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. Visto isso, os diferentes níveis de transcrição em conjunto poderiam oferecer um panorama da atividade do genoma ao longo das bases[1].

Procura-se, portanto, investigar 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*[2] 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 através 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.

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"[3]. Nota-se, contudo, a existência, em muitos casos, 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[4].

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[5], en-

[1] Isso é um pivô do projeto?
[2] referência à WormBase
[3] Talvez venha a convir uma mudança de nome.
[4] Discursar mais sobre a região repetitiva?
[5] citar.
[6] citar.

quanto que na contagem de reads alinhados por região empregou-se o software HTSeq[6]. 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.[7]

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.[8]

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, na fase 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 à jusante do 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.

Avalia-se, portanto, a correlação entre os coeficientes de transcrição de cada cópia e o coeficiente de seu gene mais próximo 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 à jusante ou montante do 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. Separando as copias na mesma fita das em fitas diferentes, espera-se tambem observar efeito da compactação da cromatina na região.

3 Resultados

3.1 As cópias são de fato diferentes?

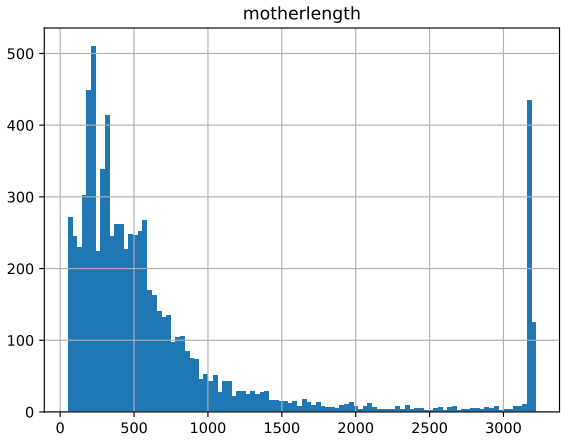
Visando constatar a unicidade das cópias de transposon pelo genoma, efetuou-se alinhamentos BLAST das sequências head entre elas mesmas, e verificou-se que aproximadamente 87,6% das sequências não foram alinhadas com nenhuma outra e por volta de 96,0% alinharam no máximo uma vez, mostrando que, de fato,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

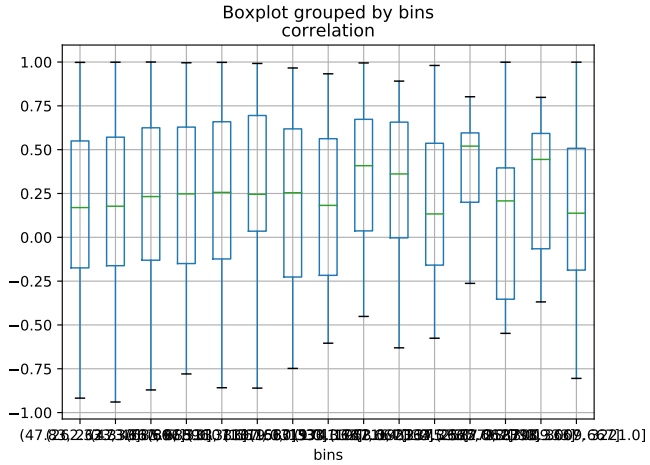
3.2 A integridade afeta os níveis de transcrição?

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.

[7] Já existe?

[8] 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.





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 3.2. Contudo, é interessante observar também um aumento significativo da quantidade de cópias íntegras ou quase íntegras, na extrema direita da figura 3.2. 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.

Em seguida, passou-se a analisar mais diretamente a influência do comprimento-mãe na transcrição: elaborou-se gráfico de dispersão entre o coeficiente de transcrição de cada head e seu comprimento mãe, e o resultado é apresentado na figura 3.2[9].

[9] melhorar esse gráfico, talvez mandar um LOWESS

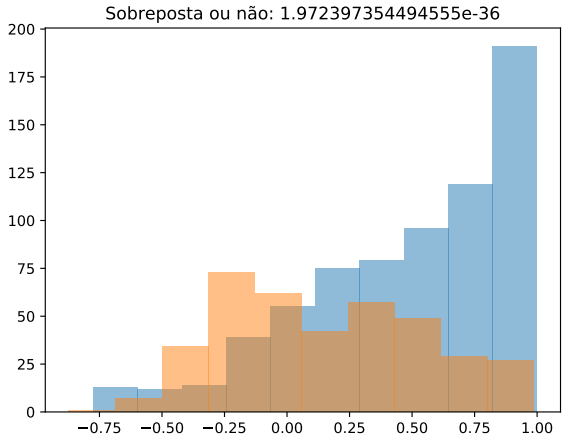
Embora a quatidade de dados seja muito maior nas extremidades direita e esquerda do gráfico, como já apontava a figura 3.2, não há diferença significativa de nível de transcrição entre as duas regiões, o que pode sugerir que, embora incompletas, as cópias de menor comprimento-mãe ainda sim façam uso 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 à jusante ou montante de algum, a fim de esclarecer efeitos de readthrough ou vazamento.

3.3 Há influência de genes próximos na transcrição das cópias?

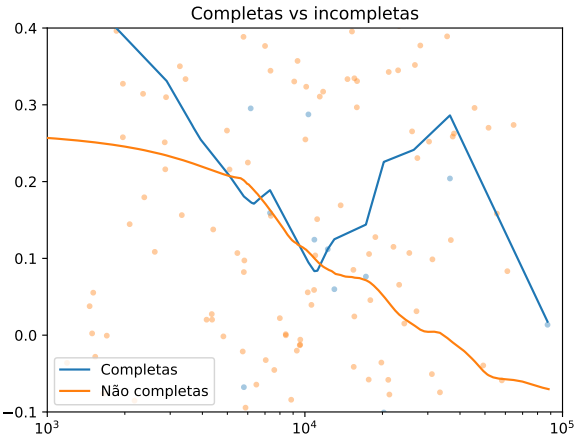
Inicialmente, dividiu-se as cópias entre as que sobrepunham alguma região gênica e as que não. Tomou-se os valores de correlação de transcrição com o gene vizinho (ou o gene sobreposto) e comparou-se a distribuição desses valores nas duas populações por meio de teste U de Mann-Whitney[10]. O p-valor obtido foi de aproximadamente $1,972 \cdot 10^{-36}$, sugerindo muito fortemente que há diferença entre os níveis de transcrição dos dois grupos, como aponta a figura 3.3.

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



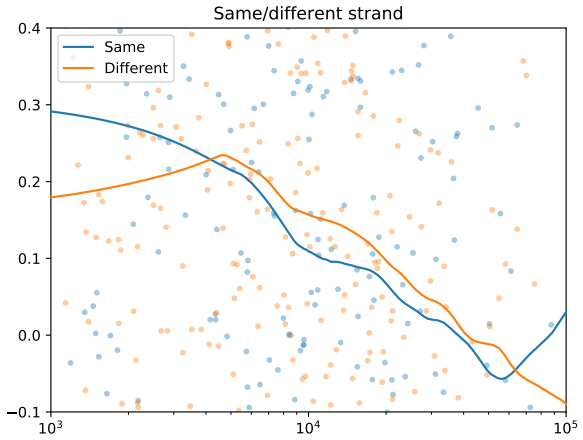
[discursar sobre efeito de outlier nas correlações próximas de 1.]

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, em um primeiro momento, cópias definidas como íntegras, com mais de 3150 pares de base, de cópias incompletas, com menos de 750 pares de base [explicar (in)conclusões]. O comportamento geral dos dados foi modelado por curvas LOWESS[11] e o resultado é apresentado na figura 3.3.



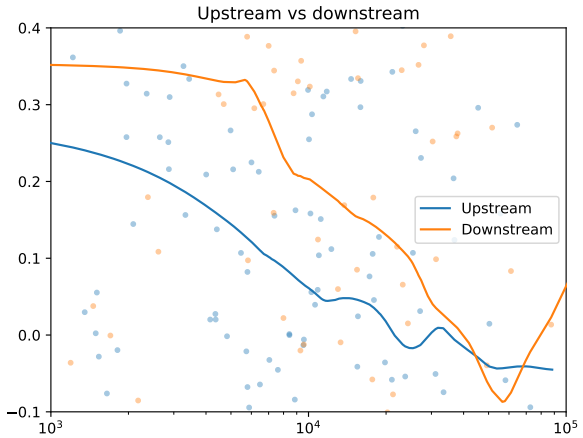
Da mesma forma, distinguiu-se cópias na mesma fita das em fita diferente de seu gene vizinho (figura 3.3), e observou-se que de fato há um comportamento decrescente na correlação de transcrição conforme a distância ao vizinho aumenta, como é natural de se esperar, principalmente entre $5 \cdot 10^3$ a $5\dot{1}0^4$ pares de base de distância. Para outras distâncias fora dessa faixa, o comportamento da correlação é ruidoso, muito provavelmente devido ao menor número de medidas.

[11] Locally Weighted Scatterplot Smoothing (ref)



Para a quantidade de dados analisada e o nível de ruído estatístico decorrente, não foi possível observar distinção de correlação de transcrição ao gene vizinho entre as populações de mesma fita ou em fita diferente a ele, obtendo p-valor aproximado de 0.1514 pelo teste U de Mann-Whitney.

O mesmo procedimento foi executado separando-se agora as heads à jusante das que estão à montante de seu gene mais próximo na mesma fita, e resultou na figura 3.3.



Há visível superioridade da correlação quando as heads estão à jusante do gene, 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. A superioridade é corroborada pelo teste U de Mann-Whitney, dado que o p-valor obtido foi de $9,452 \cdot 10^{-3}$ ao se comparar as duas populações.