UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS

EDUARDO CHEROBIN MARTINS

Análise da influência de elementos de transposição do clado CR1 na arquitetura do genoma do *Schistosoma mansoni*

São Carlos
2022

EDUARDO CHEROBIN MARTINS

Análise da influência de elementos de transposição do clado CR1 na arquitetura do genoma do *Schistosoma mansoni*

Dissertação apresentada no programa de pós graduação do Instituto de Física de São Carlos, Universidade de São Paulo a fim de obter o título de Mestre em Ciências

Área de concentração: Física Biomolecular
Orientadores: Prof. Dr. Ricardo De Marco (*in memoriam*) e Profa. Dra. Ana Paula Ulian de Araújo.
Coorientadora: Profa. Dra. Gisele Strieder Philippsen.

Versão corrigida
(Versão original disponível na Unidade que aloja o programa)

São Carlos
2022

Martins, Eduardo Cherobin
Análise da influência de elementos de transposição do clado CR1 na arquitetura do genoma do *Schistosoma mansoni* / Eduardo Cherobin Martins; orientadora Ana Paula Ulian de Araujo; co-orientadora Gisele Strieder Philippsen - versão corrigida -- São Carlos, 2022.
98 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Física Biomolecular) -- Instituto de Física de São Carlos, Universidade de São Paulo, 2022.

1. Retrotransposons. 2. Schistosoma mansoni. 3. Bioinformática. I. Araujo, Ana Paula Ulian de, orient. II. Philippsen, Gisele Strieder, co-orient. III. Título.

Ao Professor Ricardo

AGRADECIMENTOS

Ao Ricardo DeMarco, por, sem me levar pela mão, ter me mostrado a beleza da ciência. Pelos esforços para tornar possível a minha permanência e, nos momentos pandêmicos difíceis, pela paciência e compreensão nos momentos em que eu mais precisava.

À Ana Paula por me adotar e, mesmo antes disso, por acreditar e me ajudar quando eu vagava sem rumo. Pelos frequentes e pacientes empurrõezinhos e conselhos, mesmo quando parecia que este trabalho jamais seria concluído. Por me dar esperança para continuar e ter me estimulado a cada pequena conquista. Pela amizade.

À Gisele Strieder, Gi, por mergulhar de cabeça no projeto, por também ter me adotado e, insistentemente e pacientemente, ter acreditado que seria possível. Por ter andado ao meu lado, inclusive nos momentos de crise e preocupação. Pelos caminhos e artigos, explicações e conversas. Pela amizade.

À Luíza Zuvanov, pela vivência e ensinamentos. Pela amizade e parceria. Por ser um exemplo a ser seguido e por, sem hesitação, sempre me receber de braços abertos. Por me estimular quando as análises não davam certo e pelas conversas inspiradoras e transformadoras.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), por fomentar o projeto e tornar esse trabalho possível.

À Gabriela Ramalho, Gabi, por me ajudar a ver que eu era capaz, ajudar-me a recuperar a autoestima e a ter coragem para tomar atitudes práticas para resolver problemas iminentes. Por me ajudar a segurar a barra.

À Larissa Miranda Morth, por ter me mostrado um novo mundo dentro de mim, e ter me ajudado a explorá-lo. Ajudando-me a buscar formas e forças para escrever, a aceitar-me como sou e meus momentos e sentimentos.

À Etefania Pavarina, Fanni, por estar ao meu lado (e me aguentar) nos momentos bons e ruins. Por todo o carinho que me ajudou a seguir em frente, por ser um exemplo de pesquisadora no quarto do lado, por todo o amor, cuidado e afeto.

À Adriana Thaís (Catatau), pelas conversas, confidências, pelas companhias, pelos rolês, e por ter cuidado tanto de mim. Por ter me ajudado a segurar as barras e sido uma presença acolhedora e amiga.

À Paula Comminato, por ter me acolhido no meu momento mais difícil da pandemia, pela presença, amizade e parceria. Por ser alguém que eu possa confiar e me sentir seguro. Por me mostrar, com suas ações, uma nova forma de ver o mundo.

Ao Matheus Pereira (Teta), por ser meu grande amigo, por ser a alegria dos dias pré-pandemia, companhia de almoços e jantas no bandejão, de risadas, de rolês, de encher a cara e chorar, e de cantar loucamente ao som de um violão. E por me estender a mão independentemente da razão.

Ao Thiago Freire (Corre), por ser meu parceiro do dia a dia, nas conversas, no videogame e por prontamente me ajudar a não ficar sobrecarregado com os diversos problemas que enfrentamos.

À Luisa (Canadá) e Igor Mesquita (Pinguim) pelo acolhimento, conversas e carinho, pelos conselhos e inspiração.

Ao grupo acaso de teatro, por ser um espaço seguro para eu viver a minha arte.

Ao alojamento, por sempre me receber de braços abertos.

À Alexandra Elbakyan, pela iniciativa visionária de democratizar o acesso ao conhecimento.

E a todos que também participaram comigo nessa jornada: Beatriz Caroline(Bia), Diogo Maciel, Alessandro Pereira, Jéssica Puppo, Everton Edesio, Giulie Cherobin Martins, Izabela Dias do Nascimento.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

“E quando tudo parecia não ter caminho
escolhas e mais escolhas foram surgindo
E as escolhas viraram códigos
e os códigos viraram gráficos

E dos gráficos nasceram teorias
e as teorias alimentaram a alma
que, agora com sede, buscava rotas

E as rotas viraram acordos
e o compromisso de ir mais fundo
e seguindo seus passos
fomos inventando um novo mundo

E encontrando padrões
em linhas genômicas
uma flor, em meio ao deserto
e fomos chegando mais e mais perto

E a flor tinha características
e era rodeada de pistas
e a cada fim, novas perguntas

Até que veio o afastamento
E um fim, seguido de luto
não terminaríamos a aventura juntos

Não soube o que fazer
É o fim? Pensei
Mas vocês chegam, com lágrimas de perda
oferecendo a mão, para quem já não tinha
destino

E eu aceitei, chorrindo
e numa labuta para cada passo
nós redescobrimos o espaço
concluímos significados

Em meio a reviravoltas e hiatos
crises, risos,
foram surgindo pedaços
E tudo tomou forma

Para que hoje, com sorriso
brandemos: Ficou bonito!
”

Autoria própria

RESUMO

MARTINS, E. C. **Análise da influência de elementos de transposição do clado CR1 na arquitetura do genoma do *Schistosoma mansoni.*** 2022. 98 p. Dissertação (Mestrado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2022.

Elementos de transposição (TEs) constituem grande parte dos genomas eucarióticos e podem influenciar de maneira significativa as atividades genômicas. Observações prévias de colinearidade entre elementos transponíveis e genes *downstream* próximos levantaram o questionamento sobre um possível padrão de inserção envolvendo TEs não-LTR da família CR1, presentes em *S. mansoni*. Neste contexto, utilizando ferramentas de bioinformática, foram analisados dados de sequenciamento em larga escala do genoma de *S. mansoni*, juntamente com dados de RNA-seq e CHIP-seq, a fim de avaliar as regiões em que as cópias desses elementos estão inseridas. Estas análises buscaram verificar a existência de tendências quanto à localização relativa entre inserções e genes, níveis de transcrição de genes próximos e localização de regiões ricas em histonas com modificações epigenéticas. Parte significativa das cópias de elementos do clado CR1 foram encontradas sendo colineares a genes *downstream* e próximas a genes com transcrição acima do esperado para inserções ao acaso, o que pode significar alguma preferência de inserção para tais regiões. Além disso, elementos do clado CR1 com região codificante de um domínio PHD foram encontrados com população significativa a uma distância de menos de 1000 pb de histonas com a modificação H3K4me3, a qual é reconhecida pelo domínio PHD. Este resultado sugere um possível direcionamento das inserções destes TEs, mediado pelo domínio PHD, embora análises de inserções *de novo* sejam necessárias para verificar esta hipótese. O entendimento dessas influências poderá ajudar a compreender o genoma do parasito e um possível mecanismo de direcionamento de inserções para regiões de histonas com a modificação H3K4me3, se comprovada a relação.

Palavras-chave: Retrotransposons. *Schistosoma mansoni.* Bioinformática.

ABSTRACT

MARTINS, E. C. **Analysis of the influence of CR1 clade transposition elements on *Schistosoma mansoni* genome architecture.** 2022. 98 p. Dissertation (Master in Science) - Institute of Physics of São Carlos, University of São Paulo, São Carlos, 2022.

Transposable elements (TEs) constitute a large part of eukaryotic genomes and can significantly influence genomic activities. Previous observations of collinearity between transposable elements and nearby downstream genes raised the question of a possible insertion pattern involving non-LTR TEs of the CR1 family, present in *S. mansoni*. In this context, using bioinformatics tools, large-scale sequencing data of the *S. mansoni* genome were analyzed, together with RNA-seq and CHIP-seq data, in order to evaluate the regions where copies of these elements are inserted. These analyzes sought to verify the existence of trends regarding the relative location between insertions and genes, transcription levels of nearby genes and location of histone-rich regions with epigenetic modifications. A significant part of the copies of elements of the CR1 clade were found to be collinear to downstream genes and close to genes with transcription above the expected for random insertions, which may indicate some insertion preference for such regions. Furthermore, elements of the CR1 clade with coding region for a PHD domain were found with a significant population at a distance of less than 1000 bp from histone modification H3K4me3, which is recognized by the PHD domain. This result suggests a possible targeting of these TEs, mediated by the PHD domain, although analyzes of insertions *de novo* are necessary to verify this hypothesis. The understanding of these influences may help to understand the genome of the parasite and a possible mechanism of targeting insertions for histone with the H3K4me3 modification.

Keywords: Retrotransposons. *Schistosoma mansoni.* Bioinformatics.

SUMÁRIO

1 INTRODUÇÃO

A compreensão dos genomas eucarióticos têm progredido ao passo que mais e mais organismos têm sua informação genética sequenciada, mostrando que grande parte dos genomas é composto de DNA repetitivo, como por exemplo cerca de 54% do genoma humano.[1]

1.1 Elementos de transposição

Dentre os elementos repetitivos do DNA, existem os elementos de transposição ou elementos transponíveis (TEs) que são sequências com tipicamente 100 a 10000 pares de bases aptas a realizar a transposição, ou seja, uma movimentação de um lugar do genoma para outro, utilizando-se de processos enzimáticos específicos, frequentemente realizado por proteínas codificadas pelo próprio elemento.[2] Essa movimentação gera, ocasionalmente ou obrigatoriamente, a replicação do elemento, que acontece de forma cumulativa no genoma. Por isso, elementos de transposição estão presentes em quase todas as espécies eucarióticas conhecidas, com expressiva representação no genoma (mais de 80% do genoma das plantas e 3% a 45% em metazoários.[3]

Elementos de transposição já foram descritos como parasitas genômicos, pois garantem a sua sobrevivência reproduzindo-se mais rápido que o hospedeiro que o carrega. [4] Com pouquíssimas exceções, todos os genomas eucarióticos conhecidos apresentam TEs, o que sugere que estes elementos sejam componentes dos genomas eucarióticos a muito tempo.[2,5] Levando-se também em consideração a sua replicação cumulativa no genoma, pode-se esperar que elementos transponíveis tenham um grande impacto na trajetória evolutiva de seus hospedeiros. Assim, estudar a abundância e a diversidade de TEs entre as diferentes espécies é uma etapa essencial para o entendimento de como esses elementos impactam e contribuem para a diversificação e a biologia das espécies.

1.1.1 Mecanismos de transposição

Os elementos de transposição eucarióticos foram primariamente divididos em duas classes: retrotransposons ou de classe I, que utilizam mecanismos de “copiar e

colar” e têm um RNA como intermediário; e TEs de DNA ou de classe II, que utilizam majoritariamente mecanismos de “recortar e colar” e têm um DNA como intermediário.[6]

Os elementos da classe I mobilizam-se por meio de um mecanismo pelo qual um intermediário de RNA é transcrito reversamente em uma cópia de cDNA que é integrada em outras partes do genoma. Dependendo do mecanismo de replicação e integração utilizado, os TEs de classe I são divididos em três subclasses principais: (a) elementos de repetição terminal longa (LTR), (b) elementos não-LTR e (c) elementos mobilizados pela recombinase de tirosina. Uma descrição detalhada sobre as subclasses pode ser encontrada em Wells e Feschotte (2020).

Elementos do tipo LTR possuem sequências LTR (*long terminal repeats*), que são sequências idênticas nas duas extremidades do elemento e são reconhecidas pela RNA polimerase II a fim de iniciar a transcrição (Figura 1). As sequências codificam as proteínas transcriptase reversa, protease, integrase e gag. Após a tradução no citoplasma celular, as proteínas gag organizam-se de forma a encapsular a transcriptase reversa, o mRNA do TE e a integrase; nessa etapa o mRNA é copiado para cDNA. Em seguida, a integrase insere o cDNA no genoma novamente.

Neste trabalho, em função dos TEs estudados serem de classe I do tipo não-LTR, a seguir será feita uma breve descrição destes. TEs do tipo não-LTR possuem, em sua maioria, dois quadros de leitura abertos (em inglês: *open reading frame* ou ORF) que são comumente chamados de ORF1 e ORF2. A ORF2 possui sequências que codificam uma proteína com funções de transcriptase reversa e endonuclease, duas enzimas envolvidas no processo de replicação do elemento de transposição. A ORF1, por sua vez, já foi reportada como codificante de proteínas com funções de chaperona[7] e de interação com RNA,[8] sendo dispensável ou ausente em alguns grupos.[2]

Após a transcrição do TE não-LTR, o mRNA segue para o citoplasma para a tradução (Figura 1). As proteínas do TE e o mRNA associam-se para formar uma ribonucleoproteína e adentrar ao núcleo.[9] No local de inserção, a endonuclease (que pode ter função de endonuclease AP ou de restrição) cliva o DNA alvo de forma que a extremidade clivada sirva de *primer* para a transcrição reversa do mRNA do TE, realizada pela transcriptase reversa. O trecho sintetizado serve de molde para a fita complementar. Existe a possibilidade da transcriptase reversa deixar o sítio no meio

do processo, resultado em uma cópia incompleta, o que ocasiona a maior representatividade da extremidade 5’ observada para TEs não-LTRs.[2]

Já os elementos transponíveis da classe II ou DNA transposons, costumam ter em suas extremidades repetições terminais invertidas, ou ITRs (*inverted terminal repeats*), que são sequências seguidas, *downstream*, de seu complemento reverso. O mecanismo de transposição mais compreendido desta classe remete aos elementos que possuem sequência codificante para a enzima transposase. A transposase liga-se ao DNA na região dos ITRs, ocasionando a quebra das ligações nucleotídicas com o restante do genoma, e essa sequência é integrada em outra região do DNA pela transposase ligada (Figura 1).

Também há a possibilidade de classificar os elementos de transposição entre autônomos e não autônomos. Os elementos autônomos são os que são capazes de codificar a maquinaria enzimática necessária para sua própria transposição. Os elementos não autônomos carecem dessa capacidade e dependem da maquinaria codificada pelos elementos autônomos para se multiplicarem.[2]

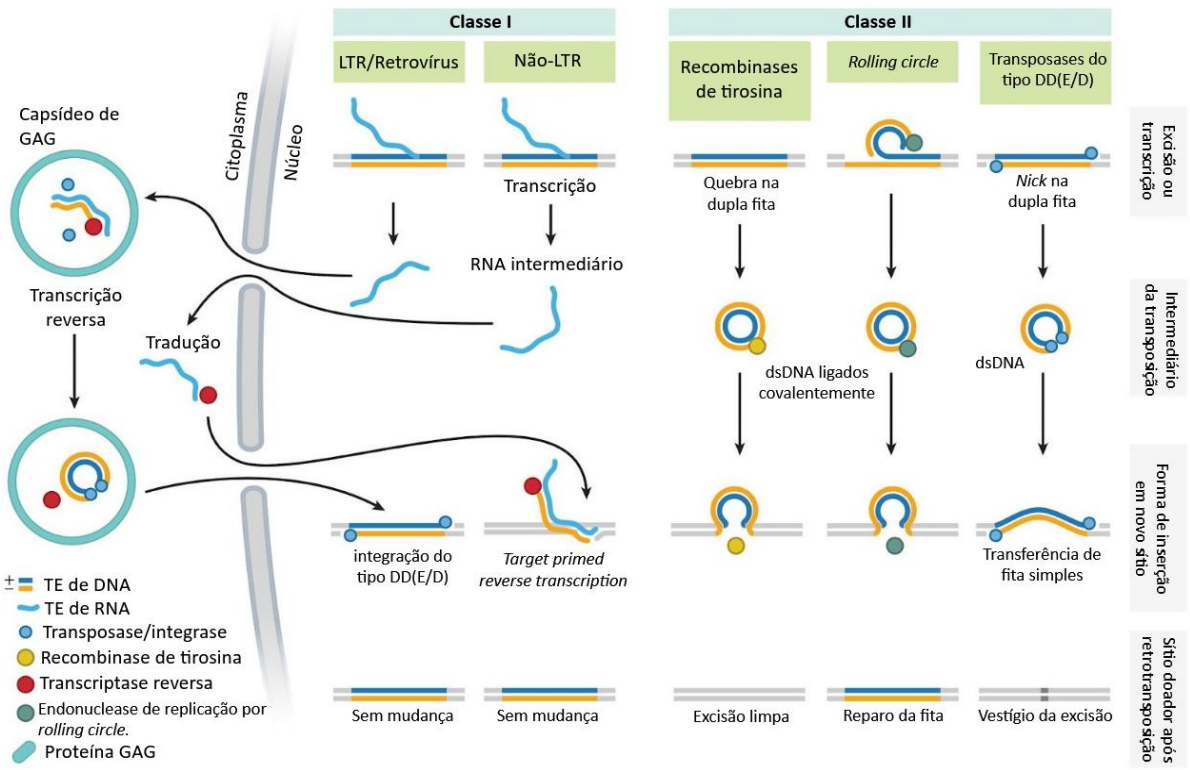

Figura 1 - Principais mecanismos de transposição.
Fonte: Adaptada de WELLS[2]

1.1.2 Impactos da transposição no genoma e na expressão gênica

Ao descobrir os elementos de transposição, Bárbara McClintock propôs que eles atuariam como elementos de controle.[10] Nas décadas posteriores, os TEs foram descritos como elementos “egoístas”, ou seja, que se replicam pela simples capacidade de replicar buscando a própria sobrevivência, sem possuir alguma cooperação com o restante das atividades genômicas. O avanço nas pesquisas conduziu ao entendimento de atividades regulatórias e codificantes nos genomas hospedeiros, derivadas de elementos de transposição, corroborando as ideias iniciais de Barbara McClintock.

Transposições podem ter efeitos neutros, negativos ou positivos no genoma hospedeiro,[11] sendo geralmente disruptivas quando ocorrem em regiões gênicas e, portanto, geralmente deletérias sob o ponto de vista do indivíduo. No entanto, modificações com efeitos positivos tendem a ser retidas, o que constitui uma vantagem evolutiva sob a perspectiva de populações.

Como exemplos de modificações possíveis, cita-se o fato de que inserções próximas do 5’ do gene podem introduzir um novo promotor e alterar o sítio de início de transcrição (Figura 2a), interferir em sequências regulatórias previamente existentes (Figura 2b) ou atuar como um novo elemento regulatório (Figura 2c)[4]. Quando a inserção ocorre em um éxon, pode ocasionar a disrupção da função gênica, podendo ocasionar o fim daquela linhagem. Inserções em íntrons podem provocar a transcrição antissenso (Figura 2d), introduzir um sítio de nucleação de heterocromatina possivelmente silenciando o gene (Figura 2e), alterar o padrão de *splicing* (Figura 2h) ou ainda instituir um novo éxon (Figura 2i).[4]

Nos casos abordados, a alteração do equilíbrio na expressão pode levar a certas doenças, como câncer, mas também pode desempenhar um efeito regulatório benéfico e possivelmente ser “recrutada” pelo organismo. Estudos indicam que aproximadamente 4% dos genes no genoma humano exibem sequências derivadas de TEs em regiões codificantes, enquanto o mesmo é observado em 24% das sequências promotoras.[12-13] Ainda no genoma humano, os retroelementos Alu e L1 estão presentes na região 5’ UTR do gene ZNF177 e aumentam os níveis de transcrição do gene. O estudo de Landry, Medstrand e Mager (2001) também sugere que 4% das regiões 5’ UTR de genes humanos aportam tais retroelementos, indicando que os TEs podem ter efeitos regulatórios complexos e possivelmente influenciarem

a transcrição/tradução de vários genes, tendo impacto relevante no organismo.[14] Algumas sequências originadas de transposons estão entre as mais conservadas do genoma, sugerindo que esses elementos são funcionais.[15]

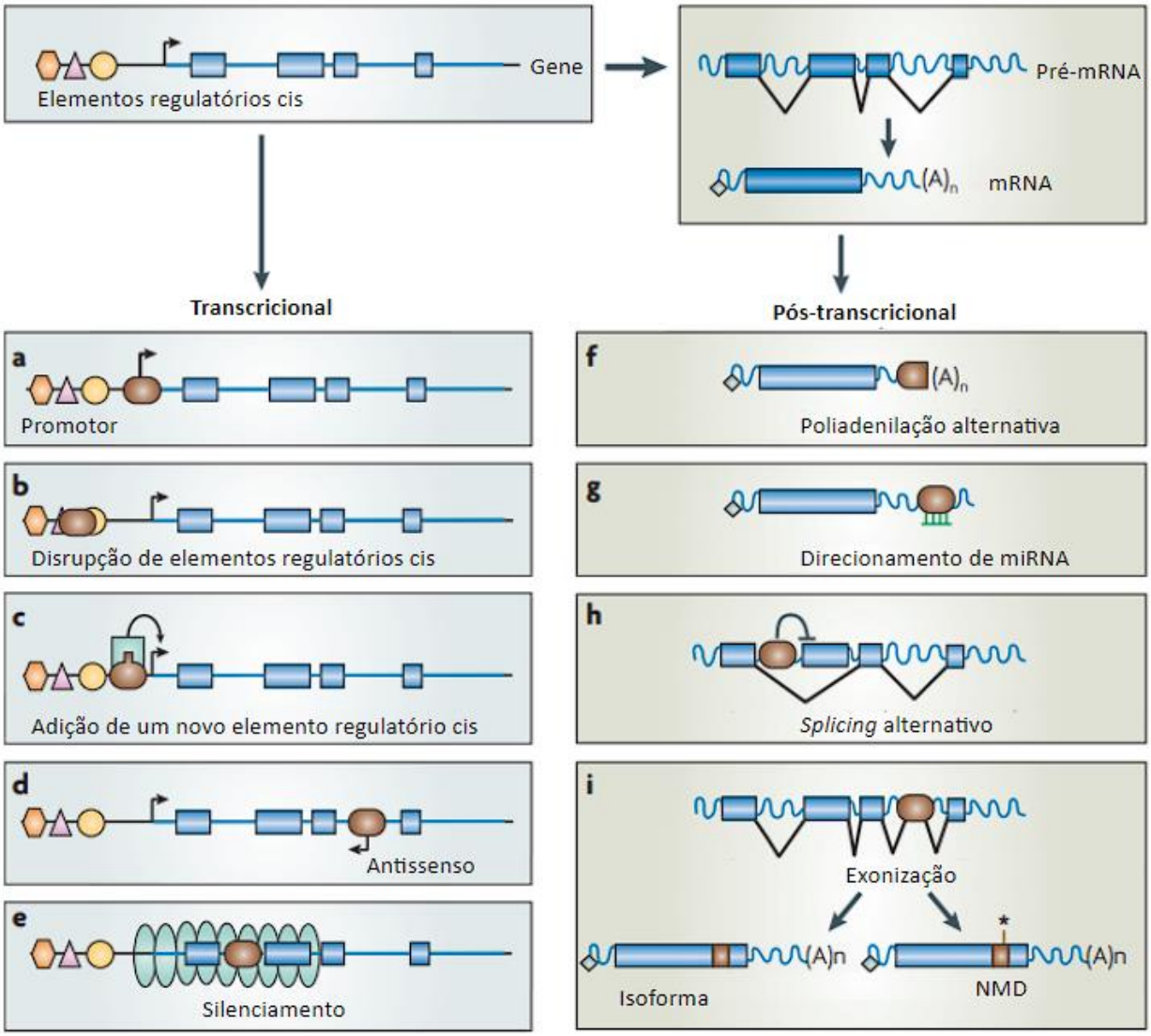

Figura 2 - Influência da transposição na expressão gênica. Possibilidades nas quais a transposição pode alterar a expressão em nível transcricional (a-e) e pós-transcricional (f- i).
Fonte: Adaptada de FESCHOTTE[4]

1.1.3 Direcionamentos de inserção

A distribuição genômica das inserções de TEs resulta da conjunção entre a integração dos elementos no genoma, que pode ser direcionada ou não, e a seleção no hospedeiro, que tende a perpetuar inserções que o beneficiem e eliminar inserções

que tragam prejuízos. Eventos de seleção incluem morte da célula ou dificuldades adaptativas do organismo, quando a inserção ocorre nas células germinativas. À parte a eliminação direta dos indivíduos, as cópias também são perdidas por decaimento ou recombinação. Para evitarem a extinção, os TEs precisam estar em constante multiplicação.[9]

A partir dos mecanismos de transposição (Figura 1) espera-se, a priori, que não exista direcionamento para tal. Porém, com o sequenciamento de nova geração e aumento do volume de dados, tem sido possível identificar padrões de inserção, desde sítios específicos até regiões amplas. Dentre os TEs não-LTR, há o elemento R2 da *Drosophila melanogaster* que se insere preferencialmente em genes de DNA ribossomal 28S.[16] Por outro lado, cópias localizadas em regiões específicas do genoma não necessariamente representam alguma preferência de inserção, como é o caso dos elementos presentes em *Arabidopsis thaliana*, que estão acumulados em regiões pericentroméricas. Inserções *de novo* desses elementos foram encontradas distribuídas uniformemente no genoma, o que leva à conclusão de que, ao invés de uma preferência de inserção para tais regiões, eventos de seleção em que restaram apenas as inserções próximas aos centrômeros devem ter ocorrido.[17]

A cromatina é o substrato natural de inserção de TEs e sua estrutura pode afetar a eficiência ou seletividade da integração em alguma região do genoma. Pode-se pensar que regiões de eucromatina seriam necessariamente preferidas por conta da disponibilidade espacial e maior atividade transcricional, mas cada TE tem preferências por propriedades distintas da cromatina.[9] O retrotransposon LTR *skipper-1* possui preferência de integração em regiões centroméricas em *D. discoideum,* que por sua vez são regiões de heterocromatina com a presença de histonas metiladas na lisina 9.[18] O mesmo TE possui região codificante para um domínio CHD (*chromodomain*) que tipicamente interage com histonas com metilação na lisina 9 (H3K9me), sugerindo que é possível que esteja envolvido em algum mecanismo de direcionamento de inserção.[19] O TE não-LTR TRE5, por outro lado, está estritamente associado com regiões de aproximadamente 50 pares de bases *upstream* de genes de RNA transportador.[18] Inserções intragênicas têm, em geral, menor probabilidade de passar para a geração seguinte. Isto porque inserções que prejudiquem a transcrição de certos genes podem inviabilizar o desenvolvimento e, consequentemente, a reprodução do organismo portador daquele genoma. Neste estudo foram abordadas principalmente as inserções intergênicas.

1.2 Modificações de histona e suas regulações na expressão gênica

Nas células eucarióticas, o DNA é compactado em cromatina no núcleo. O núcleo básico da cromatina chama-se nucleossomo (Figura 3) e contém cerca de 146 pares de bases de DNA em torno de um octâmero de histonas, formado por um tetrâmero H2 (H2A e H2B) e dois dímeros H3 e H4 (Figura 3).[20] Histonas são proteínas que interagem com o DNA e possibilitam a sua compactação, servindo como elementos de controle de sua disponibilidade.

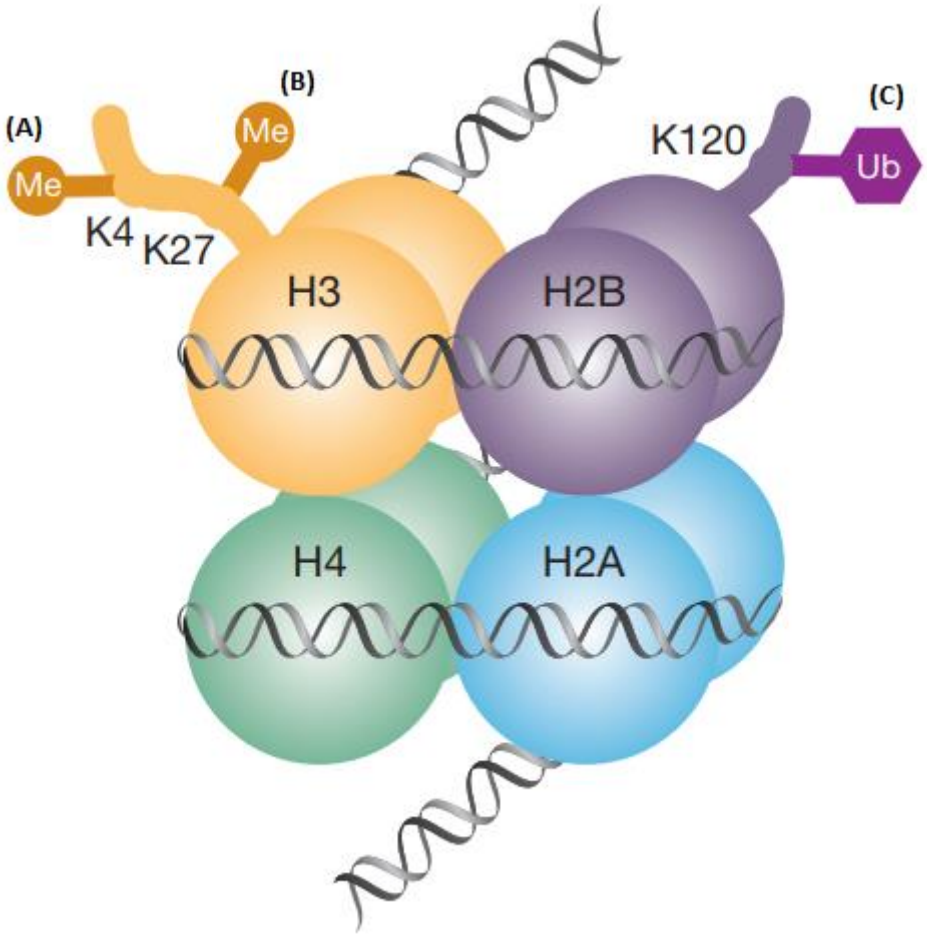

Figura 3 - Nucleossomo com as modificações H3K4me (A), H3K27me e H2BK120ub.
Fonte: Adaptada de ZAIDI *et al*. [21]

As caudas N-terminais e C-terminais das histonas podem sofrer modificações pós transcricionais, como acetilação, fosforilação e metilação.[22] Essas modificações podem ocasionar mudanças conformacionais na cromatina e, consequentemente,

alterar a expressão gênica,[23] podendo possibilitar ou não a transcrição dos genes envolvidos.

As metilações, que geralmente ocorrem na lisina 4 das histonas H3 e H4, estão entre as mais importantes modificações pós-traducionais.[24] Os resíduos de lisina podem ser mono, di e tri metilados. Metilações em H3K4, H3K36 e H3K79 são consideradas marcações de ativação da transcrição e cromatina aberta; enquanto em H3K9, H3K27 e H4K20 remetem à repressão da transcrição e cromatina condensada.[25] Pode-se consultar algumas modificações e seus efeitos transcricionais na tabela 1. As acetilações podem reduzir a carga positiva dos resíduos de lisina, enfraquecendo a interação destes com o DNA, deixando o DNA mais exposto. Por isso, geralmente são consideradas marcações de ativação de transcrição.[26]

Tabela 1 - Modificações de histona e seus efeitos transcricionais.

Modificação de histona	Efeito transcricional	Referência
H3K4me3	Elongação da transcrição, ativação de eucromatina.	DI CERBO *et al.*[27]
H3K9me3	Repressão da transcrição	
H3K27me3	Silenciamento transcricional	
H3K9Ac	Ativação da transcrição	SUKA *et al.*[28]

Fonte: Adaptada de LAWRENCE *et al*.[29]

1.3 O domínio PHD

Elementos de transposição do tipo não-LTR como já comentado, possuem, em sua maioria, dois quadros de leitura aberta (ORFs) que são comumente chamados de ORF1 e ORF2 (Figura 4). As duas enzimas envolvidas no processo de replicação do elemento de transposição situam-se na ORF 2 enquanto a função da ORF1 é ainda pouco compreendida.

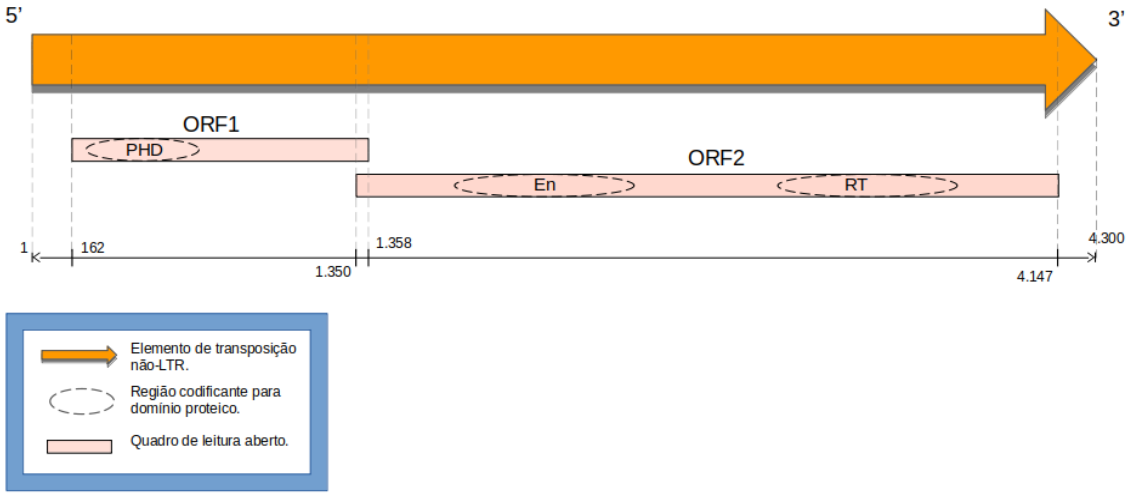

Figura 4 - Quadros de leitura aberta de TEs não LTR e regiões codificantes para domínios proteicos. Distâncias em pares de bases. EN = Endonuclease, RT = Transcriptase reversa e PHD = Plant homeodomain. As distâncias, são um exemplo e correspondem às do TE Perere-6.
Fonte: Adaptada de DEMARCO[30]

Dentre as sequências codificantes encontradas na ORF1 de elementos não-LTR, cita-se o PHD (*plant homeodomain*), presente em parte dos TEs analisados neste trabalho.[30] O PHD é um domínio proteico que possui a propriedade de reconhecer estados de metilação e acetilação de histonas, principalmente na lisina 4 da histona 3,[31] e também outras modificações como os estados de acetilação e metilação da lisina 14 e da arginina 2.[31-32]

Estruturalmente, o domínio PHD consiste em duas folhas beta antiparalelas seguidas de uma alfa hélice carboxi-terminal. O enovelamento é estabilizado por dois átomos de zinco (Figura 5).[31]

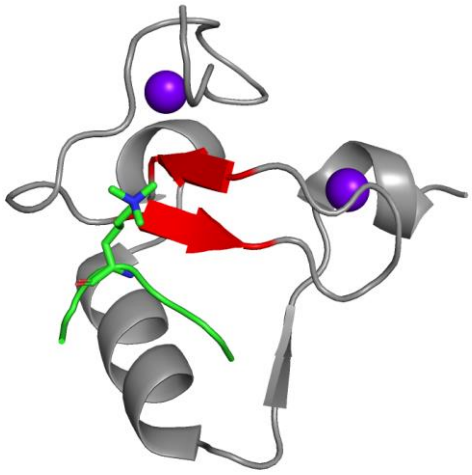

Figura 5 - Exemplo de domínio PHD interagindo com a lisina 4 trimetilada pertencente à histona H3 (H3K4) (PDBid: 2YYR). Em vermelho as folhas beta, em roxo os átomos de zinco e em verde a lisina trimetilada proveniente da histona H3.
Fonte: Molécula depositada no repositório PDB (Protein Data Bank).[33]

Os domínios PHD possuem alta variabilidade estrutural e apresentam afinidade por modificações diversas de histonas, inclusive com mais de uma modificação simultaneamente.[34] Em humanos, os domínios PHD podem estar associados a modificações de histonas,[32] atuando como reguladores epigenéticos. Nesses casos, mutações no domínio PHD estão associadas a doenças como leucemia mielogênica aguda,[35] câncer de mama e síndrome de Omenn.[36-37]

1.4 Schistosoma mansoni

Schistosoma mansoni é um trematódeo parasita e está entre os principais causadores da esquistosomose, juntamente com as espécies *S. japonicum* e *S. hematobium*. As espécies são consideradas endêmicas em 78 países localizados na América, Ásia e África e a doença é um grave problema de saúde pública.[38] Por estar relacionada com contato com água contaminada, é prevalente em populações em condições de pobreza e que não têm acesso ao saneamento básico.[39] A doença é inicialmente assintomática, mas evolui para quadros graves podendo levar o paciente a óbito. Para o tratamento é utilizado o praziquantel, que é eficaz para todas as espécies de *Schistosoma* e pode ser administrado em humanos adultos e crianças.[40]

O parasita possui um ciclo de vida envolvendo dois hospedeiros (Figura 6): o homem, como hospedeiro definitivo, e o caramujo do gênero *Biomphalaria,* como hospedeiro intermediário.

O ciclo se inicia com ovos de *S. mansoni* sendo liberados nas fezes humanas. Na presença de água, os ovos liberam o miracídio, uma larva ciliada capaz de nadar e penetrar em um caramujo do gênero *Biomphalaria*, dentro do qual desenvolve-se em esporocisto e passa a liberar formas conhecidas como cercárias, via reprodução assexuada. A cercária é um outro tipo de larva capaz de nadar que, ao contato com a pele humana, é capaz de penetrá-la e de se transformar em esquistosômulo. O esquistosômulo trafega até o fígado, no qual amadurece em adulto.[39] Os indivíduos adultos do *Schistosoma spp*, ao contrário da maioria dos trematódeos, são gonocóricos, ou seja, possuem sexos separados.[41] Durante a cópula, o macho e a fêmea prendem-se e migram para o intestino, onde a fêmea põe os ovos que são liberados com as fezes e o ciclo se reinicia. Como a reprodução é sexuada, o conteúdo genético de cada prole é diferente, o que aumenta a adaptabilidade da espécie mediante pressões evolutivas.

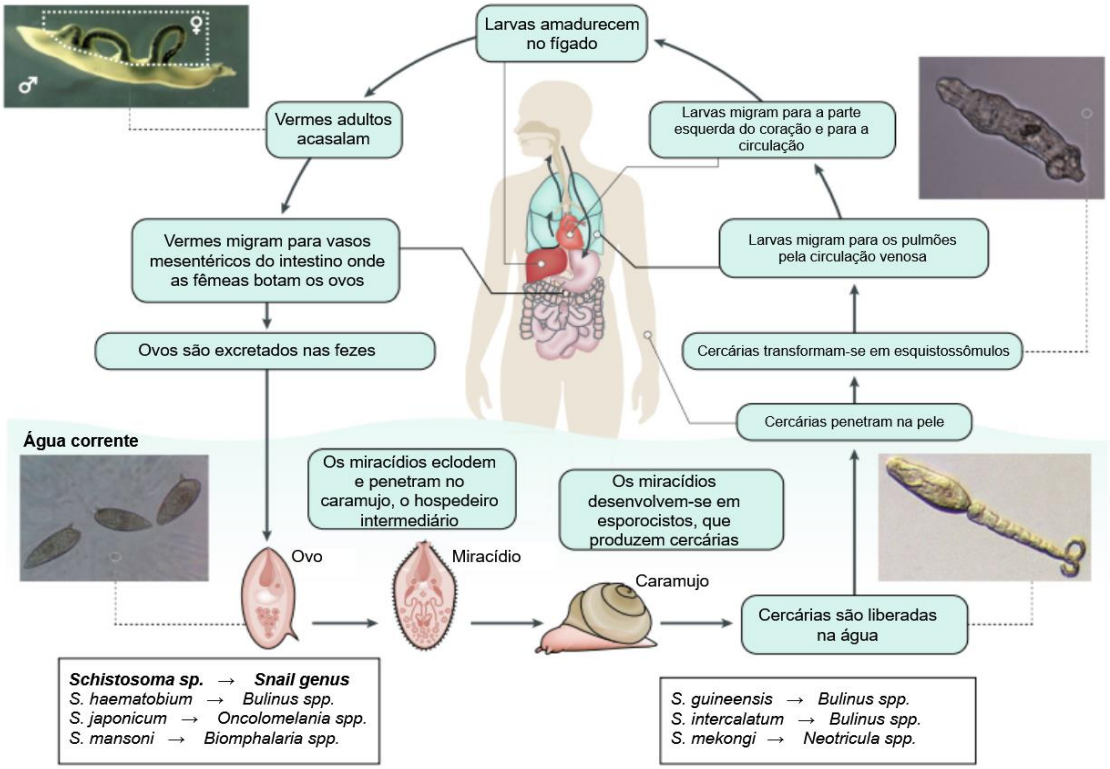

Figura 6 - Ilustração do ciclo de vida de Schistosoma sp.
Fonte: Adaptada de MCMANUS *et al*.[39]

As espécies *S. mansoni* e *S. japonicum* são prevalentes na África e presume-se que o *S. mansoni* tenha sido trazido para o Brasil e América do Sul, possivelmente, pelo tráfico de povos escravizados.[42] A espécie *S. japonicum*, por outro lado, é prevalente na Ásia. Acredita-se que houve a dispersão de uma espécie ancestral ao *S. mansoni* através da migração de mamíferos da Ásia para a África há aproximadamente 12-19 milhões de anos.[43]

O genoma de *S. mansoni* contém em torno de 270 milhões de pares de bases,[44] divididas entre 7 cromossomos autossômicos e um sexual. O organismo é diplóide (2n = 16) e possui sistema de diferenciação de sexos ZW, ou seja, o macho possui dois cromossomos Z e as fêmeas um cromossomo Z e outro W. Possui em torno de 10.000 genes codificantes e 14.500 transcritos.[45] Foram identificadas 72 famílias para elementos de transposição não-LTR e 72 famílias para elementos LTR, compreendendo 15% e 5% do genoma, respectivamente.[46] Os TEs do tipo LTR são dos clados Ty3/Gypsy e BEL, enquanto os TEs do tipo não-LTR estão restritos aos clados RTE, CR1 e R2.

Dois TEs do tipo não-LTR da classe RTE (SR2 e Perere-3) parecem ter tido um aumento significativo na transposição após a divergência de *S. mansoni* e *S. japonicum*, ocasionando uma representatividade muito maior desses elementos em *S. mansoni* (15% contra 8% em *S. japonicum*).[42] É possível que esses elementos estejam relacionados com a adaptabilidade do organismo e especiação em situações de estresse evolutivo e possam também desempenhar um papel na resistência ao tratamento tradicional com praziquantel.

2 MOTIVAÇÃO

Observações prévias de colinearidade entre elementos transponíveis (TEs) e genes *downstream* próximos levantaram o questionamento sobre um possível padrão de inserção envolvendo TEs não-LTR da família CR1, presentes em *S.mansoni*.

3 OBJETIVOS

O objetivo geral foi centrado em análises genômicas de *S. mansoni* visando esclarecer padrões de inserção de elementos transponíveis da família CR1. Para isso, os seguintes objetivos específicos foram traçados:

- Comparar a proporção de inserções colineares ao gene *downstream* mais próximo com a de inserções esperadas ao acaso, para avaliar se existe alguma tendência.
- Avaliar a distribuição das distâncias das inserções, com o gene *downstream* mais próximo, a fim de verificar se existe alguma distância preferencial para as inserções.
- Avaliar a distância entre as inserções e marcadores epigenéticos, bem como suas frequências, para verificar se as inserções estão localizadas em regiões ricas com determinados marcadores.
- Avaliar níveis de transcrição dos genes próximos às inserções para verificar se eles possuem transcrição significativamente abaixo/acima de uma amostragem aleatória, bem como os níveis de transcrição entre as fases de vida do parasito, o que poderia indicar se há alguma tendência de inserção.
- Realizar uma análise filogenética dos domínios PHD dos elementos não-LTR do clado CR1, frente a outros domínios com interação conhecida com histonas modificadas, a fim de caracterizar tais domínios.

4 METODOLOGIA E RESULTADOS

4.1 Mapeamento das inserções no genoma

Para as análises, foram utilizados todos os elementos da classe não-LTR presentes no genoma de *S. mansoni* (Tabela 2), com exceção do elemento Perere-9, do clado R2. Os elementos pertencentes aos clados RTE e CR1 foram todos incluídos. As sequências foram obtidas a partir do banco de dados de nucleotídeos disponível no repositório do NCBI e também de Laha *et al*. (2005).

Retrotransposons do clado CR1 (*chicken repeat 1*) têm como característica, além dos atributos comuns em TEs não-LTR, a presença de terminações 3' com uma sequência repetitiva (GATTCTRT) ao invés do trecho de poliadenina, e uma tendência, acima da usual, de truncamento na extremidade 5'.[47] Os elementos CR1 estudados foram classificados por,[30] por uma análise filogenética utilizando a transcriptase reversa como parâmetro de proximidade.

Membros do clado RTE são caracterizados por possuírem um 3' UTR incomumente curto, predominantemente composto de repetições AT em trímero, tetrâmero e pentâmero.[48] Os elementos RTE SR3/Perere-3 e SR2 foram caracterizados por uma análise filogenética do domínio da transcriptase reversa, presente em todos os elementos não-LTR.[45-49]

Com o intuito de mapear as cópias de cada retrotransposon no genoma de *S. mansoni*, foram definidos trechos de 100 pares de base para representar cada elemento de transposição. É esperado que sequências próximas da extremidade 3' de elementos não-LTR sejam mais frequentes no genoma, quando comparado à sequências da extremidade 5', devido à baixa processividade da enzima transcriptase reversa no mecanismo de transposição destes elementos.[50]

Arquivos no formato fasta relativos ao genoma de *S. mansoni* (versão 7) – primariamente depositados por Berriman *et al*. (2009) e Protasio *et al*. (2012) – foram obtidos no repositório WormBase.*[1] Neste mesmo repositório, foi obtido o arquivo em formato GFF com as coordenadas gênicas, na versão correspondente. No arquivo GFF foram selecionados apenas os genes que possuíam região codificante (*protein coding*) e regiões UTRs descritas. A seleção foi realizada através de um código escrito

* Disponível em: https://wormbase.org//#012-34-5.

(*script*) em *Python*. Os dados da posição de cada gene, bem como sua orientação no genoma, foram armazenados.

Tabela 2 - Retrotransposons utilizados neste trabalho. Foram escolhidos os clados CR1 e RTE.

TE	Identificação	Fonte	Clado	Trecho Utilizado(pb)	Tamanho(pb)
SR1	U66331.1	Genbank	CR1	2237 - 2337	2337
SR2	AF025681.2	Genbank	RTE	3480 -3580	3652
SR3		LAHA[45]	RTE	3108 - 3208	3208
Perere	BK004067.1	Genbank	CR1	4775 - 4875	4875
Perere-2	BN000793.1	Genbank	CR1	4240 - 4340	4544
Perere-3	BN000794.1	Genbank	RTE	3096 - 3196	3196
Perere-4	BN000795.1	Genbank	CR1	5011 - 5111	5111
Perere-5	BN000796.1	Genbank	CR1	4330 - 4430	5057
Perere-6	BN000797.1	Genbank	CR1	4090 - 4190	4300
Perere-7	BN000798.1	Genbank	CR1	4914 - 5014	5014

Fonte: Elaborada pelo autor.

A fim de mapear as inserções dos TEs no genoma, foi realizado um *blastn* dos trechos de 100 pb indicados na tabela 2, representando cada retrotransposon, contra o genoma do *S. mansoni*. Foram aceitos apenas resultados com *e-value* abaixo de 10^{-10} e com mais de 90% de identidade, o que indica que as cópias selecionadas são conservadas. Esses foram considerados como as inserções de cada TE.

Devido ao fato que os retrotransposons Perere-3 e SR3 apresentam um percentual de identidade proteica superior a 80% na região relativa à transcriptase reversa, optou-se pela consideração conjunta destes dois elementos (SR3/Perere-3), estratégia também utilizada por Venancio et al. (2010). As populações das inserções estão indicadas na Tabela 3.

Tabela 3 - Populações das inserções de cada TE. Os retrotransposons Perere-3 e SR3 foram considerados em conjunto por apresentarem superposição nos alinhamentos.

TE	População das cópias
Perere	646
Perere-2	839
Perere-4	172
Perere-5	520
Perere-6	348
Perere-7	309
SR1	1043
SR2	1500
SR3/Perere-3	16375

Fonte: Elaborada pelo autor.

4.2 Análise de colinearidade de inserções intergênicas

Através de um *script* em *Python,* foram comparadas as coordenadas das inserções, encontradas como resultado do *blastn,* com as coordenadas dos genes selecionados no arquivo GFF. Nesta comparação, buscou-se especificamente pelo gene *downstream* mais próximo a cada inserção, como exemplificado na Figura 7. O gene *downstream* em relação à inserção foi considerado como sendo a próxima sequência codificante à 3', enquanto o gene *upstream* foi o mais próximo codificado à 5' com relação à inserção. Foram contempladas as quatro possibilidades de orientação das inserções em relação ao gene.

Considerou-se como região intergênica aquela entre o término da UTR 3' do gene e o início da UTR 5' do próximo gene (Figura 7). Inserções em que pelo menos uma extremidade do TE estava localizada nesta região foram denominadas inserções intragênicas. Por sua vez, foram consideradas inserções intergênicas aquelas em que nenhuma extremidade do TE estava em região intragênica, sendo anotados o gene *downstream* e o gene *upstream* dessa inserção. As inserções que não atendiam a nenhum dos casos foram desconsideradas das análises e a população de cópias decorrente desta anotação foi registrada na Tabela 4. A fração de TEs intergênicos foi calculada como a razão entre as inserções intergênicas e a população de cópias (Tabela 4).

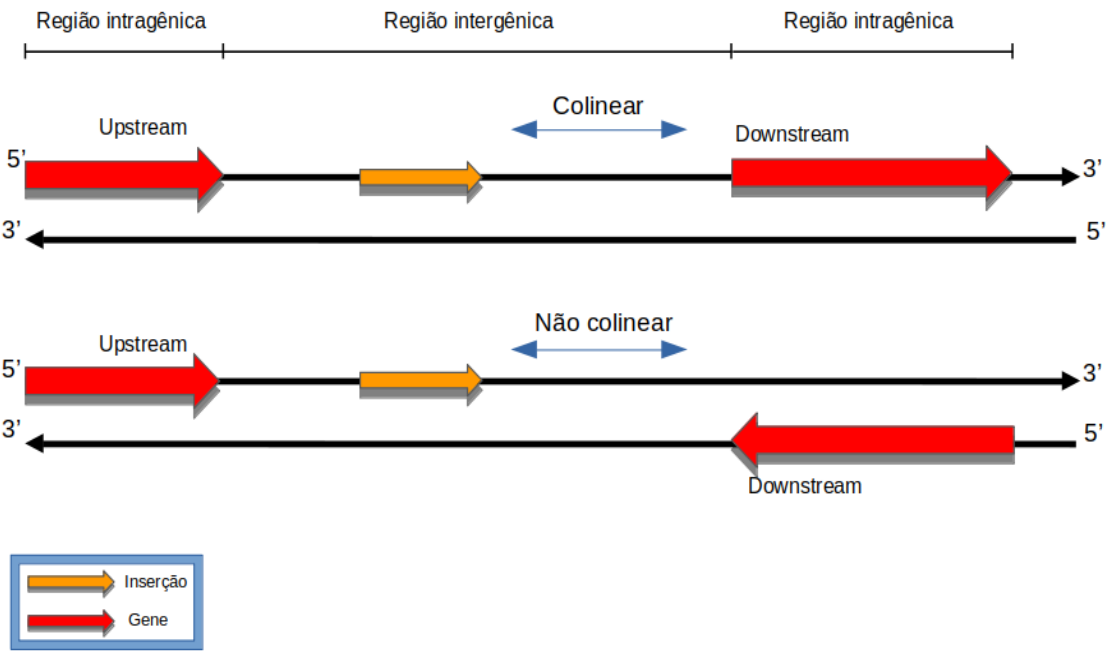

Figura 7 - Representação dos casos avaliados no estudo da fração de TEs intergênicos e colineares. Dimensões ilustrativas.
Fonte: Elaborada pelo autor.

A colinearidade dos TEs foi analisada para inserções intergênicas e intragênicas: para as intergênicas, foi considerada colinear a inserção cujo gene *downstream* mais próximo possuía a mesma orientação que a inserção (fig. 7). Para as intragênicas, foi considerada colinear a inserção que estava compreendida em um gene de mesma orientação. A fração de TEs colineares consistiu na razão entre a contagem de inserções colineares e a contagem de inserções intergênicas, o análogo para as intragênicas (Tabela 4).

Tabela 4 - Fração das inserções intergênicas e colineares das cópias de cada TE.

TE	Clado	Número de cópias	Fração de intergênicos	Fração de colineares (intergênicos)	Fração de colineares (intragênicos)
Perere	CR1	618	0.87	0.57	0.26
Perere-2	CR1	816	0.71	0.75	0.4
Perere-4	CR1	163	0.93	0.77	0.33
Perere-5	CR1	498	0.67	0.81	0.4
Perere-6	CR1	341	0.64	0.83	0.36
Perere-7	CR1	299	0.94	0.76	0.47
SR1	CR1	1013	0.80	0.85	0.73
SR2	RTE	1471	0.23	0.41	0.46
SR3/Perere-3	RTE	15952	0.53	0.47	0.29

Fonte: Elaborada pelo autor.

Analisando a Tabela 4, foi observada uma tendência das cópias dos elementos do clado CR1 estarem em regiões intergênicas e colineares ao gene *downstream* mais próximo. Considerando-se que a soma das regiões intergênicas corresponde a 46% do total do genoma, seria esperado para uma distribuição aleatória das cópias no genoma que a distribuição das inserções ocorresse nessa proporção. No entanto, inserções em regiões gênicas possuem maior probabilidade de interferirem negativamente no funcionamento do gene, tendendo a ser selecionadas negativamente, o que pode explicar a observação de que as inserções de elementos não-LTR sejam prevalentes em regiões intergênicas.[51]

É esperado que cópias de TEs não-LTR em regiões intrônicas estejam não colineares com o respectivo gene.[51] Isso ocorre porque existe uma forte seleção contra elementos com a mesma orientação do gene devido à maior chance desses elementos prejudicarem a sua transcrição. O retrotransposon do clado RTE SR3/Perere-3 possui a fração de cópias intergênica/intragênica mais próxima da proporção dos tamanhos das regiões (53% contra 46%), ou seja, tem uma quantidade considerável de cópias em regiões intragênicas. Isto pode ser justificado pela baixa colinearidade (29%) nas regiões intragênicas, considerando que se espera encontrar a maior parte das cópias intragênicas em regiões intrônicas. O TE SR2 possui a maior parte das cópias em regiões intragênicas e colinearidade próxima dos 50% nessas regiões (46%), o que poderia indicar que esse TE tenha alguma característica que o impeça de prejudicar a transcrição gênica mesmo quando em senso com o gene.

É possível que ainda aconteça um evento de seleção especificamente no clado CR1 que desprivilegie inserções em íntrons. Considerando que as inserções *de novo* tenham orientação senso com o gene *downstream*, pode-se especular que essa tendência de orientação ocasionou uma baixa população intragênica devido à possibilidade de prejuízo à transcrição. Nesse caso, as duas características estariam relacionadas.

Há ainda a possibilidade das cópias terem sido selecionadas positivamente por exercerem algum efeito regulatório benéfico ao organismo, como a adição de um novo promotor ou elemento regulatório *cis.*[4] No tópico “Análise dos níveis de transcrição de genes *downstream* a inserções de TEs” é realizada uma análise a fim de verificar se haveria diferenças de transcrição entre genes com inserções de TEs CR1 *upstream* e

o restante dos genes, indicando possivelmente que os TEs estudados teriam papel regulatório nos genes *downstream*.

Assim, com o objetivo de avaliar se a fração de cópias colineares observada para cada TE poderia refletir um padrão de inserção preferencial, foi realizada uma simulação (utilizando-se a linguagem *Python*) em que se considerou a distribuição aleatória das inserções nas regiões intergênicas do genoma de *S. mansoni*, bem como o fato de que as duas orientações (fita positiva ou negativa do genoma) são igualmente prováveis.

Para tanto, a cada região intergênica foi definido um parâmetro denominado peso (Figura 8) que corresponde à razão entre a extensão da região em questão e a soma da extensão de todas as regiões intergênicas. Assim, o parâmetro peso representa a probabilidade de certa região ser sorteada para a inserção, sob a consideração de uma distribuição uniforme de probabilidades. Em cada iteração da simulação, reproduziu-se o número observado de inserções de cada elemento de transposição: para cada inserção foi realizado o sorteio da região intergênica, através da função *choices* disponível na biblioteca *random*, levando-se em consideração o peso de cada região; uma vez definido o sítio da inserção, foi realizado o sorteio da fita – positiva ou negativa – sendo que ambas possuem a mesma probabilidade de ocorrência. Em seguida, a inserção simulada foi avaliada quanto à sua possível colinearidade em relação ao gene *downstream*. Este procedimento foi realizado para cada inserção intergênica observada naturalmente, de modo que em cada iteração da simulação o total de cópias, para cada TE, foi mimetizado. No total, foram realizadas dez mil iterações na simulação, sendo então definido o valor médio de inserções colineares para cada TE (Tabela 5).

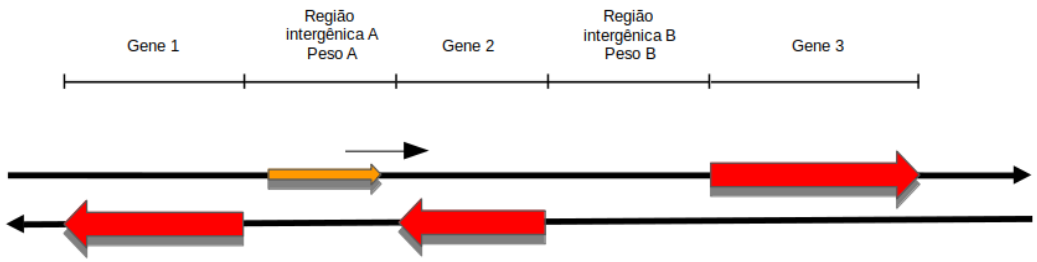

Figura 8 - Esquema da representação das regiões intergênicas para a simulação que objetivou o estudo da colinearidade. Os pesos são proporcionais ao tamanho de tais regiões. Dimensões ilustrativas.
Fonte: Elaborada pelo autor.

Nesta análise foram considerados apenas os oito maiores *contigs*, que representam os oito cromossomos de *S. mansoni,* correspondentes a cerca de 96% da informação genômica disponível.

Tabela 5 - Colinearidade das inserções observadas e simuladas no genoma. Os TEs SR2 e SR3/Perere-3 pertencem ao clado RTE, enquanto os demais ao clado CR1.

TE	col obs.	col. sim.	desvio padrão	intergênicos	p-valor
Perere	0.57	0.523	0.004	536	0.01
Perere-2	0.75	0.524	0.004	576	4.38E-27
Perere-4	0.77	0.523	0.027	151	3.98E-10
Perere-5	0.81	0.523	0.008	333	2.11E-26
Perere-6	0.83	0.523	0.015	218	3.01E-20
Perere-7	0.76	0.523	0.011	280	3.71E-15
SR1	0.85	0.524	0.002	815	2.11E-77
SR2	0.41	0.524	0.008	338	3.77E-05
SR3/Perere-3	0.47	0.524	0.000	8406	2.48E-19

Fonte: Elaborada pelo autor.

Considerando-se uma distribuição uniforme quanto ao aspecto da colinearidade, o esperado seria um percentual de aproximadamente 52% das inserções sendo colineares ao gene *downstream*; no entanto, são observados valores significativamente maiores em todos os elementos do clado CR1 (Tabela 5). Estes resultados sugerem uma tendência quanto à colinearidade das inserções desses

elementos, que pode ser resultado do fato que estas inserções possam exercer algum papel regulatório nos genes *downstream*.

4.3 Análise de distâncias entre inserções de TEs e genes *downstream*

Também foi possível avaliar a tendência de localização das inserções nas regiões intergênicas. Como exemplificado na Figura 9, foi medida a distância em pares de bases (pb) entre a extremidade 3’ das inserções dos TEs e o início da UTR 5’ do gene *downstream* mais próximo. Em seguida, foi definida a razão entre esta distância e a extensão da região intergênica. Por meio desta grandeza, foi possível avaliar o posicionamento da inserção nessa região: quanto menor o valor da razão, mais deslocada a inserção está em direção ao gene *downstream*. Por fim, foi calculada a razão média para cada TE (Tabela 6), sendo as medidas separadas entre aquelas cujo gene *downstream* foi colinear ou não à cópia.

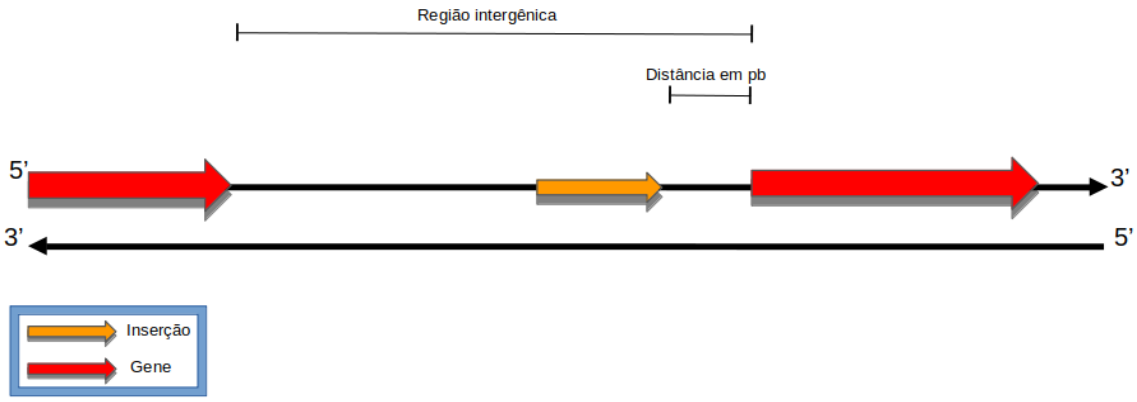

Figura 9 - Representação esquemática da medida da distância entre a inserção e o gene *downstream*.
Fonte: Elaborada pelo autor.

Tabela 6 - Distâncias relativas médias entre inserções e o respectivo gene *downstream*.

TE	Distância média relativa (Colineares)	Distância média relativa (Não colineares)	Número de cópias (Colineares)	Número de cópias (Não colineares)	Codificam PHD	Clado
Perere	0.43	0.48	308	228		CR1
Perere-2	0.26	0.50	432	144	X	CR1
Perere-4	0.43	0.49	107	34		CR1
Perere-5	0.25	0.53	270	63	X	CR1
Perere-6	0.24	0.52	182	36	X	CR1
Perere-7	0.51	0.55	212	68		CR1
SR1	0.18	0.51	691	124	X	CR1
SR2	0.49	0.43	139	199		RTE
SR3/Perere-3	0.47	0.47	3984	4422		RTE

Fonte: Elaborada pelo autor.

Como pode ser observado (Tabela 6), no caso de inserções colineares, os TEs que possuíam uma região codificante para o domínio PHD apresentaram distâncias relativas médias menores do que os que não possuíam essa região. Assim, com o objetivo de avaliar em maiores detalhes as cópias intergênicas que possuíam colinearidade com o gene *downstream*, foi analisada também a distribuição de frequências relativas de TEs em relação à distância ao gene respectivo (figuras 10 e 11). Interessantemente, os TEs que possuem região codificante para o domínio PHD apresentaram uma porção considerável da população a menos de 1000 pares de bases do gene *downstream*, diferentemente dos demais.

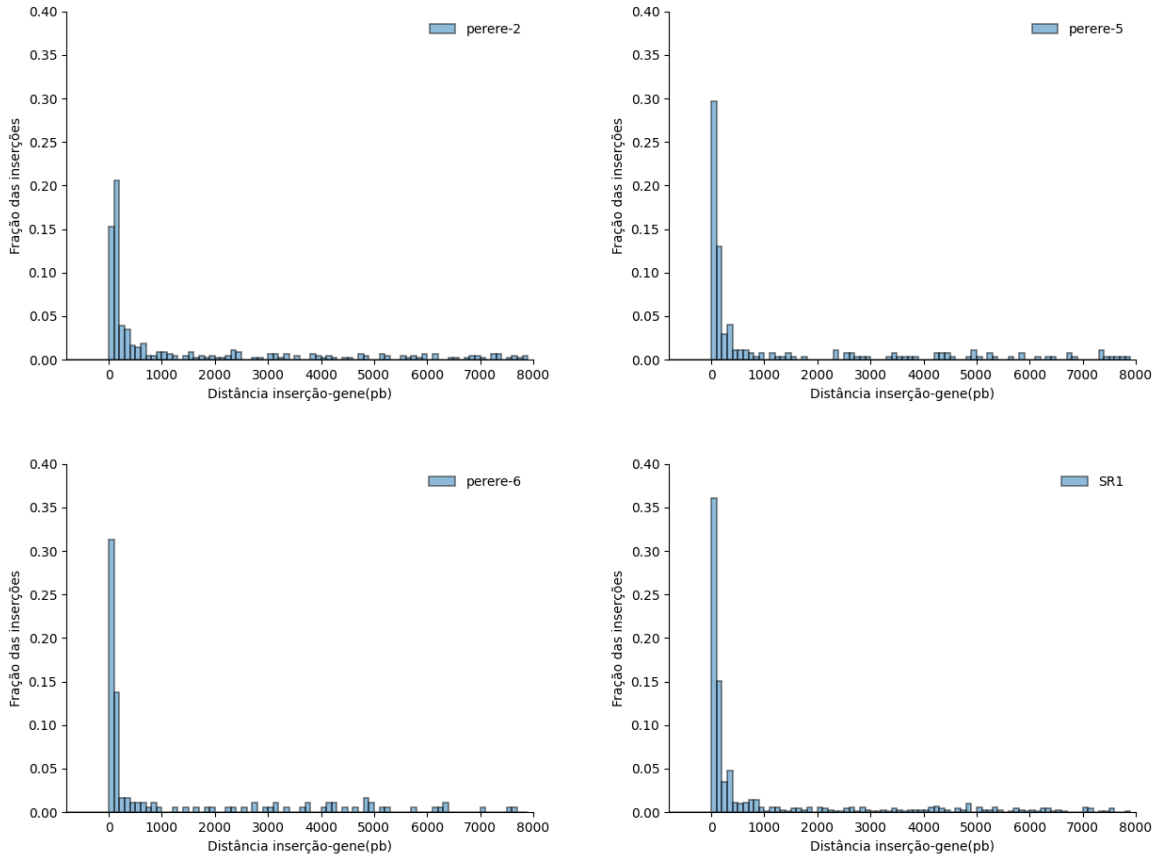

Figura 10 - Histogramas de distribuição das cópias - intergênicas e colineares com o gene *downstream* - de TEs com região codificante para o domínio PHD, em relação à distância do gene *downstream* mais próximo. As populações são Perere-2: 432 cópias; Perere-5: 270 cópias; Perere-6: 182 cópias e SR1: 691 cópias.

Fonte: Elaborada pelo autor.

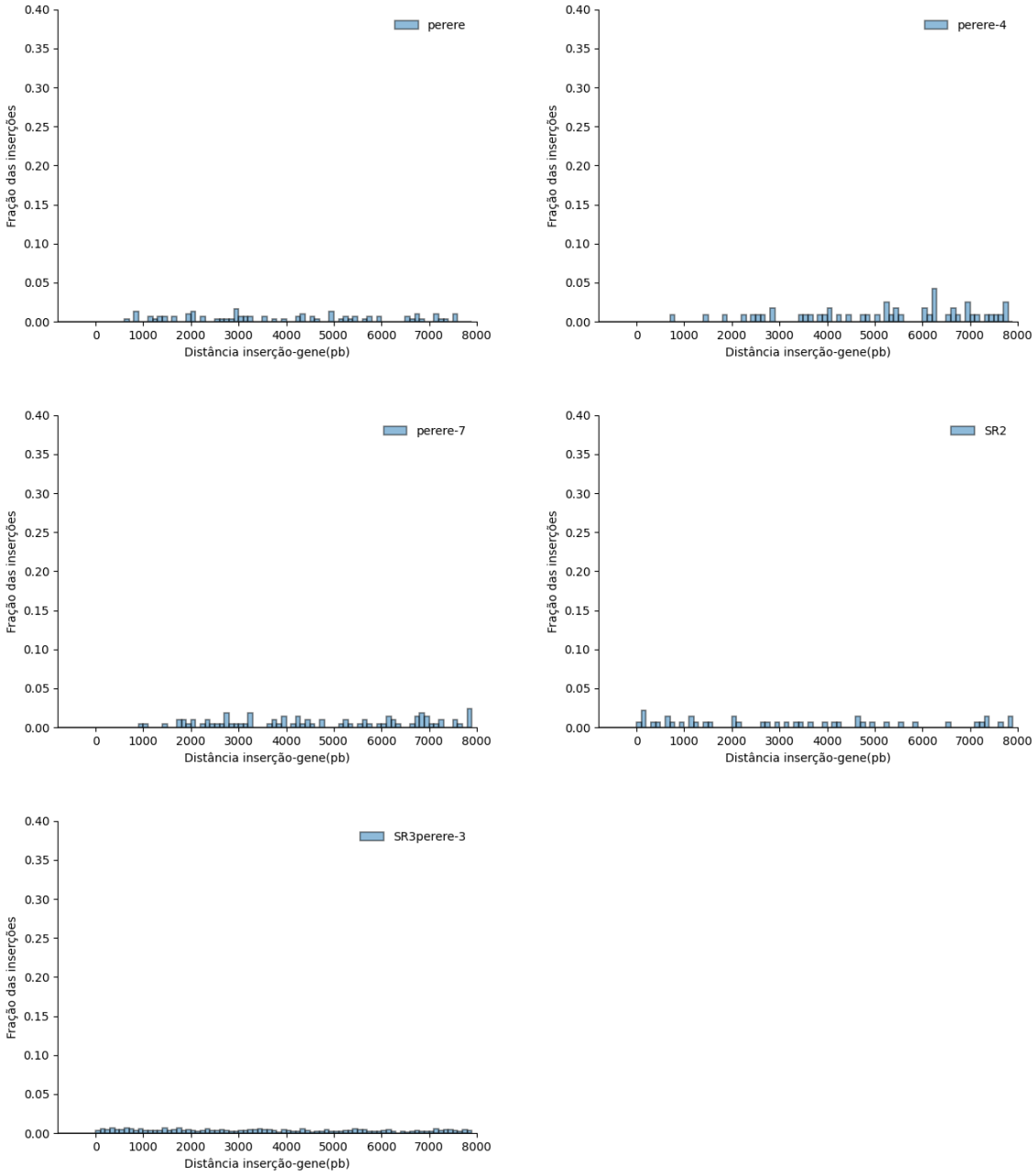

Figura 11 - Histogramas de distribuição das cópias intergênicas e colineares com o gene *downstream* de TEs sem região codificante para o domínio PHD, de acordo com a proximidade em relação ao gene *downstream*. As populações são Perere: 308 cópias; Perere-4: 117 cópias; Perere-7: 212 cópias; SR2: 139 cópias; SR3/Perere-3: 3984 cópias.
Fonte: Elaborada pelo autor.

Como pode ser observado na Figura 10, elementos com o domínio PHD apresentaram significativa parcela das inserções intergênicas colineares a menos de 500pb do gene *downstream*. Demais elementos da família CR1 sem esse domínio não apresentaram esse comportamento. Este resultado sugere um direcionamento das inserções de elementos CR1 com o domínio PHD para regiões *upstream* muito

próximas do local de início de transcrição de certos genes. Possivelmente esse domínio exerça papel nesse direcionamento, o que poderia ser vantajoso para a sobrevivência dos elementos por evitar o comprometimento do gene no caso de uma inserção intragênica. Além disso, regiões próximas a genes possivelmente têm maior probabilidade de serem transcricionalmente ativas, o que facilitaria a replicação do TE. Resultados semelhantes foram encontrados em *D. melanogaster* para TEs de DNA, em que aproximadamente 73% das inserções ocorreram a menos de 500 pares de bases *upstream* de sítios de início de transcrição.[52]

4.4 Análise de inserções de TEs em regiões *downstream* de interação com histonas modificadas.

Considerando que domínios PHD interagem com histonas modificadas,[34] foram realizadas análises para verificar relações entre sítios de interação de histona com a modificação H3K4me3 e as inserções dos TEs estudados que possuíam região codificante para o domínio PHD. Para tanto, foram acessados resultados de CHIP-seq (Tabela 7) no repositório NCBI, referentes à histona com modificação H3K4me3 em *S. manoni* das quatro fases de vida do parasito. Os arquivos foram obtidos no formato fastq e mapeados no genoma de *S. mansoni* utilizando-se o *software* HISAT2.[53] Em seguida, foi empregado o *software* HOMER para identificar regiões de enriquecimento de sequências alinhadas (*reads*);[54] foi definido que cada região de interação com histona H3K4me3 teria um comprimento de mil pares de bases, tal como ilustrado na figura 12.

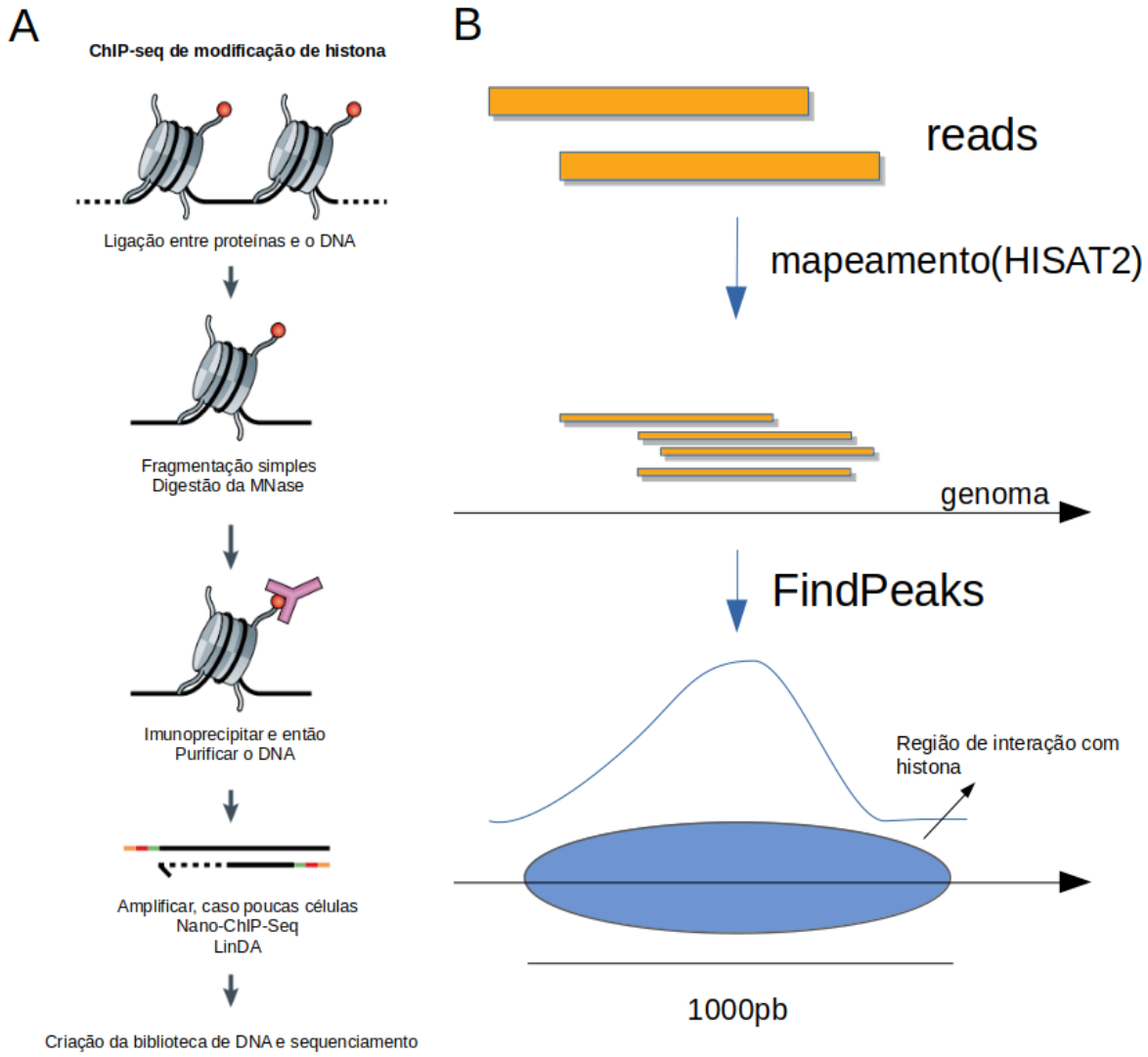

Figura 12 - Esquema ilustrativo do experimento de CHIP-seq (A) e do processamento dos dados de CHIP-seq objetivando a identificação de regiões com histona modificada (B).
Fonte: Adaptada de FUREY.[55]

Tabela 7 - Acesso NCBI dos experimentos de CHIP-seq das fases de vida de S.mansoni.

Fase de vida	Acesso NCBI
Miracídio	SRR6307186
Esporócito	SRR6307188
Adulto	SRR1136063
Cercária	SRR1087939

Fonte: Elaborada pelo autor.

Para cada TE foi selecionado apenas o grupo de inserções intergênicas e colineares ao respectivo gene *downstream* para a análise. Foram medidas as distâncias (Figura 13), em pares de bases, das inserções dos TEs em relação ao pico de histona mais próximo, sendo que valores positivos indicam picos *downstream* e negativos, *upstream*. Casos em que não houve qualquer região de histona no *contig*

da inserção foram desconsiderados. As figuras 14 e 15 ilustram a distribuição de frequências relativas do número de cópias de TEs em relação às regiões de histona.

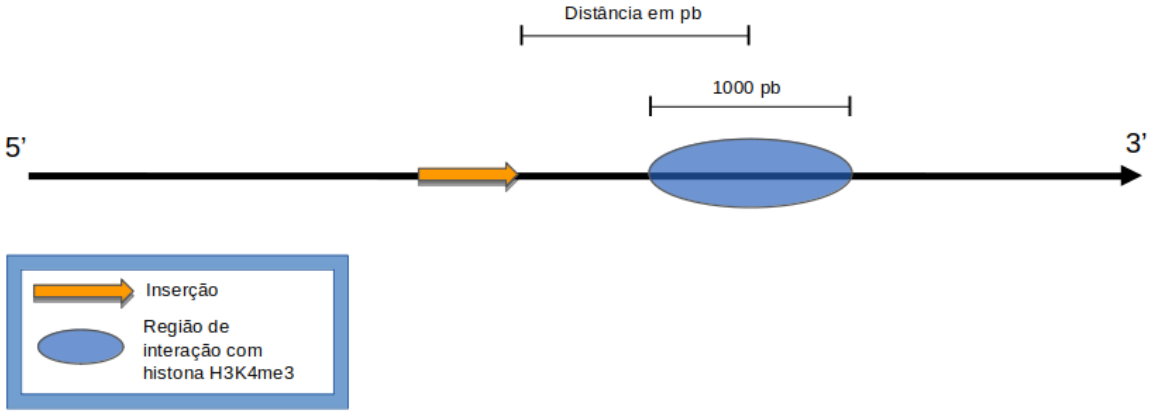

Figura 13 - Esquema ilustrativo da medida de distância entre as inserções de TEs e as regiões de interação com a histona com a modificação H3K4me3.
Fonte: Elaborada pelo autor.

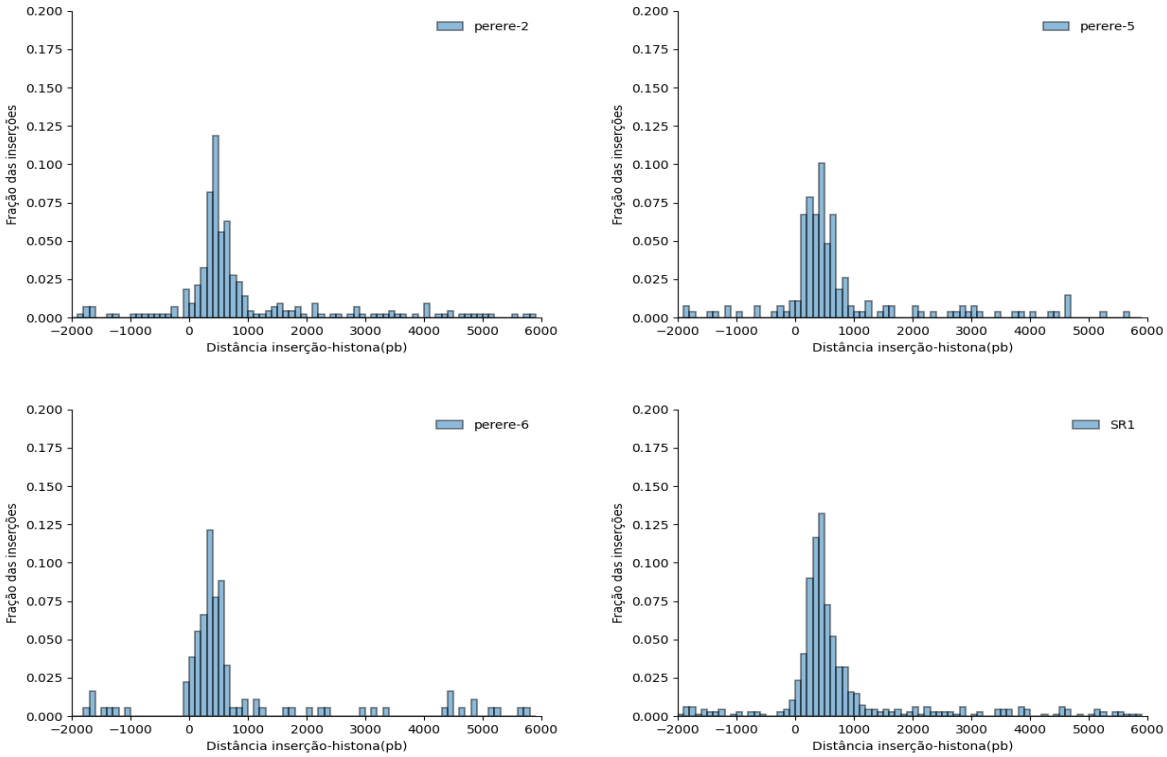

Figura 14 - Histogramas de distribuição das cópias - intergênicas e colineares com o gene downstream - dos TEs com região codificante para o domínio PHD, em relação à distância da histona com modificação H3K4me3 mais próxima. As populações são Perere-2: 429 cópias; Perere-5: 268 cópias; Perere-6: 181 cópias e SR1: 688 cópias.
Fonte: Elaborada pelo autor.

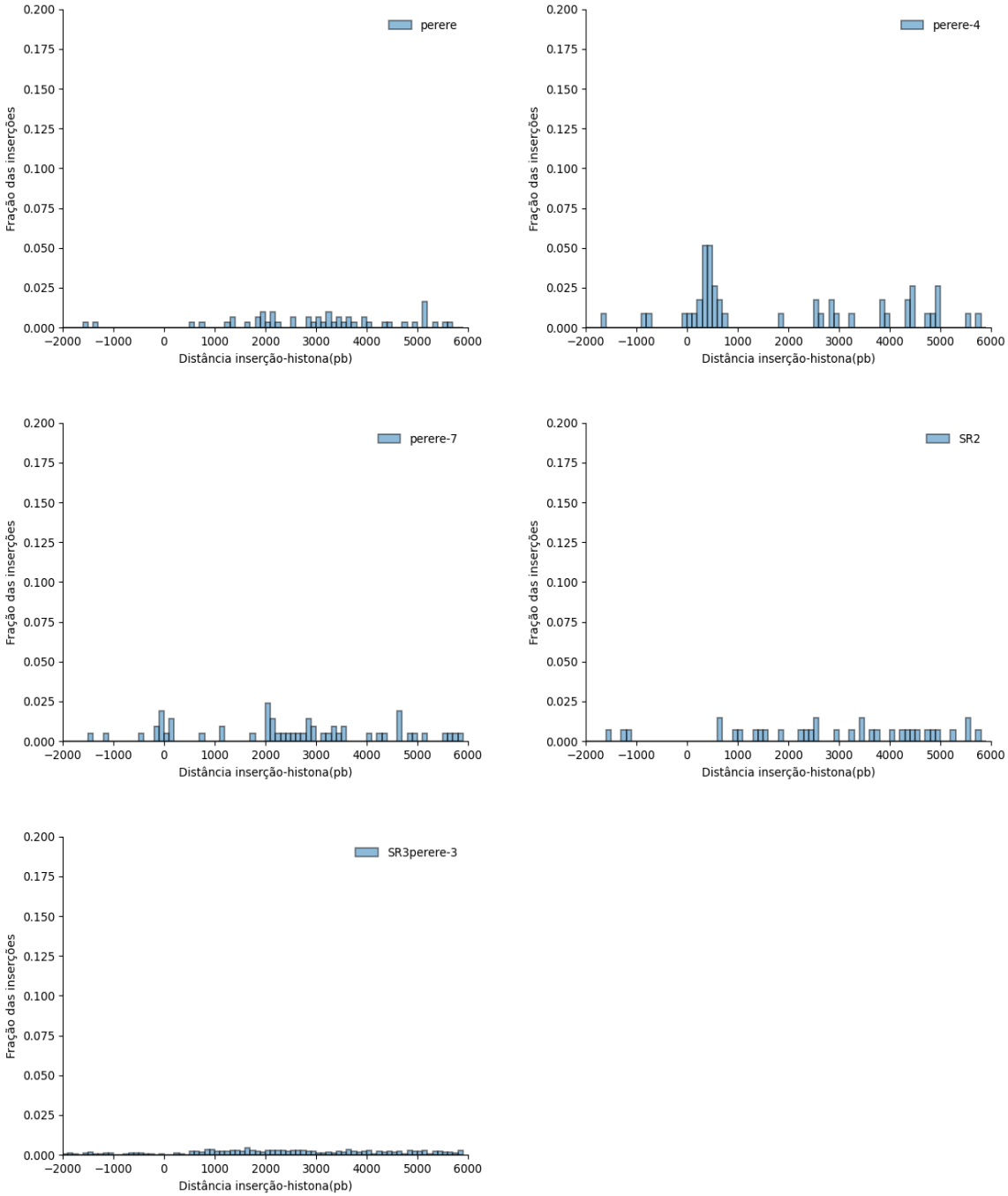

Figura 15 - Histogramas de distribuição das cópias - intergênicas e colineares com o gene downstream - dos TEs sem região codificante para o domínio PHD, em relação à distância da histona com modificação H3K4me3 mais próxima. As populações são: Perere, 304 cópias; Perere-4, 116 cópias; Perere-7, 209 cópias; SR2, 138 cópias; SR3/Perere-3, 3968 cópias.

Fonte: Elaborada pelo autor.

Considerando-se as inserções intergênicas e colineares a genes *downstream*, observa-se que os elementos com região codificante para domínio PHD tiveram parte

significativa das inserções adjacentes a regiões de histona com a modificação H3K4me3, enquanto aqueles, mesmo os do mesmo clado, que não possuem esse domínio, não mostraram essa configuração. Considerando que domínios PHD já foram descritos como reconhecedores de modificações específicas de histona,[34] esse resultado sugere um direcionamento de inserção específico mediado pelo domínio PHD.

Este resultado apresenta semelhança com o retrotransposon LTR Ty3/gypsy, que possui um cromodomínio (CHD) no C-terminal de sua integrase. Cromodomínios são motivos de 40 a 50 aminoácidos que são descritos como interagentes com histonas com modificação H3K9me, marcação epigenética característica de heterocromatina.[19] O elemento Ty3/gypsy, em *A. thaliana*, mostra significante associação com regiões de heterocromatina.[56] Mecanismos epigenéticos têm evoluído, nas células eucarióticas, supostamente a fim de silenciar a expressão e mobilidade dos elementos de transposição, como parte de uma resposta do organismo para frear a multiplicação desses elementos.[57] Entretanto, é possível que os elementos do clado CR1 que possuem o PHD tenham encontrado uma “zona segura” ao inserirem-se em regiões com histonas com modificações associadas à ativação de eucromatina (H3K4me3).

Também existem registros de elementos de transposição que atuam na regulação do genoma em nível epigenético.[57] Nesse caso, existiria a possibilidade de que os TEs do clado CR1 com PHD estejam atuando de forma regulatória nas histonas modificadas, consequentemente também atuando na regulação de genes próximos.

4.5 Análise dos níveis de transcrição de genes *downstream* a inserções de TEs.

Em busca de avaliar os genes que foram encontrados próximos de inserções de TEs, foram analisados dados de RNA-seq a fim de comparar os níveis de transcrição entre populações de genes *downstream* de inserções de TEs e o restante dos genes. Para isso, os dados experimentais foram obtidos no repositório do NCBI de acordo com a Tabela 8.

Tabela 8 - Resultados de RNA-seq utilizados nas análises

Fase de vida	Acesso NCBI
Miracídio	SRR922067
Esporocisto	SRR922068
Cercária (4 h)	ERR022872
Cercária (4 h)	ERR022877
Cercária (4 h)	ERR022878
Esquistossômulo (3 h)	ERR022874
Esquistossômulo (3 h)	ERR022876
Esquistossômulo (3 h)	ERR022879
Esquistossômulo (24 h)	ERR022880
Esquistossômulo (24 h)	ERR022881
Adulto (7 sem)	ERR022873
Cauda	ERR022875

Fonte: Elaborada pelo autor.

O experimento de RNA-seq consiste na fragmentação de fitas de RNA maduro, transcrição reversa dos fragmentos para cDNA e sequenciamento (Figura 16). Posteriormente, os fragmentos sequenciados são alinhados contra o genoma a fim de se obter uma contagem de fragmentos para cada gene, que espera-se que seja proporcional à quantidade de RNAs maduros provenientes daquele gene,[58] o que permite a inferência do nível de transcrição do mesmo. Os valores absolutos dessas contagens, para cada gene, variam entre experimentos e são proporcionais à quantidade total de fragmentos sequenciados (tamanho da biblioteca) e ao tamanho do gene, visto que um gene maior terá como transcrito um RNA maior, que será fragmentado em mais pedaços.

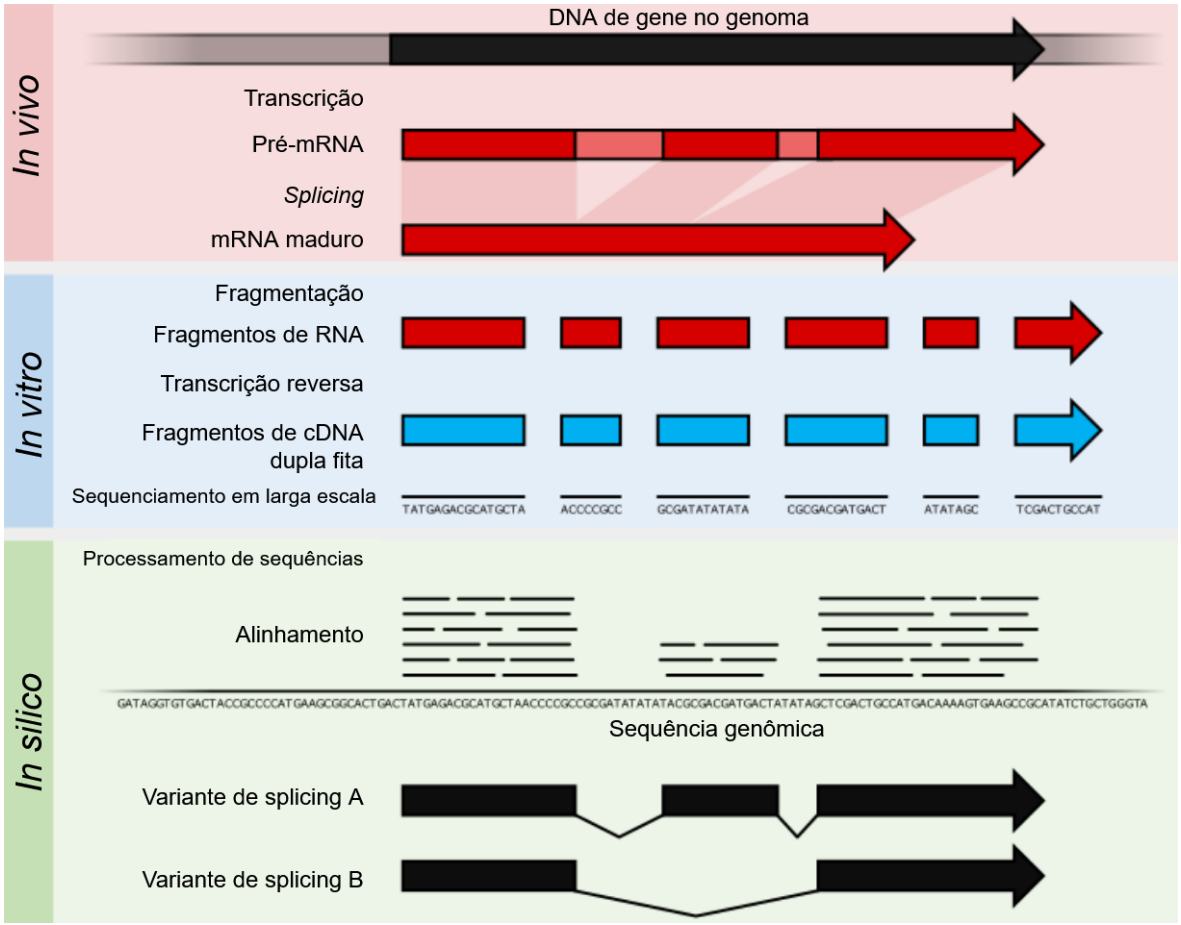

Figura 16 - Esquema do experimento de RNA-seq.
Fonte: Adaptada de LOWE *et al*.[59]

Considerando o mapeamento dos *reads* no genoma, primeiramente foram retirados os adaptadores residuais do sequenciamento (*trimming*) através do *software* Trimmomatic.[60] Novamente, foi utilizado o *software* HISAT2 para o mapeamento dos *reads* processados no genoma. Para a contagem das sequências alinhadas para cada gene foi utilizado o *software* HTSEQ-count,[61] com os parâmetros *union* e *nonunique all*.

As contagens de transcritos de cada gene do RNA-seq são proporcionais ao número total de sequências mapeadas e ao tamanho do gene. Para a comparação entre genes no mesmo experimento e entre experimentos, é necessário um procedimento de normalização. A métrica utilizada foi a FPKM (fragmentos por quilobase por milhão) (Equação 1).[62] O cálculo da normalização pelo número total de fragmentos foi realizado pela biblioteca *edgeR,* da linguagem de programação R, e a normalização pelo tamanho do gene foi realizada em um *script* em *Python*.

$$FPKM = \frac{sequências\ mapeadas\ para\ determinado\ gene}{\left(\frac{número\ total\ de\ sequências\ mapeadas}{10^6}\right)\left(\frac{comprimento\ gene}{10^3}\right)} \quad (1)$$

Foram comparados os níveis de transcrição dos genes *downstream* de inserções (intergênicas e colineares ao gene *downstream*) de cada TE com o restante da população genômica (controle), nas diversas fases de vida analisadas para *S. mansoni* (Figura 17 e Tabela 9). A fim de avaliar se as duas populações possuíam diferença significativa em termos de níveis de transcrição, foi aplicado o teste Mann-Whitney U.[60-61] Este teste foi realizado por meio da função *add_stat_annotation* da biblioteca *statannot* do Python; a mesma função também realizou a correção de Bonferroni para múltiplas comparações.

A Figura 17 ilustra a distribuição de frequências do nível de expressão de genes controle (box-plot em azul) e a distribuição de frequências do nível de expressão de genes *downstream* a cópias intergênicas e colineares (box-plot em laranja) para o elemento SR1, considerando-se a fase de vida cercária (4 horas). De forma análoga, a comparação entre o conjunto de genes *downstream* e o conjunto controle foi realizada para todos os TEs analisados neste trabalho, em cada fase do ciclo de vida do parasita *S. mansoni*, incluindo a análise estatística (teste Mann-Whitney U). As análises individuais constam no Apêndice A e o resultado geral é apresentado na Tabela 9.

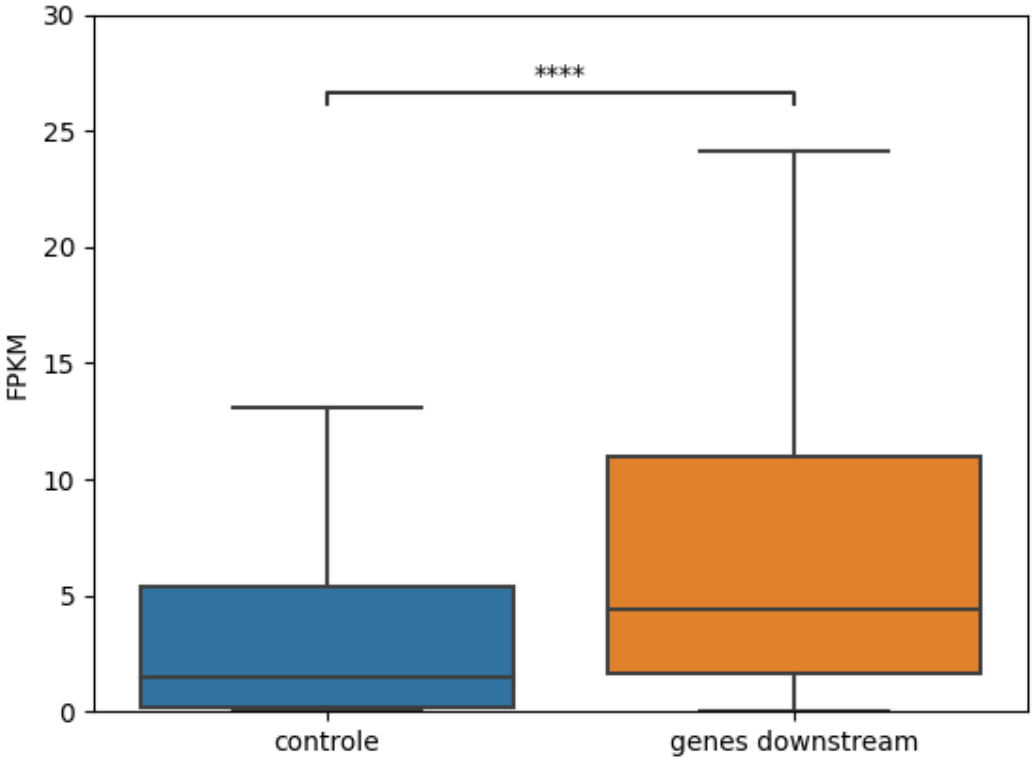

Figura 17 - Exemplo de comparação entre a população de genes downstream das inserções (intergênicas e colineares) do elemento SR1 e o restante da população de genes (controle) para a fase de vida cercária 4h. ****: p-valor <= 10-4.

Fonte: Elaborada pelo autor.

Tabela 9 - Comparação de populações de genes *downstream* (a inserções intergênicas e colineares) com o restante da população gênica (controle) em cada fase do ciclo de vida. Células em vermelho indicam valores de expressão gênica significativamente acima do esperado (p-valor<0,01), enquanto células em azul indicam valores significativamente abaixo do esperado; células em azul claro valores abaixo do esperado mas sem significância estatística e células em branco valores próximos ao esperado. Indicações de (1), (2) e (3) referem-se a réplicas amostrais.

	Miracídio	esporocisto	Cercária 4h(1)	Cercária 4h(2)	Cercária 4h (3)	Esquistossômulo 3h (1)	Esquistossômulo 3h (2)	Esquistossômulo 3h (3)	Esquistossômulo 24h (1)	Esquistossômulo 24h (2)	Adulto 7 semanas	Cauda
Perere												
Perere-2												
Perere-4												
Perere-5												
Perere-6												
Perere-7												
SR1												
SR2												
SR3/Perere-3												

Fonte: Elaborada pelo autor.

Na Tabela 9 pode-se notar que, com exceção do Perere, as populações de genes colineares e *downstream* de inserções de TEs da família CR1 tiveram níveis de transcrição significativamente acima do esperado em todas as fases do ciclo de vida avaliadas. A mesma análise foi realizada para o conjunto de inserções não colineares, mas não foi possível constatar diferença significativa de expressão gênica em relação ao controle. Conjuntamente com a análise de colinearidade, é possível supor que as inserções estejam desempenhando um papel regulatório no gene downstream mais próximo.

Observou-se que o TE Perere apresenta duas características desviantes do restante dos elementos da família CR1: (1) Possui a menor proporção de genes colineares entre os intergênicos (57% das inserções)(Tabela 4) e (2) os genes *downstream* às suas inserções intergênicas e colineares tiveram nível de expressão abaixo do esperado (Tabela 9), indicando que esse TE pode ter um mecanismo de inserção diferente dos demais.

Em relação ao clado RTE (Tabela 9), foi possível observar que genes *downstream* associados a elementos SR2 apresentam expressão similar ao grupo controle, enquanto que genes *downstream* associados a elementos SR3/Perere-3 apresentaram expressão gênica abaixo do observado para o grupo controle, em todas as fases de vida. Embora a maioria das inserções do TE SR2 seja intragênica (77%),

a população dos intergênicos colineares ainda não é tão baixa (139 elementos) a ponto de poder se atribuir a ausência de tendência à baixa população.

4.6 Análise da prevalência das inserções de TEs *upstream* a regiões de interação com histonas modificadas.

Segundo os dados anteriores, têm-se indícios de que TEs do clado CR1 contendo a sequência codificante para o domínio PHD apresentam inserções próximas a regiões com histona modificada. Restando saber ainda, dentre as modificações possíveis, qual seria a mais prevalente, se é que existiria prevalência. Para essa análise foram selecionados no banco NCBI os resultados de CHIP-seq de histonas com as modificações H3K27me3, H3K9me3, H3K9Ac e H3K4me3 (Tabela 9) para as fases de vida adulto e cercária do *S. mansoni*, que são todos os dados disponíveis para esse experimento e parasito. A trimetilação na lisina 4 em histonas (H3K4me3) é característica de eucromatina e facilitação da transcrição,[65] bem como a acetilação na lisina 9 (H3K9Ac).[66] Por outro lado, a trimetilação nas lisinas 9 e 27 está associada à heterocromatina e à repressão da transcrição.[26]

Tabela 10 - Modificações de histona e acessos dos dados de CHIP-seq retirados para análise.

Fase de vida	Modificação de histona	Acesso NCBI
Cercária	H3K27me3	SRX423996
	H3K9me3	SRX425465
	H3K9Ac	SRX424549
	H3K4me3	SRX424005
Adulto	H3K27me3	SRR1131932
	H3K9me3	SRR1138589
	H3K9Ac	SRR1138588
	H3K4me3	SRR1136063

Fonte: Elaborada pelo autor.

O processamento dos dados de sequenciamento provenientes do CHIP-seq para a identificação de regiões de interação com histona modificada foi realizado da mesma forma que no tópico “Análise de inserções de TEs em regiões *downstream* de interação com histonas modificadas”. Foram utilizados os dados de localização das inserções no genoma comuns a todas as análises deste trabalho. O objetivo da análise foi quantificar o número de inserções associadas a histonas com cada modificação e também avaliar qual a fração de inserções eram associadas a mais de um sítio de

histona modificada. Para a quantificação dos pares inserção – sítio de histona modificada, foram contabilizadas inserções com o 3’ a 1000pb *downstream* ou *upstream* do centro da região de histona (Figura 18).

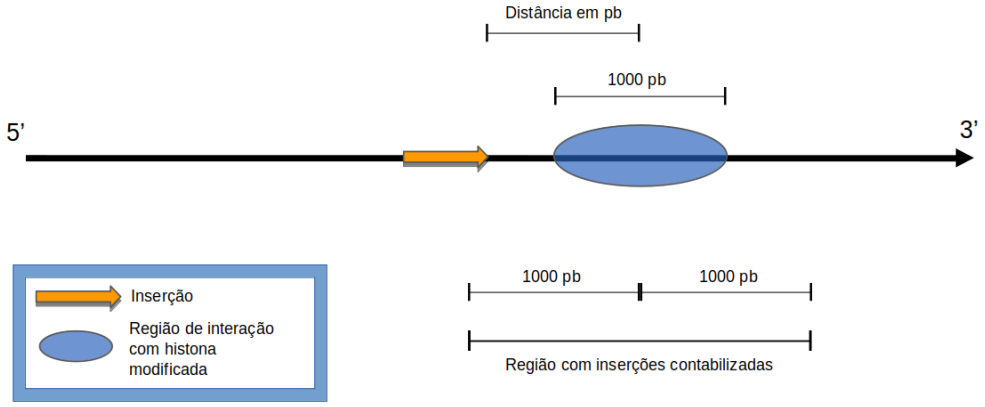

Figura 18 - Esquema ilustrativo da medida de distância entre as inserções de TEs e as regiões de interação com a histona com a modificação H3K4me3. Inserções com a extremidade 3’ a menos de 1000pb downstream e ou upstream da região de histona foram contabilizadas como próximas.
Fonte: Elaborada pelo autor.

Foram registradas as ocorrências de inserções próximas a cada modificação (tabelas 11 e 12), bem como a razão das inserções próximas a cada modificação frente a todas as inserções próximas a alguma modificação. Considerando uma distribuição uniforme das cópias de TEs no genoma, seria esperado que inserções sem nenhum tipo de direcionamento, mas próximas a algum sítio de histona modificada, fossem proporcionais à população de sítios de cada histona modificada frente a todos os sítios possíveis. Para avaliar se maiores ocorrências de inserção não refletiam apenas uma maior população de histonas de certa modificação, também foram calculadas as razões entre as contagens de cada histona modificada frente ao total. O teste z entre proporções foi utilizado para a estimativa do p-valor associado à razão esperada (inserções proporcionais às populações de histona com cada modificação) e observada.[67]

Na Tabela 11, referente à fase de vida adulta do *S. mansoni*, percebe-se que inserções próximas a histonas com modificação H3K4me3 (e H3K9Ac em alguns

casos) foram mais numerosas que o esperado caso não houvesse preferência por nenhuma modificação. Por outro lado, inserções em regiões de histonas com as modificações H3K27me3 e H3K9me3 estiveram menos representadas que o esperado, ainda que sejam modificações menos numerosas.

No caso da Tabela 12, que é referente aos dados de CHIP-seq de cercária, inserções próximas a histonas com modificação H3K4me3 continuam com representação acima do esperado e, analogamente para a modificação H3K27me3, houve representação abaixo do esperado.

Tabela 11 - Ocorrências de inserções de TEs (com região codificante para o domínio PHD) próximas a região de interação com histonas de diversas modificações, na fase de vida de adulto do *S. mansoni.* O p-valor foi calculado a partir do teste z de proporções entre a razão obtida e a esperada.

(adulto)		H3K27me3	H3K9me3	H3K9Ac	H3K4me3
Perere-2	Ocorrências	6	6	64	181
	Razão obtida	0.02	0.02	0.25	0.70
	Razão esperada	0.08	0.15	0.19	0.58
	p-valor	3.9e-09	1e-38	0.03	2.6e-05
Perere-5	Ocorrências	4	10	37	126
	Razão obtida	0.02	0.06	0.21	0.71
	Razão esperada	0.08	0.15	0.19	0.58
	p-valor	4.9e-07	2.6e-07	0.55	0.00019
Perere-6	Ocorrências	4	5	34	90
	Razão obtida	0.03	0.04	0.26	0.68
	Razão esperada	0.08	0.15	0.19	0.58
	p-valor	0.001	5.1e-11	0.086	0.023
SR1	Ocorrências	15	19	143	403
	Razão obtida	0.03	0.03	0.25	0.69
	Razão esperada	0.08	0.15	0.19	0.58
	p-valor	9.6e-16	6.4e-53	0.0018	8.2e-09
Contagem CHIP-seq		858	1589	2076	6366

Fonte: Elaborada pelo autor.

Tabela 12 - Ocorrências de inserções de TEs (com região codificante para o domínio PHD) próximas a região de interação com histonas de diversas modificações, na fase de vida de cercária do *S. mansoni.* O p-valor foi calculado a partir do teste z de proporções entre a razão obtida e a esperada.

(cercaria)		H3K27me3	H3K9me3	H3K9Ac	H3K4me3
Perere-2	Ocorrências	69	126	170	185
	Razão obtida	0.13	0.23	0.31	0.34
	Razão esperada	0.21	0.23	0.32	0.24
	p-valor	1.5e-09	0.87	0.59	9.5e-07
Perere-5	Ocorrências	55	83	121	129
	Razão obtida	0.14	0.21	0.31	0.33
	Razão esperada	0.21	0.23	0.32	0.24
	p-valor	9.7e-05	0.38	0.74	7.3e-05
Perere-6	Ocorrências	40	54	82	86
	Razão obtida	0.15	0.21	0.31	0.33
	Razão esperada	0.21	0.23	0.32	0.24
	p-valor	0.0089	0.3	0.82	0.0018
SR1	Ocorrências	153	235	377	405
	Razão obtida	0.13	0.20	0.32	0.35
	Razão esperada	0.21	0.23	0.32	0.24
	p-valor	4.8e-16	0.0077	0.85	6e-15
Contagem CHIP-seq		5410	5957	8203	6099

Fonte: Elaborada pelo autor.

Na análise das tabelas 11 e 12, observa-se que as cópias de todos os TEs CR1 com PHD foram encontradas em regiões de histona com modificações H3K4me3 em proporção significativamente acima do esperado, caso a distribuição fosse proporcional à quantidade de regiões de cada modificação encontrada no genoma. Em adultos, os TEs SR1 e Perere-2 tiveram também inserções que indicariam a mesma tendência para H3K9Ac. Histonas com modificação H3K4me3 estão relacionadas com ativação da eucromatina e histonas com modificação H3K9Ac com ativação da transcrição.[66-67] Sob a hipótese que o PHD reconheceria as modificações de histona e mediaria as inserções para regiões próximas dessas modificações, pode-se especular que regiões com tais propriedades facilitariam a transcrição do TE e sua replicação no genoma.

Ainda há a possibilidade de que a mesma região seja alvo de dois ou mais tipos de modificação de histona, podendo ocasionar super-representação de modificações que acontecem em regiões próximas. Foram construídos gráficos de Venn (figuras 19 e 20) a fim de avaliar como as regiões de histonas são compartilhadas entre as quatro

modificações estudadas. Como apenas H3K4me3 e H3K9Ac apresentaram populações acima do esperado, são apresentados apenas os diagramas referentes a essas modificações, estando os demais disponíveis no Apêndice B. No caso de duas regiões de interação com histonas de diferentes modificações estarem vinculadas à mesma cópia de TE, então essas regiões foram contabilizadas na intersecção do diagrama de Venn.

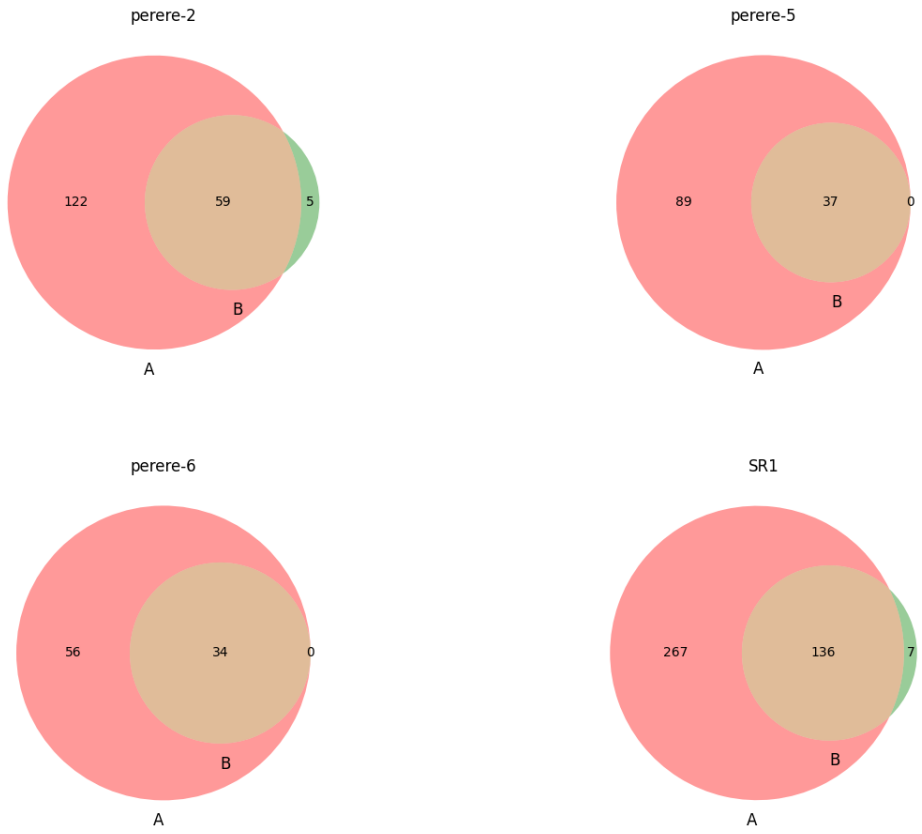

Figura 19 - Diagramas de Venn das regiões compartilhadas entre as histonas com modificação H3K4me3(A) e H3K9Ac(B), em adulto, para os TEs com domínio PHD.
Fonte: Elaborada pelo autor.

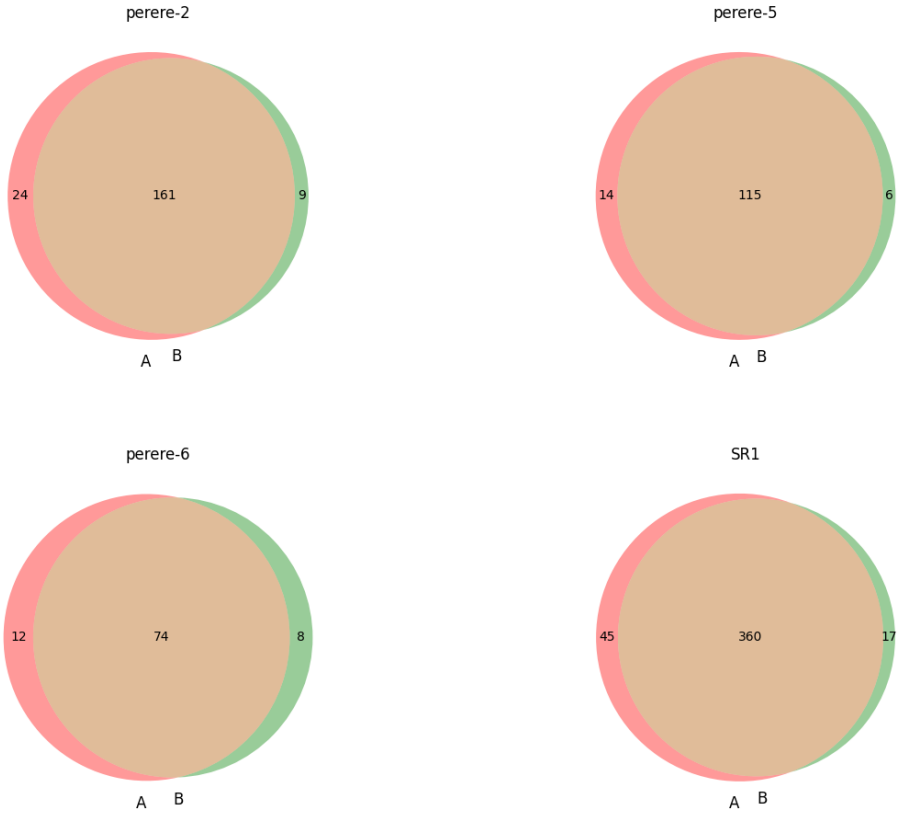

Figura 20 - Diagramas de Venn das regiões compartilhadas entre as histonas com modificação H3K4me3(A) e H3K9Ac(B), em cercária, para os TEs com domínio PHD.
Fonte: Elaborada pelo autor.

Em cercárias (Figura 20), os sítios são quase na totalidade compartilhados, embora as regiões únicas de H3k4me3 sejam levemente mais numerosas. Em adultos (Figura 19), inserções próximas de H3K9Ac também estiveram próximas de H3K4me3, mas o contrário não pode ser dito, o que indica que as contagens acima da média de H3K9Ac podem ser resultado de regiões comuns com H3K4me3, o que também indica que, se existe algum direcionamento da transposição para histonas modificadas, a H3K4me3, dentre as estudadas parece ser preferencial. Experimentos análogos para as outras fases de vida, assim como experimentos especificamente de células germinativas, no futuro poderão enriquecer e ajudar a esclarecer os resultados apresentados.

4.7 Análise filogenética dos domínios PHDs

Algumas interações de domínios PHD com histonas modificadas foram descritas na literatura, primeiramente o reconhecimento do estado de metilação da H3K4 (H3K4me0 e H3K4me3/2) e,[30-68] em seguida, estados de metilação de H3R2,[71] H3K36 e de acetilação em H3K14.[72-73]

Através de uma análise filogenética é possível comparar as sequências de aminoácidos dos domínios PHD descritos acima com as sequências dos domínios relativos aos TEs do clado CR1 presentes nesse trabalho. Para tanto, a princípio foi necessária a reconstrução do domínio PHD do TE SR1, cuja sequência depositada apresentava apenas cerca de metade do número de nucleotídeos descritos para os outros TEs do mesmo clado. Nesta sequência parcial, não foram encontrados indícios, via *softwares* de predição de domínios, das sequências codificadoras de domínios essenciais para a sobrevivência para TEs do clado, como da transcriptase reversa e endonuclease. Além disso, como os TEs possuem a extremidade 3' mais conservada devido à baixa processividade da transcriptase reversa no momento da inserção, também se supôs que a parte faltante era no 5'.

A reconstrução seguiu os moldes descritos em Philippsen, Gisele S., and Ricardo DeMarco, 2020 (Figura 21).[74] A *priori,* realizou-se um blastn da extremidade 5' do trecho disponível do TE SR1, contra o banco de ESTs do NCBI. Foram utilizados 300pb, que correspondem a cerca de um oitavo da sequência disponível do SR1 (2337pb). Era selecionado o melhor alinhamento que alinhasse na extremidade 5' da sequência alvo, então, a parte faltante era anexada na sequência alvo. Em seguida, uma nova sequência alvo era selecionada na extremidade 5' para um novo alinhamento. Eventualmente, o processo culminava em uma sequência que não seria mais ampliada, finalizando a busca. Por fim, a sequência reconstruída foi submetida novamente ao *blasn* contra o banco de ESTs para a busca manual de nucleotídeos consenso nas regiões de *gaps*. Ao final, foi possível reconhecer as duas ORFs principais dos elementos CR1, além das sequências codificadoras para os domínios da transcriptase reversa, endonuclease e PHD.

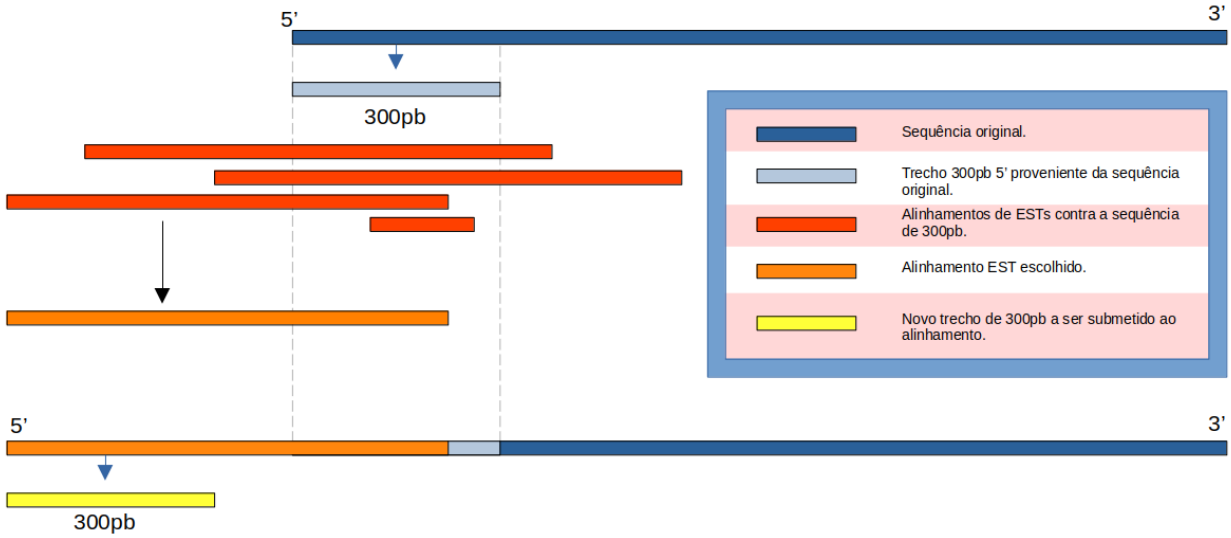

Figura 21 - Esquema do método utilizado para alongamento da extremidade 5' faltante do TE SR1. Um trecho de 300pb é submetido ao BLAST, o alinhamento mais promissor é selecionado e as partes são juntadas. A seguir, um novo query de 300pb é submetido ao BLAST e assim por diante até um comprimento satisfatório.
Fonte: Elaborada pelo autor.

Para dar prosseguimento à análise filogenética, sequências de aminoácidos de domínios PHD distintos foram obtidas em Sanchez *et al*. 2011. Os domínios dos TEs com PHD presentes neste trabalho (SR1, perere-2, perere-5 e perere-6) foram traduzidos e reunidos com as demais sequências utilizadas (Tabela 13) para realizar um alinhamento múltiplo com o *software muscle*.[75] Em seguida, foi criada a árvore com o *software MrBayes*, utilizando-se os parâmetros apropriados para aminoácidos. O resultado foi visualizado com o *software figtree* (http://tree.bio.ed.ac.uk/software/figtree/) e é mostrado na Figura 22.

Tabela 13 - Domínios PHD utilizados na análise filogenética e o tipo de modificação de histona preferencial para interação, bem como a referência na qual a interação foi reportada.

Nome da sequência	Referência	Interação primária com mod. De histona
BPTF	(LI et al., 2006)	H3K4me3/2
PHF2	(WEN et al., 2010)	H3K4me3/2
Yng1	(TAVERNA et al., 2006)	H3K4me3/2
ING4	(HUNG et al., 2009)	H3K4me3/2
TAF3	(VAN INGEN et al., 2008)	H3K4me3/2
RAG2	(RAMÓN-MAIQUES et al., 2007)	H3K4me3/2
Pygo	(FIEDLER et al., 2008)	H3K4me3/2
MLL1	(WANG et al., 2010)	H3K4me3/2
JARID1A	(WANG et al., 2009)	H3K4me3/2
BHC80	(LAN et al., 2007)	H3K4me0
AIRE	(CHIGNOLA et al., 2009)	H3K4me0
DNMT3L	(OOI et al., 2007)	H3K4me0
TRIM24	(TSAI et al., 2010)	H3K4me0
DPF3B2	(ZENG et al., 2010)	H3K4me0
DPF3B1	(ZENG et al., 2010)	H3K14Ac
phd_perere2		?
phd_perere-5		?
phd_perere6		?
phd_SR1		?

Fonte: Elaborada pelo autor.

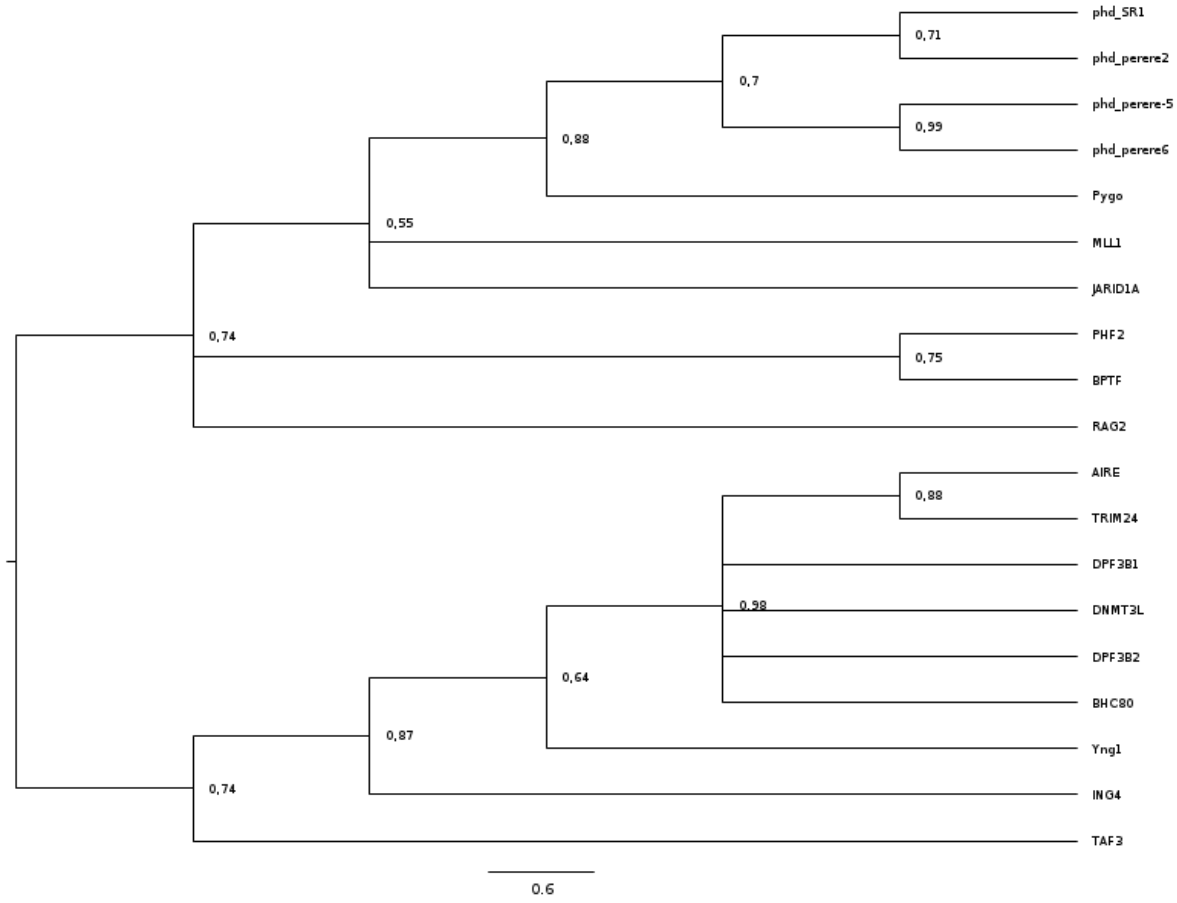

Figura 22: Árvore filogenética de domínios PHD diversos, calculada com o software MrBayes e visualizada com o software FigTree.
Fonte: Elaborada pelo autor.

Analisando em conjunto a Figura 15 e a Tabela 12, percebemos que os domínios PHD dos elementos Pererê e SR1 estiveram próximos de domínios em que foi reportada interação preferencial com as histonas metiladas (H3K4me3/2), o que reforça a possibilidade do domínio PHD transcrito por esses elementos possuir interação preferencial com esse tipo de modificação de histona.

4.8 Análise de especificidade de transcrição dos genes próximos a TEs

Também é interessante avaliar se os genes próximos de TE possuem transcritos concentrados em fases de vida ou se estes aparecem de forma contínua durante todo o ciclo de vida de *S. mansoni*. No caso da primeira opção, haveria a suspeita da inserção estar relacionada com os processos enzimáticos envolvidos em fases de vida específicas. No caso da segunda opção, poder-se-ia supor que a fase

de vida não é determinante para a inserção dos TEs ou que genes continuamente expressos se beneficiam de sua propagação de alguma maneira.

Para isso foi utilizada a métrica de tecido-especificidade (τ).[76] Segundo essa métrica, o τ é definido, para cada gene, como:

$$\tau = \frac{\sum_{i=1}^{N} (1-x_i)}{N-1} \qquad (2)$$

Em que x_i é o nível de transcrição do gene para cada tecido e N o número de tecidos. Um τ = 0 significa que a transcrição em todos os tecidos foi idêntica, ao passo que um τ próximo de 1 significaria grande variedade nos níveis de transcrição. É necessário que os valores de x sejam normalizados pelo valor máximo. Neste trabalho, N foi interpretado como o número de fases de vida de *S. mansoni*, x_i como o nível de transcrição, normalizado, em determinada fase de vida e o τ como uma métrica para avaliar a especificidade de transcrição de cada gene nas fases de vida do parasito.

Os níveis de transcrição de cada gene, em cada fase de vida, foram os mesmos da seção anterior “Análise dos níveis de transcrição de genes downstream a inserções de TEs”, mais especificamente na Tabela 7. Também foi realizado o mesmo tratamento dos dados, desde a trimagem até a normalização.

Para cada TE, os genes de *S. mansoni* foram separados em duas populações: aqueles que foram considerados *downstream* segundo a análise citada anteriormente e a população total de genes do indivíduo. Foi medido o τ de cada gene e o objetivo dessa comparação foi avaliar se os genes *downstream* de TEs possuem tendências de especificidade de transcrição em alguma fase de vida, sendo o restante da população genômica considerado como controle. Foi avaliado se as duas populações são estatisticamente diferentes com o teste de Mann-Whitney-Wilcoxon com correção de Bonferroni. Os resultados foram apresentados na forma de boxplot e divididos entre TEs do clado CR1 (Figura 23) e RTE (Figura 24).

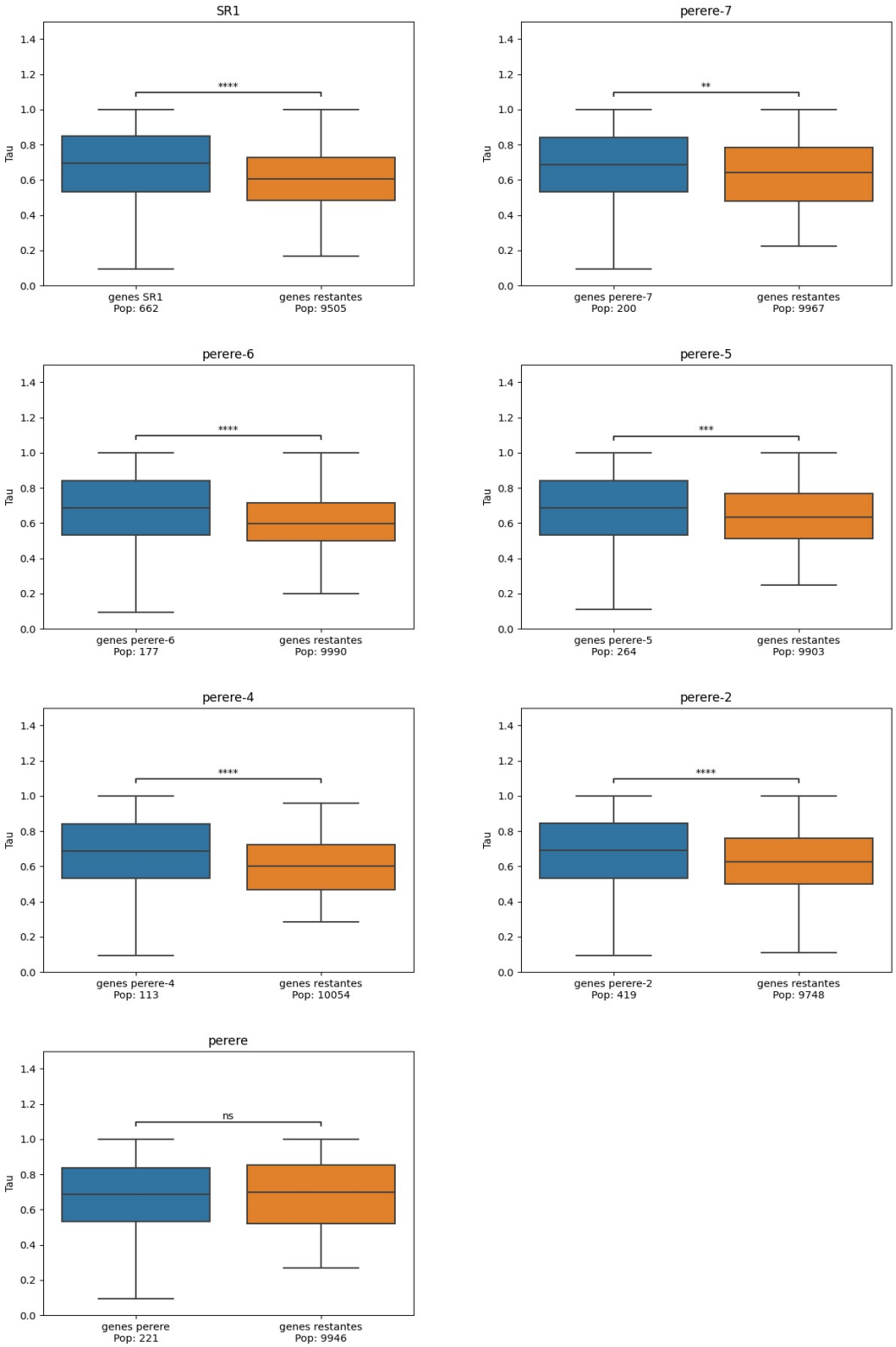

Figura 23: Análise do tau entre as populações genômicas de S.mansoni para TEs do clado CR1. O teste estatístico utilizado foi o Mann-Whitney-Wilcoxon com correção de Bonferroni (ns: 5.00e-02 < p <= 1.00e+00; *: 1.00e-02 < p <= 5.00e-02; **: 1.00e-03 < p <= 1.00e-02; ***: 1.00e-04 < p <= 1.00e-03; ****: p <= 1.00e-04).

Fonte: Elaborada pelo autor.

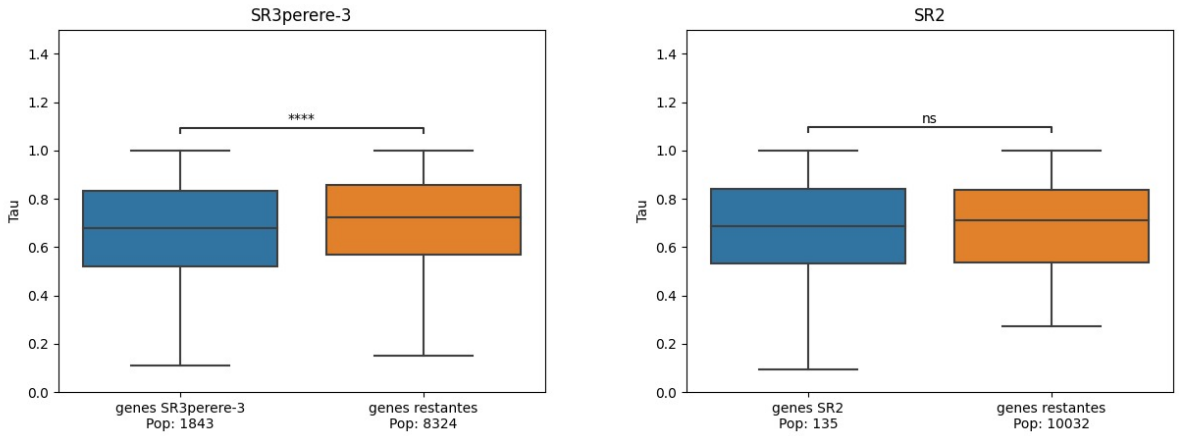

Figura 24: Análise do tau entre as populações genômicas de S.mansoni para TEs do clado RTE. O teste estatístico utilizado foi o Mann-Whitney-Wilcoxon com correção de Bonferroni (ns: 5.00e-02 < p <= 1.00e+00; *: 1.00e-02 < p <= 5.00e-02; **: 1.00e-03 < p <= 1.00e-02; ***: 1.00e-04 < p <= 1.00e-03; ****: p <= 1.00e-04).
Fonte: Elaborada pelo autor.

Analisando a Figura 23, pode-se observar que os genes *downstream* aos elementos do clado CR1 apresentaram a mediana para o τ acima do valor desta estatística para a população restante de genes de *S. mansoni*, com exceção do TE Perere. Nos TEs do clado RTE (Figura 24) tal comportamento não foi observado. Um valor de τ alto indica que os níveis de transcrição não se mantêm de forma homogênea nas fases de vida, ou seja, existe especificidade de transcrição. Esse valor pode indicar que o gene é transcrito para gerar proteínas envolvidas com processos mais específicos de certas fases de vida, como os genes envolvidos nas transições de uma fase para outra. Entretanto, mesmo que as medianas entre os genes *downstream* do clado CR1 e o grupo controle exibam diferenças significativas do ponto de vista estatístico, a distribuição dos dados é semelhante para os dois conjuntos avaliados, o que sugere que o impacto biológico não seja relevante.

5 CONCLUSÃO

As inserções de elementos CR1 no genoma de *S. mansoni* foram encontradas como majoritariamente colineares, intergênicas e *upstream* a genes com transcrição acima da média. Uma vez que inserções em regiões gênicas possuem maior probabilidade de interferirem negativamente no funcionamento do gene, tendendo a ser selecionadas negativamente, seria esperado que os elementos apresentassem maior população de cópias em regiões intergênicas. Porém, o fato de serem colineares à genes *downstream* com níveis de transcrição acima da média sugere um direcionamento senso-específico para tais regiões ou a possibilidade de seleção positiva proveniente de um papel regulatório de TEs desse clado, como por exemplo a adição de um novo promotor ou elemento regulatório *cis*. Ainda, haveria a possibilidade de se considerar eventos de seleção negativa, mas isso não explicaria a distribuição observada, já que os elementos do clado RTE tiveram comportamentos diferentes em todos os casos. Observou-se que genes *downstream* de TEs do clado CR1 apresentaram uma leve tendência a possuírem níveis de transcrição diferentes nas várias fases de vida do parasita, o que poderia significar que as inserções estejam relacionadas com processos mais específicos. O elemento Perere, mesmo sendo classificado com o clado CR1, apresentou um comportamento divergente do restante do clado.

Os elementos CR1 contendo a região codificante do domínio PHD apresentaram, além das características já apresentadas para o clado CR1, parte significativa da população de cópias a menos de 500 pares de bases do gene *downstream* e de regiões de histonas com modificação H3K4me3. Demais elementos da mesma família, mas sem o domínio PHD, não apresentaram tal comportamento. Considerando que domínios PHD são descritos como interagentes com histonas modificadas, principalmente a modificação H3K4me3, este resultado sugere um direcionamento das inserções de CR1, mediado pelo domínio PHD, para regiões *upstream* muito próximas do local de início de transcrição de certos genes.

Ainda, cópias de todos os TEs CR1 com PHD foram encontradas em regiões de histona com a modificação H3K4me3 em proporção significativamente acima do esperado, frente às outras possíveis modificações de histona. Sabendo que H3K4me3 é uma modificação relacionada à ativação da eucromatina, pode-se especular que regiões com tais propriedades facilitariam a transcrição do TE e sua replicação no

genoma. É possível que os elementos do clado CR1 que possuem o PHD tenham encontrado uma “zona segura” de mecanismos de repressão epigenéticos ao se inserirem em regiões com histonas com modificações associadas à ativação de eucromatina (H3K4me3). A análise filogenética desse domínio também apontou que os domínios PHD em estudo são próximos de domínios PHD descritos com interações preferenciais com histonas com modificação H3K4me3 do que outras modificações, corroborando os demais resultados.

Em conjunto, os resultados deste trabalho apontam para a influência de TEs na dinâmica genômica do parasita *S. mansoni,* embora experimentos ainda sejam necessários para comprovar a existência dessas relações. Se as relações forem comprovadas, o resultado será de grande importância para o entendimento da influência de elementos de transposição na arquitetura do genoma do parasito *Schistosoma mansoni*. Além disso, resultados experimentais que confirmem o direcionamento dos elementos mediados pelo domínio PHD trarão luz para esclarecer o papel da ORF1 de elementos de transposição não-LTR, bem como para revelar um novo mecanismo de transposição mediado por esse domínio.

REFERÊNCIAS

1 NURK, S. *et al.* The complete sequence of a human genome. **Science**, v. 376, n. 6588, p. 44–53, Apr. 2022.

2 WELLS, J. N.; FESCHOTTE, C. A field guide to eukaryotic transposable elements. **Annual Review of Genetics**, v. 54, p. 539–561, Nov. 2020.

3 WICKER, T. *et al.* A unified classification system for eukaryotic transposable elements. **Nature Reviews Genetics**, v. 8, n. 12, p. 973–982, Dec. 2007.

4 FESCHOTTE, C. Transposable elements and the evolution of regulatory networks. **Nature Reviews. Genetics**, v. 9, n. 5, p. 397–405, May 2008.

5 WICKER, T. *et al.* A unified classification system for eukaryotic transposable elements. **Nature Reviews Genetics**, v. 8, n. 12, p. 973–982, Dec. 2007.

6 FINNEGAN, D. J. Eukaryotic transposable elements and genome evolution. **Trends in genetics: TIG**, v. 5, n. 4, p. 103–107, Apr. 1989.

7 MARTIN, S. L. *et al.* LINE-1 retrotransposition requires the nucleic acid chaperone activity of the ORF1 protein. **Journal of Molecular Biology**, v. 348, n. 3, p. 549–561, May 2005.

8 KOLOSHA, V. O.; MARTIN, S. L. High-affinity, non-sequence-specific RNA binding by the open reading frame 1 (ORF1) protein from long interspersed nuclear element 1 (LINE-1). **Journal of Biological Chemistry**, v. 278, n. 10, p. 8112–8117, Mar. 2003.

9 SULTANA, T. *et al.* Integration site selection by retroviruses and transposable elements in eukaryotes. **Nature Reviews Genetics**, v. 18, n. 5, p. 292–308, Maio 2017.

10 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, June 1950.

11 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, p. 21–42, 2012.

12 JORDAN, I. K. *et al.* Origin of a substantial fraction of human regulatory sequences from transposable elements. **Trends in Genetics: TIG**, v. 19, n. 2, p. 68–72, Feb. 2003.

13 NEKRUTENKO, A.; LI, W. H. Transposable elements are found in a large number of human protein-coding genes. **Trends in genetics: TIG**, v. 17, n. 11, p. 619–621, Nov. 2001.

14 LANDRY, J. R.; MEDSTRAND, P.; MAGER, D. L. Repetitive elements in the 5’ untranslated region of a human zinc-finger gene modulate transcription and translation efficiency. **Genomics**, v. 76, n. 1–3, p. 110–116, Aug.2001.

15 KAMAL, M.; XIE, X.; LANDER, E. S. A large family of ancient repeat elements in the human genome is under strong selection. **Proceedings of the National Academy of Sciences**, v. 103, n. 8, p. 2740–2745, Feb. 2006.

16 BURKE, W. D.; CALALANG, C. C.; EICKBUSH, T. H. The site-specific ribosomal insertion element type II of Bombyx mori (R2Bm) contains the coding sequence for a reverse transcriptase-like enzyme. **Molecular and Cellular Biology**, v. 7, n. 6, p. 2221–2230, June 1987.

17 QUADRANA, L. *et al.* The Arabidopsis thaliana mobilome and its impact at the species level. **eLife**, v. 5, p. e15716, June 2016.

18 SPALLER, T. *et al.* Convergent evolution of tRNA gene targeting preferences in compact genomes. **Mobile DNA**, v. 7, n. 1, p. 17, Aug. 2016.

19 GAO, X. *et al.* Chromodomains direct integration of retrotransposons to heterochromatin. **Genome Research**, v. 18, n. 3, p. 359–369, Mar. 2008.

20 HAN, J. **Histone mutations and cancer**.Singapure: Springer,2021.

21 ZAIDI, S. *et al.* De novo mutations in histone-modifying genes in congenital heart disease. **Nature**, v. 498, n. 7453, p. 220–223, June 2013.

22 STRAHL, B. D.; ALLIS, C. D. The language of covalent histone modifications. **Nature**, v. 403, n. 6765, p. 41–45, Jan. 2000.

23 KOUZARIDES, T. Chromatin modifications and their function. **Cell**, v. 128, n. 4, p. 693–705, Feb. 2007.

24 GREER, E. L.; SHI, Y. Histone methylation: a dynamic mark in health, disease and inheritance. **Nature Reviews Genetics**, v. 13, n. 5, p. 343–357, Apr. 2012.

25 BLACK, J. C.; VAN RECHEM, C.; WHETSTINE, J. R. Histone lysine methylation dynamics: establishment, regulation, and biological impact. **Molecular Cell**, v. 48, n. 4, p. 491–507, Nov. 2012.

26 BANNISTER, A. J.; KOUZARIDES, T. Regulation of chromatin by histone modifications. **Cell Research**, v. 21, n. 3, p. 381–395,Mar.2011.

27 DI CERBO, V. *et al.* Acetylation of histone H3 at lysine 64 regulates nucleosome dynamics and facilitates transcription. **eLife**, v. 3, p. e01632, Mar. 2014.

28 SUKA, N. *et al.* Highly specific antibodies determine histone acetylation site usage in yeast heterochromatin and euchromatin. **Molecular Cell**, v. 8, n. 2, p. 473–479, Aug. 2001.

29 LAWRENCE, M.; DAUJAT, S.; SCHNEIDER, R. Lateral thinking: how histone modifications regulate gene expression. **Trends in Genetics**, v. 32, n. 1, p. 42–56, Jan. 2016.

30 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,July 2005.

31 LI, H. *et al.* Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. **Nature**, v. 442, n. 7098, p. 91–95, July 2006.

32 BAKER, L. A.; ALLIS, C. D.; WANG, G. G. PHD fingers in human diseases: Disorders arising from misinterpreting epigenetic marks. **Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis:** Epigenetics of Development and Human Disease. v. 647, n. 1, p. 3–12, Dec. 2008.

33 BERMAN, H. M. *et al.* The protein data bank. **Nucleic Acids Research**, v. 28, n. 1, p. 235–242, Jan. 2000.

34 SANCHEZ, R.; ZHOU, M.-M. The PHD finger: a versatile epigenome reader. **Trends in Biochemical Sciences**, v. 36, n. 7, p. 364–372,July 2011.

35 VAN ZUTVEN, L. J. C. M. *et al.* Identification of NUP98 abnormalities in acute leukemia: JARID1A (12p13) as a new partner gene. **Genes, Chromosomes & Cancer**, v. 45, n. 5, p. 437–446, May 2006.

36 PEÑA, P. V. *et al.* Molecular mechanism of histone H3K4me3 recognition by plant homeodomain of ING2. **Nature**, v. 442, n. 7098, p. 100–103, July. 2006.

37 MATTHEWS, A. G. W. *et al.* RAG2 PHD finger couples histone H3 lysine 4 trimethylation with V(D)J recombination. **Nature**, v. 450, n. 7172, p. 1106–1110, Dec. 2007.

38 ABOU-EL-NAGA, I. F. Schistosoma mansoni sarco/endoplasmic reticulum Ca2+ ATPases (SERCA): role in reduced sensitivity to praziquantel. **Journal of Bioenergetics and Biomembranes**, v. 52, n. 5, p. 397–408, Oct. 2020.

39 MCMANUS, D. P. *et al.* Schistosomiasis. **Nature Reviews:** disease primers, v. 4, n. 1, p. 13, Aug. 2018.

40 ZWANG, J.; OLLIARO, P. L. Clinical efficacy and tolerability of praziquantel for intestinal and urinary schistosomiasis-a meta-analysis of comparative and non-comparative clinical trials. **PLoS neglected tropical diseases**, v. 8, n. 11, p. e3286, 2014.

41 REY, O. *et al.* Population genetics of African Schistosoma species. infection, genetics and evolution: **Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases**, v. 89, p. 104727, Apr. 2021.

42 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, May 2010.

43 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, Mar. 2003.

44 SIMPSON, A. J.; SHER, A.; MCCUTCHAN, T. F. The genome of Schistosoma mansoni: isolation of DNA, its size, bases and repetitive sequences. **Molecular and Biochemical Parasitology**, v. 6, n. 2, p. 125–137, Aug.1982.

45 LAHA, T. *et al.* Characterization of SR3 reveals abundance of non-LTR retrotransposons of the RTE clade in the genome of the human blood fluke, Schistosoma mansoni. **BMC genomics**, v. 6, p. 154,Nov. 2005.

46 BERRIMAN, M. *et al.* The genome of the blood fluke Schistosoma mansoni. **Nature**, v. 460, n. 7253, p. 352–358, July 2009.

47 HAAS, N. B. *et al.* Chicken repeat 1 (CR1) elements, which define an ancient family of vertebrate non-LTR retrotransposons, contain two closely spaced open reading frames. **Gene**, v. 197, n. 1–2, p. 305–309, Sept. 1997.

48 MALIK, H. S.; EICKBUSH, T. H. The RTE class of non-LTR retrotransposons is widely distributed in animals and is the origin of many SINEs. **Molecular Biology and Evolution**, v. 15, n. 9, p. 1123–1134, Sept.1998.

49 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, Sept.1999.

50 LEVIN, H. L.; MORAN, J. V. Dynamic interactions between transposable elements and their hosts. **Nature Reviews Genetics**, v. 12, n. 9, p. 615–627, Aug. 2011.

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, May 2011.

52 BELLEN, H. J. *et al.* The Drosophila gene disruption project: progress using transposons with distinctive site specificities. **Genetics**, v. 188, n. 3, p. 731–743, July 2011.

53 KIM, D.; LANGMEAD, B.; SALZBERG, S. L. HISAT: a fast spliced aligner with low memory requirements. **Nature Methods**, v. 12, n. 4, p. 357–360, Apr. 2015.

54 HEINZ, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. **Molecular Cell**, v. 38, n. 4, p. 576–589, May 2010.

55 FUREY, T. S. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. **Nature Reviews. Genetics**, v. 13, n. 12, p. 840–852, Dec. 2012.

56 PEREIRA, V. Insertion bias and purifying selection of retrotransposons in the Arabidopsis thaliana genome. **Genome Biology**, v. 5, n. 10, p. R79, 2004.

57 SLOTKIN, R. K.; MARTIENSSEN, R. Transposable elements and the epigenetic regulation of the genome. **Nature Reviews. Genetics**, v. 8, n. 4, p. 272–285, Apr. 2007.

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

59 LOWE, R. *et al.* Transcriptomics technologies. **PLoS computational biology**, v. 13, n. 5, p. e1005457, 2017.

60 BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics (Oxford, England)**, v. 30, n. 15, p. 2114–2120, Aug. 2014.

61 ANDERS, S.; PYL, P. T.; HUBER, W. HTSeq--a Python framework to work with high-throughput sequencing data. **Bioinformatics (Oxford, England)**, v. 31, n. 2, p. 166–169, Jan. 2015.

62 MORTAZAVI, A. *et al.* Mapping and quantifying mammalian transcriptomes by RNA-Seq. **Nature Methods**, v. 5, n. 7, p. 621–628, July 2008.

63 MANN, H. B.; WHITNEY, D. R. On a test of whether one of two random variables is stochastically larger than the other. **Annals of Mathematical Statistics**, v. 18, n. 1, p. 50–60, 1947.

64 WILCOXON, F. Individual comparisons by ranking methods**.** *In*: KOTZ, S., JOHNSON, N.L. (eds) **Breakthroughs in statistics.** New York: Springer,1992. p.80-83.Series in Statistics.

65 SANTOS-ROSA, H. *et al.* Active genes are tri-methylated at K4 of histone H3. **Nature**, v. 419, n. 6905, p. 407–411, sept. 2002.

66 POKHOLOK, D. K. *et al.* Genome-wide map of nucleosome acetylation and methylation in yeast. **Cell**, v. 122, n. 4, p. 517–527, Aug.2005.

67 SPRINTHALL, R. C. **Basic statistical analysis**. 9. ed. New York: Pearson, 2011.

68 DI CERBO, V. *et al.* Acetylation of histone H3 at lysine 64 regulates nucleosome dynamics and facilitates transcription. **eLife**, v. 3, p. e01632, Mar. 2014.

69 SUKA, N. *et al.* Highly specific antibodies determine histone acetylation site usage in yeast heterochromatin and euchromatin. **Molecular Cell**, v. 8, n. 2, p. 473–479, Aug. 2001.

70 TAVERNA, S. D. *et al.* Yng1 PHD finger binding to H3 trimethylated at K4 promotes NuA3 HAT activity at K14 of H3 and transcription at a subset of targeted ORFs. **Molecular Cell**, v. 24, n. 5, p. 785–796, Dec. 2006.

71 CHAKRAVARTY, S.; ZENG, L.; ZHOU, M.-M. Structure and site-specific recognition of histone H3 by the PHD finger of human autoimmune regulator. **Structure**, v. 17, n. 5, p. 670–679, May 2009.

72 SHI, X. *et al.* Proteome-wide analysis in Saccharomyces cerevisiae identifies several PHD fingers as novel direct and selective binding modules of histone H3 methylated at either lysine 4 or lysine 36. **Journal of Biological Chemistry**, v. 282, n. 4, p. 2450–2455, Jan. 2007.

73 ZENG, L. *et al.* Mechanism and regulation of acetylated histone binding by the tandem PHD finger of DPF3b. **Nature**, v. 466, n. 7303, p. 258–262,July 2010.

74 PHILIPPSEN, G. S.; DEMARCO, R. Identification of transposable elements in Schistosoma mansoni. **Methods in Molecular Biology,** v. 2151, p. 135–144, 2020.

75 EDGAR, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. **Nucleic Acids Research**, v. 32, n. 5, p. 1792–1797, Mar. 2004.

76 YANAI, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. **Bioinformatics,** v. 21, n. 5, p. 650–659, Mar. 2005.

APÊNDICE A - Análise dos níveis de transcrição de genes *downstream* a inserções de TEs

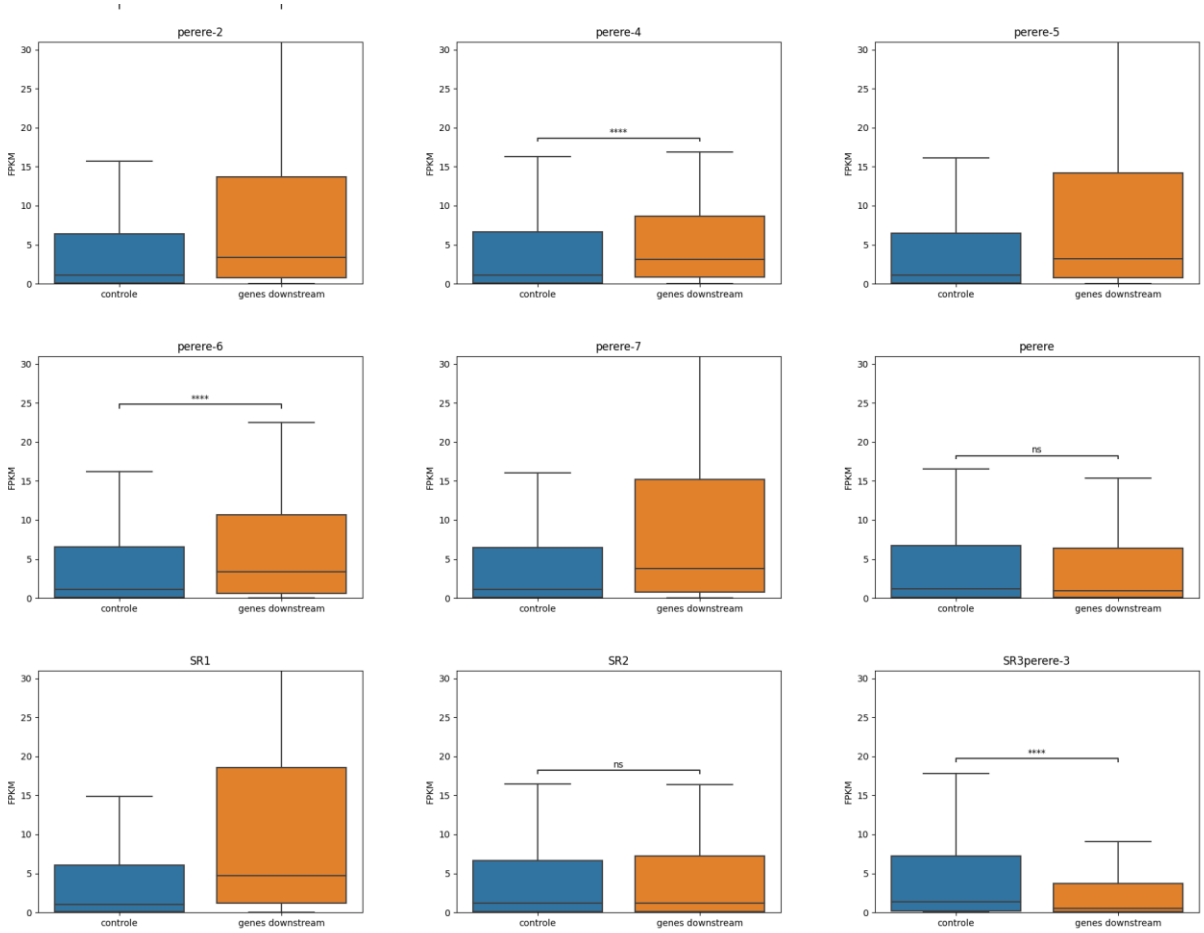

Figura A.1 - Comparação entre a população de genes downstream das inserções (intergênicas e colineares) dos TEs avaliados e o restante da população de genes (controle) para a fase de vida miracídio. ****: p-valor <= 10-4.

Fonte: Elaborada pelo autor.

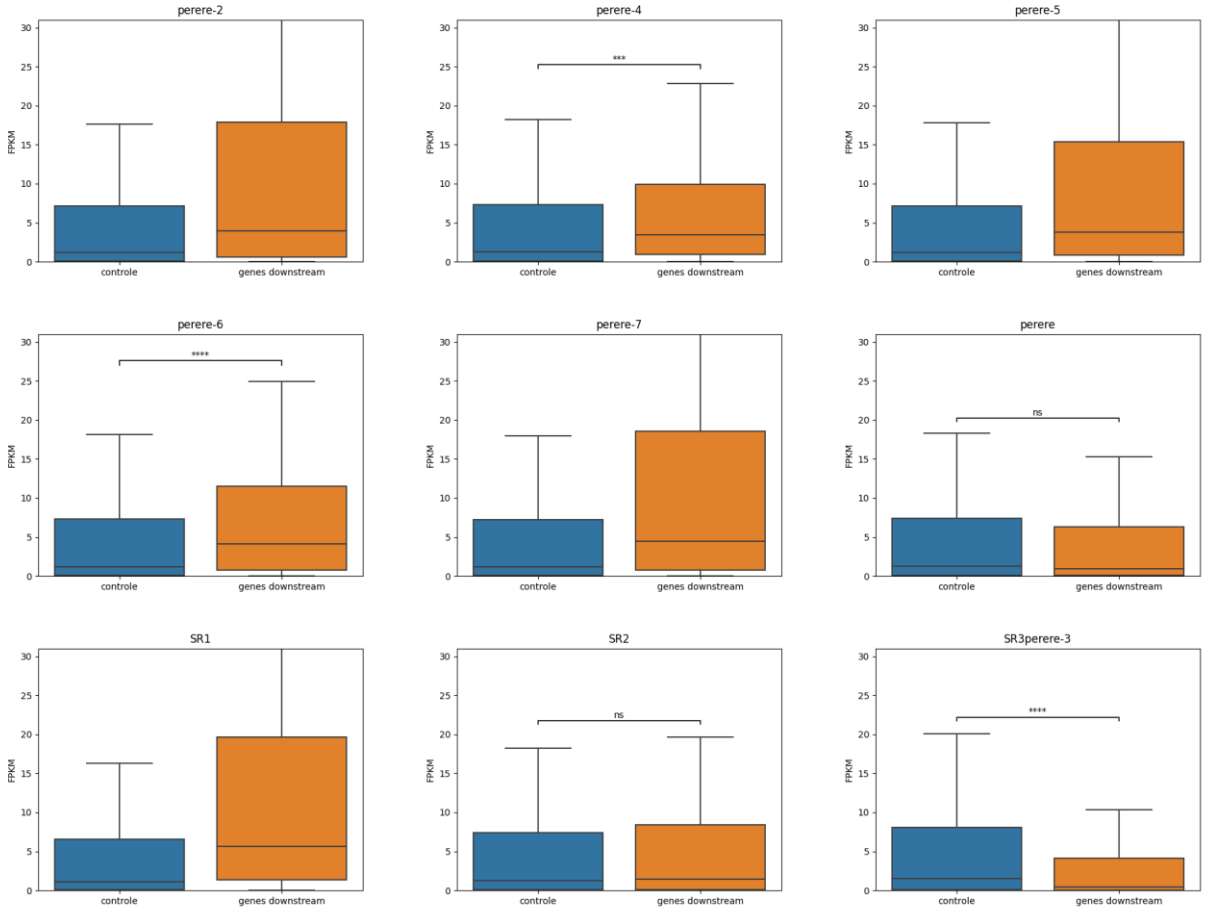

Figura A.2 - Comparação entre a população de genes downstream das inserções (intergênicas e colineares) dos TEs avaliados e o restante da população de genes (controle) para a fase de vida esporocisto. ****: p-valor <= 10-4.

Fonte: Elaborada pelo autor.

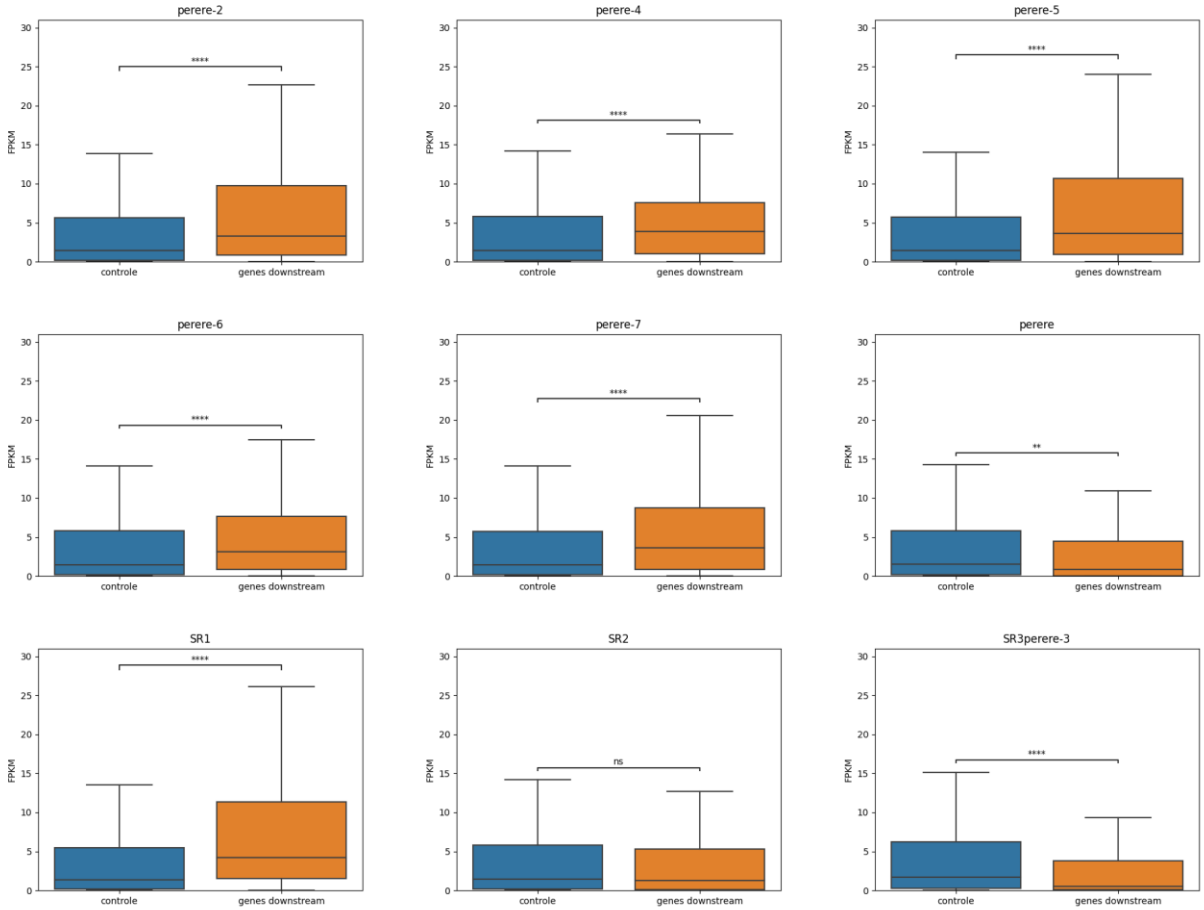

Figura A.3 - Comparação entre a população de genes downstream das inserções (intergênicas e colineares) dos TEs avaliados e o restante da população de genes (controle) para a fase de vida cercária 4h(1). ****: p-valor <= 10-4.

Fonte: Elaborada pelo autor.

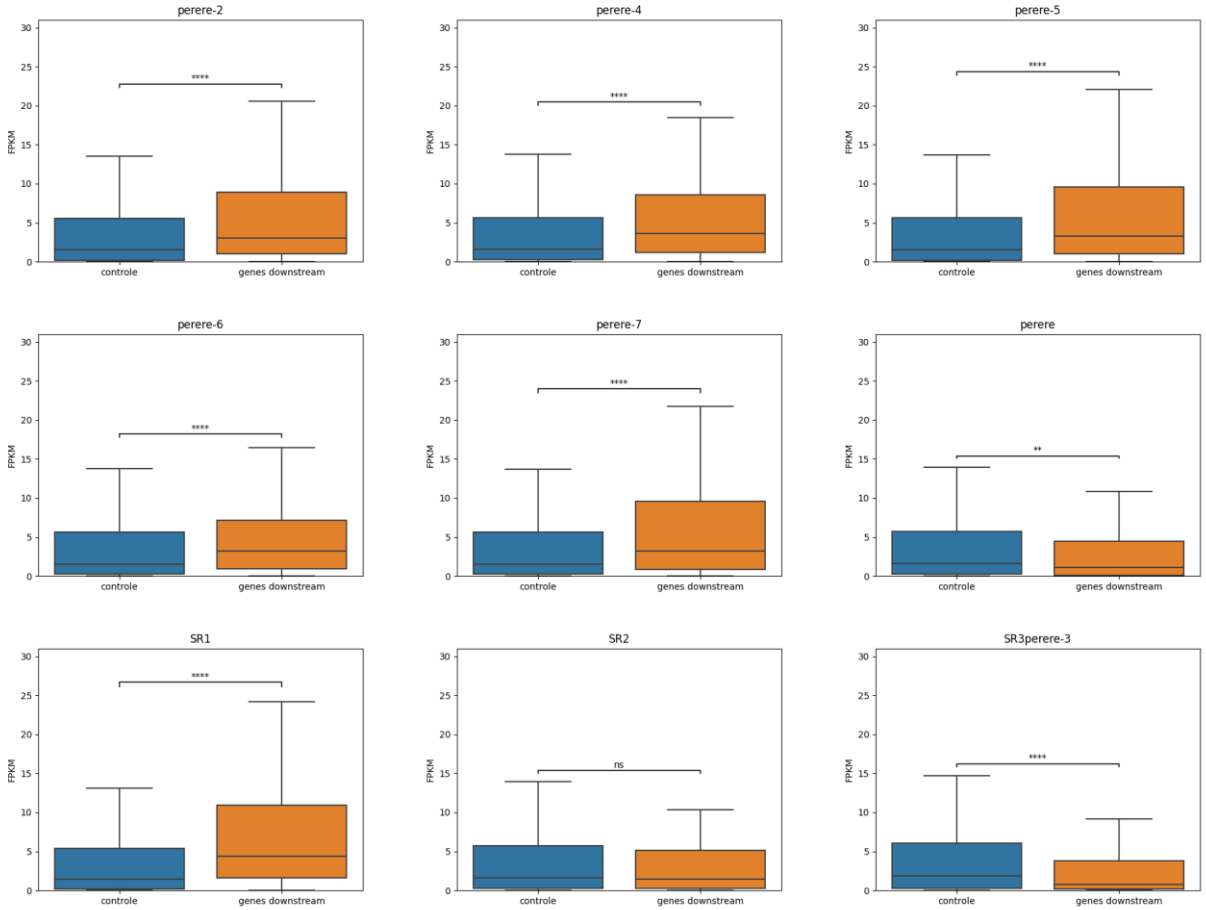

Figura A.4 - Comparação entre a população de genes downstream das inserções (intergênicas e colineares) dos TEs avaliados e o restante da população de genes (controle) para a fase de vida cercária 4h(2). ****: p-valor <= 10-4.

Fonte: Elaborada pelo autor.

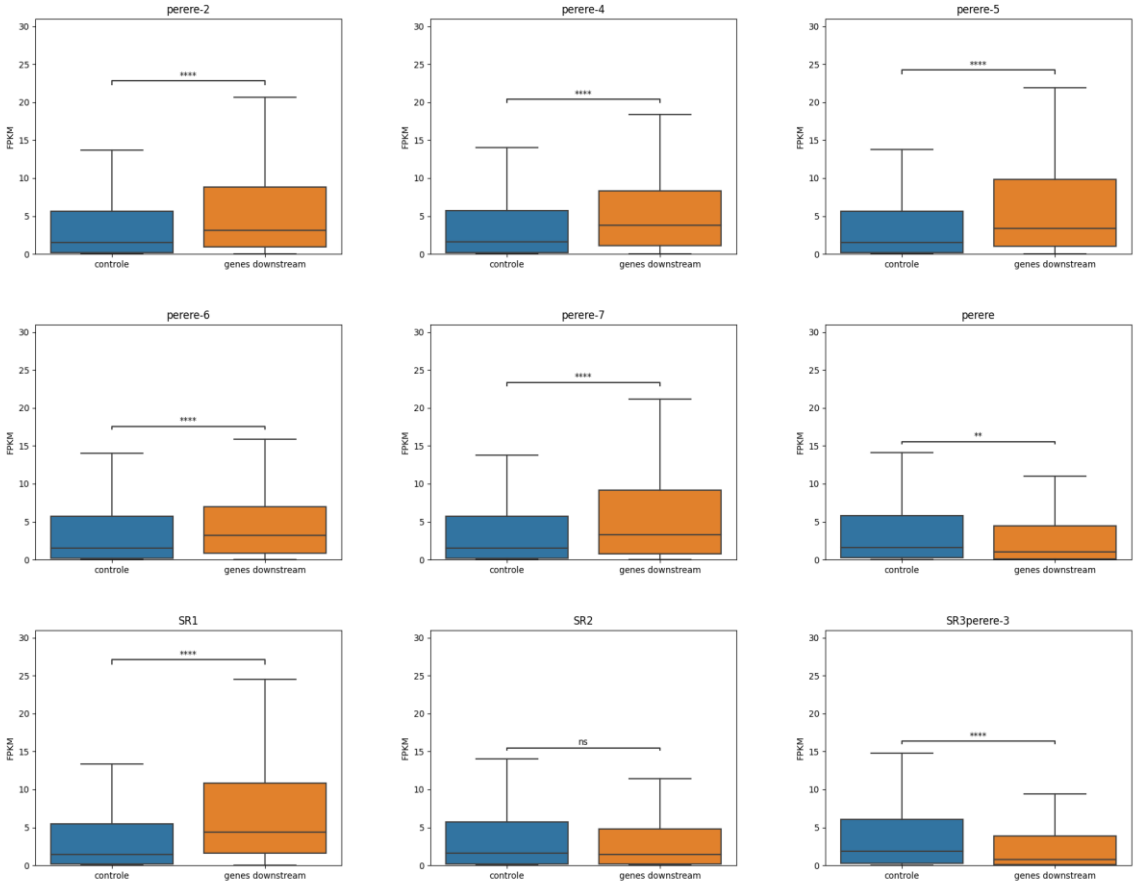

Figura A.5 - Comparação entre a população de genes downstream das inserções (intergênicas e colineares) dos TEs avaliados e o restante da população de genes (controle) para a fase de vida cercária 4h(3). ****: p-valor <= 10-4.

Fonte: Elaborada pelo autor.

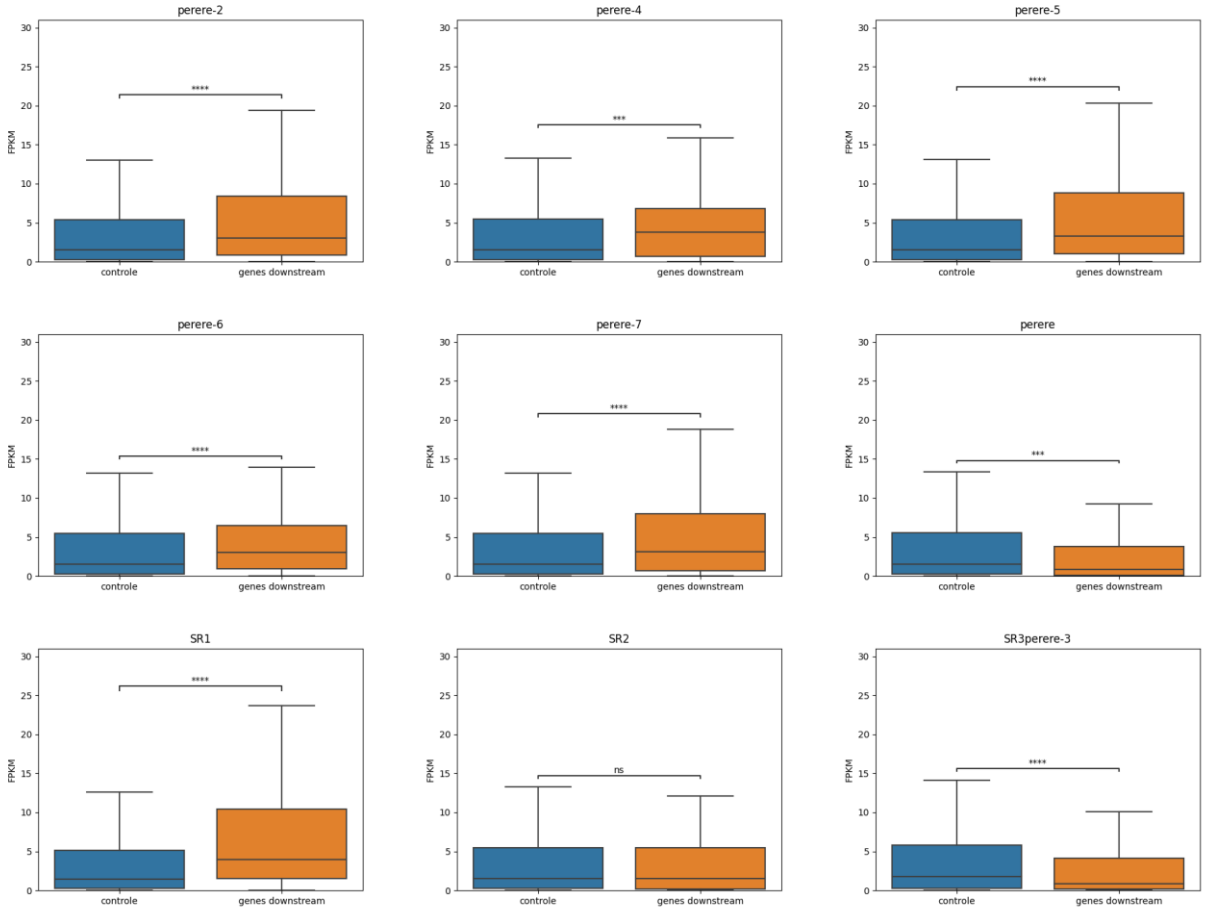

Figura A.6 - Comparação entre a população de genes downstream das inserções (intergênicas e colineares) dos TEs avaliados e o restante da população de genes (controle) para a fase de vida esquistossômulo 3h(1). ****: p-valor <= 10-4.

Fonte: Elaborada pelo autor.

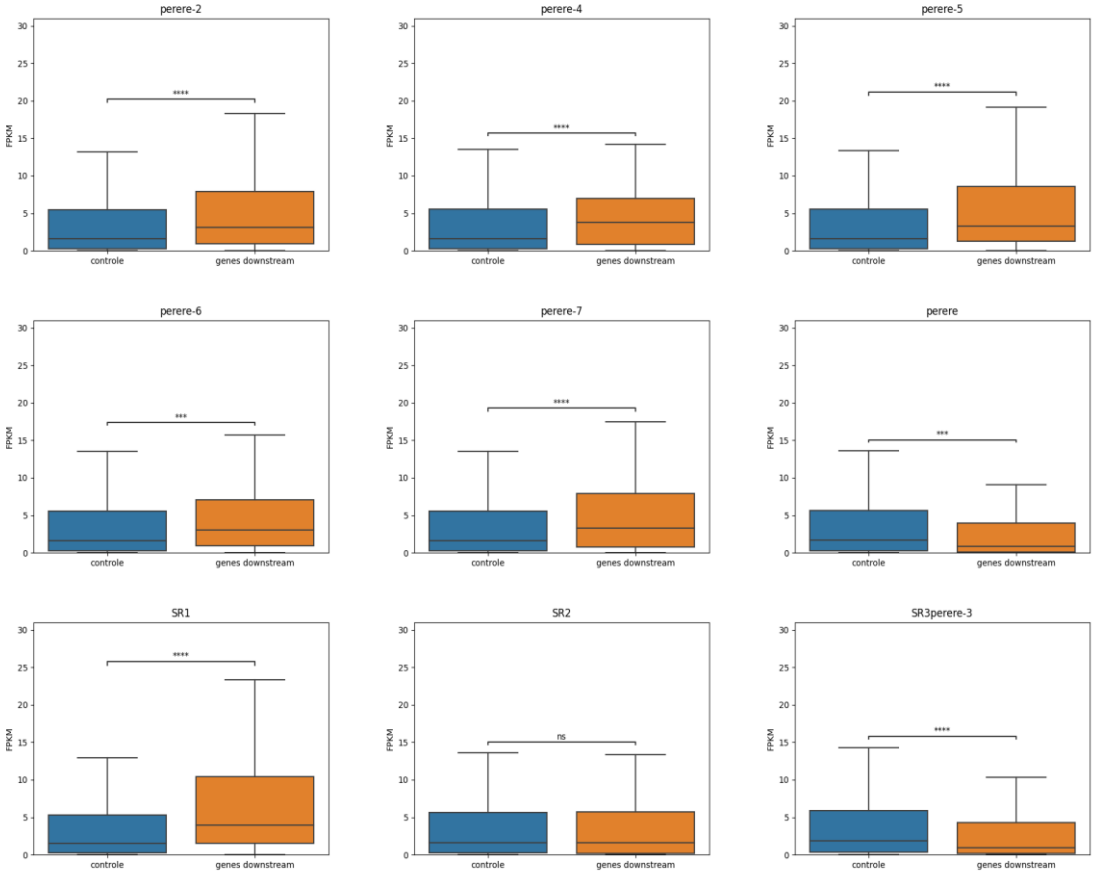

Figura A.7 - Comparação entre a população de genes downstream das inserções (intergênicas e colineares) dos TEs avaliados e o restante da população de genes (controle) para a fase de vida esquistossômulo 3h(2). ****: p-valor <= 10-4.

Fonte: Elaborada pelo autor.

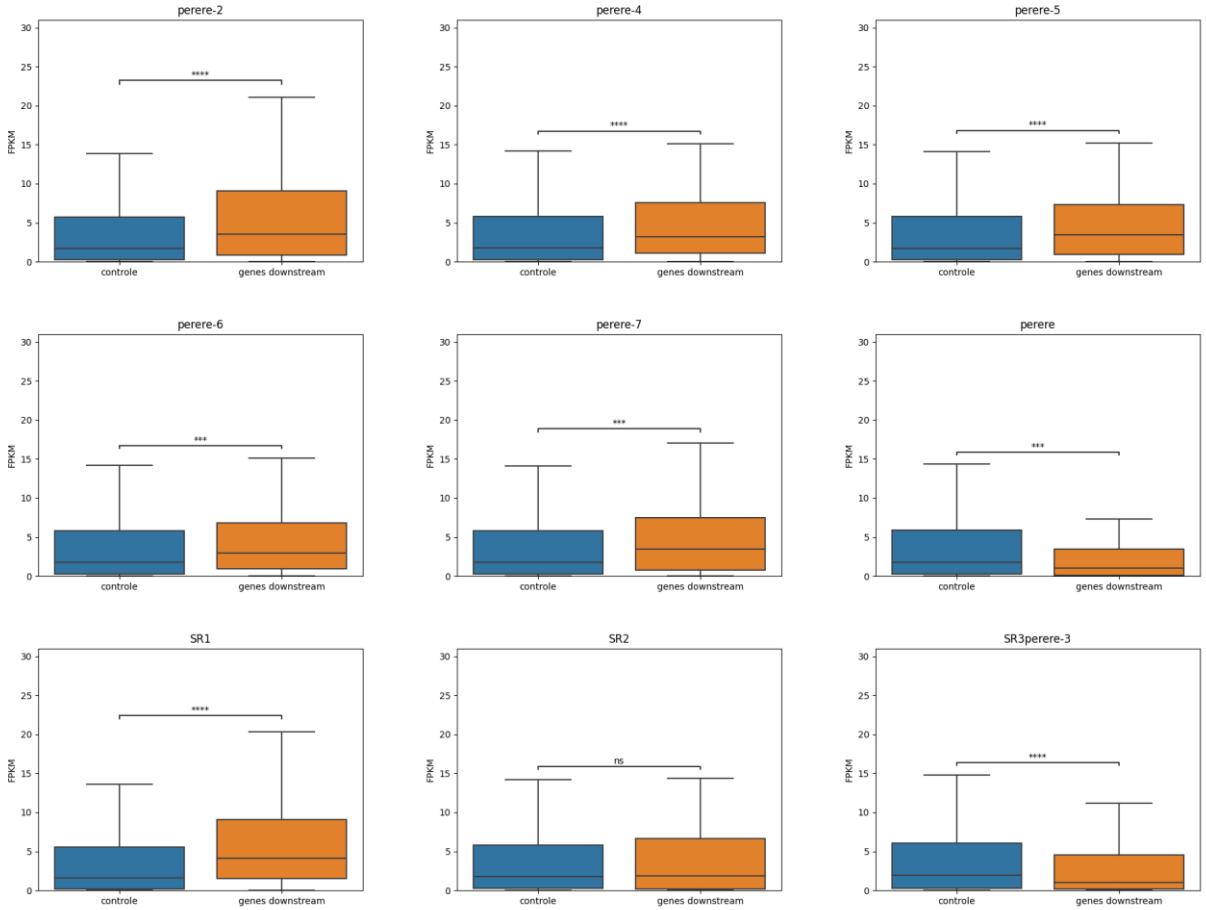

Figura A.8 - Comparação entre a população de genes downstream das inserções (intergênicas e colineares) dos TEs avaliados e o restante da população de genes (controle) para a fase de vida esquistossômulo 3h(3). ****: p-valor <= 10-4.

Fonte: Elaborada pelo autor.

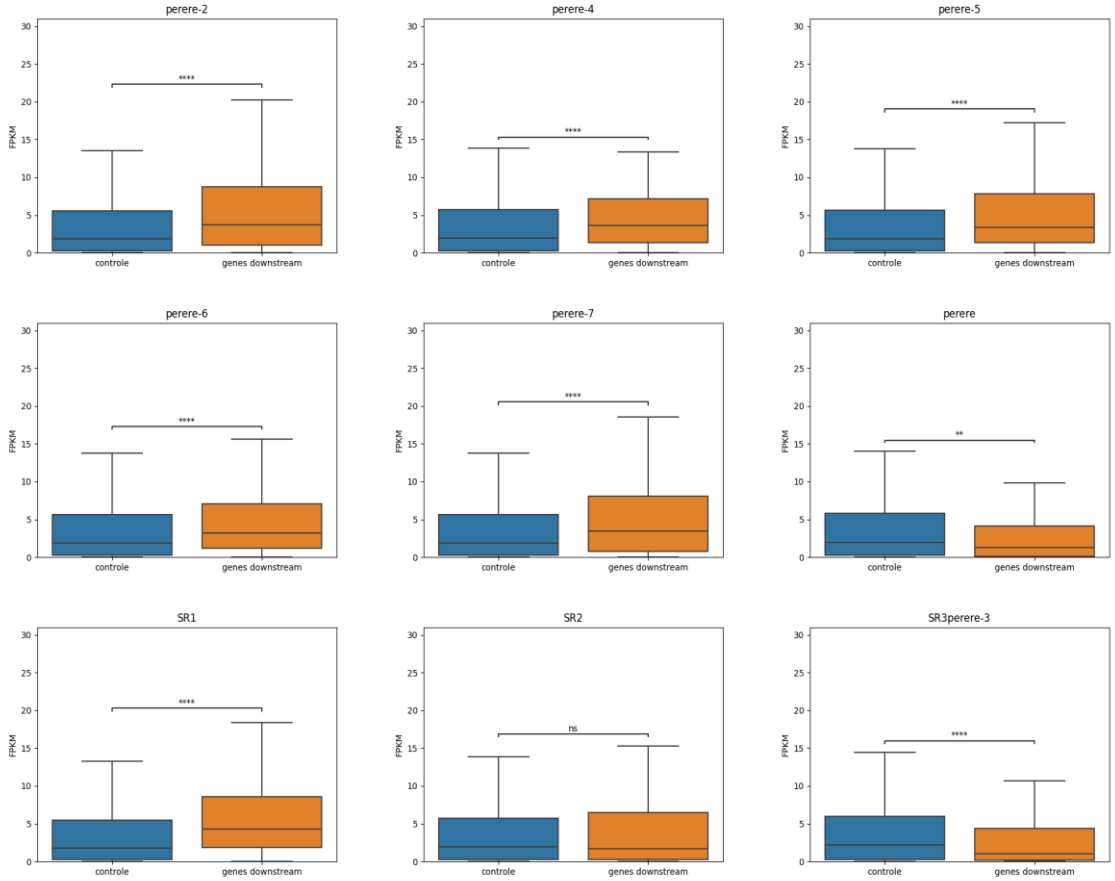

Figura A.9 - Comparação entre a população de genes downstream das inserções (intergênicas e colineares) dos TEs avaliados e o restante da população de genes (controle) para a fase de vida esquistossômulo 24h(1). ****: p-valor <= 10-4.

Fonte: Elaborada pelo autor.

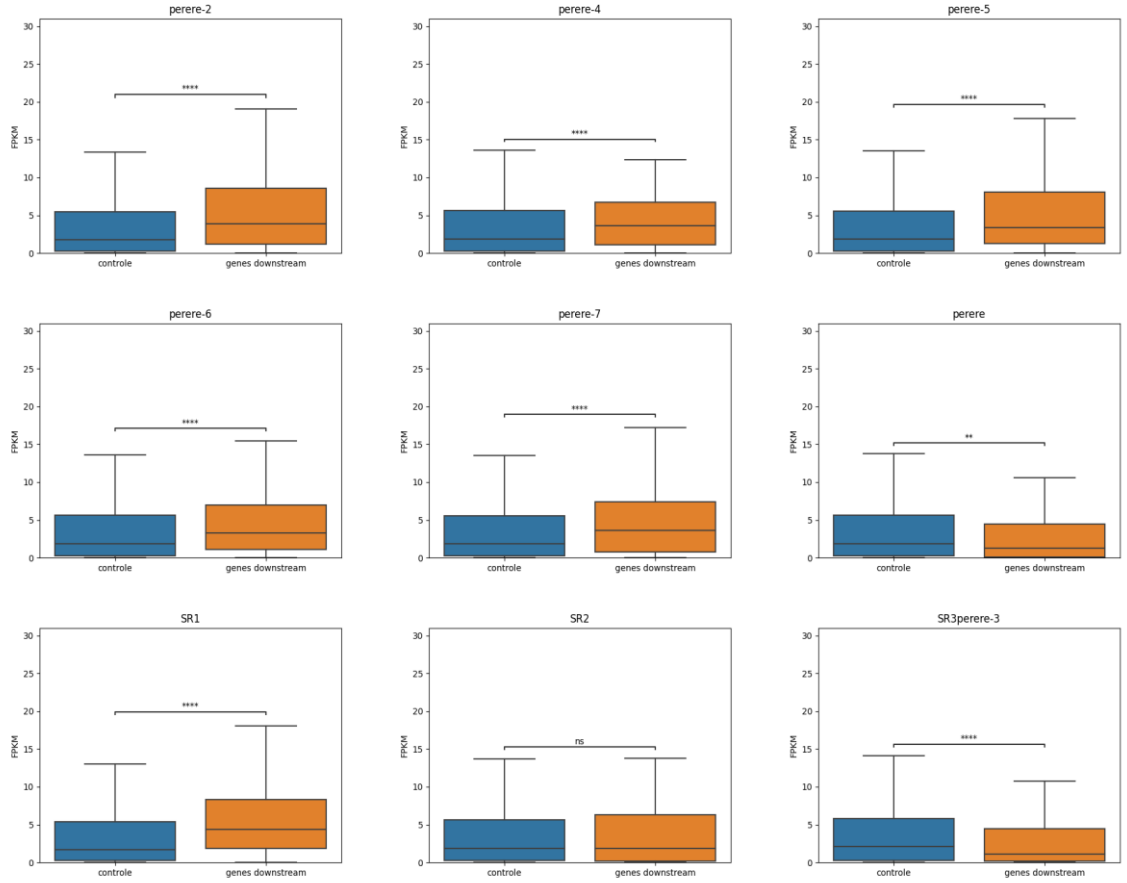

Figura A.10 - Comparação entre a população de genes downstream das inserções (intergênicas e colineares) dos TEs avaliados e o restante da população de genes (controle) para a fase de vida esquistossômulo 24h(2). ****: p-valor <= 10-4.

Fonte: Elaborada pelo autor.

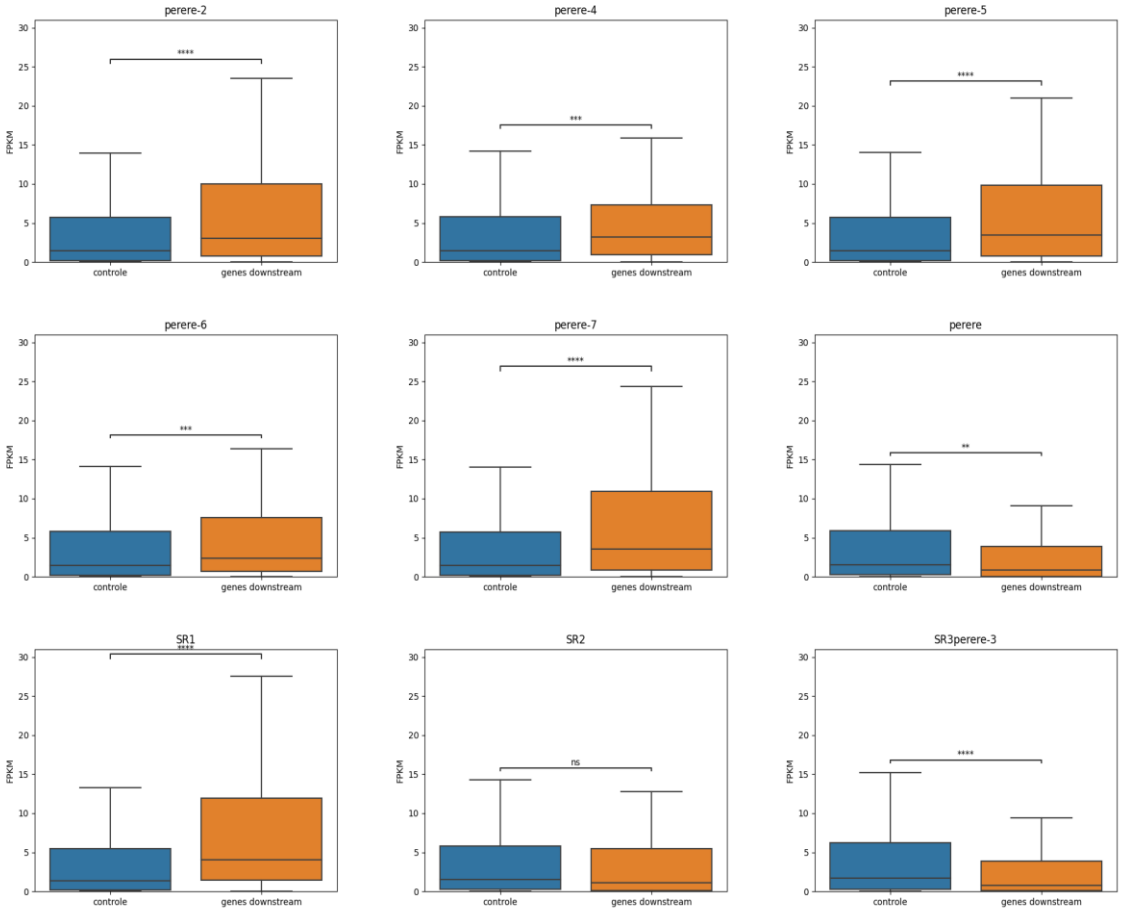

Figura A.11 - Comparação entre a população de genes downstream das inserções (intergênicas e colineares) dos TEs avaliados e o restante da população de genes (controle) para a fase de vida adulto 7 semanas. ****: p-valor <= 10-4.

Fonte: Elaborada pelo autor.

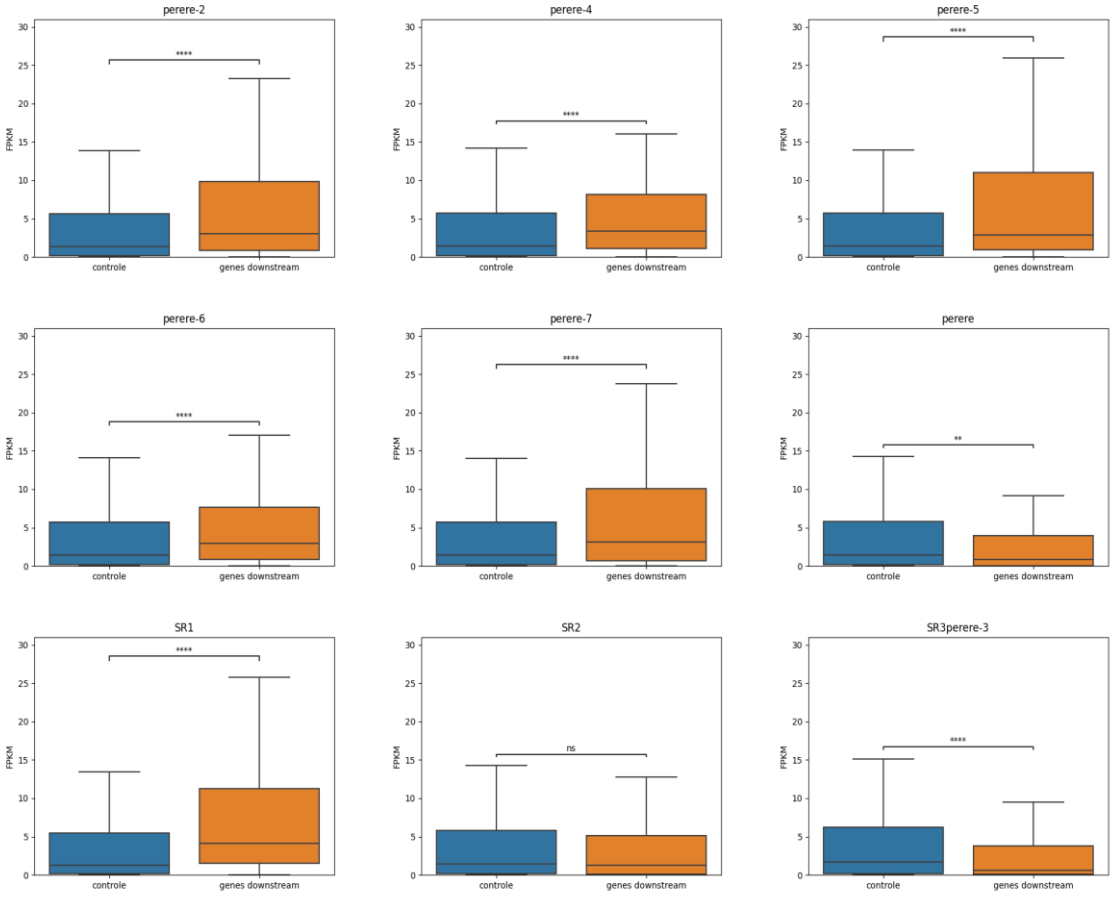

Figura A.12 - Comparação entre a população de genes downstream das inserções (intergênicas e colineares) dos TEs avaliados e o restante da população de genes (controle) para a fase de vida cauda. ****: p-valor <= 10-4.

Fonte: Elaborada pelo autor.

APÊNDICE B - Análise da prevalência das inserções de TEs upstream a regiões de interação com histonas modificadas.

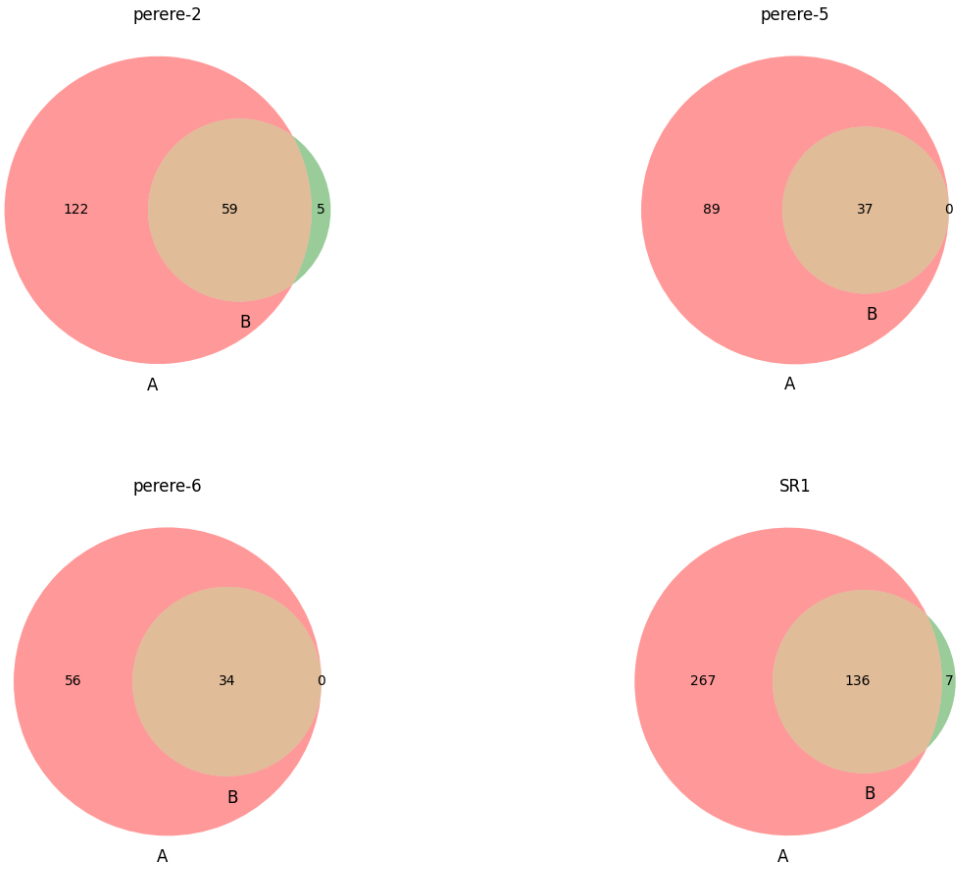

Figura B.1 - Diagramas de Venn das regiões compartilhadas entre as histonas com modificação H3K4me3(A) e H3K9Ac(B), em adulto, para os TEs com domínio PHD.
Fonte: Elaborada pelo autor.

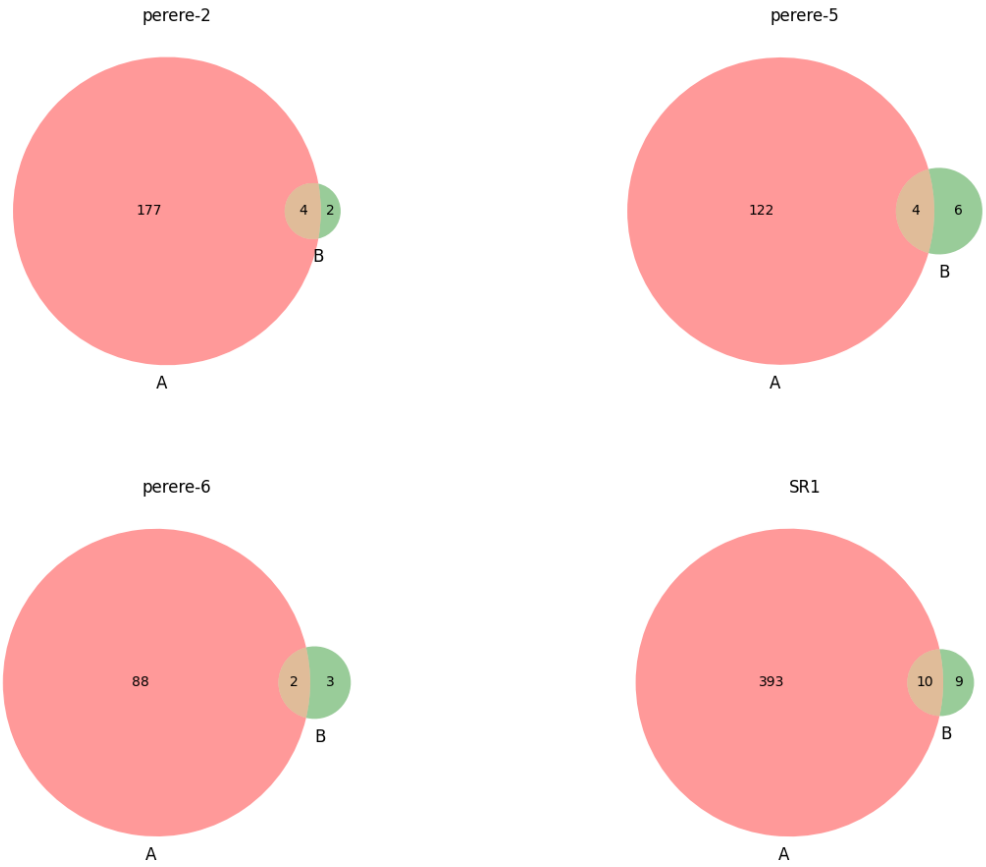

Figura B.2 - Diagramas de Venn das regiões compartilhadas entre as histonas com modificação H3K4me3(A) e H3K9me3(B), em adulto, para os TEs com domínio PHD.
Fonte: Elaborada pelo autor.

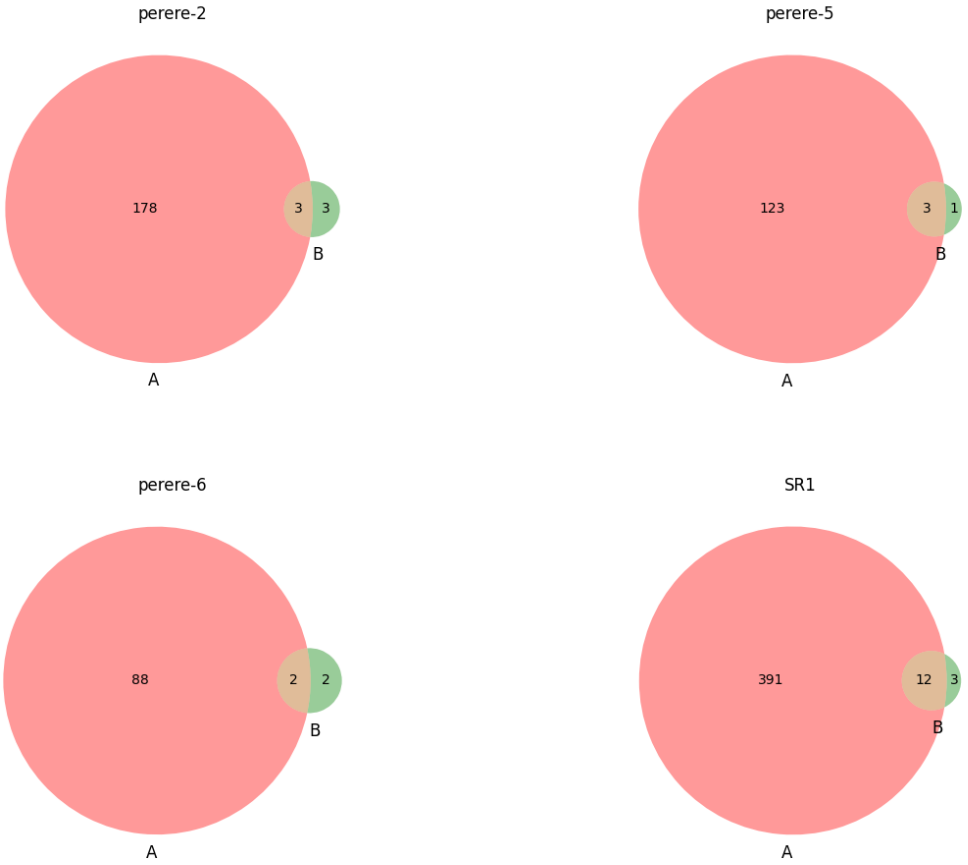

Figura B.3 - Diagramas de Venn das regiões compartilhadas entre as histonas com modificação H3K4me3(A) e H3K27me3(B), em adulto, para os TEs com domínio PHD.
Fonte: Elaborada pelo autor.

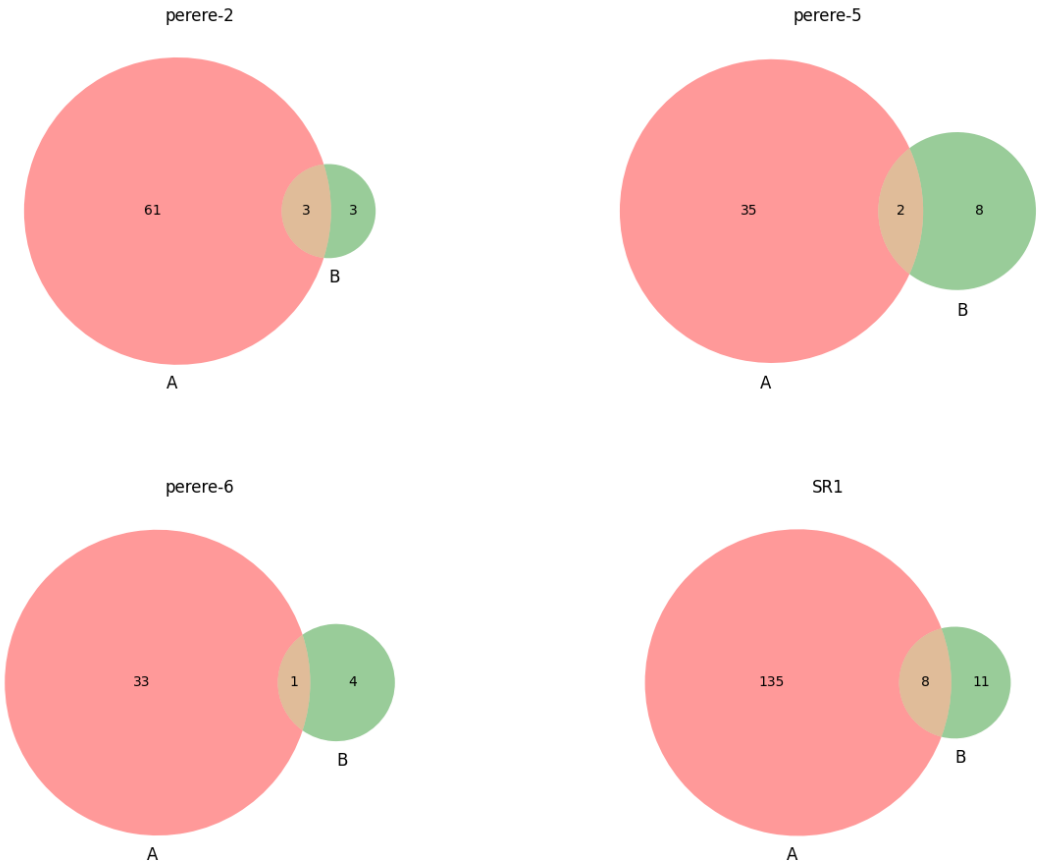

Figura B.4 - Diagramas de Venn das regiões compartilhadas entre as histonas com modificação H3K9Ac(A) e H3K9me3(B), em adulto, para os TEs com domínio PHD.
Fonte: Elaborada pelo autor.

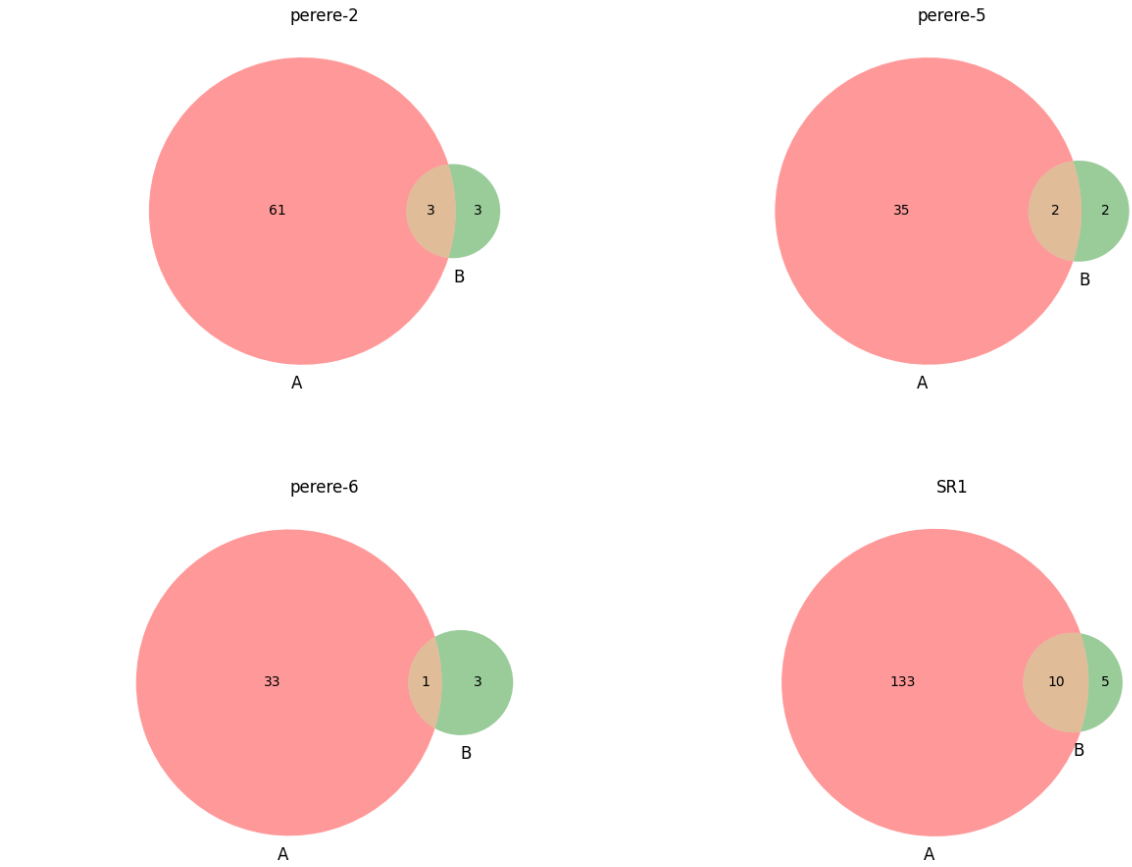

Figura B.5 - Diagramas de Venn das regiões compartilhadas entre as histonas com modificação H3K9Ac(A) e H3K27me3(B), em adulto, para os TEs com domínio PHD.
Fonte: Elaborada pelo autor.

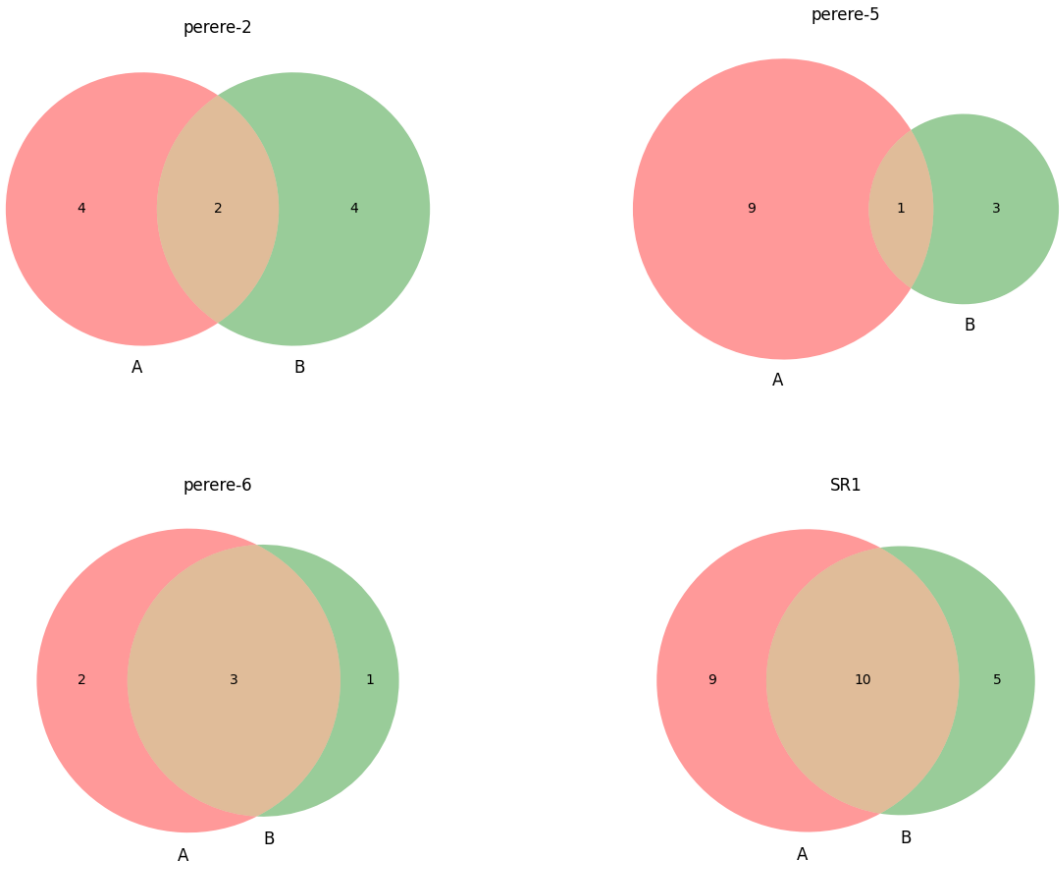

Figura B.6 - Diagramas de Venn das regiões compartilhadas entre as histonas com modificação H3K4me3(A) e H3K27me3(B), em adulto, para os TEs com domínio PHD.
Fonte: Elaborada pelo autor.

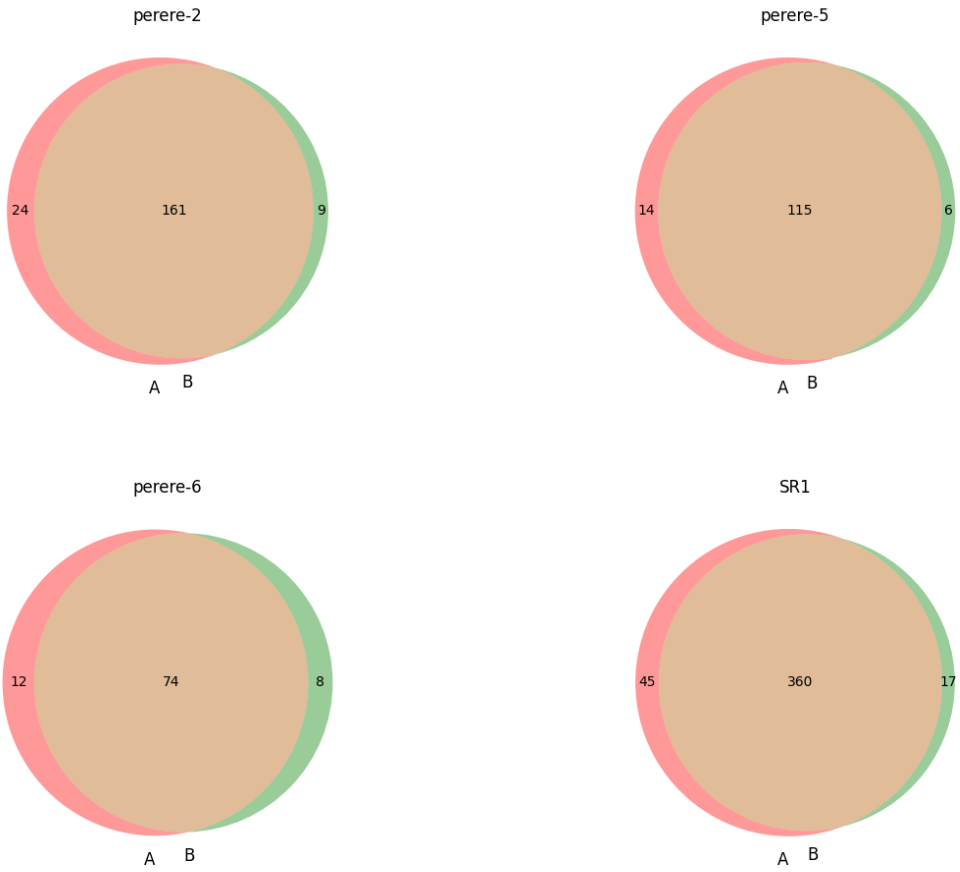

Figura B.7: Diagramas de Venn das regiões compartilhadas entre as histonas com modificação H3K4me3(A) e H3K9Ac(B), em cercária, para os TEs com domínio PHD.
Fonte: Elaborada pelo autor.

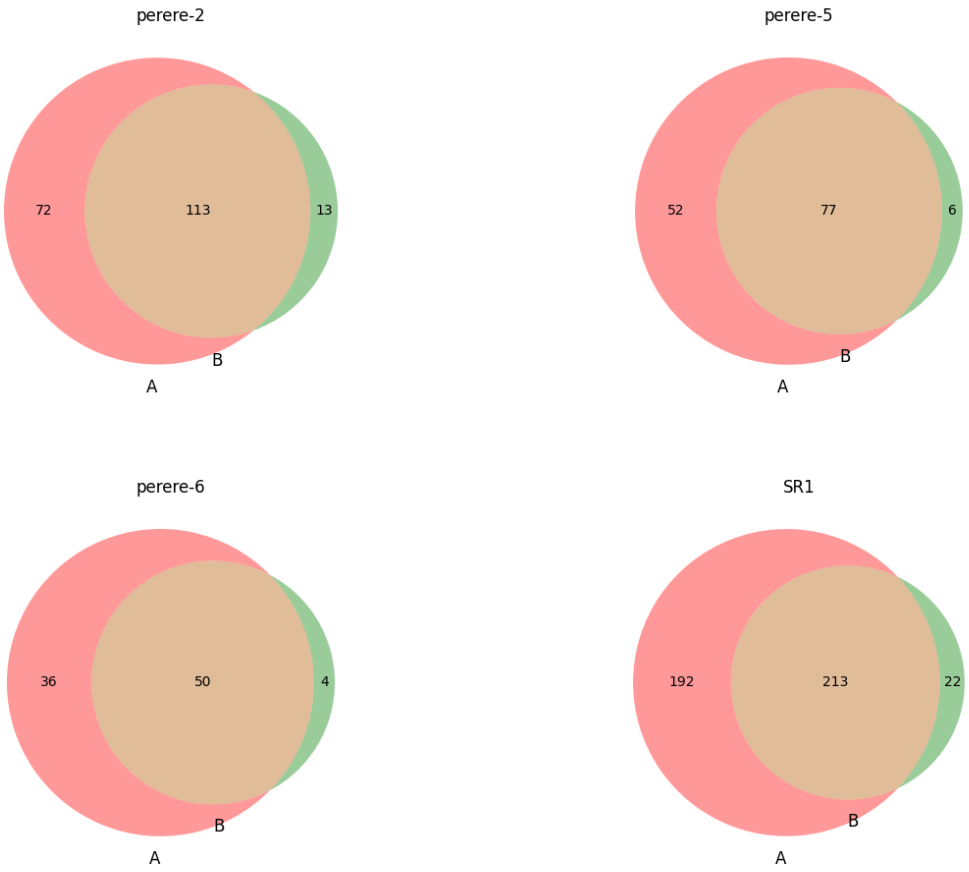

Figura B.8: Diagramas de Venn das regiões compartilhadas entre as histonas com modificação H3K4me3(A) e H3K9me3(B), em cercária, para os TEs com domínio PHD.
Fonte: Elaborada pelo autor.

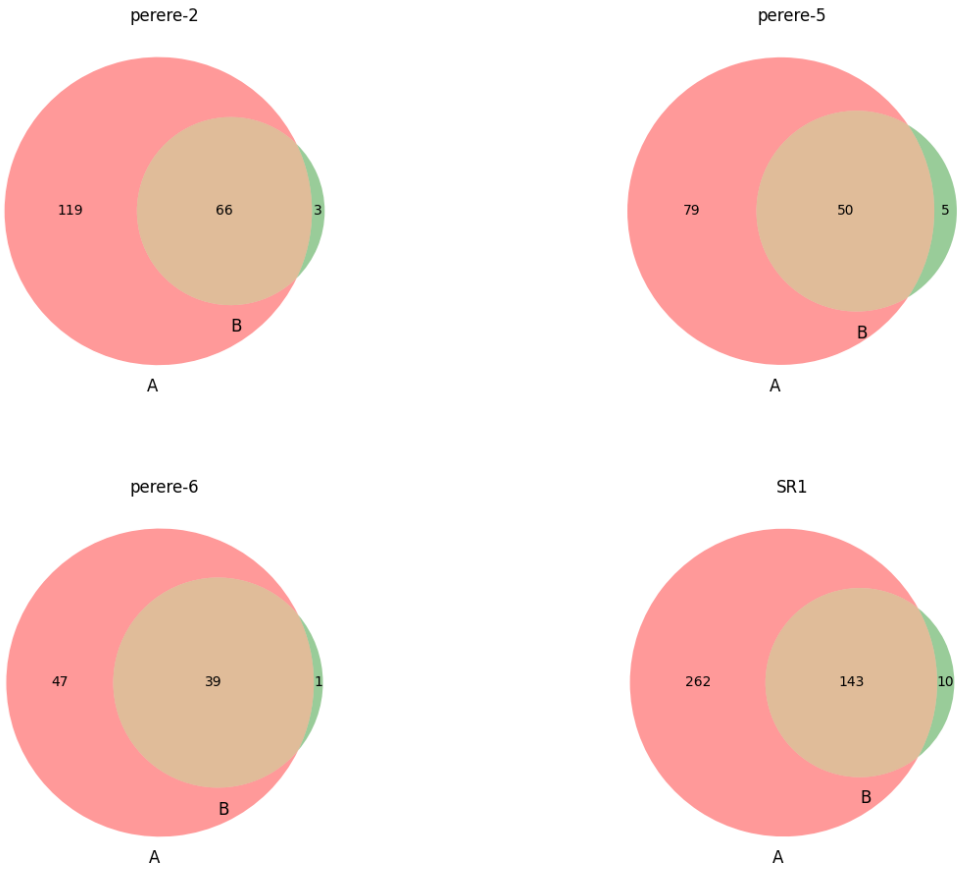

Figura B.9: Diagramas de Venn das regiões compartilhadas entre as histonas com modificação H3K4me3(A) e H3K27me3(B), em cercária, para os TEs com domínio PHD.
Fonte: Elaborada pelo autor.

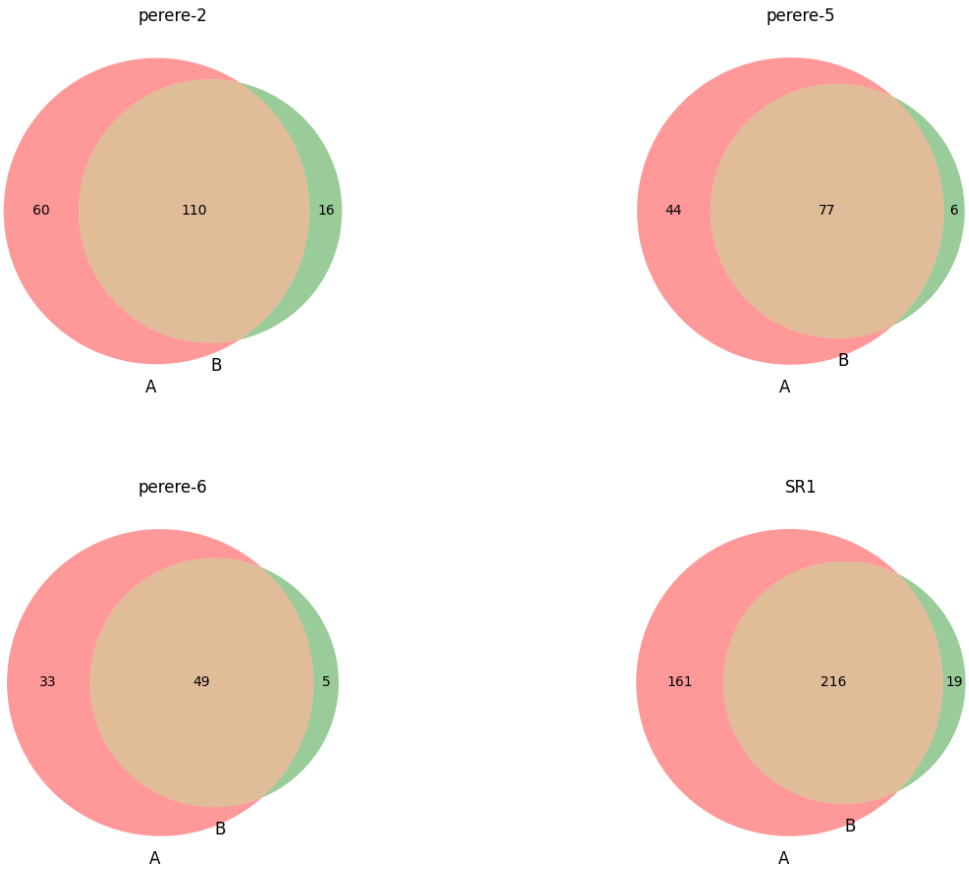

Figura B.10: Diagramas de Venn das regiões compartilhadas entre as histonas com modificação H3K9Ac(A) e H3K9me3(B), em cercária, para os TEs com domínio PHD.
Fonte: Elaborada pelo autor.

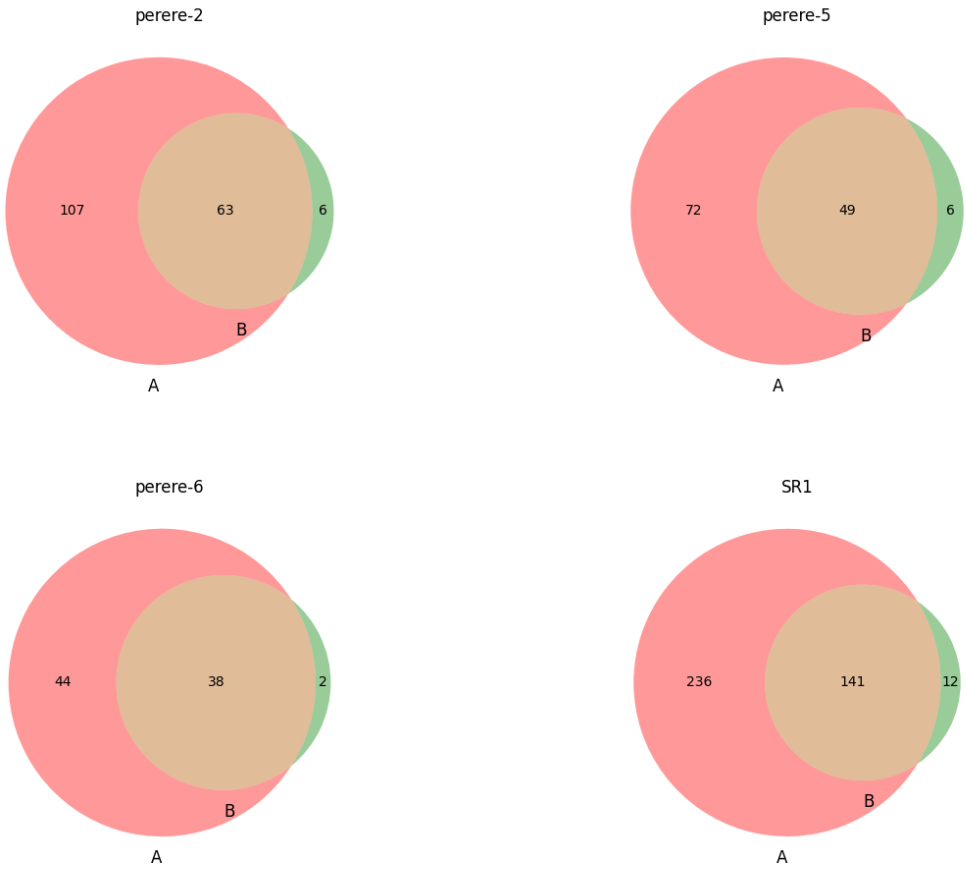

Figura B.11: Diagramas de Venn das regiões compartilhadas entre as histonas com modificação H3K9Ac(A) e H3K27me3(B), em cercária, para os TEs com domínio PHD.
Fonte: Elaborada pelo autor.

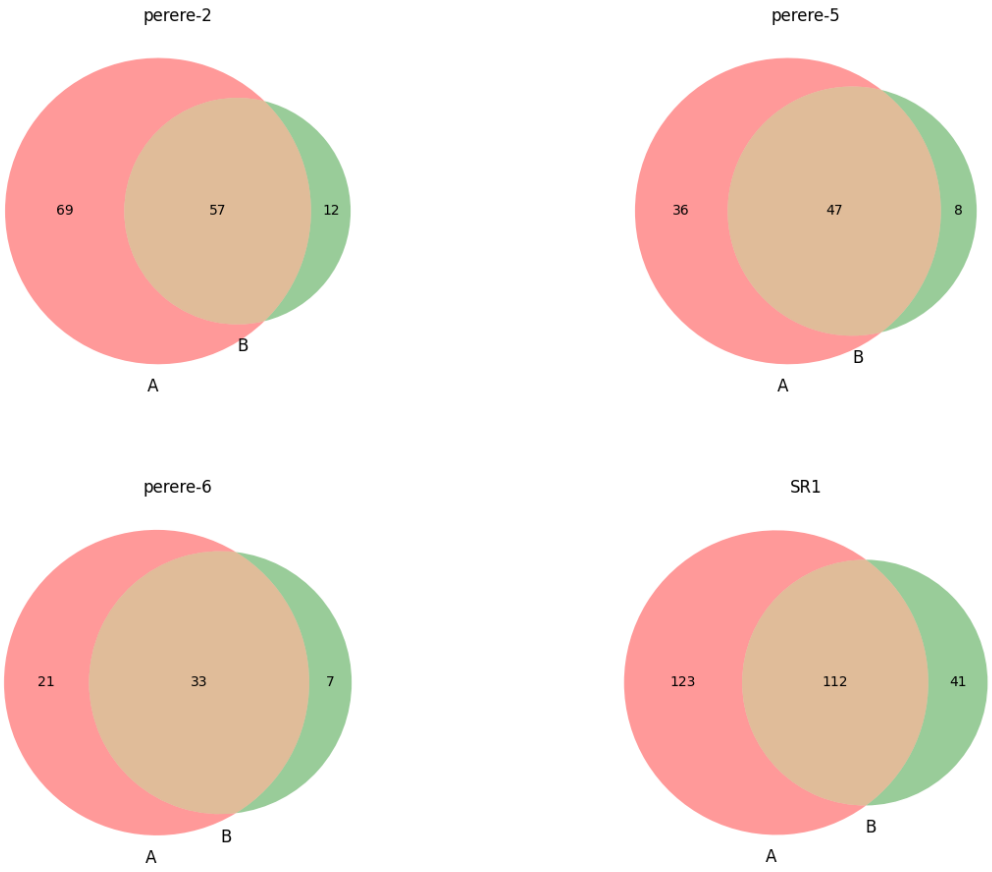

Figura B.12: Diagramas de Venn das regiões compartilhadas entre as histonas com modificação H3K9me3(A) e H3K27me3(B), em cercária, para os TEs com domínio PHD.
Fonte: Elaborada pelo autor.