VII CONGRESSO BRASILEIRO DE SOJA
MERCOSOJA 2015

GENOMIC ANALYSIS OF SOYBEAN ACCESSIONS FOR ALLELIC VARIATIONS IDENTIFICATION IN BRAZILIAN GERMPLASM

MALDONADO DOS SANTOS, J.V.[1,2]; JOSHI, T.[3]; KHAN, S.M.[3]; LIU, Y.[3]; WANG, J.[3]; VUONG, T.D.[3]; MARCELINO-GUIMARÃES, F.C.[1]; OLIVEIRA, M.F.de[1]; VALLIYODAN, B.[3]; XU, D.[3]; NGUYEN, H.T.[3]; ABDELNOOR, R.V.[1,2]. [1]Brazilian Corporation of Agricultural Research (Embrapa Soja), Londrina, PR, Brazil, jv_maldonado@hotmail.com; [2]Londrina State University (UEL), Londrina, PR, Brazil; [3]University of Missouri, Columbia, MO, USA.

Soybean [*Glycine max* (L.) Merrill] is one of the most important leguminous crop of the world, due to its importance in human food, animal feed and oil production. Actually, Brazil is the second largest soybean producer with potential to become the largest one, with an estimation of 95.919 million of tons, from 31.621 million hectares for the 2014/15 Brazilian growing season (COMPANHIA NACIONAL DE ABASTECIMENTO, 2014). The success of soybean for the Brazilian agribusiness is the direct result of applied technologies for tropical conditions and the increase of the production in traditional areas and advancement to new agricultural frontiers (EMBRAPA SOJA, 2014).

The Brazilian soybean breeding program has a very recent history, with the first cultivar dating from the 40s, and a large number of cultivars, adapted to different environmental conditions, have been released over the last decades. The development of tools that can help breeding programs to keep the demand for cultivars with high yield and adapted to different stress conditions are essential for the increasing demand of food in the world. Molecular biology techniques have emerged as important tools for plant breeding assistance, especially after the sequencing of the reference genome. Among these tools, the new high-throughput sequencing platforms arise as important alternatives in genetics and plant breeding studies. In this study, we resequenced 28 Brazilian soybean lines released over last 50 years and belonging to different maturity groups to evaluate the modifications among the genomes.

Young leaf tissue sample of each 28 Brazilian cultivars were collected during growth stage V3. The genomic DNA was isolated for each sample through the Qiagen Mini Plant DNeasy kit (QIAGEN INC., VALENCIA, CA, USA), following the manufacturer's instructions The DNA samples were sent to FASTERIS Company, Switzerland, for sequencing, on the Illumina Hiseq 2000 platform.

The resequenced reads from the soybean accessions were mapped with the new version of the soybean reference genome (Gmax_275_Wm82.a2.v1) through the BWA software (LI; DURBIN, 2009). The aligned reads were processed through Picard tools version 1.107 and a binary file, representing the assembled genome of each resequenced cultivar was generated. For SNPs/InDels detection, the Genome Analysis Toolkit (GATK) (MCKENNA et al., 2010), version 3.0, was used to make a local realignment into InDels regions to generate a new binary file with fewer errors. Therefore, the new assembled files for each cultivars were used for SNPs/InDels calling. For this analysis, the HaplotypeCaller module of GATK was selected due to accuracy for SNPs/InDels detection compared to UnifiedGenotype module.

The analysis was conducted using bioinformatics Next-Generation Sequencing (NGS) data analysis workflow (LIU et al., 2014) developed in SoyKB for SNP and Indel calling and was conducted using XSEDE as the computing infrastructure, iPlant as the data and cloud infrastructure (GOFF et al. (2011), and the Pegasus workflow systems (DEELMAN et al., 2005) to control and coordinate the data management and computational tasks. For a genetic annotation and functional classification, where allelic variations had been detected, we used snpEff program (CINGOLANI et al., 2012). An enrichment analysis of these modified genes detected through snpEff were made through the website AgriGO (China Agricultural University, 2014).

The resequencing effort, for the 28 cultivars, generated around 5.5 billion of paired-end reads (100 bp long) with an average of depth of 14.78x the soybean genome. The average percentage of mapped reads on the soybean genome for these cultivars was 94.31%, which means the resequencing covered most of the soybean genome regions. Around 5.9 million of SNPs and 1.3 million of InDels were found on the resequenced genomes, distributed over all the 20 chromosomes (Table 1). Cultivars Santa Rosa and Doko presented the highest number of SNPs, while in contrast, Anta 82 and VMAX RR showed the lowest number.

Furthermore, we were able to identify over 500,000 SNPs that are exclusive for the lines used in our study (Table 2). These findings can be useful in breeding programs, allowing the selection of unique alleles for specific lines and for cultivar fingerprinting. In this study, we also identified 720 SNPs that share the same allele in all Brazilian lines and are divergent to the reference genome. These SNPs are located in 369 genes and causing putative modifications in start/stop codons, coding sequences and splice site regions. According to our results, some of these genes are associated with some important biological processes. Whether these genes were involved in tropical adaptation of soybean, it still needs to be clarified.

References

CHINA AGRICULTURAL UNIVERSITY. ANALYSIS TOOLKIT FOR THE AGRICULTURAL COMMUNITY (agriGO). Disponível em: <http://bioinfo.cau.edu.cn/agriGO/analysis.php>. Acesso em: 14/11/2014.

CINGOLANI, P.; PLATTS, A.; WANG, L. L.; et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118 ; iso-2; iso-3. **Landes Bioscience**, v. 6, n. 2, p. 80–92, 2012.

COMPANHIA NACIONAL DE ABASTECIMENTO. Séries Históricas de Área Plantada, Produtividade e Produção, Relativas às Safras 1976/77 a 2014/15 de Grãos, 2001 a 2014 de Café, 2005/06 a 2014/15 de Cana-de-Açúcar. Disponível em: <http://www.conab.gov.br/conteudos.php?a=1252&>. Acesso em: 14/11/2014.

DEELMAN, E.; SINGH, G.; SU, M.; et al. Pegasus : A framework for mapping complex scientific workflows onto distributed systems. **Scientific Programming**, v. 13, n. January, p. 219–237, 2005.

EMBRAPA SOJA. EMBRAPA SOJA. História: Histórico no Brasil. Disponível em: <https://www.embrapa.br/en/soja/cultivos/soja1/historia>. Acesso em: 14/11/2014.

GOFF, S. A; VAUGHN, M.; MCKAY, S.; et al. The iPlant Collaborative: Cyberinfrastructure for Plant Biology. **Frontiers in plant science**, v. 2, n. July, p. 34, 2011.

LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows-Wheeler transform. **Bioinformatics (Oxford, England)**, v. 25, n. 14, p. 1754–60, 2009.

LIU, Y.; KHAN, S. M.; WANG, J.; et al. Large Scale NGS resequencing data analysis workflow for soybean germplasm using iPlant, XSEDE and SoyKB framework. **Bioinformatics (Oxford, England)**, v. in press, 2014.

MCKENNA, A.; HANNA, M.; BANKS, E.; et al. The Genome Analysis Toolkit : A MapReduce framework for analyzing next-generation DNA sequencing data. , p. 1297–1303, 2010.

Table 1. Distribution of SNPs/InDels per gene regions in the 28 soybean cultivar genomes compared to the reference genome.

Type	Intergenic	UTR	Intron	Exon	Other	Total
SNPs	2,684,448	112,790	287,414	218,671	2,565,021	5,868,344
Indels	463,106	40,105	79,721	25,861	721,051	1,329,844

*Intergenic, the variant is in a intergenic region; UTR, variant hits 5' and 3' UTR region; Intron, variant hits an intron; Exon; variants hits an exon; Other: all the remaining genome regions.

Table 2. Exclusive SNPs for each Brazilian soybean cultivar used in this study

Accession name	Number of SNPs
Anta 82	3,586
BR 16	7,036
BRS/GO 8360	5,328
BRS/GO 8660	20,388
BRS/GO Chapadões	74,314
BRS 232	3,651
BRS 284	62,279
BRS 360 RR	3,731
BRS 361	10,778
BRS Sambaíba	31,811
BRS Valiosa RR	344
BRSMG 850G RR	318
BRSMT Pintado	3,116
BRSMT Uirapuru	10,662
CD 201	11,050
Conquista	1,486
Doko	42,826
Embrapa 48	1,882
Emgopa 301	12,590
FT Abyara	36,447
FT Cristalina	458
IAC 8	41,325
IAS 5	8,918
NA 5909 RG	22,691
P98Y11	18,590
Paraná	6,835
Santa Rosa	96,105
VMAX RR	3,215
Total	541,760