PEDRO DE CARVALHO BRAGA ILÍDIO SILVA

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

Trabalho de conclusão do curso de Ciências Físicas e Biomoleculares apresentado ao Instituto de Física de São Carlos da Universidade de São Paulo como requisito parcial à obtenção do título de Bacharel.

Orientador: Ricardo De Marco

SÃO CARLOS

2019

Resumo

A quantificação da transcrição individual de sequências repetitivas, a exemplo de elementos de transposição, no genoma de um organismo geralmente se mostra uma tarefa complicada, visto que a alta taxa de ambiguidade não permite alinhamentos de *reads* de RNA-Seq específicos a cada uma delas. Em 2005, foi descrito um transposon do tipo não-LTR no genoma de *Schistosoma mansoni*, denominado Perere-3, com a incomum propriedade de produzir uma região 3'-UTR que se apresentava distinta em cada cópia do elemento (DEMARCO *et al*., 2015). Tal efeito deve-se a ausência de um sinal endógeno de término de transcrição, que leva à transcrição de uma região *downstream* a cada cópia. Nossas análises sugerem que, apesar da transcrição dessa região variável, ela não é reincorporada ao genoma no processo de reinserção do transposon, visto que tais regiões tendem a não se repetir na extremidade das cópias do Perere-3. Desta maneira, é possível utilizar essa região variável como uma sonda para monitorar individualmente o nível de transcrição (RPKM) de cada cópia, por meio do mapeamento e contagem de *reads* de RNA-Seq. Procurou-se, então, no presente trabalho, avaliar de que forma estão distribuídos os valores de RPKM entre as cópias, bem como investigar a influência de genes próximos a elas em sua atividade transcricional. Observou-se que ao menos 15% das cópias aparentam ser inativas, não apresentando indícios de transcrição, e que a grande maioria delas se mostra incompleta. Há também fortes evidências de que cópias sobrepostas a genes têm transcrição mais intensa e coordenada com eles, sugerindo que uma grande proporção das cópias, na verdade, faz parte de transcritos gênicos. Nota-se também coordenação da transcrição entre cada cópia e seu gene vizinho de maneira dependente da distância entre os dois. Isso sugere que efeitos de remodelamento de cromatina derivados da expressão dos genes vizinhos podem estar influenciando a expressão de elementos de transposição.

Palavras-chave: Perere-3. Transposon não-LTR. *Schistosoma mansoni*.

1 Introdução

Transposons, ou elementos de transposição, são sequências de DNA amplamente presentes no genoma dos organismos vivos, de forma que chegam a compor 85% do genoma de certas espécies (SCHNABLE *et al*., 2009) e estima-se que representa ao menos 45% do genoma humano (INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM *et al*., 2001). O que as caracteriza é a capacidade de codificar enzimas responsáveis por extrair ou replicar a própria sequência do transposon em si, e reintegrá-la ao genoma hospedeiro em uma outra localidade. São, portanto, uma forma essencial de mutabilidade genômica, tendo papel importante na variabilidade e evolução dos organismos ao longo das gerações.

Caso a reintegração ao genoma envolva produção de uma fita complementar de DNA a partir de molde intermediário de RNA, o transposon é denominado retrotransposon, e codifica, portanto, também uma transcriptase reversa.

Os retrotransposons são comumente divididos em duas classes, os transposons com repetições terminais longas (LTR, do inglês *long terminal repeats*) e os não-LTR, que não possuem tais extremidades repetitivas. A diferença fundamental entre as classes se dá na forma da reintegração de suas cópias ao genoma. No caso dos transposons LTR, a dupla fita de DNA é sintetizada externamente ao genoma do organismo e futuramente integrada a ele por mecanismos de recombinação, enquanto que os transposons não-LTR integram seu transcrito de RNA ao genoma temporariamente, e o utilizam já na região alvo como molde à produção da fita complementar de DNA. O transcrito é removido posteriormente pela ação de enzimas exonucleases e é substituído por fita análoga de DNA utilizando-se a extremidade 3’ aberta do genoma como primer na sintetização (BEAUREGARD, 2008).

Estudos anteriores de sequenciamento de pontas de transcritos do transposon não-LTR Perere-3 com a técnica de RACE demonstraram que transcritos deste elemento possuíam uma ponta 3’ variável logo após seu sinal de parada de tradução (DEMARCO *et al*., 2005). Mapeamentos desses transcritos no genoma de *Schistosoma mansoni* demonstraram que esse fenômeno deve-se ao fato de que as cópias de Perere-3 não possuem um sinal de parada de transcrição endógeno, e de que, portanto, os transcritos produzidos a partir deste transposon

possuem uma cauda 3' que tem como molde a região *downstream* do transposon. Devido à alta similaridade entre as cópias dos elementos de transposição, geralmente não é possível atribuir sequências de RNASeq a uma determinada cópia. No entanto, no caso do Perere-3, seria possível mapear sequências nas regiões *downstream* de cada uma delas e, desta maneira, obter uma medida cópia-específica da atividade transcricional. Em decorrência desse novo fato, novas questões podem ser levantadas, e abre-se um novo campo de exploração.

Seria possível avaliar, a partir dos níveis de transcrição de cada cópia separadamente, se temos poucas cópias que são responsáveis pela maior parte da transcrição do Perere-3, além de que, se cada réplica do transposon carrega o mesmo promotor, as diferenças desses níveis devem se originar de fatores externos à sequência do transposon em si, a exemplo da compactação da cromatina ou falhas na terminação da transcrição 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 Perere-3 em seus níveis de transcrição, fazendo uso de técnicas de bioinformática.

2 Metodologia

Inicialmente, utilizou-se a sequência da região 3' conservada do transposon (GenBank: BN000794.1 até a base 3196) para mapear as regiões candidatas a sequência molde da região variável de transcritos de Perere-3, a partir do que foi exposto por DeMarco, 2005. Efetuou-se busca por regiões semelhantes no genoma do *S. mansoni* por meio de alinhamento BLASTn (ALTSCHUL *et al*., 1990), no intuito de encontrar todas as cópias do Perere-3 inseridas no material genético do platelminto.

Já sabia-se de antemão que um outro transposon, denominado SR-3, possuía semelhança o suficiente com o transposon de nosso interesse para fazer com que o Perere-3 alinhasse também com as cópias de SR-3 no processo de busca ao longo do genoma. Toma-se então o cuidado de repetir o procedimento descrito antes com a região do SR-3 análoga à região do Perere-3 com que estamos trabalhando e prosseguir apenas com os alinhamentos que obtiveram maior *score* quando

realizados com o Perere e, portanto, que apresentam maior probabilidade de serem de fato uma cópia deste último.

A análise das distribuições de *reads* de RNASeq *downstream* do final do elemento Perere-3 sugere que, mesmo a uma distância de 1.000 pb, ainda há quantidade razoável de *reads* alinhados, nos levando a inferir que, na maioria das cópias, a interrupção da transcrição ocorre em comprimentos maiores que o mencionado. Caso contrário, esperaríamos observar decrescimento da contagem a partir de alguma posição (Figura 1).

Nota-se que o aumento do números médio de *reads* próximo da ponta 3' real do transcrito provavelmente reflete vieses posicionais comumente observados em dados de RNASeq (TUERK *et al*., 2017). Ainda mostra-se também um decaimento brusco da contagem nas extremidades do gráfico da Figura 1, em decorrência da diminuição da quantidade de alinhamentos possíveis que incluam o par de base.

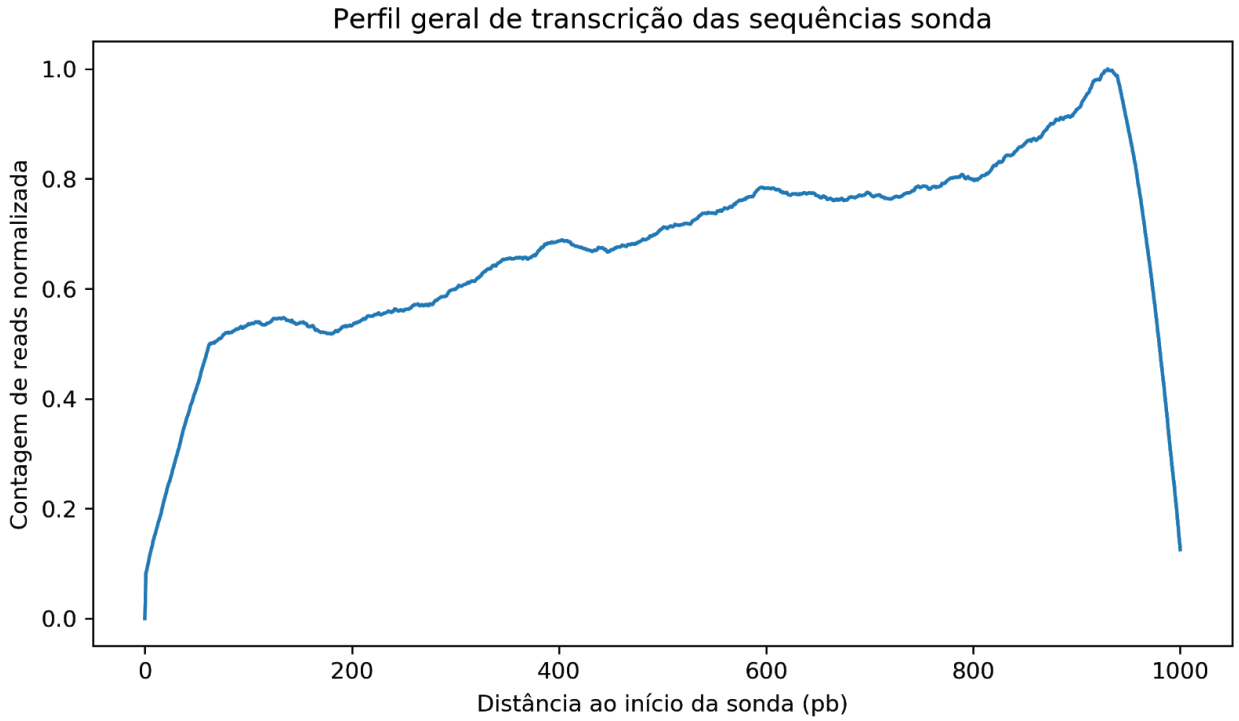

Figura 1: Soma das contagens de *reads* em cada posição ao longo das sondas, dividindo-se a contagem em cada par de base da sonda pela contagem total na região ocupada por ela e dividindo-se ainda os valores finais das somas em cada posição pelo maior valor encontrado.

Desta maneira, para a realização de medidas de expressão de cada cópia, utilizamos uma janela de 1.000 pares de base à imediata jusante de cada alinhamento encontrado com o BLASTn no procedimento anterior. A essas regiões, específicas de cada cópia, daremos o nome de sequências **sonda**. Em muitos casos, nota-se a existência de uma pequena região com a sequência GTAA-repetitiva, entre cada sonda e a região constante de sua cópia-mãe, a região comum a todas as cópias, de forma que consideramos como sonda as 1.000 bases posteriores às repetições. Nota-se que a presença desse tipo de sequência repetitiva é bastante comum em pontas dos elementos de transposição. (HAN *et al.*, 2010)

A fim de determinar se as regiões sonda representavam regiões únicas do genoma, realizamos um alinhamento com o programa BLAST de cada uma destas sequências contra o conjunto total de regiões sonda e verificamos que a sequência dessas regiões não apresentava redundância na grande maioria dos casos.

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 Perere-3, descritos a seguir.

2.1 Contagem de *reads*

Para medidas de níveis de transcrição, faz-se uso das bibliotecas de RNA-Seq com os seguintes códigos de acesso no SRA: ERR022872, ERR022873, ERR022874, ERR022875, ERR022876, ERR022877, ERR022879 e ERR022881, 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 mapeamento de *reads* no genoma, utilizou-se a ferramenta HiSat2 (KIM; LANGMEAD; SALZBERG, 2015) que gerou um arquivo de mapeamento do tipo SAM. Os dados de mapeamento foram cruzados com as coordenadas de genes ou sequências sonda em arquivo GFF utilizando o software HTSeq (ANDERS; PYL; HUBER, 2015), permitindo assim deduzir a quantidade de *reads* em cada região ou gene. Posteriormente, esses dados foram convertidos para a métrica de RPKM a partir da divisão do valor final da contagem pelo comprimento da região em que se procura o alinhamento dos *reads* e pelo tamanho de cada biblioteca, a fim de se obter um valor normalizado em relação a esses fatores.

2.2 Avaliação do comprimento das cópias de Perere-3

A fim de avaliarmos a integridade de cada cópia, recuperamos o comprimento do alinhamento BLAST da cópia completa do elemento Perere-3 com as várias regiões do genoma. Levamos em conta apenas alinhamentos que continham a extremidade 3' da cópia de referência, de forma que sequência sonda estava sempre adjacente à ponta 3' da sequência conservada de Perere-3 no genoma.

2.3 Correlação com genes vizinhos

Espera-se que grande parcela das cópias encontradas no genoma sejam transcritas por efeito de *readthrough*, ou transcrição passiva, em que as cópias do transposon são transcritas por fazerem parte de outros transcritos já produzidos (principalmente em UTRs) 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 gênico. Supõe-se que esses casos resultem em coeficientes de transcrição independentes da integridade da cópia de Perere-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 essa correlação nos referiremos daqui em diante como correlação com o gene vizinho, e espera-se que seja maior no casos em que a transcrição da sequência sonda é passiva.

2.4 Distância ao gene vizinho

Ainda sob a questão da influência dos genes mais próximos no coeficiente de transcrição das sequências sonda, outro fator a se considerar é a distância em si entre a sonda 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 sonda 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 regiões sonda das cópias são de fato diferentes?

Conforme descrito na introdução, devido a produção de transcritos de Perere-3 que possuíam região 3'-UTR variável, decidimos utilizar uma região de 1000 pb *downstream* do sinal de parada de tradução do Perere-3 como uma região sonda. Visando verificar a similaridade entre regiões sonda do transposon no genoma, efetuou-se alinhamentos BLAST das sequências sonda 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 Perere-3, e a temática de análise proposta mantém-se válida.

O fato de não encontrarmos múltiplas cópias das regiões sonda sugere que, apesar da presença de regiões variáveis na região 3' do transcrito de Perere-3, elas não devem sofrer o processo de transcrição reversa durante a integração do elemento no genoma.

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 RPKMs menores que 5, sendo que por volta de 15,648% apresentaram nenhuma transcrição. A distribuição em escala logarítmica é apresentada na Figura 2.

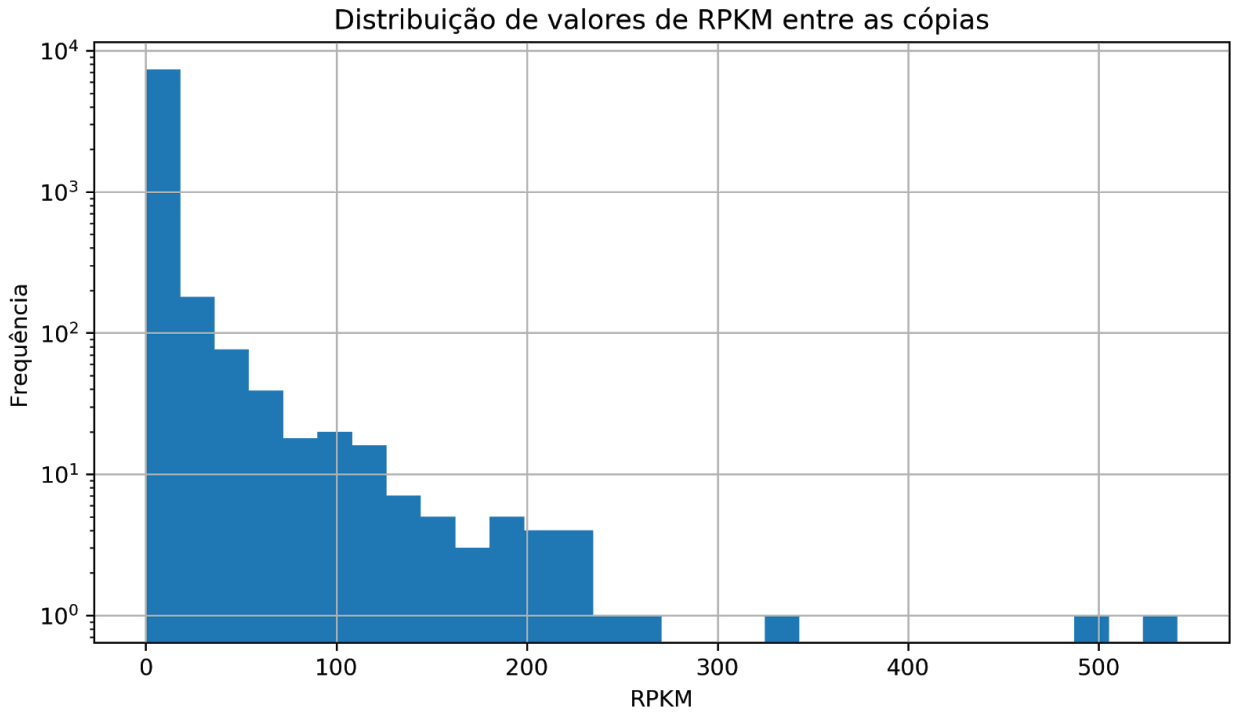

Figura 1: Histograma de distribuição de valores de RPKM relativos às regiões sonda das cópias de Perere-3 presentes no genoma.

Conclui-se assim, que, muito provavelmente, uma grande parcela (por volta de ao menos 15%) das cópias de Perere-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 tamanhos das regiões invariantes das diversas cópias encontradas (Figura 2).

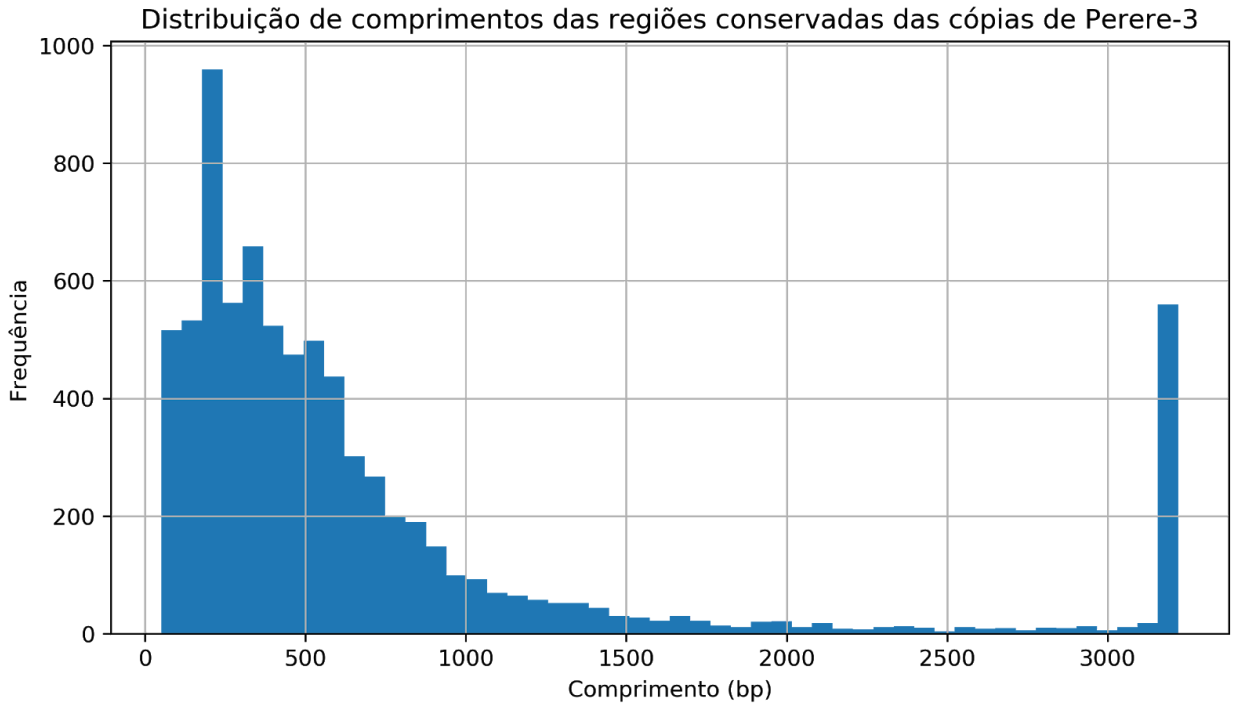

Figura 2: Histograma da distribuição dos comprimentos da região invariante das cópias de Perere-3.

Já nesse ponto, há uma clara predominância das cópias menores, principalmente com menos de 750 pares de base de comprimento, representadas mais à esquerda na Figura 2. Contudo, é interessante observar também um aumento significativo da quantidade de cópias íntegras ou quase íntegras, na extrema direita da Figura 2. Esse aumento sugere fortemente que a integridade das cópias é um fator determinante para sua disseminação (ou ao menos sobrevivência) e que ainda existem, de fato, cópias ativas no genoma do S. mansoni.

A fim de caracterizar a dependência entre os níveis de transcrição e a integridade das cópias, divide-se as cópias de Perere-3 em três populações, estipuladas a partir da distribuição de comprimentos exposta na Figura 2: uma região de cópias com alto grau de completude, de comprimento da região invariante maior que 3150 pb; uma região de cópias incompletas, com menos de 750 pb de comprimento da região constante; e uma região intermediária ruidosa, devido a seu menor volume de dados. As distribuições de valores de RPKM, para cada população é então disposta na forma de um boxplot e apresentada pela Figura 3.

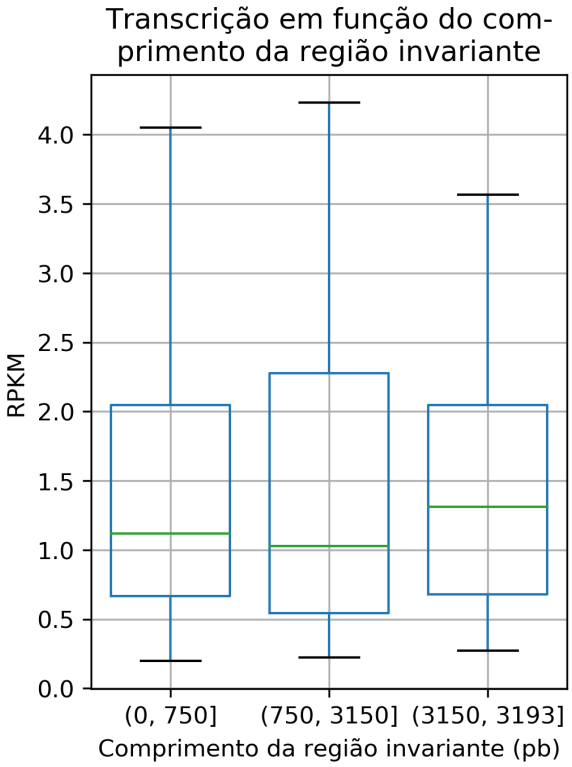

Figura 3: Comparação das distribuições de RPKM sob forma de *boxplot* para diferentes faixas de comprimento da região invariante de cada cópia. A caixa da direita representa a população de cópias íntegras ou quase íntegras, enquanto que predominam cópias incompletas na caixa da esquerda. A caixa no centro da Figura representa a região intermediária com menor quantidade de dados.

Nota-se ligeira elevação da mediana das regiões extremais em relação à região central, sugerindo maiores níveis de transcrição das cópias em conjunto nessas populações. A diferença, contudo, não se mostra bem delineada pelo teste estatístico (p-valores 0.0980 e 0.284 para as comparações entre incompletas e intermediárias e entre completas e intermediárias, respectivamente, teste U de Mann-Whitney). Uma possível causa seria a ausência de valores suficientes na região central, evidenciada pela Figura 2, de forma a tornar o dado ruidoso e requerer futuras análises para esclarecimento.

Embora a quantidade de dados seja muito maior nas extremidades direita e esquerda do gráfico, como já apontava a Figura 2, também não há diferença significativa de nível de transcrição entre as duas regiões (p-valor ≃ 0.433, teste U de Mann-Whitney), o que pode sugerir que, embora incompletas, as cópias de menor comprimento 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 exigido na etapa de escolha dos alinhamentos BLAST do Perere-3 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.

Desta maneira, espera-se que ocorra transcrição autônoma apenas nas cópias completas, sendo que a alta expressão de cópias incompletas pode ser derivada do fato de várias delas estarem sendo transcritas de modo passivo, devido a pertencerem a UTRs de genes próximos, ou fenômenos de vazamento na terminação de transcrição, um fenômeno no qual transcritos de maior tamanho são produzidos a partir de falhas no processo de terminação (GUO *et al*., 1991).

3.4 Como os níveis de transcrição de cada cópia são afetados pela presença de genes próximos?

A fim de observar a influência de genes próximos na atividade transcricional das cópias de Perere-3, divide-se a população de cópias encontradas em grupos e compara-se suas distribuições de valores de RPKM.

Em um primeiro momento, separa-se cópias sobrepostas a alguma região gênica das que não sobrepõem. Em seguida, divide-se também de outra forma a população total: entre as cópias que se encontram *downstream* a seu gene mais próximo e aquelas que se encontram *upstream* a ele. Os resultados são apresentados na Figura 4.

As cópias que se encontram em regiões gênicas possuem níveis de transcrição visivelmente mais elevados em relação às que não se encontram em regiões já transcritas (Figura 4a, p-valor ≃ 1,573 . 10^{-63}, teste U de Mann-Whitney), o que é explicado por fazerem uso da maquinaria não própria de transcrição, utilizando o promotor gênico e, portanto, não necessitando estarem completas. Trata-se, portanto, de uma evidência da ocorrência de *readthrough*.

Comparando as cópias *downstream* com as *upstream* de seu gene vizinho (Figura 4b), não se observa diferença clara na atividade transcricional (p-valor ≃ 0,164, teste U de Mann-Whitney), de forma que não se pode concluir satisfatoriamente se há ocorrência de vazamento de transcrição, em que uma cópia *downstream* a algum gene seria transcrita por falha do terminador de transcrição deste último.

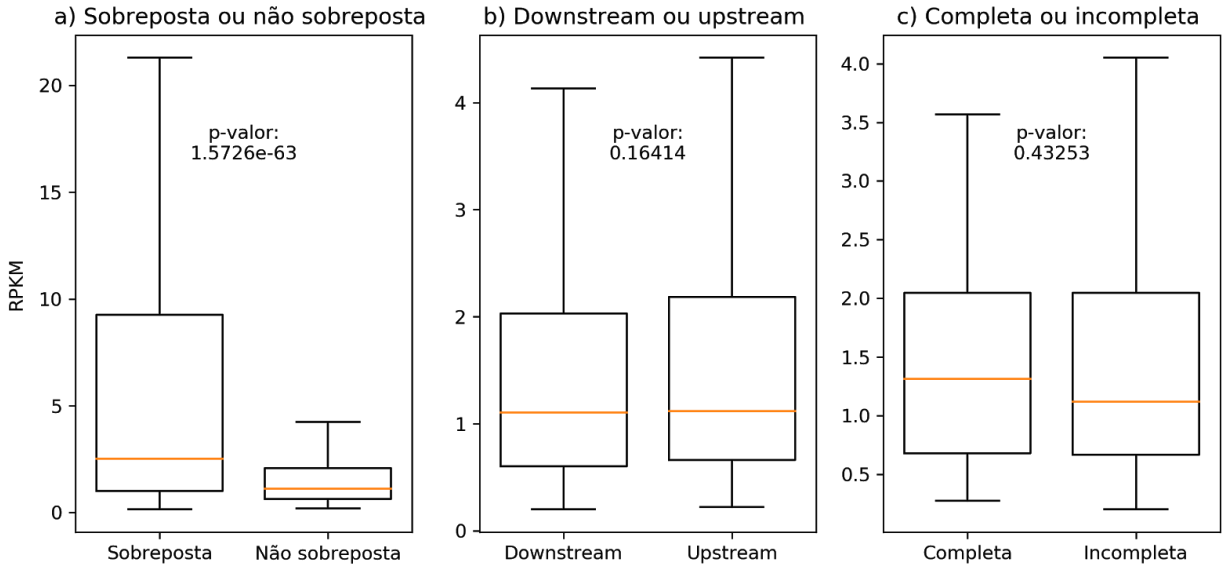

Figura 4: *Boxplots* das distribuições dos valores de RPKM das cópias de Perere-3 com seus respectivos genes vizinhos, segundo diferentes formas de amostragem das populações de acordo com características diversas. p-valores calculados pelo teste U de Mann-Whitney entre cada par de populações. a) Comparação entre as distribuições de RPKM no caso em que as cópias sobrepõem algum gene e no caso em que estão em regiões distintas do gene mais próximo. b) Comparação entre o caso em que as cópias se encontram *downstream* do gene mais próximo e o caso em que ela se encontra *upstream* de seu gene vizinho. c) Comparação revisitada entre as cópias completas, com mais de 3150 pb de extensão de sua região invariante, e as incompletas, com região invariante de comprimento menor a 750 pb, já exibida na Figura 3.

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

Nesta seção procura-se dividir a população de sondas entre diferentes posições relativas ao seu gene mais próximo. Nota-se que, em alguns poucos casos, não havia nenhum gene no mesmo contig da sonda, 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 sonda 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 sondas sobre as quais menos de quatro bibliotecas de *reads*, das oito bibliotecas

utilizadas, possuíam contagem de *reads*, de forma a permitir uma análise de correlação menos ruidosa. Isso impactou em perda considerável de dados: mantiveram-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 sobreposto) e comparou-se a distribuição desses valores nas duas populações por meio de teste U de Mann-Whitney, também conhecido como teste da soma dos postos de Wilcoxon. O p-valor obtido foi de aproximadamente $4{,}2 \cdot 10^{-78}$, sugerindo muito fortemente que de fato há diferença entre os níveis de transcrição dos dois grupos, como aponta a Figura 5a.

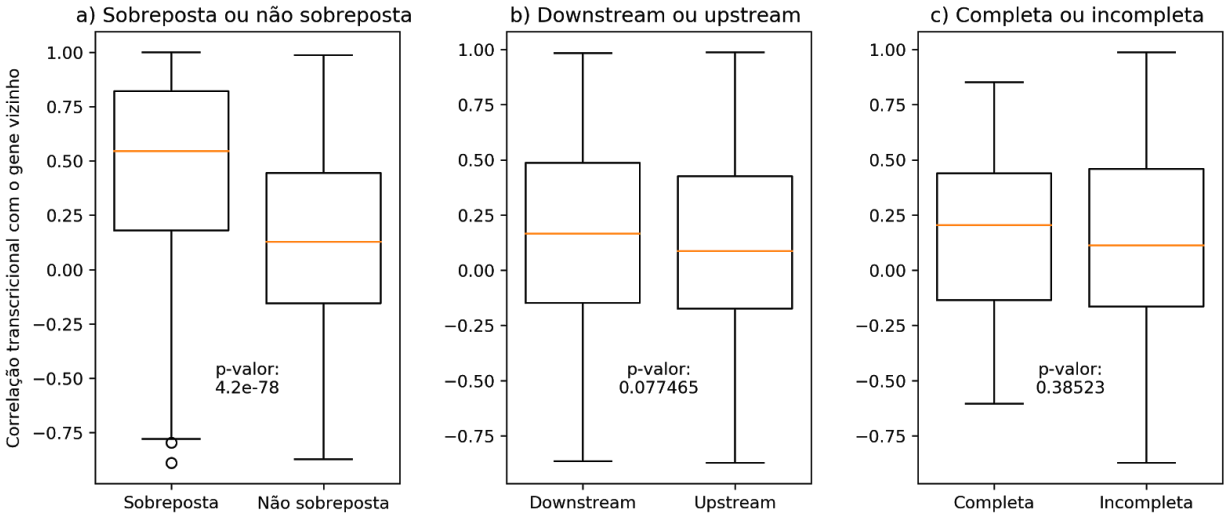

Figura 5: *Boxplots* das distribuições dos valores de correlação transcricional das cópias de Perere-3 com seus respectivos genes vizinhos, segundo diferentes formas de amostragem das populações de acordo com características diversas. p-valores calculados pelo teste U de Mann-Whitney entre cada par de populações. a) Comparação entre a correlação com o gene no caso em que as cópias os sobrepõem e no caso em que estão em regiões distintas do gene mais próximo. b) Comparação entre o caso em que as cópias se encontram *downstream* do gene mais próximo e o caso em que ela se encontra *upstream* de seu gene vizinho. c) Comparação entre as cópias completas, com mais de 3150 pb de extensão de sua região invariante, e as incompletas, com região invariante de comprimento menor a 750 pb.

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 sondas 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 da região invariante, de cópias incompletas, com menos de 750 pares de base nessa região. O comportamento geral dos dados foi modelado por curvas LOWESS (Locally Weighted Scatterplot Smoothing) e o resultado é apresentado na Figura 6.

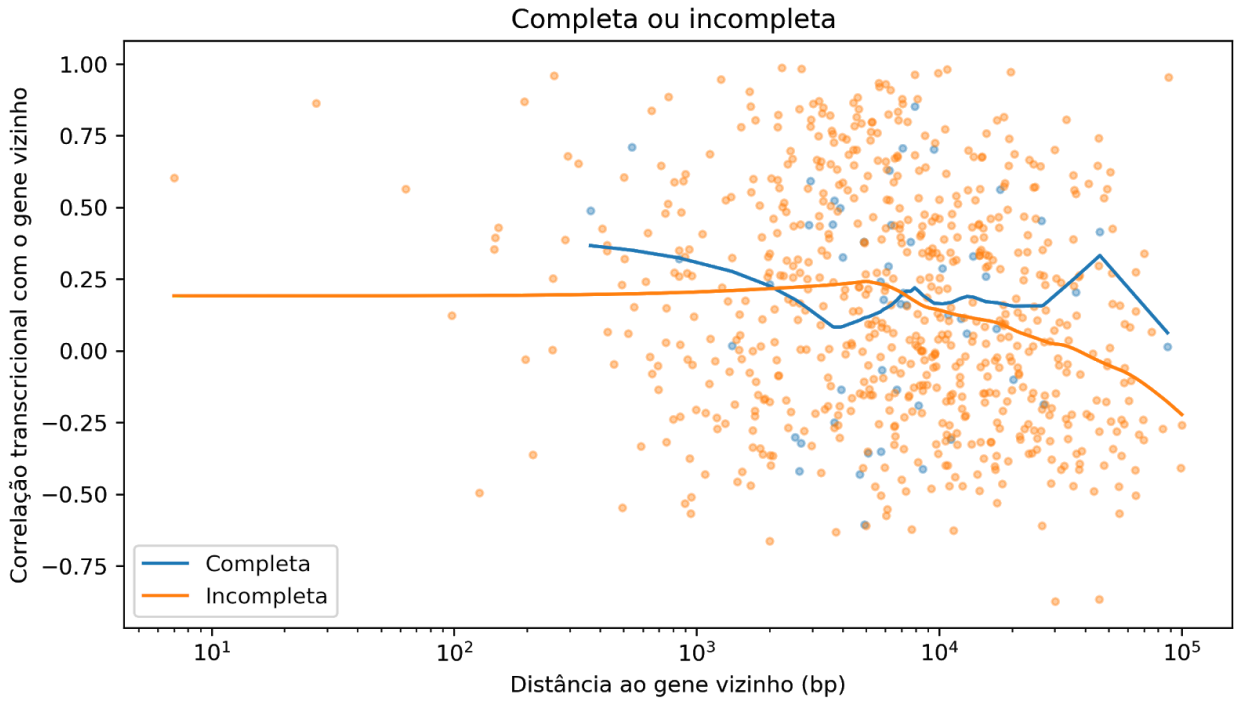

Figura 6: Gráfico de dispersão do coeficiente de correlação entre a expressão da sequência sonda e a expressão do gene vizinho mais próximo em função da distância entre eles. Pontos laranjas representam sondas de cópias incompletas do transposon e pontos azuis sondas de cópias completas. Linhas de tendência para cada um dos conjuntos utilizando a metodologia de regressão não-paramétrica de LOWESS foram calculadas. Foram consideradas cópias completas e incompletas aquelas com comprimento maior que 3150 bp e menor que 750, respectivamente.

É possível notar que, no caso das cópias incompletas, há uma clara tendência de queda de correlação a altas distâncias, o que parece indicar existência de um efeito espacial na influência

do gene na expressão do transposon. Isso seria compatível tanto com um fenômeno de vazamento de transcrição, como com um efeito de de abertura de cromatina na região de um promotor gênico que esteja intensificando também a atividade de um promotor próprio de alguma cópia próxima.

No caso da cópias completas é possível observar um formato irregular da curva de tendência, e concluímos que isso seria uma consequência do pequeno número de pontos para essa população, impedindo uma conclusão clara sobre a dependência distância-correlação transcricional nessa amostra.

No entanto, é possível notar que, mesmo para essa população, a linha de tendência se mantém acima do zero em todas as distâncias, o que parece sugerir a existência de uma coordenação de expressão entre cópias completas e os genes vizinhos. Neste caso, visto que estas cópias devem possuir um promotor independente, imagina-se que a principal explicação seria o efeito de abertura de cromatina que influenciaria a expressão do transposon.

A comparação entre as distribuições das duas populações em questão, de cópias íntegras ou não íntegras, mostra uma distribuição semelhante, com medianas bastante próximas (Figura 5c). De fato, uma análise estatística revela que não há diferenças significativas entre as duas populações (p-valor ≃ 0,385, teste U de Mann-Whitney).

3.5.3 Maior correlação de cópias *downstream* de um gene não se mostra suficientemente evidente

Devido ao fato de não sabermos se os efeitos de correlação da expressão entre o transposon e o gene vizinho se devem a eventos de vazamento ou efeitos de abertura de cromatina, desenhamos novos experimentos a fim de tentar inferir a influência dos diferentes processos. O mesmo procedimento da subseção anterior foi executado separando-se das demais agora as cópias *downstream* de seu gene mais próximo (Figura 7). 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, se aproximando de zero a maiores distâncias.

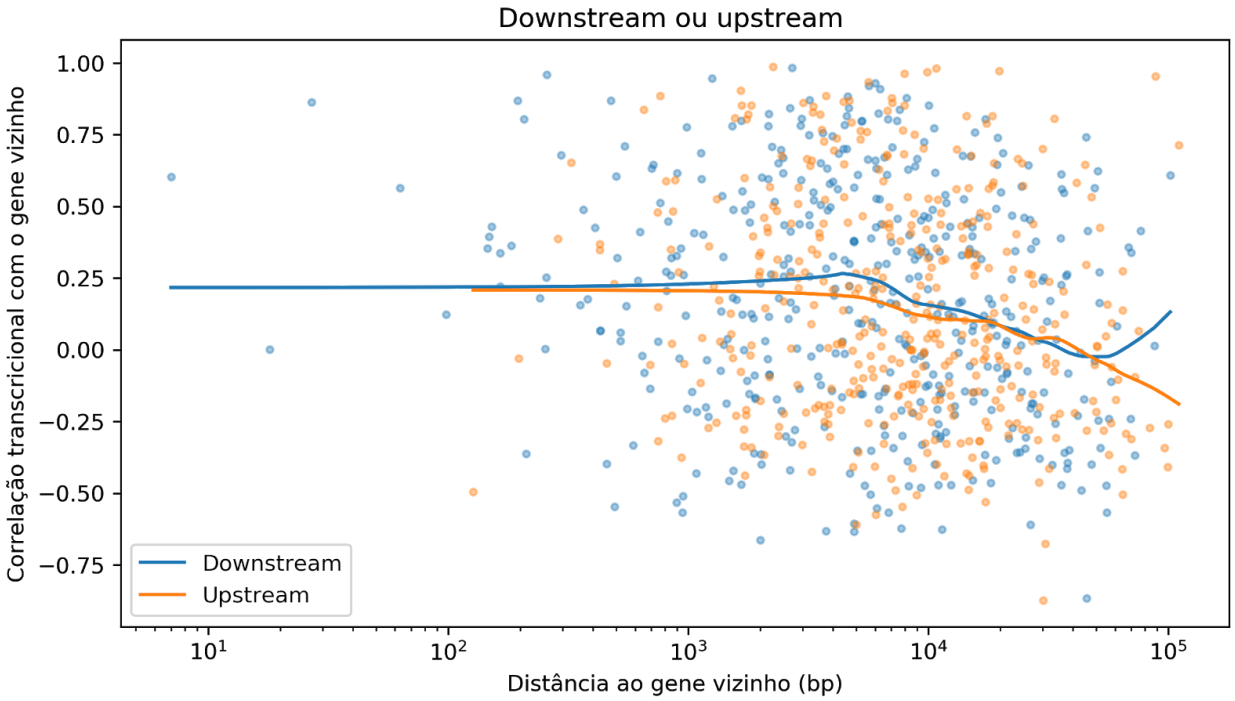

Figura 7: Gráfico de dispersão do coeficiente de correlação entre a expressão da sequência sonda e a expressão de seu gene mais próximo em função da distância entre eles. Pontos azuis representam sondas de cópias que apresentam-se em posição *downstream* do gene mais próximo, e pontos laranjas as sondas remanescentes. Linhas de tendência para cada um dos conjuntos utilizando a metodologia de regressão não-paramétrica de LOWESS foram calculadas.

Pela Figura 7, as cópias *downstream* parecem ainda revelar certa superioridade da correlação com o vizinho no gráfico apresentado, dado que sua curva LOWESS permanece superior à das demais cópias principalmente na região com maior número de amostras, com distância ao vizinho entre 10^3 e 10^4 pares de base, reforçando a hipótese de que há vazamento nas transcrições gênicas, ou seja, que o nível de transcrição das sondas pode ser incrementado se ela estiver próxima à extremidade 3' de uma região codificante do genoma, provavelmente em decorrência de falha na terminação da transcrição.

A análise estatística das correlações das populações *upstream* e *downstream*, contudo, não demonstra diferença clara de correlação, com as populações *downstream* e *upstream* apresentando (Figura 8) medianas aproximadas de 0.16463 e 0.08582, respectivamente (p-valor ≈ 0.0775, teste U de Mann-Whitney). Se selecionarmos apenas as subpopulações com distância abaixo de 20.000 pb, tal diferença se torna bem pouco mais evidente, com medianas de

aproximadamente 0.20990 e 0.12773, respectivamente (p-valor ≈ 0.07614, teste U de Mann-Whitney).

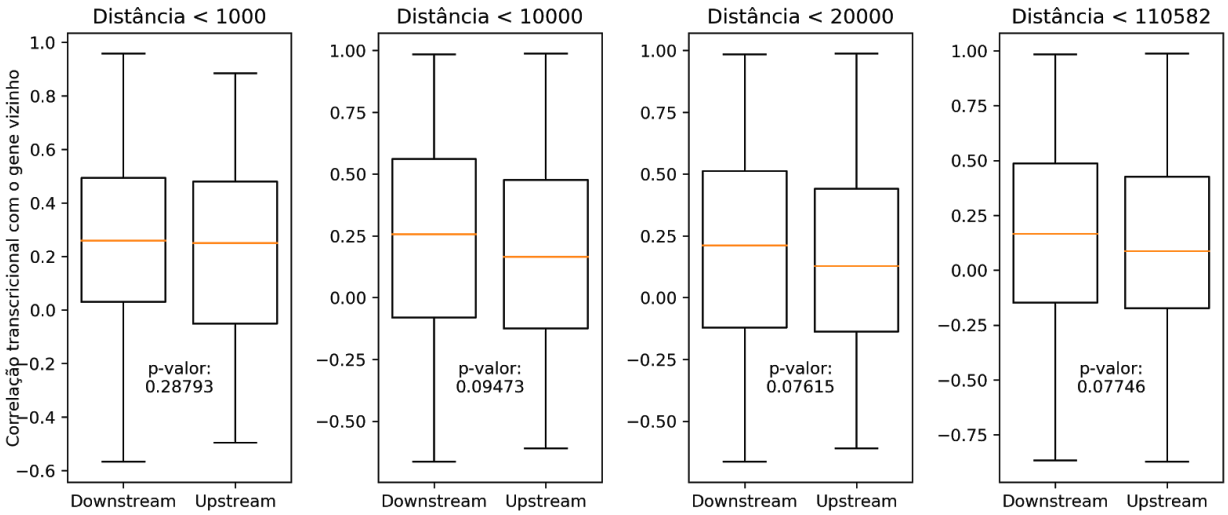

Figura 8: Comparações entre as populações de cópias *upstream* e *downstream* do seu gene mais próximo, para diferentes limiares da distância cópia-gene vizinho em pares de base.

4 Conclusões

Nossa análise demonstrou que diversas cópias do transposon Perere-3 possuem transcrição na região *downstream* da extremidade 3' da região conservada. Há forte evidência de que cópias inseridas sobrepostas a um gene apresentam transcrição em 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 grande maior volume de comprimentos menores que 750 pb, o que sugere que a maioria desses transposons devem fazer parte do transcrito gerado pelo gene e não apresenta transcrição independente dele.

Além disso, vimos uma tendência de que a transcrição desses elementos seja realizada de forma coordenada com o gene vizinho, ainda que de maneira modesta. Isso sugere influência da atividade transcricional dos genes vizinhos na transcrição do elemento. Análises adicionais, realizando simulações com a randomização dos dados de pares transposon-gene permitirão avaliar a robustez desse achado. Deve-se notar, no entanto, que já foi descrito, em genomas de eucariotos, que genes vizinhos tendem a apresentar uma coordenação de expressão devido a efeitos relacionados ao remodelamento de cromatina (BATADA *et al*., 2007).

Uma outra possibilidade seria um efeito de vazamento da terminação de transcrição, em que uma pequena fração dos transcritos de um gene não apresentaria terminação adequada, criando um transcrito que avançaria sobre a sequência do elemento de transposição adjacente. Ambos os casos seriam adequados para explicar o efeito da distância na diminuição da correlação de expressão. No entanto, se a hipótese de vazamento fosse correta, esperaríamos uma correlação consideravelmente maior em transposons localizados *downstream* do gene em relação àqueles *upstream*, e que cópias incompletas (que possuem menor probabilidade de possuir um promotor independente) apresentassem correlação maior que cópias completas. A ausência destas tendências parece favorecer a hipótese de coordenação de expressão devido a remodelamento de cromatina, sugerindo que a maioria dos elementos de transposição apresentando transcrição o fazem com um promotor fisicamente independente do gene vizinho.

Referências

ALTSCHUL, Stephen F. *et al.* Basic local alignment search tool. **Journal of molecular biology**, v. 215, n. 3, p. 403-410, 1990.

ANDERS, Simon; PYL, Paul Theodor; HUBER, Wolfgang. HTSeq—a Python framework to work with high-throughput sequencing data. **Bioinformatics**, v. 31, n. 2, p. 166-169, 2015.

BATADA, Nizar N.; URRUTIA, Araxi O.; HURST, Laurence D. Chromatin remodelling is a major source of coexpression of linked genes in yeast. **Trends in genetics**, v. 23, n. 10, p. 480-484, 2007.

BEAUREGARD, Arthur; CURCIO, M. Joan; BELFORT, Marlene. The take and give between retrotransposable elements and their hosts. **Annual review of genetics**, v. 42, p. 587-617, 2008.

DEMARCO, Ricardo *et al*. Identification of 18 new transcribed retrotransposons in *Schistosoma mansoni*. **Biochemical and biophysical research communications**, v. 333, n. 1, p. 230-240, 2005.

GUO, Wentong *et al*. Leaky transcription termination produces larger and smaller than genome size hepatitis B virus X gene transcripts. **Virology**, v. 181, n. 2, p. 630-636, 1991.

HAN, Jeffrey S. Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent developments, and unanswered questions. **Mobile Dna**, v. 1, n. 1, p. 15, 2010.

HUNTER, John D. Matplotlib: A 2D graphics environment. **Computing in science & engineering**, v. 9, n. 3, p. 90, 2007.

INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM *et al*. Initial sequencing and analysis of the human genome. **nature**, v. 409, n. 6822, p. 860, 2001.

KIM, Daehwan; LANGMEAD, Ben; SALZBERG, Steven L. HISAT: a fast spliced aligner with low memory requirements. **Nature methods**, v. 12, n. 4, p. 357, 2015.

MCKINNEY, Wes *et al*. Data structures for statistical computing in python. In: **Proceedings of the 9th Python in Science Conference**. 2010. p. 51-56.

SCHNABLE, Patrick S. *et al*. The B73 maize genome: complexity, diversity, and dynamics. **science**, v. 326, n. 5956, p. 1112-1115, 2009.

TUERK, Andreas; WIKTORIN, Gregor; GÜLER, Serhat. Mixture models reveal multiple positional bias types in RNA-Seq data and lead to accurate transcript concentration estimates. **PLoS computational biology**, v. 13, n. 5, p. e1005515, 2017.

VIRTANEN, Pauli *et al*. SciPy 1.0--Fundamental Algorithms for Scientific Computing in Python. **arXiv preprint arXiv:1907.10121**, 2019.