218

INFLUÊNCIA DO TAMANHO DE READS E PRESENÇA DE DUPLICATAS NÃO NATURAIS EM BIBLIOTECAS DE RNA-Seq DE SOJA SOB DEFICIT HÍDRICO

MOLINARI, M.D.C.[1]; BARBOSA, D.A.[1]; FUGANTI-PAGLIARINI, R.[2]; ANDREATA, E.C.[3]; CARANHATO, A.L.H. [1]; MARIN, S.R.R.[4]; MERTZ-HENNING, L.M.[4]; NEPOMUCENO, A.L.[4]

[1]Universidade Estadual de Londrina - UEL, Londrina-PR, maylamolinari@hotmail.com; [2]Bolsista CNPq, Embrapa Soja; [3]UNOPAR; [4]Embrapa Soja.

Em metodologias de sequenciamento em larga escala como o RNA-Seq, as duplicatas não naturais podem causar erros no mapeamento e gerar dados de expressão diferencial de qualidade duvidosa. As duplicatas não naturais de PCR (Reação da polimerase em cadeia) surgem de produtos de PCR múltiplos da mesma molécula modelo, que se ligam à célula de fluxo e a sua remoção destina-se a reduzir o ruído e minimizar os falsos positivos (Li et al., 2009; Dozmorovv et al., 2015; Ebbert et al., 2016). Os pacotes de *softwares* mais utilizados na literatura para identificação e remoção destas duplicatas, são as ferramentas MarkDuplicates, do pacote Picard e rmdup do pacote SAMTools. A diferença entre elas está na forma como os algoritmos utilizados tratam os dados. O MarkDuplicates não exclui as duplicadas, apenas as sinaliza/marca, para que se possa escolher levar em consideração as duplicatas ou não nas análises seguintes, além de exigir mais memória RAM e demorar mais tempo para rodar os dados. O *software* rmdup, por outro lado, remove de forma rápida e com uso de pouca memória RAM. O SAMTools remove as duplicatas não naturais com eficiência de bibliotecas *single-end* (Ebbert et al., 2016). Para bibliotecas *paired-end* existem outras ferramentas mais adequadas para esta etapa, como o Depud, uma ferramenta do pacote EAGER (Peltzer et al., 2016).

A relação entre o tamanho dos *reads* e as duplicatas não naturais é direta, uma vez que uma das etapas de análise é a remoção de sequências muito curtas, abaixo de 40 pb que ocorrem várias vezes dentro da sequência alvo e, portanto, fornecem apenas informações ambíguas (Bolger; Lohse; Usabel, 2014). Quanto menor o tamanho dos fragmentos de entrada, menores os fragmentos após a etapa de limpeza. Assim, fragmentos sequenciados de 50 pb após a limpeza ficarão com 35-45 pb e os fragmentos de 100 pb ficarão, após a limpeza, com fragmentos entre 40-90 pb. Considerando esta informação, é importante, durante o delineamento experimental, realizar a escolha correta do tamanho dos *reads* que se deseja obter, dependendo da plataforma de sequenciamento e do objetivo do trabalho (Conesa et al., 2016). Bibliotecas geradas por exemplo, por plataformas Illumina podem gerar *reads* que variam de 50 a 200 pb (Liu et al., 2012; Quail et al., 2012; Buermans; Den Dunnen 2014; Vincent et al., 2016).

Portanto, como o tamanho final dos *reads* sequenciados relaciona-se diretamente com a qualidade do alinhamento no genoma referência, objetivou-se neste trabalho avaliar as taxas de ocorrência de duplicação não naturais em bibliotecas com diferentes tamanhos de *reads*, gerados por plataformas de sequenciamento Illumina.

Para realização deste trabalho foram utilizadas 24 bibliotecas de RNA-Seq gerados pela plataforma Illumina 1.9, obtidas a partir de experimentos de soja sob déficit hídrico. Todas as bibliotecas *single-end* avaliadas apresentavam uma cobertura de 1X o genoma referência (Soja – cultivar Williams 82). Doze destas bibliotecas foram sequenciadas com *reads* de 50 pb em média e as outras 12 bibliotecas sequenciadas com *reads* finais de 100 pb em média. O RNA total foi obtido a partir do reagente

Trizol, com resgate do RNAm e eliminação do RNAr. As amostras que apresentaram melhor qualidade e integridade (RIN≥ 8.0) (*Rna Integrity Number*) foram selecionadas para a síntese das bibliotecas. Os arquivos brutos FASTQ tiveram sua qualidade avaliada antes e depois da limpeza pelo *software* FastQC versão 0.11.5 (Andrews et al, 2010; Patel; Jain, 2012). A remoção dos adaptadores e de sequências de baixa qualidade foi realizada com o *software* Trimmomatic versão 0.36 (Bolger; Lohse; Usabel, 2014). O alinhamento das bibliotecas também foi realizado através do *software* HISAT2 v.2.1.0, e somente os *reads* com alinhamentos únicos foram utilizados. As taxas de duplicatas não naturais foram obtidas a partir da ferramenta rmdup do *software* O SAMTools V.1.5 (Li et al., 2009; Dozmorov et al., 2015; Ebbert et al., 2016). Os dados apresentaram distribuição normal segundo o teste de Shapiro-Wilk. Por se tratar de dados paramétricos foram realizadas análise de variância ANOVA e teste de separação de médias Tukey, ao nível de 5% de significância (Canteri, et al., 2001).

Os resultados obtidos mostraram que as bibliotecas que foram sequenciadas em fragmentos de 50 pb apresentaram maior número de duplicações não naturais quando comparada às bibliotecas geradas a partir de *reads* de 100 pb. Para algumas bibliotecas, foram identificados de 9 a 15% mais duplicatas não naturais. Bibliotecas sequenciadas em fragmentos de 100 pb apresentaram, ao contrário, menores taxas de duplicação não naturais. Embora consideradas artefatos provenientes da etapa de PCR (Ebbert et al., 2016), altas taxas de duplicações não naturais podem gerar um aumento no número de genes falsos positivos. Assim, foi possível concluir que a taxa de duplicatas não naturais é inversamente proporcional ao tamanho dos *reads* finais gerados pela plataforma Illumina, sugerindo que, para uma análise de dados mais acurada e confiável, deve-se dar preferência para bibliotecas de RNA-Seq sequenciadas em fragmentos maiores, principalmente quando o objetivo é a expressão diferencial de genes, onde genes falsos positivos podem superestimar os resultados.

Referências

ANDREWS, S. **FastQC: a quality control tool for high throughput sequence data** (2010). Disponível em: http://www.bioinformatics.babraham.ac.uk/projects/fastqc. Acesso em: Mar. 2018.

BOLGER, A.M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics,** v. 30, n. 15, p. 2114-2120, 2014.

BUERMANS, H.P.J.; DEN DUNNEN, J.T. Next generation sequencing technology: advances and applications. **Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease,** v. 1842, n.10, p. 1932-1941, 2014.

CANTERI, M.G. et al. SASM-Agri: Sistema para análise e separação de médias em experimentos agrícolas pelos métodos Scott-Knott, Tukey e Duncan. **Revista Brasileira de Agrocomputação**, v. 1, n. 2, p. 18-24, 2001.

CONESA, A.; MADRIGAL, P.; TARAZONA, S.; GOMEZ-CABRERO, D.; CERVERA, A.; MCPHERSON, A.; MORTAZAVI, A. A survey of best practices for RNA-seq data analysis. **Genome biology,** v. 17, n. 1, p. 13, 2016.

DOZMOROV, M.G.; ADRIANTO, I;, GILES, C.B.; GLASS, E.; GLENN, S.B.; MONTGOMERY, C.; LESSARD, C.J. Detrimental effects of duplicate *reads* and low complexity regions on RNA-and ChIP-seq data. **BMC bioinformatics,** v. 16, n. 13, p. S10, 2015.

EBBERT, M.T.; WADSWORTH, M.E.; STALEY, L.A.; HOYT, K.L.; PICKETT, B.; MILLER, J.; RIDGE, P.G. Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. **BMC bioinformatics**, v. 17, n. 7, p. 239, 2016.

LI, H.; HANDSAKER, B.; WYSOKER, A.; FENNELL, T.; RUAN, J.; HOMER, N.; DURBIN, R. The sequence alignment/map format and SAMTools. **Bioinformatics,** v. 25, n. 16, p. 2078-2079, 2009.
LIU, L.; LI, Y.; LI, S.; HU, N.; HE, Y.; PONG, R.; LAW, M. Comparison of Next-Generation Sequencing Systems. **Journal of Biomedicine and Biotechnology,** v. 2012, 2012.
PATEL, Ravi K.; JAIN, Mukesh. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. **PloS one**, v. 7, n. 2, p. e30619, 2012.
PELTZER, A.; JÄGER, G.; HERBIG, A.; SEITZ, A.; KNIEP, C.; KRAUSE, J.; NIESELT, K. EAGER: efficient ancient genome reconstruction. **Genome biology**, v. 17, n. 1, p. 60, 2016.
QUAIL, M.A.; SMITH, M.; COUPLAND, P.; OTTO, T.D.; HARRIS, S.R.; CONNOR, T.R.; GU, Y. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. **BMC genomics,** v. 13, n. 1, p. 341, 2012.
VINCENT, A.T.; DEROME, N.; BOYLE, B.; CULLEY, A.I.; CHARETTE, S.J. Next-generation sequencing (NGS) in the microbiological world: how to make the most of your money. **Journal of microbiological methods,** v. 138, p. 60-71, 2016.

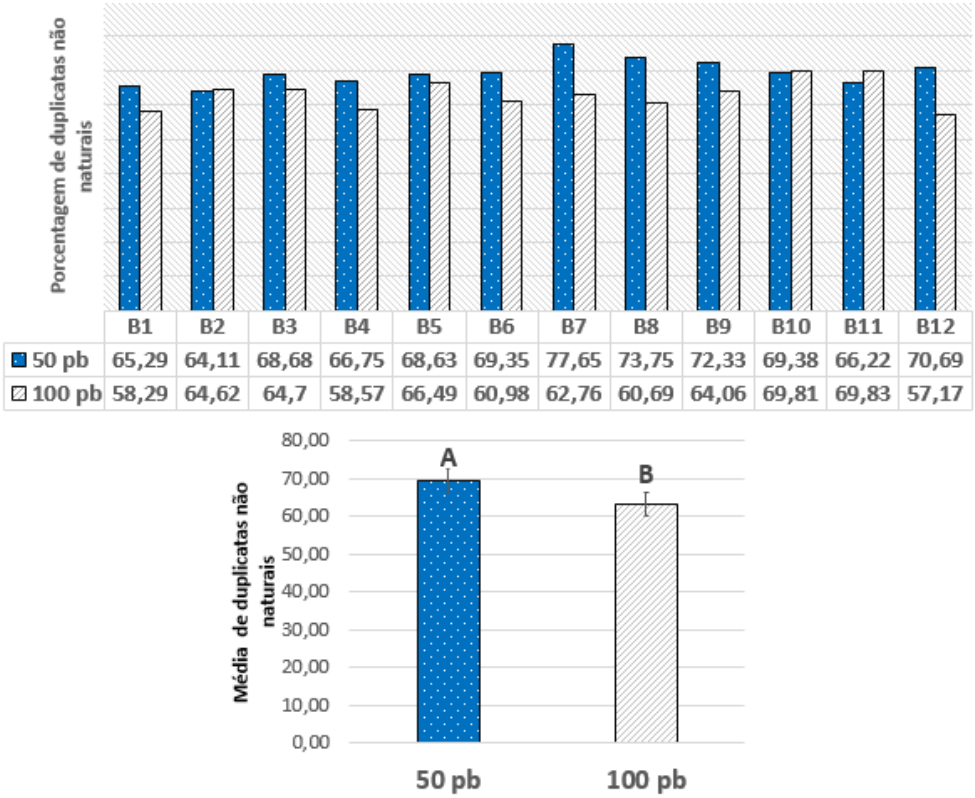

Figura 1. Taxa de duplicatas não naturais em bibliotecas de RNA-Seq com *reads* de 50 e 100 pb. Teste de Tukey ao nível de 5% de significância. Em azul escuro estão representadas as bibliotecas com *reads* de 50 pb e cinza, bibliotecas com *reads* de 100 pb. Letras maiúsculas comparam a média das bibliotecas geradas a partir de fragmentos de 50 e 100 pb através do teste de separação de médias Tukey ao nível de 5% de significância. Letras diferentes são estatisticamente diferentes. (Azul - *reads* com 50 pb e cinza – *reads* com 100 pb).