AgBioForum, 6(1&2): 4-7. ©2003 AgBioForum.

The Status of Soybean Genomics and Its Role in the Development of Soybean Biotechnologies

Randy C. Shoemaker
USDA ARS, Iowa State University

Jessica A. Schlueter
Department of Zoology and Genetics, Iowa State University

Perry Cregan
USDA ARS Soybean Genomics and Improvement Laboratory

Lila Vodkin
Department of Crop Science, University of Illinois

The soybean is a major world source of edible oil and high-quality protein. It has an interesting and complex genome structure. It has a rich repertoire of genomic tools and resources that include a vast expressed sequence tag (EST) collection, a densely populated genetic map, a developing physical map, microarray resources, and an efficient transformation system. It also has a large and active research community. The array of resources available would be improved with a finished physical map and a better understanding of the gene space and chromosomal topography of the species. Still, the soybean is clearly the model crop legume.

Key words: biotechnology, genetics, genome, Glycine max, legume, model crop.

Introduction

Advances in molecular and computational technologies have irreversibly changed the course of genetic investigation for almost all organisms. Although "single-gene" and "single investigator" studies will continue to provide detail and resolution to our understanding of biological processes, the greatest impact in this field of investigation will likely come from the high-throughput discovery tools of genomics.

Advances in soybean genomics in the last 15 years promise to revolutionize soybean genetics. What follows is a brief report on key components of soybean genomics.

The Genome and Genetic Linkage Maps

The soybean genome is only average in size compared to that of many other plants. It comprises about 1,115 million base pairs per haploid genome (Mbp/C; Arumuganathan & Earle, 1991). Approximately 40-60% of the soybean genome sequence can be defined as repetitive (Gurley, Hepburn, & Key, 1979; Goldberg, 1978). Approximately 90% of all restriction fragment length polymorphism (RFLP) probes detect duplicated loci in soybean (Shoemaker et al., 1996). Nearly 60% detect three or more loci. Hybridization-based mapping has resolved many duplicated regions of the genome. These homoeologous regions reflect segmental and whole-genome duplication events and can provide much information about the evolution of the genome.

In 1990, the first RFLP-based map of the soybean genome was published (Keim, Diers, Olson, & Shoemaker, 1990). The genetic map saw further expansion during the 1990s with the addition of more than 350 RFLP loci (Shoemaker & Olson, 1993). The development and mapping of a large set of soybean simple sequence repeat (SSR) markers were initiated in 1995 with the support of the United Soybean Board (USB). As a result of this effort, more than 600 SSR loci were developed. The 600 SSR markers developed to date were mapped in three different mapping populations (Cregan et al., 1999). As a result, the 20-plus linkage groups derived from each of the three populations were aligned into a consensus set of 20 homologous groups presumed to correspond to the 20 pairs of soybean chromosomes.

Molecular marker development in soybean has progressed from RFLPs to SSRs and now to single nucleotide polymorphisms (SNPs). The frequency of SNPs in soybean is somewhat low—the SNP frequency in coding and noncoding DNA of approximately 1.98/kbp and 4.68/kbp, respectively, as estimated from the analysis of 25 soybean genotypes (Zhu et al., in press). However, SNPs are already in use in industrial soybean breeding programs (Cahill, 2000) using allele specific hybridization (ASH) for SNP detection similar to the procedure described by Coryell, Jessen, Schupp, Webb, and Keim (1999).

Comparative Genomics

The substantial rearrangements that have occurred within the soybean genome, probably as part of the process of diploidization, make it difficult to identify lengthy stretches of syntenic chromosome segments between soybean and related legumes. It has been determined that linkage groups of mung bean and common bean are comprised generally of mosaics of short soybean linkage blocks (Boutin et al., 1995). However, Lee, Bush, Specht, and Shoemaker (1999) showed that homoeologous segments of soybean linkage groups showed a higher degree of synteny with chromosomes of common bean and mung bean than previously thought. Lee et al. (1999) also showed that homologous regions among the legumes were also homologous with duplicated regions of *Arabidopsis*. Grant, Cregan, and Shoemaker (2000) used sequences of mapped soybean RFLP probes and *Arabidopsis* genomic sequence to demonstrate synteny between *Arabidopsis* and soybean. These findings were surprising, given the millions of years since the divergence of their lineages. In contrast, only three of 50 soybean contigs (6%) were shown to possess microsynteny with *Arabidopsis* (Yan et al., 2003), whereas 54% showed microsynteny with *Medicago truncatula*. Clearly, cross-referencing to model legumes (*Medicago* and *Lotus*) will speed soybean genomics advances.

Physical Mapping

Bacterial artificial chromosome (BAC) libraries (Marek & Shoemaker 1997; Danesh et al., 1998; Tomkins et al., 1999; Salimath & Bhattacharyya, 1999; Meksem, Ruben, Zobrist, Zhang, & Lightfoot, 2000) have been produced which together cover the soybean genome many times over. Detailed physical contigs have already been developed and reported (Marek & Shoemaker, 1997). These libraries have been made from different genotypes and with a variety of enzymes and almost all have been made available to the public. Yeast artificial chromosomes have also been created for the purpose of chromosome walking and in situ hybridization (Zhu, Shi, Gresshoff, & Keim, 1996).

A genome-wide physical map was recently constructed from more than 78,000 BAC clones (Wu et al., 2003). This map consisted of approximately 2,900 contigs. The total contig length exceeded the predicted size of the soybean genome, suggesting that many contigs overlapped. More than half of the physical length of the physical map was anchored to the genetic map using RFLP and SSR markers (Wu et al., 2003).

ESTs and Gene Discovery

As a result of funding from the soybean commodity boards (North Central Soybean Research Program [NCSRP] and USB), soybean has amassed more than 300,000 ESTs representing over 80 different cDNA libraries (Shoemaker et al., 2002). The cDNA libraries giving rise to those ESTs represent a wide range of organs, developmental stages, genotypes, and environmental conditions. This resource is providing much information on differences in gene expression of members of multigene families (Granger et al., 2002). The soybean EST collection provides a large resource of publicly available genes and gene sequences and provides valuable insight into structure, function and evolution of this model crop legume. This resource is also drawing much needed bioinformatic expertise into legume research.

Genome Sample Sequencing

Genome sequencing is fundamental to understanding the genetic composition of an organism. However, the entire genome need not be sequenced before critical information on the topography of the chromosomes can be obtained. Genomic sampling of nearly 2,700 DNA sequences from more than 600 mapped loci has provided a glimpse of the composition and general structure of the soybean genome (Marek et al., 2001). These contigs were identified and developed at genetically anchored SSR and RFLP loci. An additional 237,000 BAC-end sequences have been made available to the Better Bean Initiative (USB) by Monsanto, Inc. These sequences will be instrumental in defining soybean gene space and in creating a soybean repeat database useful for whole-genome sequencing efforts.

Functional Genomics

High-density expression arrays containing 18,000 cDNAs arrayed on a filter have been developed (Vodkin et al., 2000) along with microarray technology. Currently, arrays of 27,000 genes for soybean have been printed containing low-redundancy genes from a wide range of organs, developmental stages, disease challenged tissues, and various stress conditions (L. Vodkin, personal communication, September 23, 2003).

Serial analysis of gene expression (SAGE) captures short 10- to 20-nucleotide "tags" near the 3' end of individual mRNA molecules. The frequency of appearance of tags in the library has been shown to accurately estimate expression levels in the mRNA source tissue. Initial analysis from 20 SAGE libraries in soybean has resulted in 132,992 SAGE tags, of which 40,121 are unique (J. Schupp, personal communication, January 13, 2002).

The applications of microarray technology to soybean are enormous. The future of functional genomics research will include arrays that will distinguish gene family members. Full-length sequencing of cDNA clones is an important step in collecting the data necessary to take this next step.

Table 1. Categorization of BAC-end sequences of BACs identified with RFLP probes or SSR primers.

Sequence category	SSR (1,416 sequences)	RFLP (1,254 sequences)
No significant	35.8	44.2
Hits[b]	64.2	55.8
Repetitive[c]	33.5	18.4
Genes[d] and rDNA	6.1	10.0
Soybean ESTs[e]	6.8	10.3
Hypothetical[f]	17.9	17.0

Note. Comparison of the distribution of BLAST hits from the SSR- and RFLP-identified BAC-end sequences is shown as a percentage of total SSR or RFLP sequences. A minimum E value of 10^{-6} was used to assign identities. Sequences were placed in only one category. From "Soybean Genomic Survey: BAC-end Sequences Near RFLP and SSR Markers," by L.F. Marek et al., 2001, Genome, 44, pp. 572-581.

[a] *Sequences with either no similarity to database sequences or with similarities above the minimum E value cutoff of 10^{-6}.*

[b] *Sequences with similarity to database sequences at or below the minimum E value cutoff of 10^{-6}.*

[c] *Sequences with similarity to database sequence annotated to be retrotransposon- or transposon-like or with similarity to novel, possibly soybean unique, repetitive-type sequence identified.*

[d] *Sequences with similarity to experimentally described gene sequences.*

[e] *Sequences with similarity to a soybean EST database.*

[f] *Sequences either identified with gene prediction software Diogenes from the University of Minnesota or show similarity to predicted genes in Arabidopsis.*

Genetic Transformation

Soybean has an efficient and well-developed genetic transformation system (Clemente et al., 2000; Xing et al., 2000; Zhang et al., 1999). Soybean transformation efficiencies are consistently greater than 5%, and one report states success rates in excess of 12%.

What Is Missing?

The soybean as a genetic system is still lacking key components. An understanding of the organization, complexity, and distribution of the gene space of an organism, including the topography of its repetitive sequences, is critical to efficient generation of whole-genome sequences. Although a glimpse of the distribution of genic and repetitive sequences in soybean has been seen (Marek et al., 2001), a detailed analysis is lacking. This gap could be filled by a combination of cytogenetic analyses (fluorescence in situ hybridization [FISH] and fiber FISH) and targeted (gene-rich or topographical) sequencing projects. Also required is a functional genomic analyses resource that distinguishes duplicated genes. Finally, phenotypical functional genomics systems—particularly gene knockout systems—are also needed in soybean. Improvements in transformation efficiencies have led to development of transposon tagging projects for soybean (T. Clemente, personal communication, May 20, 2003). Viral-induced gene silencing systems and TILLING (targeted induced local lesions in genomes) populations for soybean are under development (N. Nielsen, personal communication, September 8, 2003). In spite of the progress in these areas, phenotypical functional analysis systems are not yet deployed for soybean.

Soybean is the number one oilseed crop in the world and provides a multi-billion-dollar source of high-quality protein. The rich genomic resources available for soybean make it a model crop legume. The gene discovery stemming from structural and functional genomics research in soybean will certainly lead to new products and to varieties with improved nutritional and agronomic characters.

References

Arumuganathan, K., & Earle, E.D. (1991). Nuclear DNA content of some important plant species. *Plant Molecular Biology Reporter*, *9*, 208-219.

Boutin, S.R.Y., Young, N.D., Olson, T.C., Yu, Z.-H., Shoemaker, R.C.. & Vallejos, E.C. (1995). Genome conservation among three legume genera detected with DNA markers. *Genome*, *38*, 928-937.

Cahill, D. (2000, August). High throughput marker assisted selection. *Proceeding of the 8th Biennial Conference on the Cellular and Molecular Biology of the Soybean.* Lexington, KY.

Clemente, T., LaValle, B.J., Howe, A.R, Ward, D.C., Rozman, R.J., Hunter, P.E., Broyles, D.L., Kasten, D.S., & Hinchee, M.A. (2000). Progeny analysis of glyphosate selected transgenic soybeans derived from Agrobacterium-mediated transformation. *Crop Science*, *40*, 797-803.

Coryell, V.H., Jessen, H., Schupp, J.M., Webb, D., & Keim, P. (1999). Allele-specific hybridization markers for soybean. *Theoretical and Applied Genetics*, *98*, 690-696.

Cregan P.B., Jarvik, T., Bush, A., Shoemaker, R.C., Lark, K.G., Kahler, A., Kaya, N., VanToai, T., Lohnes, D.G., Chung, J., & Specht, J.E. (1999). An integrated genetic linkage map of the soybean genome. *Crop Science*, *39*, 1464-1490.

Danesh, D., Penuela, S., Mudge, J., Denny, R. Nordstrom, H., Martinez, J., & Young, N.D. (1998). A bacterial artificial chromosome library for soybean and identification of clones near a major cyst nematode resistance gene. *Theoretical and Applied Genetics*, *96*, 196-202.

Goldberg, R.B. (1978). DNA sequence organization in the soybean plant. *Biochemical Genetics*, *16*, 45-68.

Granger, C., Coryell, V., Khanna, A., Keim, P., Vodkin, L., & Shoemaker, R.C. (2002). Identification, structure, and differential expression of members of a BURP domain containing protein family in soybean. *Genome*, *45*(4), 693-701.

Grant, D., Cregan, P., & Shoemaker, R.C. (2000). Genome organization in dicots: Genome duplication in Arabidopsis and synteny between soybean and Arabidopsis. *Proceedings of the National Academy of Sciences, USA, 97*(8), 4168-4173.

Gurley, W.B., Hepburn, A.G., & Key, J.L. (1979). Sequence organization of the soybean genome. *Biochimica et Biophysica Acta*, *561*, 167-183.

Keim, P., Diers, B.W., Olson, T.C., & Shoemaker, R.C. (1990). RFLP mapping in soybean: Association between marker loci and variation in quantitative traits. *Genetics*, *126*, 735-742.

Lee, J.M., Bush, A., Specht, J.E., & Shoemaker, R.C. (1999). Mapping of duplicate genes in soybean. *Genome*, *42*, 829-836.

Marek, L.F., Mudge, J., Darnielle, L., Grant, D., Hanson, N., Paz, M., Huihuang, Y., Denny, R., Larson, K., Foster-Hartnett, D., Cooper, A., Danesh, D., Larsen, D., Schmidt, T., Staggs, R., Crow, J.A., Retzel, E., Young, N.D., & Shoemaker, R.C. (2001). Soybean genomic survey: BAC-end sequences near RFLP and SSR markers. *Genome*, *44*, 572-581.

Marek, L.F., & Shoemaker, R.C. (1997). BAC contig development by fingerprint analysis in soybean. *Genome*, *40*, 420-427.

Meksem, K., Ruben, E., Zobrist, K., Zhang, H.-B., & Lightfoot, D. (2000). Two large insert libraries for soybean: Application in cyst nematode resistance and genome wide physical mapping. *Theoretical and Applied Genetics*, *101*, 747-755.

Salimath, S., & Bhattacharyya, M.K. (1999). Generation of a soybean BAC library, and identification of DNA sequences tightly linked to the Rps1-k disease resistance gene. *Theoretical and Applied Genetics*, *98*, 712-720.

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, E.G., 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., & Olson, T. (1993). Molecular linkage map of soybean. In S. O'Brien (Ed.), *Genetic maps: Locus maps of complex genomes* (6th ed.). Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Shoemaker, R.C., Polzin, K., Labate, J., Specht, J., Brummer, E.C., Olson, T., Young, N., Concibido, V., Wilcox, J., Tamulonis, J.P., Kochert, G., & Boerma, H.R. (1996). Genome duplication in soybean (Glycine subgenus soja). *Genetics*, *144*, 329-338.

Tomkins, J.P, Mahalingam, R., Smith, H., Goicoechea, J.L., Knap, H.T., & Wing, R.A. (1999). A bacterial artificial chromosome library for soybean PI 437654 and identification of clones associated with cyst nematode resistance. *Plant Molecular Biology*, *41*(1), 25-32.

Vodkin, L.O., Khanna, A., Clough, S., Shealy, R., Philip, R., Erplending, J., Paz, M., Shoemaker, R., Coryell, V., Schupp, J., Keim, P., Rodriquez-Huete, A., Zeng, P., Polacco, J., Mudge, J., Denny, R., Young, N., Raph, C., Shoop, L., & Retzel, E. (2002). Structural and functional genomics projects in soybean. *Plant Molecular Biology Reporter Supplement*, *18*(2), S1.

Wu, C., Sun, S., Nimmakayala, P., Santos, F., Springman, R., Ding, K., Meksem, K., Lightfoot, D., & Zhang, H.-B. (2003). *A BAC and BIBAC-based physical map of the soybean genome*. Manuscript submitted for publication.

Xing, L., Ge, C., Zeltser, R., Maskevitch, G., Mayer, B.J., & Alexandropoulos, K. (2000). c-Src signaling induced by the adapters Sin and Cas is mediated by Rap1 GTPase. *Molecular Cell Biology*, *20*(19), 7363-77.

Yan, H., Mudge, J., Kim, D.J., Shoemaker, R., Cook, D., & Young, N. (in press). Estimates of conserved microsynteny among the genomes of Glycine max, Medicago truncatula and Arabidopsis thaliana. *Theoretical and Applied Genetics*.

Zhang, H., Yu, L., Mao, N., Fu, Q., Tu, Q., Gao, J., & Zhao, S. (1999). Cloning, characterization, and chromosome mapping of RPS6KC1, a novel putative member of the ribosome protein S6 kinase family, to chromosome 12q12-q13.1. *Genomics*, *61*(3), 314-8.

Zhu, T., Shi, I., Gresshoff, P., & Keim, P. (1996). Characterization and application of soybean YACs to molecular cytogenetics. *Molecular and General Genetics*, *252*, 483-488.

Zhu, Y.-L, Song, Q.-J., Hyten, D.L., Van Tassell, C.P., Matukumalli, L.K., Grimm, D.R., Hyatt, S.M., Fickus, E.W., Young, N.D., & Cregan, P.B. (2003). Single nucleotide polymorphisms (SNPs) in soybean. *Genetics*, *163*, 1123-1134. Available on the World Wide Web: http://bldg6.arsusda.gov/~pooley/soy/cregan/Zhu_et_al_20031.pdf.