Genetics and Molecular Biology, 35, 1 (suppl), *i-iv* (2012)

www.sbg.org.br

Editorial

GENOSOJA – The Brazilian Soybean Genome Consortium: High throughput omics and beyond

Ana M. Benko-Iseppon[1], Alexandre L. Nepomuceno[2] and Ricardo V. Abdelnoor[2]

[1]*Laboratório de Genética e Biotecnologia Vegetal, Departamento de Genética, Universidade Federal de Pernambuco, Recife, PE, Brazil.*
[2]*Embrapa Soja, Rodovia Carlos João Strass, Distrito de Warta, Londrina, PR, Brazil.*

Plant genomes are among the most complex and large ones of our planet, with high levels of redundancy when compared to other eukaryotic groups, leading to intricate processes for gene regulation and evolution. Such a complexity demands interdisciplinary and multidimensional approaches in order to allow a better understanding of the processes able to exploit the whole potential of the existing genes in different species, including crop plants. Among cultivated plants, soybean [*Glycine max* (L.) Merr.] occupies an outstanding position due to its importance as source of protein and oil that may also be converted into biodiesel. The seeds are rarely consumed *in natura*, but many traditional food products have been consumed, as soymilk, and tofu, as well as fermented products as soy sauce, and soy paste among others, besides its wide use for animal feed.

Soybean cultivation has been highly concentrated geographically, with only four countries (USA, Brazil, Argentina and China) accounting for almost 90% of world output. Asia (excluding China) and Africa, the two regions where most of the food insecure countries are located, account for only 5% of production. Among countries classified as 'undernourished', only India and Bolivia are significant producers of soybeans (FAO, 2009).

Available evidences indicate that the cultivated soybean was domesticated from its wild relative *Glycine soja* (Sieb. and Zucc.) in China about 5,000 years ago (Carter *et al.*, 2004). Since then, soybean has been grown primarily in temperate regions for thousands of years, first in Northern Asia and in more recent years in North America and countries of the Southern Cone of Latin America (FAO, 2009). The remarkable success of this crop in temperate zones is well known, but the crop presents also an important role in cropping systems of the tropics and subtropics, also in low fertile regions, as the Brazilian cerrado savannah (Spehar, 1995). The actual area cultivated worldwide with soybean is estimated to cover 103.5 millions of hectares, from which 24.2 only in Brazil, with considerable increases in the production achieved without significant increase in the cultivated area (Embrapa Soybean, 2011).

Send correspondence to Ana M. Benko-Iseppon. Laboratório de Genética e Biotecnologia Vegetal, Departamento de Genética, Universidade Federal de Pernambuco, 50.670-420 Recife, PE, Brazil. E-mail: ana.iseppon@gmail.com.

As a legume, soybean is able to develop symbiotic interactions with rhizobia, allowing the fixation and assimilation of atmospheric N_2, bearing quite specific mechanisms to coordinate this complex interaction (Oldroyd and Downie, 2008), absent in most angiosperm groups. Besides this peculiarity, soybean presents 2n = 40 chromosomes and was early characterized as an ancient polyploid (paleopolyploid) through genetic mapping studies that identified homeologous chromosome regions based upon duplicate RFLP markers (Shoemaker *et al.*, 1996; Lee *et al.*, 1999; 2001). Due to its allopolyploid nature, the first approaches regarded the generation of expressed sequences from different library tissues and conditions, including mainly ESTs (Expressed Sequence Tags; Nelson *et al.*, 2005) partially in annotated databases, including ca. 40.000 full length cDNA clones available (Umezawa *et al.*, 2008, see also RIKEN Soybean Full-Length cDNA Database), besides analyses regarding RNAseq under different tissues and development stages, as well as under different stressing situations (*e.g.* Libault *et al.*, 2010; Severin *et al.*, 2010). Also a complete shotgun genome sequence of the soybean cultivar Williams 82 was released, comprising 1.1-gigabase genome size allowing the integration of physical and high-density genetic maps, including 46,430 predicted protein-coding genes (Schmutz *et al.*, 2010).

The total amount of data publicly available at GenBank (NCBI) includes more than 120,000 nucleotide sequences (mainly mRNA), ~1,460,000 ESTs, ~368,000 genome sequences, ~80,000 proteins, 118 deposited structures and more than 6,2 million SNPs. Such numbers show that working with soybean is a very challenging task. By the other hand, despite of the wide data availability, most data regard cultivars from temperate regions (as Williams 82), not adapted for cultivation under tropical conditions, as it is the case of central Brazil and many other tropical countries that are subjected to distinct environmental stresses.

The proposition of creating the GENOSOJA consortium was submitted in 2007 to the National Council for Scientific and Technological Development (CNPq), an agency

linked to the Brazilian Ministry of Science and Technology (MCT), starting its activity in March 2008 with the participation of nine Brazilian institutions from different regions (Figure 1).

The proposal aimed to study the soybean genome from its organization into the structural level, seeking to characterize and sequence important genomic regions and their products, thus contributing to the identification of genes using transcriptional and proteomic methods, especially considering the plant response to different biotic and abiotic stresses that affect the culture in the Southern hemisphere. Still, the GENOSOJA network aimed to approach not only whether a gene is induced or suppressed under a given condition, but also to determine the levels at which it is expressed, the interactions with other genes, their physical location and products, allowing the identification of important genes and metabolic pathways, vital for the development and study of plants tolerant to challenging situations.

The GENOSOJA project is still in course and is structured into six Project Components (Figure 2), including management and addressing of different aspects of the soybean genome:

- I. Project management – responsible for the project administration, organization of meetings, group integration and research reports, among others.
- II. Structural Genomics – includes research activities related to the genomic physical architecture, including BAC anchoring (in the cultivar Conquista), promoter analysis and sequencing of gene-rich regions, also in comparison with other wild relatives of the genus *Glycine*, allowing studies of synteny and indication of regions important for ressequencing. This component is also responsible for the identification of single base polymorphisms (SNPs), very important for mapping purposes and marker assisted selection.
- III. Transcriptomics - comprises the largest research group, responsible for various expression profiling approaches using different strategies to access transcripts generated under different biotic (Asian rust:

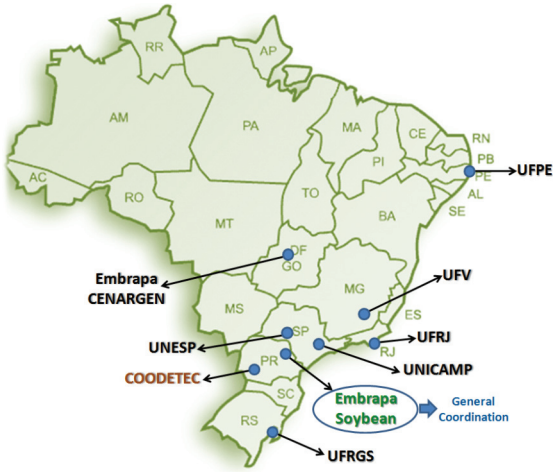

Figure 1 - The members of the GENOSOJA consortium and their geographic distribution in Brazil, including nine institutions: Embrapa Soja (Londrina, Paraná), Embrapa Recursos Genéticos e Biotecnologia (CENARGEN, Brasília, Federal District), Universidade Estadual de Campinas (UNICAMP, Campinas, São Paulo), Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP, Botucatu, São Paulo), Universidade Federal de Pernambuco (UFPE, Recife, Pernambuco), Universidade Federal do Rio Grande do Sul (UFRGS, Porto Alegre, Rio Grande do Sul), Universidade Federal de Viçosa (UFV, Viçosa, Minas Gerais) and COODETEC (Cooperativa Central de Pesquisa Agrícola, Tecnologia da Nossa Terra), a representative of the private initiative.

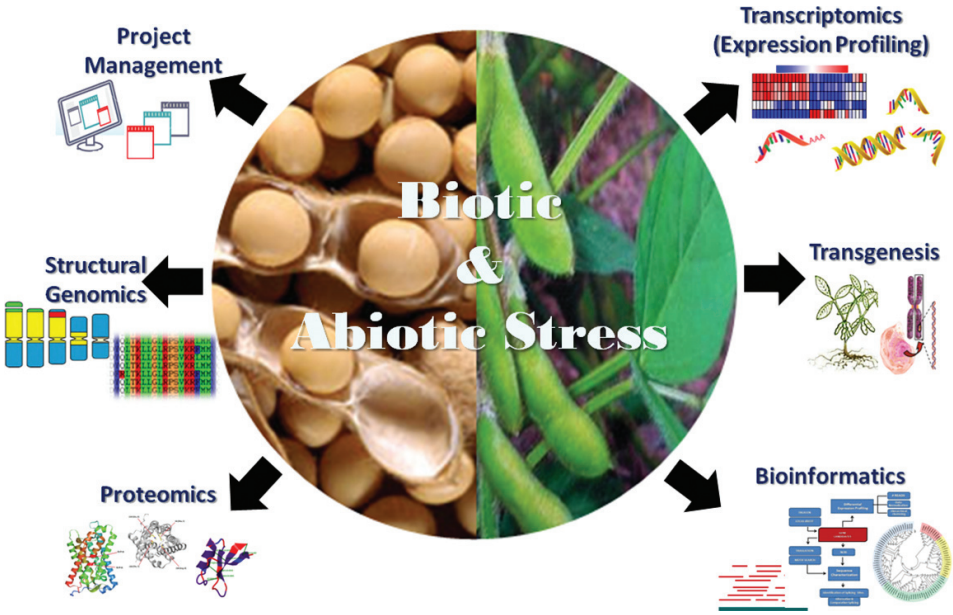

Figure 2 - Functional organization of the GENOSOJA consortium, including six project components, all composed by interinstitutional groups.

Phakopsora pachyrhizi, CPMMV: Cowpea mild mottle virus, nematodes: *Meloydogyne javanica* and *Pratylenchus brachyurus*) and abiotic (water deficit) stresses. In this workgroup different strategies were used, including:

a) Subtractive cDNA libraries (76 bp tags, Solexa Illumina® sequencing) using contrasting materials submitted to biotic interactions, including diseases (~40 million tags; Asian rust and virus inoculation) and beneficial interactions (~10 million tags; inoculation with *Bradyrhizobium japonicum),* as well as water deficit (~42 million tags, comparing tolerant and susceptible accessions).

b) SuperSAGE comprising ~3,2 Solexa Illumina® 26-bp tags distributed in six libraries generated under biotic (water deficit) and abiotic (Asian rust) stress comparisons.

c) MicroRNA libraries (Solexa Illumina®, 19-24 bp) including four libraries regarding water deficit (~4,8 millions miRNAs) and other four regarding Asian rust (~7,9 million miRNAs).

d) cDNA sequences (2,112 sequences, Sanger method) from roots infested with the nematode *M. javanica* compared with non stressed control.

The three first above mentioned experiments were carried out using the same experimental conditions, generating an extensive comparable dataset to allow the understanding of the gene expression dynamic (Subtractive cDNA and SuperSAGE libraries), including biotic and abiotic cross-talk responses as well as the post transcriptional control (miRNA).

- IV. Proteomics - aimed to study the protein profile of soybean plants, low-mass protein and peptides identification and protein-protein interactions, using the same accessions and biological conditions established for the transcriptomic analyses to ensure complementarity and reduction of experimental variability, and thus, allowing the integration of both datasets in the functional characterization of the soybean genome.
- V. Expression Assays (transgenesis) - considering the results of transcriptomics and proteomics, most valuable gene candidates are being transformed in order to infer about their effects or biological function. Members of this group are also evaluating the vicinity of genes (UTRs) for the identification of regulatory regions (promoters, enhancers, *cis*-elements, etc.) that control their expression.
- VI. Bioinformatics - this workgroup developed the GENOSOJA database (see web resource) that includes a set of tools integrating the entire project data as compared with available sequences from other public data banks.

The present issue represents the starting point of an extensive catalogue of products generated by the GENOSOJA consortium, since all members agree that many additional inferences will be soon mature for publication and application to breeding projects. Thousands of candidate genes differentially expressed have been identified and are being validated using quantitative real time PCR, many regarding strongly induced genes in contrasting libraries (*e.g.* stressed against control or tolerant against sensible in the same condition). Many of them refer to uncharacterized genes, with no given function, representing relevant data to be worked out in future functional studies, since they may represent not yet described genes, some possibly unique to legumes and important for plant breeding.

Finally, the present volume does not represent a milestone for completion of the GENOSOJA project, but an announcement of its birth, crowned with solid growth, integration and consolidation prospects. The data generated by the GENOSOJA consortium will also join the worldwide effort to study the soybean genome through the participation in the International Soybean Genome Consortium (ISGC). In this sense, the next step involves the public release of the generated data, which shall be available for the world community, allowing the effective integration with other networks throughout the world.

References

Carter TE, Nelson R, Sneller CH and Cui Z (2004) Genetic diversity in soybean. In: Boerma HR and Specht JE (eds) Soybeans: Improvement, Production, and Uses. American Society of Agronomy, Vol. 16, Madison, pp 303-416.

Lee JM, Bush A, Specht JE and Shoemaker RC (1999) Mapping duplicate genes in soybean. Genome 42:829-836.

Lee JM, Grant D, Vallejos CE and Shoemaker RC (2001) Genome organization in dicots. II. *Arabidopsis* as a 'bridging species' to resolve genome evolution events among legumes. Theor Appl Genet 103:765-773.

Libault M, Farmer A, Joshi T, Takahashi K, Langley RJ, Franklin LD, He J, Xu D, May G and Stacey G (2010) An integrated transcriptome atlas of the crop model *Glycine max*, and its use in comparative analyses in plants. Plant J 63:86-99.

Nelson RT, Grant D and Shoemaker RC (2005) ESTminer: A suite of programs for gene and allele identification. Bioinformatics 21:691-693.

Oldroyd GE and Downie JA (2008) Coordinating nodule morphogenesis with rhizobial infection in legumes. Annu Rev Plant Biol 59:519-546.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J. *et al.* (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178-183.

Severin AJ, Woody JL, Bolon YT, Joseph B, Diers BW, Farmer AD, Muehlbauer GJ, Nelson RT, Grant D, Specht JE *et al.* (2010) RNA-Seq Atlas of *Glycine max*: A guide to the soybean transcriptome. BMC Plant Biol 10:e160.

Shoemaker R, Polzin K, Labate J, Specht J, Brummer EC, Olson T, Young N, Concibido V, Wilcox J, Tamulonis J *et al.*

(1996) Genome duplication in soybean (*Glycine* subgenus soja). Genetics 144:329-338.

Spehar CR (1995) Impact of strategic genes in soybean on agricultural development in the Brazilian tropical savannahs. Field Crops Res 41:141-146.

Umezawa T, Sakurai T, Totoki Y, Toyoda A, Seki M, Ishiwata A, Akiyama K, Kurotani A, Yoshida T, Mochida K *et al.* (2008) Sequencing and analysis of approximately 40,000 soybean cDNA clones from a full-length-enriched cDNA library. DNA Res 15:333-346.

Internet Resources

Embrapa Soybean, Soybean in Numbers (in Portuguese) http://www.cnpso.embrapa.br/ (March 2012).

FAO, Food and Agriculture Organization (2009). The role of soybean in fighting world hunger. Commodities and Trade Division, Basic Foodstuffs Service http://www.fao.org/es/esc/common/ecg/125/en/The_role_of_soybeans.pdf (March 2012).

GENOSOJA, The Brazilian Soybean Genome Consortium, http://www.lge.ibi.unicamp.br/soybean (March 2012).

NCBI, *National Center for Biotechnology Information, GenBank*, http://www.ncbi.nlm.nih.gov/Genbank/index.html (March 2012).

RIKEN Soybean Full-Length cDNA Database, http://rsoy.psc.riken.jp/ (March 2012).