217

CORRELAÇÃO ENTRE A PORCENTAGEM DE ALINHAMENTOS ÚNICOS EM BIBLIOTECAS DE RNAseq DE SOJA SOB DEFICIT HIDRICO COM DIFERENTES TAMANHOS DE FRAGMENTO

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

[1]Bolsista CNPq, Embrapa Soja, Londrina, PR, renatafuganti@gmail.com; [2]Universidade Estadual de Londrina - UEL; [3]Embrapa Soja.

Nas últimas décadas, o sequenciamento em larga escala trouxe novas informações sobre o transcriptoma celular, o conjunto de todos os transcritos expressos e de suas quantidades, em uma célula em um dado estágio de desenvolvimento ou condição fisiológica. Analisar e entender os dados biológicos de um transcriptoma permite aos pesquisadores interpretar os elementos funcionais do genoma, revelar componentes moleculares de células e tecidos bem como fornece informações sobre as respostas a condições adversas, sejam bióticas ou abióticas (Martin; Wang, 2011; Pavlovich, 2017). A análise correta desta grande quantidade de dados depende, no entanto, do uso de ferramentas de bioinformática robustas, em um *pipeline* que forneça informações confiáveis e em respostas aos objetivos de cada pesquisa individualmente.

Assim, a partir dos dados brutos, que podem ser de diferentes tamanhos de fragmentos, dependendo da plataforma escolhida, deve-se inicialmente realizar a limpeza das sequências, pela identificação e remoção de sequências de adaptadores e filtragem de qualidade de cada posição de base sequenciada. Um dos softwares mais utilizado para esta etapa é o Trimmomatic (versão 0.36) (Bolger; Lohse; Usabel, 2014) e a seguir, para se visualizar a qualidade dos dados obtidos pós-limpeza pode-se utilizar o *software* FastQC (Andrews et al, 2010). Finalmente, após a limpeza das bibliotecas e a conferência da qualidade das sequências é realizada a etapa de alinhamento dos fragmentos sequenciados (*reads*) no genoma referência. Vários softwares estão disponíveis para realizar o alinhamento. Cabe ressaltar que as etapas iniciais de análise de bibliotecas de RNASeq são fundamentais para se obter dados robustos e confiáveis, que serão, muitas vezes, utilizados como informação básica na seleção de genes promissores e no desenho de estratégias de melhoramento, utilizando ferramentas de biotecnologia, visando o desenvolvimento de cultivares com caraterísticas de interesse agronômico, nutricional, de tolerância a estresse abióticos, resistência à estresses bióticos, entre outros.

Considerando-se que bibliotecas de RNASeq podem ser sequenciadas em fragmentos de diferentes tamanhos, objetivou-se com este trabalho, analisar comparativamente, bibliotecas de RNASeq de soja sequenciadas em fragmentos de 50 e 100 pb, para a recuperação de dados pós etapa de limpeza e alinhamento únicos no genoma referência. Busca-se com esta análise, avaliar se o tamanho dos *reads* pode apresentar relação com estes parâmetros.

Assim, bibliotecas de RNASeq obtidas a partir de experimentos de soja sob déficit hídrico, sequenciadas pela plataforma Illumina 1.9, com fragmentos *single-end* de 50 pb e de 100 pb, com uma cobertura de sequenciamento de 1X foram utilizadas nestas análises. A extração do RNA foi realizada utilizando o reagente Trizol® (Invitrogen, Califórnia, EUA) e a eliminação de possível DNA remanescente foi realizada utilizando-se o Kit DNAse I (Invitrogen, Califórnia, EUA). Em seguida, foi utilizado o Kit Ribominus™ plant (Invitrogen Califórnia, EUA) para retirada do rRNA (ribossômico). A qualidade do material genético extraído foi avaliada pelo *software* BioAnalyzer (Agilent Califórnia, EUA) e a concentração obtida pelo Qubit

(Thermofisher, Massachusetts, EUA). As amostras que apresentaram melhor qualidade e integridade (RIN≥ 8.0) (*Rna Integrity Number*) foram selecionadas para a síntese das bibliotecas. Os dados brutos foram disponibilizados para *download* e análise na plataforma *GeneSifter* (https://login.genesifter.net/). A qualidade das sequências foi avaliada usando o Software FastQC versão 0.11.5 (Andrews et al, 2010; Patel;Jain, 2012). A remoção dos adaptadores e das sequências de baixa qualidade foi realizada com o *software* Trimmomatic versão 0.36 (Bolger; Lohse; Usabel, 2014) padronizando os cortes da extremidade 3´ de 4 em 4 bases toda vez que a média de qualidade fosse inferior a 20 (*Phred Quality Score*, Q ≥ 20). Sequências com comprimento menor que 40 pb e maior que 90 pb também foram eliminadas para as bibliotecas sequenciadas em fragmentos de 100 pb e; para as bibliotecas sequenciadas em fragmentos de 50 pb foram mantidos fragmentos entre 35 a 45 pb. Os arquivos sem adaptadores e sequências de baixa qualidade foram reavaliados com o *software* FastQC afim de se verificar a eficiência da filtragem.

O genoma referência Willians 82 versão 2 utilizado nas análises foi obtido a partir do banco de dados Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Gmax). Para as análises de alinhamento, mapeamento e anotação gênica das bibliotecas foram utilizados o arquivo do genoma em formato FASTA (Gmax_275_v2.0.fa.gz) e o arquivo de anotação (Gmax_275_Wm82.a2.v1.gene.gff3.gz) no formato GFF3. O genoma referência foi indexado através do uso da ferramenta hisat2-build do *software* HISAT2 v.2.1.0. O alinhamento das bibliotecas também foi realizado através do *software* HISAT2 v.2.1.0, e somente os *reads* com alinhamentos únicos foram utilizados.

Os dados apresentaram distribuição normal segundo o teste de Shapiro-Wilk, sendo, portanto, realizada posteriormente a análise de variância ANOVA e teste de separação de médias Tukey ao nível de 5% de significância utilizando o *software* SASM-AGRI (Canteri et al., 2001). Foi realizado também o teste de correlação de Pearson ao nível de 5% de significância) utilizando o *software* Rstudio (Racine, 2012).

Os resultados indicaram uma maior taxa de recuperação de dados pós limpeza para as bibliotecas sequenciadas em fragmentos de 50 pb, em comparação às bibliotecas sequenciadas em fragmentos maiores, de 100 pb (Figura 1A). No entanto, a taxa de alinhamentos únicos para *reads* de 50pb foi menor se comparada à taxa obtida a partir de bibliotecas sequenciadas em fragmentos maiores (100 pb). Após a etapa de limpeza, nestas bibliotecas, os fragmentos resultantes possuem entre 40 e 90 pb, e são mais informativos (maior taxa de alinhamento único) apresentando até 20% mais alinhamentos únicos (Figura 1B). Alinhamentos únicos são importantes, principalmente, em análises de expressão diferencial, pois dão acurácia e diminuem informações ambíguas que podem levar a interpretação errôneas sobre o transcriptoma de uma célula em um dado estágio de desenvolvimento ou condição fisiológica (Busby et al., 2013; Volker; Small, 2017). O nível de correlação de Pearson entre a taxa de recuperação dos *reads* após a limpeza e a taxa de alinhamento únicos dos *reads* no genoma referência mostrou-se fortemente negativa (-0,95), indicando que a maior recuperação de dados após a limpeza é inversamente proporcional a qualidade do alinhamento. Estes dados corroboram a importância da escolha do tamanho dos fragmentos a serem sequenciados e das etapas inicias de análise de bibliotecas de RNASeq, determinantes para a obtenção de dados finais robustos e confiáveis.

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.
CANTERI, Marcelo 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.
MARTIN, J.A.; WANG, Z. Next-generation transcriptome assembly. **Nature Reviews Genetics,** v. 12, n. 10, p. 671-682, 2011.
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.
PAVLOVICH, M. Computing in Biotechnology: Omics and Beyond. **Trends Biotechnology**, v. 35, n. 6, p. 479-480, 2017.
RACINE, J.S. RStudio: A Platform-Independent IDE for R and Sweave. **Journal of Applied Econometrics**, v. 27, n. 1, p. 167-172, 2012.

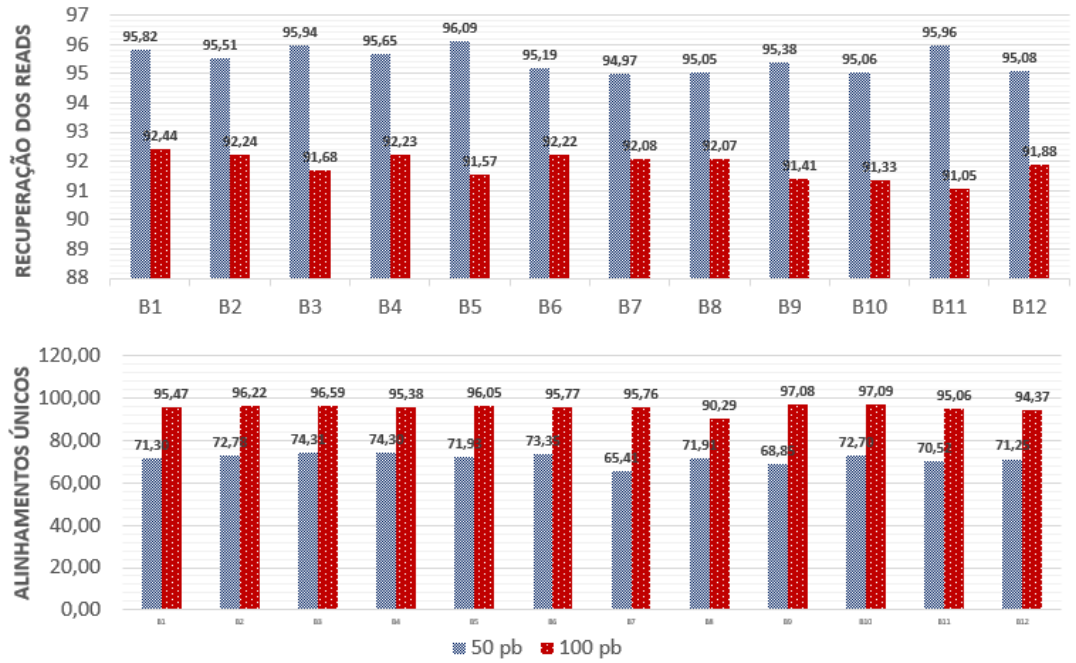

Figura 1: A) Percentagem de recuperação de fragmentos, em 12 bibliotecas sequenciadas em fragmentos de 50 pb (em azul) e 12 bibliotecas sequenciadas em fragmentos 100 pb (em vermelho), após a etapa de limpeza realizada pelo *software* Trimmomatic (versão 0.36). B) Percentagem de alinhamento únicos no genoma referência em 12 bibliotecas sequenciadas em fragmentos de 50 pb (em azul) e 12 bibliotecas sequenciadas em fragmentos 100 pb (em vermelho).