SoyMap: an integrated map of soybean for resolution and dissection of multiple genome duplication events.

Scott A. Jackson[1], Rod A. Wing[2], Gary Stacey[3], Gregory May[4] and Randy C. Shoemaker[5]

Contact information:

1. Department of Agronomy, Purdue University, 915 W. State St., West Lafayette, IN 47907

Email: sjackson@purdue.edu

Fax: 765-496-7255

2. Arizona Genomics Institute, University of Arizona, Tucson, Arizona 85721

3. National Center for Soybean Biotechnology, Divisions of Plant Science and Biochemistry, Department of Molecular Microbiology and Immunology, 271E Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211

4. National Center for Genome Resources, 2935 Rodeo Park Drive East, Santa Fe, New Mexico 87505

5. Corn Insect and Crop Genetics Research Unit, USDA-ARS and Iowa State University, Ames, Iowa, 50011

ABSTRACT

SoyMap is an NSF-funded project to create an integrated genetic and physical map of soybean as a community resource for gene cloning, evolutionary/structural genome studies and as a predicate to determine an intelligent genome sequencing approach. Although determination of sequencing approach has to some extent been obviated by the DOE-JGI soybean genome initiative (whole genome shotgun approach), the integrated map is important nonetheless to anchor the sequence map to the chromosome and linkage maps—making the sequence map functional for the community. The SoyMap project is an outgrowth of a commodity board-funded physical map of soybean: two BAC libraries were fingerprinted using high information content fingerprinting (HICF) and genetic markers were placed on the physical map. SoyMap continues this effort by expanding the library coverage with a third library, additional makers placed on the physical map to further integrate the genetic and physical maps, and preliminary sequencing to understand genome structure at duplicated regions and as quality control for the current shotgun sequencing approach.

GOALS AND PROGRESS

I. Physical mapping of soybean genome

Integration of physical map with genetic map

In order to make the physical map immediately useful for cloning genes or QTLs and to tie the eventual genome sequence to a genetic framework, the physical map consisting of contiged BAC clones is being anchored to the genetic map via hybridization of genetic markers to BAC clones (Hass-Jacobus et al., 2006) and screening of pooled BAC clones via a PCR approach. In total, more than 3,000 markers, many on the genetic map and

others derived from ESTs, will be placed on the physical map via this approach. This will result in a high-density of markers and transcripts on the physical map that tie it to the genetic map so that researchers can easily shuttle between the two maps.

Fluorescence in situ Hybridization (FISH) mapping of BAC clones and repeat elements

Another component of integrating the various soybean maps is to anchor linkage groups to chromosomes via FISH mapping of genetically anchored BAC clones to soybean chromosomes. Several chromosomes have been associated to linkage groups via trisomic analysis (Zou et al., 2003), but this is a slow and tedious approach. We have already developed a set of BAC clones that mark 15 out of the 20 chromosome arms and are associated with genetic markers (Pagel et al., 2004; Walling et al., 2005). Moreover, repeat elements (http://www.soymap.org/data/misc/soy_repeats.fasta) that were identified computationally from a random set of shotgun sequences (Lin et al., 2005) are being mapped to chromosomes via FISH to learn more about the structure of the soybean genome—find the genic space by exclusion by mapping the repeat elements.

Sequence Tag Connector (STC) database

The final part of developing and refining the physical map is the development of a STC database consisting of BAC end sequences (BES) derived from fingerprinted BAC clones from the three BAC libraries used in development of the physical map. This provides a rich resource for (1) sequence analysis of the genome, (2) anchoring the sequence map to the physical map via sequence alignments; (3) spanning gaps between sequence contigs from the WGSS (whole genome shotgun sequence); (4) anchoring to related genomes to understand chromosome evolution (OMAP paper); and (5) marker development to anchor 'floating', unanchored, contigs to the genetic map by either SSRs (simple sequence

repeat) or SNPs (single nucleotide polymorphism). The STC will result in a sampling of ~15% of the soybean genome sequence in paired reads (~130 kb) from either ends of the BAC clones.

II. Sequencing of test regions to understand genome structure

Tests regions for assembly of WGSS

At least five megabase-size regions will be selected from the physical map and sequenced from the cognate BACs. This contiguous sequence will be used by the genome sequencing group to test assemblies of the WGSS for quality assurance. The megabase-sized contigs will be selected from the FPC map focusing on contigs that are well-anchored and the underlying assembly is not ambiguous; thereby, resulting in regions that will provide good instances to test the quality of the WGSS assemblies. In the course of this activity approximately 50 BACs will be sequenced to Phase III quality.

Sequencing of Duplicated regions

A major concern is how duplicated regions in the soybean genome (Shoemaker et al., 1996; Schlueter et al., 2004; Walling et al., 2005; Schlueter et al., 2006) may confound the genome assembly due to misplacement of sequences due to homology between paralogous sequences. Therefore, in combination with United Soybean Board (USB) funding, several duplicated regions are being, or have been, sequenced to determine the level and range of sequence conservation between duplicated regions (Schlueter et al., 2006). Preliminary computational and structural analyses of these duplicated regions indicates that these will cause few, if any, problems for the genome assemblers due to the level of nucleotide variation between duplicated sequences.

Heterochromatin structure and transition

Approximately 40-50% of the soybean genome consists of repeated DNA elements. Thus, for producing a sequenced genome, we are really only interested in capturing 500-600 Mbps. Indeed, estimates derived from enrichment of hypomethylated (with reduced repeat content) soybean sequences suggest that the gene space may occupy only 340 Mbps (Nunberg et al., 2006). However, we are only now beginning to gain an understanding of the organization of the repetitive sequences within the soybean genome. FISH has shown that most of the high-copy sequences tend to be localized in pericentromeric regions (Lin et al., 2005; Walling et al., 2005). We still know little about the transition of euchromatin (gene-rich) to heterochromatin (gene-poor) regions at the sequence level. Moreover, we cannot exclude the possibility of genic islands buried in heterochromatin or even heterochromatic knobs in euchromatin. WGSS assemblers exclude repetitive sequences as they can interfere with proper assembly due to misalignment of repeats. Thus, if there are high-copy, dispersed repeats in a genome, these can result in misassembled sequence contigs. Randomly chosen BACs and BACs chosen due to repeat content are being sequenced to assess the distribution of repeats across the genome and to characterize the repeat structure of the soybean genome. To date, seven randomly chosen BACs have been sequenced and are being characterized for repeat content and structure.

III. Display of data for community at the Legume Information System (LIS)

All data will be integrated into the Legume Information System (LIS) (Gonzales et al., 2005). For more information see May et al. in this issue for more information on data structure and display at LIS.

IV. Integration with DOE effort

The NSF SoyMap project described herein works closely and in parallel with the DOE-JGI soybean genome sequencing effort headed by D. Rokhsar. Coordination of the SoyMap project is done in conjunction with JGI to complement the sequencing effort so that at the end of the day we can deliver a genome sequence that is linked to the genetic and physical maps; thereby, making it as useful as possible for the soybean biologist.

OUTLOOK

As the community embarks on a genome project we should step back and redress a few questions. The broader legume community indicated that soybean should be a reference legume genome for the Phaseoloid clade of legumes. First, what is implied by the 'reference genome'? One definition is that this is a genome to which other, probably less complete, genomes are compared. It follows, therefore, teleologically, that the reference genome should, to the extent possible, be complete. The second question is what will the soybean genome look like upon completion of this phase of sequencing? The answer to this question, based on other whole genome shotgun sequences, is that, although it will computationally capture most of the genes, it will be fragmentary and largely unanchored. Thus, the third question is how should the soybean community gear up to obtain a reference genome? This is an area that the community should ponder and begin to make plans on how to finish the genome shotgun sequence that will be produced in 2008 and also how to annotate it in a functional sense: gene annotations, functional annotation of genes via mutagenesis, microarrays and so forth. Prioritization of activities should be done always keeping in mind the specific goal of crop improvement.

Acknowledgements

Funding was provided by the National Science Foundation (DBI-0501877) and the United Soybean Board.

Gonzales, M.D., Archuleta, E., Farmer, A., Gajendran, K., Grant, D., Shoemaker, R., Beavis, W.D., and Waugh, M.E. (2005). The Legume Information System (LIS): an integrated information resource for comparative legume biology. Nucl. Acids Res. **33,** D660-665.

Hass-Jacobus, B.L., Futrell-Griggs, M., Abernathy, B., Westerman, R., Goicoechea, J.L., Stein, J., Klein, P., Hurwtiz, B., Zhou, B., Rakhshan, F., Sanyal, A., Gill, N., Lin, J.Y., Walling, J.G., Luo, M.Z., Ammiraju, J.S., Kudrna, D., Kim, H.R., Ware, D., Wing, R.A., SanMiguel, P., and Jackson, S.A. (2006). Integration of hybridization-based markers (overgos) into physical maps for comparative and evolutionary explorations in the genus Oryza and in Sorghum. BMC Genomics **7,** 199.

Lin, J.Y., Jacobus, B.H., SanMiguel, P., Walling, J.G., Yuan, Y., Shoemaker, R.C., Young, N.D., and Jackson, S.A. (2005). Pericentromeric regions of soybean (Glycine max L. Merr.) chromosomes consist of retroelements and tandemly repeated DNA and are structurally and evolutionarily labile. Genetics **170,** 1221-1230.

Nunberg A, Bedell JA, Budiman MA, Citek RW, Clifton SW, Fulton L, Pape D, Cai Z, Joshi T, Nguyen H, Xu D, Stacey G (2006) Survey sequencing of soybean elucidates the genome structure and composition. Functional Plant Biol. **33**, 765-773.

Pagel, J., Walling, J.G., Young, N.D., Shoemaker, R.C., and Jackson, S.A. (2004). Segmental duplications within the Glycine max genome revealed by fluorescence in situ hybridization of bacterial artificial chromosomes. Genome **47,** 764-768.

Schlueter, J.A., Scheffler, B., Schlueter, S.D., and Shoemaker, R.C. (2006). Sequence conservation of homeologous BACs and expression of homeologous genes in soybean (Glycine max L Merr). Genetics**,** genetics.105.055020.

Schlueter, J.A., Dixon, P., Granger, C., Grant, D., Clark, L., Doyle, J.J., and Shoemaker, R.C. (2004). Mining EST databases to resolve evolutionary events in major crop species. Genome **47,** 868-876.

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., and Boerma, H.R. (1996). Genome duplication in soybean (Glycine subgenus soja). Genetics **144,** 329-338.

Walling, J.G., Shoemaker, R.C., Young, N.D., Mudge, J., and Jackson, S.A. (2005). Chromosome level homeology in paleopolyploid soybean (Glycine max) revealed through integration of genetic and chromosome maps. Genetics**,** genetics.105.051466.

Zou, J., Lee, J., Singh, R., Xu, S.S., Cregan, P.B., and Hymowitz, T. (2003). Assignment of molecular linkage groups to the soybean chromosomes by primary trisomics. Theor Appl Genet **107,** 745-750.