ANÁLISE COMPARATIVA DE FERRAMENTAS DE BIOINFORMÁTICA PARA MONTAGEM DE GENOMAS COM TECNOLOGIA DE SEQUENCIAMENTO DE NOVA GERAÇÃO

DANIEL DA ROSA FARIAS[1]; LEOMAR GUILHERME WOYANN[1]; LUCIANO CARLOS DA MAIA[2]; ANTONIO COSTA DE OLIVEIRA[2]

[1]Centro de Genômica e Fitolhemelhoramento (CGF) - Faculdade de Agronomia Eliseu Maciel / UFPel. e-mail: fariasdr@gmail.com
[2]Centro de Genômica e Fitolhemelhoramento (CGF) - Faculdade de Agronomia Eliseu Maciel / UFPel. e-mail: acostol@terra.com.br

1. INTRODUÇÃO

A disponibilização de dados obtidos com o sequenciamento de genomas ou de regiões genômicas e a análise de transcriptomas possibilitou um grande avanço nas pesquisas em melhoramento genético de plantas (VARSHNEY, 2009). Desta forma, é possível a detecção e utilização da variabilidade genética de forma mais eficiente. Neste sentido, várias espécies de plantas apresentam os genomas completamente sequenciados, entre os quais podemos citar o arroz (*Oryza sativa* L.) (IRGSP, 2005) e *Arabidopsis thaliana* (Arabidopsis Genome Initiative, 2000).

Devido ao crescente interesse no estudo de genomas, uma nova geração de tecnologias de sequenciamento emergiu, e entre estas, a plataforma Solexa, da Illumina, já possui ampla utilização em todo o mundo. Devido à redução de custos e o aumento da capacidade de sequenciamento, essas novas plataformas são eficazes para a utilização rotineira em projetos de sequenciamento e resequenciamento de genomas individuais, para detecções de variações entre genomas-alvo e de referência (SERVICE, 2006). Estas plataformas geram leituras de pequenas regiões do genoma e com isso, a etapa de montagem destas, em sequências contíguas (*contigs*) torna-se um dos maiores desafios para estas novas tecnologias. Isto ocorre devido à presença de regiões repetitivas nos genomas (SCHATZ et al., 2010) e principalmente, às regiões repetidas em *tandem* (NAVAJAS-PÉREZ & PATERSON, 2009).

Segundo Carvalho e Silva (2010), as novas plataformas de sequenciamento, em um futuro próximo, revolucionarão o conhecimento sobre o genoma das plantas, principalmente no que concerne o estudo de variantes alélicas, SNPs (*Single Nucleotide Polymorphism*), desenvolvimento de marcadores para seleção assistida e a clonagem baseada em mapeamento, representando assim, uma importante ferramenta no melhoramento vegetal. Porém, devido à dificuldade de montagem das leituras produzidas por essas plataformas, analisar, compreender e determinar qual programa se comportará melhor com determinados grupos de dados, é uma tarefa importante para que seja possível aperfeiçoar o processo de sequenciamento de genomas.

Desta forma, com a finalidade de fornecer informações para futuros trabalhos relacionados com sequenciamento genético utilizando uma plataforma Illumina, o objetivo desse trabalho é avaliar o desempenho de programas montadores de genomas, e inferir sobre sua cobertura e qualidade das sequências geradas.

2. MATERIAL E MÉTODOS

Um total de seis programas montadores de genomas foram testados: Velvet (www.ebi.ac.uk/~zerbino/velvet), MIRA (www.chevreux.org/projects_mira.html), AbySS (www.bcgsc.ca/platform/bioinfo/software/abyss/), SOAPdenovo (soap.genomics.org.cn/soapdenovo.html), QSRA (mocklerlab.org/tools/2) e Ray (denovoassembler.sourceforge.net/). Para testá-los foi utilizado como sequência de referência 5.000.000 de pares de bases (pb) do cromossomo 1 de arroz, que foi obtido do sítio RAP-DB (http://rapdb.dna.affrc.go.jp/). A partir dessa sequência foram geradas 670.000 leituras *pair-end* com tamanho de 75 pb e inserto máximo de 350 pb. Foi simulado um sequenciamento com cobertura de 20x e foi realizada através do desenvolvimento de um programa em linguagem PERL, com alguns módulos do pacote BioPerl (http://www.bioperl.org), desenvolvidos especificamente para análises em bioinformática.

Para comparar o desempenho de cada ferramenta de montagem foram calculados e analisados o tamanho total da montagem (em pb), o número total de contigs gerados, o tamanho médio dos contigs (em pb), o comprimento do maior contig (em pb), valor N50 (em pb) designa que 50% do total de bases montadas estão contidas em contigs de tamanho N ou maiores, e L50 que é o número dos maiores contigs em que estão contidos 50% do total de bases montadas. O programa Mauve (DARLING et al., 2010) foi utilizado para identificar e mensurar quanto do total dos *contigs* gerados foram incorretamente ordenados em relação a sequência de referência, através dos algoritmos *progressiveMauve* e *Mauve Contig Mover* (RISSMAN et al., 2009).

3. RESULTADOS E DISCUSSÃO

Para muitos estudos biológicos, sequências de maiores tamanhos são necessárias. Na prática, os *contigs* gerados pelas ferramentas de montagem são separados por espaços (*gaps*), devido à presença de fragmentos repetidos no genoma.

Através dos resultados gerados pelas diferentes ferramentas utilizadas, foi possível identificar diferenças entre os programas, através do número e tamanho dos contigs, cobertura obtida e homologia posicional com a sequência de referência. Apenas os *contigs* de tamanhos iguais ou superiores a 100 pb foram considerados nas análises (Tabela 1). Quanto ao número total de bases geradas, três programas se destacaram com valores mais elevados, sendo que o programa SOAPdenovo foram geradas 4.980.933 pb, pelo programa MIRA e AbySS foram geradas 4.878.809 pb e 4.887.009 pb, respectivamente. Os programas QSRA e Ray apresentaram os menores valores. A ferramenta Velvet obteve um valor intermediário em relação às demais. Esses valores são importantes porque fornecem um indicativo da cobertura do genoma que foi atingida pela montagem realizada pelos programas.

Com relação ao tamanho e número de *contigs*, identificados pelo N50 e L50, o programa MIRA obteve destaque sobre as demais ferramentas apresentando para estas variáveis valores de 51.602 pb e 23 *contigs*, respectivamente. Em contraposição, o programa QSRA apresentou para N50 um resultado de 386 pb e L50 de 3680 contigs. Pode-se observar que o programa MIRA produziu contigs maiores e em menores quantidades. O programa QSRA apresentou uma tendência inversa. Esses cálculos normalmente são utilizados para avaliar a

qualidade das montagens de genomas porque indicam quanto do genoma está coberto por contigs relativamente grandes (SCHATZ et al., 2010), e de uma maneira geral, o ideal é que um programa de montagem de genomas produza *contigs* maiores e em menor número.

Tabela 1- Estatística para os *contigs* (≥ 100pb) para cada ferramenta de montagem.

Estatísticas	Velvet	MIRA	AbySS	SOAPdenovo	QSRA	Ray
Número total	1141	733	1748	3943	12617	5151
Total de pb geradas	4557559	4878809	4887009	4980933	4322462	4319173
Tamanho médio (bp)	3.994,4	6.655,9	2.795,8	1.263,2	342,6	838,5
Maior contig (bp)	158.967	349.914	51.619	16.958	3.414	6.685
N50	27.727	51.602	7.961	3.061	386	1.161
L50	48	23	173	490	3680	1173

Após a ordenação dos *contigs* de acordo com a sequência de referência por meio do programa Mauve (Tabela 2) pode-se verificar que os programas SOAPdenovo e QRSA apresentaram os melhores resultados quanto a porcentagem de pb alinhadas sem conflito de ordem e/ou posição, com 100% das sequências sem conflito. O programa MIRA, que apresentou os melhores resultados em relação a N50 e L50, apresentou altos níveis de conflitos em relação à ordem e/ou posição dos *contigs* gerados quando alinhado com a sequência de referência. Esta análise é um indicativo da qualidade das sequências geradas, bem como da acurácia do programa de montagem de genomas. Esta diferença está relacionada à como os algoritmos utilizados pelos programas montadores analisam as sequências repetitivas no genoma.

Tabela 2 – Resultado de um alinhamento realizado pelo programa Mauve, entre a sequência de referência e *contigs* gerados.

Programa	Pares de bases com conflito de posição	% pb corretos
MIRA	3391802	30,48
SOAP	0	100,00
AbySS	2822575	42,24
QSRA	0	100,00
Velvet	3763444	17,42
Ray	1870634	56,69

4. CONCLUSÕES

Dentre os programas testados nesse estudo a ferramenta SOAPdenovo é o programa montador de genomas que melhores resultados apresentou em relação a quantidade de bases montadas e ao correto posicionamento destas sequências em relação à sequência de referência utilizada neste trabalho.

5. REFERÊNCIAS BIBLIOGRÁFICAS

Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. **Nature**, v.408, n.6814, p.796–815, 2000.

CARVALHO, M.C.C.G.; SILVA, D.C.G. Sequenciamento de DNA de nova geração e suas aplicações na genômica de plantas. **Ciência Rural**, v.40, p.735-744, 2010.

DARLING, A.C.E.; MAU, B.; PERNA, N. T. progressivemauve: Multiple genome alignment with gene gain, loss and rearrangement. **PLoS ONE**, v.5, n.06, 2010.

INTERNATIONAL RICE GENOME SEQUENCING PROJECT. The map-based sequence of the rice genome. **Nature**, v.1, n.7052,p.793-800, 2005.

NAVAJAS-PEREZ, R.; PATERSON, A.H. Patterns of tandem repetition in plant whole genome assemblies. **Molecular Genetics and Genomics**, v.281, n.6, p.579–590, 2009.

RISSMAN, A.I.; MAU, B.; BIEHL, B.S.; DARLING, A.E.; GLASNER, J. D.; PERNA, N.T. Reordering contigs of draft genomes using the mauve aligner. **Bioinformatics**, v.25, n.16, p.2071-2073, 2009.

SCHATZ, M.C.; DELCHER, A.L.; SALZBERG, S.L. Assembly of large genomes using second-generation sequencing. **Genome Research**, v.20, p.1165–1173, 2010.

SERVICE, R.F. The race for the $1000 genome. **Science**, v.311, p.1544–1546, 2006.

VARSHNEY, R.K.; NAYAK, S.N.; MAY, G.D.; JACKSON, S.A. Next-generation sequencing technologies and their implications for crop breeding. **Trends in Biotechnology**, v.27, n.9, p.522-530, 2009.