# An Update on Generation and Use of Microarray Resources for Soybean

Lila Vodkin, Delkin Orlando Gonzalez, Gracia Zabala, Sarah Jones
Department of Crop Sciences, University of Illinois, Urbana, Illinois 61801

## Abstract

Both spotted cDNAs and 70-mer oligo arrays are available for global gene expression analyses and represent over 36,000 members of the soybean (*Glycine max*) unigene set. The cDNA resources have been described (Vodkin et al., 2004; http://soybeangenomics.cropsci.uiuc.edu) and used to investigate many biological questions including the processes of somatic embryogenesis (Thibaud-Nissen et al., 2003), soybean seed development and germination (Jones et al., 2006; Gonzalez et al., 2006), and the responses to pathogen challenge (Zho et al., 2005) and to elevated carbon atmospheric conditions (Ainsworth et al., 2006). In addition, they proved useful for identifying single gene differences between isogenic lines (Zabala and Vodkin, 2005). Both cDNA and oligo microarrays are distributed to the community on a cost recovery basis. Contact Lila Vodkin, Department of Crop Sciences, University of Illinois for availability of slides of either the GmcDNA or GmOLIGO microarrays sets for *Glycine max*.

## Development of cDNA microarray resources for soybean

The "Public EST Project for Soybean" was a multi-university and team-oriented research project funded by soybean grower check-off funds. This project stimulated development of a large public database of soybean Expressed Sequence Tags (ESTs). The number of ESTs in this database rose from less than 100 in 1998 to over 300,000 in 2005. As part of this project, a large number of cDNA libraries have been made from the mRNAs extracted from numerous tissue and organ systems of the soybean plant (Shoemaker, et al., 2002, 2004; Vodkin et al., 2004). Over

80 different cDNA libraries were constructed from which the 300,000 ESTs were generated. The soybean EST resource data reside in public databases maintained by the National Center for Biotechnology Information (NCBI) and also in the databases of The Institute of Genomic Research (TIGR) as do EST resources developed in recent years for other plant species (Walbot, 1999; Cullis, 2004).

One of the goals of the NSF-sponsored "A Functional Genomics Program for Soybean" was to select 36,000 unique ESTs from the collection of over 300,000 primary ESTs that represented a wide array of cDNAs made to mRNAs expressed in various tissue and organ systems, and physiological stages under pathogen and stress challenges (Vodkin et al., 2004). The ESTs were compared by computer programs such as PHRAP (Green, 2001) or CAP3 (Huang and Madan, 1999) and assembled into overlapping clones that have sequence similarity. These assemblies are known as contigs (contiguous segments). In this way, longer sequences representing expressed genes were assembled and identical sequences representing redundant clones were recognized. The number of sequences in the contigs in a non-normalized cDNA library is a rough approximation of the relative abundance of the mRNAs within that tissue, and this information can be used as a “virtual” RNA blot (Stromvik, et al., 2004).

The EST with the most 5’ end sequence was chosen to represent a contig and its identity was verified by sequencing, generally at the opposite or 3’ end. This created a tentative “unigene” set in which each contig was represented by a member of the cluster and many singletons in the EST collection were also identified. The complex process of preparing the soybean cDNA microarrays (Vodkin et al., 2004) involved physically picking or “reracking” the 36,000 selected soybean cDNAs that represented the “unigene” set from among the 300,000 cDNA clones that are maintained as individual recombinant *E. coli* cultures stored in 384-well plates. The plasmid DNA templates were then purified and the 3' end of each cDNA was sequenced to verify the identity of each clone and to obtain additional sequence information for each of the unique genes. Afterwards, the inserts from the 36,000 selected cDNA clones were amplified by PCR (polymerase chain reaction), the PCR products purified, and nanoliter volumes of each were spotted onto hundreds of glass microscope slides.

As a community resource, we print two microarray slide sets each consisting of 18,432 single-spotted PCR products derived from the low redundancy cDNA sets. The GmcDNA18kA set (representing sequence-driven unigene clone libraries Gm-r1021, Gm-r1083, and Gm-r1070) is highly representative of genes expressed in the developing flowers and buds, young pods, developing seed coats, and immature cotyledons, as well as from roots of seedlings and adult plants, including roots infected with the nodulating bacterium, *Bradyrhizobium japonicum*. The GmcDNA18kB microarray slide (unigene clone libraries Gm-r1088 and Gm-r1089) is highly representative of clones selected from libraries derived from tissue-culture embryos, germinating cotyledons, and seedlings subjected to various stresses including some challenged by pathogens. Completion of both sets brings the total number of cDNAs represented to 36,864. Since the “unigene” sets are low redundancy (estimated at about 20% redundancy), this means that approximately 30,000 unique transcripts can be assayed with the set of two microarray slides.

### Development of 70-mer “long oligo” arrays

As recommended by the Soybean Genomics Executive Committee (SoyGEC) and several workshops sponsored by USB or NSF (Stacey et al., 2004), a proposal for the synthesis of 70-mer long oligos representing the soybean EST collection was funded by the United Soybean Board. The long oligos are uniformly of 70 bases and were preferentially chosen to represent the 3’ region of the cDNAs where possible. The 3' region generally has more sequence variability and can be used to design oligonucleotides that distinguish among gene family members. Clustering analysis and oligo design and synthesis were performed by Illumina, Inc., San Diego, CA. A total of 38,000 unique oligos have been synthesized. We print those also on two sets of slides (GmOLIGO19KA and GmOLIGO19KB) containing 19,200 spots each.

## Uses of the soybean microarrays for global expression analyses

A number of studies using the soybean microarrays for global expression analyses have already been published. A detailed analysis of the global expression patterns during somatic embryogenesis in soybean revealed many aspects of the events that occur during reprogramming of the cotyledon cells during the induction process (Thibaud-Nissen et al., 2003). For example, the data illustrated that auxin, which induces dedifferentiation of the cotyledon, also provoked a surge in the mRNAs involved in cell division and oxidative burst. The results also indicated that the formation of somatic globular embryos was accompanied by the accumulation of storage protein transcripts and transcripts for the synthesis of gibberellic acid.

The soybean microarrays have been used also to investigate the early response to challenge by the pathogen *Pseudomonas syringae* (Zou et al, 2005), the effect of elevated $CO_2$ atmospheric conditions (Ainsworth et al., 2006), and the transcript profiles of soybean seed development and germination (Jones et al., 2006; Gonzalez et al., 2006). The cDNA arrays have been used also in hybridizations with genomic DNA to determine gene copy number for the amplified *Hps* locus that encodes a cysteine rich hydrophobic protein that causes asthma in persons allergic to soybean dust (Gijzen et al., 2006).

## Gene identification in soybean isolines using arrays

We have used the cDNA microarrays to examine isogenic lines in order to define a small list of candidate genes that may cause a mutant phenotype. For example, as a test of this approach, we compared RNA from seed coats of two isogenic lines differing at the *T* (tawny) locus, which encodes a flavonoid 3'-hydroxylase (F3’H) enzyme, and found that the levels of the F3’H transcripts varied repeatedly by more than two-fold among the isolines (Vodkin et al., 2004). The *T* locus is responsible for the tawny or gray color of the soybean trichome hairs on the plant and was the first genetic locus to be defined in soybean by crossing and segregation analysis (Woodworth, 1921). The microarray data agree with RNA blots comparing the isogenic lines at the *T* locus, one carrying an unstable allele of the *T* locus (Zabala and Vodkin, 2003). Other criteria including cosegregation data and sequencing of alleles had previously shown that F3’H is encoded by the *T* locus (Toda et al., 2002; Zabala and Vodkin, 2003).

The use of soybean microarrays was essential to the discovery that the molecular basis of the pink flower (*wp*) locus in soybean is a mutation in the *GmF3H1* gene that encodes a flavanone 3-hydrolyase (Zabala and Vodkin, 2005). A novel gene fragment rich insertion of the CACTA family of elements (designated *Tgm-Express1*) that interrupts the *GmF3H1* gene is the reason for the reduced mRNA expression in pink flowers with the homozygous *wp* genotype compared to the normal purple flowers that carry the standard *Wp1* allele.

For further information about availability of soybean cDNA microarrays contact Lila Vodkin (l-vodkin@uiuc.edu) or see http://soybeangenomics.cropsci.uiuc.edu.

## Acknowledgements

Research supported by NSF; USDA/ARS, Illinois Soybean Association, United Soybean Board.

## References

Ainsworth, E.A., Rogers, A., Vodkin, L.O., Walter, A. and Schurr, U. 2006. The effects of elevated $CO_2$ on soybean gene expression: An analysis of growing and mature leaves. Plant Physiol. 142: 135-147.

Cullis, C.A. 2004. Plant Genomics and Proteomics. Wiley-Liss, Hoboken, NJ.

Jones, S., Gonzalez, D.O., and Vodkin, L.O. 2006. Global gene expression profiles of early soybean seed development using microarrays. Am. Soc. Plant Biologists, Boston, MA, p. 255.

Gijzen, M., Kuflu, K., and Moy, P. 2006. Gene amplification of the *Hps* locus in *Glycine max*. BMC Plant Biol. 6:6.

Gonzalez, D.O. and Vodkin, L.O. 2006. Clustering analysis of transcript abundance in soybean cotyledons during germination and emergence. Plant and Animal Genome XIV, San Diego, CA, p. 295.

Shoemaker, R., Keim, P., Vodkin, L., Retzel, E., Clifton, S.W., Waterston, R., Smoller, D., Coryell, V., Khanna, A., Erpelding, J., Gai, X., Brendel, V., Raph-Schmidt, C., Shoop, EG., Vielweber, C.J., Schmatz, M., Pape, D., Bowers, Y, Theising, B., Martin, J., Dante, M., Wylie, T., Granger, C. 2002. A compilation of soybean ESTs: generation and analysis. Genome 45:329-338.

Shoemaker, R.C., Cregan, P.B., and Vodkin, L.O. 2004. Soybean Genomics. Pages 235-263 in Soybeans: Improvement, Production and Uses, 3rd edition, ASA monograph, J. Specht (ed). American Society of Agronomy, Madison, WI.

Stacey, G., Vodkin, L. O., Parrott, W., and Shoemaker, R.C., 2004. Draft plant for soybean Genomics. Plant Physiol 135: 59-70.

Stromvik, M. V., Thibaud-Nissen, F., and Vodkin, L.O., 2004. Mining soybean expressed sequence tag and microarray data. Ann Rev. Phytochemistry 38: 177-195.

Thibaud-Nissen, F., Shealy, R. T., Khanna, A., and Vodkin, L.O. 2003. Clustering of microarray data reveals transcript patterns associated with somatic embryogenesis in soybean. Plant Physiol 132:118-136.

Toda, K., Yang, D., Yamanaka, N., Watanabe, S., Harada, K., and Takahashi, R. 2002. A single-base deletion in soybean flavonoid 3’-hydroxylase gene is associated with gray pubescence color. Plant Mol. Biol. 50: 187-196.

Vodkin, L.O., Khanna, A., Shealy, R., Clough, S.J., Gonzalez, D.O., Philip, P., Zabala, G., Thibaud-Nissen, F., Sidarous, M., Strömvik, M.V., Shoop, E., Schmidt, C., Retzel, E., Erpelding, J., Shoemaker, R.C., Rodriguez-Huete, A.M., Polacco, J.C., Coryell, V., Keim, P., Gong, G., Liu, L., Pardinas, J., Schweitzer, P. 2004. Microarrays for global expression constructed with a low redundancy set of 27,500 sequenced cDNAs representing an array of developmental stages and physiological conditions of the soybean plant. BMC Genomics 5:73.

Walbot, V. 1999. Genes, genomes, genomics. What can plant biologists expect from the 1998 National Science Foundation Plant Genome Research Program. Plant Physiol. 119:1151-1155.

Zabala, G.C. and Vodkin, L.O. 2003. Cloning of the pleiotropic *T* locus in soybean and two recessive alleles that differentially affect structure and expression of the encoded flavonoid 3' hydroxylase. Genetics 163: 295-309.

Zabala, G. and Vodkin, L.O. 2005. The *wp* mutation of *Glycine max* carries a gene-fragment-rich transposon of the CACTA superfamily. Plant Cell 17: 2619-2632.

Zou, J., Rodriguez-Zas, S., Aldea, M., Li, M., Zhu, J., Gonzalez, D.O., Vodkin, L.O., DeLucia, E., Clough, S., J. 2005. Expression profiling soybean response to *Pseudomonas syringae* reveals new defense-related genes and rapid down regulation of photosynthesis. Mol Plant Microbe Interact. 18: 1161-1174.