MILHARES DE SNPS GENOTIPADOS POR SEQUENCIAMENTO DE ALTO DESEMPENHO (GBS - "GENOTYPING BY SEQUENCING") EM ESPÉCIES DE *EUCALYPTUS*

Danielle Assis de Faria
Pos-doutoranda, Embrapa Recursos Genéticos e Biotecnologia, Brasília, DF;
danyafp@gmail.com;

Dario Grattapaglia
Pesquisador-Laboratório de Genética Vegetal, Embrapa Recursos Genéticos e Biotecnologia, Brasília, DF dario.grattapaglia@embrapa.br

RESUMO – Marcadores moleculares têm sido usados em diversas áreas, constituindo de variações na sequência de DNA, detectadas por meio de diferentes técnicas, como a reação em cadeia da polimerase (PCR), ou corte com enzimas de restrição, seguidas de eletroforese, ou o sequenciamento direto da molécula de DNA. Estas técnicas são analisadas no presente trabalho, em contraposição à metodologia de genotipagem por sequenciamento GbS ("genotyping by sequencing"). A GbS abre perspectivas fenomenais para a análise genética em qualquer espécie e foi testada em duas espécies de *Eucalyptus* L'Hér., em que apresentou um bom desempenho. Os resultados obtidos demonstraram o sucesso da construção da biblioteca e a capacidade de análise simultânea de indivíduos. A GbS se caracteriza por não demandar o desenvolvimento prévio de marcadores, se basear em reagentes universais, permite o sequenciamento de centenas de amostras simultaneamente e a quantidade inicial de DNA necessário para cada amostra é pequena.

Palavras-chave: código de barras, genotipagem por sequenciamento, marcador molecular, "multiplex", NGS.

Marcadores moleculares podem ser definidos como variações existentes na sequência de DNA entre indivíduos, detectadas por meio de diferentes técnicas analíticas, como a reação em cadeia da polimerase (PCR, "polymerase chain reaction"), ou corte com enzimas de restrição, seguidas de eletroforese, ou o sequenciamento direto da molécula de DNA. Marcadores moleculares vêm sendo utilizados com sucesso há pelo menos duas décadas, para responder a diversas questões nas áreas de genética de populações, evolução, ecologia, filogenia, taxonomia, melhoramento genético, entre outras (Bernardo 2008; Jones *et al.* 2009; Siol *et al.* 2010; Perez-de-Castro *et al.* 2012).

Entre os vários tipos de marcadores moleculares, os marcadores baseados na análise direta de polimorfismos de base individual, chamados SNP ("single nucleotide polymorphism"), vêm sendo cada vez mais utilizados em plantas (Ganal *et al.* 2009). Isto ocorreu à medida que o sequenciamento de DNA se tornou mais acessível, o que permitiu o descobrimento de um grande número de SNPs entre indivíduos. SNPs são abundantes, distribuídos por todo genoma de um organismo e geralmente bialélicos, ou seja, na grande maioria das vezes, a base polimórfica apresenta uma dos dois pares de bases A, C, T ou G. Marcadores SNPs vêm substituindo a utilização de outras técnicas indiretas de detecção de polimorfismos, tais como RFLP ("amplified fragment length polymorphism") e AFLP ("restriction fragment length polymorphism"), aumentando consideravelmente a capacidade de genotipagem proporcionada por microssatélites. Conjuntos de SNPs vêm sendo desenvolvidos para várias espécies, embora este processo de descoberta e validação de marcadores SNPs ainda seja relativamente caro e demorado. Além disso, em vista da elevada especificidade das posições polimórficas selecionadas para análise, um conjunto de marcadores SNPs desenvolvido para uma espécie tende a ser pouco informativo em outra espécie próxima (Grattapaglia *et al.* 2011).

Nos últimos anos, o avanço das novas tecnologias de sequenciamento, também conhecidas como sequenciamento de próxima geração, (NGS, "next generation sequencing"), tornaram o sequenciamento de DNA ainda mais accessível e econômico (Metzker 2010), abrindo novas perspectivas para o estudo da biologia de plantas (Brautigam & Gowik 2010; Egan *et al.* 2012). Este avanço fez com que o custo do descobrimento de marcadores moleculares, principalmente SNPs, caísse vertiginosamente ao se utilizar técnicas que permitem sequenciar, de forma localizada, trechos específicos do genoma (van Orsouw *et al.* 2007; Baird *et al.* 2008; van Tassell *et al.* 2008; Hyten *et al.* 2010; Davey *et al.* 2011; Elshire *et al.* 2011).

As metodologias NGS se baseiam na redução da complexidade do genoma, ou seja, a amostragem de apenas uma pequena porção do mesmo, a qual é sequenciada, em paralelo, em dezenas ou centenas de indivíduos, permitindo detectar posições polimórficas nas sequências amostradas. Esta amostragem é realizada de maneira fácil, cortando o genoma de vários indivíduos simultaneamente com enzimas de restrição, seguida do sequenciamento massal dos fragmentos gerados após a ligação de adaptadores nas suas extremidades e, finalmente, da análise das sequências com programas computacionais específicos que permitem a detecção de SNPs entre os indivíduos. A variedade

de enzimas disponíveis e suas sensibilidades ou não a regiões hiper-metiladas do genoma fazem com que essa abordagem seja versátil. Para uma mesma espécie, o pesquisador pode utilizar diferentes enzimas para amostrar distintas regiões do genoma e, assim, descobrir milhares ou centenas de milhares de SNPs ao longo de todo o genoma.

Com a recente evolução na quantidade de sequências produzidas por NGS, esta abordagem de descobrimento de SNPs via re-sequenciamento massal de trechos do genoma permite hoje não apenas o descobrimento de SNPs, mas a própria genotipagem de dezenas de indivíduos simultaneamente. Esta metodologia é denominada GbS ("genotyping by sequencing"), ou genotipagem por sequenciamento (Davey *et al.* 2011; Elshire *et al.* 2011), e abre perspectivas fenomenais para a análise genética em qualquer espécie. GbS se caracteriza por não demandar o desenvolvimento prévio de marcadores e se basear em reagentes universais, assim como a técnica de RAPD, que permitiu, na sua época, um avanço notável na capacidade de realizar análises genéticas moleculares em qualquer espécie.

GbS envolve a digestão do DNA genômico do organismo alvo com apenas uma enzima de restrição (Elshire *et al.* 2011) NY 14853 USA
Cornell Univ, Inst Genom Divers, Ithaca, NY 14853 USA
Cornell Univ, Computat Biol Serv Unit, Ithaca, NY USA
ARS, Hard Winter Wheat Genet Res Unit, USDA, Manhattan, KS USA
ARS, Plant Soil & Nutr Res Unit, USDA, Ithaca, NY USA</ auth-address><titles><title>A robust, simple genotyping-by-sequencing (GbS, ou uma combinação delas (Poland *et al.* 2012; Sansaloni *et al.* 2011), gerando uma biblioteca de fragmentos com tamanho entre 200 e 400 bp (pares de bases). Essa metodologia traz como principal diferença em relação às técnicas anteriores, que visavam apenas a descoberta de SNPs, o fato de que, para cada indivíduo a ser genotipado, uma sequência indexadora ("barcode", código de barras) de 4 a 9 pares de base de DNA é incluída no adaptador. Desta forma, cada amostra a ser genotipada terá uma sequência "barcode" única, o que permite que sejam sequenciadas centenas de amostras simultaneamente, em um processo denominado "multiplex" (múltiplo). As sequências de 75 a 100 bases, geradas para cada amostra, podem ser identificadas pelo seu "barcode" e, assim, recuperadas individualmente, permitindo a genotipagem de SNPs. A GbS fornece marcadores do tipo presença ou ausência, resultantes do corte ou não pela enzima de restrição, gerando ou não um fragmento sequenciado. Além disso, também são obtidos marcadores SNPs codominantes, quando todos os indivíduos apresentarem o fragmento sequenciado, mas houver variação em uma ou mais bases na sequência entre os indivíduos.

O DNA a ser utilizado no protocolo de genotipagem por GbS deve ter alta qualidade, estando livre de contaminação por compostos orgânicos. Entretanto, sendo o sistema em "multiplex" com muitas amostras analisadas simultaneamente, a quantidade inicial de DNA de cada uma delas é pequena. São necessários apenas 100 ng de DNA de cada indivíduo para a digestão com enzimas e apenas uma pequena fração deste DNA é efetivamente utilizada no final, para o sequenciamento. Nos trabalhos publicados até o momento utilizando essa metodologia, foram genotipadas, simultaneamente, até 96 amostras diferentes de linhagens de cevada e milho (Elshire *et al.* 2011; Poland *et al.* 2012).

Visando testar e validar a metodologia de GbS para espécies do gênero *Eucalyptus* L'Hér., realizamos um estudo envolvendo duas espécies (*E. grandis* W. Hill e *E. globulus* Labill.), utilizando o procedimento descrito por Elshire *et al.* (2011) e sequenciamento realizado no Genômica-DF (Centro de Sequenciamento de Alto desempenho do Distrito Federal). Neste experimento piloto, foram analisados 24 indivíduos de cada uma das duas espécies. A biblioteca dos indivíduos de *E. grandis* foi analisada em réplica, com a finalidade de estimar a consistência e repetibilidade de marcadores entre diferentes preparações de GbS. Nas análises, foram adotados alguns parâmetros específicos para a declaração de genótipos, que podem variar de acordo com a espécie e número de indivíduos analisados. Foi estabelecido um mínimo de 6 e posteriormente, um mínimo de 12 sequências alinhadas na posição do SNP para declarar o genótipo do SNP no indivíduo. Foi estabelecido também que um SNP, para ser considerado como marcador molecular, deveria ser genotipado em, pelo menos, 75% dos indivíduos ("call rate" 75%) e apresentar uma frequência do alelo mais raro (MAF, "minor allele frequency") maior ou igual a 5%.

Os resultados obtidos no experimento demonstraram o sucesso da construção da biblioteca e a capacidade de análise simultânea de indivíduos (Faria *et al.* 2012). Em uma canaleta de sequenciamento da biblioteca de fragmentos dos indivíduos de *E. grandis*, foram obtidas cerca de 45.700.000 sequências, das quais apenas cerca de 4,3% não possuíam a construção correta com a presença de um dos 24 "barcodes" utilizados. Para a biblioteca de *E. globulus*, os resultados foram equivalentes com apenas 4,75% das sequências geradas sem barcode. Uma análise preliminar revelou um número expressivo de marcadores SNPs polimórficos atendendo a todas as condições colocadas anteriormente, consistente com o elevado nível de diversidade nucleotídica no gênero *Eucalyptus* (Grattapaglia *et al.* 2012). Em linhas gerais, após todo o processo de análise e filtragem de sequências e marcadores, foram genotipados 89.098 SNPs em *E. globulus* e 87.570 SNPs em *E. grandis*. Mesmo aumentando o rigor analítico para um "call rate" de 95%, ou seja, permitindo apenas 5% de dados faltantes, 38.415 e 42.305 SNPs foram genotipados para as duas espécies respectivamente.

O número médio de marcadores SNPs genotipados em cada um dos 11 cromossomos foi de 8.099 para *E. globulus* e 8.042 para *E. grandis*. Nas duas espécies o cromossomo 8 foi o que apresentou maior número de marcadores (cerca de 11.000) e o 4 (cerca de 5.000) foi o que apresentou o menor número. Quando as duas réplicas de *E. grandis* foram analisadas, 75% dos marcadores puderam ser genotipados, consistentemente, em ambas as análises, mostrando que a repetibilidade de amostragem de sequência entre canaletas independentes no sequenciador é próxima à estimativa de 80% informada pelo fabricante.

Como conclusão, o GBS apresentou um bom desempenho em *Eucalyptus*, um gênero com espécies de elevada heterozigosidade, permitindo a análise de um número muito alto de marcadores distribuídos ao longo de todo o genoma. Em comparação com técnicas anteriores, que em geral permitiam a análise de poucas centenas de marcadores microssatélites ou AFLP, a técnica de GbS aumentou em três ordens de magnitude o número de marcadores analisados, com uma repetibilidade consistente com o que se espera com as tecnologias NGS de sequenciamento. Esta técnica abre, assim, perspectivas formidáveis para estudos de diversidade genética em populações naturais e bancos de germoplasma, seleção genômica em programas de melhoramento, estudos de genômica populacional, filogenia e filogeografia, com uma resolução analítica incomparavelmente superior àquelas utilizada até o momento.

REFERÊNCIAS BIBLIOGRÁFICAS

BAIRD N.A.; ETTER P.D.; ATWOOD T.S.; CURREY M.C.; SHIVER A.L.; LEWIS Z.A.; SELKER E.U.; CRESKO W.A.; JOHNSON E.A. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. **Plos One** 3(10):e3376, 2008.

BERNARDO R. Molecular markers and selection for complex traits in plants: Learning from the last 20 years. **Crop Science** 48:1649-1664, 2008.

BRAUTIGAM A.; GOWIK U. What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research. **Plant Biology** 12:831-841, 2010.

DAVEY J.W.; HOHENLOHE P.A.; ETTER P.D.; BOONE J.Q.; CATCHEN J.M.; BLAXTER M.L. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. **Nature Reviews Genetics** 12:499-510, 2011.

EGAN A.N.; SCHLUETER J.; SPOONER D.M. Applications of Next-Generation Sequencing in Plant Biology. **American Journal of Botany** 99:175-185, 2012.

ELSHIRE R.J.; GLAUBITZ J.C.; SUN Q.; POLAND J.A.; KAWAMOTO K.; BUCKLER E.S.; MITCHELL S.E. A robust, simple genotyping-by-sequencing (GbS) approach for high diversity species. **Plos One** 6 (5):e19379, 2011.

FARIA D.A.; TANNO P.; REIS A.; MARTINS A.; FERREIRA M.E.; GRATTAPAGLIA D. Genotyping-by-Sequencing (GbS) the highly heterozygous genome of *Eucalyptus* provides large numbers of high quality genome-wide SNPs. **Plant; Animal Genomes XX Conference**, San Diego, p Abstract P0521, 2012.

GANAL M.W.; ALTMANN T.; RODER M.S. SNP identification in crop plants. **Current Opinion in Plant Biology** 12:211-217, 2009.

GRATTAPAGLIA D.; SILVA O.B.; KIRST M.; DE LIMA B.M.; FARIA D.A.; PAPPAS G.J. High-throughput SNP genotyping in the highly heterozygous genome of *Eucalyptus*: assay success, polymorphism and transferability across species. **Bmc Plant Biology** 11:65, 2011.

GRATTAPAGLIA D.; VAILLANCOURT R.; SHEPHERD M.; THUMMA B.; FOLEY W.; KULHEIM C.; POTTS B.; MYBURG A. Progress in Myrtaceae genetics and genomics: *Eucalyptus* as the pivotal genus. **Tree Genetics; Genomes** d.o.i. 10.1007/s11295-012-0491-x, 2012.

HYTEN D.L.; CANNON S.B.; SONG Q.J.; WEEKS N.; FICKUS EW.; SHOEMAKER R.C.; SPECHT J.E.; FARMER A.D.; MAY G.D.; CREGAN P.B. High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. **Bmc Genomics** 11:38, 2010.

JONES N.; OUGHAM H.; THOMAS H.; PASAKINSKIENE I. Markers and mapping revisited: finding your gene. **New Phytologist** 183:935-966, 2009.

METZKER M.L. Applications of Next-Generation Sequencing Sequencing Technologies - the Next Generation. **Nature Reviews Genetics** 11:31-46, 2010.

PEREZ-DE-CASTRO A.M.; VILANOVA S.; CANIZARES J.; PASCUAL L.; BLANCA J.M.; DIEZ M.J.; PROHENS J.; PICO B. Application of Genomic Tools in Plant Breeding. **Current Genomics** 13:179-195, 2012.

POLAND J.A.; BROWN P.J.; SORRELLS M.E.; JANNINK J.L. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. **PLoS One** 7:e32253, 2012.

SANSALONI C.; PETROLI C.; JACCOUD D.; CARLING J.; DETERING F.; GRATTAPAGLIA D.; KILIAN A. Diversity Arrays Technology (DArT) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of *Eucalyptus*. **BMC Proceedings** 5:P54, 2011.

SIOL M.; WRIGHT S.I.; BARRETT S.C.H. The population genomics of plant adaptation. **New Phytologist** 188:313-332, 2010.

VAN ORSOUW N.J.; HOGERS R.C.J.; JANSSEN A.; YALCIN F.; SNOEIJERS S.; VERSTEGE E.; SCHNEIDERS H.; VAN DER POEL H.; VAN OEVEREN J.; VERSTEGEN H.; VAN EIJK M.J.T. Complexity Reduction of Polymorphic Sequences (CRoPS (TM)): A Novel Approach for Large-Scale Polymorphism Discovery in Complex Genomes. **Plos One** 2(11):e1172, 2007.

VAN TASSELL C.P.; SMITH T.P.L.; MATUKUMALLI L.K.; TAYLOR J.F.; SCHNABEL R.D.; LAWLEY C.T.; HAUDENSCHILD C.D.; MOORE S.S.; WARREN W.C., SONSTEGARD T.S. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. **Nature Methods** 5:247-252, 2008.