COPY-NUMBER VARIATIONS REVEAL DIVERGENCE AMONG BRAZILIAN SOYBEAN CULTIVARS

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.

Copy-number variations (CNVs) refer to structural modifications that produce changes in the copy number in a specific region of the genome. Such modifications may vary in size, and may be generated through non-allelic homologous recombination (NAHR), Fork Stalling and Template Switching models (FoSTeS) and non-homologous end-joining (NHEJ) (GU et al., 2008; ZMIEŃKO et al., 2014). Recently, CNVs have been linked to many types of traits and some diseases, such as Alzheimer's (ROVELET-LECRUX et al., 2006), autism (SEBAT et al., 2010) and Parkinson (SIMON-SANCHEZ et al., 2008) in humans.

In soybean, it is also observed that a significant number of CNVs are associated to important traits, as the resistance against soybean cyst nematode (COOK et al., 2014) and the fitness avirulence of *Phytophthora sojae* for soybean infection (QUTOB et al., 2009). Therefore, the identification of CNVs on soybean genome can be very useful on soybean breeding programs. In this study, we analyzed the genome of 28 Brazilian soybean lines to find CNVs that could be used for diversity studies. The Brazilian cultivars used in this study were selected based on the date of commercial release in a 50-year span of soybean breeding program in Brazil. In addition, materials with different maturity groups, adapted to different regions of the country, were chosen, representing a large diversity set of cultivars.

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 soybean genomes 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 the realign of indels regions of each cultivar was made by Genome Analysis Toolkit (MCKENNA et al., 2010). A bioinformatics Next-Generation Sequencing data analysis workflow were developed to generate the analysis (LIU et al., 2014), using XSEDE as the computing infrastructure, iPlant as the cloud infrastructure (GOFF et al., 2011), and the Pegasus workflow systems (DEELMAN et al., 2005) to coordinate the data management and computational tasks. For CNVs detection on soybean genome, we used Copy Number estimation by a Mixture Of Poissons (cn.MOPS), version 1.10.0 (KLAMBAUER et al., 2012).

A large number of CNVs regions, spreaded through the 20 soybean chromosomes, were detected on the 28 soybean genomes. Chromosomes 14 and 17 showed the highest number of CNVs regions, while the lowest number was observed on chromosome 16. Most of the cultivars had more deletions than duplications. The cultivar BRS 284 had the highest number of CNVs, with a virtually an equal proportion of deletions and duplications in the genome. BRS 232, Doko, and Santa Rosa were the cultivars with the highest number of deletions and VMAX RR had the lowest number of

deleted regions. Additionally, cultivars BRS 284, BRS/GO 8360, and VMAX RR had the highest number of duplications, which contrast with BRS Valiosa and FT Cristalina. A summary of the number of CNVs detected for the cultivars is shown in Table 1.

A depth analysis on chromosome 16 found CNVs sub-regions with average of 10 kb that distinguished 80% of the latest material (after 2000) from six materials developed before 1990. One of these was not detected in any cultivar released after 2000 and about 70% of the material prior to 1999. These findings suggest the predominance of inserted regions on chromosome 16 in the more recent soybean cultivars. Other important CNVs regions that were able to distinguish most of the cultivars developed before 1990 from the large majority of latest ones were detected on chromosomes 6, 7, 8, 9, 13, 15, and 17.

CNVs analysis showed as an important tool to check significant modifications in soybean genome. The existence of a large amount of CNVs regions that allow the differentiation among the Brazilian soybean germplasm also emerge as a potential target for studies of important agronomic traits. Proving the importance of these CNV regions, they could become an important tool for breeding and may help the breeders to achieve a jump in productivity over the time, as well as the larger adaptability to different environments.

References

COOK, D. E.; BAYLESS, A. M.; WANG, K.; et al. Distinct Copy Number, Coding Sequence, and Locus Methylation Patterns Underlie Rhg1-Mediated Soybean Resistance to Soybean Cyst Nematode. **Plant physiology**, v. 165, n. 2, p. 630–647, 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.

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. GU, W.; ZHANG, F.; LUPSKI, J. R. Mechanisms for human genomic rearrangements. **PathoGenetics**, v. 1, p. 4, 2008.

KLAMBAUER, G.; SCHWARZBAUER, K.; MAYR, A.; et al. cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. **Nucleic acids research**, v. 40, n. 9, p. e69, 2012. 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.

QUTOB, D.; TEDMAN-JONES, J.; DONG, S.; et al. Copy number variation and transcriptional polymorphisms of Phytophthora sojae RXLR effector genes Avr1a and Avr3a. **PLoS ONE**, v. 4, n. 4, 2009.

ROVELET-LECRUX, A.; HANNEQUIN, D.; RAUX, G.; et al. APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. **Nature genetics**, v. 38, n. 1, p. 24–6, 2006.

SEBAT, J.; LAKSHMI, B.; MALHOTRA, D.; et al. Strong Association of De Novo Copy Number Mutations with Autism. **Science**, v. 316, n. 5823, p. 445–449, 2010.
SIMON-SANCHEZ, J.; SCHOLZ, S.; MATARIN, M. DEL M.; et al. Genomewide SNP Assay Reveals Mutations Underlying Parkinson Disease. **Human Mutation**, v. 29, n. 2, p. 315–322, 2008.
ZMIEŃKO, A.; SAMELAK, A.; KOZŁOWSKI, P.; FIGLEROWICZ, M. Copy number polymorphism in plant genomes. **Theoretical and Applied Genetics**, v. 127, p. 1–18, 2014.

Table 1. Total of Copy-Number Variations (CNVs) regions observed for each cultivar used in this study.

Accession Name	Decade	Deletions	Insertions	Total
Santa Rosa	1961-1970	770	332	1102
Doko	1971-1980	717	401	1118
IAC 8	1971-1980	657	289	946
IAS 5	1971-1980	621	545	1166
Paraná	1971-1980	476	396	872
Emgopa 301	1981-1990	586	348	934
FT Abyara	1981-1990	597	509	1106
FT Cristalina	1981-1990	691	170	861
BR 16	1991-2000	609	304	913
BRSMT Pintado	1991-2000	613	229	842
BRSMT Uirapuru	1991-2000	702	259	961
CD 201	1991-2000	657	285	942
Conquista	1991-2000	661	221	882
Embrapa 48	1991-2000	556	342	898
Anta 82	2001-2010	434	615	1049
BRS 232	2001-2010	730	365	1095
BRS 284	2001-2010	670	667	1337
BRS Sambaíba	2001-2010	687	291	978
BRS Valiosa RR	2001-2010	658	187	845
BRS/GO 8360	2001-2010	493	643	1136
BRS/GO 8660	2001-2010	699	202	901
BRS/GO Chapadões	2001-2010	516	292	808
BRSMG 850G RR	2001-2010	635	304	939
NA 5909 RG	2001-2010	537	504	1041
P98Y11	2001-2010	676	222	898
VMAX RR	2001-2010	542	374	916
BRS 360 RR	2011-2014	539	396	935
BRS 361	2011-2014	593	401	994