

UNIVERSIDADE DE SÃO PAULO  
INSTITUTO DE FÍSICA DE SÃO CARLOS

DANIELE SANTINI JACINTO

Influência de dois elementos de transposição na arquitetura  
do genoma de *Schistosoma mansoni*

São Carlos  
2014



DANIELE SANTINI JACINTO

Influência de dois elementos de transposição na arquitetura  
do genoma de *Schistosoma mansoni*

Tese apresentada ao Programa de Pós-  
Graduação em Física do Instituto de Física  
de São Carlos da Universidade de São  
Paulo para obtenção do título de Doutor em  
Ciências

Área de concentração: Física Aplicada  
Opção: Física Computacional

Orientador: Prof. Dr. Ricardo De Marco

Versão Corrigida

(Versão original disponível na Unidade que aloja o Programa)

São Carlos

2014

AUTORIZO A REPRODUÇÃO E DIVULGAÇÃO TOTAL OU PARCIAL DESTE TRABALHO, POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO PARA FINS DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

Ficha catalográfica elaborada pelo Serviço de Biblioteca e Informação do IFSC,  
com os dados fornecidos pelo(a) autor(a)

Jacinto, Daniele Santini  
Influência de dois elementos de transposição na  
arquitetura do genoma de *Schistosoma mansoni* /  
Daniele Santini Jacinto; orientador Ricardo De  
Marco - versão corrigida -- São Carlos, 2014.  
186 p.

Tese (Doutorado - Programa de Pós-Graduação em  
Física Aplicada Computacional) -- Instituto de Física  
de São Carlos, Universidade de São Paulo, 2014.

1. *Schistosoma*. 2. Transposons. 3.  
Bioinformática. I. De Marco, Ricardo , orient. II.  
Título.

## Folha de Aprovação



A todos que me incentivaram e me auxiliaram.





## **AGRADECIMENTOS**

Agradeço ao Prof. Dr. Ricardo De Marco por ter me dado a oportunidade de estudar algo que eu tanto desejava. Por sua compreensão com relação às limitações decorrentes do fato de eu trabalhar e estudar. Pela dedicação e atenção que sempre teve com minhas dúvidas e com este trabalho.

Agradeço à minha amiga pessoal e de grupo de pesquisa Gisele Strieder Philippsen, pelas conversas, trocas de informações e de aprendizado. E também à nossa amiga Débora Corrêa, que embora esteja em outra área de pesquisa, sempre está presente.

Agradeço a todos os funcionários da secretaria da Pós-Graduação, da Biblioteca e da Gráfica do IFSC. Sempre me atenderam com muita atenção e dedicação.

Agradeço aos meus amigos de trabalho. Cada um, a sua maneira, me apoiou e me ajudou a conciliar os horários de trabalho e de estudo.

Agradeço aos meus pais, Norivaldo (em memória) e Maria, pela educação que me deram e por terem me ensinado a persistir sempre.

Agradeço aos meus familiares, ao meu afilhado Pedro Repenning de Almeida, e a todos os meus amigos, que souberam compreender minha ausência em muitas datas, mas sempre me apoiaram.

Agradeço imensamente a Deus, por mais essa oportunidade, por tudo que consegui aprender e por todas as pessoas que conheci durante essa fase.



## RESUMO

JACINTO, D. S. **Influência de dois elementos de transposição na arquitetura do genoma de *Schistosoma mansoni***. 2014. 176p. Tese (Doutorado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2014.

Elementos transponíveis são elementos genéticos capazes de transpor para diferentes locais em um genoma hospedeiro. Na sua descoberta, considerou-se que tais elementos não apresentavam funções celulares úteis, classificando-os como genes parasitas. Atualmente, além do carácter deletério, reconhece-se que eles contribuem para a evolução dos genomas e, em alguns casos, podem realizar algumas funções celulares. Utilizando recursos de bioinformática, realizamos estudos para verificar a influência de duas famílias de retrotransposons non-LTR (Perere-3 e SR2) no genoma do *Schistosoma mansoni*. Estudos preliminares indicam que após a divergência entre *S. japonicum* e *S. mansoni*, esses elementos tiveram uma grande expansão em seu número de cópias em *S. mansoni*, sem paralelo em *S. japonicum*. Análises das regiões intrônicas que contêm inserções de qualquer uma destas duas famílias de retrotransposon em *S. mansoni*, mostrou que houve aproximadamente 30% de aumento no tamanho dos íntrons e aumento do conteúdo GC, quando comparado com os íntrons ortólogos de *S. japonicum*. As inserções foram diferencialmente representadas ao longo das estruturas dos genes com a acumulação preferencial nos íntrons localizados nas regiões terminais dos genes. As inserções dos dois elementos de transposição tendem a orientar-se na direção oposta da transcrição dos genes. As inserções de trechos do elemento SR2 enriquecidos em motivos CpG foram observados com maior frequência do que o esperado, sugerindo que estas inserções podem contribuir nas funções de genes. Nas regiões intergênicas, foi possível prever sítios para ligação de fatores de transcrição ao longo das sequências de ambos os retrotransposons. Também foi observado que elementos SR2 tendem a se fixar em regiões que flanqueiam genes codificando proteínas transmembranares, as quais podem estar envolvidas na relação hospedeiro-parasita. Usando dados de transcrição de *S. mansoni* disponíveis publicamente, foram detectados 94 casos possíveis de exonização de

inserções dos retrotransposons, produzindo mudanças do produto proteico. Estes resultados sugerem que os elementos Perere-3 e SR2 podem promover mudanças funcionais e estruturais relevantes nos genes de *S. mansoni* e pode ter contribuído significativamente para a diferenciação entre *S. mansoni* e *S. japonicum*.

Palavras-chave: Schistosoma. Transposons. Bioinformática.

## ABSTRACT

JACINTO, D. S. **Influence of two transposable elements in the genome architecture of *Schistosoma mansoni***. 2014. 176p. Tese (Doutorado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2014.

Transposable elements are genetic elements capable of transpose to different locations at a host genome. At their discovery, it was considered that such elements had no useful cellular functions, leading to their classification as parasitic genes. Currently, in addition to the deleterious character, it is recognized that they contribute to the evolution of genomes and, in some cases, may perform some cellular functions. Using bioinformatics resources, we have conducted studies to verify the influence of two families of non-LTR retrotransposons (Perere-3 and SR2) in the *Schistosoma mansoni* genome. Preliminary studies indicate that after the divergence between *S. japonicum* and *S. mansoni*, these elements had a great expansion in their copy number in *S. mansoni*, without parallel expansion in *S. japonicum*. The analysis of the intron regions containing insertions from either of these two families of transposons in *S. mansoni*, showed that there was approximately 30% of increase in the intron size and GC content when compared to orthologous introns from *S. japonicum*. Insertions were differentially represented along the gene structures with preferential accumulation in introns located at the terminal regions of the genes. Insertions of both transposon elements tended to orientate themselves in the opposite direction of gene transcription. The insertions of SR2 transposon regions enriched in CpG motifs were observed in higher frequency than expected, suggesting that these regions might be contributing in the gene functions. In the intergenic regions, it was possible to predict transcription factors binding sites along the sequences of both retrotransposons and also observed that SR2 elements were preferentially fixed at regions flanking genes coding for transmembrane proteins, which may be involved in parasite-host relationship. Using publicly available transcript data from *S. mansoni*, we detected 94 possible cases of exonization of transposon insertion, producing changes of the protein product. These results suggest that Perere-3 and SR2 insertions may promote relevant functional and structural changes in the *S. mansoni*

genes and may have significantly contributed to the differentiation between *S. mansoni* and *S. japonicum*.

Keywords: Schistosoma. Transposons. Bioinformatics.

## Lista de Figuras

Figura 1 - Classificação dos elementos de transposição em relação aos mecanismos de transposição e organização estrutural.....	37
Figura 2 - Hipótese sobre o processo de evolução de uma família de TE no genoma. ....	38
Figura 3 - Organização estrutural de retrotransposon do tipo não-LTR.....	39
Figura 4 - Etapas do processo de retrotransposição.....	40
Figura 5 - Exemplos de diferentes genes de TE que foram domesticados pelo genoma do hospedeiro .....	41
Figura 6 - Alterações que podem ocorrer em decorrência da inserção de elementos de transposição na fase de transcrição do gene .....	43
Figura 7 - Hipótese sobre a origem Asiática e dispersão do <i>Schistosoma</i> . ....	45
Figura 8 - Ciclo da esquistossomose.....	46
Figura 9 - Representação das diferentes classes de retrotransposons.....	49
Figura 10 - Distribuição da distância (par a par) entre os pares de bases do domínio da transcriptase reversa de membros de uma mesma família.....	50
Figura 11 - Árvore filogenética das sequências de nucleotídeos equivalentes ao trecho do domínio RT inseridos pelos elementos Perere-3/SR3 em <i>S. mansoni</i> .....	51
Figura 12 - Estrutura do retrotransposon SR2. ....	52
Figura 13 - Estrutura do retrotransposon Perere-3.....	52
Figura 14 - Metodologia utilizada para implementar as análises para verificar a possível influência das inserções dos elementos de transposição Perere-3 e SR2 no genoma de <i>S. mansoni</i> . ....	57
Figura 15 - Exemplificação da estruturação de um pipeline.....	58
Figura 16 - Nomenclatura utilizada para definir as regiões intergênicas .....	60
Figura 17 - Ilustração gráfica da metodologia utilizada para representar uma possível incidência de trechos dos TEs em íntrons específicos.....	69
Figura 18 - Representação esquemática do método de fração utilizado para definir a posição do TE no íntron e a distância na qual o TE ocorreu em relação ao éxon situado a 5´ .....	70
Figura 19 - Representação esquemática da metodologia de <i>sliding windows</i> .....	71

Figura 20 - Padrão de distribuição do tamanho dos íntrons de <i>S. mansoni</i> sem inserções dos elementos estudados e seus ortólogos de <i>S. japonicum</i> .....	73
Figura 21 - Padrão de distribuição do tamanho dos íntrons de <i>S. mansoni</i> com elementos Perere-3 e seus ortólogos de <i>S. japonicum</i> .....	75
Figura 22 - Padrão de distribuição do tamanho dos íntrons de <i>S. mansoni</i> com elementos SR2 e seus ortólogos de <i>S. japonicum</i> .....	75
Figura 23 - Distribuição do número de ocorrências do elemento Perere-3 e SR2 em uma mesma região intrônica.....	77
Figura 24 - Frequência encontrada para as diferentes porções do retrotransposon nas cópias encontradas no genoma de <i>S. mansoni</i> . .....	77
Figura 25 - Estrutura do elemento SR2 autônomo e não autônomo (SR2 DEL)..	78
Figura 26 - Boxplot ilustrando o comprimento dos trechos do elemento Perere-3 e SR2 nas regiões intrônicas.....	78
Figura 27 - Percentual de íntrons contendo elementos Perere-3 em relação ao número total de íntrons dos genes do conjunto ORTO-ql. ....	79
Figura 28 - Percentual de íntrons contendo elementos SR2 em relação ao número total de íntrons dos genes do conjunto ORTO-ql. ....	80
Figura 29 - Verificação da posição do elemento Perere-3 dentro dos íntrons analisados.....	81
Figura 30 - Verificação da posição do elemento SR-2 dentro dos íntrons analisados. ....	81
Figura 31 - Distribuição da quantidade de trechos mais completos dos elementos Perere-3 e SR2 .....	83
Figura 32 - Representação gráfica do conteúdo GC de um íntron utilizando a metodologia de <i>sliding windows</i> .....	85
Figura 33 - Distribuição do percentual médio de bases GC dos íntrons de <i>S. mansoni</i> com elementos Perere-3, íntrons ortólogos de <i>S. japonicum</i> e íntrons de <i>S. mansoni</i> que não apresentaram inserções .....	86
Figura 34 - Distribuição do percentual médio de bases GC dos íntrons de <i>S. mansoni</i> com elementos SR2, íntrons ortólogos de <i>S. japonicum</i> e íntrons de <i>S. mansoni</i> que não apresentaram inserções .....	86



Figura 35 - Percentual de íntrons com trechos do elemento SR2 contendo ilhas CpG .....	87
Figura 36 - Verificação da posição do elemento SR2 contendo ilhas CpG dentro dos íntrons analisados. ....	88
Figura 37 - Representação do número de elementos SR2 reais (observados) e simulados (esperados) cujos trechos correspondem a ilhas CpG presentes nos íntrons. ....	89
Figura 38 - Resultados da análise do programa Ontologizer .....	90
Figura 39 - Nomenclatura utilizada para definir as regiões intergênicas.....	98
Figura 40 - Método de fração utilizado para definir a posição do elemento na região intergênica e a distância na qual o elemento ocorreu da região promotora do gene, ou seja, da extremidade 5' .....	101
Figura 41 - Percentual de regiões intergênicas com elementos Perere-3 e SR2 .....	103
Figura 42 - Boxplot representando a distribuição de comprimentos das regiões intergênicas com a presença dos elementos Perere-3, SR2 e sem retrotransposons. ....	104
Figura 43 - Distribuição do número de ocorrências do elemento Perere-3 em uma mesma região intergênica. ....	105
Figura 44 - Distribuição do número de ocorrências do elemento SR2 em uma mesma região intergênica. ....	106
Figura 45 - Boxplot ilustrando o comprimento dos elemento Perere-3 e SR2 nas regiões intergênicas .....	106
Figura 46 - Distribuição das bases dos retrotransposons Perere-3 e SR2 inseridas nas regiões intergênicas identificadas entre os genes ortólogos.....	108
Figura 47 - Distribuição da frequência de sítios preditos através do programa Jaspar, para a ligação de fatores de transcrição.....	108
Figura 48 - Posicionamento dos sítios preditos para a ligação de fatores de transcrição na sequência do elemento Perere-3.....	110
Figura 49 - Posicionamento dos sítios preditos para a ligação de fatores de transcrição na sequência do elemento SR2.....	111
Figura 50 - Distribuição e proximidade das extremidades 5' das regiões intergênicas dos elementos Perere-3.....	112

Figura 51 - Distribuição e proximidade das extremidades 5' das regiões intergênicas dos elementos SR2.....	113
Figura 52 - Boxplot ilustrando a distância, em bases, na qual se encontram os elementos Perere-3 e SR2 das extremidades 5' nas regiões intergênicas .....	113
Figura 53 - Percentual das regiões intergênicas entre os genes ortólogos e percentual das regiões com trechos de SR2 que equivalem à ilhas CpG .....	115
Figura 54 - Boxplot ilustrando a distância, em bases, na qual se encontram os trechos dos elementos SR2 equivalentes à ilhas CpG, em relação a extremidade 5' das regiões intergênicas. ....	115
Figura 55 - Representação do número de elementos SR2 reais (observados) e simulados (esperados) cujos trechos correspondem a ilhas CpG presentes nas regiões intergênicas. ....	116
Figura 56 - Resultados da análise do programa Ontologizer para Perere-3.....	117
Figura 57 - Resultados da análise do programa Ontologizer para SR2.....	118
Figura 58 - Distribuição da quantidade de hélices transmembranares preditadas através do programa TMHMM .....	119
Figura 59 - Ilustração das etapas implementadas para análise dos elementos de transposição nas regiões de mRNAs. ....	126
Figura 60 - Distribuição de trechos similares aos elementos de transposição Perere-3 e SR2 nas diferentes regiões de mRNAs maduros.....	128
Figura 61 - Frequência das bases do retrotransposon Perere-3 em todas as EST do banco de dados e nas EST representando transcritos derivados de genes contendo presença de elementos na UTRs ou regiões codificantes .....	130
Figura 62 - Frequência das bases do retrotransposon SR2 em todas as EST do banco de dados e nas EST representando transcritos derivados de genes contendo presença de elementos na UTRs ou regiões codificantes. ....	130
Figura 63 - Boxplot ilustrando o comprimento dos elementos Perere-3 e SR2 encontrados em sequências de transcritos derivados de genes do <i>S. mansoni</i> . ....	132

Figura 64 - Visualização gráfica do software Spidey para verificar a predição dos éxons da EST. ....	181
Figura 65 - Visualização gráfica do software Spidey para verificar a predição dos éxons da EST CD202813. ....	182
Figura 66 - Visualização gráfica do software Spidey para verificar a predição dos éxons da EST CF495598 .....	183
Figura 67 - Visualização gráfica do software Spidey para verificar a predição dos éxons da EST CF495807.....	184
Figura 68 - Visualização gráfica do software Spidey para verificar a predição dos éxons da EST CF500292.....	185
Figura 69 - Visualização gráfica do software Spidey para verificar a predição dos éxons da EST CD112384 .....	186



## Lista de tabelas

- Tabela 1 - Lista de proteínas do banco de dados NR do NCBI representando o melhor alinhamento com TCs com presença dos elementos Perere-3.. 133
- Tabela 2 - Lista de proteínas do banco de dados NR do NCBI representando o melhor alinhamento com TCs com presença dos elementos SR2..... 133
- Tabela 3 - Lista de genes, com elementos Perere-3, que apresentaram enriquecimento com o processo biológico "single-organism process"... 173
- Tabela 4 - Lista de genes, com elementos Perere-3, que apresentaram enriquecimento com o processo biológico "aromatic compound biosyntetic proc"..... 175
- Tabela 5 - Lista de genes, com elementos SR2, que apresentaram enriquecimento com componente celular "intrinsic to membrane" e predição para mais do que 5 hélices transmembranares ..... 179



## Lista de quadros

Quadro 1 - Pipeline implementado para a organização dos dados iniciais.....	149
Quadro 2 - Pipeline implementado para a identificação dos genes ortólogos.....	150
Quadro 3 - Pipeline implementado para o mapeamento dos elementos de transposições SR2 e Perere-3 no genoma de <i>S. mansoni</i> .....	151
Quadro 4 - Pipeline implementado para o mapeamento dos elementos de transposição SR2 e Perere-3 nos íntrons dos genes ortólogos.....	153
Quadro 5 - Pipeline implementado para análise sobre o tamanho dos íntrons com e sem inserções.....	154
Quadro 6 - Pipeline para definição da posição da inserção no gene.....	155
Quadro 7 - Pipeline para definição da distância da inserção em relação aos éxons.....	156
Quadro 8 - Pipeline para definição do %GC dos íntrons com inserções de SR2 e Perere-3.....	158
Quadro 9 - Pipeline da análise para definir as inserções que correspondiam à ilhas CpG.....	159
Quadro 10 - Pipeline do mapeamento das regiões intergênicas entre os genes ortólogos.....	161
Quadro 11 - Pipeline do mapeamento das inserções nas regiões intergênicas dos genes ortólogos.....	162
Quadro 12 - Pipeline para definir as regiões dos elementos de transposição que mais se inseriram nas regiões intergênicas flanqueadas pelos genes ortólogos.....	162
Quadro 13 - Pipeline para descrever as características das inserções.....	163
Quadro 14 - Pipeline da análise realizada para verificar enriquecimento dos genes que flanqueiam regiões intergênicas com inserções.....	164
Quadro 15 - Pipeline para realizar a predição da possível topologia das proteínas que podem ser transcritas a partir dos genes que flanqueiam regiões intergênicas.....	165

Quadro 16 - Pipeline para verificar proximidade das inserções das extremidades 5' das regiões intergênicas.....	166
Quadro 17 - Pipeline das análises realizadas sobre as inserções nas regiões intergênicas que correspondem à ilhas CpG.....	167
Quadro 18 - Pipeline da análise para definir as inserções nas regiões codificantes.....	171
Quadro 19 - Resumo das análises para verificação da possível ocorrência de splicing entre o trecho similar ao elemento de transposição e o restante da sequência da EST CF495598.....	183
Quadro 20 - Resumo das análises para verificação da possível ocorrência de splicing entre o trecho similar ao elemento de transposição e o restante da sequência de EST CF495807.....	184
Quadro 21 - Resumo das análises para verificação da possível ocorrência de splicing entre o trecho similar ao elemento de transposição e o restante da sequência de EST CF500292.....	185
Quadro 22 - Resumo das análises para verificação da possível ocorrência de splicing entre o trecho similar ao elemento de transposição e o restante da sequência de EST CD112384.....	186



## LISTA DE ABREVIATURAS E SIGLAS

<b>bp</b>	pares de bases ( <i>base pair</i> )
<b>cDNA</b>	DNA complementar ( <i>Complementary DNA</i> )
<b>DNA</b>	Ácido desoxirribonucleico ( <i>Deoxyribonucleic acid</i> )
<b>dsRNA</b>	RNA fita dupla ( <i>double-stranded RNA</i> )
<b>LINE</b>	<i>Long Interspersed Nuclear Elements</i>
<b>Mb</b>	Mega bases
<b>miRNA</b>	micro RNA
<b>mRNP</b>	<i>messenger Ribonucleoprotein</i>
<b>Non-LTR</b>	<i>Non-Long Terminal Repeats</i>
<b>ORF</b>	Quadro de leitura aberto ( <i>Open Reading Frame</i> )
<b>RNA</b>	Ácido Ribonucleico ( <i>Ribonucleic acid</i> )
<b>RNP</b>	Ribonucleoproteico
<b>SINE</b>	<i>Short Interspersed Nuclear Elements</i>
<b>sRNA</b>	<i>small RNA</i>
<b>TE</b>	Elementos de transposição ( <i>Transposable Elements</i> )
<b>TIR</b>	Terminais repetidos invertidos ( <i>Terminal Inverted Repeats</i> )
<b>TPRT</b>	<i>Target Primed Reverse Transcription</i>
<b>TSD</b>	<i>Target Site Duplications</i>
<b>TSS</b>	Sítio de início da transcrição ( <i>Transcription Start Site</i> )
<b>UTR</b>	<i>UnTranslated Region</i>



# Sumário

<b>1</b>	<b>Introdução.....</b>	<b>29</b>
<b>2</b>	<b>Contextualização.....</b>	<b>35</b>
2.1	Elementos de transposição.....	35
2.2	O <i>Schistosoma mansoni</i> e o <i>Schistosoma japonicum</i> .....	44
2.3	Os elementos Perere-3 e SR2 de <i>S. mansoni</i> .....	48
2.4	Análise de genomas em larga escala .....	52
<b>3</b>	<b>Organização dos dados iniciais.....</b>	<b>57</b>
3.1	Metodologia.....	59
3.1.1	Definição dos genes ortólogos.....	60
3.1.2	Mapeamento dos elementos SR2 e Perere-3 no genoma.....	61
3.2	Resultados .....	61
<b>4</b>	<b>Análises nas regiões intrônicas.....</b>	<b>65</b>
4.1	Considerações iniciais.....	65
4.2	Metodologia.....	66
4.2.1	Tamanho dos íntrons em <i>S. mansoni</i> e <i>S. japonicum</i> .....	67
4.2.2	Posição das inserções no gene.....	67
4.2.3	Distância das inserções em relação aos éxons.....	68
4.2.4	Conteúdo CG .....	70
4.2.5	Inserções de ilhas CpG.....	71
4.2.6	Enriquecimento de termos de Gene Ontology no conjunto de genes com inserções.....	72
4.3	Resultados e discussão.....	73
4.3.1	Posição das inserções no gene e no íntron.....	79
4.3.2	Influência de elementos SR2 e Perere-3 no conteúdo GC dos íntrons..	84
<b>5</b>	<b>Regiões não traduzidas de genes, elementos cis-regulatórios e regiões intergênicas.....</b>	<b>95</b>
5.1	Considerações Iniciais.....	95
5.2	Metodologia.....	97
5.2.1	Características dos elementos no genoma .....	99
5.2.2	Proximidade dos elementos em relação às extremidades 5'.....	100
5.2.3	Inserções de ilhas CpG.....	101

5.2.4 Enriquecimento dos genes que flanqueiam regiões intergênicas com retrotransposons.....	102
5.3 Resultados e Discussão.....	102
5.4 Características dos elementos no genoma .....	105
5.5 Proximidade das extremidades 5'.....	112
5.6 Inserções de ilhas CpG.....	114
5.7 Enriquecimento dos genes que flanqueiam as regiões intergênicas com retrotransposons.....	117
6 Análises nas regiões presentes em mRNAs.....	123
6.1 Considerações Iniciais.....	123
6.2 Metodologia.....	124
6.3 Resultados e Discussão.....	127
7 Conclusões.....	137
Referências.....	139
Anexo I-Pipelines das análises para organização dos dados iniciais.....	149
Anexo II-Pipelines das análises nas regiões intrônicas.....	153
Anexo III-Pipelines das análises nas regiões não traduzidas de genes, elementos cis-regulatórios e regiões intergênicas.....	161
Anexo IV-Pipelines das análises nas regiões presentes em mRNA.....	169
Anexo V-Genes com elementos P3 nos íntrons e com enriquecimento .....	173
Anexo VI-Genes com elementos P3 nas regiões intergênicas e com enriquecimento.....	175
Anexo VII-Genes com elementos SR2 nas regiões intergênicas e com enriquecimento.....	179
Anexo VIII-Análise das inserções presentes em mRNA para trechos inseridos de forma não contínua em relação a outros trechos das ESTs .....	181

# Capítulo 1

## Introdução

---



## 1 Introdução

Elementos de transposição são sequências de DNA ou RNA que têm a capacidade de se translocar no genoma da célula de origem. São classificados mediante os elementos utilizados em seu processo de transposição e mediante a sua organização estrutural.

Os elementos de transposição foram identificados na década de 50 por Barbara McClintock, que sugeriu que tais elementos desempenhavam um papel importante na regulação do genoma, sem desconsiderar os possíveis efeitos danosos da ocorrência de um DNA com mobilidade.(1) Posteriormente, Doolittle e Sapienza (2), Orgel e Crick (3), na década de 60 e 70, propuseram que esses elementos poderiam ser considerados parasitas colonizando o genoma. Observaram que tais elementos apresentavam capacidade de replicação autônoma, utilizavam recursos da própria maquinaria celular para finalizar sua integração no genoma e não produziam proteínas que contribuía com o metabolismo celular. Atualmente, além desse aspecto deletério também já foram observadas a cooptação de alguns trechos inseridos nas funções celulares (4), exercendo assim um impacto positivo na célula. Esse processo de conversão de um elemento de transposição em um componente genético integrado a sistemas do genoma hospedeiro, denomina-se domesticação.(5)

As inserções desses elementos de transposição, exercem grande influência na arquitetura de genomas, pois podem alterar a expressão gênica dos organismos, promover o silenciamento e dar origem a novos genes.(6,7) Tais alterações sugerem que esses elementos influenciam a evolução genômica das espécies, desempenhando um papel importante em eventos de adaptação e especiação dos organismos.(8)

Com a recente disponibilização das sequências dos genomas das espécies *Schistosoma mansoni* (*S. mansoni*) (9) e *Schistosoma japonicum* (*S. japonicum*) (10), surgiram novas possibilidades para o estudo da influência de elementos de transposição na evolução dos genomas de organismos dessa família.

*S. mansoni* é membro da família *Schistosomatidae*, classe *Trematoda*, subclasse *Digenea*. Das espécies de *Schistosoma* que habitualmente parasitam o homem, somente *S. mansoni* mantém ciclo de vida na América Central e do Sul devido à existência do caramujo da família *Planorbidae* do gênero *Biomphalaria*. Também é encontrada na África e no Oriente Médio. *S. japonicum* se concentra no Sudeste Asiático e no Pacífico Ocidental. Ambos são os principais agentes etiológicos da esquistossomose intestinal, uma doença crônica que atinge humanos. Estima-se que a divergência entre estas duas espécies ocorreu entre 70-148 milhões de anos.(11,12)

Estudos preliminares indicam que após essa divergência, a atividade de duas famílias de retrotransposons não-LTR, da linhagem RTE, denominados SR2 (13) e Perere-3 (14), apresentam recentemente grande expansão em seu número de cópias no organismo de *S. mansoni* sem paralelo em *S. japonicum*.(15)

A comparação de genomas nos permite examinar os processos de evolução molecular. Essas comparações, quando realizadas entre espécies muito próximas, permitem a identificação de mudanças e rearranjos genômicos que ocorreram recentemente no contexto evolucionário. Focando essas análises sobre sequências de genes ortólogos, é possível verificar a existência de elementos que estejam contribuindo para mudanças funcionais desses genes.(16)

Utilizando esse método de comparação, realizamos estudos mais detalhados sobre as inserções dos retrotransposons Perere-3 e SR2 em 2.752 pares de genes ortólogos de *S. mansoni* e *S. japonicum*. Utilizando códigos escritos em linguagem Perl (17), em conjunto com a biblioteca Bioperl (18), foram analisadas as posições das inserções desses elementos nas regiões intrônicas, intergênicas e presentes em mRNA.

Os resultados que serão apresentados sugerem que inserções desses elementos possuem características que podem promover alterações funcionais e estruturais dos genes, bem como fatores para atuar sobre elementos epigenéticos .

Este trabalho está organizado em diversos capítulos sendo que o capítulo 2 introduz conceitos e definições sobre os elementos de transposição. Também apresenta especificações sobre os genomas de *S. mansoni* e *S. japonicum* e sobre



as sequências dos retrotransposons Perere-3 e SR2. Finaliza contextualizando a análise de genomas em larga escala.

O capítulo 3 descreve a metodologia utilizada para organizar os dados iniciais, resultando na identificação dos genes ortólogos. Esses dados foram utilizados nas análises apresentadas nos próximos capítulos.

O capítulo 4 apresenta considerações gerais sobre as regiões intrônicas, a metodologia utilizada nas análises dessas regiões, bem como os resultados obtidos e as discussões desses resultados. Os capítulos 5 e 6 apresentam as mesmas informações para as análises das regiões intergênicas e presentes em mRNA, respectivamente.

No capítulo 7 são apresentadas as conclusões baseadas nos resultados descritos nos capítulos anteriores.



## Capítulo 2

### Contextualização

---



## 2 Contextualização

Este capítulo apresenta as definições dos elementos de transposição, a classificação desses elementos e as influências de suas inserções no genoma. Como o estudo se baseia em elementos do tipo não-LTR, são expostas as etapas que compõem o mecanismo de transposição desse tipo de elemento. Também são descritos os genomas dos organismos em estudo, bem como, dos elementos de transposição Perere-3 e SR2, e são apresentados conceitos e métodos correlacionados com a análise de genomas em larga escala.

### 2.1 Elementos de transposição

Transposons, ou elementos de transposição (*Transposable Elements – TE*), são segmentos de DNA ou RNA capazes de serem reproduzidos e inseridos no genoma da célula de origem. Essas inserções podem ser encontradas no genoma na forma de fragmentos de DNA imediatamente adjacentes uns aos outros, em uma orientação cabeça-cauda (*tandem*) ou, de forma dispersa, em posições que aparentam ser selecionadas aleatoriamente (*Long Interspersed Nuclear Elements – LINE*; *Short Interspersed Nuclear Elements – SINE*).<sup>(19)</sup>

Considerando o mecanismo utilizado para a transposição desses segmentos, os TEs podem ser divididos em duas classes principais. A classe I reúne os retrotransposons, os quais apresentam um RNA intermediário utilizado no processo de transposição. A classe II agrupa os transposons de DNA, que não utilizam um RNA intermediário e se caracterizam por apresentarem terminais com repetições invertidas (*Terminal Inverted Repeats – TIR*). Esses elementos terminais são reconhecidos pela proteína transposase, responsável em promover a remoção do elemento do sítio doador.<sup>(20)</sup>

Alguns elementos podem produzir uma cópia adicional enquanto outros elementos deixam o sítio doador para integrar-se em outras regiões do genoma. Para ambas as formas de inserção são geradas repetições diretas flanqueando as inserções, denominadas *Target Site Duplications* (TSDs). Essas repetições não são partes integrantes dos transposons e sim, geradas em decorrência da clivagem do genoma. Elas funcionam como “pegadas” das inserções realizadas pelos TEs, principalmente nos casos de deleção dessas inserções.(8)

A classificação também é realizada com base na organização estrutural que os elementos apresentam e na filogenia da proteína transcriptase reversa ou transposase, como ilustra a Figura 1.

Os retrotransposons podem ser divididos em 4 subclasses: *Non-Long Terminal Repeats* (não-LTR), LTR, Penelope e DIRS. Uma das diferenças entre os retrotransposons mais representados em genomas, não-LTR e LTR, é proveniente do modo como a integração do RNA ocorre no genoma. Os elementos LTR geram um cDNA (*complementary DNA*) a partir do RNA intermediário e durante esse processo são reconstituídas as repetições terminais do elemento. Os elementos não-LTR utilizam o processo denominado *Target Primed Reverse Transcription* (TPRT) que promove a transcrição reversa do RNA diretamente no sítio receptor do genoma. Muitas cópias dos elementos não-LTR apresentam a extremidade 5' truncadas, provavelmente decorrentes da baixa processividade da proteína transcriptase reversa. Esse processo produz inserções sem as sequências promotoras dos elementos o que, conseqüentemente, compromete sua propagação. (21)

Os transposons de DNA promovem sua inserção no genoma através da proteína transposase. Dentre os retrotransposons, os LTR utilizam a proteína integrase, os elementos da sub-classe DIR utilizam a tirosina recombinase e os não-LTR a proteína endonuclease. Os elementos da sub-classe Penelope também utilizam uma endonuclease, a qual apresenta características um pouco diferentes das utilizadas pelos elementos não-LTR e em decorrência disso, são classificados em uma outra ordem de elementos.

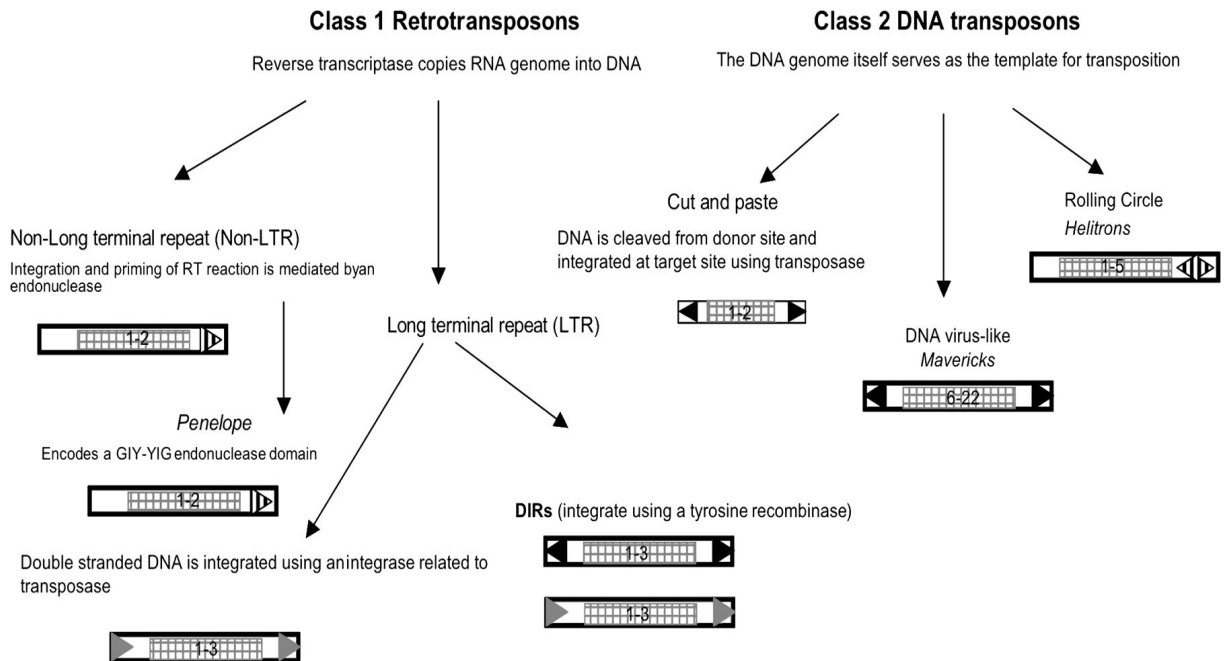


Figura 1-Classificação dos elementos de transposição em relação ao mecanismos de transposição e organização estrutural. Elementos da classe I utilizam um RNA intermediário enquanto elementos da classe II utilizam um segmento de DNA. As setas indicam a presença e orientação das regiões repetitivas que flanqueiam as inserções. Setas pretas identificam as TIRs e as cinzas as repetições diretas. As setas listradas indicam sequências repetitivas ou palíndromos no DNA. As caixas hachuradas em cinza indicam o número de ORFs e as proteínas codificadas pelos elementos autônomos. Fonte: PRITHAM (21)

A capacidade que os elementos têm de produzir as proteínas necessárias ao processo de transposição, classificam esses elementos em autônomos e não autônomos. No caso dos elementos não autônomos, os mesmos não apresentam capacidade de produzir as proteínas necessárias para o processo de transposição e para promoverem sua transposição, esses elementos utilizarão as proteínas produzidas pelos elementos autônomos.

Supõe-se que a evolução de uma família de TE no genoma seja similar ao ilustrado na Figura 2. Inicia-se com uma rápida fase de invasão, visando atingir um equilíbrio estável, o qual, segundo estudos teóricos, pode levar um longo período para ser atingido. Alterações demográficas, ambientais, distúrbios genômicos e características do hospedeiro, como a habilidade para eliminar sequências degeneradas, podem influenciar a obtenção desse equilíbrio.(22)

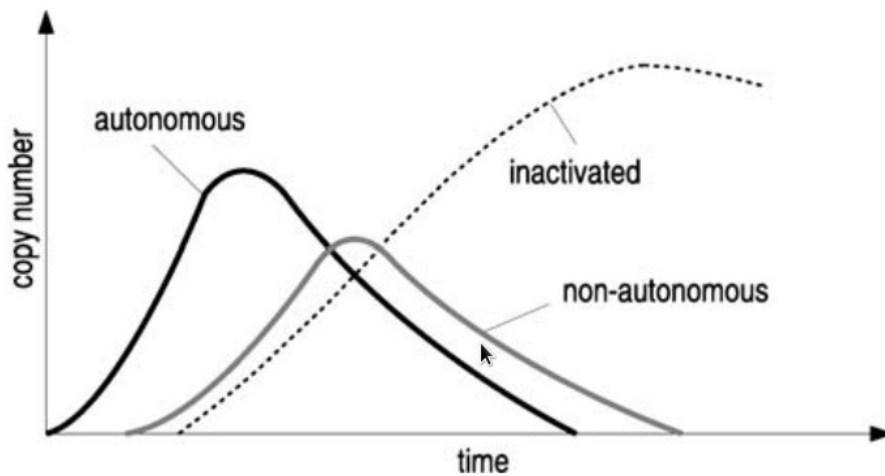


Figura 2-Hipótese sobre o processo de evolução de uma família de TE no genoma. Elementos autônomos (linha preta grossa), cópias mutantes não autônomas (linha cinza) e sequências derivadas de TE que permanecem no genoma como elementos inativos (linha pontilhada). Fonte:LE ROUZIC:CAPY (22)

As cópias desses elementos tendem a ser eliminadas progressivamente através dos processos de seleção natural e de mutações recorrentes. Cópias de elementos não autônomos podem surgir e supõe-se, que no final desse processo, permaneçam no genoma apenas sequências derivadas de TEs, as quais equivalem a elementos inativos e serão lentamente eliminadas.

A organização estrutural de elementos autônomos do tipo não-LTR é apresentada na Figura 3. Esses elementos apresentam região 5' UTR (*UnTranslated Region*) e 3' UTR, ORF 1 (*Open Reading Frame*) e ORF 2. A ORF 2 codifica a proteína endonuclease, responsável em clivar o genoma. Também codifica a transcriptase reversa, responsável em fazer a transcrição reversa do RNA transcrito para DNA complementar (cDNA). Para o elemento L1, foi observado que a ORF 1 codifica para a proteína chaperona\*.(6)

\* As proteínas da família chaperona auxiliam no enovelamento proteico e em caso de irregularidades, também encaminham a proteína que está sendo enovelada para ser degradada.





Figura 3-Organização estrutural de retrotransposon do tipo não-LTR. Os retângulos representam as proteínas que o elemento codifica. EN = endonuclease e RT = *reverse transcriptase*. Fonte: BEAUREGARD et al. (6)

O processo de transposição de elementos não-LTR têm início com a transcrição do RNA do elemento no núcleo da célula e sua passagem para o citoplasma, como ilustram os passos 1 e 2 da Figura 4.

No citoplasma, as proteínas codificadas pelo elemento são expressas e, em conjunto com o RNA do elemento formam um complexo ribonucleoproteico\* (RNP) o qual auxilia no controle da expressão gênica em nível de RNA. O complexo é transportado para dentro do núcleo onde o sítio de destino do elemento é identificado pela endonuclease, a qual promove a clivagem do DNA em uma das fitas do cromossomo. A extremidade exposta por essa clivagem é utilizada como *primer* para a transcriptase reversa iniciar a síntese do DNA complementar usando como molde o RNA do elemento.(6)

Ainda não é claro se o processo de clivagem da segunda fita é realizado pela endonuclease expressa na ORF 2 do elemento ou se é realizada por uma nuclease do hospedeiro. Também não é claro o processo no qual o RNA molde é removido.(6)

Finalizando o processo de transposição, diferentes complexos específicos para reparos, da própria maquinaria celular, são recrutados para completar o processo de integração do elemento no genoma.

Os elementos de transposição podem se tornar ativos mediante condições ambientais de *stress* físico ou químico pelo qual as células são atingidas. Por exemplo, os efeitos genotóxicos de venenos, radiações, temperaturas mais elevadas que um organismo está habitualmente acostumado, infecções virais ou metais

\* Complexo composto por um conjunto de proteínas associadas a uma molécula de RNA

pesados podem induzir a expressão de elementos LINE ou SINE.(7)

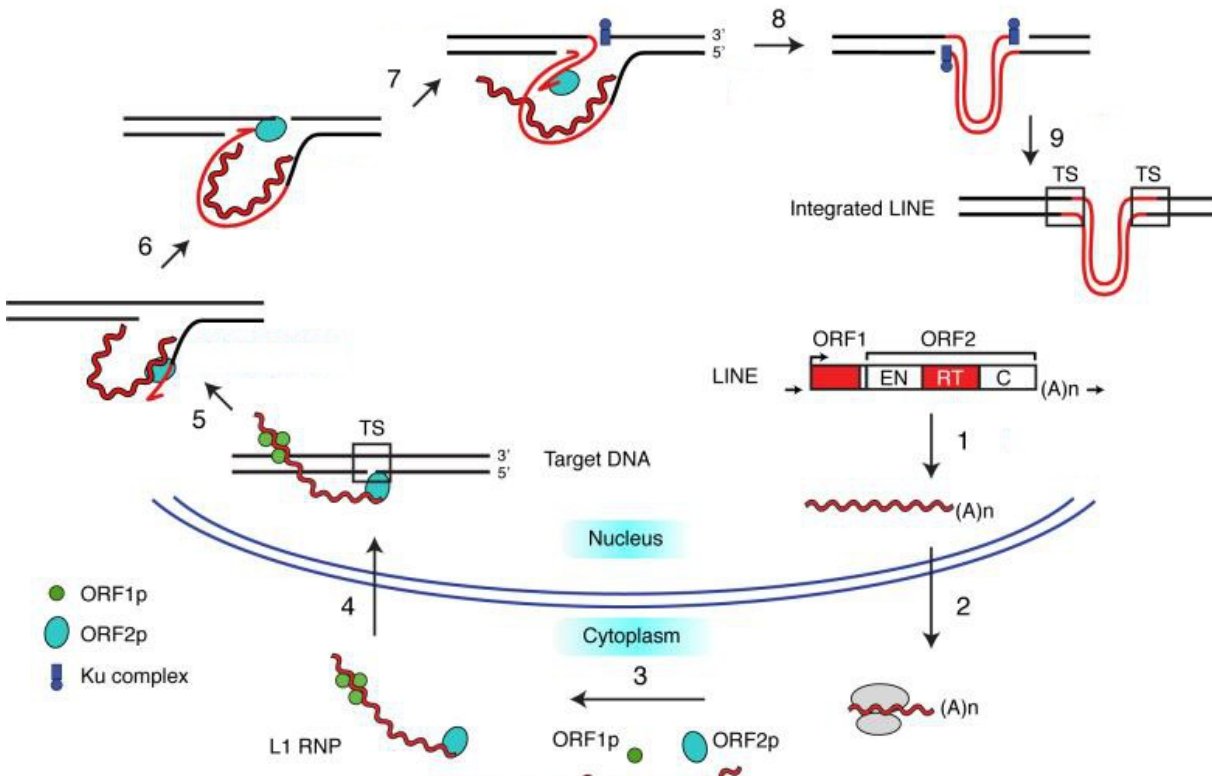


Figura 4-Etapas do processo de retrotransposição: (1) transcrição, (2) RNA é exportado para o citoplasma e ocorre a tradução, (3) formação do complexo RNP, (4) clivagem da primeira fita, (5) síntese do cDNA na primeira fita, (6) clivagem da segunda fita, (7) síntese do cDNA na segunda fita, (8) reparo do DNA clivado, (9) inserção do elemento no genoma finalizada. Fonte: BEAUREGARD et al. (6)

Alguns TEs integram-se preferencialmente em regiões ricas em genes, apesar das inserções nessas regiões tenderem a ter um aspecto mais deletério. Estudos apontam que o elemento Tn7, um transposon de DNA em bactéria (*E. coli*), insere-se em alta frequência em sítios específicos denominados attTn7 e com baixa frequência em sítios aparentemente aleatórios. As inserções nesses sítios específicos parecem não afetar o genoma da bactéria.(23,24,25)

Outros elementos de transposição podem integrar-se a heterocromatina ou nos finais dos telômeros, aumentando a probabilidade de inserções não deletérias. Elementos da classe Penelope possuem sequências complementares as sequências

de DNA teloméricas. Diversos transposons não possuem especificidade por um sítio de integração. Em estudos realizados por Levin e Moran, observou-se que o elemento L1s e os elementos não autônomos movidos por ele, estão dispersos de forma aleatória no genoma.(23)

Em uma revisão realizada recentemente, são descritos 32 eventos de domesticação nos organismos de humanos, ratos, aves, peixes, drosófilas, dentre outros. Para todos esses casos, os trechos inseridos pelos elementos evoluíram e passaram a desempenhar funções celulares.(4) A Figura 5 ilustra alguns desses eventos, identificando o gene do TE que deu origem ao processo de domesticação e o novo gene, o qual desempenha funções celulares.

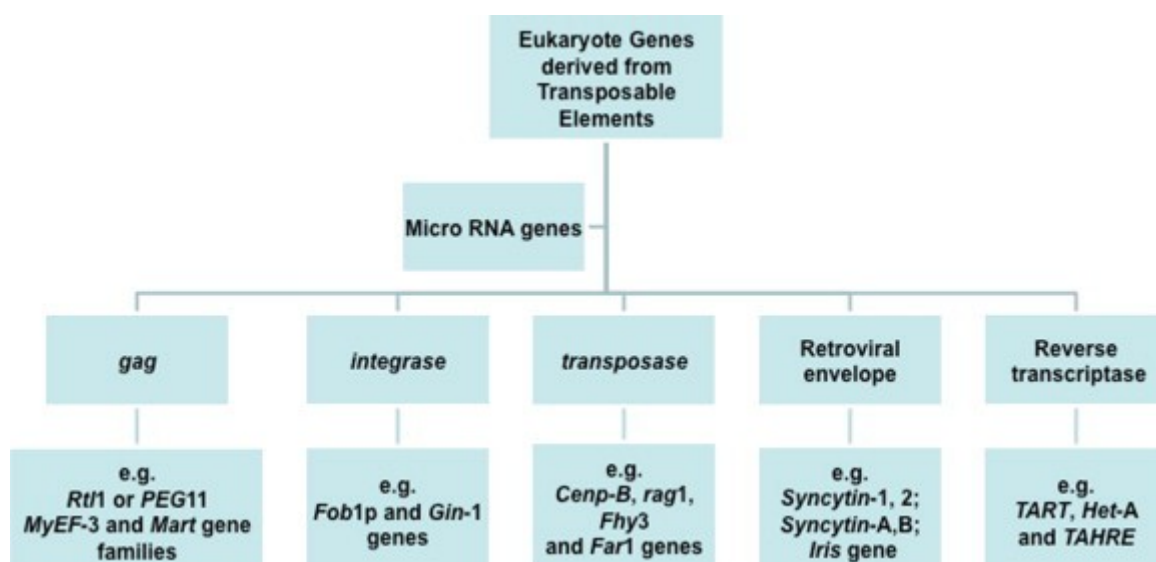


Figura 5-Exemplos de diferentes genes de TE (gag, integrase, transposase, retroviral envelope e transcriptase reversa) que foram domesticados pelo genoma do hospedeiro e evoluíram para novas funções celulares (ex: Rtl1, Fob1p,Cenp-B,Syncytin-1, TART). Fonte: ALZOHAIY et al. (4)

Os genes da família Mart, evoluíram a partir do gene GAG dos retrotransposons LTR *Sushi* (*Ty3 / Gypsy*), em peixes e anfíbios onde desempenham funções relacionadas com o desenvolvimento embrionário e com o controle de proliferação da célula e apoptose. O gene Gin-1, derivado da integrase dos elementos 412 e Mdg1, em humanos, ratos e vacas é expresso durante a embriogênese e em vários tecidos humanos adultos e tumor. Também são relatados

exemplos de telomerasas que apresentam relações filogenéticas com a transcriptase reversa de retrotransposons não-LTR.

As inserções dos elementos de transposição podem gerar alterações na transcrição do gene e também no processamento e na tradução do transcrito (26), como ilustra a Figura 6.

Durante a fase de transcrição as inserções podem influenciar de diversas formas:

a) O elemento pode estar inserindo um trecho que corresponde ao promotor de sua sequência, na região responsável pela transcrição do gene. A sequência inserida pode atuar como um elemento promotor durante a transcrição do gene;

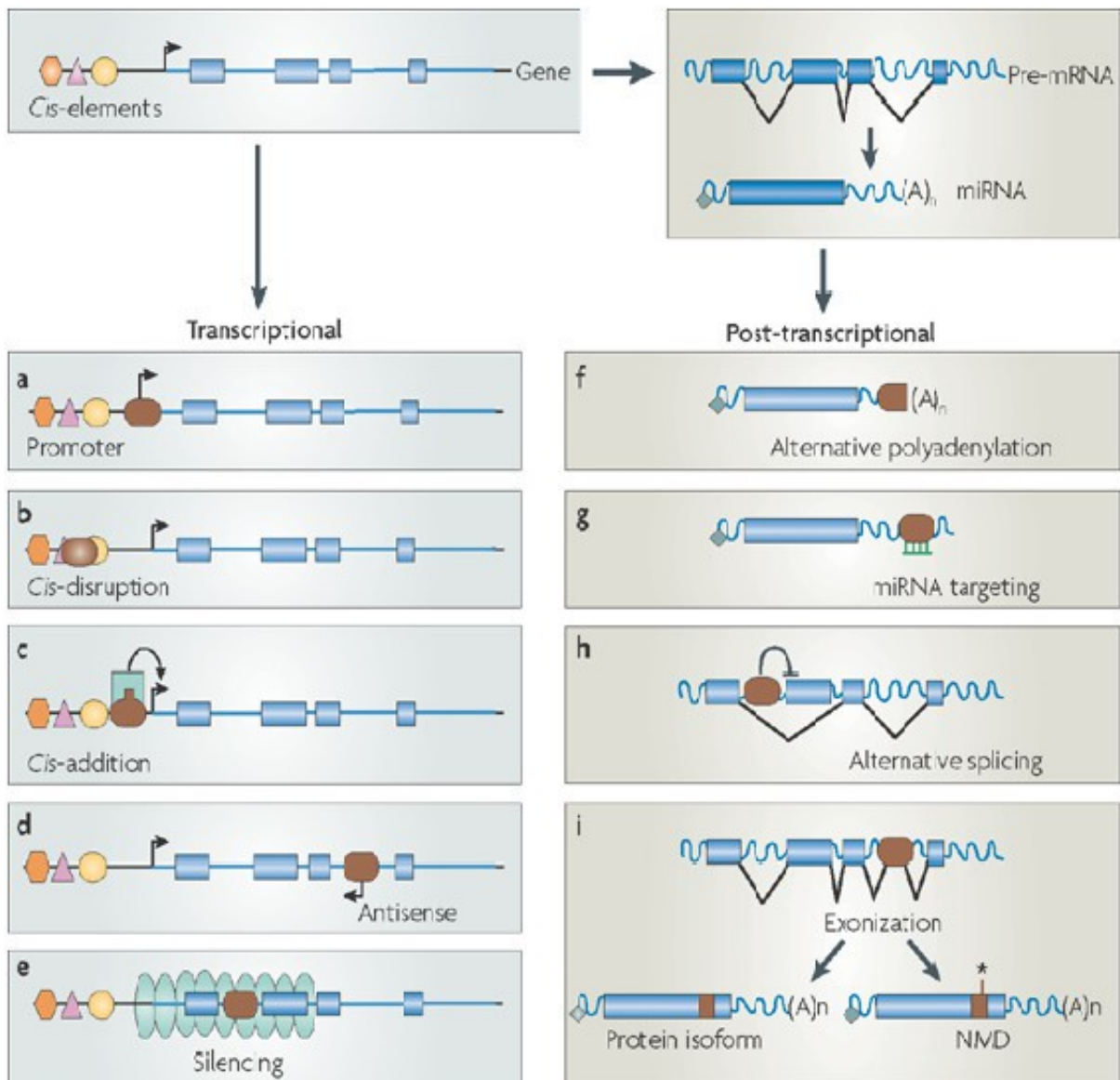
b) Os elementos reguladores do gene, os quais podem ser repressores ou ativadores da transcrição, podem ser interrompidos pela inserção do elemento e com isso afetar a regulação da expressão do gene;

c) O elemento pode inserir trechos que contenham sítios para fatores de transcrição, ativando ou não a transcrição do gene;

d) O elemento pode inserir uma sequência promotora mas no sentido contrário ao da transcrição do gene. Nesse caso, a expressão do gene pode ser afetada de duas formas: em um primeiro momento, durante a transcrição do gene pois, se ambas as extremidades promotoras iniciarem o processo de transcrição, as duas maquinarias utilizadas nessa transcrição, uma estando em um sentido e a outra em outro sentido, podem se encontrar e a transcrição ser interrompida. De outra forma, a interferência pode ser em decorrência do promotor do gene expressar o gene em um sentido, e o promotor do elemento expressar o gene em sentido oposto. Esses RNAs podem se unir formando um RNA dupla fita (*double-stranded* RNA – dsRNA) e dessa forma promover a degeneração do RNA,

e) Também pode ser que a maquinaria celular identifique a ocorrência da inserção e promova o silenciamento do trecho inserido. Esse procedimento também pode promover o silenciamento do gene.

Após a transcrição, durante o processamento do pré-mRNA para mRNA, essas inserções podem alterar a estabilidade do mRNA e também o produto proteico gerado a partir desse mRNA:



Nature Reviews | Genetics

Figura 6: Alterações que podem ocorrer em decorrência da inserção de elementos de transposição na fase de transcrição do gene (exemplos de a até e) e também na fase pós-transcrição (exemplos de f até i). Os retângulos em azul representam os éxons, as linhas entre os éxons representam os íntrons, os triângulos, círculos e hexágonos os elementos reguladores ou promotores do gene e as marcações em cor marrom as inserções do elemento de transposição. Fonte: FESCHOTTE (26)

f) Inserções na posição 3' UTR podem incluir sítios de poliadenilação\* os

\* A poliadenilação é o processo de ligação de caudas poli(A) a uma molécula de RNA mensageiro. Esta cauda terminal protege a molécula de RNA das exonucleases e é importante para a terminação da transcrição, para a exportação do RNA a partir do núcleo e para a tradução.

quais podem comprometer o tempo de vida útil desse mRNA;

g) Ainda na posição 3' UTR, os elementos podem inserir sítio para ligação de um micro RNA (miRNA<sup>\*\*</sup>) e com isso promover a degeneração do RNA;

h) As inserções nos íntrons podem produzir alterações nos padrões de *splicing* produzindo a exclusão do éxon o que possivelmente produzirá um outro tipo de proteína,

i) Também podem ser inseridas sequências auxiliares de *splicing* que facilitariam a incorporação do íntron como um éxon alternativo (exonização), gerando assim uma isoforma da proteína. A inserção de um *stop codon* nesse processo de exonização pode produzir uma proteína truncada prematuramente.

Embora alguns TEs apresentem mecanismos que minimizam os danos que as inserções podem causar ao genoma, os organismos hospedeiros apresentam mecanismos de restrição para limitar a atividade desses elementos. Dentre esses mecanismos estão a metilação dos segmentos inseridos e como consequência o silenciamento do elemento. A expressão de *small RNA* (sRNA) pode promover interrupções nas sequências dos transposons que foram transcritas. Enzimas envolvidas no processo de metabolismo dos nucleotídeos e/ou nos mecanismos de reparação do DNA também podem atuar sobre os elementos de transposição. Para elementos LTR foi observado que a proteína APOBEC3<sup>\*\*\*</sup> promove a degradação da primeira fita de cDNA gerada pelos mecanismos de transposição.(23)

## 2.2 O *Schistosoma mansoni* e o *Schistosoma japonicum*

Apesar de ambos os parasitas *S. mansoni* e *S. japonicum* infectarem humanos, são descritas características diferentes entre essas espécies. Supõe-se que o organismo ancestral teve origem no sudeste da Ásia, onde se dissemina

\*\* Os miRNAs direcionam a clivagem e reprimem a tradução de mRNAs com os quais têm complementaridade, ficando ligados a eles. Também podem degradar esses mRNA retirando a cauda poli-A, atuando desta forma na repressão da expressão gênica.

\*\*\* As proteínas da família APOBEC são conhecidas por terem a capacidade de inativar um retrovírus, gerando um elemento não infeccioso.

através da espécie *S. japonicum*. Provavelmente, em decorrência do processo de migração de mamíferos para o continente africano, ocorreu o estabelecimento do gênero *Schistosoma* neste território. Nesse novo ambiente, ocorreram processos de diversificação e especiação que permitiram o surgimento da espécie *S. mansoni* entre outras espécies. Acreditasse que a espécie *S. mansoni* por sua vez, migrou para a América Central e do Sul em decorrência da movimentação de escravos infectados provenientes da África, como ilustra a Figura 7.(12)



Figura 7-Hipótese sobre a origem Asiática e dispersão do *Schistosoma*. O *schistosoma* ancestral asiático irradia pela Ásia através da espécie *S. japonicum* (1). Entretanto, o *schistosoma* ancestral também se dispersa para África e Índia, dando origem a espécie *S. mansoni* (2). Uma divisão dessa espécie agrupa a *S. haematobium*, a qual também se dispersa pela África (3), e a espécie *S. indicum*, a qual reinvasa o continente da Índia (4). *S. mansoni* dispersa para América do Sul (5). Fonte: LOCKYER et al. (12)

Para compreender algumas diferenças entre essas espécies, é necessário analisar um pouco do ciclo do parasita, o qual é ilustrado na Figura 8. Esse ciclo tem início com o depósito dos ovos na água. Esses ovos se transformam em miracídeos, os quais infectam os caramujos e se desenvolvem na forma de esporócitos. Esses esporócitos são expelidos na água em forma de cercárias, as quais penetram a pele humana. Durante esse processo de penetração, a cercária perde a cauda, se

transforma em esquistossômulos e migra através do organismo utilizando a corrente sanguínea. No fígado o parasita se desenvolve em macho e fêmea, ou seja, em vermes adultos. Os vermes de ambos os sexos se emparelham e migram para as veias do intestino (*S. mansoni* e *S. japonicum*) ou da bexiga (*S. haematobium*), dependendo da espécie.(27)

## ESQUISTOSSOMOSE

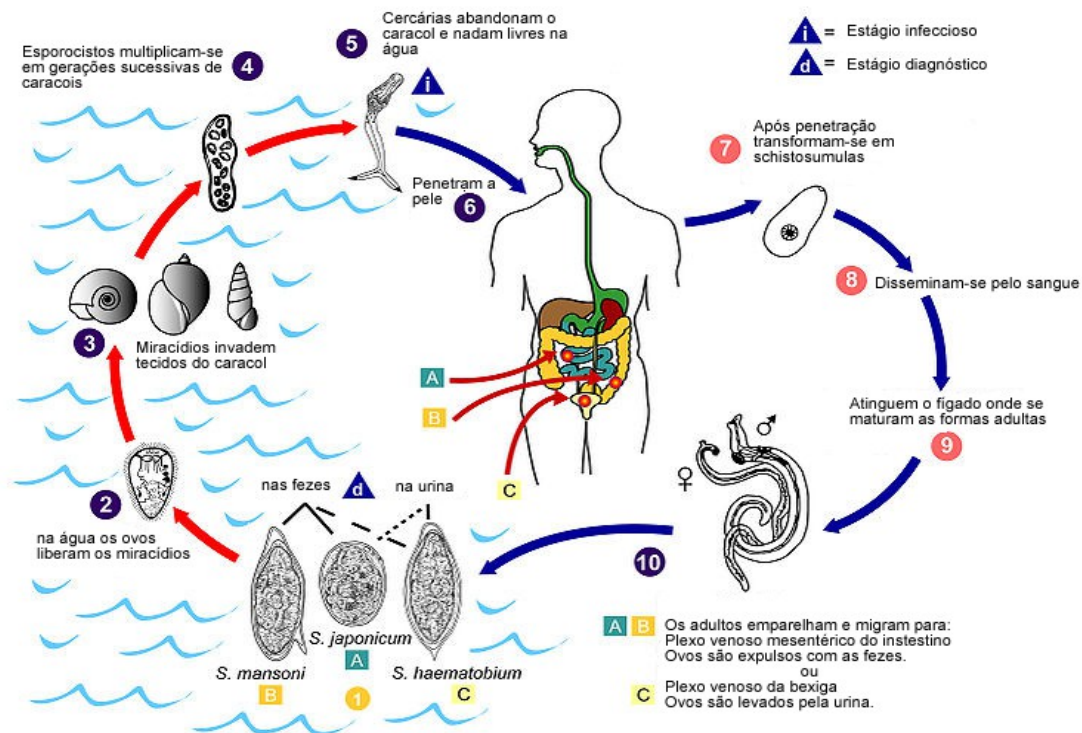


Figura 8-Ciclo da esquistossomose. Os passos de 1 a 5 retratam o processo de desenvolvimento do parasita na água, onde há o hospedeiro intermediário (caramujo), e os passos de 6 a 10 no organismo humano. Fonte: CDC Home (27)

Dentre as diferenças entre as espécies *S. mansoni* e *S. japonicum* pode-se observar o formato dos ovos, como ilustrado na Figura 8. Os ovos de *S. mansoni* apresentam um espinho lateral enquanto que em *S. japonicum* ocorre a ausência do espinho. O caramujo, que é o hospedeiro intermediário, varia entre as espécies de *Schistosoma*. Para manter seu ciclo de vida, o parasita *S. mansoni* necessita do caramujo da família *Planorbídeo*, do gênero *Biomphalaria*. *S. japonicum* utiliza como hospedeiro intermediário o caramujo do gênero *Oncomelania* da família



*Pomatiopsidae.*(12,28)

Embora os vermes adultos de *S. mansoni* e *S. japonicum* migrem ambos para o intestino, *S. japonicum* é encontrado com maior frequência na veia mesentérica superior que drena o intestino grosso. *S. mansoni* é mais frequentemente encontrado em veia similar mas que drena o intestino delgado. Os ovos são depositados no início dos vasos sanguíneos e migram para o intestino onde são eliminados nas fezes.(27)

Nos estudos realizados por Rheinberg e colaboradores (29), ficou demonstrada a diferença percentual de infecção das cercárias e a grande variação no modo com que essas cercárias se desenvolviam em vermes adulto. Também foram observadas as divergências nos padrões de tempo para chegada dos esquitossômulos ao órgão infectado (no caso desse estudo, pulmão de ratos) e o período de residência do verme nesse órgão. *S. japonicum* apresentou maiores índices de infecção e um padrão de migração/desenvolvimento mais simples e mais rápido do que *S. mansoni*.

Em 2009 foram liberados os primeiros rascunhos dos genomas dessas duas espécies. Baseando-se nas sequências desses rascunhos, os genes foram preditos de forma computacional. Também foi realizado um estudo mais superficial dos elementos repetitivos que esses genomas continham. Em 2012, o rascunho do genoma de *S. mansoni* foi aprimorado e 81% das bases organizadas em cromossomos.(30) Para *S. japonicum*, conjuntos fragmentários de sequências sem designação de cromossomos permanecem até hoje.

O genoma de *S. japonicum* é composto por 8 pares de cromossomos sendo 7 pares de autossomos e um par de cromossomo sexual. O processo de sequenciamento utilizou vermes adultos e ovos obtidos a partir de ratos infectados. Foram identificadas 397 Mb cobrindo aproximadamente 90% do genoma com um total de 13.469 genes codificando para proteínas. *S. japonicum* têm um genoma grande e uma baixa densidade de genes (34 genes por Mb) em comparação com outros invertebrados. Foram identificadas 657 famílias/elementos repetitivos constituindo 159 Mb relativos a 40.1% do genoma sendo 12.6% correspondente a elementos não-LTR. Apresentou maior frequência de genes ortólogos com

vertebrados como *H. sapiens* (4.324 pares) do que com ecdysozoas\*, como *C. Elegans* (3.292 pares), mesmo este último sendo filogeneticamente adjacente. O conteúdo GC nas regiões codificantes é de aproximadamente 36% e das regiões não codificantes de 33.8%. As regiões intergênicas apresentam conteúdo GC de 34.7%.(10)

O genoma de *S. mansoni* foi obtido a partir de cercárias que foram expelidas pelo caramujo *Biomphalaria glabrata* em Porto Rico. Apresenta 363 Mb com cerca de 11.809 genes codificando 13.197 transcritos com uma distribuição de íntrons não usual. Os genes apresentam tamanho médio de 4.7 Kb com íntrons grandes com aproximadamente 1.692 pb e com éxons com comprimento médio de 217 pb. Os genes apresentam em média 7 éxons. Nas regiões codificantes o conteúdo GC é de 36.3% e nas regiões não codificantes é de 35.2%. As regiões metiladas apresentam 37.3% (31) sendo que a média do conteúdo GC desse genoma é de 35.3%.(9) Aproximadamente 40% do genoma corresponde a 72 famílias de elementos repetitivos sendo, 15% equivalentes a elementos não-LTR e 5% a elementos LTR.

### 2.3 Os elementos Perere-3 e SR2 de *S. mansoni*.

Venâncio e colaboradores (15) realizaram estudos mais detalhados sobre a inserção de elementos de transposição no genoma de *S. mansoni* e de *S. japonicum*. Os resultados demonstram que para ambas as espécies, predominam as inserções dos elementos não-LTR, como ilustra a Figura 9.

Analisando 72 tipos de retrotransposons não-LTR, para *S. mansoni* foi observado que os elementos da família RTE estavam em maior frequência no genoma (12.2%), em especial o elemento SR2 (com 2.94%) e o Perere-3/SR3\*\* (com 4,84%). Em *S. japonicum*, dentre os 64 tipos de retrotransposons, a mesma família

---

\* Ecdysozoa é um clado de animais protostômios (no desenvolvimento embrionário a boca se forma antes que o ânus) que reúne os artrópodes, nemátodes e outros sete filos.

\*\* Esses elementos foram agrupados pois apresentam domínios de transcriptase reversa muito semelhantes.

RTE observada em *S. mansoni*, também apresentava maior concentração (7%). Os elementos Sj2 (com 0.10%) e o Sj5 (3.38%), elementos esses equivalentes ao SR2 e Perere-3/SR3 de *S. mansoni*, respectivamente.

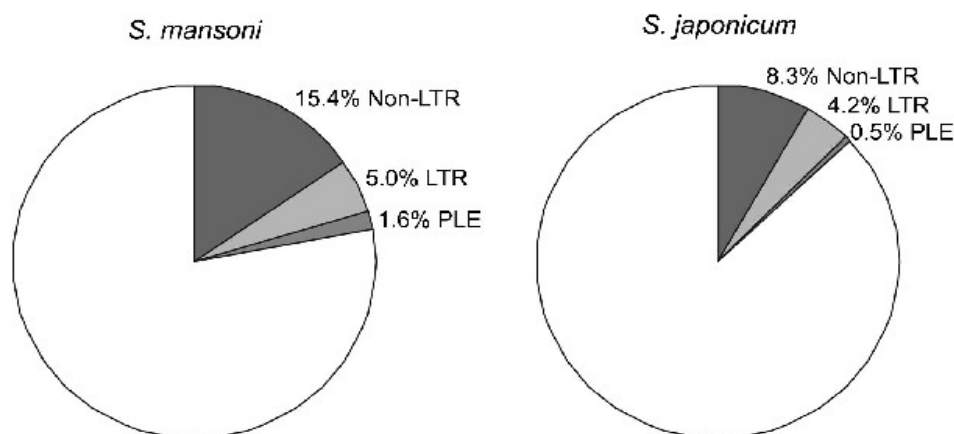


Figura 9-Representação das diferentes classes de retrotransposons não-LTR, LTR e Penelope-like (PLE), analisadas nos genomas de *S. mansoni* e *S. japonicum* Fonte: VENANCIO et al. (15)

Para as inserções dos elementos mais frequentes, realizou-se análises medindo a distância par a par das sequências inseridas, considerando apenas o trecho que correspondia ao domínio da transcriptase reversa, como ilustra a Figura 10.

O elemento SR2 em *S. mansoni* apresentou uma quantidade maior de sequências inseridas com uma pequena distância entre si, ou seja, a divergência entre essas sequências é pequena. O mesmo resultado não foi observado para o elemento Sj2, equivalente em *S. japonicum*. As sequências inseridas por esse elemento apresentam um maior grau de divergência entre si do que o grau observado nas inserções do elemento SR2 em *S. mansoni*. Supostamente, as inserções de Sj2 em *S. japonicum* não são tão recentes quando comparadas com as inserções de SR2 em *S. mansoni*.

Para o elemento Perere-3/SR3, os resultados apontam um padrão diferente. A maior parte das sequências inseridas apresentam um grau de divergência maior entre si, em ambos os organismos. Esses dados permitem sugerir que a

transposição dessa família de retrotransposons provavelmente foi muito ativa em um ancestral comum a essas espécies. Com relação a quantidade de sequências inseridas que apresentam pequena divergência entre elas, *S. mansoni* apresenta quantidades mais significativas do que *S. japonicum*, salientando novamente que *S. mansoni* contém inserções mais recentes do que *S. japonicum*.

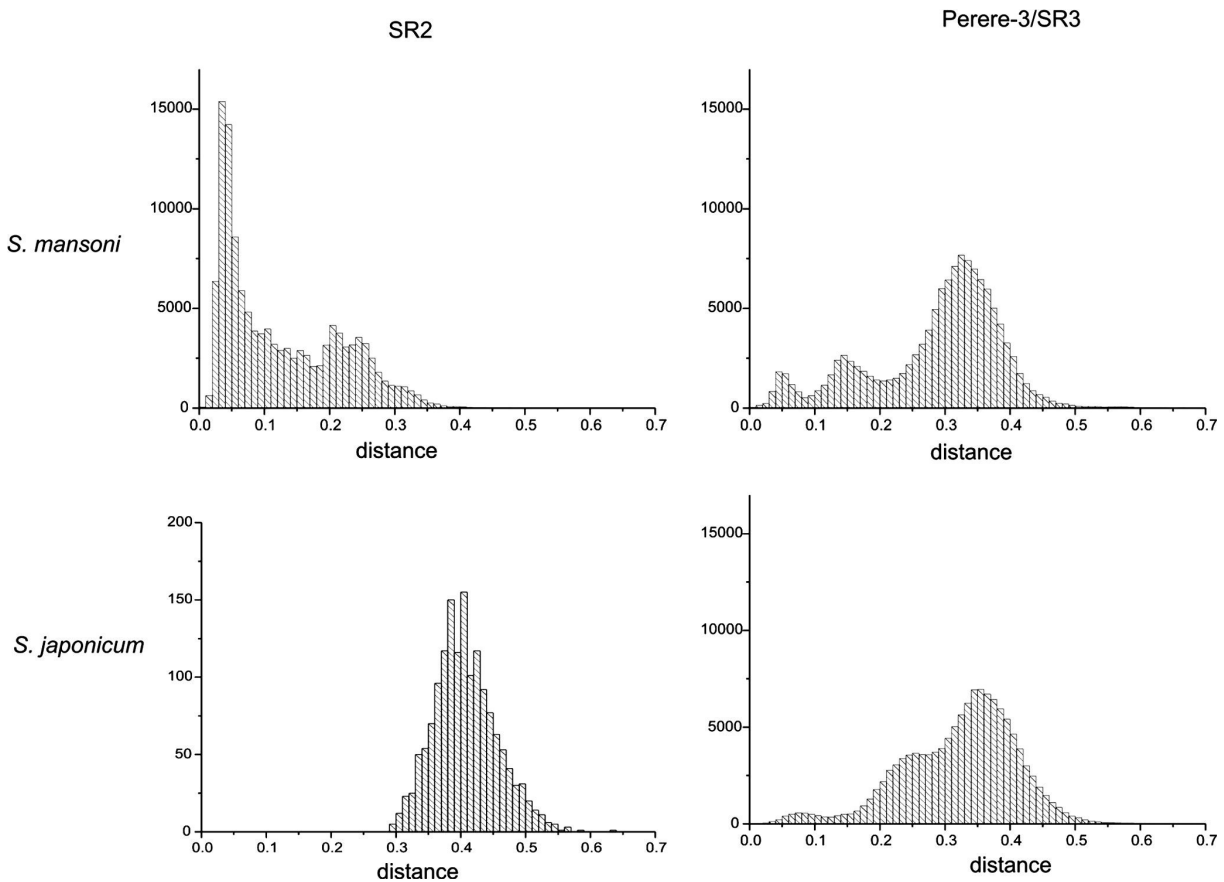


Figura 10-Distribuição da distância (par a par) entre os pares de bases do domínio da transcriptase reversa de membros de uma mesma família. Elementos da família SR2 e Perere-3 de *S. mansoni* e os elementos equivalentes em *S. japonicum* (RTE-Sj2 e RTE-Sj5, respectivamente). Fonte: VENANCIO et al.(15)

Para detalhar melhor as inserções dos elementos Perere-3/SR3, foi realizada uma análise filogenética das sequências inseridas, como ilustra a Figura 11.

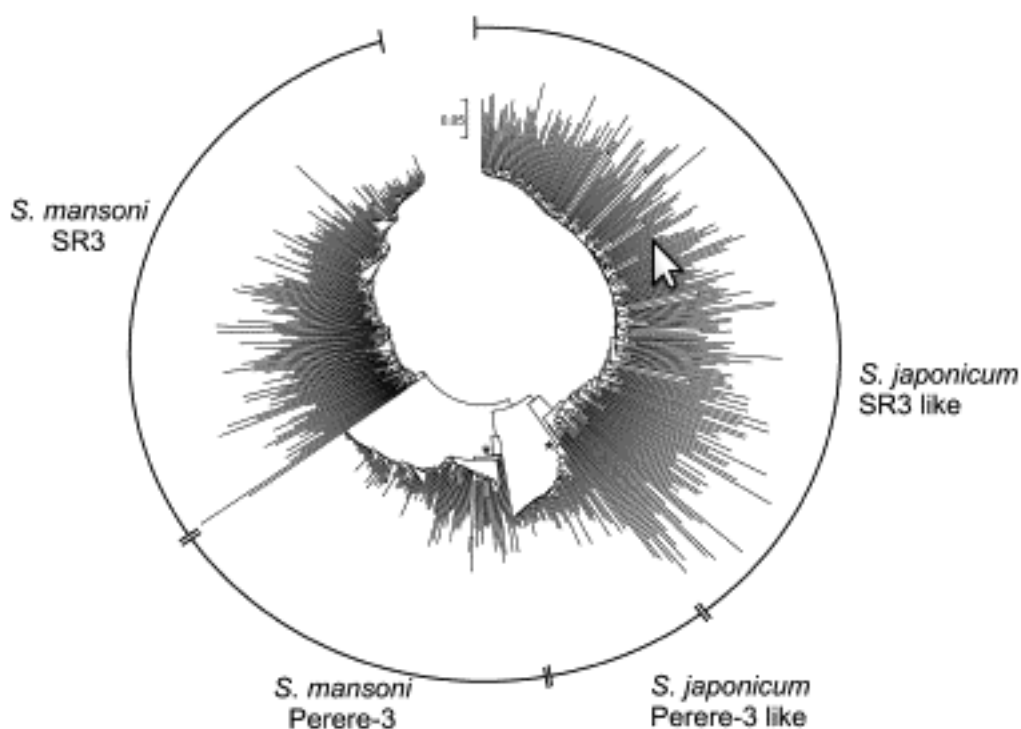


Figura 11-Árvore filogenética das sequências de nucleotídeos equivalentes ao trecho do domínio RT inseridos pelos elementos Perere-3/SR3 em *S. mansoni* e o pelo elemento correspondente em *S. japonicum*. O \* representa a posição inicial do grupo monofilético de elementos, os quais podem ter sido os agentes da recente expansão. Fonte: VENANCIO et al. (15)

O elemento SR3 em *S. mansoni* e o elemento equivalente em *S. japonicum*, apresentam uma maior quantidade de sequências representadas através de ramos mais longos, demonstrando dessa forma que a divergência desses trechos inseridos é maior quando comparados com os trechos inseridos pelos elementos Perere-3 e o elemento equivalente em *S. japonicum*. Desta maneira, é aparente que os elementos Perere-3 são mais recentes nos genomas destes parasitas quando comparados com os elementos SR3.

Entender a dinâmica dos retrotransposons SR2 e Perere-3 é importante para verificarmos a real contribuição desses elementos na arquitetura do genoma de *Schistosoma*.

O elemento SR2 apresenta aproximadamente 3.9 kb incluindo uma 5' UTR com 450 bp, 2 ORFs, que são separadas por 6 bp, de 78 e 1018 aminoácidos. A

extremidade 3' UTR apresenta 155-173 bp e terminais variáveis com *motifs* ((C/T)<sub>1-4</sub>). A ORF2 apresenta o domínio da *endonuclease* (ENDO), uma região com 210 aminoácido e o domínio da *reverse transcriptase* (RT) com 272 aminoácidos, como ilustra a Figura 12. Vários elementos SR2 apresentaram um *stop codon* que define o término dessa ORF.



Figura 12-Estrutura do retrotransposon SR2. Fonte: DREW et al. (13)

Nos estudos realizados por Drew e colaboradores (13), em 6 inserções foram encontrados TSD (*Target Site Duplication*) flanqueando as inserções com aproximadamente 8 a 12 bases.

O elemento Perere-3 apresenta uma região bem conservada com 3196 bp, e uma região de conteúdo e comprimento variável (de 40 a 450 bp) além de uma cauda poli A conservada. Sendo um elemento da classe RTE, assim como o SR2, apresenta na ORF2 domínios de ENDO e RT, como ilustra a Figura 13. Entre as bases 3194 a 3196 foi identificado um *stop codon* conservado.(14)



Figura 13-Estrutura do retrotransposon Perere-3. Fonte: DEMARCO et al. (14)

## 2.4 Análise de elementos de transposição em larga escala

Em decorrência do crescente número de sequências de genomas completos que estão sendo geradas, a área da genômica comparativa visa o aprimoramento dos métodos utilizados para realizar o alinhamento de genomas inteiros, predição de genes e predição de regiões regulatórias.(32)

A comparação de genomas fornece à Biologia informações para compreender de forma mais profunda os organismos e seus processos de evolução. Esse tipo de comparação baseia-se no fato de que esses dois genomas apresentam um ancestral em comum então, suas sequências representam as sequências do ancestral e a ação do processo de evolução.(32)

Nesse processo de evolução, as forças mutacionais geram modificações aleatórias no genoma dos organismos, as quais podem sofrer pressões seletivas negativas ou positivas. Comparar genomas evolucionariamente relacionados permite obter informações sobre as diferentes pressões atuando na estrutura dos genes desses organismos. Correlacionada a comparação de genomas e aos processos de evolução, a influência dos elementos de transposição constitui uma nova e crescente área para pesquisas.(23)

Os genomas apresentam uma série de TEs que não codificam proteínas e identificar essas sequências no genoma caracteriza um desafio que envolve a descoberta dos TEs, a classificação e categorização, o mascaramento e as anotações, as análises da dinâmica das populações e finalmente a geração do banco de dados de TEs.(33)

O estudo da influência dos TEs no genoma dos organismos envolve uma sequência de análises específicas. Os dados observados em um primeiro momento passam por consecutivas análises até que os resultados sejam suficientes para uma argumentação ou formulação de uma hipótese sobre as possíveis influências observadas. Dentre os recursos da bioinformática que podem ser utilizados na elaboração dessas análises, encontram-se disponíveis na *Internet* muitas ferramentas direcionadas às áreas mais genéricas de análises como alinhamento de sequências, filogenia, predições de determinadas regiões, dentre outras. Análises que necessitam estudar detalhes específicos de determinadas regiões,

principalmente quando envolvem dados em larga escala, necessitam de ferramentas específicas, as quais são implementadas de acordo com a necessidade dessas análises.



# Capítulo 3

## Organização dos dados iniciais

---



### 3 Organização dos dados iniciais

Considerando como dados de entrada as sequências dos genomas e dos genes preditos para *S. mansoni* e *S. japonicum*, e como base para o desenvolvimento da lógica de programação os conceitos e definições de biologia molecular, as análises foram implementadas utilizando procedimentos computacionais definidos como *pipelines*. Esses procedimentos se caracterizam por apresentarem uma cadeia de elementos de processamento, dispostos de modo que a saída de cada elemento é a entrada do próximo, como ilustra a Figura 14.

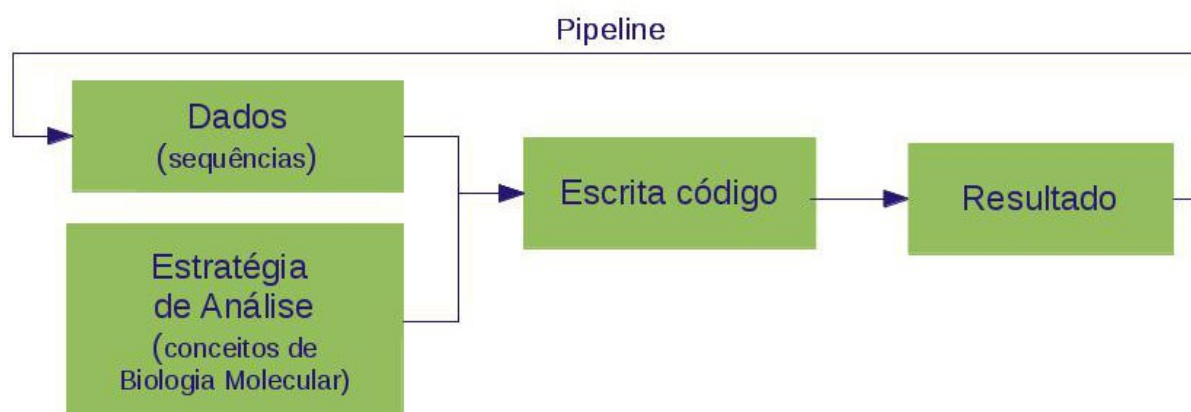


Figura 14-Metodologia utilizada para implementar as análises para verificar a possível influência das inserções dos elementos de transposição Perere-3 e SR2 no genoma de *S. mansoni*. Fonte: Elaborada pela autora.

*Pipelines* bem organizados e estruturados também auxiliam de forma significativa quando os dados de entrada, utilizados nas análises, são aprimorados e o reprocessamento dessas análises é necessário. Em 2012, após ser disponibilizado o genoma de *S. mansoni* organizado em cromossomos, foi necessário realizar o reprocessamento de todas as análises implementadas neste trabalho a partir de 2009.

A estruturação dos *pipelines* neste trabalho iniciou-se com a escrita dos códigos, evitando a utilização de constantes no corpo do programa. Todos os dados

como nomes de arquivos, siglas, coordenadas, dentre outros, foram informados utilizando passagem de parâmetros. Essa passagem de parâmetros foi realizada por um *script shell* que além dessa função, também recebeu em sua nomenclatura um número identificando a ordem de execução do código dentro do *pipeline*. Dessa forma, os ajustes dos dados de entradas são realizados nos *scripts shell* e o *pipeline* já está preparado para ser executado manual. Para automatizar essa execução, a criação de um outro *script shell* com chamada para todos os *scripts shell* numerados é necessária. A Figura 15 ilustra uma exemplificação dessa estruturação.

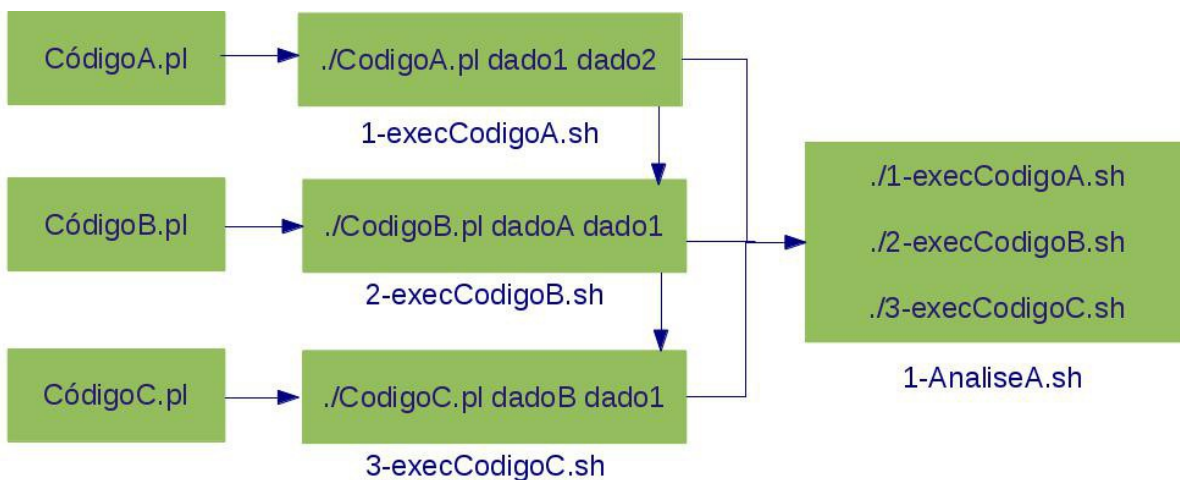


Figura 15-Exemplificação da estruturação de um *pipeline* onde os retângulos da primeira coluna representam os códigos escritos em linguagem Perl, os da segunda coluna representam os *scripts shell* que realizam as chamadas dos códigos Perl e a passagem de dados para esses códigos, e o retângulo da terceira coluna representa o *script shell* que realiza a execução automática de todo o *pipeline*. Fonte: Elaborada pela autora.

Os *pipelines* implementados para organização dos dados iniciais e para as análises das regiões intrônicas, intergênicas e presentes em mRNA, encontram-se nos Anexos I, II, III e IV, respectivamente.

A primeira etapa do estudo consistiu na organização dos dados iniciais e na identificação dos genes ortólogos de *S. mansoni* e *S. japonicum*.

O estudo foi focado no conjunto dos genes ortólogos pois em 2009, quando o estudo teve início, os genomas dessas espécies estavam na forma de rascunhos. Selecionar um conjunto de dados mais consistente permite obter resultados mais confiáveis e elaborar hipóteses mais plausíveis. Além disso, a análise restrita a este

conjunto de genes ortólogos facilita comparações de diversos aspectos da estrutura gênica destes dois organismos, sem a introdução de vieses, pois um conjunto equivalente de genes está sob análise. A ortologia auxilia na definição dessa característica pois estabelece que as sequências têm um único e mesmo ancestral comum.(34)

### 3.1 Metodologia

As coordenadas dos éxons dos genes preditos para *S. mansoni* foram obtidas a partir dos arquivos em formato *Genbank* (.gff) disponíveis no *site* do ENA (*European Nucleotide Archive*). Esses arquivos continham as sequências parciais dos 7 autossomos e do cromossomo sexual W, bem como sequências de *supercontigs* que não foram incorporadas na estrutura principal dos cromossomos.

Devido a falta de um arquivo do tipo .gff contendo as coordenadas dos genes preditos para *S. japonicum* no *site* contendo as informações do genoma, foi realizado o alinhamento das sequências dos genes preditos contra o genoma do organismo utilizando o *software* BlastN.(35) A partir deste alinhamento, os dados foram organizados de forma sequencial, com uma estrutura de dados contendo:

- identificação do gene;
- identificação do cromossomo/*scaffold*;
- fita onde o elemento se encontra (senso e antisenso);
- coordenadas de início e fim do elemento (éxon1, éxon2, etc),
- coordenadas de início e fim do elemento no cromossomo/*scaffold*.

Nessa organização sequencial dos dados do cromossomo/*scaffold*, as regiões intergênicas foram classificadas em 4 tipos, representando regiões entre genes localizados na mesma fita (5'-3',3'-5') e genes em fitas diferentes (3'-3',5'-5'), como ilustra a Figura 16.

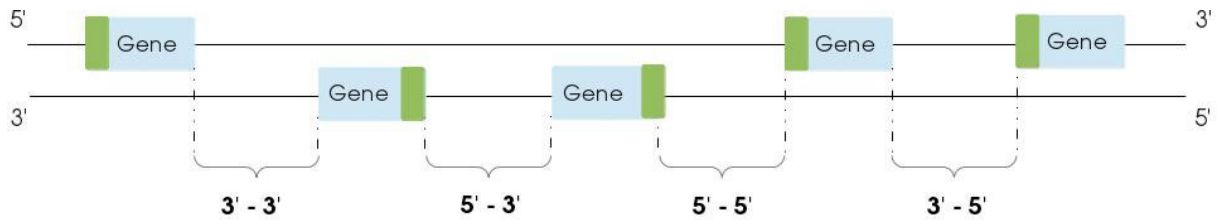


Figura 16-Nomenclatura utilizada para definir as regiões intergênicas identificando regiões com genes localizados na mesma fita ou em fitas diferentes. Os retângulos azuis representam os genes e a extremidade em verde as regiões promotoras desses genes. Fonte: Elaborada pela autora.

### 3.1.1 Definição dos genes ortólogos

Foram considerados pares de genes ortólogos aqueles vindos dos dois organismos que possuíam, de forma recíproca, o alinhamento com maior escore com o seu par, quando todos os produtos proteicos codificados por um organismo eram comparados contra o banco de proteínas do outro organismo.(36) Para realizar esse alinhamento foi utilizado o *software* BlastP considerando como alinhamentos relevantes os resultados com *evaluate* inferior a  $10^{-3}$ .

Como a proposta era definir os éxons/íntrons equivalentes entre esses pares de genes, foram selecionados apenas os pares que apresentaram um único trecho alinhado de forma contínua, com média de similaridade superior a 85%.

A estrutura de dados definida por essa etapa contém:

- identificação da proteína de *S. mansoni* (Sm);
- identificação da proteína de *S. japonicum* (Sj);
- identidade entre as proteínas;
- coordenadas do gene correspondente a proteína de Sm cujas bases são semelhantes em Sj,
- coordenadas do gene correspondente a proteína de Sj cujas bases são semelhantes em Sm.

As coordenadas dos éxons desses genes foram comparadas com as coordenadas resultantes do alinhamento entre as proteínas, e o conjunto éxons/íntrons equivalentes foi definido.

Posteriormente, esses pares de genes (ORTO) foram divididos em 2 grupos. Um grupo contendo genes ortólogos que apresentaram a mesma quantidade de éxons resultantes do alinhamento (ORTO-ql) e outro, com quantidade diferente (ORTO-qD). Os dados obtidos com esse procedimento foram:

- identificação do gene de Sm e o equivalente em Sj;
- identificação do cromossomo/*scaffold*;
- coordenadas de início e fim do éxon no cromossomo/*scaffold*;
- posição do éxon no gene (éxon1, éxon2, etc),
- fita onde o elemento se encontra.

Com o objetivo de simplificar a descrição das análises que serão apresentadas a seguir, será adotada a nomenclatura acima descrita (ORTO, ORTO-ql e ORTO-qD) para descrever o conjunto de dados que foi utilizado.

### **3.1.2 Mapeamento dos elementos SR2 e Perere-3 no genoma**

Utilizando o *software* BlastN com *evaluate*  $10^{-3}$ , foi realizado o alinhamento das sequências dos elementos SR2 e Perere-3 contra o genoma de *S. mansoni* para definir a posição das inserções no genoma. Para selecionar os dados mais significativos, além do valor de corte mencionado acima, foram considerados apenas os resultados com comprimento maior que 50 bases e identidade superior a 85%.

## **3.2 Resultados**

A partir do conjunto de 11.607 genes de *S. mansoni* e 12.657 genes de *S.*

*japonicum* deduzidos a partir da sequência do genoma destes organismos, foram identificados 7.124 pares de genes ortólogos, representando 61% e 56% dos genes de *S. mansoni* e *S. japonicum*, respectivamente. Dentre esses genes, 4.595 apresentaram um único alinhamento entre suas sequências, sendo que apenas 2.752 desses genes apresentaram a mesma quantidade de éxons alinhados (ORTO-ql). Sobre esses últimos dados foram realizadas as análises para verificar a influência das inserções dos elementos SR2 e Perere-3.

Considerando todo o genoma dos organismos, foram obtidos 16.082 alinhamentos significativos para o elemento Perere-3, correspondendo à aproximadamente 1,83% do genoma de *S. mansoni* e para o elemento SR2, 27.334 alinhamentos correspondendo à aproximadamente 2,61% do genoma.



# Capítulo 4

## Análises nas regiões intrônicas

---



## 4 Análises nas regiões intrônicas

### 4.1 Considerações iniciais

Os íntrons constituem a parte não codificante dos genes e possuem sítios acceptor e doador, utilizados no processo de *splicing*. A maior parte dos íntrons apresentam as sequências consenso GT, para o sítio acceptor, e AG, para o sítio doador. Entretanto, foram observados íntrons com sequências GC-AG e AT-AC, para os respectivos sítios.(37) Além dos sítios acceptor e doador, os íntrons apresentam um sítio de ramificação (*branch site*) e um trato de polipirimidina, ambos situados *upstream* do sítio receptor e também utilizados no processo de *splicing*. Esse processo é necessário para a maturação do RNA, ou seja, a conversão do pré-mRNA em mRNA maduro, que é então transportado do núcleo para o citoplasma. Inicia-se com reconhecimento do sítio acceptor pelo snRNA (*small nuclear RNA*) U1. Essa primeira ligação visa promover a ligação do snRNA U2 no sítio de ramificação, para que ocorra a interação entre esses snRNA. Para isso, é necessário que o trato de polipirimidina seja reconhecido pela proteína U2AF, a qual interagirá com o sítio doador AG e promoverá a ligação do snRNA U2 no sítio de ramificação. Após essas interações, o íntron apresenta a forma de um laço e, após a união das extremidades dos éxons que estão flanqueando o íntron, o mesmo será eliminado como produto dessa reação.(37) Após o processo de *splicing*, o mRNA será transportado para o citoplasma para ser sintetizado.

Estudos indicam que o processo de *splicing* e transporte do mRNA estão correlacionados. Foi observado que complexos mRNP (*messenger ribonucleoprotein*) são unidos nas junções éxons-éxons durante o processo de *splicing*. Esse complexo por sua vez, contribuí de forma mais ativa para o transporte do mRNA para o citoplasma devido ao fato de recrutar o fator de exportação ALY. (38,39,40)

Também foi observado que durante o processo de *crossing-over*, íntrons longos aumentam a probabilidade desse evento ocorrer entre sequências codificantes separadas por esses íntrons, quando comparadas com as sequências codificantes sem a presença de íntrons. (38,41) Por sua vez, em alguns animais, íntrons curtos com potencial formação para *hairpin* (mirtons), são clivados e transformados em pré-miRNA e posteriormente, utilizados no complexo de silenciamento.(38,42,43)

Elementos de regulação, como os envolvidos na formação da estrutura da cromatina, também podem ser encontrados em íntrons. Ainda não se sabe se em quantidade significativa quando comparada com a quantidade desses elementos em outras regiões não codificantes. Quando há um enriquecimento desses elementos de regulação e o íntron está posicionado na extremidade 5' do gene, provavelmente estará sujeito à seleção purificadora. (38) Também podem conter sítio receptor de *trans-splicing*, que é uma forma especial de processamento do pré-mRNA, muito observada nos tripanossomas e nematóides.(38,44)

O evento de *trans-splicing* une éxons de dois pré-mRNA diferentes. Um pré-mRNA que contém o sítio receptor de *trans-splicing* e o SL RNA (*spliced-leader RNA*). Nesse tipo de processamento, o íntron que é liberado apresenta a forma de um Y.

As análises que serão apresentadas a seguir têm como objetivo verificar a distribuição dos elementos Perere-3 e SR2 nos íntrons de *S. mansoni* e verificar quais podem ser as possíveis influências destas inserções nas regiões intrônicas.

## 4.2 Metodologia

As coordenadas dos elementos das famílias Perere-3 e SR2 foram determinadas através do alinhamento das sequências modelo dos elementos transponíveis com o genoma utilizando o programa BlastN (Conforme descrito no item 3.1.2). Comparando as coordenadas dos trechos dos retrotransposons no

genoma com as coordenadas dos íntrons do conjunto ORTO-ql de *S. mansoni*, foram definidos os elementos localizados nas regiões intrônicas desses genes. Os íntrons que apresentaram inserções conjuntas dos elementos Perere-3 e SR2 foram separados, bem como o conjunto de íntrons sem inserções. Sobre o conjunto de íntrons que apresentaram inserções do elemento SR2 ou do elemento Perere-3 foram realizadas análises mais detalhadas que são descritas a seguir.

#### **4.2.1 Tamanho dos íntrons em *S. mansoni* e *S. japonicum***

A partir dos pares de genes ortólogos definidos anteriormente, foram identificados os íntrons de *S. japonicum* equivalentes aos íntrons de *S. mansoni* que apresentaram ou não elementos de transposição.

Dessa forma, foram gerados três conjuntos de dados para cada organismo. Um contendo os íntrons com elementos Perere-3, outro com os elementos SR2 e o último, com os íntrons sem a presença dos retrotransposons. Para esses dados foi verificado o percentual de íntrons com elementos de transposição, o comprimento médio desses íntrons e qual organismo apresentava o maior íntron. Também foi verificada a distribuição do tamanho dos íntrons e gerada uma representação gráfica.

#### **4.2.2 Posição das inserções no gene**

Também foram realizadas análises para verificar se os diferentes íntrons de um gene possuem a mesma frequência de inserções de elementos transponíveis ou se existem diferenças indicativas de favorecimento de determinadas posições.

Dentre todos os genes do conjunto de dados ORTO-ql, foi verificado qual desses genes apresentava a maior quantidade de íntrons, ou seja, genes com

apenas 1 íntron, 2 íntrons até, no máximo, 32 íntrons. Representando essas quantidades na forma de um vetor, para cada posição do vetor foi acumulada a quantidade de genes que apresentava a mesma quantidade de íntrons que a posição atual do vetor. Com base nessas informações, foi calculado o número total de íntrons do conjunto ORTO-ql e o percentual que cada quantidade de íntrons representava. Por exemplo, para os genes com 4 íntrons, foram acumulados 276 genes, totalizando 1.104 íntrons que representam 9.49% dentro o total de 11.636 íntrons pertencentes aos genes do conjunto ORTO-ql.

Para cada observação de elemento de transposição em um íntron, foi registrada a posição do íntron onde o elemento estava localizado e o número total de íntrons do gene associado. Uma vez registrada a posição de todos os elementos, foi calculado o percentual de representatividade de íntrons com elementos de transposição em relação ao número total de íntrons dos genes do conjunto ORTO-ql (por exemplo, os íntrons na posição X, dos genes que contém Y íntrons, representam Z% dentro todos os íntrons dos genes do conjunto ORTO-ql)

A Figura 17 representa graficamente a metodologia utilizada na análise que verificou a proporção das inserções nas diversas posições intrônicas de um gene. No eixo X são representados os íntrons dos genes e no eixo Y o total de íntrons que o gene possui. A diagonal indica a quantidade máxima de íntrons dos genes e indica o percentual de representação de cada íntron do gene mediante todos os íntrons do conjunto de dados ORTO-ql.

### **4.2.3 Distância das inserções em relação aos éxons**

Para definir o posicionamento das inserções em relação ao comprimento total do íntron e, conseqüentemente, a distância em relação aos éxons adjacentes, os íntrons foram divididos em dois segmentos, como ilustra a Figura 18. Um deles representando a distância entre o início do íntron e uma extremidade do elemento de transposição, e o outro segmento representando a distância que compreende a

outra extremidade do elemento e o término do íntron. Dividindo um destes intervalos pela soma de ambos, pode-se observar a posição relativa ao comprimento total na qual a inserção ocorreu no íntron primitivo, ou seja, sem a presença do elemento.

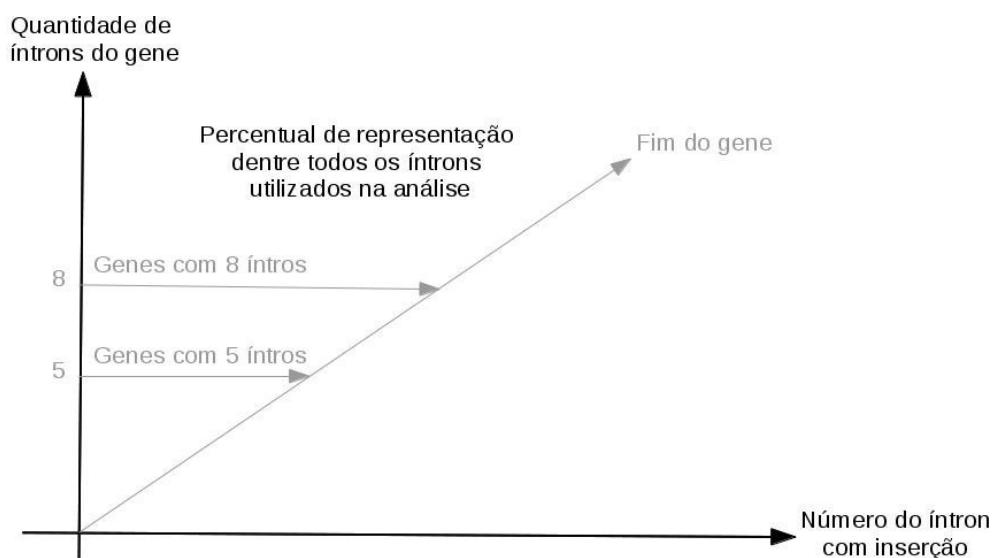


Figura 17-Illustração gráfica da metodologia utilizada para representar uma possível incidência de trechos dos TEs em íntrons específicos. Os dados foram representados em uma matriz onde o eixo X representa a posição do íntron no qual foi verificado o TE e o eixo Y a quantidade total de íntrons dos genes analisados. A diagonal indica a quantidade máxima de íntrons dos genes e limita as linhas horizontais. A indicação do percentual de representação de cada íntron do gene em relação a todos os íntrons do conjunto de dados ORTO-ql foi representada em escala de cores ao longo dos eixos. Fonte: Elaborada pela autora.

Os resultados apresentados consideram a distância entre o início do íntron, próximo do éxon posicionado na região 5', e a extremidade do elemento mais próxima a esse ponto.

Também foram verificados os sentidos das inserções com relação ao sentido da transcrição do gene. Os dados foram acumulados em intervalos de 0.1 e representados na forma de histograma.

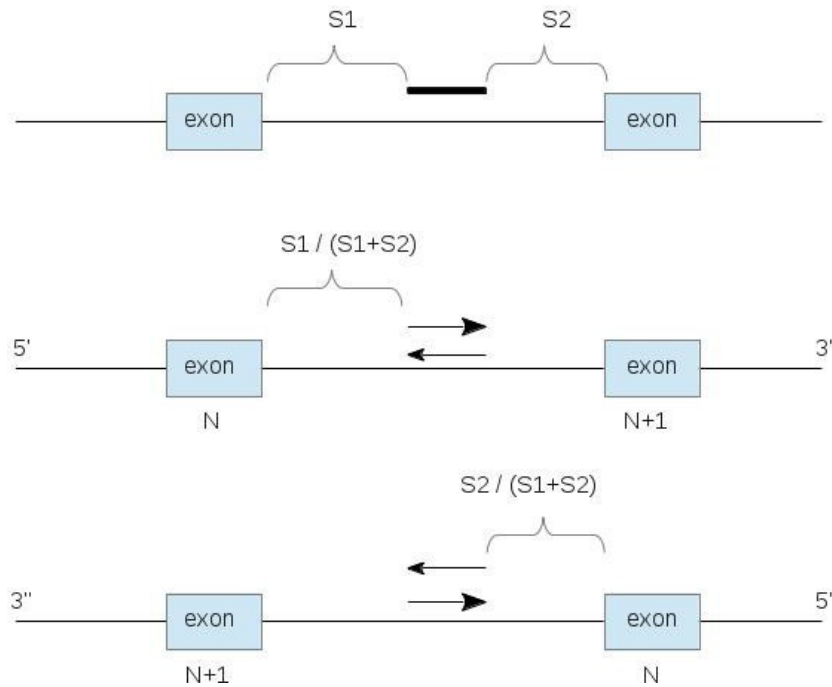


Figura 18-Representação esquemática do método de fração utilizado para definir a posição do TE no íntron e a distância na qual o TE ocorreu em relação ao éxon situado a 5'. Os retângulos representam os éxons e as setas as inserções e seu sentido. Fonte: Elaborada pela autora.

A mesma análise também foi realizada considerando apenas as inserções que continham cópias mais completas dos elementos Perere-3 e SR2, ou seja, com capacidade de produzir as proteínas para a sua transcrição. Considerando que a sequência completa desses elementos apresenta 3.327 bases e 3.913 bases, respectivamente, foram selecionadas as inserções de Perere-3 com mais de 3.000 bases e de SR2 com mais de 3.300 bases.

#### 4.2.4 Conteúdo CG

Foram realizadas verificações com relação ao conteúdo CG dos íntrons contendo inserções. Em um primeiro momento, a análise foi pontual. Considerando uma janela de 10% do tamanho do íntron, a qual se desloca ao longo do íntron de



uma em uma base. Para cada janela são contadas as quantidades de nucleotídeos G e C e calculado o percentual GC dessas janelas. Esse processo é repetido até a última base do íntron, como ilustra a Figura 19. Dessa forma, é possível observar o conteúdo GC ao longo de todo o íntron.(45)

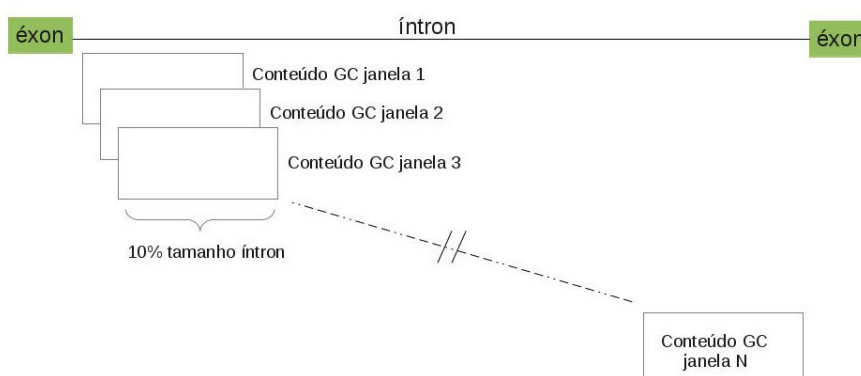


Figura 19-Representação esquemática da metodologia de *sliding windows* implementada para verificar o conteúdo GC ao longo dos íntrons com TEs e o íntron equivalente em *S. japonicum*. Fonte: Elaborada pela autora.

Uma segunda análise verificou o conteúdo GC médio de 3 conjuntos de íntrons: os de *S. mansoni* que apresentaram trechos dos retrotransposons, seus íntrons equivalentes em *S. japonicum* e, íntrons de *S. mansoni*, sem trechos dos elementos TEs. Os íntrons do último conjunto mencionado foram selecionados de forma aleatória e na mesma quantidade de íntrons do conjunto observado para *S. mansoni* e *S. japonicum*. Nessa seleção também foi considerado como tamanho mínimo de íntron, a menor quantidade de bases apresentada pelos íntrons com a presença dos retrotransposons.

Esses dados foram agrupados em intervalos de percentuais GC e para cada intervalo foi calculado o percentual de representatividade desses íntrons em relação a todos os íntrons contendo trechos dos elementos Perere-3 ou SR2.

#### 4.2.5 Inserções de ilhas CpG

Especificamente em relação ao elemento SR2, foram identificadas duas regiões com padrões similares a ilhas CpG.(46,47) São consideradas ilhas CpG as sequências com mais de 300 bp, com percentual GC acima de 55% e a relação entre o valor observado e o valor estimado é de 0.65.

Todas as inserções em íntrons que continham as regiões de ilhas CpG foram selecionadas e as análises descritas nos itens 4.2.2 (posição do elemento no gene) e 4.2.3 (distância das inserções em relação aos éxons) foram executadas novamente apenas sobre esse conjunto de dados.

Para identificar a ocorrência de algum viés de representatividade de trechos do elemento contendo ilhas CpG, foram realizadas simulações para verificar se a distribuição de regiões do elemento era próxima a esperada, caso não houvesse favorecimento na fixação de nenhuma região do retrotransposon.

Para cada elemento em íntrons, foi selecionado aleatoriamente um trecho do elemento SR2, utilizando o mesmo tamanho do dado real. Era verificado se o trecho selecionado correspondia ou não a um trecho de ilha CpG. Foi calculada a relação entre os valores observado e esperados. Para essa análise foram desconsiderados os elementos do genoma correspondentes ao elemento SR2 não autônomo uma vez que esse elemento, não contém os trechos equivalentes à ilhas CpG. Esses elementos não autônomos equivalem a 25% (468 trechos) dentre todos os elementos de SR2 nos íntrons.

#### **4.2.6 Enriquecimento de termos de Gene Ontology no conjunto de genes com inserções**

A análise para verificar o enriquecimento de termos ontológicos no conjunto de genes contendo inserções em íntrons foi realizada utilizando o programa *Ontologizer*.(48) O arquivo contendo as associações entre os genes de *S. mansoni* e os termos GO foi obtido no *site* do Instituto Sanger.(49)

Foram utilizadas as opções de método estatístico *Parent-Child*, método

*Bonferroni* para correção de erros das múltiplas comparações e nível de significância 0.05.

### 4.3 Resultados e discussão

Foram verificados os tamanhos de íntrons de três conjuntos de dados: os íntrons sem a presença dos elementos estudados, com inserções do elemento Perere-3 e com inserções do elemento SR2.

Dos 8.279 íntrons do conjunto de dados analisado (ORTO-qi), 6.706 não apresentaram inserções e um padrão similar de distribuição de tamanhos para ambos os organismos, como ilustra a Figura 20. Dentre os íntrons restantes, 138 apresentaram trechos de ambos os elementos.

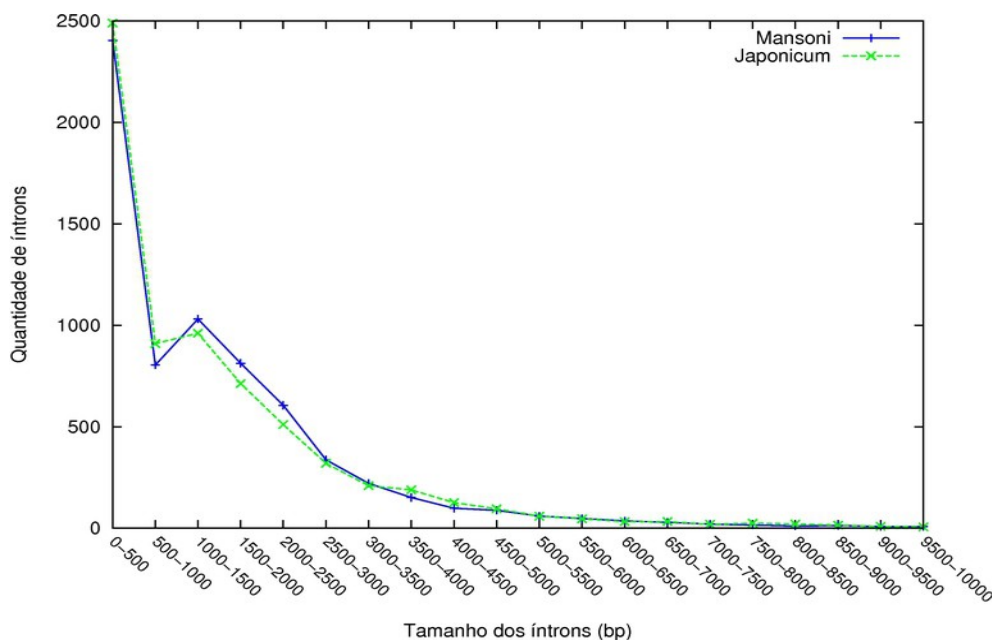


Figura 20-Padrão de distribuição do tamanho dos íntrons de *S. mansoni* sem inserções dos elementos estudados (linha azul) e seus ortólogos de *S. japonicum* (linha verde). Fonte: Elaborada pela autora.

O elemento Perere-3 apresentou inserções em 255 genes distintos de *S.*

*mansoni* (9,3% do conjunto ORTO-ql), totalizando inserções em 304 íntrons. Desses 304 íntrons, 235 são maiores em *S. mansoni* quando comparados com o ortólogo de *S. japonicum*. O tamanho médio destes íntrons é de 4.528 bases e 2.748 bases em *S. mansoni* e *S. japonicum*, respectivamente.

De forma similar, o elemento SR2 apresentou inserções em 794 genes distintos (28.9% do conjunto ORTO-ql), totalizando as inserções em 1.133 íntrons. Desses 1.133 íntrons, 818 são maiores em *S. mansoni* quando comparados com o ortólogo de *S. japonicum*. O tamanho médio dos íntrons é de 3.916 bases e 2.671 bases em *S. mansoni* e *S. japonicum*, respectivamente.

As inserções de ambos os retrotransposons se concentram em 941 genes distintos que corresponde a 34% dos genes do conjunto ORTO-ql.

O padrão observado para os íntrons sem a presença dos elementos transponíveis estudados não é o mesmo quando são analisadas as distribuições de tamanhos dos íntrons com inserções dos elementos Perere-3 e SR2. Para ambos os casos, *S. mansoni* apresenta uma quantidade maior de íntrons mais longos, quando comparados com os íntrons equivalente em *S. japonicum* (Figura 21 e Figura 22).

Considerando todos os íntrons de *S. mansoni* com elementos Perere-3 e SR2, e os íntrons equivalentes em *S. japonicum*, através do teste estatístico de Wilcoxon\*, disponível em R (50), foi obtido p-valor de 2.2e-16 para o conjunto de ambos os retrotransposons. Esse valor, sendo menor que o nível de significância (0.05 por padrão), indica que as populações de íntrons com elementos de transposição de fato possuem tamanhos significativamente diferentes da população de respectivos íntrons ortólogos de *S. japonicum*.

---

\* O teste de Wilcoxon é utilizado para comparar se as medidas de posição de duas amostras são iguais no caso em que as amostras são dependentes. Disponível em <http://www.r-tutor.com/elementary-statistics/non-parametric-methods/wilcoxon-signed-rank-test>

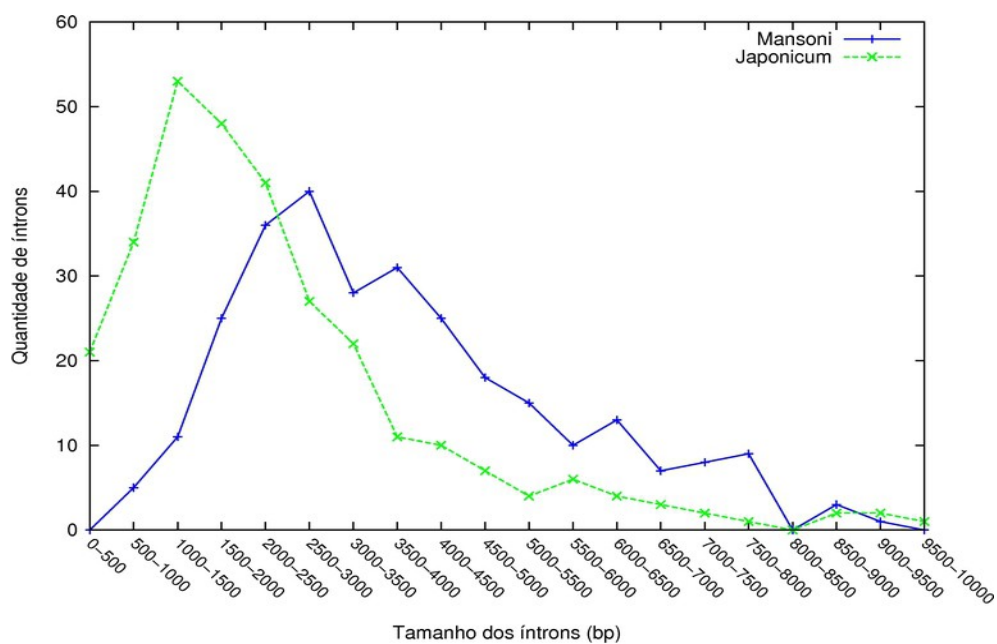


Figura 21-Padrão de distribuição do tamanho dos íntrons de *S. mansoni* com elementos Perere-3 (linha azul) e seus ortólogos de *S. japonicum* (linha verde). Fonte: Elaborada pela autora.

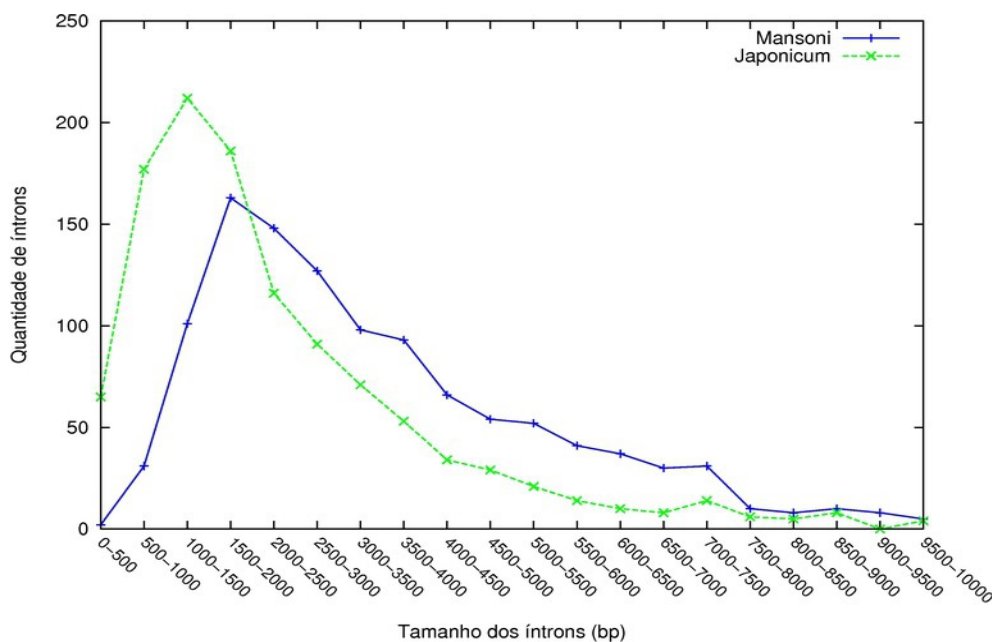


Figura 22-Padrão de distribuição do tamanho dos íntrons de *S. mansoni* com elementos SR2 (linha azul) e seus ortólogos de *S. japonicum* (linha verde). Fonte: Elaborada pela autora.

Um fator que pode contribuir para esta observação é que uma fração considerável destes elementos deve ter se inserido no genoma em um período relativamente recente.(15) Isso faz com que processos de recombinação, para eliminação de porções do elemento, tenham ocorrido de maneira relativamente limitada.

As inserções de elementos transponíveis em íntrons podem provocar perturbações na transcrição do gene mas ainda não são conhecidos todos os fatores que levam a essa perturbação. Dessa forma, a avaliação do impacto desses elementos através de análises computacionais é dificultada.(51)

Foi analisada a frequência que os retrotransposons em estudo apresentavam em uma determinada região intrônica. A maior parte dos elementos, estão posicionados de forma individual em uma determinada região intrônica, como ilustra a Figura 23.

Para o elemento SR2 foram identificados aproximadamente 4.500 trechos que apresentaram junções, entre as bases 448 e 3838, alinhadas de forma contínua, com intervalo de 10 bases entre elas, como ilustra a Figura 24.

Este perfil é decorrente do fato de que o elemento SR2 apresenta um elemento não autônomo, o qual é caracterizado por apresentar o trecho inicial e final do elemento autônomo, separados por 10 bases, como ilustra a Figura 25.

Dos elementos SR2 analisados em todo o genoma, os elementos autônomos correspondem à 1.12% das bases do genoma e os elementos não autônomos à 1.26%.

Para ambos os retrotransposons em estudo também foi verificado o tamanho médio dos trechos desses elementos. Enquanto Perere-3 apresenta uma tendência central de trechos com 133 bases, o elemento SR2 apresenta trechos em torno de 317 bases, como ilustra a Figura 26.

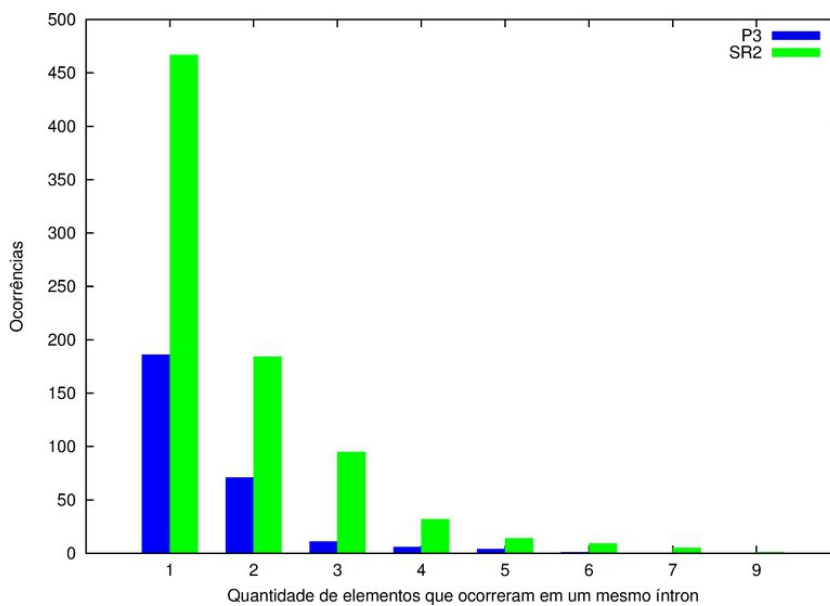


Figura 23-Distribuição do número de ocorrências do elemento Perere-3 (azul) e SR2 (verde) em uma mesma região intrônica. Fonte: Elaborada pela autora.

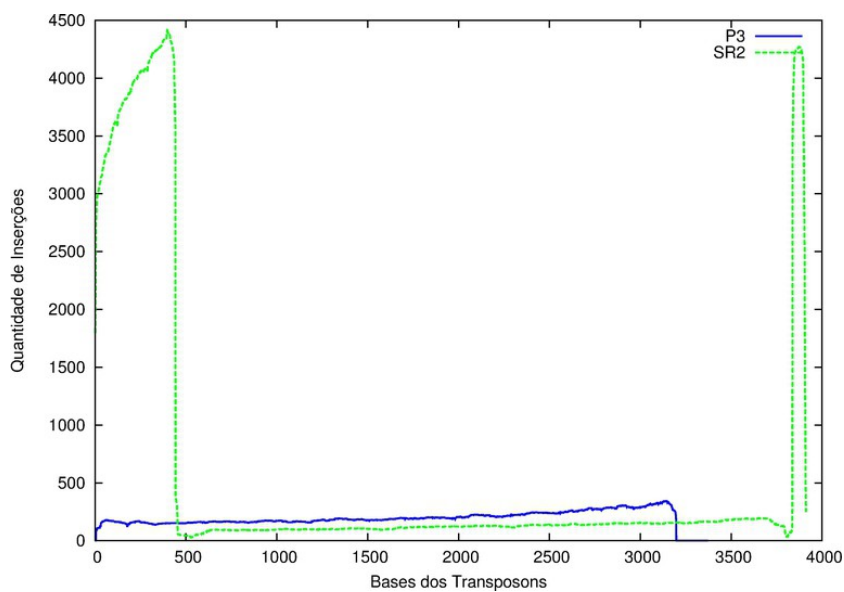


Figura 24-Frequência encontrada para as diferentes porções do retrotransposon nas cópias encontradas no genoma de *S. mansoni*. A linha em azul e verde representam os dados obtidos para os elementos Perere-3 e SR2, respectivamente. Fonte: Elaborada pela autora.

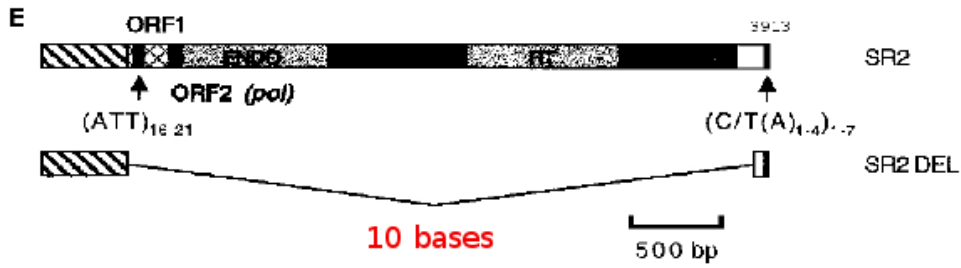


Figura 25-Estrutura do elemento SR2 autônomo e não autônomo (SR2 DEL).  
Fonte: Figura adaptada de DREW et al. (13)

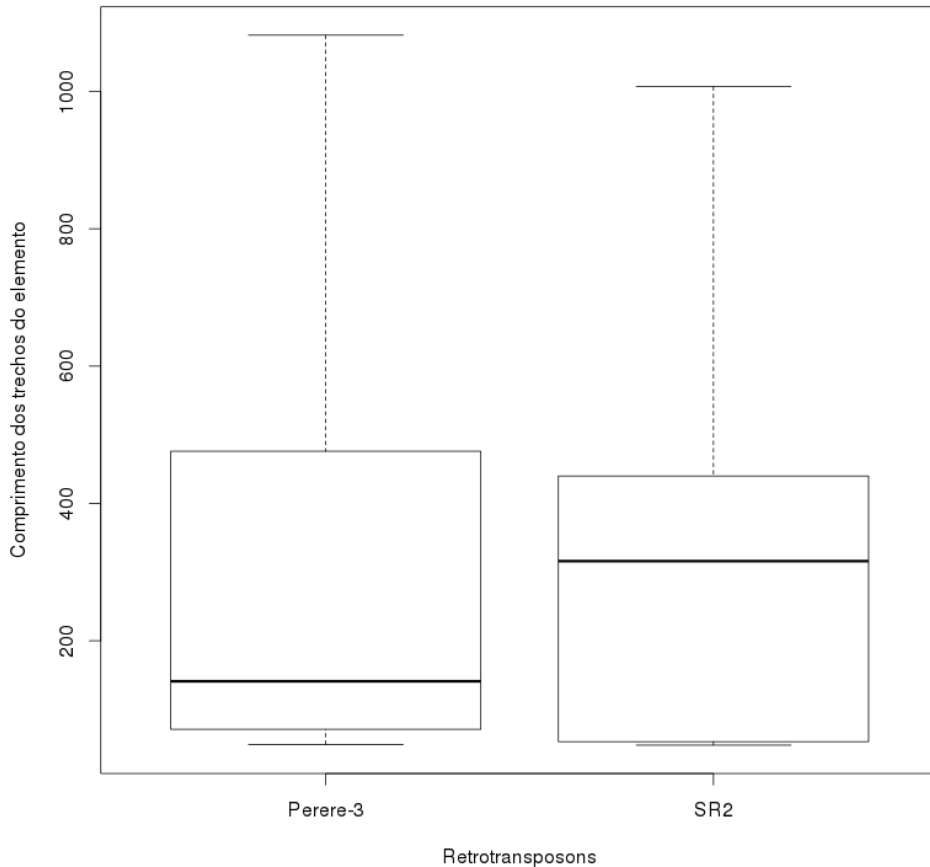


Figura 26-Boxplot ilustrando o comprimento dos trechos do elemento Perere-3 e SR2 nas regiões intrônicas. A caixa representa 50% de todos os valores observados, concentrados na tendência central dos valores (mediana: 133 bases para Perere-3 e 317 bases para SR2). A base da caixa representa o quartil inferior dos menores valores, e o topo da caixa o quartil superior dos valores observados. Os segmentos de reta conectam a base e o topo da caixa ao ponto de dados mais extremo que não é superior a 1.5 vezes o intervalo interquartil da caixa (método Tukey). Fonte: Elaborada pela autora.



### 4.3.1 Posição das inserções no gene e no íntron

Foram realizadas análises para verificar se os diferentes íntrons de um gene possuem a mesma frequência de inserções de elementos transponíveis ou se existem diferenças indicativas de favorecimento para determinadas posições.

Observando-se a Figura 27 nota-se que há enriquecimento de inserções em íntrons próximos a extremidade 3' dos genes (quadrados próximos a diagonal do gráfico), que tendem a apresentar maior frequência de elementos Perere-3, principalmente em genes com baixo número de íntrons. Um padrão muito semelhante pode ser observado na Figura 28, onde estão representados os íntrons contendo o elemento SR2.

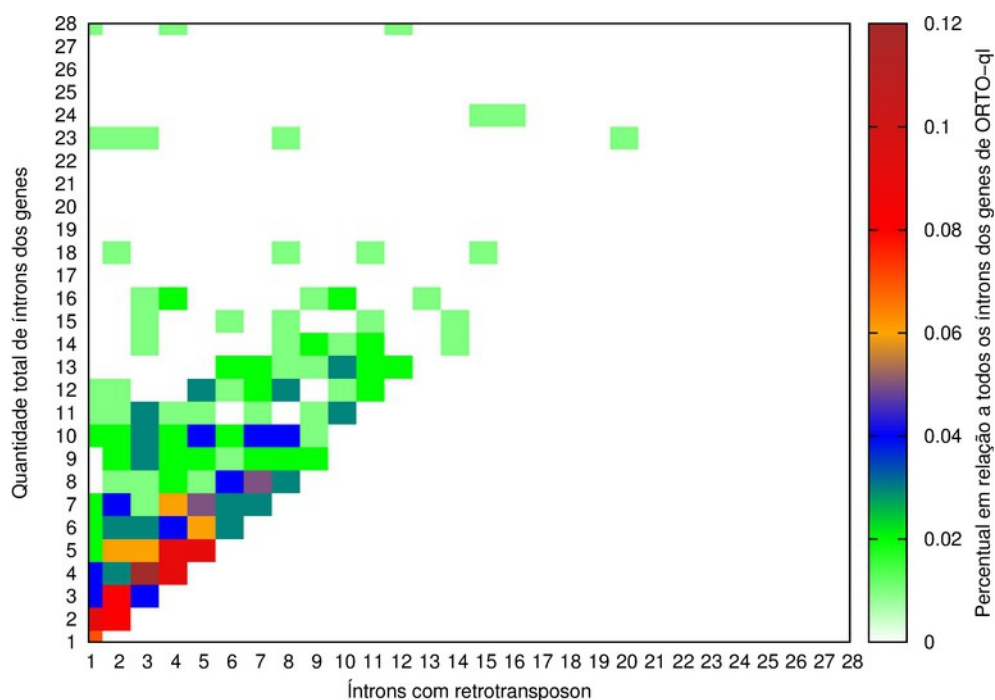


Figura 27-Percentual de íntrons contendo elementos Perere-3 em relação ao número total de íntrons dos genes do conjunto ORTO-ql. O gráfico mostra em diferentes cores a frequência relativa de inserções do retrotransposon Perere-3 nos íntrons conforme escala mostrada na direita da figura. Os dados foram organizados separando as diferentes posições dos íntrons ao longo do gene (eixo X – íntrons ordenados do início para o final do gene) e de tal modo que cada linha indica um grupo de genes contendo o número de íntrons indicando o eixo Y. Fonte: Elaborada pela autora.

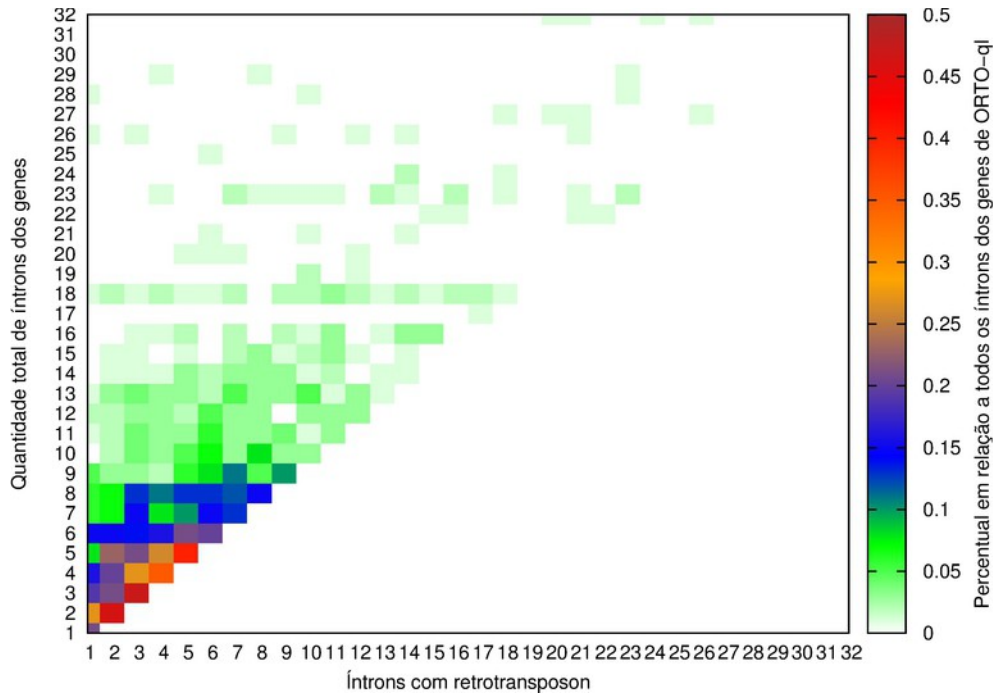


Figura 28-Percentual de íntrons contendo elementos SR2 em relação ao número total de íntrons dos genes do conjunto ORTO-qI. O gráfico mostra em diferentes cores a frequência relativa de inserções do retrotransposon SR2 nos íntrons conforme escala mostrada na direita da figura. Os dados foram organizados separando as diferentes posições dos íntrons ao longo do gene (eixo X – íntrons ordenados do início para o final do gene) e de tal modo que cada linha indica um grupo de genes contendo o número de íntrons indicando o eixo Y. Fonte: Elaborada pela autora.

A diminuição no percentual de elementos transponíveis em íntrons próximos às regiões 5' dos genes, pode estar relacionada ao fato de que inserções nestas regiões apresentam maior probabilidade de afetar elementos de controle da transcrição, os quais são encontrados nesta região. A maior fração das inserções localizadas nessa extremidade dos genes seriam selecionadas negativamente.

Considerando que inserções próximas aos éxons possuem maior chance de interferir nos padrões de *splicing*, devido à proximidade do sítio acceptor ou doador, foi verificada a distância dos elementos em relação aos éxons adjacentes, como ilustra a Figura 29, para as inserções do elemento Perere-3 e a Figura 30, para o elemento SR2 em *S. mansoni*.

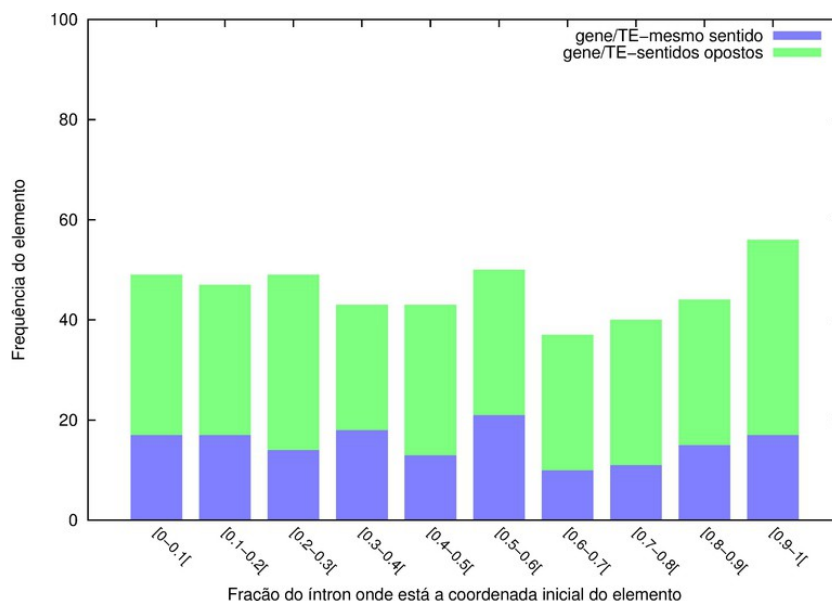


Figura 29-Verificação da posição do elemento Perere-3 dentro dos íntrons analisados. Frações representadas no eixo x indicam a distância relativa do elemento em relação a extremidade 5' do íntron. As barras na cor verde representam elementos orientados no sentido oposto ao da transcrição do gene e as barras na cor azul representam as inserções que ocorreram no mesmo sentido que a transcrição do gene. Fonte: Elaborada pela autora

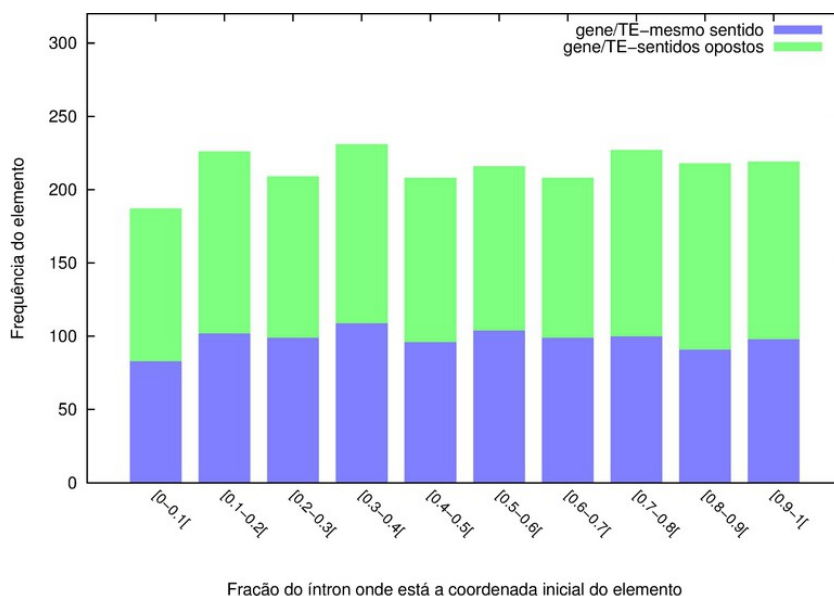


Figura 30-Verificação da posição do elemento SR-2 dentro dos íntrons analisados. Frações representadas no eixo x indicam distância relativa do elemento em relação a extremidade 5' do íntron. Fonte: Elaborada pela autora.

Nenhuma dessas análises revelou existir qualquer tendência bem definida na distância dos elementos em relação aos éxons.

Considerando o fato de que um elemento tem a mesma probabilidade de se inserir no genoma no sentido senso ou antisenso, foi utilizada a função teste binomial exata\*, disponível no ambiente R, para verificar a probabilidade de se observar uma frequência antisenso mais extrema do que a observada na análise (p-valor).

Para Perere-3 foram observados 271 casos no sentido antisenso e 137 casos no sentido senso, o que retrata um número de inserções antisenso significativamente maior (p-valor=3.058e-11).

Para o elemento SR2 foram observados 1.018 casos no sentido antisenso e 863 casos no sentido senso, também indicando que o número observado de inserções antisenso é maior do que o esperado com significância estatística (p-valor=0.0003814).

Elementos que se encontram no mesmo sentido que a transcrição do gene podem conter trechos direcionais que apresentam diferentes influências. A inserção de sítios de poliadenilação, por exemplo, produziria transcritos truncados se este sítio possuísse a mesma orientação do transcrito produzido.(52) Evidente que em decorrência de tal fato, sequências com tais características serão deletérias e terão uma probabilidade menor de serem fixadas.

Considerando que a sequência dos elementos Perere-3 e SR2 apresentam 3.327 bases e 3.913 bases, respectivamente, definimos como sequências mais completas de ambos os elementos, as inserções com mais de 3.000 bases, para o elemento Perere-3, e com mais de 3.300 bases, para o elemento SR2.

Analisando as sequências mais completas inseridas pelos elementos Perere-3 e SR2, foi possível verificar que as cópias completas do elemento Perere-3, apresentam, quase que em sua totalidade, elementos no sentido contrário da transcrição do gene (p-valor=0.02148), como ilustra a Figura 31. Para as cópias completas do elemento SR2 há um equilíbrio entre os elementos no mesmo sentido

---

\* Um teste binomial compara o número de sucessos observados num certo número de ensaios com uma probabilidade de sucesso hipotético. Disponível em <http://www.instantr.com/2012/11/06/performing-a-binomial-test/>

da transcrição do gene e em sentido oposto.

Novamente, é possível considerar que em decorrência do período relativamente recente em que ocorreram as inserções do elemento SR2, a ação de processos de recombinação e de seleção natural podem ter atuado de maneira mais limitada do que os eventos que atuaram sobre os elementos Perere-3. A maior divergência entre as sequências desses últimos elementos, indica que os mesmos estão expostos a mais tempo a ação de processos de mutação. O fato dos elementos mais completos, que apresentam todas as características para promover a transposição do elemento, serem encontrados em uma quantidade pequena, principalmente no mesmo sentido que da transcrição do gene, pode ser um reflexo da atuação desses processos.

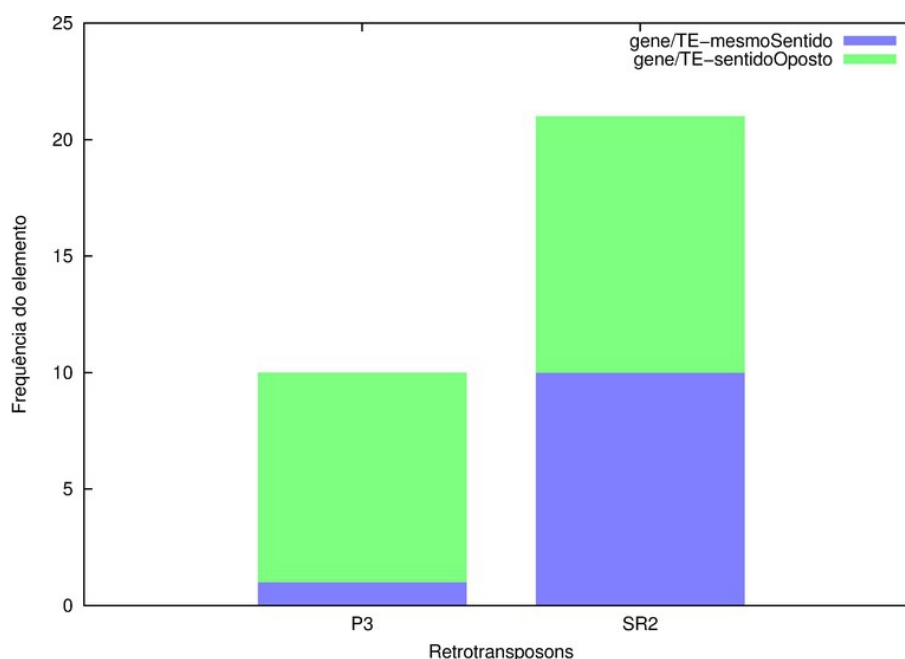


Figura 31-Distribuição da quantidade de trechos mais completos dos elementos Perere-3 e SR2 e o sentido que esse elemento se encontra com relação ao mesmo sentido da transcrição do gene (azul) e no sentido oposto ao da transcrição (verde). Fonte: Elaborada pela autora.

### 4.3.2 Influência de elementos SR2 e Perere-3 no conteúdo GC dos íntrons

O organismo de *S. mansoni* apresenta um conteúdo GC médio de 35.2% nas regiões não codificantes, 36.3% nas codificantes (9) e de 37.3% em regiões que apresentaram sítios de metilação.(31) Deste modo, pode-se considerar que este é um genoma relativamente pobre nessa classe de bases.

Genes que possuem ilhas CpG, são frequentemente, altamente expressos em múltiplos tecidos e tais sequências, em humanos e ratos, aparecem enriquecendo aproximadamente metade dos promotores, sugerindo um importante papel para as ilhas CpG no processo de regulação da transcrição.(53)

Análises do conteúdo GC ao longo dos íntrons, com a presença de elementos Perere-3 e SR2, permitem concluir que eles alteram o conteúdo GC na região da inserção, como ilustra a Figura 32. O elemento Perere-3 apresenta conteúdo GC médio equivalente a 45.3% e o elemento SR2 a 48.5%.

Também é possível observar que a presença dos elementos eleva o conteúdo GC médio dos íntrons. A Figura 33 ilustra o percentual GC dos íntrons com o elemento Perere-3 e a maioria desses íntrons apresentam conteúdo GC acima de 35%. A Figura 34 retrata a mesma distribuição para os íntrons com o elemento SR2 e também é possível notar um aumento do conteúdo GC médio dos íntrons. Mais do que 25% dos íntrons apresentam percentual GC entre 37% e 39%.

Considerando o percentual GC de todos os íntrons de *S. mansoni* com elementos Perere-3 e SR2, e os íntrons equivalentes em *S. japonicum*, através do teste estatístico de Wilcoxon, foi obtido p-valor de 2.2e-16 para o conjunto de ambos os retrotransposons, demonstrando que as populações de íntrons com elementos de transposição apresentam percentual GC significativamente diferentes.

Em estudos descritos na literatura (54), observou-se que inserções que promovem o alongamento dos íntrons e alteram o conteúdo GC entre os éxons e íntrons podem contribuir para a alteração dos padrões de *splicing*. Supõe-se que o conteúdo GC elevado dos éxons, em relação aos íntrons que o flanqueiam, parece

ser o sinal que permite a identificação dos éxons.(55)

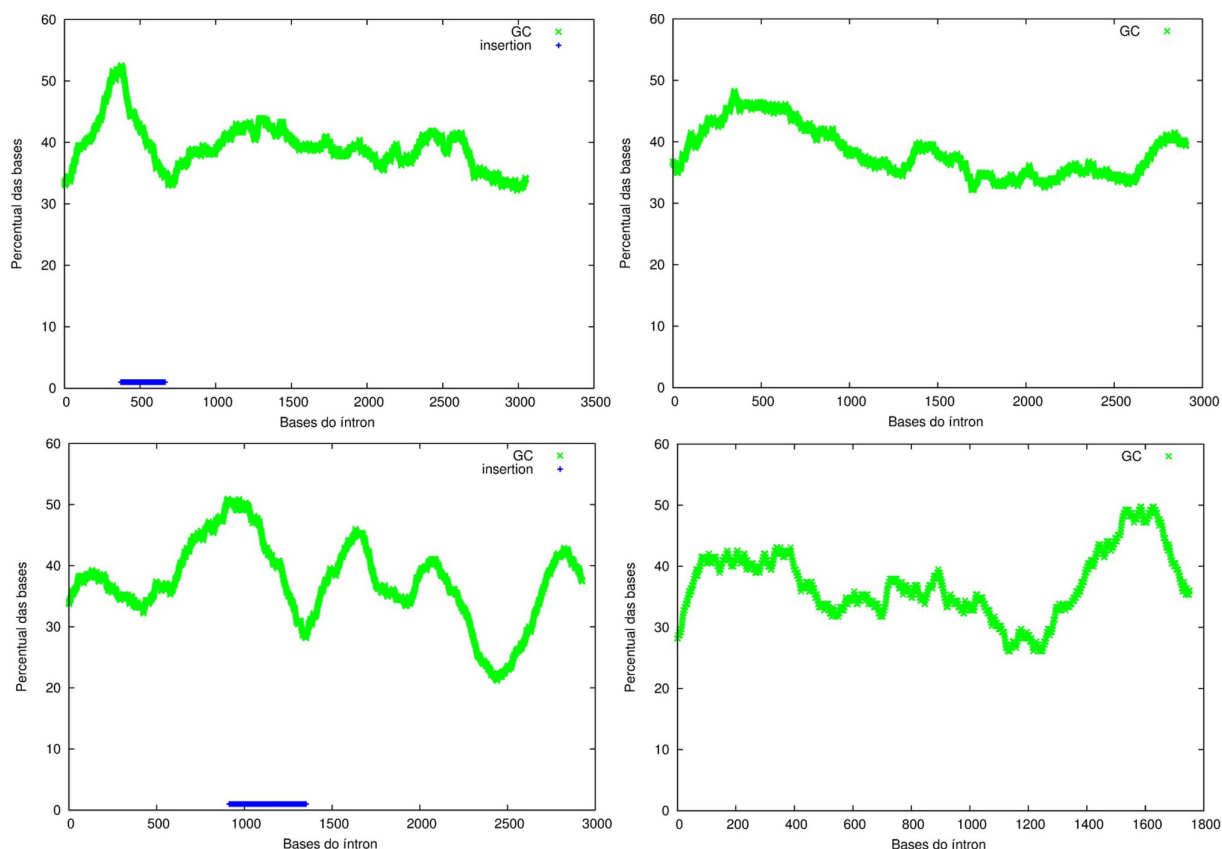


Figura 32-Representação gráfica do conteúdo GC de um íntron utilizando a metodologia de *sliding windows* (overlapping) com 10% do tamanho do íntron. A linha em verde representa o conteúdo GC ao longo do íntron e a linha em azul a posição do trecho inserido pelo elemento SR2. Os gráficos da coluna à esquerda, representam os íntrons de *S. mansoni* com trechos do elemento SR2 (Smp\_0099930-1 (gene Smp\_0099930 íntron 1) e Smp\_0177250-2) e os gráficos da coluna da direita, representam os íntrons equivalente em *S. japonicum* (Sjc\_0083430-1 e Sjc\_0060510-2). Fonte: Elaborada pela autora.

Análises do conteúdo GC ao longo da extensão dos elementos completos permitiu a identificação de dois trechos na sequência do elemento SR2 correspondentes a trechos de ilhas CpG. O primeiro trecho entre as coordenadas 1.314 e 1.707 e o segundo trecho entre as coordenadas 1.846 e 2.315. Para o elemento Perere-3 não foram identificados trechos correspondentes à ilhas CpG.

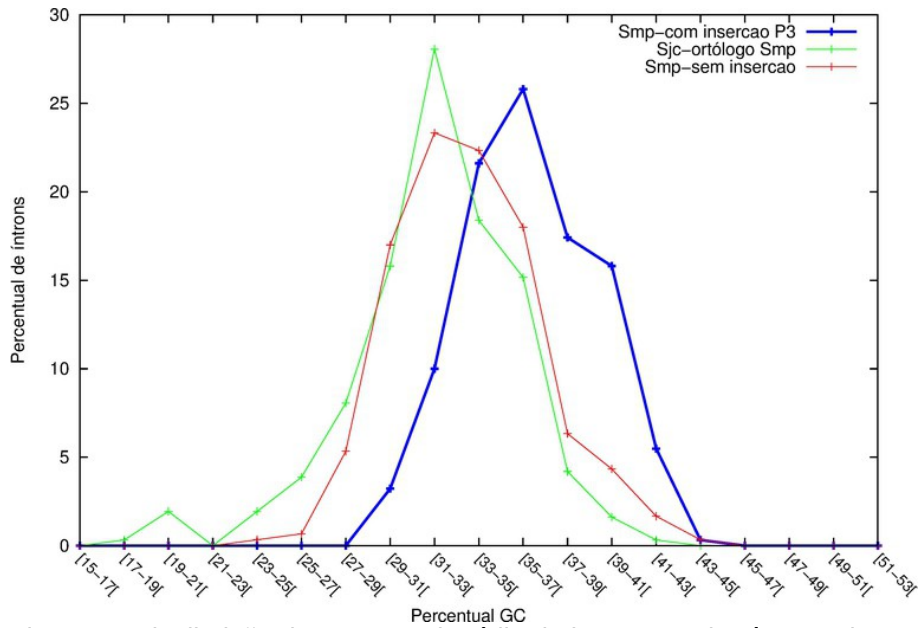


Figura 33-Distribuição do percentual médio de bases GC dos íntrons de *S. mansoni* com elementos Perere-3 (linha azul), íntrons ortólogos de *S. japonicum* (linha verde) e íntrons de *S. mansoni* que não apresentaram inserções do elemento em análise (linha vermelha). Fonte: Elaborada pela autora

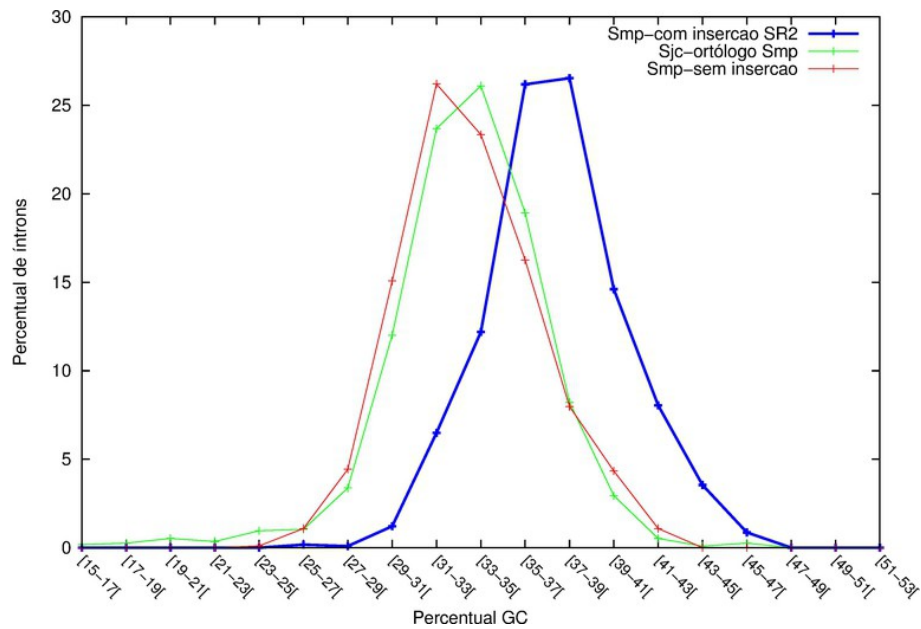


Figura 34-Distribuição do percentual médio de bases GC dos íntrons de *S. mansoni* com elementos SR2 (linha azul), íntrons ortólogos de *S. japonicum* (linha verde) e íntrons de *S. mansoni* que não apresentaram inserções do elemento em análise (linha vermelha). Fonte: Elaborada pela autora



Para os elementos contendo esses trechos foram realizadas análises para verificar a posição de íntron no gene que os elementos se concentravam (Figura 35), e a proximidade destes elementos com relação ao éxon posicionado na extremidade 5' do íntron (Figura 36).

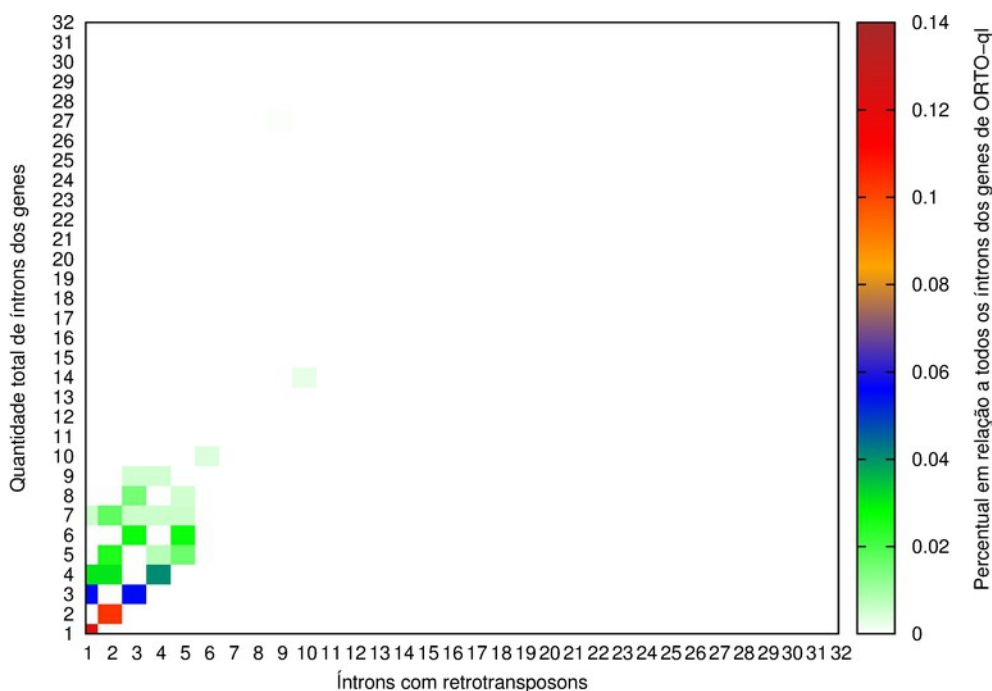


Figura 35-Percentual de íntrons com trechos do elemento SR2 contendo ilhas CpG em relação ao número total de íntrons dos genes do conjunto ORTO-ql. O gráfico mostra em diferentes cores a frequência relativa de inserções de retrotransposon SR2 nos íntrons conforme escala mostrada na direita da figura. Os dados foram organizados separando as diferentes posições dos íntrons ao longo do gene (eixo X - íntrons ordenados do início para o final do gene) e de tal modo que cada linha indica um grupo de genes contendo o número de íntrons indicado no eixo Y. Fonte: Elaborada pela autora.

O padrão de concentração no final do íntron, verificada na análise de todos os trechos do elemento, de forma geral permaneceu. Essa tendência provavelmente reflete o fato de que inserções próximas às extremidades 5' dos genes poderiam produzir o silenciamento gênico e, conseqüentemente, estarem sujeitas a um maior processo de seleção.

Também é possível observar uma concentração de elementos contendo regiões de ilhas CpG nas pontas dos éxons (Figura 36), indicando provável atuação

de uma pressão seletiva negativa para elementos com trechos de ilhas CpG nas regiões centrais dos íntrons.

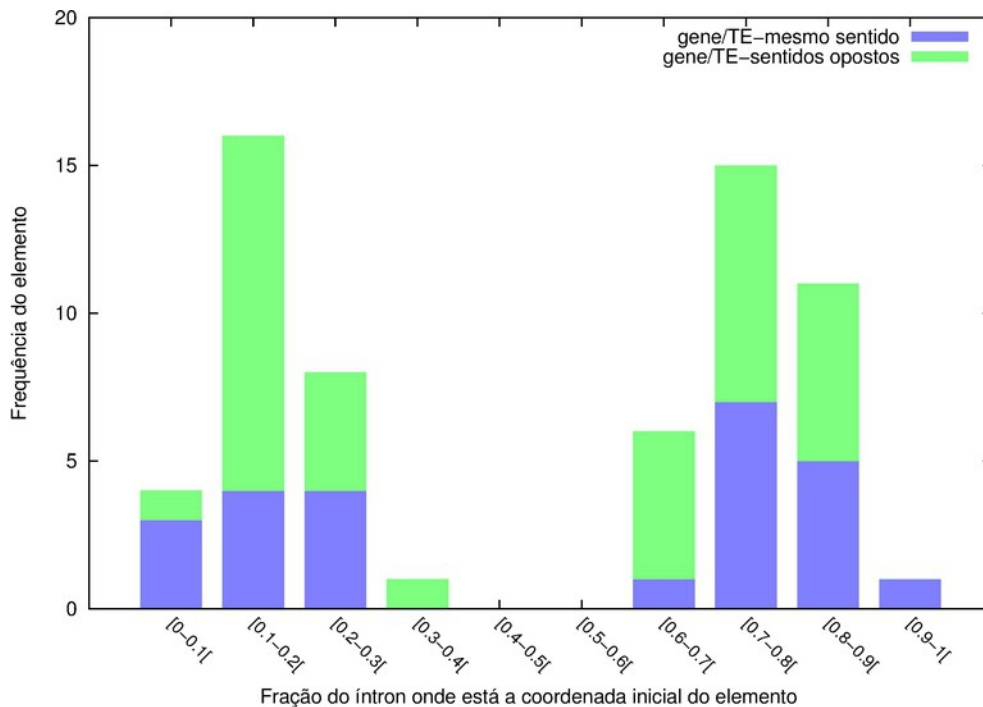


Figura 36-Verificação da posição do elemento SR2 contendo ilhas CpG dentro dos íntrons analisados. Frações representadas no eixo x indicam distancia relativa do elemento em relação a extremidade 5' do íntron. As barras na cor verde representam elementos orientados no sentido oposto ao da transcrição do gene e as barras na cor azul representam as inserções que ocorreram no mesmo sentido que a transcrição do gene. Fonte: Elaborada pela autora.

Este resultado representa uma tendência diferente daquela encontrada quando todas as regiões do elemento foram pesquisadas. No entanto, é necessário cautela ao interpretar este resultado devido ao pequeno número de regiões amostradas.

Para verificar se esses trechos de ilhas CpG estavam ocorrendo na frequência esperada, foi realizada uma simulação na qual, para cada trecho do elemento observado, foi selecionado aleatoriamente um trecho no retrotransposon, com o mesmo tamanho que o do trecho real, e verificado se o mesmo correspondia ou não a um dos trechos de ilha CpG.

Os resultados demonstram que essas regiões apresentam frequência maior

do que seria esperado em um processo randômico, como ilustra a Figura 37. Dentre 1.881 trechos do elemento SR2 em íntrons, foram observadas 54 inserções.

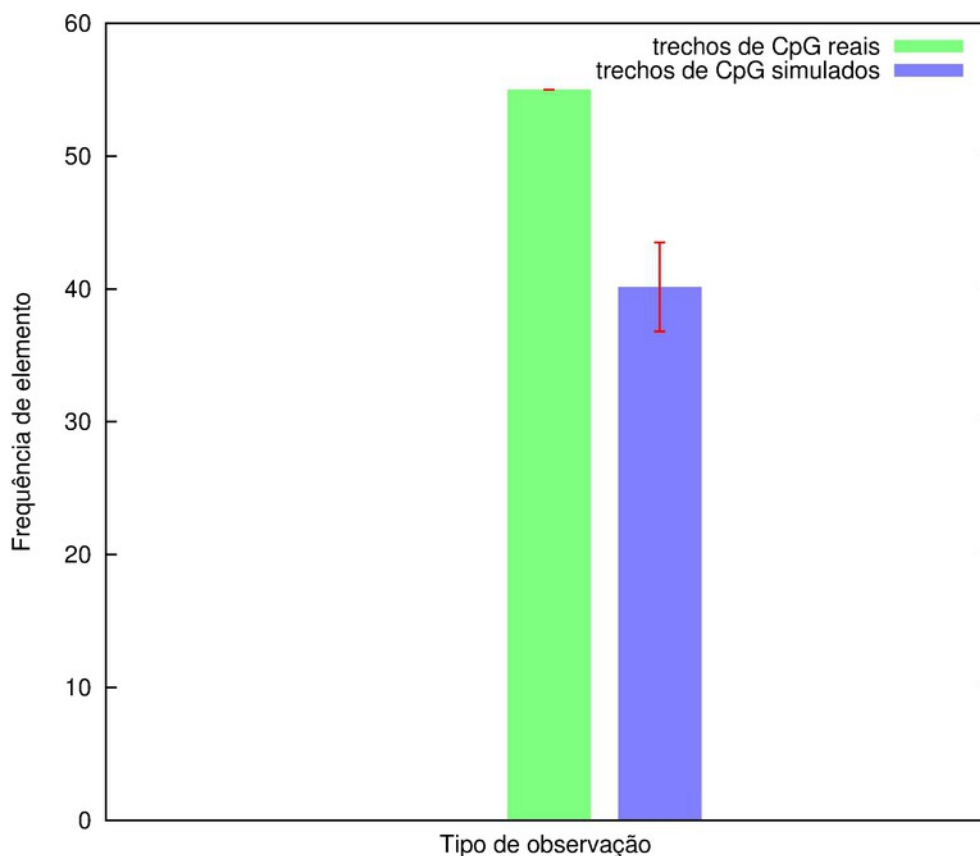


Figura 37-Representação do número de elementos SR2 reais (observados) e simulados (esperados) cujos trechos correspondem a ilhas CpG presentes nos íntrons. Considerando o tamanho e a frequência real dos elementos SR2 nos íntrons de *S. mansoni*, foram realizadas simulações utilizando seleções aleatórias de regiões do elemento SR2 distribuídas no mesmo arranjo observado no dado real. Fonte: Elaborada pela autora.

Para realizar a simulação foram considerados apenas os trechos que não representavam elementos SR2 não autônomos (1.413 elementos). Foram esperadas 39 ocorrências de trechos equivalentes à ilhas CpG com desvio padrão de 3.12. O valor real de 54 inserções é significativamente maior ( $p\text{-valor}=0.0044$ ) do que aquele da simulação (hipótese nula) em um teste de hipótese para uma proporção\*

O enriquecimento observado para regiões com ilhas CpG pode indicar uma

\* Disponível em: <http://stattrek.com/hypothesis-test/proportion.aspx>

potencial seleção positiva destes elementos em íntrons de genes. Tal enriquecimento é surpreendente visto que regiões CpG podem induzir o silenciamento gênico e portanto, espera-se que em grande parte dos casos as inserções de tais elementos possa ser deletéria.

Já foi descrito que motivos CpG estão sujeitos a metilação em *S. mansoni* e que o fenômeno de metilação de DNA é importante na regulação da ovoposição em parasitas adultos.(31) A redução da metilação fez com que ovos fossem produzidos com anomalias. Nesse estudo também são observadas evidências de que as modificações epigenéticas de *S. mansoni* apresentam relação com elementos repetitivos no genoma.

Para finalizar as análises das regiões intrônicas, foi examinado se os genes, contendo elementos transponíveis das famílias em estudo nos íntrons, poderiam apresentar-se vinculados a um determinado componente celular, processo biológico ou função molecular. O resultado dessa análise é apresentado na Figura 38.

genes-com-insercao-unico-P3

Display terms emanating from Gene Ontology

GO ID	Name	NSP	P-Value	Adj. P-Value	Rank	Pop. Count	Study Count
<input checked="" type="checkbox"/> GO:0044699	single-organism process	B	2,36e-05	0,0203	1	452	77
<input type="checkbox"/> GO:0022610	biological adhesion	B	0,000196	0,169	2	24	10
<input type="checkbox"/> GO:0006914	autophagy	B	0,000454	0,391	3	5	4
<input type="checkbox"/> GO:0007155	cell adhesion	B	0,00104	0,897	4	24	10
<input type="checkbox"/> GO:0005509	calcium ion binding	M	0,00138	1,00	5	80	20
<input type="checkbox"/> GO:0030551	cyelic nucleotide binding	M	0,00207	1,00	6	3	3

1 (None) / 1 / 942

Threshold (lower is more important) 0,1000

Browser

molecular transducer activity (GO:0060089)

Parents:

[molecular\\_function](#)

Figura 38-Resultados da análise do programa Ontologizer onde foi verificado o enriquecimento dos genes com o elemento Perere-3 em seu(s) íntron(s) em uma determinada classe dentro da ontologia do Gene Ontology. O conjunto de estudo continha 255 genes e a população 2138. Foi utilizado o método estatístico Parent-Child-Union e o método Bonferroni para correção de erros das múltiplas comparações. Fonte: Elaborada pela autora.

Dos 255 genes com presença do elemento Perere-3 em íntrons, 77 apresentaram concentração em relação a uma classe de processo biológico denominada “*single-organism process*” (Anexo V), sendo definida como um processo biológico relacionado de forma específica ao organismo.(56)

Os genes com elementos SR2 não apresentaram enriquecimento com nenhuma das classes tratadas pelo programa.



# **Capítulo 5**

## **Regiões não traduzidas de genes, elementos cis-regulatórios e regiões intergênicas**

---





## 5 Regiões não traduzidas de genes, elementos cis-regulatórios e regiões intergênicas

### 5.1 Considerações Iniciais

Extremidades não traduzidas de genes (*UnTranslated Region* – UTR) são sequências de DNA localizadas adjacentes a porção codificante, as quais são transcritas e embora não sejam traduzidas, são as principais regiões, em parceria com os íntrons, envolvidas na regulação do gene o qual elas flanqueiam.(57)

A região 5' UTR desempenha essa regulação no início do processo da tradução através de elementos como (57):

- Estrutura *Cap* 5', a qual é adicionada no final do pré-mRNA e é essencial para a tradução da proteína pois serve de sítio para a ligação de diversos fatores de iniciação eucarióticos (*eukaryotic initiation factor-eIF*) e estes por sua vez, promovem a formação do complexo pré-inicial (PIC). Essa estrutura também promove a estabilidade do mRNA pois sua remoção permite a degradação da extremidade 5' do mRNA.

- Formação de estruturas secundárias, que estão relacionadas com a estruturação da região 5' UTR. A maioria das regiões 5' UTR não são altamente estruturadas e quando são, estão relacionadas com genes de desenvolvimento e que são poucos expressos. Essa alta estruturação da região 5' UTR decorre do fato da região apresentar maior comprimento, conteúdo GC elevado e um alto grau de predição para estruturas secundárias.

- Quadro de leitura aberto *upstream* (*Upstream Open Reading Frame* - UORF) ocorre quando há um códon de parada no quadro, seguido de um códon AUG, antes do códon de iniciação principal. Está relacionado com a redução da expressão da proteína.

- Sítios de entrada de ribossomo interno (*Internal Ribosome Entry Sites* -

IRES) são motivos regulatórios de mRNA que permitem o início da tradução através de um mecanismo independente da *Cap*.

Já a região 3' UTR apresenta elementos que geralmente influenciam a regulação do gene na fase pós-transcricional. Apresenta sítios para proteínas regulatórias e também para miRNA. Imediatamente após essa região é inserida a cauda Poli-A. Essa cauda, atua na regulação da expressão gênica mediante o controle do transporte do mRNA do núcleo para o citoplasma, pela estabilidade e pela degradação do mRNA. As múltiplas cópias de 5 nucleotídeos no motivo AUUUU na região 3' UTR permite que ocorra uma estabilidade do mRNA e a expressão do gene é controlada sem alterar a taxa de tradução.(57)

Além das regiões UTR, os dados utilizados nas análises descritas nesse capítulo também podem verificar regiões que contém elementos cis-regulatórios.

Esses elementos são sítios de ligação para proteínas envolvidas na transcrição ou na regulação da expressão gênica. Dentre esses elementos se encontram os promotores centrais, utilizados para posicionar a RNA polimerase (RNAP), responsável pela transcrição dos genes. Outras regiões próximas a esses promotores regularão a transcrição: em procariotos os *operadores e ativadores* estão envolvidos; em eucariotos, as regiões ao redor dos promotores, os potenciadores (*enhancers*), silenciadores e isoladores (*insulators*) estão presentes. (58)

Esses potenciadores, são elementos situados distantes dos TSS (*Transcription Start Site*) e podem ser localizados *upstream* ou *downstream* do sítio de iniciação, nas regiões de íntrons, éxons ou nas UTR's dos genes. Já foram observados, distantes do gene regulado, a uma distância de 10.000bp em *Drosophila* ou 100.000 bp em humanos e ratos.(58)

Com relação aos elementos silenciadores, estes podem ser localizados a *upstream* ou *downstream* do TSS e através da ligação com proteínas repressoras, ou evitando a ligação dos fatores de transcrição, interferem na formação do PIC.

Por sua vez, os elementos isoladores (*insulators*) desempenham a função de bloquear a interação dos potenciadores e dos silenciadores. Há os isoladores situados entre o potenciador e o promotor e os que atuam como barreiras que

evitam a propagação da heterocromatina.(58)

Regiões intergênicas podem ser transcritas e assim como os transcritos das regiões intrônicas, gerar RNAs não codificantes, os quais podem desempenhar papel importante na regulação da expressão gênica.

Em conjunto com as regiões UTR's, os RNAs não codificantes exercem um ajuste fino na regulação da expressão gênica aumentando a complexidade desse sistema regulatório.(57)

Longos RNAs não codificantes intergênicos (*long intergenic non-coding RNA* – lincRNA) são transcritos com mais de 200 bases e são conhecidos por desempenhar inúmeras funções que abrangem desde a regulação de fatores epigenéticos e expressão gênica até a utilização desses transcritos como suporte para complexos de sinalização de proteínas.(59)

Estudos revelam que muitos lincRNA apresentam relação com elementos LTR em humanos. Nesses estudos, aproximadamente 41.9% dos transcritos lincRNA são derivados de TEs.(60,61)

As análises apresentadas a seguir pretendem verificar se os elementos de transposição Perere-3 e SR2 apresentam potencial para influenciar as regiões não traduzidas e intergênicas, as quais flanqueiam os genes ortólogos identificados nesse estudo.

## 5.2 Metodologia

Devido ao fato de predições gênicas, a partir da sequência do genoma, contemplarem apenas as regiões codificantes, não é possível realizar uma análise mais detalhada nas regiões UTRs e promotoras destes genes, devido ao fato de não haver um modo confiável de prever as mesmas. Deste modo, optou-se por realizar o estudo das regiões situadas entre dois genes como uma única entidade, a qual nos referiremos neste trabalho como “regiões intergênicas”, mas que na verdade englobam regiões UTR e promotoras dos genes estudados.

Como descrito anteriormente no item 3.1.1, foram definidos 4 tipos de regiões intergênicas para representar regiões entre genes localizados na mesma fita (5'-3',3'-5') e genes em fitas diferentes (3'-3',5'-5'), como ilustra a Figura 39.

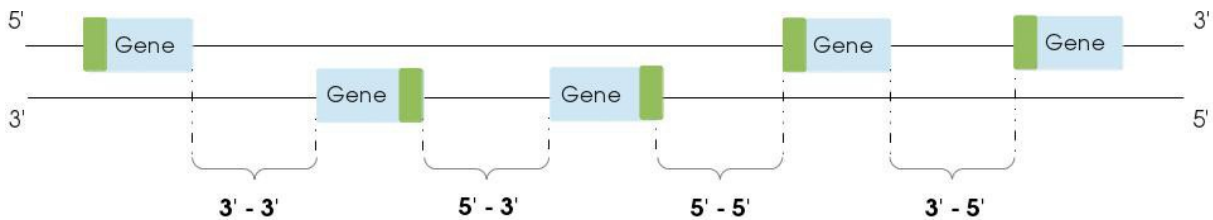


Figura 39-Nomenclatura utilizada para definir as regiões intergênicas identificando regiões com genes localizados na mesma fita ou em fitas diferentes. Os retângulos azuis representam os genes e a extremidade em verde as regiões promotoras desses genes. Fonte: Elaborada pela autora.

Percorrendo de forma sequencial o arquivo que continha a estrutura de dados com a identificação de cada elemento do cromossomo (início do cromossomo, éxons e íntrons de cada gene, regiões intergênicas e fim do cromossomo), foi verificado se dois genes, posicionados no cromossomo de forma adjacentes, ambos eram genes identificados como ortólogos. Nos casos positivos, a região entre os dois genes era selecionada como intergênica.

Um arquivo no formato fasta (62), contendo a sequência de nucleotídeos das regiões intergênicas, foi gerado para verificar a ocorrência de elementos Perere-3 e SR2 nessas regiões.

Utilizando o programa BlastN, foi realizado o alinhamento das sequências dos elementos Perere-3 e SR2 contra as sequências das regiões intergênicas situadas entre os genes ortólogos. Para selecionar os dados mais significativos, foram considerados todos os resultados de alinhamento com identidade superior a 85%, com comprimento maior que 50 bases e com *e-value* inferior a  $10^{-3}$ .

Utilizando o arquivo com as definições de todas as regiões intergênicas entre os genes ortólogos, e o arquivo resultante do alinhamento dessas regiões com os elementos de transposição, foi possível calcular o percentual de regiões intergênicas com retrotransposons. A análise foi realizada individualizando os dados para cada um dos elementos de transposição e para cada tipo de região intergênica.

Também foram analisados os tamanhos das regiões intergênicas com elementos Perere-3, com elementos SR2 e sem os retrotransposons em estudo. Para cada região, pertencente ao conjunto de todas as regiões intergênicas entre os genes ortólogos, foi verificada a ocorrência ou não de alinhamentos com os elementos de transposição. Dessa forma foram selecionadas as regiões com inserção e sem inserção. A distribuição do tamanho dessas regiões foi realizada utilizando a representação gráfica de *boxplot* através de um *script shell* e da linguagem R.

### 5.2.1 Características dos elementos no genoma

Os dados resultantes do alinhamento com os elementos de transposição também permitiu calcular a quantidade de trechos dos retrotransposons nas regiões intergênicas.

Quando o arquivo fasta dessas regiões foi criado, o cabeçalho de cada uma das sequências recebeu identificação única, contendo a descrição do cromossomo ou *scaffold* da região, as coordenadas de início e fim, o tipo de região intergênica, o gene posicionado *upstream* e *downstream* (ex: CABG01000007-16597-97386-intergene55-Smp\_173660-Smp\_094060). Dessa forma, utilizando o comando *awk* e um *script shell*, as regiões com elementos foram filtradas e suas sequências quantificadas. Os dados foram exibidos através de um histograma utilizando o programa *gnuplot*.

Também foram verificados os trechos dos elementos de transposição que mais se inseriram. Para cada inserção resultante do alinhamento das regiões intergênicas com os retrotransposons, as bases inseridas foram acumuladas uma a uma.

Aprimorando as análises com relação aos trechos dos elementos presentes nas regiões intergênicas, foi realizada a predição de possíveis sítios para a ligação de fatores de transcrição nas sequências desses elementos.

Foi utilizado o banco de dados JASPAR \* (63), que disponibiliza interface *on-line* para a predição de sítios de fatores de transcrição. São utilizadas matrizes modelos de diversos organismos, distintas para cada fator de transcrição. Foram utilizados os modelos do organismo *Caenorhabditis elegans*, o qual pertencente ao mesmo filo que o organismo *S. mansoni*, ou seja, *Nematoda*.

Os dados resultantes do processamento do programa JASPAR foram coletados e a partir deles, foi gerado um histograma contendo a frequência para as predições dos sítios de ligação. Utilizando a biblioteca BioGraphics (64), foi gerada uma representação gráfica do posicionamento desses sítios na sequência dos elementos Perere-3 e SR2.

## 5.2.2 Proximidade dos elementos em relação às extremidades 5'

Para verificar o posicionamento das sequências dos retrotransposons em relação as regiões promotoras dos genes, ou seja, extremidade 5', foi definido o posicionamento dos elementos em relação ao comprimento total da região intergênica, como ilustra a Figura 40. Essa região foi dividida em dois segmentos. Um desses segmentos representando a distância entre o início da região intergênica (foi considerada a menor coordenada no genoma) e uma extremidade da inserção, e o outro segmento representando a distância que compreende a outra extremidade da inserção e o término da região intergênica (maior coordenada no genoma). Dividindo um destes intervalos pela soma de ambos, pode-se observar a fração na qual a inserção ocorreu.

Os resultados apresentados consideram a distância entre a extremidade 5' da região intergênica e a extremidade mais próxima da inserção.

Essa distância entre a inserção e a extremidade 5' também foi calculada em bases. Utilizando os dados resultantes do alinhamento foi possível verificar qual era a extremidade 5' da região intergênica e utilizando a coordenada do alinhamento do

---

\* O banco de dados JASPAR CORE contém um conjunto de perfis curados e não-redundantes, derivados de coleções publicadas sobre sítios de fatores de transcrição experimentalmente definidos para eucariotos.

retrotransposon, calcular a distância em bases.

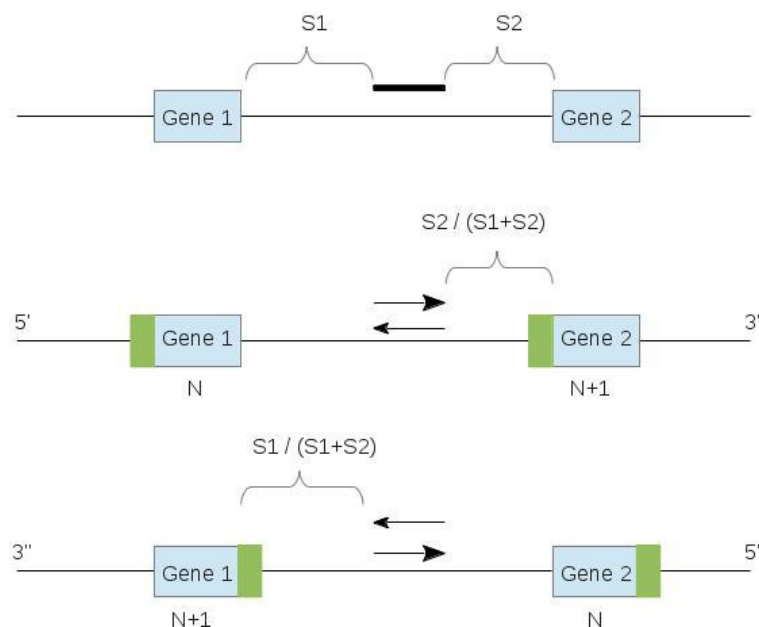


Figura 40-Método de fração utilizado para definir a posição do elemento na região intergênica e a distância na qual o elemento ocorreu da região promotora do gene, ou seja, da extremidade 5'. Os retângulos azuis representam os genes, as setas os elementos de transposição e seu sentido e, a parte verde representa a região promotora dos genes. Fonte: Elaborada pela autora.

### 5.2.3 Inserções de ilhas CpG

A metodologia utilizada para essa análise é a mesma descrita no item 4.2.5 do capítulo 4, realizadas sobre os dados dos elementos nas regiões intrônicas.

Foi verificada a distância desses trechos em relação à extremidade 5' das regiões intergênicas e a distância em bases. Também foram realizadas simulações para verificar a relação entre o valor observado/esperado para esses trechos do elemento SR2.

#### **5.2.4 Enriquecimento dos genes que flanqueiam regiões intergênicas com retrotransposons**

Com o objetivo de identificar uma possível classe de genes mais suscetível a presença dos elementos Perere-3 e SR2, foi realizada uma análise utilizando os recursos de ontologia disponíveis pelos programas Gene Ontology (GO) e Ontologizer.

Em conjunto, essas 2 ferramentas demonstram se é possível traçar um número igual ou maior de anotações para um mesmo termo, quando são comparados o conjunto de genes em estudo e outros conjuntos formados aleatoriamente a partir da população de genes.(48) Nesse caso os genes em estudos foram os que flanqueavam as regiões intergênicas com retrotransposons e a população, o conjunto de todos os genes entre as regiões intergênicas dos genes ortólogos.

O arquivo contendo as anotações que correlacionam os genes de *S. mansoni* com as identificações GO foram obtidos através do *site* do Instituto Sanger. (49)

Em decorrência das regiões com elementos SR2 serem flanqueadas por genes que apresentaram composição celular com enriquecimento para proteínas intrínsecas à membrana, utilizando o programa TMHMM (65) foi realizada análise para verificar a predição das possíveis topologias dessas proteínas de membrana.

### **5.3 Resultados e Discussão**

Foram analisadas 2.088 regiões intergênicas do genoma de *S. mansoni* entre os genes que possuíam ortólogos definidos em *S. japonicum*, sendo que 47% dessas regiões apresentaram presença dos elementos SR2 e/ou Perere-3. Dentre



esse total, aproximadamente 18% das regiões intergênicas apresentaram trechos de ambos os elementos. Em 13,5%, ocorreram apenas trechos dos elementos Perere-3 e, em 15,5% apenas trechos dos elemento SR2. As sequências dos elementos SR2 correspondem a 1,65% das bases dessas regiões, e as dos elementos Perere-3 a 1,86%.

Considerando o fato de que as análises visam examinar os elementos próximos as regiões promotoras ou terminais dos genes, as regiões identificadas como 3'-5' e 5'-3' terão seus resultados considerados como uma única classe. Essas regiões representam dois genes posicionados de forma consecutiva em uma mesma fita, sendo a orientação 3'-5' ou 5'-3' meramente consequência do posicionamento em uma ou outra fita do cromossomo. A diferenciação utilizada na identificação dessas regiões foi necessária em decorrência da metodologia utilizada para organizar os dados iniciais, descrita no capítulo 3.

Verificando o percentual de regiões intergênicas com a presença de elementos Perere-3 ou SR2, aparentemente não é possível definir uma tendência de inserção em uma região específica, como ilustra a Figura 41. As regiões 3'-5' e 5'-3' apresentam um maior percentual de regiões e consequentemente, um maior percentual de elementos de transposição.

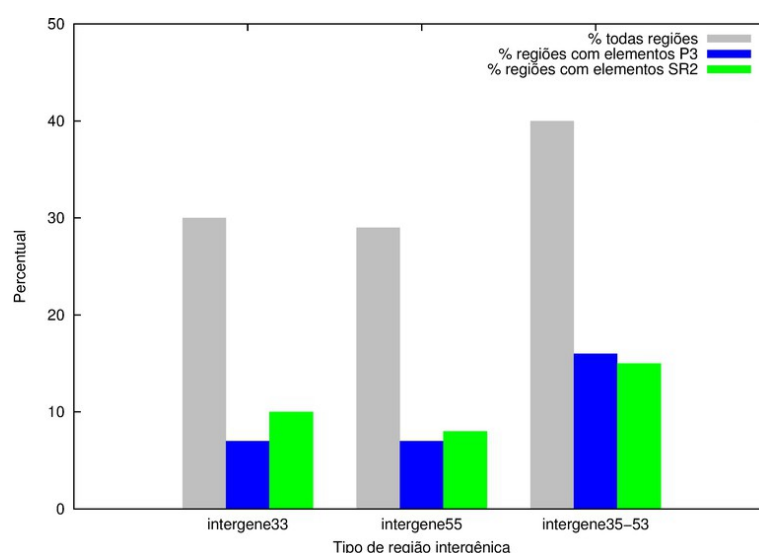


Figura 41-Percentual de regiões intergênicas com elementos Perere-3 (azul) e SR2 (verde) com relação a todas as regiões intergênicas identificadas entre os genes ortólogos (cinza). Fonte: Elaborada pela autora.

Com relação ao tamanho das regiões intergênicas com retrotransposons e sem, é possível observar uma diferença significativa no tamanho das regiões com TEs e sem (teste Wilcoxon –  $p$ -valor  $< 2.2e-16$ ). A Figura 42, ilustra os dados para as todas as regiões intergênicas com elementos Perere-3 (P3), SR2 e sem elementos.

Para essa análise foram consideradas apenas as regiões intergênicas que apresentaram elementos Perere-3 ou SR2. As regiões com ambos os elementos não foram consideradas.

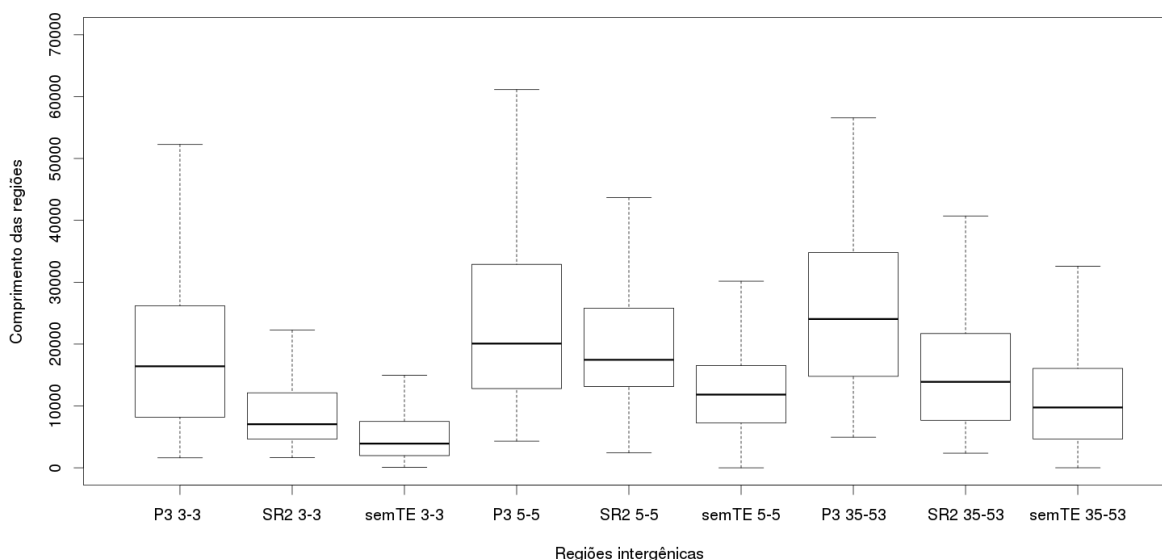


Figura 42-Boxplot representando a distribuição de comprimentos das regiões intergênicas denominadas 3-3,5-5, 3-5 e 5-3, com a presença dos elementos Perere-3 (P3), SR2 e sem retrotransposons. Fonte: Elaborada pela autora.

As regiões com elementos Perere-3, apresentam uma variação mais semelhante de tamanhos de seus trechos quando comparada com a variação dos trechos dos elementos SR2. Também são relativamente maiores que as regiões com a presença dos elementos SR2. Para Perere-3, foram observadas as medianas de 16.413 bases, 20.079 e 24.047, para as regiões 3-3, 5-5, 3-5 e 5-3, respectivamente. As regiões com elementos SR2, para a mesma sequência de regiões intergênicas, apresentam medianas com 7.046 bases, 17.458 e 13.895 bases.

Outro fato que pode ser observado é que as regiões que apresentam

extremidades promotoras (5') apresentam tamanho médio maior do que as regiões que contém apenas extremidades terminais (3'). Essa diferença é mais acentuada para as regiões sem os retrotransposons mas também pode ser observada nas regiões com elementos SR2.

## 5.4 Características dos elementos no genoma

Nas regiões intergênicas contendo elementos Perere-3 ou SR2 predominam ocorrências de um único elemento da família pesquisada (Figura 43 e Figura 44).

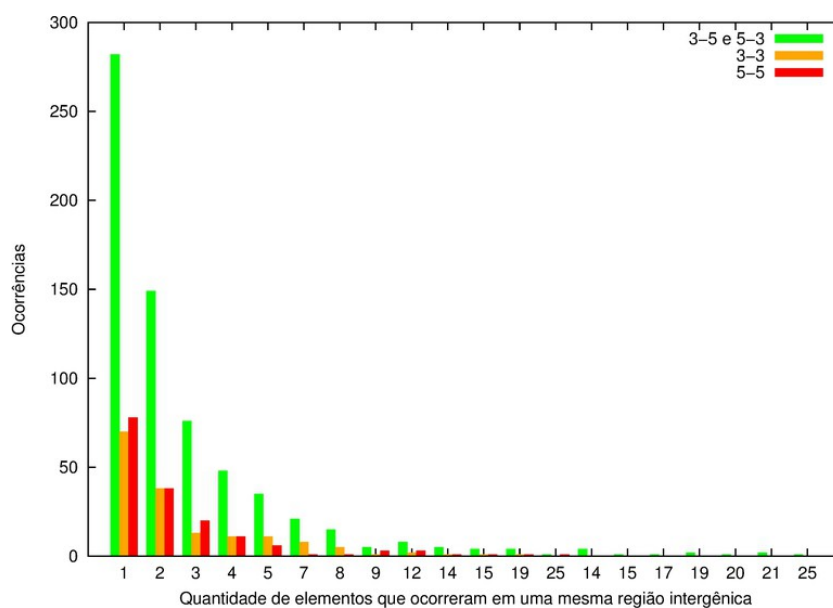


Figura 43-Distribuição da número de ocorrências do elemento Perere-3 em uma mesma região intergênica. Fonte: Elaborada pela autora.

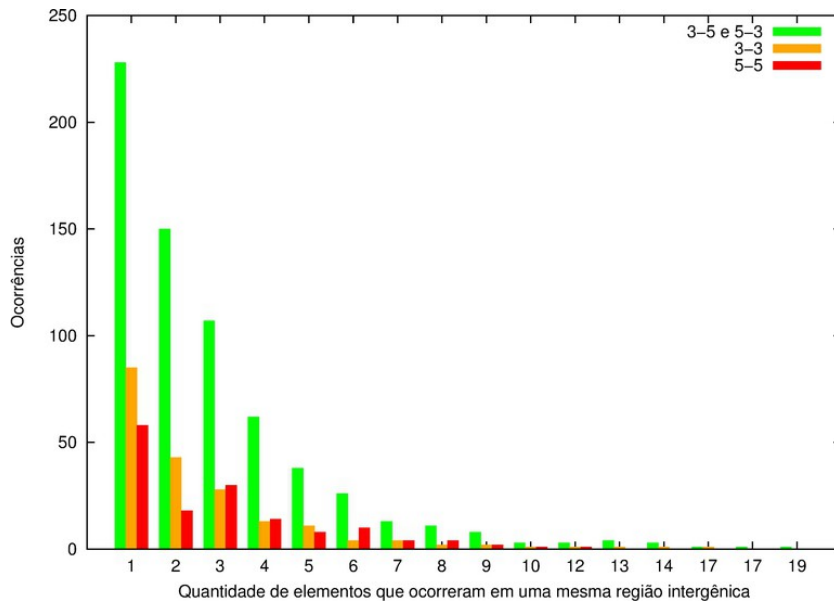


Figura 44-Distribuição da número de ocorrências do elemento SR2 em uma mesma região intergênica. Fonte: Elaborada pela autora.

Foi verificado o tamanho médio dos elementos Perere-3 e SR2, como ilustra a Figura 45.

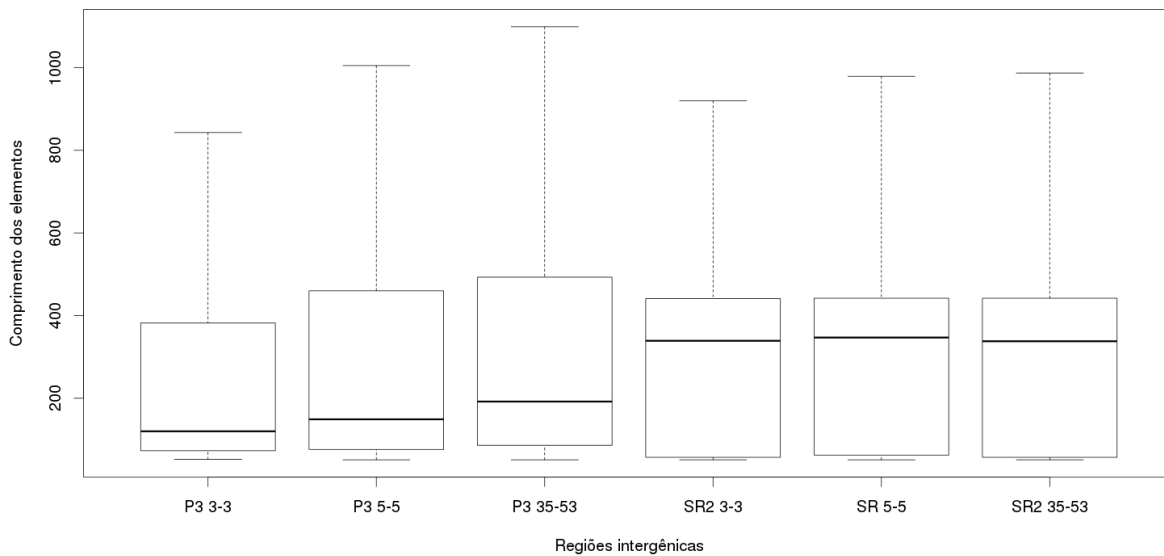


Figura 45-Boxplot ilustrando o comprimento dos elementos Perere-3 (3 primeiros dados) e SR2 (3 últimos dados) nas regiões intergênicas 3-3,5-5, 3-5 e 5-3. Fonte: Elaborada pela autora.

Comparando todos os trechos dos elementos Perere-3 com os dos elementos SR2, é possível detectar uma pequena variação de tamanho que não é estatisticamente relevante.

Para ambos os elementos foram verificadas ocorrências de trechos que equivalem a sequência de cópias mais completa dos elementos, ou seja, trechos com mais de 3.000 bases para o elemento Perere-3 (50 elementos) e mais de 3.300 bases para SR2 (17 elementos). Esses dados correspondem a valores *outline* e não são exibidos no gráfico.

Aparentemente, esses dados apresentam resultado oposto ao observado nas análises das regiões intrônicas onde foram observados mais elementos completos do elemento SR2 em relação ao Perere-3.

Com relação aos trechos dos retrotransposons que apresentam maior ocorrência, ilustrados na Figura 46, o elemento Perere-3 apresenta uma maior concentração na extremidade 3', provavelmente em decorrência da forma como a inserção dos elementos da classe non-LTR se dá no genoma, iniciando-se na ponta 3' do elemento. Devido a baixa processividade da transcriptase reversa, algumas transcrições não são concluídas totalmente, gerando elementos truncados nos quais a região 5' está ausente.

As regiões mais representadas do elemento SR2 provavelmente refletem o grande número de elementos não autônomo, o qual contém os trechos iniciais e finais do elemento autônomo e resultam nos dois picos observados na Figura 46.

Analisando o tamanho das sequências dos elementos SR2, ilustrada na Figura 45, é possível verificar que 50% dos valores observados estão limitados ao teto de aproximadamente 440 bases, que corresponde ao trecho inicial do elemento não autônomo, antes do intervalo de 10 bases com o trecho final.

Utilizando o banco de dados JASPAR, foi realizada a predição de sítios para a ligação de fatores de transcrição, baseados em 15 modelos do organismo *Caenorhabditis elegans*. A frequência desses sítios nas sequências dos elementos Perere-3 e SR2 é ilustrada na Figura 47.

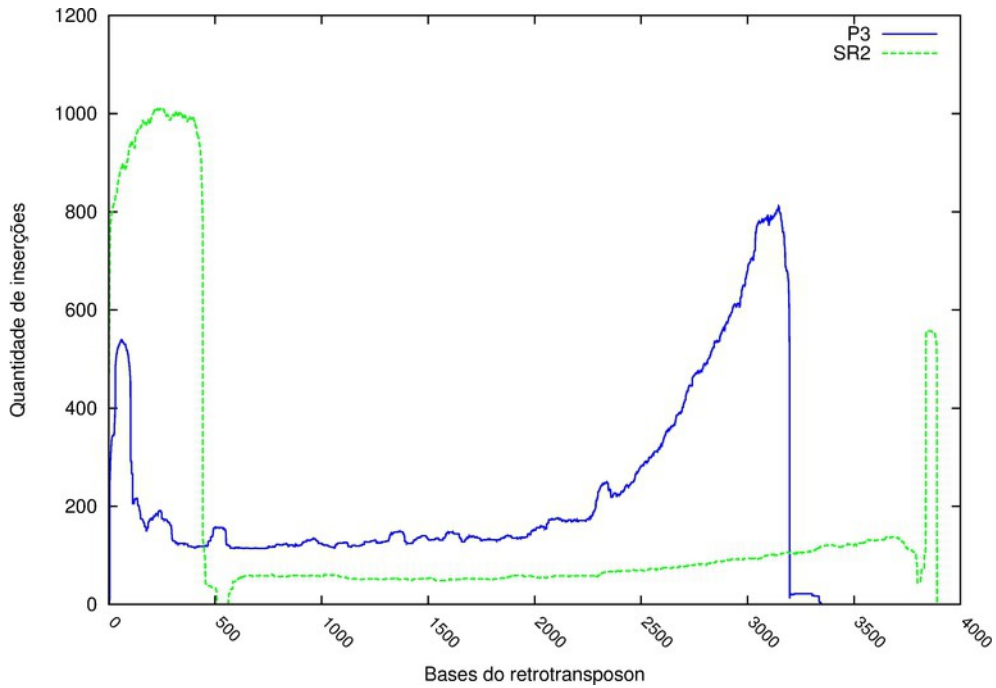


Figura 46-Distribuição das bases dos retrotransposons Perere-3 (azul) e SR2 (verde) inseridas nas regiões intergênicas identificadas entre os genes ortólogos. Fonte: Elaborada pela autora.

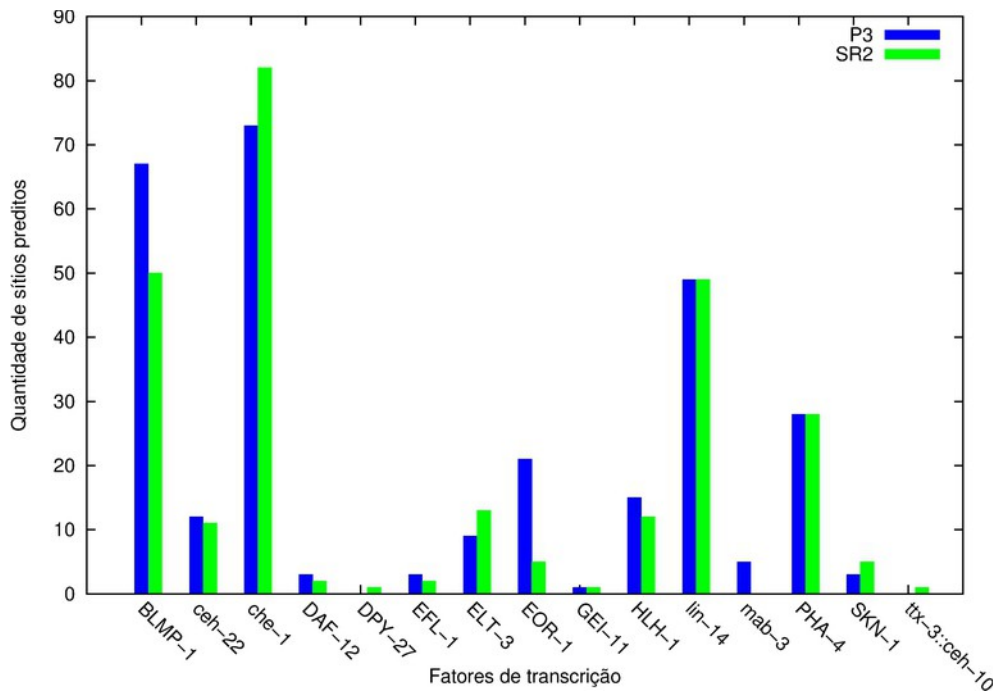


Figura 47-Distribuição da frequência de sítios preditos através do programa Jasper, para a ligação de fatores de transcrição. Fonte: Elaborada pela autora.

Para o trecho inicial do elemento SR2 (1-442 bases), o qual equivale a aproximadamente 80% do elemento SR2 não autônomo, foram preditos 27 sítios para 8 fatores de transcrição distintos ( 5 BLMP-1, 1 ceh-22, 11 che-1, 2 EOR-1, 1 GEI-11, 1 HLH-1, 5 lin-14, 1 PHA-4).

Considerando a mesma proporção de bases inicial que foi utilizada para verificar as predições dos sítios de ligação no elemento SR2 (1-440), em Perere-3 ocorrem a predição de 31 sítios (10 BLMP-1, 11 che-1, 1 EFL-1, 1 ELT-3, 1 HLH-1, 5 lin-14, 1 PHA-4, 1 SKN-1). Observando a região do elemento que apresentou maior frequência no genoma, ou seja, a região terminal, sendo considerada nessa análise o trecho acima das 3.000 bases, são preditos 59 sítios ( 23 BLMP-1, 1 ceh-22, 7 che-1, 1 ELT-3, 8 EOR-1, 1 GEI-11, 1 HLH-1, 9 lin-14, 1 mab-3, 7 PHA-4).

O posicionamento dos demais sítios preditos, tanto para Perere-3 como para o elemento SR2 estão ilustrados na Figura 48 e Figura 49, respectivamente.

Diferente das demais predições, para o fator de transcrição GEI-11, foi predito apenas um sítio de ligação, em cada um dos elementos. Curiosamente, a predição para esse sítio está nas regiões onde os elementos mais apresentaram trechos no genoma, ou seja, região 3' para o elemento Perere-3 e região 5' para o elemento SR2 (Figura 48 e Figura 49).

Para investigar o provável motivo dessa especificidade resultante da predição, foi pesquisado e encontrado na literatura que esse fator está envolvido preferencialmente com a regulação de RNAs não-codificantes.(66) GEI-11 é ortólogo do componente SNAP190 de ligação ao DNA, do complexo SNAPc, necessário para que a RNA polimerase II e III realize a transcrição de genes snRNA (*small nuclear RiboNucleic Acid*). (67) Pode ser viável verificar se essa funcionalidade pode apresentar alguma correlação com os elementos de transposição, principalmente com os elementos não autônomos que não codificam nenhum tipo de proteína. Na literatura é descrito que metade dos transcritos derivados de TEs permanecem no núcleo, indicando que podem desempenhar funções de RNA não codificantes.(68)

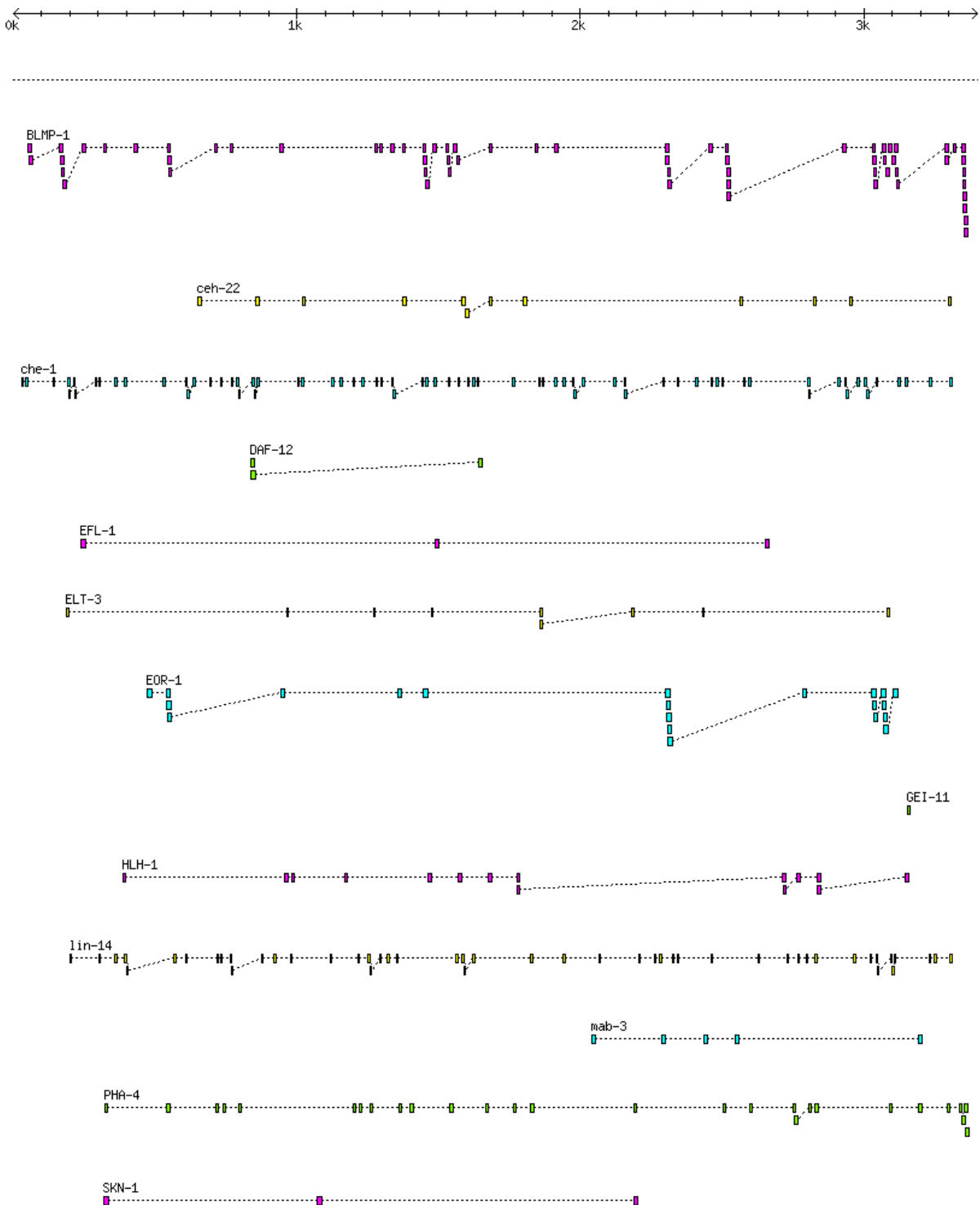


Figura 48-Posicionamento dos sítios preditos para a ligação de fatores de transcrição na sequência do elemento Perere-3. Fonte: Elaborada pela autora.



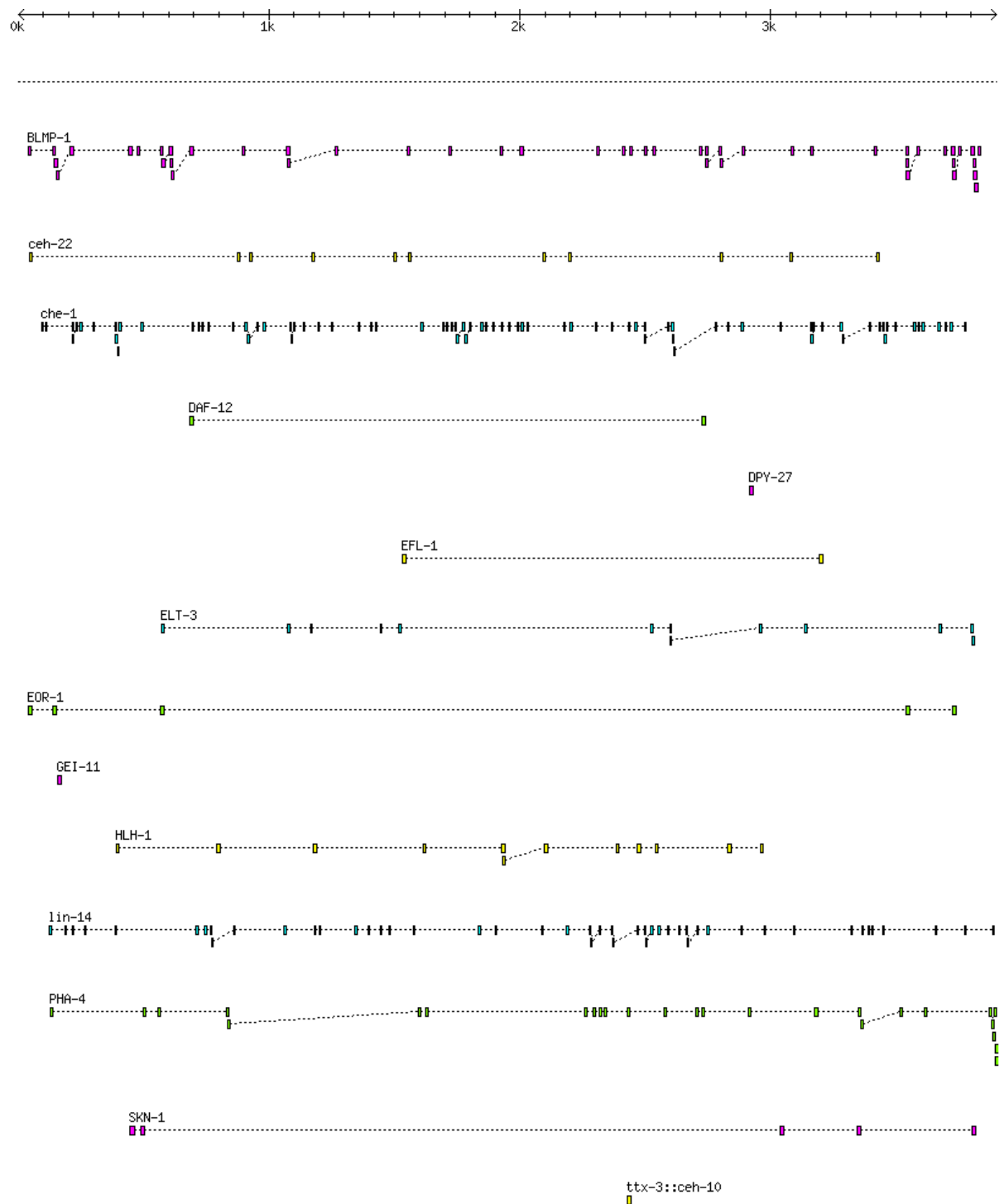


Figura 49-Posicionamento dos sítios preditos para a ligação de fatores de transcrição na sequência do elemento SR2. Fonte: Elaborada pela autora.

## 5.5 Proximidade das extremidades 5'

A análise da distância dos elementos em relação as extremidades das porções codificadoras dos genes, permitiu determinar que elementos SR2 apresentam uma pequena tendência a estarem mais próximos das extremidades 5', enquanto que para os elementos Perere-3 tal tendência não é observada, como ilustram os gráficos Figura 50 e Figura 51.

Para detalhar mais esses resultados, foram verificadas as distâncias dos elementos, considerando as bases das sequências. A Figura 52 ilustra os resultados para os elementos Perere-3 e para os elementos SR2.

Comparando todas as distâncias dos elementos Perere-3 com as distâncias dos elementos SR2, foi possível determinar uma significância estatística para a variação da distância entre esses conjuntos de dados (teste Wilcoxon –  $p$ -valor < 2.2e-16).

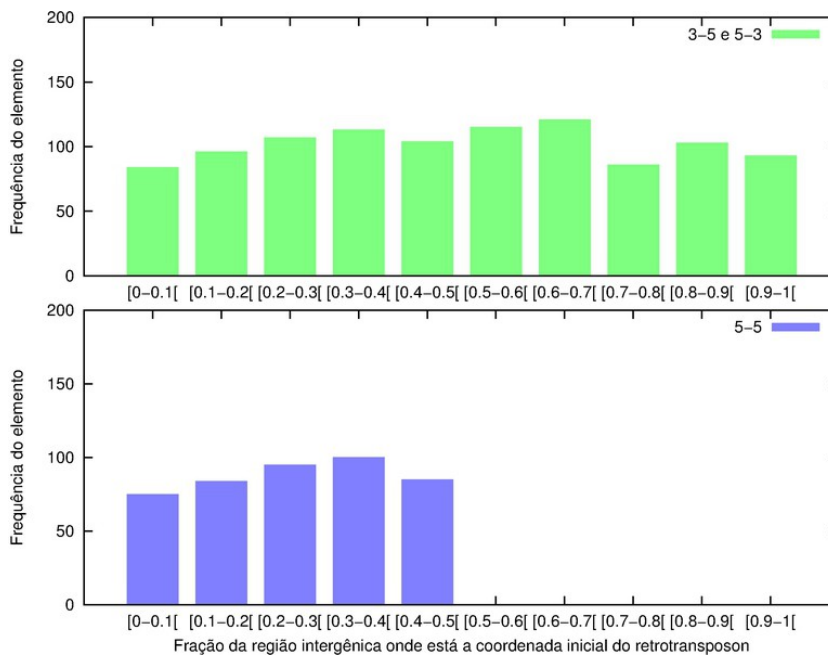


Figura 50-Distribuição e proximidade das extremidades 5' das regiões intergênicas 5'-3' e 3'-5' (verde) e 5'-5' (azul) dos elementos Perere-3. Fonte: Elaborada pela autora.

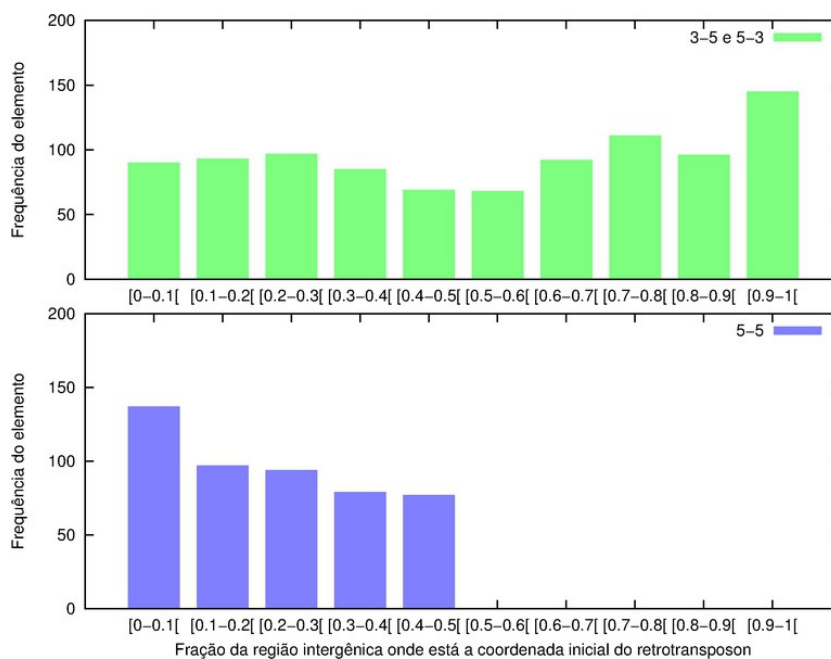


Figura 51-Distribuição e proximidade das extremidades 5' das regiões intergênicas 5'-3' e 3'-5' (verde) e 5'-5' (azul) dos elementos SR2. Fonte: Elaborada pela autora.

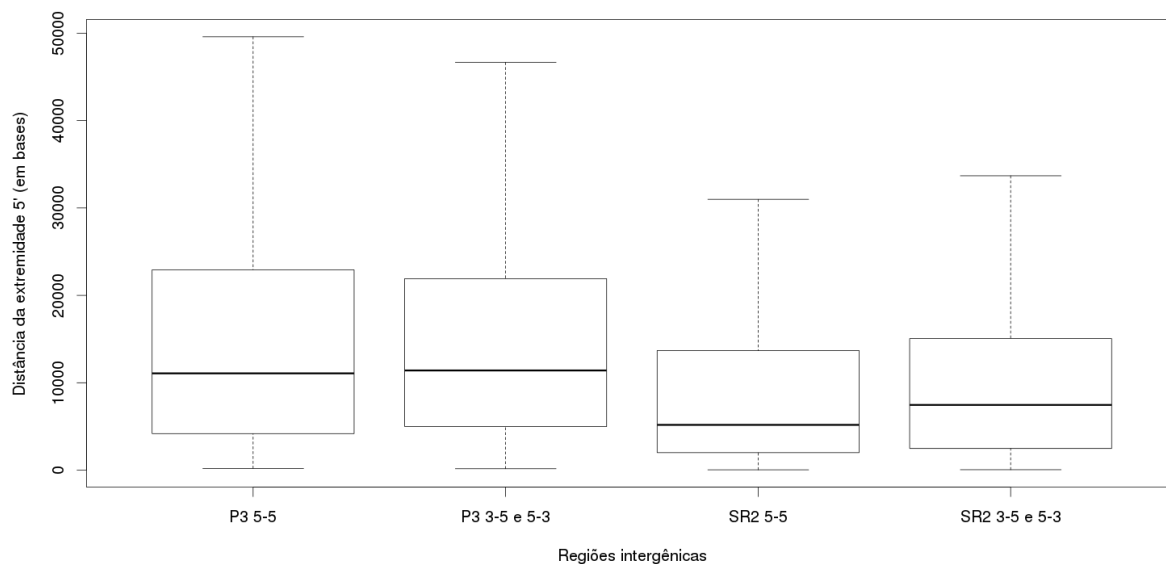


Figura 52-Boxplot ilustrando a distância, em bases, na qual se encontram os elementos Perere-3 (2 primeiros dados – medianas de 11.078 e 11.409) e SR2 (2 últimos dados – medianas de 5.186 e 7.465) das extremidades 5' nas regiões intergênicas 5-5, 3-5 e 5-3, respectivamente. Fonte: Elaborada pela autora.

Os dados revelam que o elemento SR2 está inserindo trechos com uma proximidade maior das extremidades 5' dos genes que flanqueiam as regiões intergênicas, quando essas distâncias são comparadas com as dos elementos Perere-3. Foi observada mediana de 11.078 bases e 11.409 bases para as regiões 5-5 e 3-5/5-3 para o elemento Perere-3 e, para SR2, mediana de 5.186 bases e 7.465 bases para a mesma sequência de regiões.

Esses resultados, em conjunto com a predição dos sítios para a ligação dos fatores de transcrição descritas anteriormente, sugerem que os elementos de transposição em estudo, apresentam características que podem interferir no processo de transcrição dos genes. Na literatura, um estudo com ênfase em humanos, descreve que a frequência observada é menor do que a esperada para as inserções de elementos LINES e LTR próximas ou internas às unidades transcricionais.(69)

## **5.6 Inserções de ilhas CpG**

Dos 1.990 trechos do elemento SR2 em regiões intergênicas, 50 são equivalentes a trechos de ilhas CpG. Para esses trechos, foi verificado em que tipo de região intergênica estavam ocorrendo, a distância em relação as extremidades 5' e a possibilidade desses trechos estarem ocorrendo de forma aleatória.

Como há uma maior quantidade de regiões intergênicas que contém extremidades promotoras, os trechos de ilhas CpG ocorreram em maior quantidade nessas regiões, como ilustra a Figura 53.

Para a distância média em bases desses trechos de ilhas CpG em relação à extremidade 5' da região intergênica, ilustradas na Figura 54, para a região 5-5, foi obtida mediana de 4.371 bases e para a região 3-5/5-3 mediana de 5.576 bases.

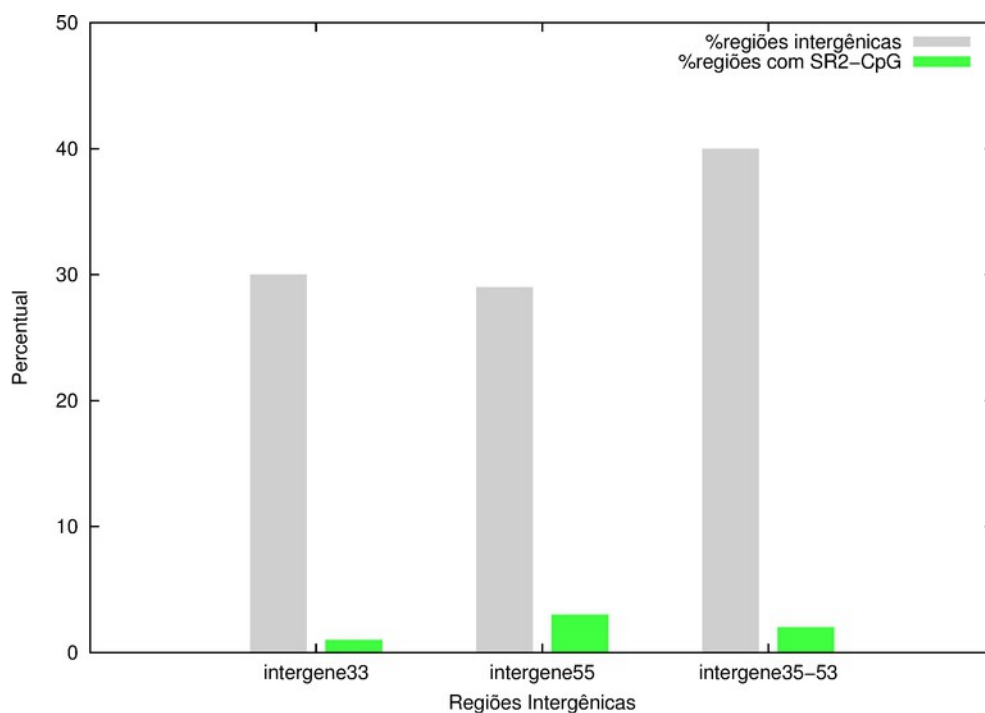


Figura 53-Percentual das regiões intergênicas entre os genes ortólogos (cinza) e percentual das regiões com trechos de SR2 que equivalem à ilhas CpG (verde). Fonte: Elaborada pela autora.

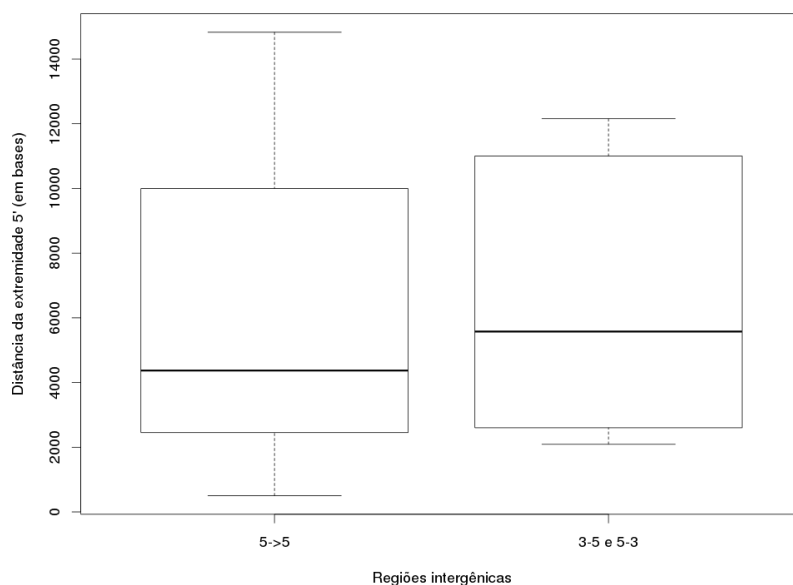


Figura 54-Boxplot ilustrando a distância, em bases, na qual se encontram os trechos dos elementos SR2 equivalentes à ilhas CpG, em relação a extremidade 5' das regiões intergênicas (medianas de 4.371 e 5.576, respectivamente). Fonte: Elaborada pela autora.

Esses trechos equivalentes às ilhas CpG podem contribuir para a constituição de promotores centrais.(58) Esses promotores, além de apresentarem TSS muito amplos ou em regiões dispersas com 100-500 bp, apresentam uma super representação de ilhas CpG.

Para verificar se a inserção desses trechos de ilhas CpG estavam ocorrendo de forma aleatória, foram realizadas simulações onde o valor estimado foi de 55 inserções, com desvio padrão de 3.3, contra o valor observado de 50 inserções, como ilustrado na Figura 55.

O valor real de 50 inserções não é significativamente diferente ( $p$ -valor=0.5684) do que aquele da simulação (hipótese nula) em um teste de hipótese para uma proporção.

Possivelmente há uma pressão seletiva contra a inserção desses trechos de ilhas CpG. Esse tipo de sequência apresenta forte relação com fatores epigenéticos e podem produzir o silenciamento gênico.(46)

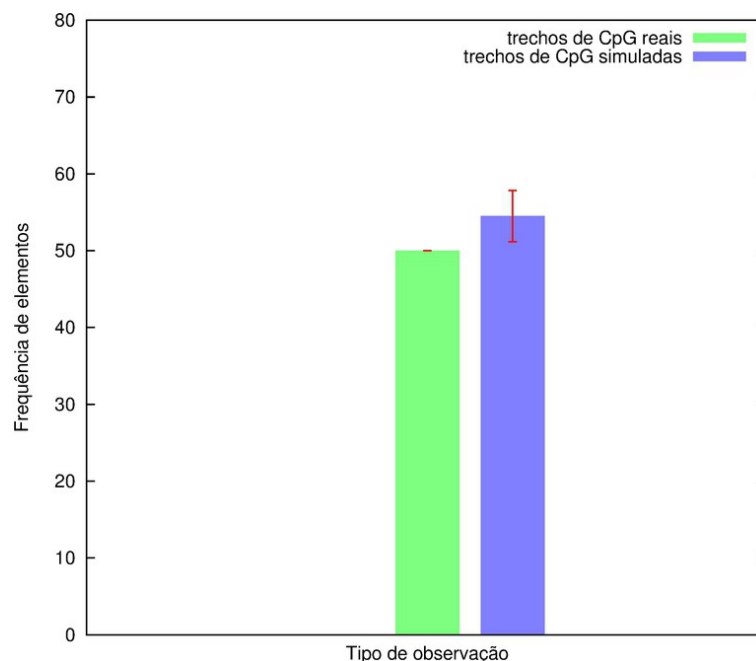


Figura 55-Representação do número de elementos SR2 reais (observados) e simulados (esperados) cujos trechos correspondem a ilhas CpG presentes nas regiões intergênicas. Considerando o tamanho e a frequência real dos elementos SR2 em *S. mansoni*, foram realizadas simulações utilizando seleções aleatórias de regiões do elemento SR2 distribuídas no mesmo arranjo observado no dado real. Fonte: Elaborada pela autora.

## 5.7 Enriquecimento dos genes que flanqueiam as regiões intergênicas com retrotransposons

As análises para verificar uma possível relação entre genes adjacentes, que apresentam elementos de transposição na região intergênica, e um termo GO, revelou um enriquecimento de determinados termos para genes associados ao elemento Perere-3, como ilustra a Figura 56.

Ontologizer - Results for todosGenesIntergeneRegion-insertion-P3 (Parent-Child-Union/Bonferroni)

todosGenesIntergeneRegion-insertion-P3

Display terms emanating from Gene Ontology

GO ID	Name	NSP	P-Value	Adj. P-Value	Rank	Pop. Count	Study Count
<input checked="" type="checkbox"/> GO:0019438	aromatic compound biosynthetic proc	B	3,17e-05	0,0476	1	258	126
<input checked="" type="checkbox"/> GO:0034654	nucleobase-containing compound bios	B	3,76e-05	0,0565	2	250	125
<input checked="" type="checkbox"/> GO:0018130	heterocycle biosynthetic process	B	4,34e-05	0,0653	3	262	127
<input checked="" type="checkbox"/> GO:0032774	RNA biosynthetic process	B	5,63e-05	0,0846	4	205	106
<input type="checkbox"/> GO:0006351	transcription, DNA-dependent	B	7,99e-05	0,120	5	205	106
<input type="checkbox"/> GO:0001262	aromatic compound biosynthetic	B	8,20e-05	0,126	6	267	127

4 (None) / 4 / 1685 Threshold (lower is more important) 0,1000

Browser

RNA biosynthetic process (GO:0032774)

Parents:

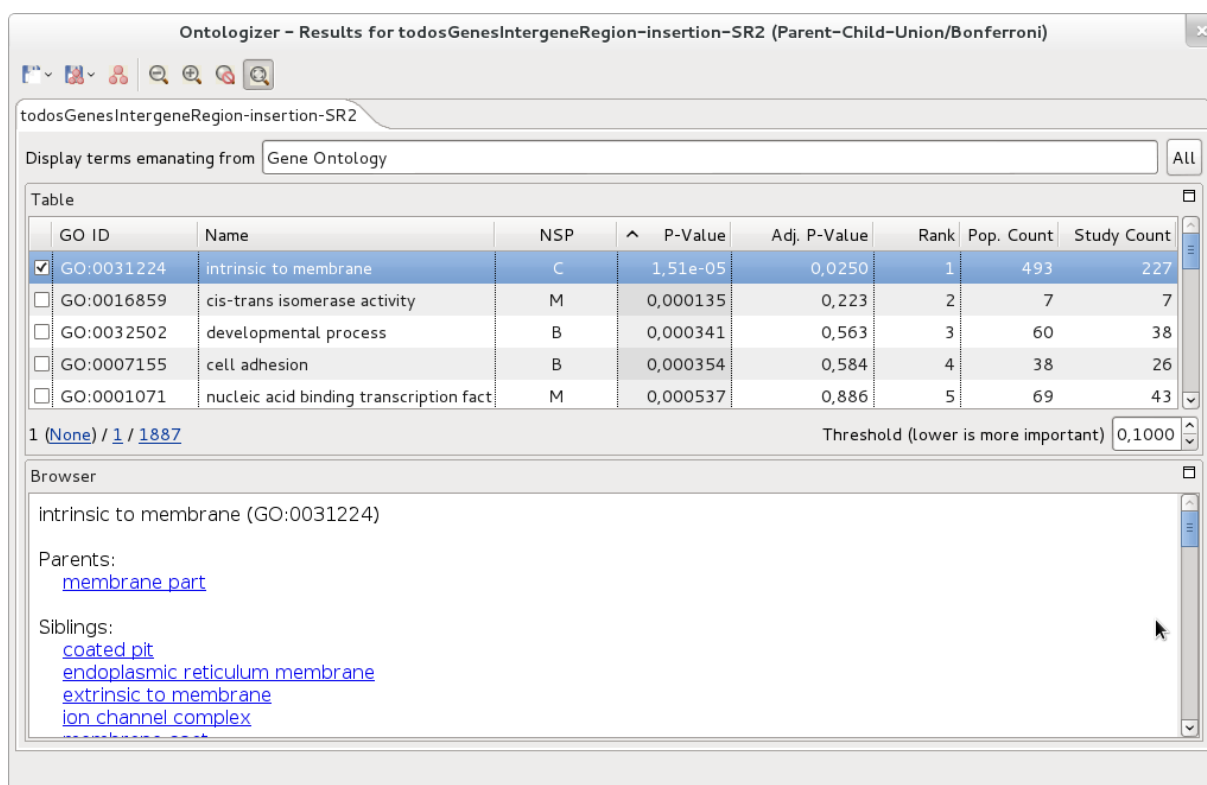
- [RNA metabolic process](#)
- [macromolecule biosynthetic process](#)
- [nucleobase-containing compound biosynthetic process\(\\*\)](#)

Figura 56-Resultados da análise do programa Ontologizer onde foi verificado o enriquecimento dos genes que flanqueiam as regiões com elementos Perere-3 em uma determinada classe dentro da ontologia do Gene Ontology. O conjunto de estudo continha 1063 genes e a população 2500. Foi utilizado o método estatístico Parent-Child-Union e o método Bonferroni para correção de erros das múltiplas comparações. Fonte: Elaborada pela autora.

Do conjunto de estudo composto por 1.063 genes, 126 apresentaram enriquecimento com relação ao processo biológico denominado “*aromatic compound*”

*biosynthetic proc*" (Anexo VI). Esse termo GO é definido como um processo biológico que está relacionado com as reações químicas e vias metabólicas (*pathways*)\*, que resultam na formação de compostos aromáticos, qualquer substância que contenha um anel de carbono aromático.(56)

Com relação aos genes flanqueando as regiões intergênicas com elementos SR2, dentre o conjunto de estudo com 1059 genes, 227 genes resultaram em enriquecimento com relação a componente celular intrínseco à membrana, como ilustra a Figura 57. Utilizando o programa TMHMM foi realizada análise para verificar a predição das possíveis topologias dessas proteínas de membrana, resultantes da transcrição/tradução dos genes que apresentaram enriquecimento.



Ontologizer – Results for todosGenesIntergeneRegion-insertion-SR2 (Parent-Child-Union/Bonferroni)

Display terms emanating from

<input type="checkbox"/>	GO ID	Name	NSP	^	P-Value	Adj. P-Value	Rank	Pop. Count	Study Count
<input checked="" type="checkbox"/>	GO:0031224	intrinsic to membrane	C		1,51e-05	0,0250	1	493	227
<input type="checkbox"/>	GO:0016859	cis-trans isomerase activity	M		0,000135	0,223	2	7	7
<input type="checkbox"/>	GO:0032502	developmental process	B		0,000341	0,563	3	60	38
<input type="checkbox"/>	GO:0007155	cell adhesion	B		0,000354	0,584	4	38	26
<input type="checkbox"/>	GO:0001071	nucleic acid binding transcription fact	M		0,000537	0,886	5	69	43

1 (None) / 1 / 1887 Threshold (lower is more important) 0,1000

Browser

intrinsic to membrane (GO:0031224)

Parents:

- [membrane part](#)

Siblings:

- [coated pit](#)
- [endoplasmic reticulum membrane](#)
- [extrinsic to membrane](#)
- [ion channel complex](#)
- [membrane part](#)

Figura 57-Resultados da análise do programa Ontologizer onde foi verificado o enriquecimento dos genes que flanqueiam as regiões com elementos SR2 em uma determinada classe dentro da ontologia do Gene Ontology. O conjunto de estudo continha 1059 genes e a população 2500. Foi utilizado o método estatístico Parent-Child-Union e o método Bonferroni para correção de erros das múltiplas comparações. Fonte: Elaborada pela autora.

\* Uma via metabólica é uma série de reações químicas onde uma reação fornece o substrato da reação seguinte sendo a reação seguinte dependente da anterior.



Dos 227 genes, 160 apresentaram pelo menos uma predição de hélice transmembranar e desses 160, 50 apresentaram predição para mais do que 5 hélices transmembranares (Anexo VII), como ilustra a Figura 58.

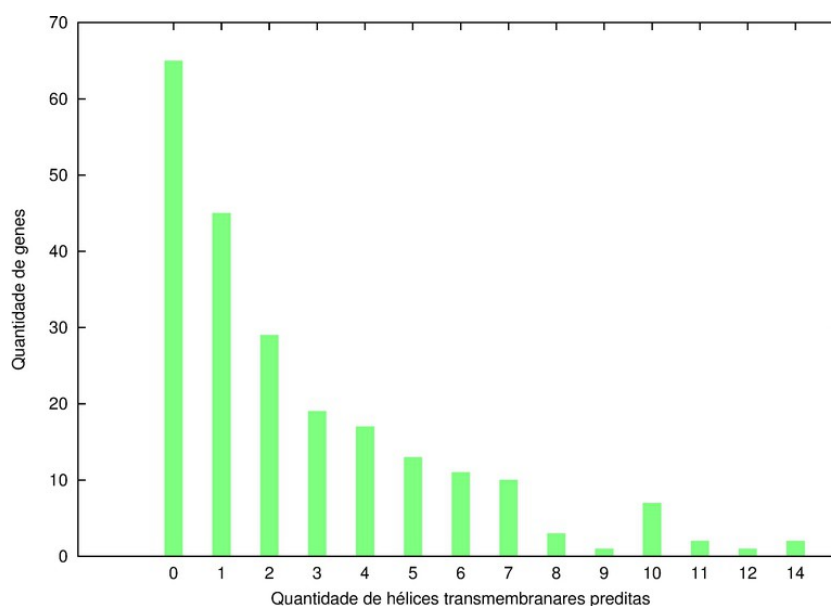


Figura 58-Distribuição da quantidade de hélices transmembranares previstas através do programa TMHMM para os 227 genes, com elementos SR2 e que apresentaram enriquecimento com relação a componente celular de membrana. Fonte: Elaborada pela autora.

Se uma parte destas proteínas transmembranares estiver localizada na interface entre o parasita e o hospedeiro, provavelmente esses genes estarão sujeitos a uma forte pressão seletiva. Neste sentido, o enriquecimento destes genes com sequências de um elemento de transposição favoreceria o desenvolvimento de circuitos genéticos mais complexos, o que permitiria uma resposta mais sofisticada a estímulos do hospedeiro e favoreceria o processo de parasitismo.

Em eucariotos os promotores são extremamente diversos e difíceis de serem caracterizados. Eles tipicamente encontram-se em sentido *upstream* dos genes e podem ter elementos regulatórios posicionados bem distantes (várias kbases) do sítio de início de transcrição. O complexo de transcrição pode promover a dobra do DNA em torno de si mesmo e possibilitar a interação entre os elementos regulatórios e os sítios de transcrição.

Nossas análises indicam que os elementos se distribuem de forma relativamente homogênea nas regiões intergênicas, incluindo regiões que podem se sobrepor as regiões promotoras e de UTRs e que não foram definidas no processo de predição gênica. Nos trechos inseridos, foram preditos sítios para a ligação de fatores de transcrição.

Observou-se também que algumas inserções que correspondem a trechos de ilhas CpG ocorreram em uma frequência menor do que seria esperado nas simulações, sugerindo que essas inserções podem estar interferindo na transcrição dos genes e conseqüentemente sofrendo pressão seletiva.

## **Capítulo 6**

### **Análises das regiões presentes em mRNA**

---



## 6 Análises nas regiões presentes em mRNAs

### 6.1 Considerações Iniciais

Para que a sequência genômica seja traduzida e dê origem as proteínas, tão essenciais aos organismos, os RNAs desempenham um papel fundamental nesse processo.(70) Mais especificamente, três formas distintas de RNA atuam nesse processo:

1) O RNA mensageiro (mRNA), que contém o trecho de nucleotídeos, que é transcrito no núcleo, a partir da sequência do DNA, transportado até o citoplasma para se associar aos ribossomos da célula e ser sintetizado;

2) O RNA de transporte (tRNA), que carrega os aminoácidos até os ribossomos para que os mesmos sejam incluídos nas cadeias de peptídeos que são sintetizadas e,

3) O RNA ribossomal (rRNA), que em conjunto com outras proteínas, formam a constituição riboproteica dos ribossomos.

As sequências codificantes originadas da transcrição do DNA, ou seja, o mRNA precursor (pré-mRNA), passam pelo processo denominado RNA *splicing*. Nessa fase, os íntrons são removidos e os éxons são unidos em uma sequência contínua. Para isso, proteínas específicas identificam sítios de ligação nas bordas dos íntrons e dessa forma definem as regiões que serão clivadas. Essas proteínas atuam como sinalizadoras que conduzem os *small nuclear riboproteins* (snRNPs) para formar a maquinaria de *splicing* denominada spliceossomo. O spliceossomo promove a aproximação das extremidades dos éxons para que a clivagem seja realizada. Dessa forma, os íntron de um mRNA são removidos e os éxons unidos.

Novos éxons podem surgir através de um processo denominado exonização. Em proporções decrescentes, esse processo pode contribuir para o enriquecimento do transcriptoma em mamíferos, vertebrados e em invertebrados. (71)

Para que ocorra a exonização é necessário que um novo sítio de *splicing* seja formado e, em mamíferos, o processo de exonização foi observado como resultado da atuação do elemento de transposição ALU.(55,72,73)

Para os genes que sofreram o processo de exonização, foram observadas novas funções sem que a função original do gene fosse alterada. (55,74) Também foram observados casos onde a exonização apresentou caráter deletério e outros casos, onde as inserções em UTR, não afetaram a proteína resultante, embora essas inserções possam atuar na regulação da expressão gênica.(55,71,75)

Considerando a quantidade crescente de TEs nas regiões intrônicas de invertebrados, vertebrados e de mamíferos, é sugerido que o tamanho dessas regiões também influencia a taxa de exonização, que também é observada de forma crescente entre esses organismos.(71)

Nos estudos realizados por Sela e colaboradores (71), foi observado que além do processo de exonização, a atuação dos TEs também pode resultar em inserções nos primeiros e últimos éxons promovendo a alteração dos mRNAs. Foi observado que em mamíferos os últimos éxons apresentam comprimento maior do que os de vertebrados e maior ainda do que os de invertebrados. Esse maior comprimento está correlacionado com a maior presença de TEs nesses éxons sugerindo que a presença de muitos TEs nessa região por ter levado ao um maior nível de regulação em organismos superiores.

As análises apresentadas a seguir visam investigar como os retrotransposons Perere-3 e SR2 podem estar interferindo nos transcritos dos genes de *S. mansoni*.

## 6.2 Metodologia

O banco de ESTs\* (*Expressed Sequence Tags*) de *S. mansoni* do NCBI, disponibiliza aproximadamente 206.000 sequências que retratam os mRNAs

---

\* EST são sequências *single-pass* parciais de cada extremidade de um clone de cDNA desenvolvida para permitir a rápida identificação de genes expressos através da análise das sequências.

transcritos em diversos ciclos de vida do parasita.

Utilizando essas sequências e o *software* BlastN, com *e-value* de  $10^{-3}$ , foi realizado o alinhamento das sequências dos retrotransposons Perere-3 e SR2 para identificar a presença de sequências derivadas desses elementos de transposição nos transcritos de *S. mansoni*. Foram considerados os alinhamentos que apresentaram pelo menos 50 bases de similaridade com os TEs.

Para as ESTs que apresentaram alinhamento com os retrotransposons, o percentual de cobertura da sequência derivada do TE nas ESTs foi calculado. Foram considerados os resultados com percentual inferior a 100% e cuja região sem mascaramento apresentava mais do que 30 bases. Essas ESTs foram alinhadas contra o genoma de *S. mansoni* após mascaramento das regiões equivalentes aos retrotransposons. Esse alinhamento permitiu avaliar se a região da EST, adjacente ao retrotransposon no transcrito, também não representava uma região repetitiva.

Para selecionar apenas os resultados que representavam sequências não repetitivas, ou seja, de natureza diversa dos elementos de transposição, foram considerados os trechos de ESTs que alinharam no máximo com 3 cromossomos/*supercontigs* distintos do genoma e com *e-value* inferior a  $10^{-3}$ .

Essas coordenadas do genoma foram resgatadas e 3.000 bases, em cada uma de suas extremidades, foram acrescentadas. O trecho da EST que apresentou similaridade com os TEs foi alinhado contra esse intervalo do genoma.

Dessa forma foram identificadas as coordenadas do genoma onde se encontravam o trecho da EST com sequência similar ao dos retrotransposons, e o outro trecho da EST, sem similaridade com os TEs.

As ESTs resultantes das análises descritas anteriormente foram comparadas com o banco de TCs\* (*Tentative Consensus*) do *Gene Indices (76)* com o objetivo de se obter sequências mais completas dos transcritos. Também foram identificadas as coordenadas das TCs equivalentes aos trechos similares as sequências dos TEs.

Utilizando o módulo *find-orfs* do *software* Ugene (77), foram preditas as ORFs dessas TCs. Para essa predição, foi considerado comprimento mínimo de 120

---

\* TCs são criadas pela união de EST em transcrições virtuais. Em alguns casos, as TCs contêm sequências de cDNA completas ou parciais (ETs – contém 5'UTR e 3'UTR), obtidos por métodos clássicos. Formas de *splicing* alternativos são construídos em TCs separadas.

bases e identificação do códon de início. Com base nessas coordenadas foi possível estimar se os elementos em estudo estavam ocorrendo em regiões codificantes ou nas regiões de UTR. A Figura 59 apresenta um resumo, de forma ilustrativa, das etapas da metodologia descrita neste item.

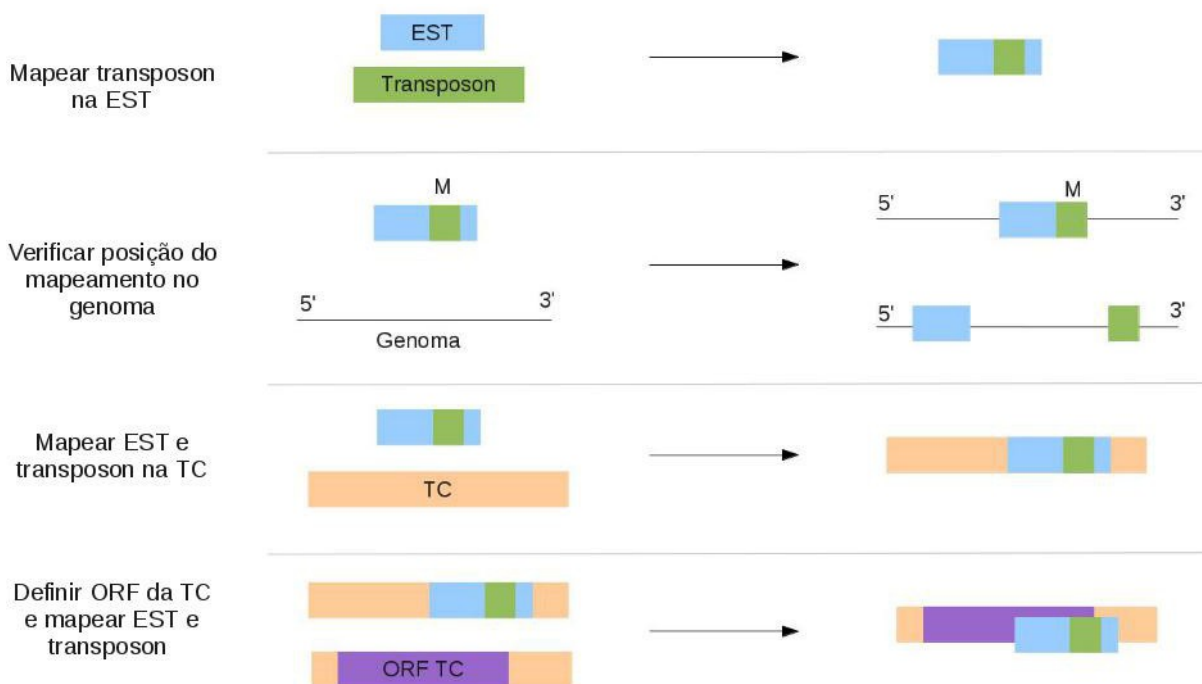


Figura 59-Ilustração das etapas implementadas para análise dos elementos de transposição nas regiões de mRNAs. Os retângulos em azul representam as sequências de ESTs e os verdes as sequências dos retrotransposons. A letra M sobre o retângulo verde indica quando a região estava mascarada ou não. O retângulo salmão representa a sequência de TC e o lilás a ORF da TC. Fonte: Elaborada pela autora.

Com o objetivo de identificar a possível função desses genes, as TCs resultantes foram alinhadas contra o banco NR (*non-redundant*) de proteínas do NCBI.(78) Foi utilizado o *software* BlastP considerando resultados com *e-value* inferiores a  $10^{-3}$ . As similaridades com as proteínas transcriptase reversa e poli-proteínas foram desconsideradas por serem proteínas típicas de retrotransposons.

Foram verificadas as regiões dos elementos de transposição que mais se inseriram em todas as ESTs do banco de dados e apenas nas ESTs que apresentaram similaridade com TCs, predição de ORF e alinhamento contra o banco



de proteínas NR. Para as ESTs analisadas com mais detalhe, também foi verificado o tamanho médio dos trechos inseridos.

### 6.3 Resultados e Discussão

Foi possível mapear no genoma um conjunto de 181 ESTs que possuíam trechos dos elementos de transposição Perere-3 e SR2 concatenados a um trecho não repetitivo, sugestivo de uma exonização de um elemento de transposição.

O alinhamento destas ESTs ao banco de *Tentative consensus* (TCs), permitiu a recuperação de 94 TC distintas (13 com elementos Perere-3 e 81 com elementos SR2) que quando mapeadas novamente no genoma apresentaram o trecho de elementos de transposição adjacente à região não repetitiva da TC no genoma. Esses dados sugerem que neste caso ocorreu a inserção de um retrotransposon dentro de uma região exônica e que o mesmo passou a fazer parte deste éxon pré-existente.

Outras 5 sequências de TCs, alinhadas às sequências de EST recuperadas, não apresentaram adjacência entre as regiões repetitivas (mascaradas) e não repetitivas (não mascaradas). Esses resultados indicam que estas sequências representam transposons que após a sua inserção adquiriram sítios de *splicing* e passaram a corresponder a um novo éxon de um gene pré-existente. A inspeção manual de cada uma dessas sequências permitiu determinar que para 5 casos, a distância entre o éxon derivado do transposon e o éxon adjacente era relativamente baixa (< 2500 bp). Isso sugere que o transcrito maduro é resultante de um processo de *cis-splicing*. Uma das 5 sequências analisadas também apresentou *stop codon* inserido pelo elemento de transposição. Os resultados dessas análises estão no Anexo VIII.

Foram preditas ORF longas (>120 pb) em 71 das 94 TCs analisadas e, baseando-se nas coordenadas das ORFs, foi possível identificar a provável região do gene onde o elemento de transposição ocorreu, como ilustra a Figura 60.

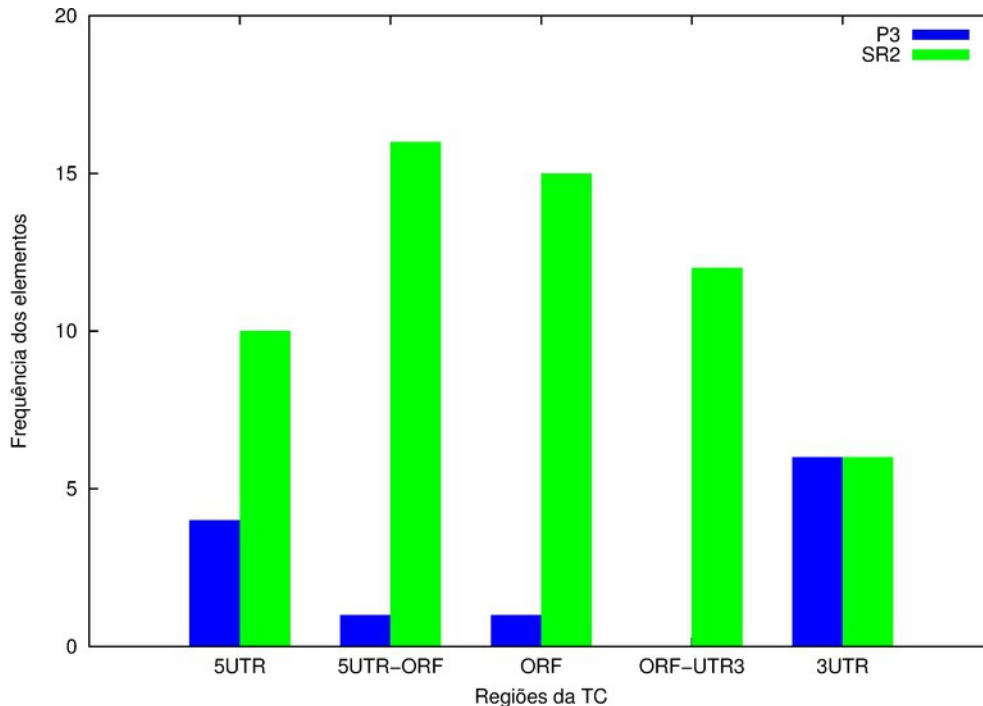


Figura 60-Distribuição de trechos similares aos elementos de transposição Perere-3 (azul) e SR2 (verde) nas diferentes regiões de mRNAs maduros. A estrutura do mRNA foi baseada em TCs construídas a partir de informações de ESTs. Fonte: Elaborada pela autora.

Os elementos Perere-3 apresentaram concentração de trechos nas regiões UTRs. Contrastando, os elementos SR2 apresentaram maior concentração de trechos nas regiões de ORF. Algumas iniciando-se nas regiões 5' UTR e estendendo-se sobre a ORF enquanto outras, iniciando-se na ORF e estendendo-se para as regiões 3' UTR.

Analisando os dados referentes aos trechos dos elementos Perere-3, os quais não ocorreram em abundância recentemente, nota-se que 6 dos 12 trechos estão na região 3'UTR. Em estudos realizados por Van de Lagemaat e colaboradores (52) foram encontrados 2 exemplos de TE servindo como sítios de poliadenilação em dois genes cujos transcritos primários terminavam em um trecho do TE. Além disso, foi verificado que em *Oryza sativa*<sup>\*</sup>, inserções de transposons são encontradas preferencialmente na região 3' UTR.(79)

Observando os elementos SR2, nota-se que seus trechos se localizam preferencialmente nas regiões codificantes e também, iniciando-se na região 5'UTR

\* *Oryza sativa* é comumente conhecida como arroz asiático. Trata-se de um organismo modelo que apresenta um genoma extenso, com aproximadamente 430Mb.

e se estendendo pela ORF.

Em estudos realizados por Sakai e colaboradores (79) em *Arabidopsis thaliana*<sup>\*\*</sup>, foi observado que o número de elementos transponíveis nas regiões de éxons ocorre em menor número do que em regiões não codificantes, refletindo o fato de que essas inserções provavelmente tem maior impacto nas funções dos genes.

Isso provavelmente reflete uma forte seleção purificadora para inserções ocorrendo em éxons codificantes.(79) Neste sentido, é curioso observar uma predominância das inserções do elemento SR2 nas regiões codificantes. Um dos fatores que poderia explicar esta aparente contradição é o fato que um grande número de ESTs, utilizadas para a construção dos TCs, são derivadas de bibliotecas ORESTES (80) que tendem a ter como alvo preferencial o centro dos genes.(81) Deste modo, uma maior representação de regiões codificantes contendo fragmentos de SR2 simplesmente refletiria uma amostragem insuficiente de regiões de UTR dos transcritos.

Considerando o fato de que foi previamente descritos que a maior parte da população de elementos Perere-3 se fixou em um período evolutivo anterior a fixação da população de SR2 (15), pode-se supor que a população de SR3<sup>\*\*\*</sup> já passou por um período maior de seleção purificadora. Esse fato pode ser um dos fatores para menor frequência do elemento Perere-3 em regiões de UTR. No entanto, isso não explica a maior concentração de fragmentos de SR2 em regiões codificantes, visto que dificilmente este tipo de inserção será neutra, devido ao impacto esperado de tais inserções na estrutura da proteína codificada pelo gene.

Também foram verificadas as frequências dos diferentes trechos dos dois elementos de transposição nas TCs analisadas. A análise foi realizada para os resultados mais significativos dos TEs contra todas as ESTs do banco de dados e, para as EST estudadas de forma mais específica, as quais apresentaram alinhamento com TCs, predição e ORFs e resultado contra o banco de proteínas NR (EST analisadas).

A Figura 61 ilustra os resultados para o elemento Perere-3, e a Figura 62

---

<sup>\*\*</sup> *Arabidopsis thaliana* é uma planta da família da mostarda utilizada como organismo modelo na área da Botânica.

<sup>\*\*\*</sup> O elemento SR3 foi estudado de forma conjunto com o elemento P3, como descrito no capítulo 2.

para o elemento SR2.

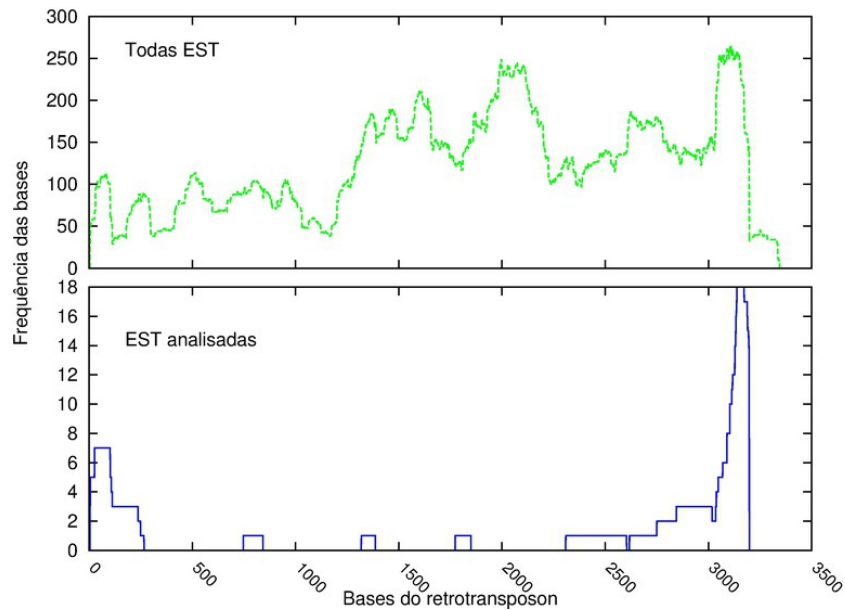


Figura 61-Frequência das bases do retrotransposon Perere-3 em todas as EST do banco de dados (verde) e nas EST representando transcritos derivados de genes contendo presença de elementos na UTRs ou regiões codificantes (azul). Fonte: Elaborada pela autora.

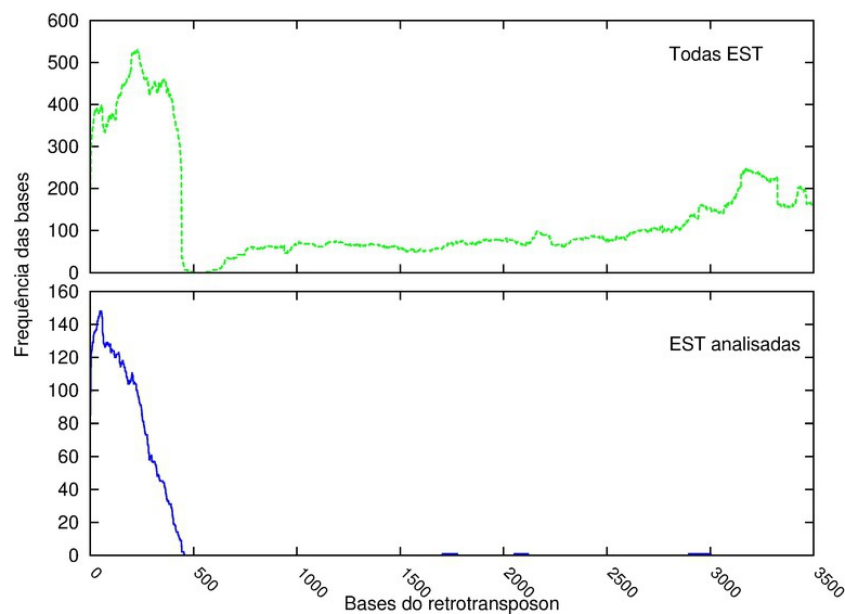


Figura 62-Frequência das bases do retrotransposon SR2 em todas as EST do banco de dados (verde) e nas EST representando transcritos derivados de genes contendo presença de elementos na UTRs ou regiões codificantes. Fonte: Elaborada pela autora.

O elemento Perere-3 apresenta uma maior representação da sua extremidade 3', provavelmente em decorrência da forma como a inserção de elementos da classe non-LTR se dá no genoma e também, devido a baixa processividade da proteína transcriptase reversa. Essa mesma observação está descrita no capítulo anterior, para as análises nas regiões intergênicas.

Também é possível observar alguns trechos correspondendo a região UTR do elemento que, conforme mencionado anteriormente, apresenta alguns sítios para a ligação de fatores de transcrição.

Há uma queda abrupta da representatividade após a base 3096, pois já foi demonstrado que somente até esta base as diferentes cópias deste elemento são conservadas, sendo que as bases *downstream* representam uma ponta variável.(14)

O trecho mais frequente do elemento SR2 nas ESTs corresponde a região 5' UTR (Figura 62). Esse trecho também corresponde a maior parte do elemento SR2 não autônomo e como mencionado no capítulo anterior, apresenta alguns sítios para a ligação de fatores de transcrição. Nota-se no entanto que há uma ausência quase que total de outras regiões, o que não é observado no conjunto total de ESTs, indicando uma possível seleção de trechos 5', levando a uma exacerbação da predominância deste trecho.

A Figura 63 ilustra o tamanho médio dos trechos presentes nas ESTs analisadas de forma mais específica. É possível detectar uma pequena variação de tamanho entre os trechos correspondentes ao elemento Perere-3 e os trechos do elemento SR2 que não é estatisticamente relevante ( $p$ -valor=0.9968). O fato das cópias dos elementos SR2 serem mais recentes que as do elemento Perere-3 (15) e, portanto, ter sofrido menor erosão devido a mutações, pode ser um dos fatores que contribuem para este maior tamanho.

A presença destes elementos em transcritos pode indicar que os elementos de transposição podem estar contribuindo como elementos regulatórios, quando são observados os sítios para a ligação de fatores de transcrição, como promotores, quando são observados em regiões 5'UTR e 5'UTR-ORF e também, como sítios de poliadenilação, quando presentes nas regiões 3'UTR. A adição de sinais de parada prematuros e adições de trechos de proteínas, derivadas dos elementos de

transposição, demonstram a influência destes elementos em genes do *S. mansoni*.

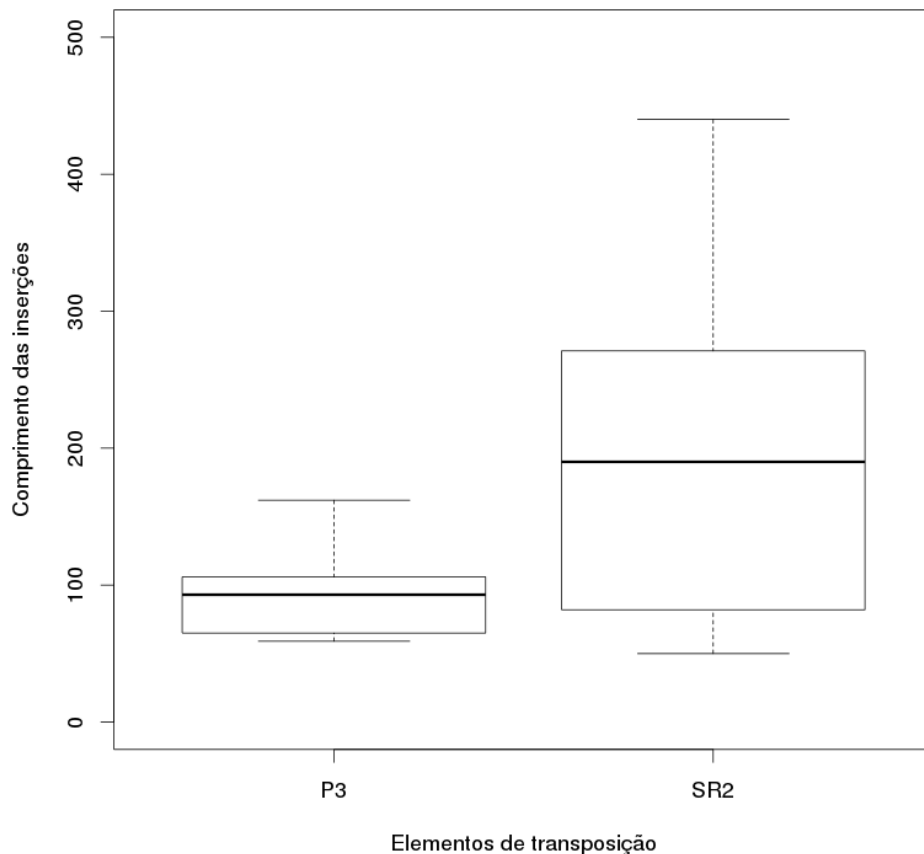


Figura 63-Boxplot ilustrando o comprimento dos elementos Perere-3 e SR2 encontrados em sequências de transcritos derivados de genes do *S. mansoni*. Fonte: Elaborada pela autora.

A Tabela 1 e Tabela 2 apresentam as proteínas do banco de dados NR do NCBI que resultaram como melhor alinhamento para as TCs com os elementos Perere-3 e SR2, respectivamente.

A maior parte dos resultados para as TCs com elementos SR2 estão vinculados às proteínas hipotéticas (51 das 59 correlações – 86%). Para o elemento Perere-3 é possível observar uma variedade maior de proteínas pois poucos casos apresentam correlação com proteínas hipotéticas (2 das 12 correlações – 16%).

Tabela 1-Lista de proteínas do banco de dados NR do NCBI representando o melhor alinhamento com TCs com presença dos elementos Perere-3

<b>Qtde de TCs</b>	<b>Região da TC onde está o elemento Perere-3</b>	<b>Descrição da proteína</b>
1	UTR5	AMP-activated protein kinase
1	UTR5	eupitriysin (M16 family)
1	UTR5	hypothetical protein
1	UTR5	SJCHGC03027 protein
1	UTR5-ORF	SJCHGC03163 protein
1	ORF	hypothetical protein
1	UTR3	glucose transport protein
1	UTR3	heme binding protein
1	UTR3	prokaryotic DNA topoisomerase
1	UTR3	serine protease inhibitors
1	UTR3	SJCHGC03947 protein
1	UTR3	tetraspanin 42 invertebrate

Fonte: Elaborada pela autora.

Tabela 2-Lista de proteínas do banco de dados NR do NCBI representando o melhor alinhamento com TCs com presença dos elementos SR2

<b>Qtde de TCs</b>	<b>Região da TC onde está o elemento SR2</b>	<b>Descrição da proteína</b>
1	UTR5	sap18
1	UTR5	bestrophin-related
1	UTR5	hypothetical protein
5	UTR5	hypothetical protein
1	UTR5	multidrug resistance protein
1	UTR5	hypothetical protein
13	UTR5-ORF	hypothetical protein
3	UTR5-ORF	hypothetical protein

continua

continuação

<b>Qtde de TCs</b>	<b>Região da TC onde está o elemento SR2</b>	<b>Descrição da proteína</b>
2	ORF	hypothetical protein
11	ORF	hypothetical protein
1	ORF	multidrug resistance protein
1	ORF	fbxI20
1	ORF-UTR3	hypothetical protein
9	ORF-UTR3	hypothetical protein
1	ORF-UTR3	multidrug resistance protein
1	ORF-UTR3	fbxI20
1	UTR3	hypothetical protein
1	UTR3	rna recognition motif
3	UTR3	hypothetical protein
1	UTR3	hypothetical protein

Fonte: Elaborada pela autora.



# Capítulo 7

## Conclusões

---



## 7 Conclusões

Os resultados apresentados nos capítulos anteriores sugerem que os elementos de transposição Perere-3 e SR2 podem estar influenciando alterações nas regiões intrônicas, intergênicas e presentes em mRNA da espécie *S. mansoni*.

Foi possível observar que a inserção de elementos das famílias estudadas contribuiu para o aumento do tamanho dos íntrons. As inserções ocorreram com maior intensidade nos últimos íntrons dos genes. As sequências desses elementos, por apresentarem conteúdo GC elevado, alteram o conteúdo GC médio desses íntrons. Essas observações permitem supor que os trechos inseridos podem estar promovendo alterações na conformação do gene, podem alterar os padrões de *splicing* e também podem aumentar os casos de exonização.(71) A observação de uma possível pressão seletiva negativa sobre os elementos de transposição posicionados no mesmo sentido da transcrição sugerem que esses elementos devem apresentar motivos que interferem na transcrição dos genes.

As inserções de ilhas CpG, apresentam um padrão diferente dos demais trechos dos elementos SR2 com relação à proximidade dos éxons. Também estão ocorrendo numa frequência maior do que seria esperada, sugerindo uma potencial seleção positiva destes elementos em íntrons de genes.

Nas análises das regiões intergênicas, foi possível prever sítios para ligação de fatores de transcrição ao longo das sequências de ambos os retrotransposons em estudo, inclusive nos trechos dos elementos que apresentaram maior frequência no genoma. Nessas regiões, o padrão de ilhas CpG não apresentou o mesmo enriquecimento observado nas regiões intrônicas. Também foi possível observar que elementos SR2 tendem a se fixar em regiões flanqueando genes que codificam proteínas transmembranares, as quais podem estar envolvidas na relação parasita-hospedeiro.

A obtenção de sequência de transcritos possuindo sequências de retrotransposons, indica a incorporação de elementos em regiões transcritas.

Também é possível prever que ocorreu a modificação da sequência de algumas proteínas devido a algumas inserções em regiões codificantes e também,

devido a observação de *stop codon* em um dos trechos dos elementos. Além disso, sequências de elementos de transposição na região 3'UTR de transcritos podem fornecer sítios de ligação de micro-RNA, visto que este é um mecanismo utilizado pelas células para controlar a abundância de transcritos de retrotransposons e, deste modo, afetar a disponibilidade do transcrito ao qual esta sequência se encontra associada.

Esses resultados permitem sugerir que as inserções dos elementos Perere-3 e SR2 contribuíram com alterações funcionais e estruturais dos genes, e também como fatores para atuar sobre elementos epigenéticos possivelmente contribuindo para a diferenciação e especiação dos organismos *S. mansoni* e *S. japonicum*.

## Referências

1 MCCLINTOCK, B. The origin and behavior of mutable loci in maize. **Proceedings of the National Academy of Sciences of the United States of America**, v. 36, n. 6, p. 344–355, 1950.

2 DOOLITTLE, W. F.; SAPIENZA, C. Selfish genes, the phenotype paradigm and genome evolution. **Nature**, v. 284, n. 5757, p. 601–603, 1980. doi: 10.1038/284601a0.

3 ORGEL, L. E.; CRICK, F. H. Selfish DNA: the ultimate parasite. **Nature**, v. 284, n. 5757, p. 604–607, 1980.

4 ALZOHAIRY, A. M. et al. Transposable elements domesticated and neofunctionalized by eukaryotic genomes. **Plasmid**, v. 69, n. 1, p. 1–15, 2013. doi: 10.1016/j.plasmid.2012.08.001.

5 REBOLLO, R.; ROMANISH, M. T.; MAGER, D. L. Transposable elements: an abundant and natural source of regulatory sequences for host genes. **Annual Review of Genetics**, v. 46, n. 1, p. 21–42, 2012. doi: 10.1146/annurev-genet-110711-155621.

6 BEAUREGARD, A.; CURCIO, M. J.; BELFORT, M. The take and give between retrotransposable elements and their hosts. **Annual Review of Genetics**, v. 42, p. 587–617, 2008. doi: 10.1146/annurev.genet.42.110807.091549.

7 GOODIER, J. L.; KAZAZIAN, H. H. Retrotransposons revisited: the restraint and rehabilitation of parasites. **Cell**, v. 135, n. 1, p. 23–35, 2008. doi: 10.1016/j.cell.2008.09.022.

8 KAZAZIAN, H. H. Mobile elements: drivers of genome evolution. **Science**, v. 303, n. 5664, p. 1626–1632, 2004. doi: 10.1126/science.1089670.

9 BERRIMAN, M. et al. The genome of the blood fluke *Schistosoma mansoni*. **Nature**, v. 460, n. 7253, p. 352–358, 2009. doi: 10.1038/nature08160.

10 THE SCHISTOSOMA JAPONICUM GENOME SEQUENCING AND

FUNCTIONAL ANALYSIS CONSORTIUM. The *Schistosoma japonicum* genome reveals features of host-parasite interplay. **Nature**, v. 460, n. 7253, p. 345–351, 2009.doi: 10.1038/nature08140.

11 LAWTON, S. P. et al. Genomes and geography: genomic insights into the evolution and phylogeography of the genus *Schistosoma*. **Parasites & Vectors**, v. 4, p. 131, 2011.doi: 10.1186/1756-3305-4-131.

12 LOCKYER, A. E. et al. The phylogeny of the Schistosomatidae based on three genes with emphasis on the interrelationships of *Schistosoma* Weinland, 1858. **Parasitology**, v. 126, n. Pt 3, p. 203–224, 2003.

13 DREW, A. C. et al. SR2 elements, non-long terminal repeat retrotransposons of the RTE-1 lineage from the human blood fluke *Schistosoma mansoni*. **Molecular Biology and Evolution**, v. 16, n. 9, p. 1256–1269,1999.

14 DEMARCO, R. et al. Identification of 18 new transcribed retrotransposons in *Schistosoma mansoni*. **Biochemical and Biophysical Research Communications**, v. 333, n. 1, p. 230–240, 2005.doi: 10.1016/j.bbrc.2005.05.080.

15 VENANCIO, T. M. et al. Bursts of transposition from non-long terminal repeat retrotransposon families of the RTE clade in *Schistosoma mansoni*. **International Journal for Parasitology**, v. 40, n. 6, p. 743–749,2010.doi: 10.1016/j.ijpara.2009.11.013.

16 FRAZER, K. A. et al. Cross-species sequence comparisons: a review of methods and available resources. **Genome Research**, v. 13, n. 1, p. 1–12,2003.doi: 10.1101/gr.222003.

17 PERL COMMUNITY, **The perl programming language**. Disponível em:<<http://www.perl.org/>>. Acesso em: 6 janeiro 2014.

18 BIOPERL COMMUNITY, B. **BioPerl**. Disponível em: <[http://www.bioperl.org/wiki/Main\\_Page](http://www.bioperl.org/wiki/Main_Page)>.Acesso em: 6 janeiro 2014.

19 JURKA, J. et al. Repetitive sequences in complex genomes: structure and evolution. **Annual Review of Genomics and Human Genetics**, v. 8, p. 241–259,

2007.doi: 10.1146/annurev.genom.8.080706.092416.

20 WICKER, T. et al. A unified classification system for eukaryotic transposable elements. **Nature Reviews Genetics**, v. 8, n. 12, p. 973–982, 2007.doi: 10.1038/nrg2165.

21 PRITHAM, E. J. Transposable elements and factors influencing their success in eukaryotes. **Journal of Heredity**, v. 100, n. 5, p. 648 –655, 2009.doi: 10.1093/jhered/esp065.

22 LE ROUZIC, A.; CAPY, P. Theoretical approaches to the dynamics of transposable elements in genomes, populations, and species. In: LANKENAU, D.H.; VOLFF, J.N. (Org.); **Transposons and the dynamic genome**. Berlin: Springer, 2011, v.4, p.1–19

23 LEVIN, H. L.; MORAN, J. V. Dynamic interactions between transposable elements and their hosts. **Nature Reviews Genetics**, v. 12, n. 9, p. 615–627, 2011.doi: 10.1038/nrg3030.

24 CRAIG, N. L. Target site selection in transposition. **Annual Review of Biochemistry**, v. 66, p. 437–474, 1997.doi: 10.1146/annurev.biochem.66.1.437.

25 KUDUVALLI, P. N.; MITRA, R.; CRAIG, N. L. Site-specific Tn7 transposition into the human genome. **Nucleic Acids Research**, v. 33, n. 3, p. 857–863, 2005.doi: 10.1093/nar/gki227.

26 FESCHOTTE, C. Transposable elements and the evolution of regulatory networks. **Nature Reviews Genetics**, v. 9, n. 5, p. 397–405, 2008.doi: 10.1038/nrg2337.

27 CDC HOME. **CDC - Schistosomiasis - Biology**. Disponível em: <<http://www.cdc.gov/parasites/schistosomiasis/biology.html>>. Acesso em: 7 janeiro 2014.

28 DESPOMMIER, D. D.; GWADZ, R. W.; HOTEZ, P. J. Schistosomes: *Schistosoma mansoni* (Sambon 1907), *Schistosoma japonicum* (Katsurada 1904), *Schistosoma haematobium* (Bilharz 1852). In: **Parasitic Diseases**. New York:

Springer,1995.p.108–121. Disponível em:  
<[http://link.springer.com/chapter/10.1007/978-1-4612-2476-1\\_18](http://link.springer.com/chapter/10.1007/978-1-4612-2476-1_18)>.Acesso em:05  
janeiro 2014.

29 RHEINBERG, C. E. et al. Schistosoma haematobium, S. intercalatum, S. japonicum, S. mansoni, and S. rodhaini in mice: relationship between patterns of lung migration by schistosomula and perfusion recovery of adult worms. **Parasitology Research**, v. 84, n. 4, p. 338–342, 1998.

30 PROTASIO, A. V. et al. A systematically improved high quality genome and transcriptome of the human blood fluke Schistosoma mansoni. **PLoS Neglected Tropical Diseases**, v. 6, n. 1, p. E1455,2012.doi: 10.1371/journal.pntd.0001455.

31 GEYER, K. K. et al. Cytosine methylation regulates oviposition in the pathogenic blood fluke Schistosoma mansoni. **Nature Communications**, v. 2, p. 424,2011.doi: 10.1038/ncomms1433.

32 URETA-VIDAL, A.; ETTWILLER, L.; BIRNEY, E. Comparative genomics: genome-wide analysis in metazoan eukaryotes. **Nature Reviews. Genetics**, v. 4, n. 4, p. 251–262,2003.doi: 10.1038/nrg1043.

33 JANICKI, M.; ROOKE, R.; YANG, G. Bioinformatics and genomic analysis of transposable elements in eukaryotic genomes. **Chromosome Research: an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology**, v. 19, n. 6, p. 787–808,2011.doi: 10.1007/s10577-011-9230-7.

34 KOONIN, E. V. Orthologs, paralogs, and evolutionary genomics. **Annual Review of Genetics**, v. 39, p. 309–338,2005.doi: 10.1146/annurev.genet.39.073003.114725.

35 ALTSCHUL, S. F. et al. Basic local alignment search tool. **Journal of Molecular Biology**, v. 215, n. 3, p. 403–410,1990.doi: 10.1016/S0022-2836(05)80360-2.

36 KRISTENSEN, D. M. et al. Computational methods for gene orthology inference. **Briefings in Bioinformatics**, v. 12, n. 5, p. 379–391,2011.doi: 10.1093/bib/bbr030.

37 LEVINE, A. **Bioinformatics approaches to RNA splicing**.2001. 60p. Master thesis ( Philosophy)-University of Cambridge and Sanger Centre, Cambridge,2001.



- 38 ROGOZIN, I. B. et al. Origin and evolution of spliceosomal introns. **Biology Direct**, v. 7, p. 11,2012.doi: 10.1186/1745-6150-7-11.
- 39 LUO, M.; REED, R. Splicing is required for rapid and efficient mRNA export in metazoans. **Proceedings of the National Academy of Sciences**, v. 96, n. 26, p. 14937–14942,1999.doi: 10.1073/pnas.96.26.14937.
- 40 ZHOU, Z. et al. The protein Aly links pre-messenger-RNA splicing to nuclear export in metazoans. **Nature**, v. 407, n. 6802, p. 401–405,2000.doi: 10.1038/35030160.
- 41 FEDOROVA, L.; FEDOROV, A. Introns in gene evolution. **Genetica**, v. 118, n. 2-3, p. 123–131, 2003.
- 42 WESTHOLM, J. O.; LAI, E. C. Mirtrons: microRNA biogenesis via splicing. **Biochimie**, v. 93, n. 11, p. 1897–1904,2011.doi: 10.1016/j.biochi.2011.06.017.
- 43 KIM, E.; GOREN, A.; AST, G. Alternative splicing: current perspectives. **BioEssays: news and reviews in molecular, cellular and developmental biology**, v. 30, n. 1, p. 38–47,2008.doi: 10.1002/bies.20692.
- 44 NILSEN, T. W. Evolutionary origin of SL-addition trans-splicing: still an enigma. **Trends in Genetics**, v. 17, n. 12, p. 678–680, 2001.
- 45 TAJIMA, F. Determination of window size for analyzing DNA sequences. **Journal of Molecular Evolution**, v. 33, n. 5, p. 470–473,1991.doi: 10.1007/BF02103140.
- 46 DEATON, A. M.; BIRD, A. CpG islands and the regulation of transcription. **Genes & Development**, v. 25, n. 10, p. 1010–1022,2011.doi: 10.1101/gad.2037511.
- 47 TAKAI, D.; JONES, P. A. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. **Proceedings of the National Academy of Sciences**, v. 99, n. 6, p. 3740–3745,2002.doi: 10.1073/pnas.052410099.

48 BAUER, S. et al. Ontologizer 2.0--a multifunctional tool for GO term enrichment analysis and data exploration. **Bioinformatics**, v. 24, n. 14, p. 1650–1651,2008.doi: 10.1093/bioinformatics/btn250.

49 SANGER INSTITUTE. **Schistosoma mansoni**. Disponível em:<<http://www.sanger.ac.uk/resources/downloads/helminths/schistosoma-mansoni.html>>.Acesso em: 12 janeiro 2014.

50 R COMMUNITY. **The R project for statistical computing**. Disponível em:<<http://www.r-project.org/>>. Acesso em: 3 fevereiro 2014.

51 ZHANG, Y.; ROMANISH, M. T.; MAGER, D. L. Distributions of transposable elements reveal hazardous zones in mammalian introns. **PLoS Computational biology**, v. 7, n. 5, p. E1002046,2011.doi: 10.1371/journal.pcbi.1002046.

52 VAN DE LAGEMAAT, L. N. et al. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. **Trends in Genetics**, v. 19, n. 10, p. 530–536,2003.doi: 10.1016/j.tig.2003.08.004.

53 SHARIF, J. et al. Divergence of CpG island promoters: a consequence or cause of evolution? **Development, Growth & Differentiation**, v. 52, n. 6, p. 545–554,2010.doi: 10.1111/j.1440-169X.2010.01193.x.

54 AMIT, M. et al. Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. **Cell Reports**, v. 1, n. 5, p. 543–556,2012.doi: 10.1016/j.celrep.2012.03.013.

55 KEREN, H.; LEV-MAOR, G.; AST, G. Alternative splicing and evolution: diversification, exon definition and function. **Nature Reviews Genetics**, v. 11, n. 5, p. 345–355,2010.doi: 10.1038/nrg2776.

56 EMBL-EBI. **QuickGO**. Disponível em:<<http://www.ebi.ac.uk/QuickGO/>>.Acesso em: 8 fevereiro 2014.

57 BARRETT, L. W.; FLETCHER, S.; WILTON, S. D. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. **Cellular and Molecular Life Sciences: CMLS**, v. 69, n. 21, p. 3613–3634,2012.doi:

10.1007/s00018-012-0990-9.

58 RIETHOVEN, J.-J. M. Regulatory regions in DNA: promoters, enhancers, silencers, and insulators. **Methods in Molecular Biology**, v. 674, p. 33–42,2010.doi: 10.1007/978-1-60761-854-6\_3.

59 NEWS, E.; V. J.; PERMALINK, More clues that intergenic DNA is functional. **Evolution News & Views**. Disponível em: <[http://www.evolutionnews.org/2013/07/more\\_clues\\_that\\_1074451.html](http://www.evolutionnews.org/2013/07/more_clues_that_1074451.html)>. Acesso em: 8 fevereiro 2014.

60 HADJIARGYROU, M.; DELIHAS, N. The intertwining of transposable elements and non-coding RNAs. **International Journal of Molecular Sciences**, v. 14, n. 7, p. 13307–13328,2013.doi: 10.3390/ijms140713307.

61 KELLEY, D.; RINN, J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. **Genome Biology**, v. 13, n. 11, p. R107,2012.doi: 10.1186/gb-2012-13-11-r107.

62 NCBI FASTA. **Fasta Format**. Disponível em:< <http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml>>. Acesso em: 19 janeiro 2014.

63 PORTALES-CASAMAR, E. et al. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. **Nucleic Acids Research**, v. 38, n. Database issue, p. D105–D110,2010.doi: 10.1093/nar/gkp950.

64 STEIN, L. D. **Bio::Graphics**. Disponível em:< <http://search.cpan.org/~lds/Bio-Graphics-2.37/lib/Bio/Graphics.pm>>. Acesso em: 5 fevereiro 2014.

65 KROGH, A. et al. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. **Journal of Molecular Biology**, v. 305, n. 3, p. 567–580,2001.doi: 10.1006/jmbi.2000.4315.

66 NIU, W. et al. Diverse transcription factor binding features revealed by genome-wide ChIP-seq in *C. elegans*. **Genome Research**, v. 21, n. 2, p. 245–254,2011.doi: 10.1101/gr.114587.110.

67 WONG, M. W. et al. The large subunit of basal transcription factor SNAPc is a Myb domain protein that interacts with Oct-1. **Molecular and Cellular Biology**, v. 18, n. 1, p. 368–377, 1998.

68 FAULKNER, G. J. et al. The regulated retrotransposon transcriptome of mammalian cells. **Nature Genetics**, v. 41, n. 5, p. 563–571, 2009. doi: 10.1038/ng.368.

69 MEDSTRAND, P. et al. Impact of transposable elements on the evolution of mammalian gene regulation. **Cytogenetic and Genome Research**, v. 110, n. 1-4, p. 342–352, 2005. doi: 10.1159/000084966.

70 CLANCY, S. **RNA Functions**. Disponível em: <<http://www.nature.com/scitable/topicpage/rna-functions-352>>. Acesso em: 22 janeiro 2014.

71 SELA, N.; KIM, E.; AST, G. The role of transposable elements in the evolution of non-mammalian vertebrates and invertebrates. **Genome Biology**, v. 11, n. 6, p. R59, 2010. doi: 10.1186/gb-2010-11-6-r59.

72 SOREK, R. The birth of new exons: mechanisms and evolutionary consequences. **RNA**, v. 13, n. 10, p. 1603–1608, 2007. doi: 10.1261/rna.682507.

73 KRULL, M.; BROSIUS, J.; SCHMITZ, J. Alu-SINE exonization: en route to protein-coding function. **Molecular Biology and Evolution**, v. 22, n. 8, p. 1702–1711, 2005. doi: 10.1093/molbev/msi164.

74 KREHLING, J.; GRAVELEY, B. R. The origins and implications of alternative splicing. **Trends in Genetics**, v. 20, n. 1, p. 1–4, 2004. doi: 10.1016/j.tig.2003.11.001.

75 LIN, L. et al. Diverse splicing patterns of exonized alu elements in human tissues. **PLoS Genetics**, v. 4, n. 10, p. E1000225, 2008. doi: 10.1371/journal.pgen.1000225.

76 COMPUTATIONAL BIOLOGY AND FUNCTIONAL GENOMICS LABORATORY, H. **Gene Index – DFCI**. Disponível em: <<http://compbio.dfci.harvard.edu/tgi/>>. Acesso em 8 fevereiro 2014.

77 OKONECHNIKOV, K. et al. Unipro UGENE: a unified bioinformatics toolkit. **Bioinformatics**, v. 28, n. 8, p. 1166–1167, 2012. doi: 10.1093/bioinformatics/bts091.

78 PRUITT, K. D.; TATUSOVA, T.; MAGLOTT, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. **Nucleic Acids Research**, v. 33, n. Database issue, p. D501–504, 2005. doi: 10.1093/nar/gki025.

79 SAKAI, H.; TANAKA, T.; ITOH, T. Birth and death of genes promoted by transposable elements in *Oryza sativa*. **Gene**, v. 392, n. 1-2, p. 59–63, doi: 10.1016/j.gene.2006.11.010.

80 VERJOVSKI-ALMEIDA, S. et al. Transcriptome analysis of the acoelomate human parasite *Schistosoma mansoni*. **Nature Genetics**, v. 35, n. 2, p. 148–157, 2003. doi: 10.1038/ng1237.

81 FIETTO, J. L. R.; DEMARCO, R.; VERJOVSKI-ALMEIDA, S. Use of degenerate primers and touchdown PCR for construction of cDNA libraries. **BioTechniques**, v. 32, n. 6, p. 1404–1408, 1410–1411, 2002.



## Anexo I-Pipelines das análises para organização dos dados iniciais

Título: Pipeline para organização dos dados iniciais				
	Dados Entrada	Estratégia de Análise	Software utilizado	Dados Saída
Fase 1	GFF Smp	Extrair as coordenadas dos éxons, de Smp utilizando as flags padrões definidas para arquivos do tipo GFF	Módulo Perl	Éxons Smp
Fase 2	Genes Preditos Sjc Genoma Sjc	Considerar alinhamentos com <i>evaluate</i> de $10^{-3}$	BlastN	Éxons Sjc – Blast
Fase 3	Genes Preditos Sjc Genoma Sjc	Realizar outro alinhamento para obter exatidão nas coordenadas iniciais e finais dos éxons/íntrons	Spidey	Éxons Sjc - Spidey
Fase 4	Éxons Sjc – Blast Éxons Sjc - Spidey	Verificar coordenadas de início e fim de éxons comparando os resultados do Blast e do Spidey	Módulo Perl	Éxons Sjc
Fase 5	Éxons Smp Éxons Sjc	Utilizando o arquivo com todas as coordenadas dos éxons, organizadas de forma crescente e por cromossomo/ <i>scaffold</i> , definir coordenadas dos íntrons e coordenadas e tipo das regiões intergênicas	Módulo Perl	ÉxonsÍntron sIntergênic as Smp ÉxonsÍntron sIntergênic as Sjc

Quadro 1-Pipeline implementado para a organização dos dados iniciais. Fonte: Elaborado pela autora.

Título: Pipeline da definição dos genes ortólogos				
Dados Entrada		Estratégia de Análise	Software utilizado	Dados Saída
Fase 1	Genes preditos Smp	Considerar alinhamentos com <i>evaluate</i> de $10^{-3}$ utilizando como <i>query</i> as sequências de Smp e como banco de dados as sequências de Sjc e depois, vice-versa	BlastP	Genes preditos de Smp X Sjc
	Genes preditos Sjc			
Fase 2	Genes preditos de Smp X Sjc	Considerar alinhamentos que de forma recíproca apresentam os mesmos resultados quando ocorre a alteração entre as sequências de <i>query</i> e de banco de dados e similaridade > 85%	Módulo Perl	Genes Ortólogos
Fase 3	Genes Ortólogos	Definir os éxons equivalentes ao trecho alinhado entre os pares de genes ortólogos	Módulo Perl	ORTO-Éxons
	Éxons Smp			
	Éxons Sjc			
Fase 4	ORTO-Éxons	Separar genes com as mesmas quantidades de éxons e com quantidades diferentes	Módulo Perl	ORTO-ql ORTO-qD
Fase 5	ORTO-Éxons	Definir íntrons equivalentes ao trecho alinhado entre os pares de genes ortólogos	Módulo Perl	ORTO-Íntrons

Quadro 2-Pipeline implementado para a identificação dos genes ortólogos. Fonte: Elaborado pela autora.



<b>Título:</b> Pipeline do mapeamento dos retrotransposons SR2 e Perere-3 no genoma				
<b>Entrada</b>		<b>Estratégia de Análise</b>	<b>Software utilizado</b>	<b>Saída</b>
Fase 1	Sequência SR2	Alinhamentos das sequências de elementos de transposições contra o Genoma de <i>S. mansoni</i> considerando <i>evaluate</i> de $10^{-3}$	BlastN	Inserções no genoma SR2
	Sequência Perere-3			Inserções no genoma Perere-3
	Genoma Smp			
Fase 2	Inserções genoma SR2	Considerar alinhamentos com mais de 50 bp e com identidade superior a 85%	Script shell	Inserções no genoma SR2
	Inserções genoma Perere-3			Inserções no genoma Perere-3

Quadro 3-Pipeline implementado para o mapeamento dos elementos de transposições SR2 e Perere-3 no genoma de *S. mansoni*. Fonte: Elaborado pela autora.



## Anexo II-Pipelines das análises nas regiões intrônicas

<b>Título:</b> Pipeline do mapeamento dos elementos de transposições SR2 e Perere-3 nos íntrons dos genes ortólogos			
<b>Entrada</b>	<b>Estratégia de Análise</b>	<b>Software utilizado</b>	<b>Saída</b>
Fase 1 Inserções no genoma SR2 Inserções no genoma Perere-3 ORTO-Íntrons-Qtdel	Verificar quais íntrons dos genes ortólogos de Smp abrangem coordenadas correspondentes as inserções que os elementos de transposições apresentaram no genoma	Módulo Perl	Inserções SR2 nos íntrons Smp Inserções Perere-3 nos íntrons Smp
Fase 2 ORTO-qi Inserções SR2 nos íntrons Smp Inserções Perere-3 nos íntrons Smp	Separar íntrons que apresentam inserções de Perere-3, de SR2, de ambos os elementos, e os íntrons sem inserção.	Módulo Perl	Íntrons Smp inserções SR2 íntrons Smp inserções Perere-3 Íntrons Smp inserções Ambos Íntrons Smp sem Inserções

Quadro 4-Pipeline implementado para o mapeamento dos elementos de transposição SR2 e Perere-3 nos íntrons dos genes ortólogos. Fonte: Elaborado pela autora.

<b>Título:</b> Pipeline da análise sobre o tamanho dos íntrons com e sem inserções			
<b>Entrada</b>	<b>Estratégia de Análise</b>	<b>Software utilizado</b>	<b>Saída</b>
<b>Etapa 1</b> Íntrons Smp inserções SR2 Íntrons Smp inserções Perere-3 Íntrons Smp sem Inserções ORTO-Íntrons -Sjc ORTO-ql	Identificar íntrons de Sjc equivalentes aos íntrons de Smp	Módulo Perl	Íntrons Sjc equivalentes Smp com inserções SR2 Íntrons Sjc equivalentes Smp com inserções Perere-3 Íntrons Sjc equivalentes Smp sem inserções
<b>Etapa 2</b> Íntrons Smp inserções SR2 Íntrons Smp inserções Perere-3 Íntrons Smp sem Inserções Íntrons Sjc equivalentes Smp com inserções SR2 Íntrons Sjc equivalentes Smp com inserções Perere-3 Íntrons Sjc equivalentes Smp sem inserções	Verificar o tamanho do íntron e agrupar em intervalos de 500 em 500 bases até o comprimento máximo de 10.000 bases. Acima desse comprimento, agrupar em um mesmo intervalo.	Módulo Perl Gnuplot	Gráfico com distribuição dos tamanhos de íntrons, com e sem inserções, de Smp e Sjc

Quadro 5-Pipeline implementado para análise sobre o tamanho dos íntrons com e sem inserções.  
 Fonte: Elaborado pela autora.

<b>Título:</b> Pipeline da análise para definição da posição da inserção no gene				
<b>Entrada</b>		<b>Estratégia de Análise</b>	<b>Software utilizado</b>	<b>Saída</b>
<b>Fase 1</b>	Íntrons com inserções SR2	Considerando a quantidade total de íntrons do gene cujo íntron apresenta inserção, calcular o percentual de representação desse íntron em relação a todos os íntrons do conjunto, grupando em uma matriz cujo eixo x representa a posição do íntron com inserção e o eixo y a quantidade de íntrons dos genes.	Módulo Perl Gnuplot	Heatmap com % de representação dos íntrons com inserção agrupados pela posição no gene
	Íntrons com inserções Perere-3			
	Íntrons Smp			
	Total genes de Smp agrupados por qtde íntrons			

Quadro 6-Pipeline para definição da posição da inserção no gene. Fonte: Elaborado pela autora.

<b>Título:</b> Pipeline da análise para definir a distância da inserção em relação aos éxons			
<b>Entrada</b>	<b>Estratégia de Análise</b>	<b>Software utilizado</b>	<b>Saída</b>
Fase 1 Íntrons Smp inserções SR2 Íntrons Smp inserções Perere-3 ORTO-Íntrons -Smp	Separar inserções por sentido senso e antisenso	Script Shell	Posição inserções SR2 Posição inserções Perere-3
Fase 2 Posição inserções SR2 Posição inserções Perere-3	Calcular fração do segmento que compreende extremidade do íntron, próxima ao éxon posicionado na região 5', e a extremidade da inserção mas próxima a esse ponto inicial. Acumular essas frações em intervalos de 0.1 para gerar histograma	Código Perl GnuPlot	Inserções íntrons – near 5' SR2 Inserções íntrons – near 5' Perere-3 Distribuição para inserções SR2 Distribuição para inserções Perere-3

Quadro 7-Pipeline para definição da distância da inserção em relação aos éxons. Fonte: Elaborado pela autora.

<b>Título:</b> Pipeline da análise para definir o %GC dos íntrons com inserções de SR2 e Perere-3			
<b>Entrada</b>	<b>Estratégia de Análise</b>	<b>Software utilizado</b>	<b>Saída</b>
Fase 1 Íntrons Smp inserções SR2 Íntrons Smp inserções Perere-3 Íntrons Smp Genoma Smp	Gerar arquivos com as sequências dos íntrons com inserção dos elementos SR2 e Perere-3 e sem inserção	Módulo Perl Bioperl	Íntrons com inserções SR2 Íntrons com inserções Perere-3
Fase 2 Íntrons com inserções SR2 Íntrons com inserções Perere-3 Íntrons sem inserção	Utilizando o método de <i>sliding window</i> , calcular o conteúdo GC ao longo do íntron considerando janela com intervalo de 10% do tamanho do íntron e com sobreposição	Módulo Perl Gnuplot	Gráficos conteúdo GC ao longo dos íntrons
Fase 3 Íntrons com inserções SR2 Íntrons com inserções Perere-3	Quantificar o total de nucleotídeos e calcular o percentual GC das sequências	Módulo Perl	%GC das inserções SR2 %GC das inserções de Perere-3
Fase 4 Íntrons com inserções SR2 Íntrons com inserções Perere-3 íntrons Sjc Genoma Sjc	Localizar o íntron de Sjc, ortólogo ao íntron de Smp com inserção e gerar arquivo fasta com a sequência do íntron de Sjc	Módulo Perl Bioperl	Íntrons ortólogos Sjc para SR2 Íntrons ortólogos Sjc para Perere-3
Fase 5 Íntrons ortólogos Sjc para SR2 Íntrons ortólogos Sjc para Perere-3	Quantificar o total de nucleotídeos e calcular o percentual GC das sequências	Módulo Perl	%GC íntrons ortólogos para SR2 %GC íntrons ortólogos para Perere-3

continua

## continuação

<b>Título:</b> Pipeline da análise para definir o %GC dos íntrons com inserções de SR2 e Perere-3				
Fase 6	Íntrons sem inserção Íntrons com inserções SR2 Íntrons com inserções Perere-3 Genoma Smp	Selecionar aleatoriamente íntrons sem inserção na mesma quantidade de íntrons que apresentaram inserções de SR2 e de Perere-3 e gerar arquivo fasta com as as sequências desses íntrons	Módulo Perl Bioperl	Íntrons sem inserção
Fase 7	Íntrons sem inserção	Quantificar o total de nucleotídeos e calcular o percentual GC das sequências	Módulo Perl	%GC íntrons sem inserção
Fase 8	%GC das inserções SR2 %GC das inserções de Perere-3 %GC íntrons ortólogos para SR2 %GC íntrons ortólogos para Perere-3 %GC íntrons sem inserção	Acumular o percentual GC dos íntrons dentro do intervalo entre 15% e 55%	Módulo Perl Gnuplot	Histograma com média %GC para íntrons com inserção, íntrons ortólogos e íntrons sem inserção

Quadro 8-Pipeline para definição do %GC dos íntrons com inserções de SR2 e Perere-3. Fonte: Elaborado pela autora.



<b>Título:</b> Pipeline da análise para definir as inserções que correspondiam à ilhas CpG				
<b>Entrada</b>		<b>Estratégia de Análise</b>	<b>Software utilizado</b>	<b>Saída</b>
Fase 1	Sequência SR2	Considerando percentual GC > 65%, valor esperado de GC $\geq$ 0.65 e comprimento > 300 bp, verificar a existências de ilhas CpG	CpG Island Searcher	CpG Island SR2
	Sequência Perere-3			
Fase 2	CpG Island SR2	Verificar quais inserções correspondem à trechos do SR2 que equivalem à ilhas CpG	Módulo Perl	Íntrons Smp inserções SR2 CpG
	Íntrons Smp inserções SR2			
Fase 3	Íntrons Smp inserções SR2 CpG	Separar inserções por sentido senso e antisenso e verificar a distância da posição inicial da inserção em relação a extremidade inicial ou final do éxon mais próximo.	Módulo Perl Gnuplot	Posição inserções SR2 CpG
	ORTO-Íntrons -Smp			
Fase 4	Íntrons com inserções SR2 CpG	Considerando a quantidade total de íntrons do gene cujo íntron apresenta inserção, calcular o percentual de representação desse íntron em relação a todos os íntrons do conjunto, agrupando em uma matriz cujo eixo x representa a posição do íntron com inserção e o eixo y a quantidade de íntrons dos genes.	Módulo Perl Gnuplot	Heatmap com % de representação dos íntrons com inserção de ilhas CpG agrupados pela posição no gene
	Íntrons Smp			
	Total genes de Smp agrupados por qtde íntrons			
Fase 5	Posição inserções SR2 CpG	Para cada inserção real de ilha CpG, obter o tamanho da inserção e selecionar aleatoriamente na sequência do elemento SR2 trecho com a mesma dimensão e verificar se é correspondente ao trecho de ilha CpG	Módulo Perl Gnuplot	Histograma sobre valor observado e valor esperado para inserções ilhas CpG
	Sequência SR2			

Quadro 9-Pipeline da análise para definir as inserções que correspondiam à ilhas CpG. Fonte: Elaborado pela autora.



## Anexo III-Pipelines das análises nas regiões não traduzidas de genes, elementos cis-regulatórios e regiões intergênicas

<b>Título:</b> Pipeline do mapeamento das regiões intergênicas entre os genes ortólogos				
<b>Entrada</b>		<b>Estratégia de Análise</b>	<b>Software utilizado</b>	<b>Saída</b>
Fase 1	Genoma de <i>S.mansoni</i> com marcações éxons, íntrons, etc  listagem Genes Ortólogos	Verificar se os gene n e n+1 da fita do genoma estão presentes na listagem dos genes ortólogos	Módulo Perl	Regiões intergênicas dos genes ortólogos
Fase 2	Regiões intergênicas dos genes ortólogos  Sequências do genoma de <i>S.mansoni</i>	Montar arquivo fasta com as sequências dessas regiões.  Para otimizar as próximas análises, criar o cabeçalho de cada sequência da seguinte forma: > cromossomo-inicio-fim-tipoRegiaoIntergênica-gene1-gene2.	Módulo Perl	Sequências das regiões intergênicas

Quadro 10-Pipeline do mapeamento das regiões intergênicas entre os genes ortólogos. Fonte: Elaborado pela autora.

<b>Título:</b> Pipeline do mapeamento das inserções nas regiões intergênicas dos genes ortólogos				
<b>Entrada</b>		<b>Estratégia de Análise</b>	<b>Software utilizado</b>	<b>Saída</b>
Fase 1	Sequências das regiões intergênicas	Considerar alinhamentos com <i>evaluate</i> de $10^{-3}$ identidade > 85% e comprimento maior 50 bp	BlastN Script Shell	Inserções elementos de transposições nas regiões intergênicas
	Sequência SR2			
	Sequência Perere-3			

Quadro 11-Pipeline do mapeamento das inserções nas regiões intergênicas dos genes ortólogos. Fonte: Elaborado pela autora.

<b>Título:</b> Pipeline para definir as regiões dos elementos de transposições que mais se inseriram nas regiões intergênicas flanqueadas pelos genes ortólogos				
<b>Entrada</b>		<b>Estratégia de Análise</b>	<b>Software utilizado</b>	<b>Saída</b>
Fase 1	Blast das Inserções de SR2 nas regiões intergênicas	Verificar o intervalo do trecho inserido pelo elemento de transposição e acumular cada posição de base que pertence à esse trecho em um vetor que contém a dimensão total de bases do elemento de transposição	Móduli Perl Gnuplot	Distribuição das bases mais inseridas pelos elementos de transposição
	Blast das Inserções de Perere-3 nas regiões intergênicas			

Quadro 12-Pipeline para definir as regiões dos elementos de transposição que mais se inseriram nas regiões intergênicas flanqueadas pelos genes ortólogos. Fonte: Elaborado pela autora.

<b>Título:</b> Pipeline para descrever as características das inserções			
<b>Entrada</b>	<b>Estratégia de Análise</b>	<b>Software utilizado</b>	<b>Saída</b>
Fase 1 Inserções SR2 nas regiões intergênicas Inserções Perere-3 nas regiões intergênicas	Utilizando os resultados da execução do programa Blast, mediante dados contidos no cabeçalho de cada sequência, como exemplo abaixo, é possível definir o tamanho das regiões intergênicas Ex: CABG01000007-16597-97386-intergene55-Smp_173660-Smp_094060	Módulo Perl	Tamanho regiões intergênicas com inserção SR2 Tamanho regiões intergênicas com inserção Perere-3
Fase 2 Tamanho regiões intergênicas com inserção SR2 Tamanho regiões intergênicas com inserção Perere-3	Separar o tamanho das regiões intergênicas utilizando os 4 tipos de regiões definidas (3->3,5->3,5->5,3->5)	Script Shell	Tamanho regiões intergênicas com inserção SR2 por tipo de região Tamanho regiões intergênicas com inserção Perere-3 por tipo de região
Fase 3 Tamanho regiões intergênicas com inserção SR2 por tipo de região Tamanho regiões intergênicas com inserção Perere-3 por tipo de região	Utilizar um boxplot para representar a distribuição do tamanho dessas regiões	Script Shell Script R	Boxplot contendo o tamanho das regiões intergênicas com inserções dos elementos SR2 e Perere-3

Quadro 13-Pipeline para descrever as características das inserções. Fonte: Elaborado pela autora.

<b>Título:</b> Pipeline da análise realizada para verificar enriquecimento dos genes que flanqueiam regiões intergênicas com inserções			
<b>Entrada</b>	<b>Estratégia de Análise</b>	<b>Software utilizado</b>	<b>Saída</b>
Fase 4 Todos os genes que flanqueiam as regiões intergênicas Genes que flanqueiam regiões intergênicas com inserções SR2 Genes que flanqueiam regiões intergênicas com inserções Perere-3 Gene_association para Smp Gene Ontology OBO	Executar o software Ontologizer utilizando método estatístico Parent-Child, método Bonferroni para correção de erros das múltiplas comparações e nível de significância 0.05	Ontologizer	Enriquecimento dos genes que flanqueiam as regiões intergênicas com inserções de SR2 Enriquecimento dos genes que flanqueiam as regiões intergênicas com inserções de Perere-3

Quadro 14-Pipeline da análise realizada para verificar enriquecimento dos genes que flanqueiam regiões intergênicas com inserções. Fonte: Elaborado pela autora.

<b>Título:</b> Pipeline para realizar a predição da possível topologia das proteínas que podem ser transcritas a partir dos genes que flanqueiam regiões intergênicas com inserções do elemento SR2 e que apresentaram enriquecimento em relação a composição celular para proteínas intrínsecas à membrana				
<b>Entrada</b>		<b>Estratégia de Análise</b>	<b>Software utilizado</b>	<b>Saída</b>
Fase 4	Interface de execução do Ontologyzer	Selecionar item com enriquecimento e copiar listagem dos genes do conjunto de estudo que apresentaram o enriquecimento	Copiar/Colar Shell Script	genesEnriquecidos-SR2.txt
Fase 2	genesEnriquecidos-SR2.txt proteínas Smp.fasta	Criar arquivo em formato fasta com a sequência de aminoácidos dos genes que apresentaram enriquecimento	Módulo Perl	genesEnriquecidos-SR2.fasta
Fase 3	genesEnriquecidos-SR2.fasta	Fazer upload do arquivo contendo as sequências fasta no site do software TMHMM e selecionar a opção "one line per protein"  Copiar e colar o resultado exibido em um arquivo em formato txt	Software TMHMM	predicaoTopologias.txt
Fase 4	predicaoTopologias.txt	Filtrar os resultados utilizando o parâmetro PredHel o qual descreve o número de transmembranas hélices que foram preditas por N-best para cada gene  Gerar gráfico ilustrando a distribuição da quantidade de transmembranas hélices preditas	Script Shell  Gnuplot	gráfico ilustrando a quantidade de transmembrana hélices preditas para os genes com enriquecimento

Quadro 15-Pipeline para realizar a predição da possível topologia das proteínas que podem ser transcritas a partir dos genes que flanqueiam regiões intergênicas. Fonte: Elaborado pela autora.

<b>Título:</b> Pipeline para verificar proximidade das inserções das extremidades 5' das regiões intergênicas			
<b>Entrada</b>	<b>Estratégia de Análise</b>	<b>Software utilizado</b>	<b>Saída</b>
Fase 1	<p>Inserções Perere-3 nas regiões intergênicas</p> <p>O cabeçalho das sequências das regiões intergênicas contém as coordenadas de início e fim da região, como proposto anteriormente.</p> <p>Inserções SR2 nas regiões intergênicas</p> <p>Utilizando essas coordenadas, definir os segmentos S1 e S2</p>	Módulo Perl	<p>Segmentos com inserções SR2</p> <p>Segmentos com inserções Perere-3</p>
Fase 2	<p>Segmentos com inserções SR2</p> <p>Calcular a fração do segmento que compreende a extremidade 5' da região intergênica e a extremidade mais próxima da inserção (segmento S1 para intergene53, S2 para intergene35 e para as regiões intergênica55, foi considerado o menor segmento)</p> <p>Segmentos com inserções Perere-3</p> <p>Acumular essas frações em intervalos de 0.1 para gerar histograma</p>	<p>Módulo Perl</p> <p>Gnuplot</p>	<p>Inserções Intergênicas-near 5' SR2</p> <p>Distribuição para inserções SR2</p> <p>Inserções intergênicas-near5' Perere-3</p> <p>Distribuição para inserções Perere-3</p>

Quadro 16-Pipeline para verificar proximidade das inserções das extremidades 5' das regiões intergênicas. Fonte: Elaborado pela autora.



<b>Título:</b> Pipeline das análises realizadas sobre as inserções nas regiões intergênicas que correspondem à ilhas CpG				
<b>Entrada</b>		<b>Estratégia de Análise</b>	<b>Software utilizado</b>	<b>Saída</b>
Fase 1	CpG Island SR2  Inserções SR2 nas regiões intergênicas	Verificar quais inserções correspondem à trechos do SR2 que equivalem à ilhas CpG	Módulo Perl	Regiões intergênicas com inserções SR2-CpG
Fase 2	Regiões intergênicas com inserções SR2-CpG	Verificar a distância dessas inserções em relação à extremidade 5' utilizando o método de fração.	Módulo Perl  Gnuplot	Inserções Intergênicas-near 5' SR2-CpG  Distribuição para inserções SR2-CpG
Fase 3	Regiões intergênicas com inserções SR2-CpG	Comprimento médio do trecho que compreende a extremidade 5' dos genes e as inserções de ilhas CpG	Módulo Perl  Script R	Boxplot ilustrando o tamanho das
Fase 4	Regiões intergênicas com inserções SR2  Regiões intergênicas com inserções SR2-CpG  Gene association para Smp  Gene Ontology OBO	Executar o software Ontologizer utilizando método estatístico Parent-Child, método Bonferroni para correção de erros das múltiplas comparações e nível de significância 0.05	Ontologizer	Enriquecimento dos genes que flanqueiam as inserções de SR2-CpG
Fase 5	Posição inserções SR2 CpG  Sequência SR2	Para cada inserção real de ilha CpG, obter o tamanho da inserção e selecionar aleatoriamente na sequência do elemento SR2 trecho com a mesma dimensão e verificar se é correspondente ao trecho de ilha CpG	Módulo Perl  Gnuplot	Histograma sobre valor observado e valor esperado para inserções ilhas CpG

Quadro 17-Pipeline das análises realizadas sobre as inserções nas regiões intergênicas que correspondem à ilhas CpG. Fonte: Elaborado pela autora.



## Anexo IV-Pipelines das análises nas regiões presentes em mRNA

<b>Título:</b> Pipeline da análise para definir as inserções nas regiões codificantes			
<b>Entrada</b>	<b>Estratégia de Análise</b>	<b>Software utilizado</b>	<b>Saída</b>
Fase 1 EST-Smp Sequência SR2 Sequência Perere-3	Considerar alinhamentos com <i>evaluate</i> de $10^{-3}$ utilizando como <i>query</i> as sequências dos elementos SR2 e Perere-3 e como banco de dados as sequências EST.	BlastN	Trechos EST com inserção SR2  Trechos EST com inserção Perere-3
Fase 2 Trechos EST com inserção SR2 Trechos EST com inserção Perere-3	Considerar apenas alinhamentos que resultaram em pelo menos 50 bases de similaridade com os elementos SR2 e Perere-3, e calcular o percentual de cobertura da EST decorrente desse alinhamento.	Módulo Perl	Trechos EST com inserção SR2  Trechos EST com inserção Perere-3
Fase 3 Trechos EST com inserção SR2 Trechos EST com inserção Perere-3 EST-Smp	Mascarar o trecho EST que corresponde a sequência inserida pelos elementos SR2 e Perere-3. Considerar apenas EST cujo percentual de cobertura seja inferior a 100% e que apresente mais do que 30 bases sem máscara	Módulo Perl BioPerl	EST com trechos mascarados SR2  EST com trechos mascarados Perere-3
Fase 4 EST com trechos mascarados Genoma Smp	Considerar alinhamentos com <i>evaluate</i> de $10^{-3}$ utilizando como <i>query</i> os trechos da EST sem inserção e como banco de dados o genoma de Smp.	BlastN	Posição no genoma do trecho da EST sem inserção
Fase 5 Posição no genoma do trecho da EST sem inserção	Selecionar os trechos que alinharam com no máximo 3 cromossomos diferentes	Módulo Perl	Posição no genoma do trecho da EST sem inserção

continua

## continuação

Título: Pipeline da análise para definir as inserções nas regiões codificantes				
Fase 6	Trechos EST com inserção SR2 Trechos EST com inserção Perere-3 Genoma Smp	Considerar alinhamentos com <i>evaluate</i> de $10^{-3}$ utilizando como <i>query</i> os trechos com inserção dos elementos SR2 e Perere-3 e como banco de dados o genoma de Smp.	BlastN	Posição no genoma do trecho da EST com inserção
Fase 7	TCs do Gene Index EST que apresentaram inserção	Considerar alinhamentos com <i>evaluate</i> de $10^{-3}$ utilizando como <i>query</i> as EST que apresentaram inserção e como banco de dados as sequências de TCs	BlastN	TC-EST com inserção
Fase 8	TC-EST com inserção Trechos EST com inserção Trechos EST sem inserção	Considerar alinhamentos com <i>evaluate</i> de $10^{-3}$ utilizando como <i>query</i> os trechos da EST e como banco de dados as sequências de TCs	BlastN	ESTs mapeadas nas TCs
Fase 9	TC-EST com inserção	Predizer ORF para TC contendo pelo menos 120 bases e com identificação do codon de início	Módulo Perl Ugene	ORF das TC-EST com inserção
Fase 10	TC-EST com inserção Banco NR do NCBI	Considerar alinhamentos com <i>evaluate</i> de $10^{-3}$ utilizando como <i>query</i> as TCs e como banco de dados as sequências do NR	BlastP	NR-TC-EST com inserção

continua

continuação

<b>Título:</b> Pipeline da análise para definir as inserções nas regiões codificantes				
Fase 11	<p>Posição no genoma do trecho da EST com inserção</p> <p>Posição no genoma do trecho da EST sem inserção</p> <p>ESTs mapeadas nas TCs</p> <p>ORF das TC-EST com inserção</p> <p>NR-TC-EST com inserção</p>	<p>Identificar os trechos sem máscara e com máscara da EST que alinham de forma contínua e não-contínua no genoma e verificar onde a inserção do elemento de transposição provavelmente ocorreu na TC considerando a predição de ORF.</p>	<p>Módulo Perl BioGraphics</p>	<p>NR-TC-EST com inserção</p> <p>Representação gráfica em HTML</p>

Quadro 18-Pipeline da análise para definir as inserções nas regiões codificantes. Fonte: Elaborado pela autora.



## Anexo V-Genes com elementos Perere-3 nos íntrons e que apresentaram enriquecimento com a classe de processo biológico "*single-organism process*"

Tabela 3-Lista de genes, com elementos Perere-3, que apresentaram enriquecimento com o processo biológico "*single-organism process*"

ID Gene	ID Proteína	Função da proteína
Smp_152190	CCD74688	basic helix-loop-helix transcription factor,hes-related
Smp_196240	CCD75954	cadherin-related
Smp_078690	CCD58662	calponin homolog, putative
Smp_124530	CCD59609	dihydropyridine-sensitive l-type calcium channel, putative
Smp_199150	CCD60909	DNA repair protein rad51 homolog 3, r51h3,putative
Smp_142290	CCD58613	hyperpolarization activated cyclic nucleotide-gated potassium channel, putative
Smp_127250	CAZ28663	kinesin-associated protein, putative
Smp_165650	CCD60369	map-kinase activating death domain protein,putative
Smp_178780	CCD81980	meso-ectoderm gene expression control protein
Smp_031680	CCD81092	nAChR subunit (ShAR1-alpha-like)
Smp_139330	CCD81091	nAChR subunit (ShAR1-beta-like)
Smp_194780	CCD76537	neurotracting/lsamp/neurotrimin/obcam related cell adhesion molecule
Smp_179490	CCD59909	partial bestrophin-related
Smp_131150	CCD79841	polymyositis/scleroderma autoantigen-related
Smp_141530	CCD80662	putative anoctamin
Smp_032310	CCD82892	putative arp2/3 complex 21 kD subunit
Smp_055240	CCD79723	putative cadherin
Smp_004070	CCD74741	putative carbonic anhydrase
Smp_134190	CCD81282	putative cell adhesion molecule
Smp_040020	CCD80185	putative centromere protein A (cenp-A) (centromere autoantigen A)
Smp_160210	CCD79520	putative ceramide glucosyltransferase
Smp_105340	CCD74625	putative cyclic nucleotide-gated ion channel
Smp_131190	CCD80759	putative diacylglycerol kinase, zeta, iota
Smp_009550	CCD76365	putative DNA mismatch repair protein MLH1
Smp_055310	CCD76627	putative histone-lysine n-methyltransferase, seto7
Smp_060750	CCD76327	putative homeobox protein
Smp_133710	CCD76127	putative importin 7

continua

## continuação

ID Gene	ID Proteína	Função da proteína
Smp_162230	CCD80270	putative microtubule-associated protein tau
Smp_161850	CCD75426	putative monocarboxylate transporter
Smp_180570	CCD77092	putative nachr subunit
Smp_196800	CCD75914	putative phd finger protein
Smp_053820	CCD79934	putative preprotein translocase secy subunit (sec61)
Smp_170710	CCD79882	putative protein C10orf118 (CTCL tumor antigen HD-CL-01/L14-2)
Smp_172910	CCD78508	putative protein phosphatase pp2a regulatory subunit B
Smp_140770	CCD77195	putative protein phosphatase-7
Smp_132500	CCD78872	putative rab
Smp_160290	CCD83040	putative sodium/potassium-dependent atpase beta subunit
Smp_124240	CCD74761	putative sodium/potassium-dependent atpase beta subunit
Smp_016600	CCD80763	putative solute carrier family 1 (glial high affinity glutamate transporter)
Smp_174710	CCD80247	putative spindle assembly checkpoint component MAD1 (Mitotic arrest deficient protein 1)
Smp_194390	CCD76523	putative suppressor of cytokine signaling
Smp_104960	CCD79710	putative syntaxin
Smp_105020	CCD75777	putative titin
Smp_130890	CCD79633	putative transient receptor potential cation channel, subfamily m, member
Smp_210770	CCD75412	putative vacuolar protein sorting 26, vps26
Smp_136440	CCD82989	putative voltage-gated potassium channel
Smp_062560	CCD79605	putative wnt inhibitor frzb2
Smp_197770	CCD59236	rab6-interacting protein 2/elks/erc/cast, putative
Smp_162250	CCD80272	rpgr-interacting protein 1 related
Smp_163020	CCD59165	RU2S , putative
Smp_159570	CCD74724	septate junction protein
Smp_124540	CCD59608	serine-rich repeat protein , putative
Smp_125060	CCD59758	serine/threonine kinase
Smp_155280	CCD58634	Sp17 protein , putative
Smp_170210	CAZ36106	tektin related
Smp_054220	CCD59988	vacuolar ATP synthase subunit g, putative

OBS: Os genes relacionados a descrição "*hypothetical protein*" não constam na tabela



**Anexo VI-Genes com elementos Perere-3 nas regiões intergênicas  
que flanqueiam esses genes e que apresentam  
enriquecimento com relação ao processo biológico  
"aromatic compound biosyntetic proc"**

Tabela 4-Lista de genes, com elementos Perere-3, que apresentaram enriquecimento com o processo biológico "aromatic compound biosyntetic proc"

<b>ID Gene</b>	<b>ID Proteína</b>	<b>Função da proteína</b>
Smp_152190	CCD74688	basic helix-loop-helix transcription factor,hes-related
Smp_072100	CCD60569	ccr4-not transcription complex, putative
Smp_194820	CCD82556	DNA repair and recombination protein rad54-related
Smp_068230	CCD81987	DNAj-like protein
Smp_150810	CCD79411	double histidine kinase DhkD-like
Smp_128330	CCD81789	ets-related
Smp_068780	CCD77680	FTZ-F1 nuclear receptor-like protein
Smp_081620	CCD79588	gsx family homeobox protein
Smp_131000	CCD79807	harp (smarcal1)-related
Smp_053520	CCD79410	homeobox protein aristaless-related
Smp_017020	CCD83066	homeobox protein vsx-1-related
Smp_172430	CCD75517	insulinoma-associated protein (ia-1)-related
Smp_132810	CCD58502	IPR001092 Basic helix-loop-helix dimerisation region bHLH,domain-containing protein
Smp_061570	CCD80211	lung cancer metastasis-related (lcmr1) protein
Smp_144380	CCD78762	msx family homeobox protein
Smp_027990	CCD58471	neural gene activation protein
Smp_072470	CCD76422	neurogenic differentiation factor
Smp_150630	CCD59712	nuclear receptor, putative
Smp_198750	CCD59555	pinn, putative
Smp_156890	CCD80132	protein kinase
Smp_168600	CCD82607	putative aryl hydrocarbon receptor
Smp_124090	CCD74790	putative b-cell lymphoma/leukemia
Smp_129500	CCD78309	putative dna-directed rna polymerase I largest subunit
Smp_040710	CCD78769	putative dna-directed rna polymerase II 13.3 kD polypeptide
Smp_140100	CCD81028	putative double-stranded rna-binding protein zn72d

continua

## continuação

ID Gene	ID Proteína	Função da proteína
Smp_127020	CCD79224	putative ets
Smp_126530	CCD75986	putative ets
Smp_145640	CCD81728	putative forkhead protein/ forkhead protein domain
Smp_086270	CCD77557	putative forkhead protein/ forkhead protein domain
Smp_054350	CCD82559	putative forkhead protein/ forkhead protein domain
Smp_070760	CCD78376	putative gas41
Smp_174700	CCD80248	putative hepatocyte nuclear factor 4-alpha (hnf-4-alpha)
Smp_027300	CCD79779	putative histone-lysine n-methyltransferase, suv9
Smp_158000	CCD78085	putative homeobox protein
Smp_136900	CCD79791	putative homeobox protein distal-less dlx
Smp_146580	CCD77339	putative homeobox protein knotted-1
Smp_147790	CCD77959	putative homothorax homeobox protein
Smp_142560	CCD75724	putative insulinprotein enhancer protein isl
Smp_060240	CCD82574	putative lipopolysaccharide-induced transcription factor regulating tumor necrosis factor alpha
Smp_060220	CCD82575	putative lipopolysaccharide-induced transcription factor regulating tumor necrosis factor alpha
Smp_060210	CCD82576	putative lipopolysaccharide-induced transcription factor regulating tumor necrosis factor alpha
Smp_174320	CCD81693	putative lozenge
Smp_144180	CCD76560	putative mixed-lineage leukemia protein, mll
Smp_168610	CCD82608	putative myelin transcription factor 1, myt1
Smp_154340	CCD76258	putative nuclear factor Y transcription factor subunit B homolog
Smp_128120	CCD77872	putative polybromo-1
Smp_089630	CCD74654	putative polyglutamine binding protein
Smp_212260	CCD80268	putative retinoblastoma-binding protein 4 (rbbp4)
Smp_040350	CCD76554	putative rfx5
Smp_165270	CCD81979	putative single-minded
Smp_157540	CCD76481	putative smad
Smp_130200	CCD77387	putative steroidogenic factor 1 (stf-1) (sf-1) (adrenal 4 binding protein) (steroid hormone receptor ad4bp) (fushi tarazu factor homolog 1)
Smp_210680	CCD78446	putative suppression of tumorigenicity
Smp_151070	CCD79922	putative suppressor of ty
Smp_180690	CCD81421	putative tiptop

continua

continuação

ID Gene	ID Proteína	Função da proteína
Smp_157840	CCD82312	putative transcription factor ap-2 gamma
Smp_040390	CCD76559	putative transcription factor sp5/buttonhead
Smp_176710	CCD80405	putative transcriptional factor nfil3/e4bp4
Smp_151810	CCD79850	putative voltage-gated potassium channel
Smp_160730	CCD76051	putative zinc finger protein
Smp_155250	CCD80213	putative zinc finger protein
Smp_068240	CCD81988	putative zinc finger protein
Smp_049580	CCD76239	putative zinc finger protein
Smp_044870	CCD77281	putative zinc finger protein
Smp_015840	CCD81874	putative zinc finger protein
Smp_155720	CCD80885	serine/threonine kinase
Smp_090980	CCD76148	serine/threonine kinase
Smp_058620	CCD59503	serine/threonine kinase
Smp_180820	CCD75778	sox transcription factor
Smp_161600	CCD59090	sox transcription factor
Smp_097730	CCD59556	srf homolog, putative
Smp_132800	CCD58503	STARP antigen , putative
Smp_062530	CCD79602	suppressor of variegation 4-20-related
Smp_154640	CCD75795	SWI/SNF-related
Smp_053140	CCD59714	tip60, putative
Smp_134790	CCD78772	transcription initiation factor iid, 28 kD subunit-related
Smp_066720	CCD58906	transcriptional adaptor 2 (ada2)-related
Smp_000830	CCD76678	homeobox protein smox-3
Smp_002640	CCD78555	hox protein Smox1
Smp_003280	CCD74820	hypothetical protein
Smp_003000	CCD79487	putative cell division protein kinase 9-B (EC 2.7.11.22) (EC 2.7.11.23) (Cyclin-dependent kinase 9-B)
Smp_004640	CCD77821	putative dna-directed RNA polymerase I
Smp_014080	CCD82804	putative lipopolysaccharide-induced transcription factor regulating tumor necrosis factor alpha
Smp_009020	CCD79212	putative rac guanyl-nucleotide exchange factor
Smp_003900	CCD74759	putative t-box transcription factor tbx20

OBS: Os genes relacionados a descrição "*hypothetical protein*" não constam na tabela



**Anexo VII-Genes com elementos SR2 nas regiões intergênicas que flanqueiam esses genes e que apresentam enriquecimento com relação a componente celular "*intrinsic to membrane*" e predição para mais do que 5 hélices transmembranares**

Tabela 5-Lista de genes, com elementos SR2, que apresentaram enriquecimento com componente celular "*intrinsic to membrane*" e predição para mais do que 5 hélices transmembranares

<b>ID Gene</b>	<b>ID Proteína</b>	<b>Função da proteína</b>
Smp_139080	CCD81324	bestrophin-related
Smp_055200	CCD79729	dolichyl-diphosphooligosaccharide-proteinglycosyl transferase-related
Smp_051810	CCD79095	elongation of fatty acids protein 1
Smp_143690	CCD80168	feline leukemia virus subgroup C receptor-related
Smp_132730	CCD58519	G-protein coupled receptor fragment, putative
Smp_171720	CCD79484	lipid phosphate phosphatase-related
Smp_162110	CCD60521	multidrug resistance pump, putative
Smp_076520	CCD60288	organic solute transporter, putative
Smp_181230	CCD59218	phospholipid-transporting atpase
Smp_181150	CCD76445	putative abc transporter
Smp_147070	CCD80736	putative amino acid transporter
Smp_180950	CCD77771	putative anion exchange protein
Smp_176940	CCD76435	putative cationic amino acid transporter
Smp_144970	CCD79673	putative copper-transporting atpase 1, 2 (copper pump 1,2)
Smp_136400	CCD78485	putative cytochrome B561
Smp_066900	CCD79263	putative innexin
Smp_129820	CCD82812	putative multidrug resistance protein 1 (ATP-binding cassette C1)
Smp_135490	CCD79304	putative multidrug resistance pump
Smp_133550	CCD81459	putative neuropeptide F-like receptor
Smp_141880	CCD81490	putative neuropeptide receptor
Smp_157640	CCD76490	putative peptide (FMRFamide/neurokinin-3)-like receptor
Smp_053820	CCD79934	putative preprotein translocase secy subunit (sec61)
Smp_090870	CCD76161	putative protein transport protein Sec13

continua

## continuação

ID Gene	ID Proteína	Função da proteína
Smp_150500	CCD78079	putative recs1 protein (responsive to centrifugal force and shear stressprotein 1 protein)
Smp_162980	CCD83008	putative rhodopsin-like orphan GPCR
Smp_129810	CCD82811	putative rhodopsin-like orphan GPCR
Smp_056080	CCD76613	putative rhodopsin-like orphan GPCR
Smp_175010	CCD77967	putative solute carrier family 17, member 7 (vesicular glutamate transporter)
Smp_164830	CCD81221	putative ssm4 protein
Smp_165170	CCD78634	putative transient receptor potential cation channel,subfamily m, member
Smp_171120	CCD77728	putative udp-galactose transporter
Smp_019980	CCD75110	putative vacuole membrane protein
Smp_081250	CCD75136	putative voltage-gated potassium channel
Smp_053020	CCD80014	putative zinc finger protein
Smp_161500	CCD59764	rhodopsin-like orphan GPCR, putative
Smp_013950	CCD58923	selectively expressed in embryonic epithelia protein-1 , putative
Smp_131350	CCD80783	sodium-bile acid cotransporter related
Smp_066150	CAZ34096	solute carrier family 35 member C2, putative
Smp_031220	CCD81323	solute carrier family 35-related
Smp_167630	CCD59038	solute carrier family, putative
Smp_123280	CCD80372	transport proetin (smap-4-related)

OBS: Os genes relacionados a descrição "*hypothetical protein*" não constam na tabela

## Anexo VIII-Análise das inserções presentes em mRNA para trechos inseridos de forma não contínua em relação a outros trechos das ESTs

Para 5 casos, onde o trecho similar a sequência dos elementos de transposição não se apresentou de forma adjacente aos demais trechos da EST, foram realizadas análises para verificar se esse intervalo entre os trechos poderiam ser fruto de um fenômeno de *splicing*, *trans-splicing* ou podem ser artefatos decorrentes da montagem do genoma.

Para a EST CD202813, as análises descritas no capítulo 5 indicam que a região do elemento compreende as bases 2 à 251 da EST, enquanto o outro trecho da EST, corresponde as bases de 263 à 330. A distância entre esse trecho corresponde a 2.232 bases.

Realizando a predição de éxons através do *software* Spidey, como ilustra a Figura 64, é possível observar que as bases correspondentes ao trecho inserido pelo retrotransposon (2-251) correspondem as bases do éxon 1. O outro trecho da EST (263-330) corresponde as bases do éxon 2.

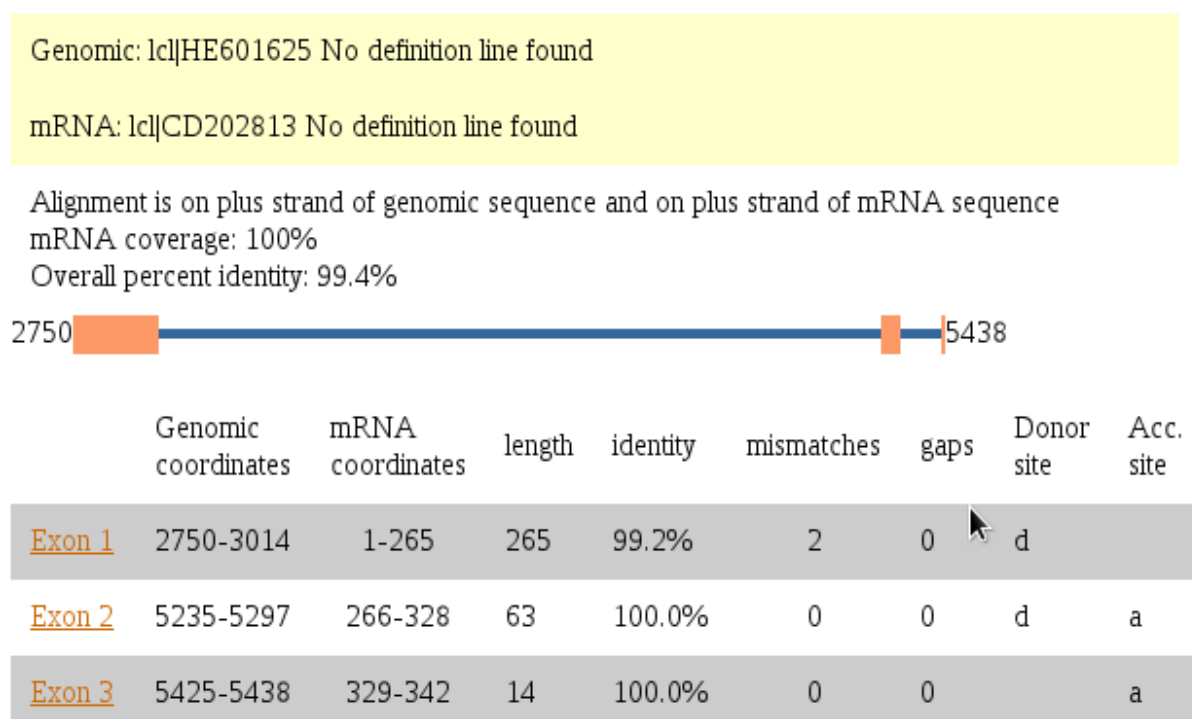


Figura 64-Visualização gráfica do *software* Spidey para verificar a predição dos éxons da EST. Para a sequência genômica, foram consideradas 3.000 bases ao redor do trecho similar ao elemento de transposição e do outro trecho da EST. Fonte: Elaborada pela autora.

Observando os dados da coluna “Genomic Coordinates” da Figura 64, é possível constatar que a distância entre esses éxons é de aproximadamente 2.221 bases (5.235-3.014), como indicado anteriormente pelas análises realizadas nesse trabalho.

Para esse caso foi possível observar que o trecho inserido pelo retrotransposon contém *stop codon*, como ilustra a Figura 65.

**Exon 1: 2750-3014 (genomic); 1-265 (mRNA)**

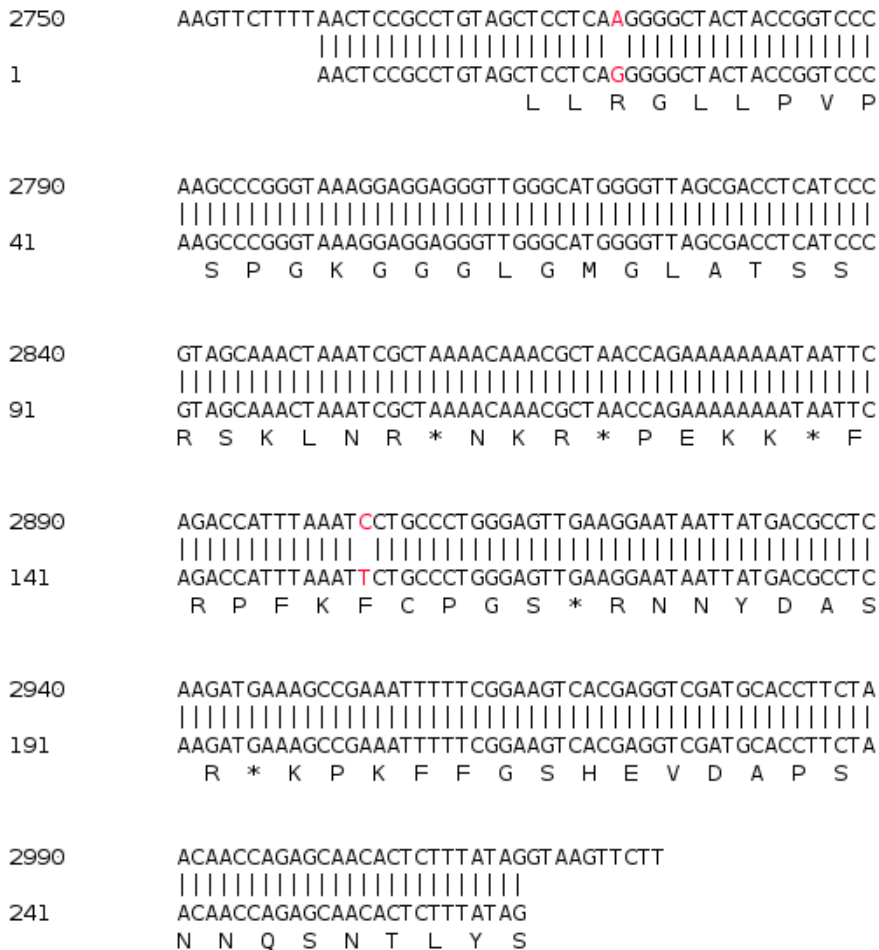


Figura 65-Visualização gráfica do *software* Spidey para verificar a predição dos éxons da EST CD202813. O trecho inserido pelo elemento de transposição, o qual corresponde ao éxon apresentada stop codons, representados nessa figura pelo símbolo \*. Fonte: Elaborada pela autora.



O mesmo tipo de análise foi realizada para as demais EST, onde os dados obtidos pelas análises desse trabalho e pelo *software* Spidey serão apresentados, a seguir, de forma mais sucinta.

	ID EST	Trecho Retrotransposon (Bases)	Outro trecho EST (Bases)	Distância entre trechos (Bases)
Análise desse trabalho	CF495598	3-92	143-167	586
Resultados Spidey	Figura 66	Éxon 1	Éxon 4	(3.583-3.001) 582

Quadro 19-Resumo das análises para verificação da possível ocorrência de *splicing* entre o trecho similar ao elemento de transposição e o restante da sequência da EST CF495598. Fonte: Elaborado pela autora.

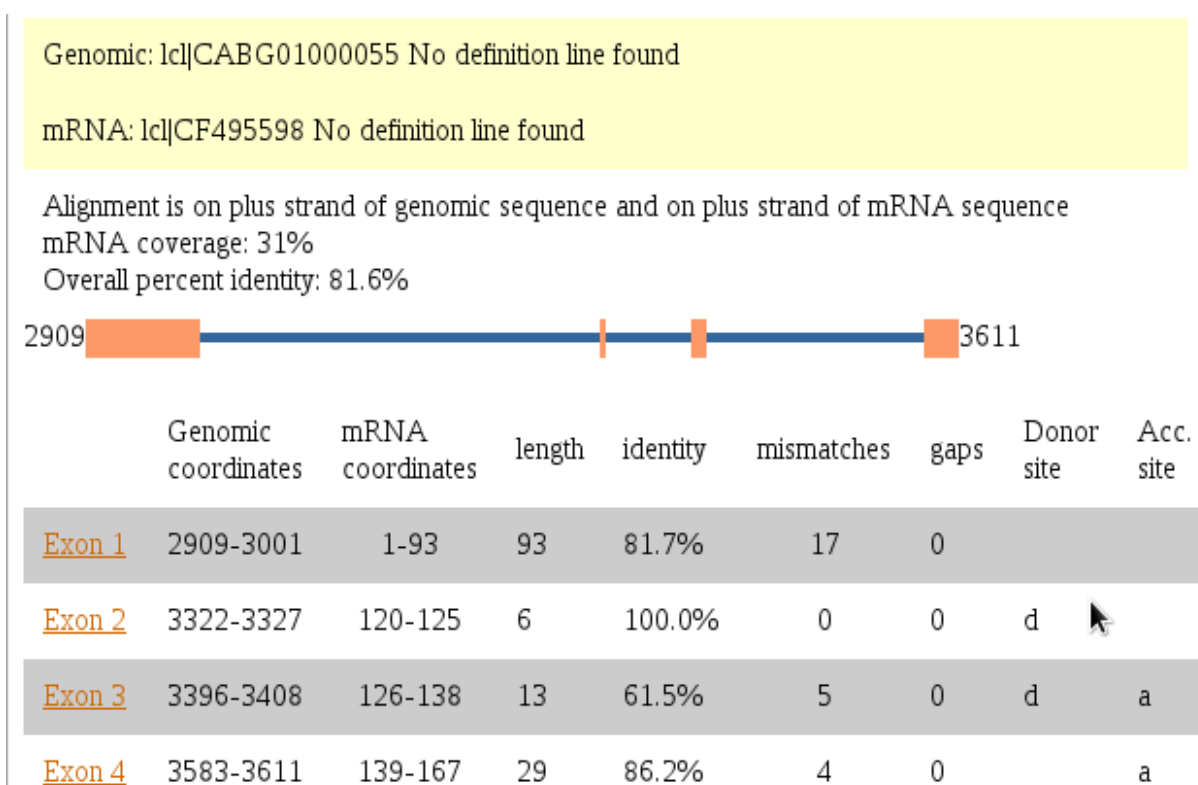


Figura 66-Visualização gráfica do *software* Spidey para verificar a predição dos éxons da EST. Para a sequência genômica, foram consideradas 3.000 bases ao redor do trecho similar ao elemento de transposição e do outro trecho da EST CF495598. Fonte: Elaborada pela autora.

	ID EST	Trecho Retrotransposon (Bases)	Outro trecho EST (Bases)	Distância entre trechos (Bases)
Análise desse trabalho	CF495807	457-529	1-433	1.267
Resultados Spidey	Figura 67	Éxon 3, 4 e 5	Éxon 1	(1.734-3.001) 1.267

Quadro 20-Resumo das análises para verificação da possível ocorrência de *splicing* entre o trecho similar ao elemento de transposição e o restante da sequência de EST CF495807. Fonte: Elaborado pela autora.

Genomic: lcl|HE601624 No definition line found

mRNA: lcl|CF495807 No definition line found

Alignment is on minus strand of genomic sequence and on plus strand of mRNA sequence

mRNA coverage: 87%

Overall percent identity: 96.6%

3433  1676

	Genomic coordinates	mRNA coordinates	length	identity	mismatches	gaps	Donor site	Acc. site
<a href="#">Exon 1</a>	3001-3433	1-433	433	99.8%	1	0	d	
<a href="#">Exon 2</a>	2858-2861	434-437	4	100.0%	0	0		a
<a href="#">Exon 3</a>	2712-2723	455-466	12	66.7%	4	0	d	
<a href="#">Exon 4</a>	2439-2455	467-483	17	58.8%	7	0	d	a
<a href="#">Exon 5</a>	1676-1734	484-542	59	89.8%	6	0		a

Figura 67-Visualização gráfica do *software* Spidey para verificar a predição dos éxons da EST. Para a sequência genômica, foram consideradas 3.000 bases ao redor do trecho similar ao elemento de transposição e do outro trecho da EST CF495807. Fonte: Elaborada pela autora.

	ID EST	Trecho Retrotransposon (Bases)	Outro trecho EST (Bases)	Distância entre trechos (Bases)
Análise desse trabalho	CF500292	10-107	136-485	305
Resultados Spidey	Figura 68	Éxon 1	Éxon 3	(3003-3306) 303

Quadro 21-Resumo das análises para verificação da possível ocorrência de *splicing* entre o trecho similar ao elemento de transposição e o restante da sequência de EST CF500292.Fonte: elaborado pela autora.

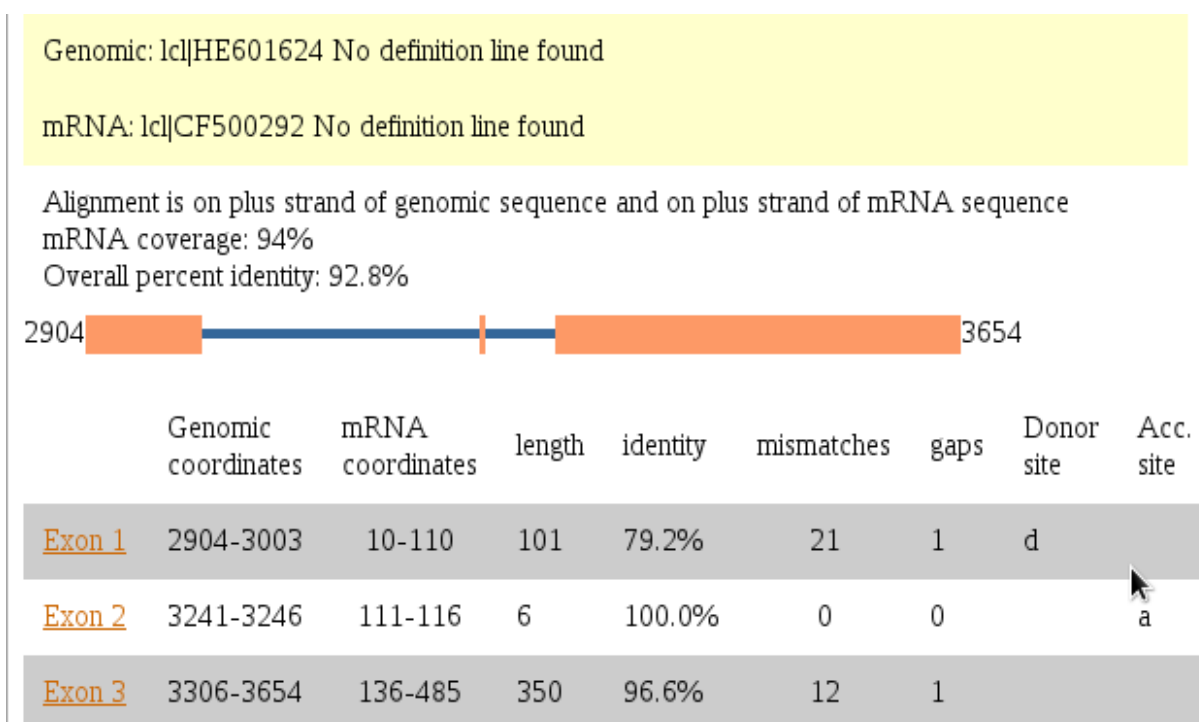


Figura 68-Visualização gráfica do *software* Spidey para verificar a predição dos éxons da EST. Para a sequência genômica, foram consideradas 3.000 bases ao redor do trecho similar ao elemento de transposição e do outro trecho da EST CF500292.Fonte: Elaborada pela autora.

Para um dos casos, embora tenha sido observada uma distância de 19 bases entre os trechos da EST, a predição de éxons pelo *software* Spidey identificou as bases dos dois trechos em um único éxon, conforme dados do Quadro 22 e da Figura 69.

	ID EST	Trecho Retrotransposon (Bases)	Outro trecho EST (Bases)	Distância entre trechos (Bases)
Análise desse trabalho	CD112384	9-59	68-121	19
Resultados Spidey	Figura 69	Éxon 1	Éxon 1	0

Quadro 22-Resumo das análises para verificação da possível ocorrência de *splicing* entre o trecho similar ao elemento de transposição e o restante da sequência de EST CD112384. Fonte: Elaborado pela autora.

Genomic: lcl|HE601626 No definition line found

mRNA: lcl|CD112384 No definition line found

Alignment is on plus strand of genomic sequence and on plus strand of mRNA sequence

mRNA coverage: 20%

Overall percent identity: 78.6%

2960  3073

	Genomic coordinates	mRNA coordinates	length	identity	mismatches	gaps	Donor site	Acc. site
<a href="#">Exon 1</a>	2960-3073	19-121	103	78.6%	22	11		

Figura 69-Visualização gráfica do *software* Spidey para verificar a predição dos éxons da EST. Para a sequência genômica, foram consideradas 3.000 bases ao redor do trecho similar ao elemento de transposição e do outro trecho da EST CD112384. Fonte: Elaborada pela autora.