VI CONGRESSO BRASILEIRO DE SOJA
CUIABÁ, MT - 2012

Genotipagem de cultivares brasileiros de soja através da detecção de SNPs

Genotyping of Brazilian soybean cultivars through SNPs detection

Nascimento, L.C.[1,2]; Vidal R.O.[1]; Mondego, J.M.C.[3]; Costa, G.G.L.[1,2]; Junior, O.R.[1,2]; Rodrigues, F.[4]; Nepomuceno, A.L.[4]; Marcelino-Guimarães, F.C.[4]; Abdelnoor, R.V.[4]; Pereira, G.A.G. [1]; Carazzolle, M.F.[1,5].

1 – Laboratório de Genômica e Expressão (LGE) – Instituto de Biologia - Universidade Estadual de Campinas (UNICAMP) – Campinas – SP – Brasil; e-mail: *l.costa.nascimento@gmail.com*
2 – Laboratório Central de Tecnologias de Alto Desempenho (LaCTAD) – Universidade Estadual de Campinas (UNICAMP) – Campinas – SP - Brasil.
3 – Centro de Pesquisa e Desenvolvimento em Recursos Genéticos Vegetais, Instituto Agronômico de Campinas (IAC) – Campinas – SP – Brasil.
4 – Centro Nacional de Pesquisas em Soja – Empresa Brasileira de Pesquisa Agropecuária – Embrapa Soja – Londrina – PR – Brasil.
5 – Centro Nacional de Processamento de Alto Desempenho (CENAPAD) – Universidade Estadual de Campinas (UNICAMP) – Campinas – SP – Brasil.

Resumo

A soja é um dos principais produtos da balança comercial brasileira, respondendo por mais de 10% do total das exportações do país. No ano de 2007, o governo brasileiro estabeleceu um consórcio de pesquisas em soja – denominado GENOSOJA – com o objetivo de identificar características genéticas que possam facilitar o processo produtivo da planta, com foco nos diversos estresses que acometem a produção nacional, como a ocorrência de secas, o ataque de pragas e a doença da ferrugem asiática. Entre os objetivos do consórcio está a geração de sequências de DNA e mRNA de cultivares brasileiros selecionados em programas de melhoramento. Polimorfismos de nucleotídeo único (SNPs; single nucleotide polymorphisms) são diferenças de uma base entre as sequências de DNA de indivíduos, cultivares ou espécies. A identificação de SNPs tem uma grande aplicação em melhoramento de plantas, pois podem ser utilizados como marcadores moleculares para genotipagem. Neste trabalho, os dados de DNA e mRNA gerados pelo consórcio GENOSOJA foram utilizados para identificação de SNPs que podem colaborar na seleção de cultivares de interesse, bem como na geração de novos cultivares resistentes aos problemas da lavoura brasileira.

Introdução

Soja (*Glycine max*) é o legume de maior importância econômica no mercado internacional, com uma produção mundial de aproximadamente duzentos e sessenta milhões de toneladas na safra 2010/2011 (Empresa Brasileira de Pesquisa Agropecuária – Embrapa Soja). O Brasil aparece como o segundo maior produtor, com aproximadamente 30% da produção mundial. A produção de soja no Brasil é fortemente influenciada pelas condições climáticas, como longos períodos de seca ou alagamento, e também por alguns ataques de pragas da lavoura, como por exemplo, o fungo *Phakopsora* pachyrhizi (causador da doença ferrugem asiática) e alguns nematoides. A solução para esses problemas está nos programas de melhoramento, aliados as descobertas de novas técnicas de plantio e prevenção. Durante vários anos de trabalho, alguns cultivares tolerantes e susceptíveis foram identificados e isolados dessas regiões. A genotipagem e o entendimento das bases moleculares desses cultivares podem contribuir para a geração de novos cultivares e ajudar na seleção assistida.

Polimorfismo de nucleotídeo único (SNP; *single nucleotide polymorphism*) é uma mutação em uma base presente em ao menos 1% de uma população. Os SNPs

estão presentes ao longo de todo o genoma e são importante fonte de variabilidade entre indivíduos, cultivares ou espécies. A identificação desses SNPs tem uma grande aplicação em melhoramento de plantas, pois os mesmos podem ser utilizados como marcadores moleculares para genotipagem. Recentes avanços nas tecnologias de sequenciamento de DNA e mRNA (chamado de RNA-seq), têm revolucionado a área de melhoramento permitindo a identificação de SNPs em larga escala e com baixo custo de produção. Esta metodologia é preferencialmente aplicada para plantas que possuem o seu genoma sequenciado, permitindo a identificação de SNPs em regiões codantes e intergênicas. Esse é o caso da soja, que possui o genoma completo sequenciado com todos os genes anotados (http://www.phytozome.net/soybean.php).

Neste contexto, o consórcio brasileiro do genoma da soja (GENOSOJA) foi criado em 2007 com o objetivo de 1) produzir dados de DNA e mRNA de cultivares brasileiros com fenótipos importantes e 2) analisar esses dados com ajuda de uma equipe multidisciplinar, distribuída por várias instituições, de forma a entender os mecanismos moleculares dos atuais cultivares e produzir novos cultivares que aumentem a produção de soja no Brasil. Todos esses dados estão integrados e podem ser acessados pelos pesquisadores através do link http://www.lge.ibi.unicamp.br/soybean (Nascimento *et al.*, 2012).

Nesse trabalho os dados de RNA-seq e resequenciamento de DNA de cultivares brasileiros gerados pelo consórcio GENOSOJA foram utilizados para identificação de SNPs utilizando o genoma da soja como referência. Esses SNPs podem ser úteis para seleção assistida de cultivares de interesse e entendimento dos mecanismos moleculares dos cultivares brasileiros de importância agrícola.

Materiais e métodos

A obtenção de um genoma ou transcriptoma de referência de alta qualidade é o principal pré-requisito necessário para a utilização da metodologia de identificação de SNPs. Após essa etapa os diferentes indivíduos, cultivares ou espécies podem ser sequenciados com baixa cobertura (~30x) utilizando alguma das novas tecnologias de sequenciamento disponíveis, por exemplo, 454 Life Sciences, Illumina Solexa e SOLID. Cada sequência produzida pelo sequenciador, chamado de *read*, representa um fragmento de DNA, ou mRNA, e o tamanho depende da tecnologia de sequenciamento utilizada (de 50 à 400 bp).

Os *reads* foram submetidos a um filtro de qualidade que descarta sequências com alguma base com qualidade menor do que a nota phred igual a 20. Os *reads* dos diferentes indivíduos foram alinhados na sequência de referência utilizando os softwares BWA (Li and Durbin, 2010) ou SOAP (Li *et al.*, 2009) para de DNA ou mRNA, respectivamente. O BWA é mais indicado para alinhamento em genoma, pois realiza um alinhamento mais robusto, portanto mais demorado, levando em consideração eventos de inserção e deleção (indels) mais frequentes em regiões intergênicas. Os resultados foram salvos no formato SAM.

A detecção de SNPs foi realizada com o pacote SAMtools programa *pileup*, que identifica as variações diretamente no arquivo SAM. O software VarScan (Koboldt *et al.*, 2009) foi utilizado para 1) filtrar as variações baseados em contagem de reads alinhados na região, qualidade das bases, ambiguidade de alinhamento e frequência alélica e 2) subtrair os SNPs através da comparação entre dois cultivares. De acordo com os filtros, SNPs com qualidade inferior a phred 15 foram descartados, assim como SNPs em regiões de cobertura inferior a 20 e frequência do menor alelo inferior a 10%. Também foram descartados SNPs em região próximas a indels (10bp para ambos os lados) e em regiões ricas em variações (com mais de 2 SNPs nos 20bp ao redor).

Para o caso de SNPs identificados dentro de regiões codantes, os genes foram classificados e agrupados em termos do banco de dados Gene Ontology usando o software blast2go. A posição do gene polimórfico no cromossomo também foi contabilizada.

Resultados e discussão

A tabela abaixo resume os dados produzidos pelo consórcio do GENOSOJA e que estão sendo utilizados para identificação de SNPs.

Tabela 1 – Sumário das bibliotecas utilizadas para identificação de Polimorfismos

Biblioteca	Cultivar	Tipo	Número de sequências
Ferrugem Asiática	PI1356- resistente	mRNA	14.886.500
Ferrugem Asiática	PI230970 – resistente	mRNA	13.894.355
Seca	BR 16 – suscetível	mRNA	11.783.331
Seca	Embrapa 48 – tolerante	mRNA	30.520.738
Alagamento	Embrapa 45	mRNA	25.281.310
Alagamento	BR 4	mRNA	22.369.635
Nematóide	MG/BR 46	mRNA	9.965.041
Nematóide	Chapadões	mRNA	115.559.517
Nermatóide	TMG115RR	mRNA	142.907.158
Ressequenciamento	Conquista	DNA	Em andamento

O pipeline de identificação de SNPs está sendo aplicado para todas as bibliotecas descritas na tabela 1. As análises para as bibliotecas de *RNA-seq* (biblioteca subtrativa) de estresse a seca (cultivares tolerantes e susceptíveis) estão finalizadas (Vidal *et al.*, 2012) e os resultados estão descritos nos parágrafos abaixo.

Após a aplicação dos filtros de qualidade um total de 12,285,871 *reads* (tamanho de 45 bp) e 30,326,963 *reads* (tamanho de 76 bp) foram obtidos para os cultivares susceptíveis e tolerantes a seca, respectivamente. O alinhamento desses *reads* nas sequências de referência dos genes resultaram em 6,317,010 *reads* do cultivar susceptível que foram alinhados em 9,698 genes, e para os cultivares tolerantes, 6,120,258 *reads* foram alinhados em 21,951 genes. Um total de 7,897 genes alinhados foram comuns para ambos os cultivares sendo utilizados na identificação dos SNPs.

A identificação dos genes polimórficos foi realizada pelos softwares SAMtools pileup e VarScan. Um total de 44,510 variações foi identificado, representando SNPs dentro dos cultivares (SNPs entre alelos) e entre os cultivares. Os SNPs entre alelos representam a maior parte dos polimorfismos identificados (~3 SNPs/gene) mas estes não são uteis como marcadores moleculares. No entanto, foram identificados 165 SNPs entre os cultivares susceptíveis e tolerantes num total de 127 genes. Estes genes foram submetidos a anotação utilizado o banco de dados do Gene Ontology e a figura 1 resume as classes identificadas.

Diversas classes relacionadas à resposta a estresses representados por vários fatores de transcrição (WRKY, MYB, Zinc Finger, Homeodomain-ZIP) foram identificados. Todos esses polimorfismos identificados são fortes candidatos a marcadores moleculares para seleção assistida.

Conclusão

A identificação de SNPs específicos entre cultivares pode ser uma ferramenta muito útil para o programa de melhoramento. O consórcio GENOSOJA está gerando dados de sequenciamento de DNA e mRNA de diversos cultivares brasileiros de interesse agrícola que estão sendo utilizados para identificação de SNPs em larga escala produzindo centenas ou milhares de marcadores moleculares. A aplicação dessa metodologia na identificação de SNPs entre cultivares susceptíveis e tolerantes a seca mostraram a presença de polimorfismos em genes relacionados à resposta ao estresse e fatores de transcrição, tornando-se bons candidatos para a produção de transgênicos e para o programa de melhoramento.

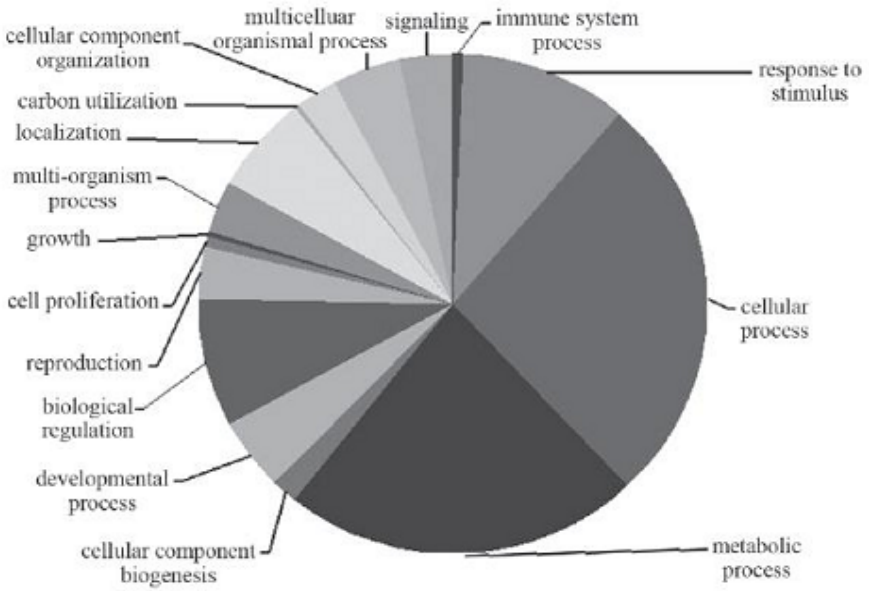

Figura 1: Classificação baseada no *Gene Ontology* dos 127 genes com polimorfismo entre cultivares susceptíveis e tolerantes a seca

Referências

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK and Ding L (2009). **VarScan: variant detection in massively parallel sequencing of individual and pooled samples.** Bioinformatics 25 (17): 2283-2285.

Li H and Durbin R (2010). **Fast and accurate long-read alignment with Burrows-Wheeler transform.** Bioinformatics 26 (5): 589-595.

Li R, Yu C, Li Y, Lam T-W, Yiu S-M, Kristiansen K and Wang J (2009). **SOAP2: an improved ultrafast tool for short read alignment**. Bioinformatics 25 (15): 1966-1967.

Nascimento LC, Lacerda GGL, Binneck E, Pereira GAG and Carazzolle MF (2012). **A web-based bioinformatics interface applied to Genosoja Project: Databases and pipelines.** *Genetics and Molecular Biology*, 35, 2 (suppl).

Vidal R, Nascimento LC, Mondego JMC, Pereira GAG and Carazzolle MF (2012). **Identification of SNPs in RNA-Seq data of two cultivars of *Glycine max* (soybean) differing in drought resistance.** *Genetics and Molecular Biology*, 35, 2 (suppl).