido foi purificado e eluído para sequenciamento no equipamento Illumina HiSeq 2000. Posteriormente as *reads* contendo a sequência SL foram filtradas por meio de ferramentas de bioinformática. Os dados gerados permitiram o estudo dos sítios de inserção da sequência SL e foram reproduzidos empregando outras metodologias cujos dados gerados apresentaram grande sobreposição em relação às bibliotecas construídas no primeiro experimento.

A técnica aumentou o limite de resolução do transcriptoma processado, permitindo a detecção de eventos raros de SLTS que ainda não haviam sido detectados. Algumas bibliotecas geradas no trabalho do grupo foram utilizadas no presente estudo, especificamente duas réplicas biológicas de *S. mansoni* nas fases de miracídio, esporocisto, esquistossômulo e verme adulto. Para estas fases uma amostra de RNA total foi purificada para captura dos mRNAs através de complementaridade com a cauda Poli-A, os cDNAs foram produzidos, purificados e as extremidades dos fragmentos foram reparadas e ligadas a adaptadores. Parte da biblioteca equalizada foi amplificada

por PCR em emulsão, ligadas a esferas magnéticas e posteriormente as esferas positivas foram enriquecidas e sequenciadas pelo equipamento Ion Torrent PGM™ System (Life Technologies), produzindo o conjunto de dados denominado SL Enriched (BORONI, 2014).

2 JUSTIFICATIVA

O entendimento de aspectos biológicos de *S. mansoni* é essencial para o desenvolvimento de estratégias futuras de controle e quimioterapia da EM. Especificamente, a compreensão do SLTS no parasito pode fornecer informações importantes para a elaboração de fármacos que podem atuar neste tipo de mecanismo.

Estudos apontam para a existência de uma regulação diferencial do SLTS entre diferentes formas de vida de *S. mansoni*. Estes mecanismos estão relacionados com a entrada alternativa do SL com geração de novas isoformas de transcritos em fases distintas ou não. Desta forma, podem estar sendo utilizados sítios *upstream*, como também sítios *downstream* do sítio original de entrada do SL, que podem viabilizar ou inviabilizar a síntese proteica. Estas características necessitam de uma investigação mais abrangente e sugerem um importante papel do SLTS na regulação da expressão gênica no parasito (BORONI, 2014).

3 OBJETIVOS

3.1 Objetivo geral

Descrever possíveis consequências do processamento por *Spliced Leader trans-splicig* no repertório proteico predito de *S. mansoni*.

3.2 Objetivos específicos

- Identificar os transcritos que sofrem SLTS obtidos de bibliotecas de RNA-seq de diferentes fases de desenvolvimento do ciclo de vida do parasito (miracídio, esporocisto, esquistossômulo e vermes adultos).

- Verificar o posicionamento dos sítios aceptores de SLTS nos transcritos;

- Pesquisar mudança na fase de leitura dos transcritos processados.

- Comparar os domínios proteicos traduzidos das sequências dos transcritos processadas com os seguimentos completos de aminoácidos das respectivas proteínas codificadas por estes transcritos.

4 METODOLOGIA

4.1 Origem dos dados de sequenciamento por RNA-seq

Para desenvolvimento do trabalho, foram utilizadas oito bibliotecas de RNA-seq compreendendo quatro diferentes fases do ciclo de vida do parasito. As bibliotecas foram geradas por Mariana Boroni durante seu doutoramento na Universidade Federal de Minas Gerais (UFMG) e pela Núbia Gonçalves durante seu mestrado no Centro de Pesquisas René Rachou (CPqRR).

O RNA total (fases miracídio, esporocisto, esquistossômulo e vermes adultos) foi purificado com esferas magnéticas Dynalbeads C1 Streptavidin (Invitrogen) e um oligo dT biotinilado foi utilizado para a captura dos mRNAs por complementaridade com a calda poli-A. Em seguida, as esferas e a solução com os RNAs foram utilizadas como substrato para a produção de DNA complementar (cDNA) com o kit SuperScript III Reverse Transcriptase (Invitrogen). Os cDNAs foram amplificados com o kit GoTaq DNA Polymerase (Promega) a partir de um iniciador complementar ao SL de *S. mansoni* e um iniciador complementar à sequência estendida do Dynalbeads oligo-(dT) (Invitrogen), de forma a produzir bibliotecas enriquecidas em transcritos processados por SLTS.

As bibliotecas foram sequenciadas no equipamento Ion Torrent PGM^{TM} System da Life Technologies no laboratório multiusuário de genômica do ICB/UFMG pelo responsável técnico Rennan Garcias e pela doutoranda Mariana Boroni com colaboração da Núbia Gonçalves. As *reads* produzidas foram do tipo *single-end* com comprimentos variados. Para cada fase do parasito analisada foram produzidas duas bibliotecas constituindo-se réplicas biológicas.

4.2 Identificação e eliminação da sequência SL das *reads*

Para selecionar apenas transcritos contendo a sequência SL, as *reads* com a sequência líder foram identificadas e filtradas. Em seguida os nucleotídeos correspondentes ao SL foram removidos das *reads*. Estas etapas foram realizadas com o programa Cutadapt (MARTIN, 2010) na versão 1.4.1. Cutadapt foi implementado em linguagem Python e desenvolvido para remoção de adaptadores de sequências por meio

de um algoritmo de alinhamento global adaptado. Para atender ao propósito do trabalho, a sequência de um possível adaptador foi substituída pela sequência SL. Assim, as *reads* foram submetidas ao processo de identificação da sequência SL e remoção dele a partir dos parâmetros descritos na tabela 1.

Tabela 1: Parâmetros selecionados para utilização do Cutadapt.

Parâmetros – Cutadapt

Parâmetro	**Descrição**
-g	Especifica a sequência a ser removida (quando presente no início da *read*).
-a	Especifica a sequência a ser removida (quando presente no final da *read*).
-O 18	Tamanho mínimo que a sequência deve conter para ser considerada SL.
-e 0.07	Taxa de erro permitido.
-- info-file	Referente à opção de gerar um arquivo de relatório.
--untrimmed-output	Referente à opção de gerar um arquivo com as *reads* não processadas.
-m 50	Elimina *reads* muito curtas (com menos de 50 nucleotídeos).
-o	Referente ao arquivo de saída contendo as *reads* processadas.

Após o término do processo foi observado que o SL estava posicionado na maioria das vezes no início das *reads,* o que era esperado. No entanto, também foram encontradas sequências do complemento reverso do SL, posicionado mais frequentemente no final da *read.* Isso acontece devido ao tipo de metodologia utilizada de sequenciamento, em que o adaptador foi inserido em qualquer uma das extremidades das *reads*. Deste modo, o parâmetro "-g", foi usado para especificar a sequência SL no início da *read* e o parâmetro "-a" para o complemento reverso no final da *read.*

O arquivo de saída foi obtido contendo apenas as sequências processadas. Foram gerados mais dois arquivos de saída, um contendo as sequências que não foram

processadas e outro com um relatório do processo realizado com cada *read*. O arquivo de relatório mostra o identificador da *read*, sua sequência SL removida e sequências anteriores (quando presentes) e posteriores ao SL. A escolha do Cutadapt foi justamente devido à grande quantidade de informação que pôde ser obtida sobre o processo (através da opção de obtenção deste arquivo de relatório e de um relatório de tela) e devido aos diversos argumentos que permitem um processamento criterioso dependendo do tipo de dado processado.

4.2.1 Seleção das *reads* com SL no início de sua sequencia

Foram desenvolvidos três *scripts* em Perl para seleção apenas das *reads* com o SL no início da sequencia nucleotídica. O primeiro *script* (denominado intersection_report_fastq.pl) percorre os arquivos fastq contendo as *reads* processadas e seus respectivos arquivos de relatório para filtrar apenas as *reads* que serão analisadas nas etapas posteriores (*reads* maiores de 50 nucleotídeos), uma vez que no arquivo de relatório gerado pelo Cutadapt, estão mostrados os cortes feitos até mesmo em *reads* com menos de 50 nucleotídeos que não estão nos arquivos fastq de *reads* processadas. O segundo *script* (denominado SL_end.pl) percorre o arquivo de relatório filtrado pelo *script* anterior e realiza um segundo filtro que seleciona apenas as *reads* que sofreram corte de SL que estava na extremidade. Por fim, o terceiro *script* (denominado fastq_SL_end.pl) recupera as *reads* (correspondentes destes relatos) do arquivo fastq gerado pelo Cutadapt. O *script* intersection_report_fastq.pl não faria diferença no resultado final, no entanto, foi construído apenas para garantir um arquivo de relatório contendo somente as *reads* de interesse (com mais de 50 nucleotídeos) antes de dar prosseguimento ao processo de filtragem.

4.3 Alinhamento das *reads* no genoma de referência

A fim de se obter a localização genômica das *reads* e consequentemente identificar os genes cujos transcritos foram processados por SLTS, foi realizado um alinhamento das *reads* no genoma de referência de *S. mansoni* na versão 5 (obtido em ftp://ftp.sanger.ac.uk/pub/pathogens/Schistosoma/mansoni/) com o programa Bowtie2 (LANGMEAD; SALZBERG, 2012) na versão 2.1.0. Antes de escolher os parâmetros, o alinhador foi testado em modo padrão. A média de *reads* alinhadas em modo padrão foi entre 60% e 70%. Assim sendo, para aumentar a quantidade de *reads* alinhadas foi feita

uma diminuição e padronização do tamanho das *reads* em 50 nucleotídeos, como estratégia de diminuir a ocorrência de *gaps* e *mismatches*. O processo foi feito a partir de um *script* desenvolvido em Perl (denominado 50_fastq.pl) que elimina nucleotídeos do final da *read* até que esta permaneça apenas com 50 nucleotídeos. A estratégia não prejudica o prosseguimento das análises, uma vez que a informação necessária é a posição do primeiro nucleotídeo alinhado, para identificação da localização do sítio de *trans-splicing* (localizado logo antes deste nucleotídeo). Para reduzir as *reads* que tinham o complemento reverso do SL, seria necessário eliminar os nucleotídeos iniciais das *reads*, no entanto, foi decidido que estas *reads* não fariam parte das análises posteriores. Isso foi decidido para facilitar o prosseguimento das análises e porque provavelmente, estas *reads* não causariam modificações nos resultados finais, representando um subconjunto muito pequeno em relação ao total de *reads* trimadas.

Após estabelecimento dos novos parâmetros, as *reads* foram novamente alinhadas no genoma de referência de *S. mansoni* com o Bowtie2. Os parâmetros selecionados estão descritos na tabela 2.

Tabela 2: Parâmetros escolhidos para utilização do Bowtie2.

Parâmetros Bowtie2	
bowtie2-build	Cria os arquivos de index de genoma de referência.
-x	Antecede os arquivos de index.
-q	Antecede arquivo de entrada no formato fastq.
-k 1	Especifica a quantidade de mapeamentos por *read*.
-un	Referente à opção de gerar um arquivo com as *reads* não mapeadas.
-S	Referente ao arquivo de saída contendo as *reads* mapeadas.

A escolha do Bowtie2 deve-se ao fato de que esta ferramenta utiliza um algoritmo que o torna rápido e eficiente, principalmente do ponto de vista de memória RAM requerida para a análise, sendo essa ferramenta capaz de nos fornecer de forma eficaz a informação de interesse que é a posição do primeiro nucleotídeo da *read* para identificação do gene a qual pertence e a posição do sítio de *trans-splicing*.

Após o alinhamento, foi possível identificar precisamente em qual gene e em qual região específica do gene (5'UTR, CDS ou íntron) se encontravam os sítios de SLTS. Esta etapa permite uma análise qualitativa para determinação dos sítios preferenciais e dos sítios menos utilizados pelo mecanismo de SLTS no que diz respeito a região onde se localizam. Para isso, foi desenvolvido um *script* em Perl (denominado indentification_local_sites.pl). Assim, também foi possível ajustar as coordenadas dos genes de acordo com a sequência de nucleotídeos que permanece após o processamento por SLTS. Estas coordenadas serão necessárias em passos posteriores para recuperação da sequência processada.

O indentification_local_sites.pl utiliza como entrada o arquivo de anotação do genoma de *S. mansoni* utilizado durante a etapa de alinhamento (obtido em ftp://ftp.sanger.ac.uk/pub/pathogens/Schistosoma/mansoni/). Este arquivo foi filtrado com o programa nativo do Unix awk, de forma que fossem mantidas somente informações de interesse, ou seja, sobre os genes codificadores de mRNAs (cromossomo, coordenada inicial e final de cada CDS, fita e identificador), mesmo existindo outros tipos de RNA não codificadores de proteínas que sofrem SLTS. Outra entrada manipulada pelo programa foi o arquivo no formato sam, gerado a partir do alinhamento. Este arquivo também foi modificado com os programas nativos do Unix awk e sed, para que passassem a conter somente as informações da coordenada de mapeamento, a fita e o cromossomo de cada *read* mapeada. Este segundo arquivo de saída modificado continha muitas coordenadas repetidas, referentes aos genes com múltiplas *reads* alinhadas na mesma posição. Sendo assim, as repetições foram eliminadas e a informação sobre a quantidade de registros de cada coordenada de mapeamento foi mantida em uma coluna adicional. Por fim, as coordenadas foram ordenadas com o programa sort para facilitar a manipulação do arquivo pelo indentification_local_sites.pl. O programa manipula os arquivos percorrendo cada coordenada de mapeamento e identificando onde elas se enquadram de acordo com as coordenadas das CDSs de cada gene. Assim, foi possível identificar as seguintes situações:

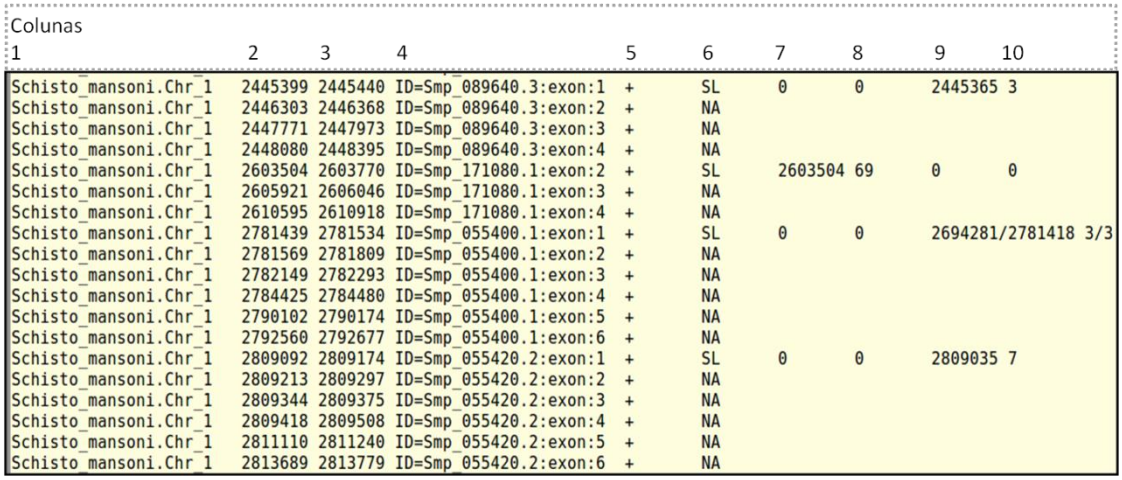
1- Coordenada posicionada pouco antes da coordenada inicial do primeiro éxon de um gene (interpretação: sítio de SLTS na região 5' UTR);
2- Coordenadas entre a coordenada inicial e final de um determinado éxon de um gene (interpretação: sítio de SLTS em CDS);

3- Coordenadas entre éxons de um gene (interpretação: sítio de SLTS em região intrônica).
4- Coordenadas entre éxons de um gene, logo no início do éxon seguinte (interpretação: sítio de SLTS idêntico ao sítio de *cis-splicing*).

O programa não é capaz de fazer registros muito específicos, de forma que as interpretações só podem ser feitas em análises com outros programas desenvolvidos posteriormente. O arquivo de saída do identification_local_sites.pl tem formato tabular, cujas colunas descrevem as seguintes informações:

- Coluna 1: cromossomo;
- Coluna 2: coordenada inicial;
- Coluna 3: coordenada final;
- Coluna 4: identificador do gene e número do éxon;
- Coluna 5: fita;
- Coluna 6: "SL" (região de entrada do SL) ou "NA" (região onde não ocorreu alteração por entrada de SL);
- Coluna 7: coordenada de mapeamento (caso situada dentro da CDS);
- Coluna 8: coordenada de mapeamento (caso situada no íntron anterior à um determinado éxon);
- Coluna 9: quantidade de *reads* mapeadas na CDS (caso a coluna 6 descrever "SL" e a coluna 7 for diferente de 0);
- Coluna 10: quantidade de *reads* mapeadas na região intrônica (caso a coluna 6 descrever "SL" e a coluna 8 for diferente de 0).

A figura 8 mostra uma pequena parte do arquivo de saída do identification_local_sites.pl para facilitar o entendimento de sua estrutura.



Colunas

1	2	3	4	5	6	7	8	9	10
Schisto_mansoni.Chr_1	2445399	2445440	ID=Smp_089640.3:exon:1	+	SL	0	0	2445365	3
Schisto_mansoni.Chr_1	2446303	2446368	ID=Smp_089640.3:exon:2	+	NA				
Schisto_mansoni.Chr_1	2447771	2447973	ID=Smp_089640.3:exon:3	+	NA				
Schisto_mansoni.Chr_1	2448080	2448395	ID=Smp_089640.3:exon:4	+	NA				
Schisto_mansoni.Chr_1	2603504	2603770	ID=Smp_171080.1:exon:2	+	SL	2603504	69	0	0
Schisto_mansoni.Chr_1	2605921	2606046	ID=Smp_171080.1:exon:3	+	NA				
Schisto_mansoni.Chr_1	2610595	2610918	ID=Smp_171080.1:exon:4	+	NA				
Schisto_mansoni.Chr_1	2781439	2781534	ID=Smp_055400.1:exon:1	+	SL	0	0	2694281/2781418	3/3
Schisto_mansoni.Chr_1	2781569	2781809	ID=Smp_055400.1:exon:2	+	NA				
Schisto_mansoni.Chr_1	2782149	2782293	ID=Smp_055400.1:exon:3	+	NA				
Schisto_mansoni.Chr_1	2784425	2784480	ID=Smp_055400.1:exon:4	+	NA				
Schisto_mansoni.Chr_1	2790102	2790174	ID=Smp_055400.1:exon:5	+	NA				
Schisto_mansoni.Chr_1	2792560	2792677	ID=Smp_055400.1:exon:6	+	NA				
Schisto_mansoni.Chr_1	2809092	2809174	ID=Smp_055420.2:exon:1	+	SL	0	0	2809035	7
Schisto_mansoni.Chr_1	2809213	2809297	ID=Smp_055420.2:exon:2	+	NA				
Schisto_mansoni.Chr_1	2809344	2809375	ID=Smp_055420.2:exon:3	+	NA				
Schisto_mansoni.Chr_1	2809418	2809508	ID=Smp_055420.2:exon:4	+	NA				
Schisto_mansoni.Chr_1	2811110	2811240	ID=Smp_055420.2:exon:5	+	NA				
Schisto_mansoni.Chr_1	2813689	2813779	ID=Smp_055420.2:exon:6	+	NA				

Figura 8: Arquivo de saída do indentification_local_sites.pl.

4.4. Preparo dos arquivos para obtenção de dados

Para que o arquivo de saída do indentification_local_sites.pl fique próprio para obtenção dos dados referentes aos sítios de SLTS foi necessário filtrar somente as linhas dos respectivos genes processados, uma vez que o arquivo possui informações do conjunto total de genes do arquivo de anotação. Não foi possível obter este filtro no código do indentification_local_sites.pl, porque a ideia foi recuperar não só as linhas com registro de entrada de SL, como também as demais linhas com as coordenadas dos outros éxons que fazem parte do conteúdo dos transcritos processados (portanto, o *script* também registrou linhas correspondentes a éxons de transcritos não processados).

Para esta filtragem foi necessário desenvolver um *script* em Python (denominado processed_genes.py) que identifica as linhas com registro de entrada de SL e verifica se a linha de baixo refere-se ao mesmo transcrito. Se isto for verdadeiro, o *script* armazena esta linha em um arquivo de saída, assim como todas as outras linhas referentes ao mesmo transcrito.

Os dados utilizados para as análises posteriores e geração dos resultados foram apenas os respectivos de genes presentes no par de bibliotecas (réplicas biológicas) do mesmo estágio do ciclo de vida do parasito, assim, genes fora da interseção entre as bibliotecas de mesma fase foram desconsiderados. Mais especificamente, em cada fase do parasito, só seria analisada a intersecção de registros tanto do gene quanto de seus sítios, ou seja, cada região (cada linha do arquivo de saída do processed_genes.py) foi verificada por meio de outro programa desenvolvido em Perl (denominado

intersection_genes_sites.py) e só foi registrada em um arquivo de saída, se estivesse presente nos arquivos referentes às duas bibliotecas de mesma fase de desenvolvimento do parasito. Além disso, só foram considerados os sítios registrados por, pelo menos, três *reads* mapeadas. Estas foram estratégias de seleção dos sítios com maior autenticidade.

4.5 Obtenção de dados dos transcritos processados por SLTS

Após o preparo adequado dos arquivos com informações da localização dos sítios, os primeiros dados foram gerados. A primeira informação obtida foi a contagem do número de transcritos correspondentes a genes processados em cada fase do parasito. Para isso, foi desenvolvido um *script* simples em Perl (denominado SLTS_genes.pl) que lista os identificadores gênicos sem repetições. Em seguida, os identificadores foram submetidos a uma contagem com o programa grep.

Foi criado um *script* (denominado total_processed_genes.pl) para obtenção da quantidade de genes processados no conjunto total de *reads* analisadas, sem redundâncias. Para isso, os identificadores dos genes dos arquivos de saída do processed_genes.py foram armazenados em um só arquivo, que serviu de arquivo de entrada para o total_processed_genes.pl.

Em seguida, foi feita uma verificação da quantidade de genes cujos transcritos sofrem entrada da sequência SL em regiões distintas. Nesta etapa, foram desenvolvidos vários *scripts*, que combinados com os programas nativos do Unix, permitiram as seguintes contagens:

1- Genes cujos transcritos apresentam um ou mais sítios somente na região 5'UTR (quantidade obtida pelo programa desenvolvido em Python "genes_5UTR.py" e com o awk e o grep);
2- Genes cujos transcritos apresentam um ou mais sítios apenas em CDSs (quantidadede obtida pelo programa desenvolvido em Perl "genes_CDS.pl" e com o grep);
3- Genes cujos transcritos apresentam um ou mais sítios apenas em íntrons (quantidade obtida pelo programa desenvolvido em Perl "genes_introns.pl");

4- Genes cujos transcritos apresentam entradas em mais de uma região distinta, como, por exemplo, em CDS e íntron (quantidade obtida por programas nativos do Unix).

4.6 Verificação de mudança de fase de leitura

Para verificar se a entrada do SL provocou uma mudança na fase de leitura dos transcritos processados, primeiramente foram selecionadas apenas isoformas de transcritos processados por SLTS a partir de entradas de SL dentro de CDSs e em regiões intrônicas. Isso porque é pouco provável que genes com sítios de *trans-splicing* na região 5' UTR sofram este tipo de dano. A análise foi feita através de curadoria manual por meio do software de visualização IGV 2.1 (de Integrative Genomics Viewer) (THORVALDSDÓTTIR; ROBINSON; MESIROV, 2013) (Broad Institute, obtido em http://www.broadinstitute.org/software/igv/download/). Foram utilizados como entrada do IGV, o arquivo de anotação, o genoma de referência e os arquivos de alinhamento do Bowtie2. O IGV permite a visualização dos genes (íntrons, éxons e regiões UTR) com as respectivas *reads* alinhadas, o sítio de *splicing* utilizado por cada gene processado, as possíveis janelas de leitura e as sequências em nucleotídeos.

4.7 Extração das sequências gênicas processadas por SLTS

Os arquivos de saída do processed_genes.py foram modificados (com a adição de colunas do arquivo de anotação original, correspondentes a cada linha com informação das coordenadas de CDSs dos genes e edição dos identificadores para diferenciar cada isoforma de *trans-splicing*) de forma que se tornassem adequados para extração da sequência de nucleotídeos dos genes já processados por SLTS (sem incluir o SL no início da sequência) com o programa gffread do pacote Cufflinks.

4.8 Verificação dos domínios das proteínas derivadas dos genes processados

Como forma de comparar os domínios proteicos das proteínas derivadas dos transcritos processados por SLTS com a sequência proteica original, primeiramente foi obtida a sequência de aminoácidos de cada proteína codificada pelos transcritos através do EMBOSS transeq (RICE; LONGDEN; BLEASBY, 2000) (European Bioinformatics

Institute, obtido em http://www.ebi.ac.uk/Tools/st/emboss_transeq/). Foram selecionadas aleatoriamente, apenas 27 proteínas processadas por SLTS para realização desta análise, estas, foram então comparadas com as sequências das proteínas originais (arquivo Schistosoma_mansoni.ASM2379v2.27.cds.all.fa obtido em http://metazoa.ensembl.org/info/website/ftp/index.html).

As sequências foram então submetidas ao InterProScan (JONES *et al.*, 2014) na versão 5 (European Bioinformatics Institute, obtido em ftp://ftp.ebi.ac.uk/pub/software/unix/iprscan/5/), uma ferramenta que combina metodologias de reconhecimento de assinaturas proteicas a partir da extração de informações de diversos bancos de dados do InterPro (HUNTER; JONES; MITCHELL, 2012). O InterPro armazena dados de proteínas como função, família e domínios proteicos. Desta forma, os algoritmos do InterPro podem ser utilizados de forma integrada pelo software InterProScan para caracterização funcional de sequências (nucleotídeos ou aminoácidos). O InterProScan está integrado a outras fontes de informação como os bancos Pfam e PDB.

Todo o *pipeline* utilizado nesse trabalho está esquematizado na figura 9.

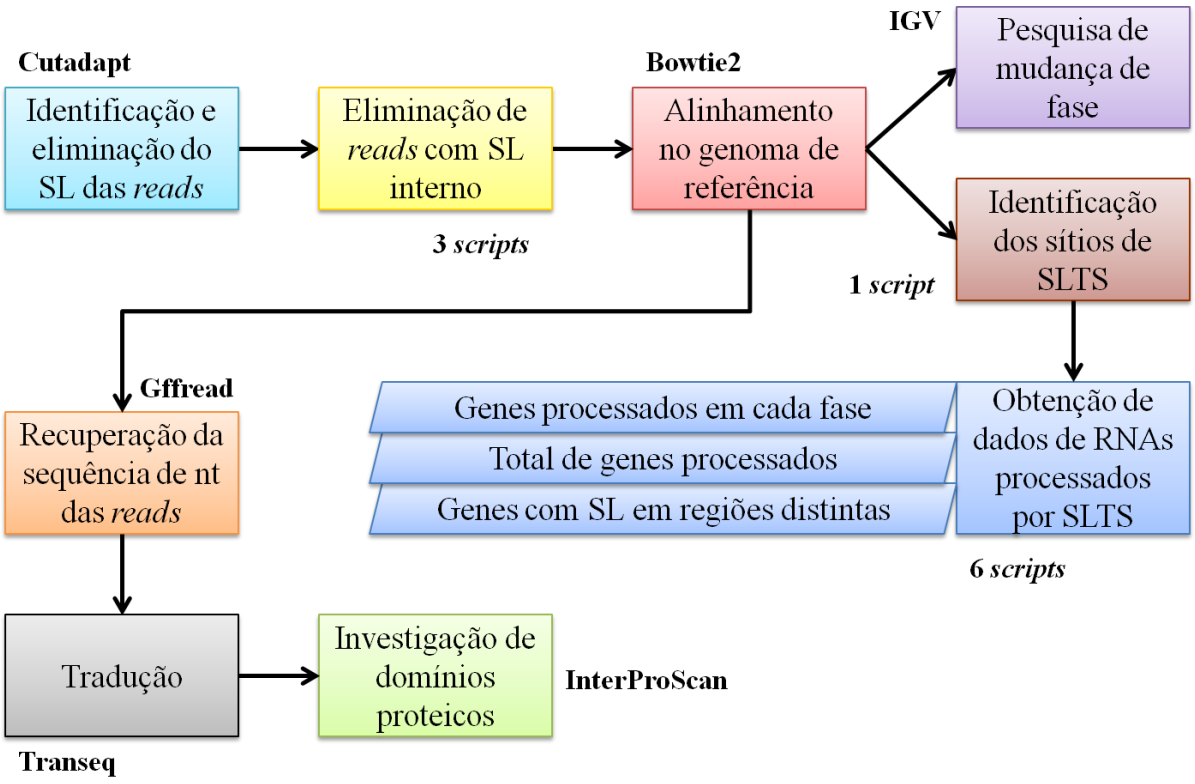


Figura 9: *Pipeline* de trabalho.

5 RESULTADOS E DISCUSSÃO

5.1 Bibliotecas utilizadas, processamento das *reads* e seu alinhamento no genoma de referência

As bibliotecas utilizadas para as análises foram sequenciadas utilizando-se cDNAs enriquecidos em transcritos que sofrem SLTS. A quantidade das *reads* geradas em cada biblioteca encontra-se na tabela 3. A média da qualidade por base das bibliotecas foi em torno de 20 na escala Phred, que de modo geral, representa o *score* de qualidade mínimo considerado aceitável e indica a probabilidade de ocorrência de um erro a cada 100 pares de bases. Todavia, os valores de escala Phred de bibliotecas de Ion Torrent PGM™ System muitas vezes são subestimados, pois ao alinhar as *reads* observa-se com frequência uma maior acurácia das bases (LOMAN *et al*., 2012; ROTHBERG *et al*., 2011).

A tabela 3 mostra o percentual de *reads* curtas (com menos de 50 nucleotídeos) eliminadas do conjunto total de dados, como estratégia de minimizar erros de alinhamento. A tabela também mostra o número de *reads* que tiveram a sequência SL identificada e posteriormente retirada e a porcentagem que estas *reads* representam em relação ao total de *reads* das bibliotecas.

Ao visualizar os arquivos de relatório gerados com o cutadapt foi possível observar que algumas *reads* continham a sequência SL em seu interior e não nas extremidades. Estes transcritos foram estudados em um trabalho prévio (BORONI, 2014) e provavelmente são transcritos processados por SLTS que foram retro-transcritos e reinseridos no genoma. Essas *reads* foram eliminadas e a porcentagem de *reads* processadas mantidas após esta eliminação (com os *scripts* intersection_report_fastq.pl, SL_end.pl e fastq_SL_end.pl) foi descrita na tabela 3.

Para alinharmos as *reads* no genoma de referência foi utilizado o programa Bowtie2, o que resultou em torno de 60% a 70% de *reads* alinhadas. Assim sendo, para aumentar a quantidade de *reads* alinhadas foi feita uma diminuição e padronização do tamanho das *reads* em 50 nucleotídeos (com o 50_fastq.pl), como estratégia de diminuir a ocorrência de *gaps* e *mismatches*. A estratégia permitiu o alinhamento de 80% a 90% das *reads*.

Tabela 3: Dados. Fase de desenvolvimento do parasito de cada biblioteca, denominação das réplicas, total de *reads* e de bases de cada biblioteca, percentual de *reads* curtas eliminadas, número e percentual de *reads* com a sequência SL, porcentagem de *reads* contendo a sequência SL que se mantiveram nas análises (sem SL interno), número e percentual de *reads* alinhadas no genoma de referência.

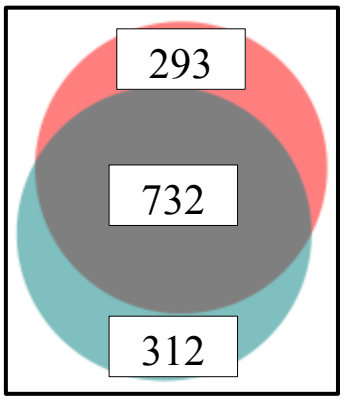
Estatísticas do processamento das *reads* e seu alinhamento no genoma de referência

Fase de desenvolvimento	**Réplica**	**Total de *reads* da biblioteca**	**Total de bases da biblioteca**	**Percentual de *reads* curtas eliminadas**	**Número de *reads* com SL**	**Percentual de *reads* com SL**	**Percentual de *reads* sem SL interno**	**Número de *reads* alinhadas**	**Percentual de *reads* alinhadas**
Miracídio	M.R1	3.402.016	625.5 Mbp	3.2%	197.027	5,8%	93,8%	166.200	89,9%
Miracídio	M.R2	2.792.890	463.4 Mbp	3,9%	181.704	6,5%	93,8%	149.057	87,4%
Esporocisto	Esp.R1	3.479.656	574.3 Mbp	8,8%	580.184	16,7%	93,6%	501.653	92,3%
Esporocisto	Esp.R2	3.127.173	490.1 Mbp	11,8%	600.584	19,2%	93,6%	530.519	94,2%
Esquistossômulo	Esq.R1	2.562.467	473.5 Mbp	3,6%	170.468	6,7%	91,3%	140.119	89,9%
Esquistossômulo	Esq.R2	3.372.875	710.0 Mbp	2.9%	172.487	5,1%	92,9%	143.269	89,4%
Verme Adulto	V.A.R1	2.586.306	448.0 Mbp	6,2%	304.787	11,8%	89,8%	254.011	92,7%
Verme Adulto	V.A.R2	2.802.032	534.0 Mbp	5,4%	189.668	5,8%	88,8%	153.724	91,1%

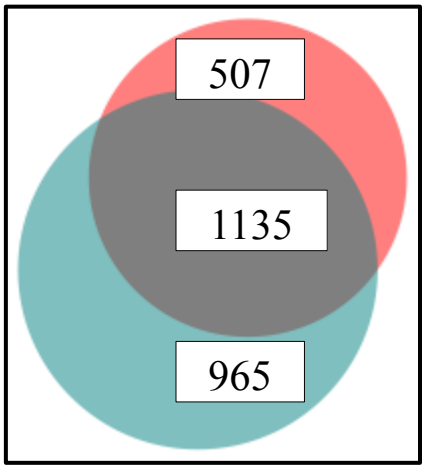
Inicialmente acreditava-se que apenas um pequeno subgrupo de mRNAs de *S. mansoni* adquiria a sequência líder por SLTS (DAVIS *et al*., 1995), mas com o avanço nos estudos de *transcriptoma* pelas técnicas de RNA-seq, foi possível estimar que ~77% dos genes codificadores de proteínas podem sofrer SLTS em alguma fase de seu desenvolvimento. No entanto, ainda não se sabe por que alguns genes monocistrônicos são processados por este mecanismo e outros não (BORONI, 2014). Em outros trabalhos, também foi visto que o mecanismo de SLTS não parece estar associado com transcritos codificadores de determinadas classes de proteínas, nem mesmo se resume a certos tipos de células, tecidos, ou são sexo específico (DAVIS *et al*., 1995; MOURÃO *et al*., 2013).

A soma de transcritos processados nas bibliotecas de mesma fase foi de 1.337 para miracidio, 2.607 para esporocisto, 1.851 para esquistossômulo e 1.879 para verme adulto. Para a obtenção do número de genes processados por SLTS em cada biblioteca, o arquivo de saída do identification_local_sites.pl (que fornece a localização dos sítios de SLTS) foi então modificado (com os *scripts* processed_genes.py, intersection_genes_sites.py e SLTS_genes.pl) e, em seguida, foi obtido o número de genes processados nas bibliotecas de mesmo estágio de vida do parasito (interseção) e a quantidade de genes particulares (fora da interseção) de cada biblioteca de mesma fase (figura 10).

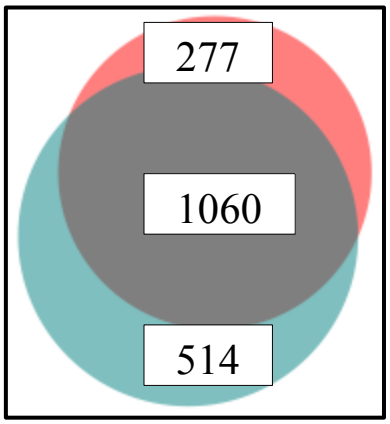
Miracídio



Esporocisto



Esquistossômulo



Verme adulto

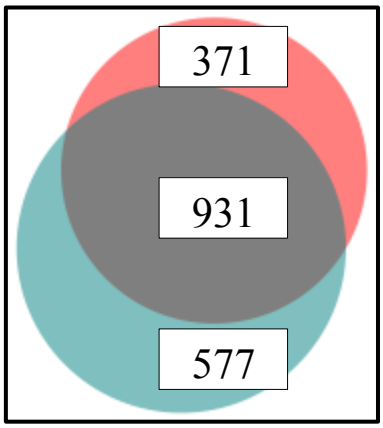


Figura 10: Genes processados por SLTS observados nas bibliotecas dos estágios do ciclo de vida do parasito. Diagramas de Venn construídos no BioVenn (HULSEN; VLIEG; ALKEMA, 2008), apresentando a distribuição dos genes processados por SLTS nas bibliotecas de mesmo estágio de *S. mansoni*. A cor rosa representa a réplica 1 de cada fase e a cor azul a réplica 2.

Para o estudo do SLTS no parasito considerou-se apenas os genes que estavam representados em ambas as bibliotecas de mesmo estágio do ciclo de vida do *S. mansoni* (identificados por pelo menos 3 *reads* mapeadas) com o SL inserido exatamente na mesma posição, seja em suas regiões não traduzíveis (5' UTR), intrônicas ou exônicas.

Os genes que apresentaram estas condições foram selecionados para as análises posteriores, sendo 469 de miracidio, 491 de esporocisto, 725 de esquistossômulo e 524 de verme adulto. Nota-se que o número de genes processados por SLTS encontrados no esquistossômulo foi maior comparado aos valores encontrados nas demais bibliotecas. Isso provavelmente ocorreu somente porque esta fase apresentou maior intersecção de transcritos processados com sítios de SLTS no mesmo local comparando as duas réplicas.

A quantidade de genes em cada cromossomo cujos transcritos receberam a sequência SL foi proporcional ao tamanho dos cromossomos, sendo o cromossomo 1 e o cromossomo sexual (W), os maiores cromossomos, os que contiveram uma quantidade mais alta de transcritos processados, respectivamente. No cromossomo 7 (o menor deles), foi encontrado um número menor de genes processados em relação aos demais cromossomos, mostrando não haver enriquecimento do mecanismo em genes presentes em um determinado cromossomo.

Nesse estudo encontramos que a soma de genes processados foi de 1.249, totalizando 1.733 isoformas geradas por SLTS (dados obtidos com o total_processed_genes.pl). Os 1.249 genes processados identificados (sem incluir redundâncias entre as bibliotecas) representam cerca de 10% dos genes anotados no arquivo GFF do genoma. Em um trabalho prévio no qual foram analisados transcritos das mesmas bibliotecas do presente estudo, foi visto que transcritos processados que exibem maior quantidade de isoformas derivadas de *cis-splicing* alternativo possuem menor incidência de SLTS, enquanto que genes com alta quantidade de éxons, mas com baixo número de isoformas de *cis-splicing* possuem uma chance maior de sofrerem SLTS, sugerindo uma maior eficiência do spliceossomo na falta de componentes de *cis-splicing* (BORONI, 2014).

5.2 Identificação precisa da localização dos sítios de SLTS

Após a obtenção do número de genes cujos transcritos possuíam sítios de SLTS nas diferentes regiões utilizando os *scripts* genes_5UTR.py (genes com sítios na região 5’ UTR), genes_introns.pl (genes com sítios em íntrons) e genes_CDS.pl (genes com sítios em CDSs), os transcritos com entrada de SL na região 5’ UTR foram divididos

em duas categorias: uma composta pelos transcritos que sofreram perda de toda região 5’ UTR e outra com os que tiveram apenas parte da 5’ UTR perdida (figura 11).

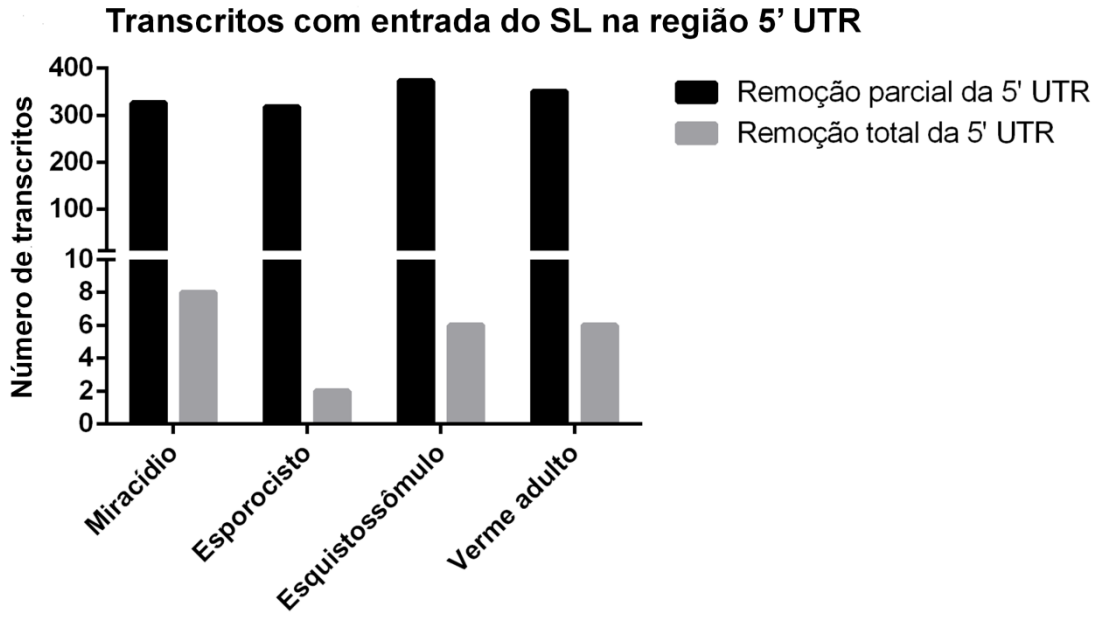


Figura 11: Transcritos com uma ou mais entradas de SL somente na região 5’ UTR.

O SLTS permite um aumento da diversidade de mRNAs, elevando a capacidade de codificação do genoma (NILSEN; GRAVELEY, 2010; MOURÃO *et al*., 2013), mas também pode estar inibindo o processo de tradução e contribuindo para a instabilidade do transcrito (PREUßER; JAÉ; BINDEREIF, 2012).

O SLTS convencional (com entrada do SL na região 5’ UTR) é claramente a forma preferencial diante das diferentes possibilidades de processamento alternativo. Do total de transcritos processados por SLTS, a porcentagem que sofre uma ou mais entradas do SL somente na região 5’ UTR é de 69% em miracídios, 65% em esporocistos e esquistossômulos e 67% em vermes adultos. Portanto, a produção de possíveis proteínas encurtadas ou peptídeos sem funcionalidade são eventos mais raros, que não fazem parte do papel principal do SLTS no parasito.

O SLTS alternativo pode ser um meio de expansão do repertório proteico de *S. mansoni* (MOURÃO *et al*., 2013). Quanto maior a distância entre o sítio de SLTS e o códon de iniciação da tradução, menor é a possibilidade da sequência modificada após o processamento conter outro AUG. Interessantemente, a sequência SL de *S. mansoni* (5’ AACCGUCACGGUUUUACUCUUGUGAUUUGUUGCAUG 3’) (RAJKOVIC *et al.*,

1990) termina com um códon AUG, porém, o tripleto não necessariamente é utilizado como códon de iniciação das ORFs do respectivo conjunto de transcritos processados por SLTS (DAVIS *et al*., 1995). O AUG no final do SL é uma característica dos platelmintos, não estando presente na sequência SL de outros organismos (CHENG *et al*., 2006).

Em todas as fases do parasito, mais de 97% dos transcritos com uma ou mais entradas de SL na 5’ UTR sofrem *splicing* no interior da região não traduzível, apresentando redução de parte destas (observação feita a partir da contagem dos genes codificadores desses transcritos, utilizando o arquivo de saída do processed_genes.py).

A porção que permanece e/ou a sequência SL podem estar sendo utilizadas como UTRs alternativas. O tamanho médio da 5' UTR na maioria das classes taxonômicas varia entre 100 e 200 bases (MIGNONE *et al*., 2002). De acordo com as ocorrências identificadas, o comprimento perdido é bastante variável, podendo ser muito pequeno ou incluindo quase toda a região não traduzível. No entanto, em mamíferos, um único nucleotídeo presente na 5 'UTR pode ser suficiente para a iniciação da tradução (HUGHES; ANDREWS, 1997).

Menos de 10 transcritos em todas as fases sofreu perda da região 5’ UTR por inteiro. Estes são casos em que a sequência SL pode estar sendo utilizada como região 5’ não traduzível, onde se localizam os sítios de ligação do ribossomo para a síntese proteica. As UTRs participam dos mecanismos de regulação pós-transcricional e influenciam na eficiência da tradução (principalmente de proteínas envolvidas no desenvolvimento, fatores de transcrição e de crescimento), podendo modular o transporte de mRNAs para fora do núcleo, localização subcelular e a estabilidade do transcrito (MIGNONE *et al*., 2002).

Assim como os transcritos com entrada de SL na 5 'UTR, os transcritos que sofrem SLTS alternativo com sítios em íntrons também foram divididos em categorias, sendo uma com os que sofreram entrada de SL no mesmo sítio aceptor de *cis-splicing* e outra com os que utilizam sítios de *trans-splicing* presentes dentro do íntron (figura 12).

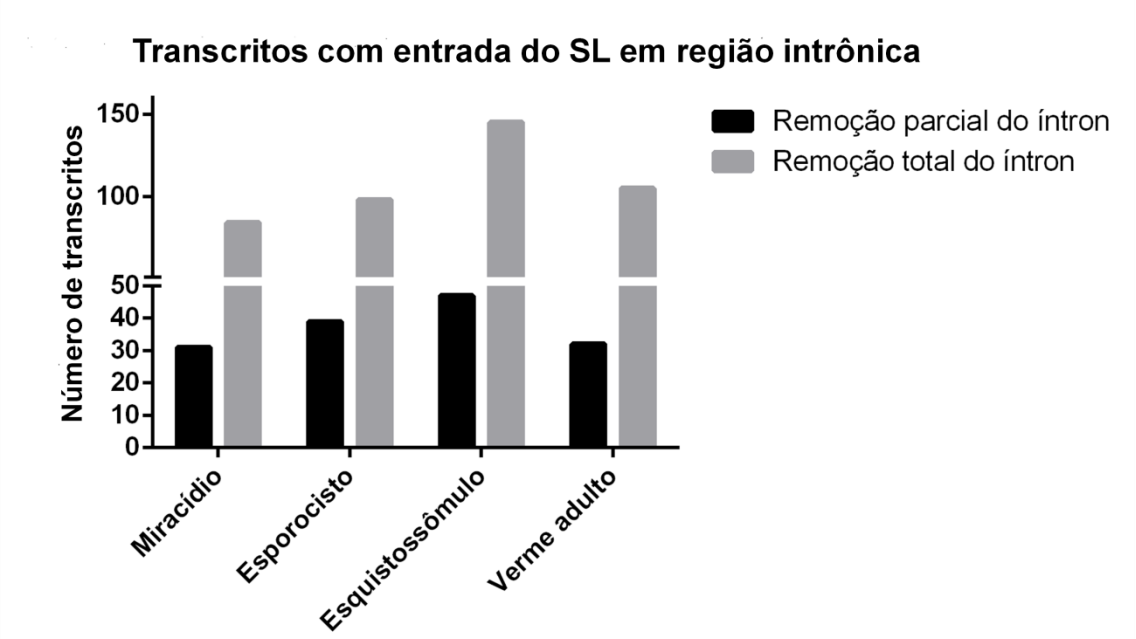


Figura 12: Transcritos com uma ou mais entradas de SL somente em região intrônica.

O dinucleotídeo AG é o sítio aceptor de *splicing* canônico, geralmente associado a um trato de polipirimidina. É utilizado no *cis-splicing* e em outras ocasiões como sítio de *trans-splicing* com alto grau de conservação (REED, 1989; NILSSON *et al.*, 2010). Uma pequena parte dos genes apresentam sítios no interior de regiões intrônicas, e mesmo estes sítios alternativos utilizados para SLTS presentes no meio do íntron apresentaram conservação da sequência AG, assim como todos os sítios nas regiões 5' UTR e em CDSs. Os íntrons sujeitos ao processamento geralmente apresentam sítios aceptores canônicos mais fortes e um trato de polipirimidina maior, o que está relacionado com a decisão da ocorrência de *splicing* em *cis* ou em *trans* (BORONI, 2014).

Um subconjunto muito pequeno dos transcritos processados possui ocorrências de sítios de SLTS no interior de regiões codificadoras (figura 13). Quanto mais próximo o sítio de entrada do SL for da região 3' UTR, menor é a possibilidade de ser produzida uma sequência proteica funcional. Em geral, transcritos que sofrem SLTS em CDS tem redução parcial do éxon alvo da adição do SL.

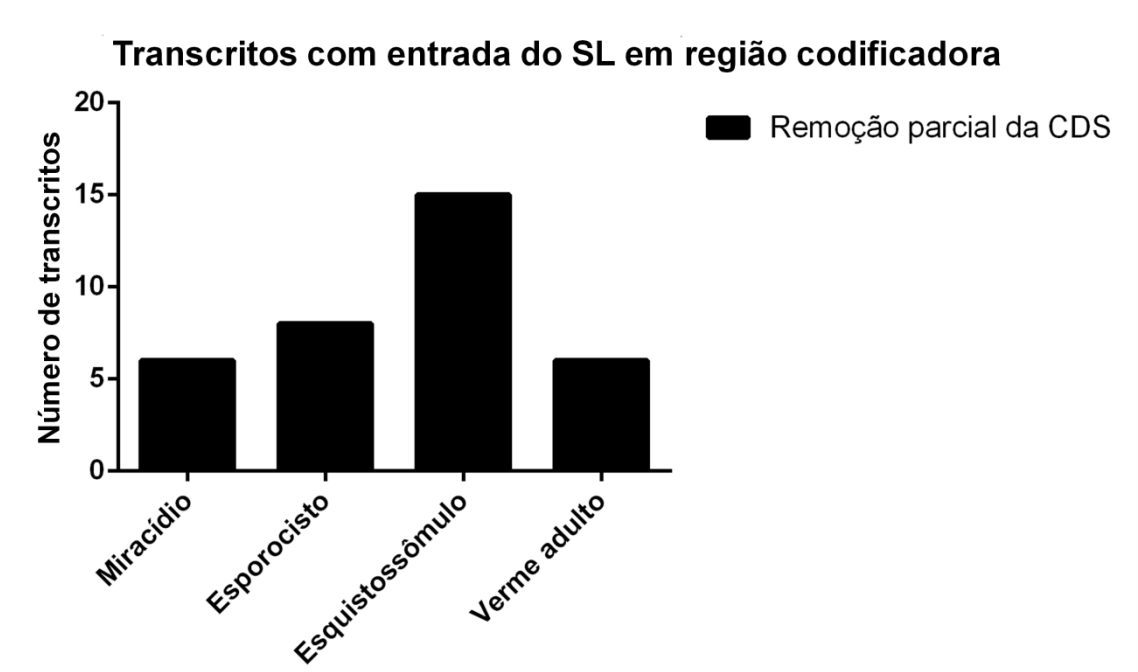


Figura 13: Transcritos com uma ou mais entradas de SL somente em CDS.

Por fim, foi obtida a quantidade de transcritos com entradas do SL em mais de uma região (figura 14).

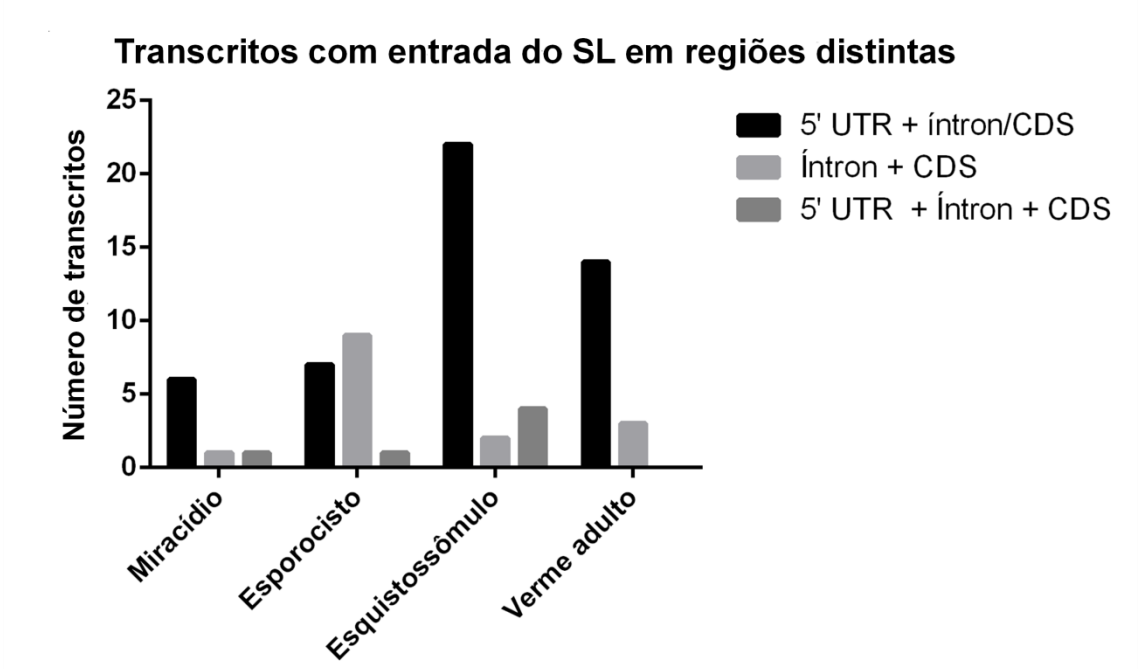


Figura 14: Transcritos com uma ou mais entradas de SL em regiões distintas.

A frequência de transcritos com sítios alternativos em mais de uma região corresponde a menos de 5% em relação ao total de transcritos processados por SLTS em todas as fases. As formas alternativas de SLTS são capazes de causar alteração na fase de leitura do transcrito, o que pode provocar a produção de peptídeos totalmente modificados em relação à sequência proteica original.

Pensando nisso, foram escolhidos 96 genes para estudo de caso, a fim de se obter informações referentes ao impacto que as entradas da sequência SL em diferentes regiões podem acarretar no conjunto de proteínas codificadas pelos transcritos dos genes processados. Nesta análise foram registradas as seguintes informações de cada gene analisado:

1- Local de entrada da sequência SL: em qual íntron ou éxon, em que posição (se foi no final de um íntron, ou seja, na mesma região do sítio de *cis-splicing* ou dentro de um éxon);
2- Total de éxons do gene;
3- Qual foi o sítio aceptor de SLTS;
4- Qual a fase de leitura verdadeira (utilizada para a síntese da proteína original);
5- Se a fase de leitura verdadeira referente ao item 4 (na região do éxon de entrada do SL ou éxon posterior ao sítio de SLTS nos casos de entrada em íntron) apresenta um AUG alternativo para síntese da proteína processada por SLTS caso não seja utilizado o AUG do SL;
6- Qual é a fase de leitura para a síntese da proteína processada por SLTS caso seja utilizado o AUG do SL como códon de início. Isso foi feito a partir da verificação da fase em que se encontra o primeiro códon formado a partir da sequência do gene que permanece íntegra após a entrada do SL (casos em que a sequência SL é usada como 5’ UTR);
7- Se a fase de leitura obtida no item 6 possui um AUG alternativo para síntese da proteína processada por SLTS caso não seja utilizado o AUG do SL (casos em que a sequência SL e a sequência entre o SL e o próximo AUG são utilizadas como 5’ UTR);
8- Se a(s) outra(s) fase de leitura diferente(s) da(s) fase(s) obtida(s) nos itens 4 e 6 possuem AUG alternativo para início da síntese proteica ou se apresentam *stop codons* próximos ou distantes do sítio de SLTS.

OBS: Estes resultados estão apresentados na tabela 4, onde a descrição de códon para metionina (M) ou *stop codon* próximo ou distante significa até 50 nucleotídeos e mais de 50 nucleotídeos do sítio aceptor de SLTS, respectivamente. Essas descrições de M/*stop codon* próximo ou distante e a descrição de "ausência de M", referem-se à região do éxon de entrada do SL ou éxon posterior ao sítio de SLTS nos casos de entrada em íntron. Caso existam outros éxons, os mesmos não foram considerados devido ao fato de estarem mais distantes do sítio de SLTS.

Foram inclusas todas as ocorrências de sítios dentro de regiões codificadoras. Não foram incluídos transcritos com entrada na região 5' UTR, pois não seria possível obter muitas informações sobre eles, uma vez que, a sequência da proteína codificada teoricamente não é alterada e por não ser possível prever se estes transcritos sofrem mudança de fase de leitura e síntese de proteínas modificadas. Transcritos com sítios de SLTS em regiões intrônicas foram escolhidos aleatoriamente, com a preocupação de selecionar parte que teve o íntron removido por inteiro e parte que teve o SL inserido no centro da sequência intrônica. A tabela 4 apresenta os resultados referentes a todos os 96 genes analisados (inclui genes presentes nos arquivos de todas as fases do parasito).

Tabela 4: Estudo de caso dos 96 genes selecionados.

Gene	Região de entrada do SL	Total de éxons do gene	Sítio aceptor de SLTS	Fase de leitura (FL)	Presença ou ausência de M na FL original	FL com uso do AUG do SL	Presença ou ausência de M ou *stop códon* na FL com uso do AUG do SL	Presença ou ausência de M ou *stop códon* na(s) outra(s) FL(s)
Smp_000630.1	**Final do íntron 1**	**3**	**AG**	**1**	**M (distante)**	**2**	***Stop c.* (próximo)**	***Stop c.* (próximo)**
Smp_003970.1	**Final do íntron 3**	**7**	**AG**	**2**	**M (distante)**	**1**	***Stop c.* (próximo)**	***Stop c.* (próximo)**
Smp_014020.1	**Final do íntron 5**	**7**	**AG**	**2**	**M (distante)**	**2**	**FL original**	***Stop c.* (próximo)** ***Stop c.* (distante)**
Smp_167000.1	**Final do íntron 1**	**2**	**AG**	**1**	**M (próxima)**	**3**	***Stop c.* (próximo)**	***Stop c.* (próximo)**
Smp_034190.1	**Final do íntron 5**	**6**	**AG**	**1**	**M (distante)**	**3**	***Stop c.* (próximo)**	***Stop c.* (próximo)**
Smp_007960.1	**Final do íntron 4**	**5**	**AG**	**1**	**M (distante)**	**3**	***Stop c.* (próximo)**	**M (próxima)**
Smp_003760.1	**Final do íntron 3**	**4**	**AG**	**3**	**M (próxima)**	**3**	**FL original**	***Stop c.* (próximo)** ***Stop c.* (próximo)**
Smp_008070.1	**Final do íntron 1**	**3**	**AG**	**1**	**Ausência de M**	**1**	**FL original**	***Stop c.* (próximo)** ***Stop c.* (próximo)**
Smp_063760.1	**Final do íntron 1**	**6**	**AG**	**1**	**M (próxima)**	**1**	**FL original**	***Stop c.* (distante)** ***Stop c.* (próximo)**
Smp_079240.1	**Final do íntron 2**	**3**	**AG**	**2**	**M (próxima)**	**2**	**FL original**	***Stop c.* (próximo)** ***Stop c.* (próximo)**

Gene	Região de entrada do SL	Total de éxons do gene	Sítio aceptor de SLTS	Fase de leitura (FL)	Presença ou ausência de M na FL original	FL com uso do AUG do SL	Presença ou ausência de M ou *stop códon* na FL com uso do AUG do SL	Gene
Smp_083020.1	**Final do íntron 3**	**5**	**AG**	**2**	**M (distante)**	**2**	**FL original**	***Stop c.* (próximo)** ***Stop c.* (próximo)**
Smp_085750.1	**Final do íntron 2**	**5**	**AG**	**1**	**Ausência de M**	**1**	**FL original**	**M (próxima)** ***Stop c.* (próximo)**
Smp_123880.1	**Final do íntron 2**	**4**	**AG**	**3**	**M (próxima)**	**3**	**FL original**	***Stop c.* (próximo)** ***Stop c.* (próximo)**
Smp_143150.1	**Final do íntron 1**	**3**	**AG**	**2**	**M (próxima)**	**2**	**FL original**	***Stop c.* (próximo)** ***Stop c.* (próximo)**
Smp_161940.1	**Final do íntron 3**	**4**	**AG**	**3**	**Ausência de M**	**3**	**FL original**	***Stop c.* (distante)** ***Stop c.* (próximo)**
Smp_042430.1	**Final do íntron 1**	**7**	**AG**	**3**	**Ausência de M**	**3**	**FL original**	***Stop c.* (próximo)** ***Stop c.* (próximo)**
Smp_045960.1	**Final do íntron 9**	**11**	**AG**	**3**	**M (próxima)**	**2**	***Stop c.* (distante)**	**M (próxima)**
Smp_017360.1	**Final do íntron 6**	**7**	**AG**	**3**	**Ausência de M**	**1**	***Stop c.* (próximo)**	***Stop c.* (próximo)**
Smp_050430.1	**Final do íntron 5**	**6**	**AG**	**3**	**Ausência de M**	**2**	***Stop c.* (próximo)**	***Stop c.* (distante)**
Smp_164770.1	**Final do íntron 4**	**10**	**AG**	**2**	**M (próxima)**	**3**	***Stop c.* (distante)**	***Stop c.* (distante)**

Gene	Região de entrada do SL	Total de éxons do gene	Sítio aceptor de SLTS	Fase de leitura (FL)	Presença ou ausência de M na FL original	FL com uso do AUG do SL	Presença ou ausência de M ou *stop códon* na FL com uso do AUG do SL	Gene
Smp_130430.1	**Final do íntron 2**	**5**	**AG**	**3**	**M (próxima)**	**2**	***Stop c.* (próximo)**	***Stop c.* (próximo)**
Smp_050830.1	**Final do íntron 10**	**11**	**AG**	**1**	**Ausência de M**	**2**	**Ausência de M**	**Ausência de M**
Smp_042680.1	**Final do íntron 1**	**6**	**AG**	**2**	**Ausência de M**	**3**	**Ausência de M**	***Stop c.* (distante)**
Smp_022500.1	**Final do íntron 1**	**3**	**AG**	**2**	**M (próxima)**	**1**	***Stop c.* (próximo)**	***Stop c.* (próximo)**
Smp_021070.1	**Final do íntron 8**	**11**	**AG**	**2**	**M (próxima)**	**2**	**FL original**	**M (distante)** ***Stop c.* (distante)**
Smp_054360.1	**Final do íntron 3**	**4**	**AG**	**2**	**M (próxima)**	**2**	**FL original**	***Stop c.* (próximo)** ***Stop c.* (próximo)**
Smp_059800.1	**Final do íntron 7**	**11**	**AG**	**1**	**M (próxima)**	**1**	**FL original**	***Stop c.* (próximo)** ***Stop c.* (próximo)**
Smp_069880.1	**Final do íntron 2**	**3**	**AG**	**1**	**M (distante)**	**1**	**FL original**	**M (distante)** ***Stop c.* (distante)**
Smp_124220.1	**Final do íntron 21**	**23**	**AG**	**1**	**M (distante)**	**1**	**FL original**	***Stop c.* (próximo)** ***Stop c.* (próximo)**

Gene	Região de entrada do SL	Total de éxons do gene	Sítio aceptor de SLTS	Fase de leitura (FL)	Presença ou ausência de M na FL original	FL com uso do AUG do SL	Presença ou ausência de M ou *stop códon* na FL com uso do AUG do SL	Gene
Smp_175750.1	**Final do íntron 2**	**3**	**AG**	**3**	**M (próxima)**	**3**	**FL original**	***Stop c.* (próximo)**
								***Stop c.* (próximo)**
Smp_156940.1	**Final do íntron 6**	**9**	**AG**	**3**	**M (distante)**	**3**	**FL original**	***Stop c.* (próximo)**
								***Stop c.* (próximo)**
Smp_143380.1	**Final do íntron 15**	**16**	**AG**	**1**	**M (próxima)**	**1**	**FL original**	**M (próxima)**
								***Stop c.* (próximo)**
Smp_014010.1	**Final do íntron 10**	**11**	**AG**	**1**	**Ausência de M**	**1**	**FL original**	***Stop c.* (distante)**
								***Stop c.* (próximo)**
Smp_169070.1	**Final do íntron 5**	**7**	**AG**	**2**	**Ausência de M**	**2**	**FL original**	**M (próxima)**
								***Stop c.* (próximo)**
Smp_027360.1	**Final do íntron 1**	**2**	**AG**	**3**	**M (distante)**	**1**	***Stop c.* (próximo)**	***Stop c.* (próximo)**
Smp_032950.1	**Final do íntron 1**	**2**	**AG**	**1**	**M (distante)**	**2**	***Stop c.* (próximo)**	**M (próxima)**
Smp_044260.1	**Final do íntron 1**	**2**	**AG**	**2**	**M (próxima)**	**1**	***Stop c.* (próximo)**	**M (próxima)**
Smp_052500.1	**Final do íntron 1**	**3**	**AG**	**2**	**Ausência de M**	**1**	***Stop c.* (próximo)**	***Stop c.* (distante)**
Smp_069890.1	**Final do íntron 4**	**6**	**AG**	**2**	**M (próxima)**	**1**	***Stop c.* (próximo)**	***Stop c.* (próximo)**

Gene	Região de entrada do SL	Total de éxons do gene	Sítio aceptor de SLTS	Fase de leitura (FL)	Presença ou ausência de M na FL original	FL com uso do AUG do SL	Presença ou ausência de M ou *stop códon* na FL com uso do AUG do SL	Gene
Smp_196620.1	**Final do íntron 3**	**6**	**AG**	**2**	**M (distante)**	**2**	**FL original**	***Stop c.* (próximo)** ***Stop c.* (próximo)**
Smp_171940.1	**Final do íntron 3**	**7**	**AG**	**2**	**Ausência de M**	**3**	***Stop c.* (próximo)**	***Stop c.* (próximo)**
Smp_029150.1	**Final do íntron 1**	**3**	**AG**	**1**	**M (próxima)**	**1**	**FL original**	***Stop c.* (distante)** **Ausência de M**
Smp_059240.1	**Final do íntron 1**	**3**	**AG**	**2**	**M (distante)**	**2**	**FL original**	***Stop c.* (próximo)** ***Stop c.* (distante)**
Smp_068110.1	**Final do íntron 1**	**5**	**AG**	**2**	**Ausência de M**	**2**	**FL original**	***Stop c.* (próximo)** ***Stop c.* (próximo)**
Smp_068480.1	**Final do íntron 2**	**5**	**AG**	**2**	**M (distante)**	**2**	**FL original**	**M (próxima)** ***Stop c.* (próximo)**
Smp_071050.1	**Final do íntron 1**	**2**	**AG**	**2**	**M (distante)**	**2**	**FL original**	**M (próxima)** ***Stop c.* (próximo)**
Smp_095130.1	**Final do íntron 1**	**4**	**AG**	**1**	**Ausência de M**	**1**	**FL original**	***Stop c.* (próximo)** ***Stop c.* (próximo)**
Smp_199400.1	**Final do íntron 1**	**2**	**AG**	**3**	**M (próxima)**	**1**	***Stop c.* (próximo)**	***Stop c.* (próximo)**

Gene	Região de entrada do SL	Total de éxons do gene	Sítio aceptor de SLTS	Fase de leitura (FL)	Presença ou ausência de M na FL original	FL com uso do AUG do SL	Presença ou ausência de M ou *stop códon* na FL com uso do AUG do SL	Gene
Smp_142990.1	**Final do íntron 1**	**14**	**AG**	**3**	**M (distante)**	**3**	**FL original**	***Stop c.* (próximo)** ***Stop c.* (próximo)**
Smp_174180.1	**Final do íntron 1**	**2**	**AG**	**3**	**M (próxima)**	**3**	**FL original**	***Stop c.* (próximo)** ***Stop c.* (próximo)**
Smp_140610.1	**Final do íntron 7**	**9**	**AG**	**1**	**M (distante)**	**1**	**FL original**	**M (próxima)** ***Stop c.* (próximo)**
Smp_048420.1	**Final do íntron 2**	**3**	**AG**	**3**	**M (distante)**	**3**	**FL original**	**M (próxima)** ***Stop c.* (próximo)**
Smp_210280.1	**Final do íntron 3**	**4**	**AG**	**1**	**M (distante)**	**3**	***Stop c.* (próximo)**	**M (próxima)**
Smp_210950.1	**Final do íntron 4**	**8**	**AG**	**3**	**M (distante)**	**1**	***Stop c.* (próximo)**	**M (próxima)**
Smp_166320.1	**Final do íntron 1**	**3**	**AG**	**3**	**M (distante)**	**1**	**M (distante)**	***Stop c.* (distante)**
Smp_026390.1	**Final do íntron 2**	**4**	**AG**	**2**	**Ausência de M**	**2**	**FL original**	**M (próxima)** **M (próxima)**
Smp_054560.1	**Final do íntron 1**	**2**	**AG**	**3**	**Ausência de M**	**3**	**FL original**	***Stop c.* (distante)** ***Stop c.* (distante)**
Smp_068990.1	**Final do íntron 1**	**2**	**AG**	**3**	**Ausência de M**	**3**	**FL original**	**M (próxima)** **Ausência de M**

Gene	Região de entrada do SL	Total de éxons do gene	Sítio aceptor de SLTS	Fase de leitura (FL)	Presença ou ausência de M na FL original	FL com uso do AUG do SL	Presença ou ausência de M ou *stop códon* na FL com uso do AUG do SL	Gene
Smp_210640.1	**Final do íntron 1**	**10**	**AG**	**2**	**Ausência de M**	**2**	**FL original**	***Stop c.* (próximo)** ***Stop c.* (próximo)**
Smp_054200.1	**Final do íntron 4**	**5**	**AG**	**2**	**Ausência de M**	**1**	***Stop c.* (próximo)**	**M (próxima)**
Smp_054780.1	**Final do íntron 1**	**3**	**AG**	**2**	**M (distante)**	**1**	***Stop c.* (próximo)**	***Stop c.* (distante)**
Smp_055600.1	**Final do íntron 3**	**4**	**AG**	**1**	**M (próxima)**	**2**	***Stop c.* (próximo)**	***Stop c.* (próximo)**
Smp_015690.1	**Final do íntron 1**	**8**	**AG**	**2**	**M (distante)**	**2**	**FL original**	***Stop c.* (próximo)** ***Stop c.* (próximo)**
Smp_016410.1	**Final do íntron 14**	**21**	**AG**	**3**	**M (distante)**	**3**	**FL original**	***Stop c.* (próximo)** ***Stop c.* (próximo)**
Smp_041600.1	**Final do íntron 21**	**23**	**AG**	**3**	**Ausência de M**	**3**	**FL original**	***Stop c.* (próximo)** ***Stop c.* (próximo)**
Smp_041650.1	**Final do íntron 1**	**4**	**AG**	**1**	**M (distante)**	**1**	**FL original**	***Stop c.* (distante)** ***Stop c.* (próximo)**
Smp_059660.1	**Final do íntron 2**	**5**	**AG**	**3**	**Ausência de M**	**3**	**FL original**	***Stop c.* (próximo)** ***Stop c.* (próximo)**

Gene	Região de entrada do SL	Total de éxons do gene	Sítio aceptor de SLTS	Fase de leitura (FL)	Presença ou ausência de M na FL original	FL com uso do AUG do SL	Presença ou ausência de M ou *stop códon* na FL com uso do AUG do SL	Gene
Smp_154640.1	**Final do íntron 1**	**2**	**AG**	**2**	**M (distante)**	**2**	**FL original**	***Stop c.* (distante)** ***Stop c.* (próximo)**
Smp_157780.1	**Final do íntron 2**	**3**	**AG**	**1**	**M (próxima)**	**3**	***Stop c.* (próximo)**	***Stop c.* (próximo)**
Smp_045810.1	**Final do íntron 2**	**6**	**AG**	**3**	**M (próxima)**	**2**	***Stop c.* (próximo)**	***Stop c.* (próximo)**
Smp_047410.1	**Final do íntron 2**	**3**	**AG**	**1**	**Ausência de M**	**1**	**FL original**	***Stop c.* (distante)** ***Stop c.* (próximo)**
Smp_097280.1	**Final do íntron 2**	**6**	**AG**	**3**	**Ausência de M**	**1**	***Stop c.* (próximo)**	***Stop c.* (próximo)**
Smp_035410.1	**Final do íntron 1**	**5**	**AG**	**1**	**Ausência de M**	**1**	**FL original**	***Stop c.* (próximo)** **M (próxima)**
Smp_155870.1	**Final do íntron 1**	**6**	**AG**	**3**	**Ausência de M**	**3**	**FL original**	**M (próxima)** ***Stop c.* (próximo)**
Smp_045810.2	**Final do íntron 2**	**7**	**AG**	**3**	**M (próxima)**	**2**	**M (próxima)**	***Stop c.* (próximo)**

Gene	Região de entrada do SL	Total de éxons do gene	Sítio aceptor de SLTS	Fase de leitura (FL)	Presença ou ausência de M na FL original	FL com uso do AUG do SL	Presença ou ausência de M ou *stop códon* na FL com uso do AUG do SL	Gene
Smp_076650.2	**Dentro do éxon 3**	**3**	**AG**	**1**	**M (próxima)**	**2**	***Stop c.* (próximo)**	***Stop c.* (próximo)**
Smp_088390.1	**Dentro do éxon 1**	**3**	**AG**	**2**	**M (próxima)**	**1**	***Stop c.* (próximo)**	***Stop c.* (próximo)**
Smp_173410.1	**Dentro do éxon 2**	**4**	**AG**	**2**	**M (próxima)**	**3**	***Stop c.* (próximo)**	***Stop c.* (próximo)**
Smp_076650.1	**Dentro do éxon 3**	**3**	**AG**	**1**	**M (distante)**	**2**	***Stop c.* (próximo)**	***Stop c.* (próximo)**
Smp_180360.1	**Dentro do éxon 1**	**4**	**AG**	**1**	**M (próxima)**	**2**	***Stop c.* (próximo)**	***Stop c.* (próximo)**
Smp_136980.1	**Dentro do éxon 1**	**7**	**AG**	**1**	**Ausência de M**	**1**	**FL original**	**M (próxima)** ***Stop c.* (próximo)**
Smp_057230.1	**Dentro do éxon 1**	**10**	**AG**	**1**	**M (próxima)**	**1**	**FL original**	***Stop c.* (próximo)** ***Stop c.* (próximo)**
Smp_092570.1	**Dentro do éxon 3**	**4**	**AG**	**2**	**M (próxima)**	**3**	**M (próxima)**	***Stop c.* (próximo)**
Smp_043070.1	**Dentro do éxon 1**	**6**	**AG**	**3**	**M (próxima)**	**2**	***Stop c.* (próximo)**	***Stop c.* (distante)**
Smp_103830.2	**Dentro do éxon 1**	**2**	**AG**	**3**	**M (próxima)**	**2**	***Stop c.* (próximo)**	***Stop c.* (próximo)**
Smp_147290.1	**Dentro do éxon 4**	**9**	**AG**	**1**	**M (próxima)**	**2**	**M (distante)**	**M (próxima)**

Gene	Região de entrada do SL	Total de éxons do gene	Sítio aceptor de SLTS	Fase de leitura (FL)	Presença ou ausência de M na FL original	FL com uso do AUG do SL	Presença ou ausência de M ou *stop códon* na FL com uso do AUG do SL	Gene
Smp_150830.1	**Dentro do éxon 1**	**14**	**AG**	**3**	**M (próxima)**	**1**	***Stop c.* (próximo)**	***Stop c.* (próximo)**
Smp_167890.1	**Dentro do éxon 1**	**5**	**AG**	**3**	**Ausência de M**	**3**	**FL original**	***Stop c.* (próximo)** ***Stop c.* (próximo)**
Smp_173660.1	**Dentro do éxon 1**	**1**	**AG**	**1**	**M (próxima)**	**2**	***Stop c.* (próximo)**	***Stop c.* (distante)**
Smp_094470.1	**Dentro do éxon 1**	**1**	**AG**	**2**	**M (distante)**	**2**	**FL original**	**M (próxima)** ***Stop c.* (próximo)**
Smp_209080.1	**Dentro do éxon 3**	**3**	**AG**	**2**	**M (distante)**	**2**	**FL original**	***Stop c.* (próximo)** ***Stop c.* (próximo)**
Smp_135240.1	**Dentro do éxon 11**	**12**	**AG**	**2**	**Ausência de M**	**3**	**M (próxima)**	***Stop c.* (próximo)**
Smp_080070.1	**Dentro do éxon 5**	**8**	**AG**	**3**	**Ausência de M**	**3**	**FL original**	***Stop c.* (próximo)** ***Stop c.* (próximo)**
Smp_180630.1	**Dentro do éxon 3**	**5**	**AG**	**2**	**Ausência de M**	**2**	**FL original**	***Stop c.* (próximo)** ***Stop c.* (próximo)**
Smp_154230.1	**Dentro do éxon 4**	**6**	**AG**	**2**	**M (distante)**	**1**	***Stop c.* (próximo)**	**M (próxima)**
Smp_201100.1	**Dentro do éxon 1**	**1**	**AG**	**2**	**M (próxima)**	**1**	***Stop c.* (próximo)**	***Stop c.* (próximo)**

Smp_068090.1	**Dentro do éxon 1**	**5**	**AG**	**2**	**M (distante)**	**2**	**FL original**	***Stop c.* (distante)** ***Stop c.* (próximo)**

Ao todo foram analisados 74 transcritos com entrada de SL em íntrons e 22 com entrada em regiões codificadoras. Dos transcritos analisados, 10 foram processados em mais de uma região, no entanto só foi analisado um sítio de cada transcrito devido ao fato de que a utilização e localização do(s) sítio(s) em um mesmo transcrito podem variar entre os estágios do parasito. Portanto, não seria possível obter um consenso sobre a proporção de transcritos com uma ou mais entradas de SL levando em consideração a localização dos sítios.

Como esperado, a maior parte dos transcritos com entrada do SL em íntrons utiliza um dos sítios de *cis-splicing*, o que pode ou não alterar a fase de leitura do transcrito processado. Foi demonstrado que apenas 28% dos AUGs do SL se enquadram na maior fase de leitura dos RNAs receptores de *S. mansoni*, enquanto que 72% dos AUGs do SL do parasito estão fora da fase de leitura principal (que normalmente é a maior) (CHENG *et al*., 2006).

No presente estudo, dos 96 transcritos analisados, 45 são traduzidos por uma ORF diferente da ORF original caso o AUG do SL seja utilizado. Com o uso do AUG do SL, 37 desses 45 transcritos passariam a ser traduzidos por uma ORF com *stop* códon antes de apresentarem algum outro AUG que não seja o do SL. Caso sejam traduzidos, a grande maioria desses transcritos (31 de 37) continuam tendo códon AUG disponível para início da tradução, diferente de quando o AUG do SL é inserido na maior fase, em que apenas 35% dos transcritos de *S. mansoni* continuam tendo outro códon disponível que especifica a metionina iniciadora (CHENG *et al*., 2006).

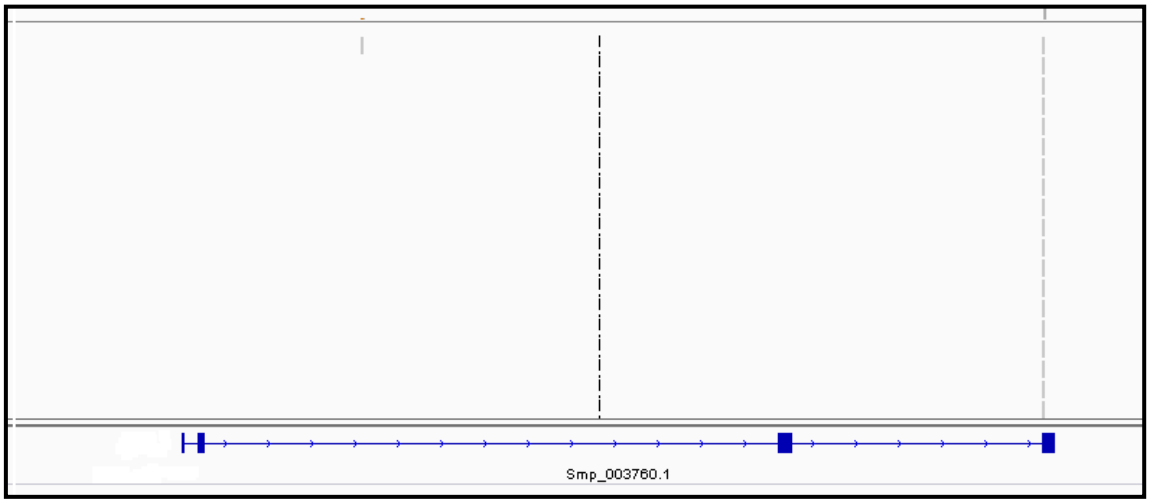
No entanto, a maioria das ORFs de mRNAs de *S. mansoni* que sofrem SLTS, não são iniciadas por uma metionina. Isso também foi visto por DAVIS e colaboradores em 1995 e sugere que o SLTS nesse organismo tenha como função (não sendo função principal) o fornecimento de uma metionina iniciadora para ORFs encurtadas (DAVIS *et al*., 1995).

Em 2006, Cheng e colaboradores provaram através de métodos de transfecção transiente de RNA que o AUG do SL pode sim ser utilizado como códon de início da tradução. Os autores identificaram cDNAs processados por SLTS em bancos de dados biológicos. Os transcritos com inserção do SL na maior fase de leitura foram então alinhados contra bancos de dados proteicos. Assim, foi observado que a maioria das ORFs iniciadas pelo AUG do SL exibem alta similaridade com outras proteínas comparando o ponto em que o SL entrou até o próximo AUG. Após a transfecção transiente em diversos estágios do *Schistosoma*, foi visto que o AUG do SL pode servir

como um códon iniciador *in vivo*. Em seguida o AUG do SL foi mutado e os autores demonstraram que a atividade protéica era perdida quase que por completo (CHENG *et al*., 2006).

No presente trabalho a análise dos genes selecionados foi realizada no visualizador IGV. Durante a visualização de dados, foi possível observar transcritos exibindo diferentes situações decorrentes do SLTS, sendo algumas destacadas nas figuras 15 a 19. A escolha de cinco transcritos para exibição das imagens do IGV foi baseada na seleção de situações distintas entre si e interessantes para exposição do texto. Um exemplo é o caso do gene Smp_135240.1, que apresenta o AUG do SL e um códon AUG alternativo disponível para início da tradução na terceira fase de leitura, enquanto não apresenta nenhum AUG nas outras fases. Porém, este gene também apresenta na mesma fase, um *stop* códon próximo ao AUG do SL e ao outro AUG alternativo. Nas letras a e b das figuras 15 a 19, as "tiras" em cinza representam algumas/todas *reads* alinhadas e as "tiras" azuis representam as regiões intrônicas (mais fina) e exônicas (mais grossa) do transcrito, com pequenas setas indicando o sentido da fita (*forward* ou reversa). As imagens de letra "A" possuem um campo de visão maior, mostrando os genes inteiros e as imagens de letra "B" elucidam a sequência nucleotídica do gene na região de entrada do SL e as três possíveis fases de leitura da fita correspondente ao gene (*forward* ou reversa). A fase de leitura correta/original do gene foi identificada como aquela que não contém *stop* códons em seu interior e a fase de leitura utilizada caso ocorra o uso do códon AUG do SL como códon de início da tradução foi identificada de acordo com a correspondência entre o primeiro códon formado e o aminoácido que ele especifica.

A



B

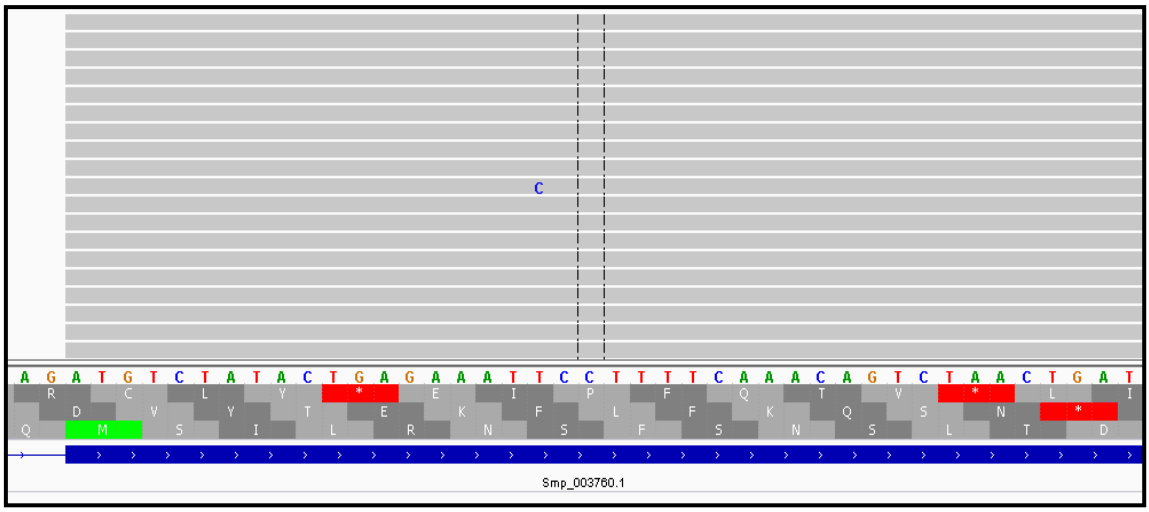


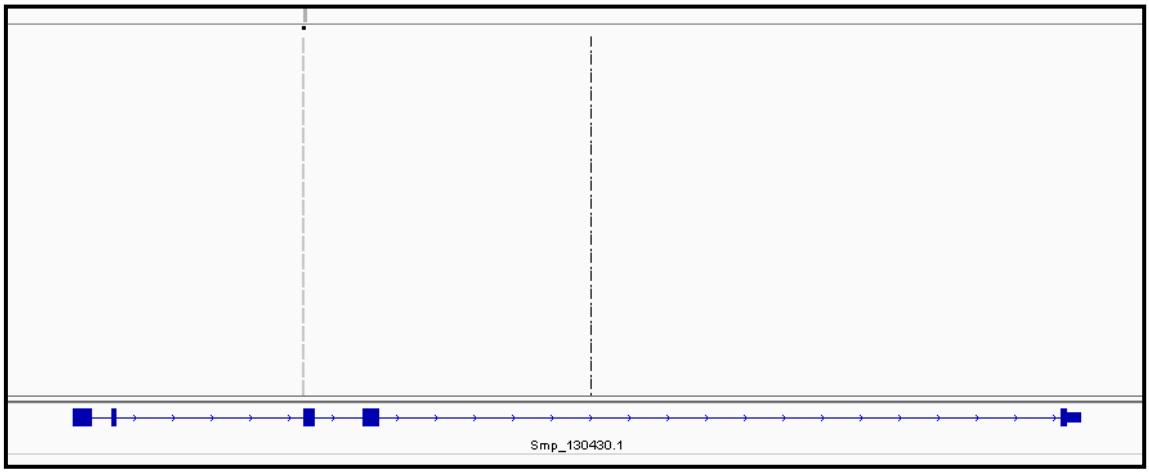
Figura 15: Gene Smp_003760.1: A) Gene apresentado em azul, composto por 4 éxons, com entrada do SL apresentada pelas tiras em cinza no final do íntron 3. **B)** Sequência nucleotídica da região do sítio de SLTS (AG) e três fases de leituras possíveis da fita *forward*. De acordo com o alinhamento do ponto de entrada do SL com relação às três fases, a utilização do AUG do SL como códon de início da tradução coincide com a terceira fase representada na figura.

Na figura 15, o gene Smp_003760.1 mantém a fase de leitura do transcrito original utilizando ou não o AUG do SL para início da tradução. O interessante é que a fase mantida é iniciada por um códon de metionina, portanto, é possível tanto o uso do AUG do SL, como do códon AUG alternativo inicial. As outras fases possuem *stop*

códons próximos ao sítio de SLTS e possivelmente dariam origem a peptídeos sem funcionalidade.

No entanto, mesmo com as possibilidades de uso do AUG alternativo e do AUG do SL, o processamento de SLTS preserva somente o éxon 4, que é o último éxon do gene, assim, o peptídeo gerado continuaria representando uma parte muito pequena na proteína original, que provavelmente também não possui atividade.

A



B

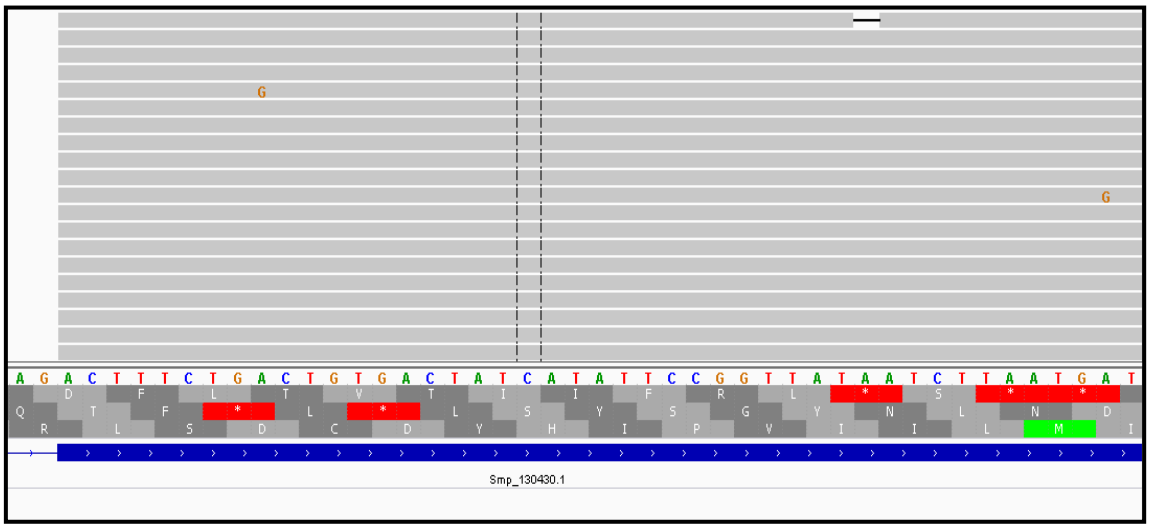
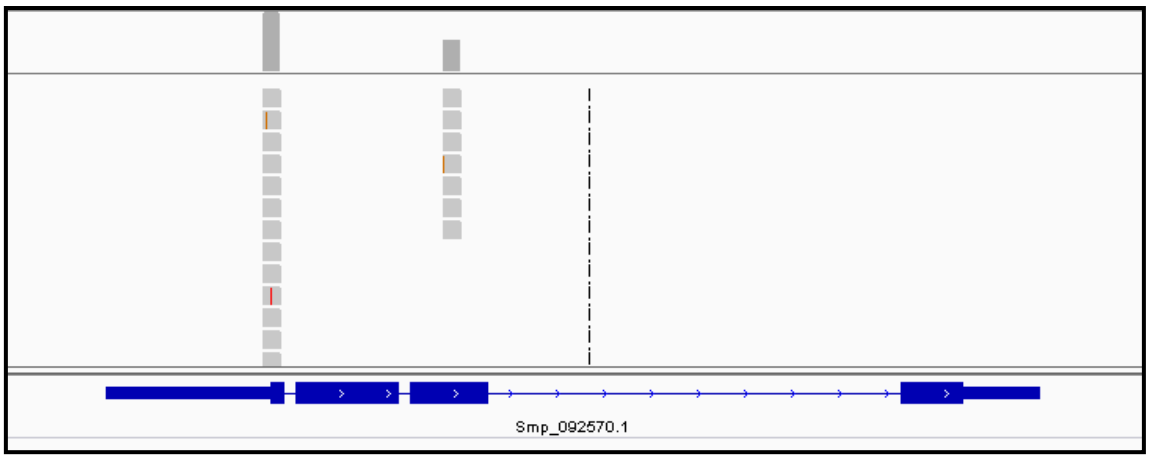


Figura 16: Gene Smp_130430.1 A) Gene apresentado em azul, composto por 5 éxons, com entrada do SL apresentada pelas tiras em cinza no final do íntron 2. **B)** Sequência nucleotídica da região do sítio de SLTS (AG) e três fases de leituras possíveis da fita *forward*. De acordo

com o alinhamento do ponto de entrada do SL com relação às três fases, a utilização do AUG do SL como códon de início da tradução coincide com a segunda fase representada na figura.

Na figura 16, o gene Smp_130430.1 não mantém a fase de leitura da proteína original (terceira fase) caso utilize o AUG do SL como códon de início da tradução. No entanto, a fase correspondente ao uso do AUG do SL e a outra fase alternativa possuem *stop* códons próximos ao sítio de inserção do SL. Provavelmente, a única possibilidade de geração de um peptídeo funcional seria a tradução utilizando a fase de leitura original. Nesse caso a sequência SL junto à sequência que vai do início do éxon 3 até o códon AUG disponível, seriam correspondentes à região 5' UTR.

A



B

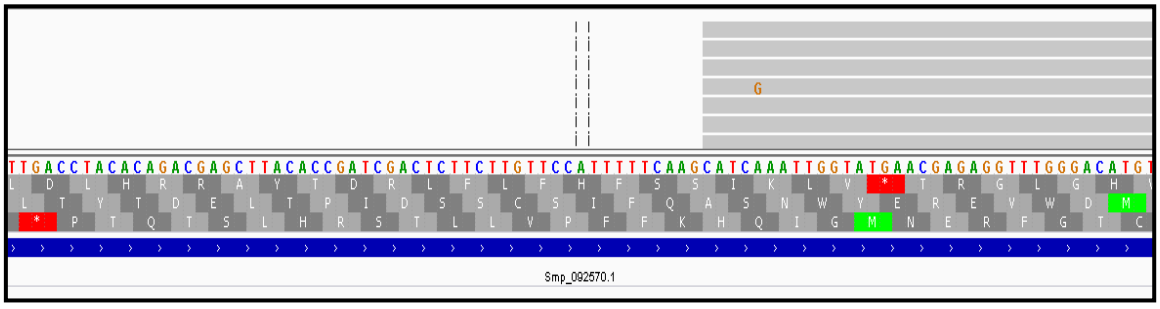
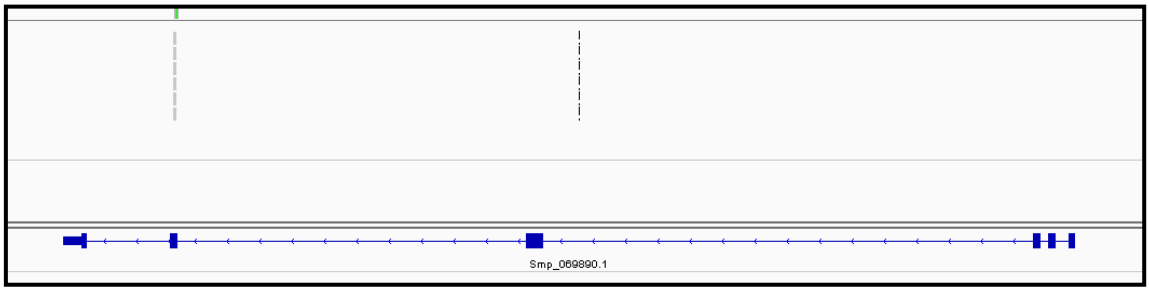


Figura 17: Gene Smp_092570.1 A) Gene apresentado em azul, composto por 4 éxons, com entrada do SL apresentada pelas tiras em cinza dentro do éxon 3. **B)** Sequência nucleotídica da região do sítio de SLTS (AG) e três fases de leituras possíveis da fita forward. De acordo com o

alinhamento do ponto de entrada do SL com relação às três fases, a utilização do AUG do SL como códon de início da tradução coincide com a terceira fase representada na figura.

Na figura 17, o gene Smp_092570.1 não mantém a fase de leitura do transcrito original (segunda fase) caso utilize o AUG do SL como códon de início da tradução. Tanto a fase referente ao uso do AUG do SL, como a fase de leitura original possuem códon AUG alternativo próximo ao sítio de SLTS. Este é um caso com maior possibilidade de geração de um peptídeo que ainda possui grande parte da região correspondente aos domínios protéicos da proteína, caso a fase verdadeira seja traduzida. O gene também apresenta uma entrada de SL na região 5' UTR.

A



B

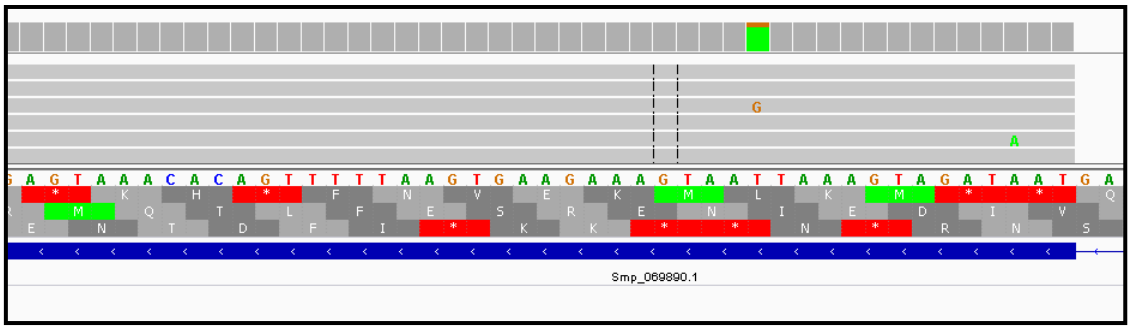


Figura 18; Gene Smp_069890.1 A) Gene apresentado em azul, composto por 6 éxons, com entrada do SL apresentada pelas tiras em cinza no final do íntron 4. **B)** Sequência nucleotídica da região do sítio de SLTS (AG) e três fases de leituras possíveis da fita reversa. De acordo com o alinhamento do ponto de entrada do SL com relação às três fases, a utilização do AUG do SL como códon de início da tradução coincide com a primeira fase representada na figura.

Na figura 18, no gene Smp_069890.1 provavelmente o AUG do SL não é utilizado para início da tradução, pois na fase correspondente existe um *stop* códon logo após o sítio de SLTS. Essa pode ser uma estratégia de inibição da tradução do transcrito. Ainda assim, a situação mais provável é que a tradução ocorra pela fase de leitura original, que por sua vez, possui um AUG alternativo próximo ao ponto de inserção do SL. Nesse caso a sequência SL junto à sequência que vai do início do éxon 5 até o códon AUG disponível, seriam correspondentes à região 5' UTR. Mesmo mantendo a fase verdadeira, a proteína perde os quatro primeiros éxons, podendo então ser produzido um peptídeo muito pequeno e afuncional.

De modo geral, grande parte dos genes apresentou sítios de inserção do SL em regiões diferentes entre os estágios do ciclo de vida do *S. mansoni*. Isso sugere que o *trans-splicing* tem consequências funcionais na regulação do desenvolvimento do parasito. Essa abundância diferencial de *splicing* alternativo também foi observada no *T. brucei* em 2010, no estudo de Nilsson e colaboradores. Neste mesmo trabalho, os autores encontraram cerca de 90 transcritos em que o *splicing* alternativo promoveu a remoção do códon de iniciação da tradução.

Como mencionado anteriormente, o *splicing* alternativo também ocorre em genes com sítios de SLTS na região 5' UTR. Existem casos em *S. mansoni* em que um único gene possui entrada de SL na 5' UTR e alternativamente em íntrons e regiões codificadoras. A figura 19 mostra um exemplo de um gene que sofreu entrada do SL na 5' UTR, no final do primeiro e do segundo íntron utilizando os sítios de *cis-splicing* como sítios de *trans-splicing* e também dentro do éxon 2.

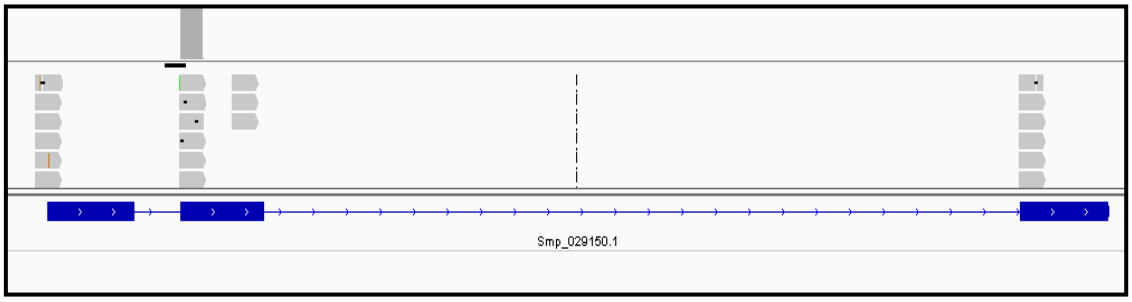
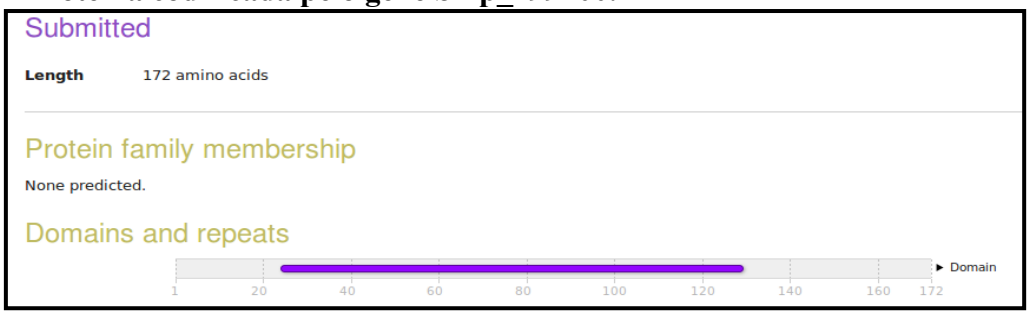


Figura 19: Gene Smp_029150.1. Gene apresentado em azul, composto por 3 éxons, com entrada do SL apresentada pelas tiras em cinza na região 5' UTR, no final do íntron 1, dentro do éxon 2, dentro do éxon 2 e no final do íntron 2.

Esta diversidade de ocorrências de SLTS sugere que não existe um padrão do impacto do processo nos genes alvo. A sequência de aminoácidos das proteínas de 27 transcritos aleatórios (excluindo transcritos de proteínas hipotéticas), dos genes analisados, foi comparada com a sequência original da região codificadora completa dos mesmos para avaliação das conseqüências do SLTS em termos de perda das regiões correspondentes aos domínios proteicos. De todas as 27 proteínas analisadas, 20 apresentaram perda parcial das regiões correspondentes aos domínios proteicos e 7 apresentaram perda total dos domínios. As figuras 20 e 21 mostram esquemas do InterProScan de duas das proteínas analisadas.

Proteína codificada pelo gene Smp_199400.1



Proteína codificada pelo gene Smp_199400.1 processada pelo SLTS

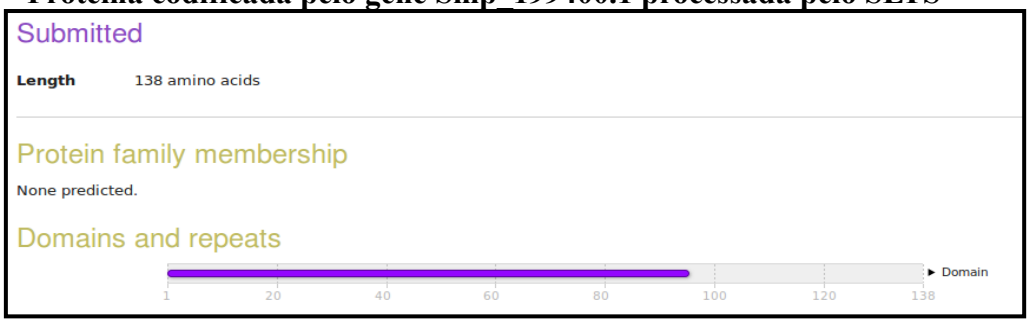
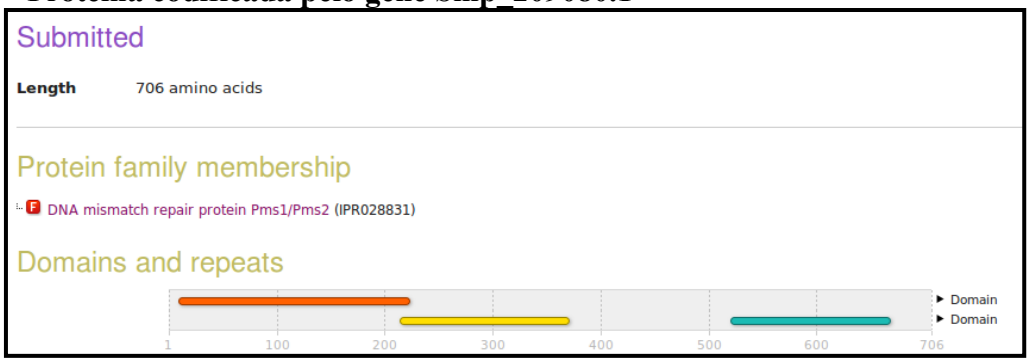


Figura 20: Imagem do InterProScan mostrando a predição de domínios da proteína codificada pelo gene Smp_199400.1 (proteína nativa e derivada do transcrito processado por SLTS).

Na figura 20 a proteína codificada pelo gene Smp_199400.1 perde parte do seu único domínio após o processamento de SLTS. De acordo com a figura, ao analisar a sequência de aminoácidos especificados pelos códons da fase verdadeira do transcrito, observa-se que a proteína do gene Smp_199400.1 perde uma porção relativamente pequena do seu domínio. A perda parcial da região de domínios foi a ocorrência mais comumente

encontrada no conjunto de proteínas analisadas, porém, algumas delas apresentaram perda de uma porção bem maior das regiões de domínios, o que aumenta as chances de produção de um polipeptídeo isento de atividade.

Proteína codificada pelo gene Smp_209080.1



Proteína codificada pelo gene Smp_209080.1 processada pelo SLTS

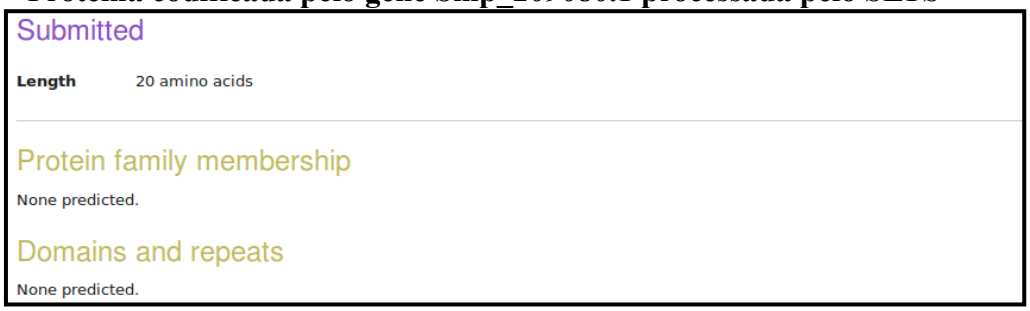


Figura 21: Imagem do InterProScan mostrando a predição de domínios da proteína codificada pelo gene Smp_209080.1 (proteína nativa e derivada do transcrito processado por SLTS).

Na figura 21 a proteína codificada pelo gene Smp_209080.1 perde todos os seus domínios após o processamento de SLTS. Isso significa que ao analisar a sequência de aminoácidos especificados pelos códons da fase verdadeira do transcrito, pode-se afirmar que o processamento de SLTS inviabiliza a síntese de um peptídeo funcional.

O impacto do SLTS no conjunto de proteínas de *S. mansoni* requer a verificação da perda de domínios em proteínas derivadas de transcritos sem mudança de fase de leitura, no sentido de confirmar a conservação da função destas proteínas. Isso só é possível através de estudos experimentais.

Em 2013, Mourão e colaboradores descreveram que o silenciamento do SLTS em esporocistos por *RNA interference* (RNAi) provocou uma redução do tamanho dos parasitos. Isso sugere que o mecanismo possui grande importância no desenvolvimento

do organismo, podendo ser fundamental para a regulação de processos metabólicos (MOURÃO *et al*., 2013).

Ainda se tratando da função do SLTS, o capeamento de mRNAs por adição de SL parece estar relacionado com a garantia de maior estabilidade da molécula em tripanosomas (HUANG; VAN DER PLOEG, 1991). Em *S. mansoni*, além de estabilidade, o SL pode ter alguma funcionalidade no mecanismo de tradução e em processos de transporte, impedindo que proteínas sejam direcionadas ao seu local de atuação (por causar perda do peptídeo sinal, por exmplo).

No que diz respeito à função de *SL-resolved operons*, cerca de apenas 2% dos genes processados por SLTS em *S. mansoni* são unidades policistrônicas, portanto necessitam desta funcionalidade (BORONI, 2014). Portanto, a necessidade do *trans-splicing* no processamento de transcritos policistrônicos em monocistrônicos levantou questões sobre a existência de outros níveis de regulação por parte deste processo (HELM; WILSON; DONELSON, 2008). Este trabalho mostrou a grande diversidade de condições em que as proteínas podem apresentar após serem processadas por SLTS. A condição mais comumente encontrada foi a entrada do SL na região 5’ UTR. Portanto, entradas alternativas de SL, ou seja, inserções em região intrônica ou exônica, gerando peptídeos menores, são menos frequentes. Mesmo com pouca representatividade no proteoma predito do parasito, as formas alternativas de SLTS podem ser um meio para o aumento da diversidade de isoformas proteicas de *S. mansoni*.

6 CONCLUSÕES

Neste trabalho foram apresentadas as possíveis consequências do SLTS no *S. mansoni*. A localização dos sítios de entrada da sequência SL foi determinada em cada transcrito processado no conjunto de dados, permitindo a determinação da frequência de uso de sítios em regiões não traduzíveis (5' UTR), intrônicas e exônicas. Alguns transcritos processados foram analisados quanto à possível ocorrência de mudança de fase de leitura consequente do uso do AUG do SL e quanto à presença do códon de iniciação da síntese proteica em regiões próximas a inserção do SL. Algumas proteínas codificadas por transcritos processados, sem considerar mudança da fase de leitura, foram analisadas quanto à perda parcial ou total de domínios funcionais. O SLTS convencional foi identificado como a forma preferencial do mecanismo no parasito, apresentando uma maioria de transcritos com sítios de SLTS na região 5' UTR. A preferência pela região 5' UTR sugere que o SLTS pode impactar na estabilidade do transcrito e eficiência da tradução. A alta utilização de sítios na 5' UTR mostra que a produção de peptídeos sem funcionalidade, ou de tamanho reduzido são eventos mais raros que, portanto, não podem ser considerados função principal do SLTS e não exercem muita influência no proteoma do parasito. Ainda assim, foi observada a produção de RNAs muito curtos após processamento que poderiam dar origem a peptídeos provavelmente sem função, devido ao seu comprimento ser muito restrito. Por outro lado, a entrada alternativa do SL com produção dessas proteínas modificadas é uma estratégia de ampliação do repertório proteico. Foram identificados genes com entradas da sequência líder em regiões codificadoras e em íntrons, que por sua vez, ocorrem mais frequentemente nos sítios canônicos de *cis-splicing*. Muitos transcritos com entrada do SL no íntron, ou dentro de um éxon não continham AUG alternativo para início da tradução na região próxima ao sitio de SLTS. Isso reforça a possibilidade de utilização do AUG do SL, um evento já relatado em outros estudos. Alguns transcritos apresentaram sítios de entrada do SL em regiões diferentes entre os estágios de *S. mansoni*. Isso nos leva a crer que o *splicing* alternativo tem um papel na regulação do desenvolvimento do parasito. Durante a comparação dos domínios dos peptídeos processados com a sequência completa de aminoácidos das respectivas proteínas, foi visto que a maioria das proteínas perde parcialmente regiões de domínios. No entanto, algumas chegam a perder toda sua sequência conservada correspondente aos domínios,

sendo reduzidas a peptídeos afuncionais. Esta pode ser uma estratégia de repressão da expressão gênica via inibição da síntese proteica por SLTS.

7 PERSPECTIVAS

Selecionar proteínas que sofreram diferentes impactos do SLTS, para testes experimentais que confirmem o dano provocado pelo processamento.

8 REFERÊNCIAS

ABUBUCKER, S.; MCNULTY, S.; N.; ROSA, B. A. *et al*. Identification and characterization of alternative splicing in parasitic nematode transcriptomes. **Parasites e Vectors**, v. 7, p. 151, 2014.

AGABIAN, N. Trans-splicing of nuclear pre-mRNAs. **Cell**, v. 61, p. 1157-1160, 1990.

ALLEN, M. A.; LADEANA; HILLIER, *et al*. A global analysis of C. elegans trans-splicing. **Genome Research**, v. 21, p. 255-264, 2011.

AMARAL, R.; TAUIL, P.; LIMA, D. An analysis of the impact of the Schistosomiasis Control Programme in Brazil . **Memórias Instituto Oswaldo Cruz**, Rio de Janeiro, v. 101, n. 1, p. 79-85, 2006.

BERGET, S.; MOORE, C.; SHARP, P. Spliced segments at the 5' terminous of adenovirus 2 late mRNA. **Proceedings of the National Academy of Sciences USA**., v. 74, n. 8, p. 3171-3175, 1977.

BERRIMAN, M.; HAAS, B.; LOVERDE, P. *et al*. The genome of the blood fluke *Schistosoma mansoni*. **Nature**, v. 460, n. 7253, p. 352-358, 2009.

BITAR, M.; BORONI, M.; MACEDO, A. *et al*. The spliced leader trans-splicing mechanism in different organisms: molecular details and possible biological roles. **Frontiers in Genetics**, v. 4, n. 199, p. 199, 2013.

BLUMENTHAL, T. Gene clusters and polycistronic transcription in eukaryotes. **BioEssays**, v. 20, p. 480-487, 1998.

BLUMENTHAL, T.; GLEASON, K. S. *Caenorhabditis elegans* operons: form and function. **Nature Reviews**. Genetics, v. 4, n. 2, p. 112–120, 2003.

BLUMENTHAL, T. Operons in eukaryotes. **Briefings in Functional Genomics and Proteomics**, v.3, n.3, p. 199-211, 2004.

BLUMENTHAL, T. Trans-splicing and operons. **WormBook**, p. 1-9, 2005.

BLUMENTHAL, T. Trans-splicing and polycistronic transcription in *Caenorbabditis elegans*. **Trends in Genetics**, v. 11, n.4, p. 132-6, 1995.

BOOTHROYD, J.; CROSS, G. Transcripts coding for variant surface glycoproteins of *Trypanosoma brucei* have a short, identical exon at their 5' end. **Gene**, v. 20, n. 2, p. 281-289, 1982.

BORONI, M. O spliced leader trans-splicing no parasito *Schistosoma mansoni*. **Tese de Doutorado** (programa de pós-graduação em bioinformática) – Universidade Federal de Minas Gerais, 2014.

BOTROS, S.; BENNETT, J. Praziquantel Resistance. **Expert Opinion Drug Discovery**, v. 2, n. 1, p. 35-40, 2007.

CAMPBELL, D.; THOMAS, S.; STURM, N. Transcription in kinetoplastid protozoa: why be normal? **Microbes and Infection**, v. 5, n. 13, p. 1231-1240, 2003.

CARVALHO, O., S.; COELHO, P. M. Z.; LENZI, H., L. *Schistosoma mansoni* e Esquistossomose: uma visão multidisciplinar. 20. ed. **Editora Fiocruz**, 2008.

CHENG, G.; COHEN, L.; NDEGWA, D. *et al*. The Flatworm Spliced Leader 3_-Terminal AUG as a Translation Initiator Methionine. **The Journal of Biological Chemistry**, v. 281, n. 2, p. 733-743, 2006.

CHOW, L.; GELINAS, R.; BROKER, T. *et al*. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. **Cell**, v. 12, n. 1, p. 1-8, 1977.

COURA, J.; CONCEIÇÃO, M. Specifc schistosomiasis treatment as a strategy for disease control. **Memórias Instituto Oswaldo Cruz**, v. 105, n. 4, p. 598-603, 2010.

DAVIS, R.; HARDWICK, C.; TAVERNIER, P. *et al*. 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, p. 21813-21819, 1995.

DENKER, J.; ZUCKERMAN, D. New components of the spliced leader RNP required for nematode trans-splicing. **Nature**, v. 417, p. 667-670, 2002.

FRANTZ, C.; EBEL, C.; PAULUS, F. *et al*. Characterization of trans-splicing in Euglenoids. **Curr Genet**, v. 37, n. 6, p. 349-55, 2000.

GANOT, P.; KALLESOE, T.; REINHARDT, R. *et al*. Spliced-Leader RNA trans-splicing in a Chordate, *Oikopleura dioica* with a compact genome. **Molecular and Cellular Biology**, v. 24, n. 17, p. 7795-7805, 2004.

GARBER, M.; GRABHERR, G.; GUTTMAN, M. *et al*. Computational methods for transcriptome annotation and quantification using RNA-seq. **Nature Methods**, v. 8, n. 6, 2011.

GROSSE, S. Schistosomiasis and water resources development: a re-evaluation of an important environment-health linkage. **Technical Working Paper**, v. 2, n. 43, p. 1-43, 1993.

GUIMARÃES, R.; FREITAS, C.; DUTRA, L. Multiple regression for the Shistosomiasis positivity índex estimates in the Minas Gerais state – Brazil at small communities and cities levels. **Open Science Open Minds**. Disponível em: http://www.intechopen.com/,. 2013.

HASTINGS, K. SL trans-splicing: easy come or easy go? **Trends in Genetics**, v. 21, n. 4, p. 240-247, 2005.

HELM, J.; WILSON, M.; DONELSON, J. Different trans RNA splicing events in bloodstream and procyclic *Trypanosoma brucei*. **Molecular and Biochem Parasitology**, v. 159, n. 2, p. 134-137, 2008.

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, 1991.

HUGHES M. G.; ANDREWS, D. W. A single nucleotide is a sufficient 5′ untranslated region for translation in an eukaryotic in vitro system. **Federation of European Biochemical Societies**, v. 414, n. 1, p. 19-22, 1997.

HULSEN, T.; VLIEG, J.; ALKEMA, W. BioVenn – a web application for the comparasion and visualization of biological lists using area-proportional Venn diagrams. **BMC Genomics**, v. 9. n. 488, 2008.

HUNTER, S.; JONES, P.; MITCHELL, A. InterPro in 2011: new developments in the family and domain prediction database. **Nucleic Acids Research**, v. 40, 2012.

JONES, P.; BINNS, D.; CHANG, H. *et al*.InterProScan 5: genome-scale protein function classification. **Bioinformatics**, v. 30, n. 9, p. 1236-1240, 2014.

KATZ, N.; COELHO, P. Clinical therapy of *Schistosomiasis mansoni*: The Brazilian contribution . **Acta Tropica**, v. 108, p. 72–78, 2008.

LANGMEAD, B.; SALZBERG, S. Fast gapped-read alignment with Bowtie 2. **Brief Communications**, v. 9, n. 4, 2012.

LASDA, E. L.; BLUMENTHAL, T. Trans-splicing. **Developmental Biology**, v. 2, n. 3, p. 417–434, 2011.

LI, H., WANG, J., MOR, G. *et al*. A Neoplastic Gene Fusion Mimics Trans-Splicing of RNAs in Normal Human Cells. **Science**, v. 321, 2008.

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.

LOMAN, N. J.; MISRA, R. V.; DALLMAN, T. J. *et al.* Performance comparison of benchtop high-throughput sequencing platforms. **Nature Biotechnology**, v. 30, n. 5, p. 434–439, 2012.

MARTIN, M. Cutadapt removes adapter sequences from high-throughput sequencing reads.**Technical Notes, EMBnot.journal**, 2010.

MATERA, G.; WANG, Z. A day in the life of the spliceosome. **Nature**, v. 15, n. 2, p. 108-121, 2014.

MATSUMOTO, J.; DEWAR, K.; WASSERSCHEID, 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.

MIGNONE, F.; GISSI, C.; LIUNI, S. *et al*. Untranslated regions os mRNAs. **Genome Biology**, v. 3, n. 3, fev., 2002.

MOURÃO, M.; BITAR, M.; LOBO, F. *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*. **Memórias do Instituto Oswaldo Cruz**, v. 108, n. 6, p. 707-717, 2013.
NAGALAKSHMI, U.; WANG, Z.; WAERN, K. *et al*. The transcriptional landscape of the Yeast genome defined by RNA sequencing. **Science**, v. 320, n. 5881, p. 1344-1349, 2008.

NILSEN, T.; GRAVELEY, B. Expansion of the eukaryotic proteome by alternative splicing. **Nature**, v.463, n.7280, p. 457-463, 2010.

NILSSON, D.; GUNASEKERA, K.; MANI, J. *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.

OLIVEIRA, F; KUSEL, J. R.; RIBEIRO, F. *et al*. Response of the surface membrane and excretory system of *Schistosoma mansoni* damage to treatment with praziquantel and other biomelecules. **Parasitology**, v. 132, p. 321–330, 2006.

PETTITT, J.; MULLER, B.; STANSFIELD, I. *et al*. Spliced leader trans-splicing in the nematode *Trichinella spiralis* uses highly polymorphic, noncanonical spliced leaders. **RNA**, v. 14, p. 760-770, 2008.

PICCA-MATTOCIA, L.; VALLE, C., BASSO, A. *et al*. Cystochalisin D abolishes the schistosomicidal activity of praziquantel. **Experimental Parasitology**, v. 115, p. 344–351, 2007.

POUCHKINA, N.; TUNNACLIFFE, A. Spliced Leader RNA – Mediated trans-splicing in phylum rotifera. **Molecular Biology and Evolution**, v. 22, n. 6, p. 1482-1489, 2005.

PREUßER, C.; BINDEREIF, A.; Exo-endo trans-splicing: a new way to link. **Cell Research**, v. 23, p. 1071-1072, 2013.

PREUßER, C.; JAÉ, N.; BINDEREIF, A. mRNA splicing in trypanosomes. **International Jornal of Medical Microbiology**, v. 302, p. 221-224, 2012.

PROTASIO, A.; TSAI, I.; BABBAGE, A. *et al*. A Systematically Improved High Quality Genome and Transcriptome of the Human Blood Fluke *Schistosoma mansoni*. **Plos neglected tropical disease**, v. 6, n. 1, p. e14552012.

RAJKOVIC, A.; DAVIS, R.; SIMONSENT, N. *et al*. A spliced leader is present on a subset of mRNAs from the human parasite *Schistosoma mansoni*. **Proceedings of the National Academy of Sciences**, v. 87, p. 8879-8883, 1990.

REED, R. The organization of 3' Splice-site sequences in mammalian introns. **Genes and Development**, v. 3, p. 2113-2123, 1989.

RICE, P.; LONGDEN, I.; BLEASBY, A. EMBOSS: The European Molecular Biology Open Software Suite. **Resource Internet**, v. 16, n. 6, 2000.

ROTHBERG, J. M.; HINZ W.; REARICK T. M. *et al*. An integrated semiconductor device enabling non-optical genome sequencing. **Nature**, v. 475, n. 7356, p. 348–352, 2011.

STALEY, J. P.; GUTHRIE, C. Mechanical Devices of the Spliceosome: Motors, Clocks, Springs, and Things. **Cell Press,** v. 92, p. 315–326, 1998.

THORVALDSDÓTTIR, H.; ROBINSON, J.; MESIROV, J. P. Integrative Genomics Viewer (IGV): high-performace genomics data visualization and exploration. **Briefings in Bioinformatics**, v. 14, n. 2, p. 178-192, 2013.

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

WHO - WORLD HEALTH ORGANIZATION. **Schistosomiasis**. Fact sheet 115. Disponível em: http://www.who.int/mediacentre/factsheets/fs115/en/., 2013.

WILL, C.; LUHRMANN, R. Protein functions in pre-mRNA splicing. **Current Opinion in Cell Biology**, v. 9, n. 3, p. 320-328, 1997.

WILSON, A.; COULSON, P. Strategies for a schistosome vaccine: can we manipulate the immune response effectively? **Microbes and Infection**, v. 1, p. 535-43, 1999.