DESENVOLVIMENTO E VALIDAÇÃO DE CHIP DE DNA PARA GENOTIPAGEM EM ESCALA DE ACESSOS DO PROGRAMA DE CONSERVAÇÃO DE GERMOPLASMA E DE MELHORAMENTO GENÉTICO DE ARROZ

Márcio Elias Ferreira[1]; Alexandre Magalhães Martins[2]; Paulo Hideo Nakano Rangel[3]

Palavras-chave: *Oryza sativa*, SNP, marcadores moleculares, germoplasma

INTRODUÇÃO

O emprego de tecnologia genômica em apoio a programas de conservação de germoplasma e de melhoramento genético de diferentes culturas agrícolas é crescente em todo o mundo. Marcadores moleculares, especialmente, têm sido aplicados em atividades de rotina na conservação da diversidade genética de espécies de interesse econômico e no desenvolvimento de novas variedades de plantas. O uso de tecnologias genômicas facilita as rotinas de conservação e desenvolvimento varietal, possibilita a seleção de plantas de interesse com base na análise simultânea do genótipo em vários locos e contribui para o aumento da eficiência dos programas.

A análise de polimorfismo de DNA em grande número de plantas de arroz a cada geração, como requer, por exemplo, os programas de melhoramento genético, demanda metodologias de genotipagem com as seguintes características: (1) capacidade multiplex, que se traduz na possibilidade de genotipar uma grande quantidade de locos de marcadores moleculares ao mesmo tempo; (2) rapidez analítica, o que possibilita a genotipagem de um grande número de plantas em pouco tempo, possibilitando a identificação de plantas com o genótipo desejado antes do florescimento, de forma a permitir a seleção, o cruzamento e a recombinação na mesma geração; (3) robustez analítica, gerando genotípicos fidedignos, com alto nível de acurácia, maximizando o sucesso da seleção de plantas com combinações alélicas desejadas; (4) baixo custo de genotipagem por planta, visto que a quantidade de plantas analisadas atinge a casa de quatro dígitos em cada geração de melhoramento.

Estas características requerem uma adaptação das rotinas de genotipagem no laboratório para uso em escala. Isto significa uma intervenção metodológica que possibilite a genotipagem de um grande número de amostras de DNA, em milhares de locos do genoma, em um curto período de tempo e a um baixo custo por amostra. Isto só é possível com a automação de etapas, desde a extração de DNA até a fase final de análise dos dados. Um ponto crítico de todo o processo é, sem dúvida, o aumento da eficiência da etapa de genotipagem de grande número de amostras em milhares de locos polimórficos, o que exige uma grande capacidade multiplex do marcador selecionado. A necessidade de automação tem nos marcadores SNP uma grande opção.

Um marcador SNP (*Single Nucleotide Polymorphism*) pode ser definido como um sítio do DNA onde foi observada a substituição de uma única base entre amostras de indivíduos de uma mesma população (Risch & Merikangas, 1996). Um SNP é entendido ainda como uma posição no DNA onde diferentes sequências alternativas (alelos) co-existem em indivíduos em uma população, e o alelo menos frequente possui frequência maior que 1% (Brookes, 1999). O grande marco para o emprego em rotina de marcadores SNP foi o desenvolvimento de novas tecnologias de sequenciamento de DNA, conhecidas como

[1] PhD em Genética e Melhoramento de Plantas, Embrapa Recursos Genéticos e Biotecnologia.
[2] Doutorando em Biologia Molecular, Universidade de Brasília - UnB
[3] Doutor em Genética e Melhoramento de Plantas, Embrapa Arroz e Feijão.

tecnologias de sequenciamento de nova geração ou NGS (*Next Generation Sequencing*). Essas tecnologias são capazes de produzir uma quantidade enorme de dados de sequência, da ordem de bilhões de bases, em um curto intervalo de tempo, em sistemas altamente automatizados, e a um baixo custo por base sequenciada (Ganal, 2009). Dessa forma, essas tecnologias aumentaram significativamente a velocidade de descoberta de SNPs e, consequentemente, estimularam a sua aplicação em diversas áreas, inclusive na conservação de germoplasma e no melhoramento genético de plantas.

Neste trabalho, dados de sequenciamento, montagem e alinhamento de genomas de oito variedades de arroz foram empregados na descoberta e seleção de milhares de sítios SNP. As informações em torno das sequências que flanqueiam os SNPs selecionados foram utilizadas para desenvolver um chip de DNA para uso em genotipagem em escala. Este chip foi avaliado e validado para a genotipagem de amostras do germoplasma de arroz.

MATERIAL E MÉTODOS

Oito variedades de arroz *japonica* tropical (Chorinho, Puteca, IAC 165, BRS Primavera, Ligeiro, Moroberekan, Azucena e Catetão) foram selecionadas para sequenciamento genômico. Amostras de DNA foram extraídas de folhas jovens usando o protocolo padrão CTAB, com modificações (Ferreira & Grattapaglia, 1998). O sequenciamento foi realizado a partir de bibliotecas genômicas de fragmentos, amplificadas para PCR em ponte (*bridge PCR*), possibilitando o sequenciamento por síntese de fragmentos de pontas pareadas através de um sequenciador Illumina GAII. As bibliotecas genômicas foram preparadas de acordo com as instruções do fornecedor (www.illumina.com). As amostras de DNA foram inicialmente fragmentadas por nebulização e as extremidades 3´ ligadas a bases A. Produtos de ligação foram separados em gel de agarose 1% e fragmentos com inserto de ~200 pb foram purificados para reações de sequenciamento. Os dados de sequenciamento das oito variedades de arroz foram inicialmente examinados para identificação de sequências de DNA cloroplástico e mitocondrial, além de potenciais sequências contaminantes (fungo, bactéria, vírus) para a exclusão de sequências não-nucleares e/ou exógenas. Adaptadores de sequenciamento Illumina e sequências de baixa qualidade foram eliminadas utilizando o software CLC Genomics Workbench 4.1, CLC Bio, Aarhus, Denmark. O programa foi empregado na montagem do genoma das oito variedades usando a sequência do genoma da cultivar Nipponbare (v. MSU 6.1) como referência. O alinhamento das pseudomoléculas dos genomas sequenciados possibilitou a identificação de sítios de polimorfismo de base única (SNPs) distribuídos por todo o genoma de arroz. Foram considerados na análise apenas SNPs com cobertura mínima 20x em cada combinação pareada. Entre os critérios de seleção de sítios SNP incluiu-se o polimorfismo em três de quatro pareamentos utilizados na análise (Chorinho x Puteca; IAC 165 x Primavera; Azucena x Catetão; Ligeiro x Moroberekan). Os SNPs selecionados foram listados e identificados em arquivo contendo informações de localização, qualidade, e sequência de região flanqueadora (mínimo de 100 pb flanqueando cada lateral do sítio) para síntese de um chip de DNA para reações em paralelo em esferas de sílica.

RESULTADOS E DISCUSSÃO

O sequenciamento NGS possibilitou a obtenção de grande quantidade de dados de sequência para as oito variedades de arroz selecionadas. Foram obtidas sequências totalizando entre 4 e 8 bilhões de pares de bases para cada variedade, possibilitando a montagem de mais de 94% do genoma em relação à referência Nipponbare, com um nível de cobertura genômica acima de 11x (Tabela 1).

Tabela 1 – Parâmetros de sequenciamento, montagem e cobertura do genoma de oito variedades de arroz *japonica* tropical submetidas a sequenciamento NGS.

Variedade	No. Fragmentos	Bases	Tamanho médio (pb) do fragmento	% Genoma montado	Cobertura
Azucena	58.832.480	4.471.268.477	76	94%	12,3
Catetão	64.782.392	4.250.579.050	66	94%	11,9
IAC 165	66.809.463	4.472.696.736	68	95%	12,4
Ligeiro	70.047.717	4.673.044.461	68	95%	12,9
Moroberekan	71.061.320	4.730.575.594	67	95%	13,2
BRS Primavera	72.805.882	4.980.823.233	69	95%	13,8
Chorinho	102.993.784	7.827.527.584	76	100%	20,4
Puteca	107.059.362	8.136.511.512	76	96%	21,2

Referência: cv. Nipponbare (372.317.567 pb)

O banco de dados de sequências das oito variedades *japonica* tropical possibilitou a montagem extensiva de regiões distribuídas nos 12 cromossomos da espécie. O alinhamento comparativo das sequências genômicas permitiu a identificação de milhares de SNPs distribuídos ao longo do genoma de arroz. Uma sequência de filtros de seleção foi aplicada ao banco de dados de SNPs para identificar uma bateria de aproximadamente 4.300 marcadores com potencial de revelar polimorfismo entre acessos de *O. sativa*. Os filtros incluíram, por exemplo, restrições de cobertura mínima no sítio do SNP igual ou maior a 20x para qualquer combinações pareadas das oito variedades, número de SNPs proporcional ao tamanho físico dos cromossomos (Tabela 2), e distância física (pb) proporcional entre sítios SNPs ao longo dos cromossomos. Tomou-se o cuidado ainda em identificar SNPs polimórficos em três de quatro combinações pareadas (Chorinho x Puteca; Primavera x IAC 165; Catetão x Azucena; Moroberekan x Ligeiro). Sempre que possível, procurou-se selecionar 50% dos SNPs em região gênicas (éxon) do genoma de arroz.

Tabela 2 – Relação entre a montagem obtida de cada cromossomo de arroz e o número de SNPs selecionados para compor o chip de DNA.

Cromossomo	Montagem Total (Kbp)	Total de SNPs selecionados
1	42.255	494
2	36.258	448
3	35.000	341
4	35.600	429
5	30.168	368
6	31.532	368
7	29.968	349
8	28.699	362
9	23.222	183
10	23.346	280
11	28.773	385
12	27.306	293
TOTAL	372.127	4.300

Para iniciar o processo de validação do chip de DNA para uso na genotipagem em

escala de acessos de arroz, amostras de DNA em regime de prova e contra-prova foram extraídas de uma mesma planta de cada uma das oito variedades de arroz usadas no sequenciamento. A comparação entre os genótipos obtidos em milhares de locos SNP entre duas amostras (prova e contra-prova) da mesma variedade permitiu estimar o grau de acurácia (reprodutibilidade) da genotipagem usando o chip de DNA. A comparação contabilizou o número de inconsistências entre a genotipagem de prova e contra-prova no conjunto de marcadores SNPs com sinal de fluorescência adequado para análise (Tabela 3). A ausência de dados em uma das réplicas (ou em ambas) naturalmente não foi considerada na análise. O resultado indica que o nível de acurácia da genotipagem de >3500 marcadores SNPs simultaneamente com o chip de DNA é extremamente elevado (>99,97%).

Tabela 3 – Estimativa da acurácia de genotipagem de milhares de marcadores SNPs simultaneamente com o chip de DNA em oito variedades de arroz, em regime de prova e contra-prova.

	Chorinho	Puteca	IAC 165	BRS Primavera	Azucena	Moroberekan	Catetão	Ligeiro
Total de SNPs	3742	3742	3742	3742	3742	3742	3742	3742
Dado faltante (p)	47	24	19	24	38	56	46	121
Dado faltante (p&cp)	161	138	145	138	148	156	151	141
Número de genótipos analisados	3.534	3.580	3.578	3.580	3.556	3.530	3.545	3.480
Número genótipos inconsistentes	1	0	1	0	0	1	0	3
Acurácia	99,972%	100%	99,972%	100%	100%	99,972%	100%	99,914%

p= prova; cp= contra-prova

CONCLUSÃO

O emprego de NGS para sequenciamento em escala de genomas de oito variedades de arroz *japonica* tropical foi altamente eficiente. A montagem do genoma e a análise comparativa da sequência de DNA destas variedades permitiu a descoberta de milhares de sítios SNP.

O emprego de filtros de seleção possibilitou a seleção de 4.300 SNPs distribuídos uniformemente nos 12 cromossomos de arroz. As informações de sequência na região flaqueadora dos SNPs selecionados possibilitou o desenvolvimento de um chip de DNA para genotipagem em escala de acessos de arroz.

O teste do chip em amostras de prova e contra-prova de oito variedades de arroz possibilitou a genotipagem de 3.742 SNPs, com acurácia acima de 99,972%.

REFERÊNCIAS BIBLIOGRÁFICAS

BROOKES, A. J. The essence of SNPs. Gene, v.234, n.2, p177-86, 1999.

FERREIRA, M., GRATTAPAGLIA, D. Introdução ao uso de marcadores moleculares em análise genética (Documentos, p. 220p). EMBRAPA-CENARGEN, 1998.

GANAL, M. W., ALTMANN, T., RÖDER, M. S. (2009). SNP identification in crop plants. Current opinion in plant biology, v.12, n2, p.211-217, 2009.

RISCH N., MERIKANGAS K. The future of genetic studies of complex human diseases. Science v.273, p.1516-1517, 1996.