Análise bioinformática das cópias do transposon não-LTR Pererê-3 no genoma de Schistosoma mansoni

Pedro de Carvalho Braga Ilídio Silva

Novembro de 2019

1 Introdução

(... Fundamentos biológicos...)
Tal propriedade nos fornece uma interessante ferramenta: as cópias de Pererê-3 observadas no genoma podem ser distinguidas umas das outras, visto que possuem uma região característica a cada uma delas. Desta forma, novas questões podem ser levantadas, e abre-se um novo campo de exploração.

É possível, por exemplo, que se avalie os níveis de transcrição de cada cópia separadamente, além de que, se cada réplica do transposon carrega o mesmo promotor, as diferenças destes níveis devem se originar de fatores externos à sequência do transposon em si, a exemplo da compactação da cromatina ou influência de regiões codificantes próximas.

Procura-se, portanto, caracterizar parâmetros como os níveis de transcrição ou a integridade das cópias como um todo, bem como os efeitos de genes próximos e da região em que se encontra cada réplica de pererê-3 em seus níveis de transcrição, fazendo uso de técnicas de bioinformática.

2 Metodologia

Inicialmente, obteve-se a região conservada do transposon a cada replicação, a partir do que foi exposto em (CITAR). Efetuou-se busca por regiões semelhantes no genoma do *S. mansoni*[1] por meio de alinhamento BLASTn, no intuito de encontrar todas as cópias do Pererê-3 inseridas no material genético do platelminto.

Já sabia-se de antemão que um outro transposon, denominado SR-3, possuia semelhança o suficiente com o transposon de nosso interesse para fazer com que o Pererê-3 alinhasse também com as cópias de SR-3 no processo de busca do longo do genoma. Toma-se então o cuidado de repetir o procedimento descrito antes com a região do SR-3 análoga à região do Pererê-3 com que estamos trabalhando, e prosseguir apenas com os alinhamentos que obtiveram maior score quando realizados com o Pererê e, portanto, apresentam maior probabilidade de serem de fato uma cópia deste último.

[1] referência à WormBase

Parte-se, então, à obtenção das sequências pertencentes às cópias mas diferentes entre elas, a "impressão digital" de cada cópia, caracterísca do Pererê-3 antes mencionada. Visto que a determinação do comprimento exato dessas regiões mostrou-se complicada, tomou-se o comprimento heurístico de 1000 pares de base à imediata jusante de cada alinhamento encontrado com o BLASTn no procedimento anterior. A essas regiões, específicas de cada cópia, dá-se o nome de sequências "head"[2]. Contudo, em muitos casos nota-se a existência de uma pequena região com a sequência GTAA-repetitiva entre cada head e a região constante de sua cópia-mãe, a região comum a todas as cópias, de forma que consideramos como head as 1000 bases posteriores às repetições[3].

No decorrer do projeto, foram também usados dados de transcriptoma e anotações gênicas, com ajuda dos quais criou-se quatro parâmetros a serem comparados e caracterizados entre as diferentes cópias do transposon Pererê-3, descritos a seguir.

2.1 Contagem de reads

Nas questões relacionadas aos níveis de transcrição, faz-se uso das bibliotecas de RNA-Seq (...) que compreendem diferentes estágios de vida do *S. mansoni* e conta-se, nas regiões de interesse do genoma, quantos reads dessas bibliotecas podem ser ali pareados. Para o pareamento, utilizou-se a ferramenta HiSat2 [citação] enquanto que na contagem de reads alinhados por região empregou-se o software HTSeq [citação]. Divide-se posteriormente o valor final da contagem pelo comprimento da região em que se procura o alinhamento dos reads, pois assume-se que cada par de base da região tem igual probabilidade de alinhar-se com cada read e, portanto, a contagem final depende linearmente do comprimento do intervalo de DNA que se analisa. Divide-se o mesmo valor também pela contagem total de reads de cada biblioteca, pois uma base de dados com mais reads naturalmente geraria maior contagem por região. As divisões são feitas, então, a fim de eliminar esses fatores para fins comparativos, e ao valor resultante dá-se o nome de coeficiente de transcriçao.[4]

2.2 Comprimento-mãe

Com curiosidade sobre os efeitos da integridade de cada cópia, chama-se de comprimento-mãe o comprimento do alinhamento BLAST que gerou cada sequência head. Toma-se cuidado para considerar apenas os alinhamentos que contém a extremidade 3' da cópia de referência, de forma a ser provável a presença da sequência head na região pós-3' no genoma.[5]

[2] Talvez venha a convir uma mudança de nome.

[3] Discursar mais sobre a região repetitiva?

[4] Já existe?

[5] Assume-se que este valor esteja fortemente correlacionado com a integridade da cópia, e não são consideradas as mutações pontuais, componentes também importantes ao se pensar no grau de integridade de uma sequência.

2.3 Correlação com genes vizinhos

Espera-se que grande parcela das cópias encontradas no genoma sejam transcritas também por efeito de readthrough, ou transcrição passiva, em que as cópias do transposon são transcritas por se encontrarem em regiões já codificantes, no quadro de leitura aberta (ORF) de um gene, por exemplo, e não fazendo uso da maquinaria própria de transcrição. Um efeito semelhante ocorre se o transposon está inserido *downstream* ao gene, de forma a poder ser transcrito por ocasionais falhas do terminador genico. Supõe-se que esses casos resultem em coeficientes de transcrição independentes da integridade da cópia de Pererê-3, uma vez que as proteínas nela codificadas são dispensáveis para a presença da cópia no transcriptoma.

Portanto, avalia-se os coeficientes de correlação de Pearson entre a transcrição (contagem de *reads* normalizada) de cada cópia do transposon em cada biblioteca e a transcrição de seu gene mais próximo em cada uma (ou do gene em que se insere, para os casos em que a cópia se encontra na região expressa de um gene). A esta correlação será referida daqui em diante como correlação com o gene vizinho, e espera-se que seja maior no casos em que a transcrição da sequência head é passiva.

2.4 Distância ao gene vizinho

Ainda sob a questão da influência dos genes mais próximos no coeficiente de transcrição das sequências head, outro fator a se considerar é a distância em si entre a head e o gene vizinho, em pares de base, pois espera-se que a correlação entre eles se intensifique quanto menor for tal distância. Avaliando os casos em que a head está *downstream* ao gene separadamente, espera-se também determinar se há atividade do promotor gênico na transcrição da cópia do transposon, isto é, verificar a ocorrência de transcrição por readthrough ou vazamento.

3 Resultados

3.1 As cópias são de fato diferentes?

Visando constatar a unicidade das cópias de transposon pelo genoma, efetuou-se alinhamentos BLAST das sequências head entre elas mesmas, e verificou-se que aproximadamente 87,6% das sequências não foram alinhadas com nenhuma outra e por volta de 96,0% alinharam no máximo uma vez, mostrando que, de fato, não há coincidências gerais claras entre entre as regiões UTR 3' do Pererê-3, e a temática de análise proposta mantém-se válida.

3.2 As cópias se transcrevem de forma igual pelo genoma?

Avaliando simplesmente a distribuição dos valores de coeficiente de transcrição, observa-se grande predominância de medidas pequenas, com aproximadamente $88,496\%$ de cópias com seu coeficiente menor que 5 (unidade), sendo que por volta de $15,648\%$ apresentaram nenhuma transcrição. A distribuição em escala logarítmica é apresentada na figura 3.2.

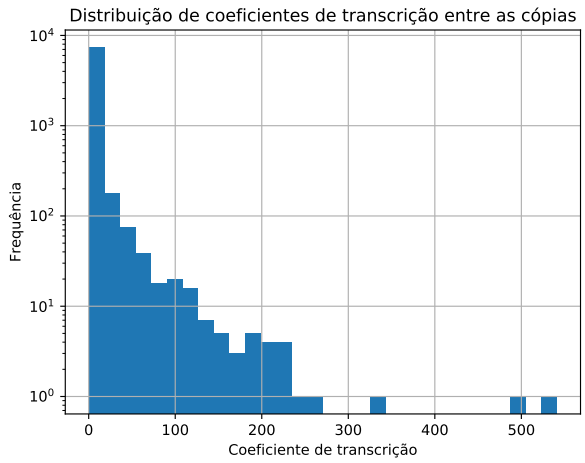

Figure 1: Histograma de coeficientes de transcrição para as cópias de Pererê-3.

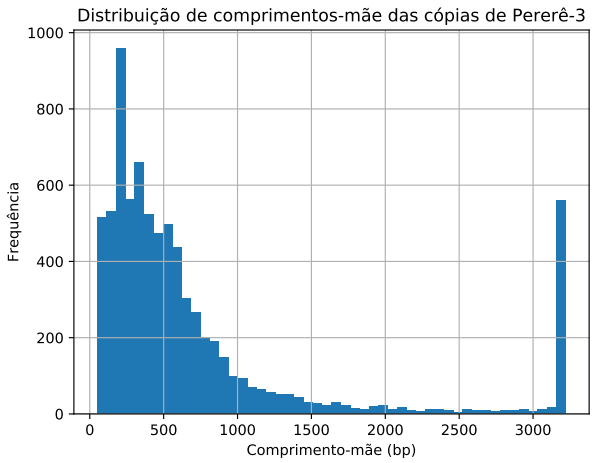

Figure 2: Histograma de comprimentos da região invariante (comprimentos-mãe) das cópias de Pererê-3.

Conclui-se assim, que, muito provavelmente, uma grande parcela (por volta de ao menos 15%) das cópias de Pererê-3 inseridas no genoma de *S. mansoni* se encontram inativas, não sendo capazes de se transpor autonomamente.

3.3 A integridade afeta os níveis de transcrição?

A fim de investigar a influência da completude das cópias em sua capacidade de produzir novos elementos, inicialmente observou-se a distribuição de comprimentos-mãe encontrados.

Já nesse ponto, há uma clara predominância das cópias menores, principalmente com menos de 600 pares de base de comprimento, representadas mais à esquerda na figura 3.3. Contudo, é interessante observar também um aumento significativo da quantidade de cópias íntegras ou quase íntegras, na extrema direita da figura 3.3. Esse aumento sugere fortemente que a integridade das cópias é um fator determinante para sua disseminação (ou ao menos sobrevivência) e corrobora a hipótese de que há grande efeito de readtrough: já previa-se que cópias truncadas, de comprimento menor, teriam menor efeito no produto final de um gene, visto que menos aminoácidos seriam adicionados, e portanto seriam mais prováveis de se instalar em um gene mantendo-o ativo e funcional, de forma que a cópia poderia tirar proveito da maior conservatividade das regiões gênicas e predominar no genoma.

Embora a quatidade de dados seja muito

maior nas extremidades direita e esquerda do gráfico, como já apontava a figura 3.3, não há diferença significativa de nível de transcrição entre as duas regiões, o que pode sugerir que, embora incompletas, as cópias de menor comprimento-mãe ainda sim façam uso da maquinaria própria. essa hipótese é contudo desfavorecida ao se observar que as cópias encontradas possuem obrigatoriamente a extremidade 3', como foi pré-requisitado na etapa de escolha dos alinhamentos blast do pererê no genoma, pois isso implica que a região faltante nas cópias incompletas é justamente a 5', onde se esperaria encontrar o promotor de cada elemento.

Se aprofundando na questão da influência dos genes próximos, procura-se então observar separadamente as cópias sobrepostas ou não a algum gene, bem como *downstream* ou *upstream* a algum, a fim de esclarecer efeitos de readthrough ou vazamento.

3.4 Há influência de genes próximos na transcrição das cópias?

Procurou-se investigar o efeito que genes próximos exercem sobre a transcrição de uma cópia

3.5 Como a configuração relativa entre a head e seu gene vizinho afeta sua correlação de transcrição?

Nesta seção procura-se dividir a população de heads entre diferentes posições relativas ao seu gene mais próximo. Nota-se que, em alguns poucos casos, não havia nenhum gene no mesmo contig da head, de forma que essas situações tiveram de ser descartadas e restaram 7645 cópias em análise. Ao se calcular a correlação com o gene vizinho, houve muitos casos em que não foram alinhados sobre a head nenhum read de alguma determinada biblioteca, o que fazia com que a maioria dos pontos utilizados na determinação da correlação fossem nulos, e apenas um ponto que distasse significativamente dos demais era necessário para elevar a correlação a valores próximos a um, dado que a distribuição assumia características lineares. Foram então removidas dos dados as heads sobre as quais menos de quatro bibliotecas de reads, das oito bibliotecas utilizadas, geravam contagem de reads não nula. Isso impactou em perda considerável de dados: mativeram-se 3269 cópias na análise.

3.5.1 Cópias sobrepostas a genes sugerem forte correlação transcricional

Inicialmente, dividiu-se as cópias entre as que sobrepunham alguma região gênica e as que não. Tomou-se os valores de correlação de transcrição com o gene vizinho (ou o gene so-

[6] também conhecido como teste da soma dos postos de Wilcoxon.

breposto) e comparou-se a distribuição desses valores nas duas populações por meio de teste U de Mann-Whitney[6]. O p-valor obtido foi de aproximadamente $1,972 \cdot 10^{-36}$, sugerindo muito fortemente que de fato há diferença entre os níveis de transcrição dos dois grupos, como aponta a figura 3.5.1.

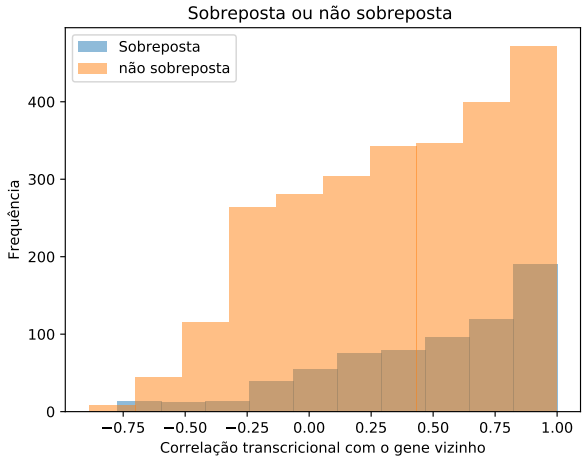

Figure 3: Comparação dos histogramas dos valores de correlação entre o coeficiente de transcrição de cada cópia de Pererê-3 e o de seu gene mais próximo, nos casos em que a cópia não sobrepõe algum gene (em laranja), ou com o gene no qual a cópia está inserida (em azul), se ela de fato estiver em algum.

Nota-se muitas correlações próximas de 1, vistas a direita da figura **??**, de forma que a transcrição das cópias é quase a mesma de seus genes próximos em uma grande quantidade de casos, o que suporta a ideia do *readthrough*.

[7]Locally Weighted Scatterplot Smoothing (ref)

3.5.2 Insuficiência de dados torna inconclusiva a diferença de correlação com vizinho entre cópias íntegras ou não

A fim de avaliar se a transcrição das heads já apresentava diferença apenas com a distância ao gene vizinho de cada cópia, graficou-se a correlação de transcrição com o gene vizinho em função da distância de cada cópia a ele, distinguindo-se cópias definidas como íntegras, com mais de 3150 pares de base de comprimento-mãe, de cópias incompletas, com menos de 750 pares de base. O comportamento geral dos dados foi modelado por curvas LOWESS[7] e o resultado é apresentado na figura 3.5.2.

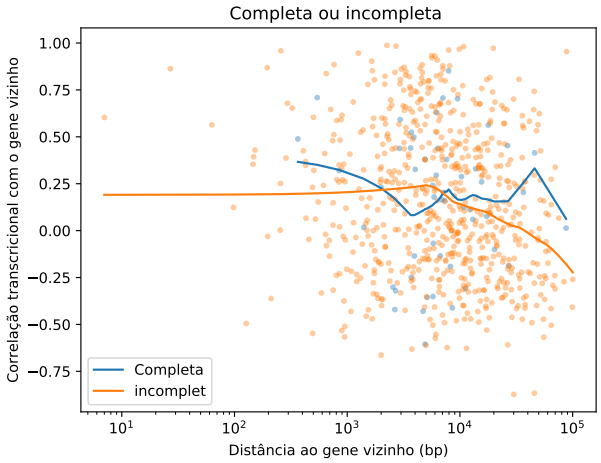

Figure 4: Correlação com o gene mais próximo de cada cópia em função da distância entre os dois, para as cópias ditas completas, com comprimento maior que 3150 bp, e ditas incompletas, cujo comprimento é inferior a 750.

O p-valor obtido na comparação destas duas populações pelo teste U de Mann-Whitney foi de 0.385, de forma que não se revela nenhuma informação sobre o sistema

analisado. Nota-se que, embora houvesse 7650 cópias identificadas no genoma de *S. mansoni*, apenas 55 cópias completas e com gene vizinho correlacionado puderam ser utilizadas, de forma que a pequena quantidade de dados pode se mostrar, em futuras análises com maior número de instâncias, o fator determinante da falta de informação.

3.5.3 Cópias *downstream* a seu gene vizinho podem de fato possuir maior correlação com ele.

O mesmo procedimento da subseção anterior foi executado separando-se das demais agora as cópias *downstream* a seu gene mais próximo, culminando-se na figura 3.5.3. Embora com dados ainda ruidosos, nota-se mais claramente uma tendência decrescente da coordenação com o gene vizinho conforme se aumenta a distância da cópia a ele.

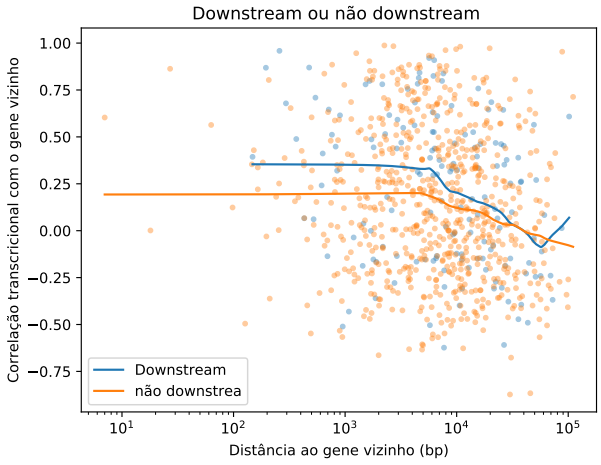

Figure 5: Correlação com o gene mais próximo de cada cópia em função da distância entre os dois, para as cópias *downstream* a seu gene mais próximo (em azul) e para o restante dos dados (em laranja).

Pela figura 3.5.3, as cópias *downstream* parecem ainda revelar visível superioridade da correlação com o vizinho no gráfico apresentado, dado que sua curva LOWESS permanece superior a das demais cópias principalmente na região com maior número de amostras (distância entre $5 \cdot 10^3$ e $2 \cdot 10^4$ pares de base), reforçando a hipótese de que há vazamento nas transcrições gênicas, ou seja, que o nível de transcrição das heads pode ser incrementado se ela estiver próxima à extremidade 3' de uma região codificante no genoma, provavelmente em decorrência de falha na terminação da transcrição.

Essas supesições, contudo, ainda tem caráter duvidoso frente à tamanha ruidez estatísca dos dados apresentados, caráter este evidenciado pelo teste U de Mann-Whitney com o resultado aproximado de 0.0318 para o p-valor. Ainda abaixo do limite heurístico de 0.05, de forma a poder ser considerada significativa a diferença entre as populações, esse resultado se beneficiaria de futuras análises mais profundas, possivelmente com mais bibliotecas de reads incluídas, de forma a aumentar a confiabilidade das correlações medidas.

3.5.4 A configuração que se esperava mais demonstrar efeitos de compactação de cromatina não se diferenciou na análise.

Traçou-se o mesmo caminho das últimas subseções, porém tomando à parte a parcela de cópias cujo sentido era diferente do de seu gene vizinho e cuja posição se dava após a terminação 3' desse último. Nessa configuração, dado que não poderia haver *readthrough* estando em fitas diferentes e os

promotores provavelmente estariam o mais próximos possível, teríamos o melhor caso para observar coordenações de transcrição decorrentes da descompactação de cromatina no local. O p-valor, contudo permanece alto demais para constatar significância estatística, com o valor arredondado de 0.1870, demonstrando que, de fato ocorrendo, a influência do grau de empacotamento do DNA é demasiado sutil para ser averiguada da forma proposta, dentro da quantidade de dados utilizada.

4 Conclusões

Algumas tendências esperadas foram aqui reforçadas. O trecho na UTR-3' do transposon Pererê-3 mostrou-se de fato distinto entre as cópias via alinhamentos BLAST. Mesmo que pouco abaixo do limiar heurístico, a correlação de transcrição entre uma cópia de transposon e seu gene mais próximo ainda mostrou-se significativamente mais alta quando a cópia está *downstream* ao gene, o que reforça hipóteses de grande atuação da transcrição passiva dos elementos de transposição por mecanismos de falha do terminador gênico. Ainda que o efeito de *outliers* nos cálculos de correlação possa ser melhor investigado, há forte evidência de que cópias inseridas na região codificante de um gene apresentam transcrição em grande maior taxa e com mais coordenação com o gene em comparação a cópias exteriores aos quadros gênicos de leitura aberta. A distribuição de comprimentos das cópias também mostrou comportamento interessante, com maiores frequências de valores ou abaixo de 750 bp, casos que em tese facilitariam a inserção das cópias nas regiões transcritas por genes, ou acima de 3150 bp, quando há mais probabilidade de existir transcrição autônoma funcional.