Published March, 2011

Soybean Genomics Research Program Strategic Plan

Implementing Research to Meet 2012–2016 Strategic Milestones

Edited by Roger Boerma (University of Georgia, Athens GA),
Richard Wilson (Oilseeds & Bioscience Consulting, Raleigh NC),
and Ed Ready (United Soybean Board, St. Louis MO)

Executive Summary

This strategic plan builds on the soybean communities' previous efforts (October, 1999; July, 2001; May, 2003; July, 2005; and May, 2007) to review progress on the development and deployment of soybean genomic resources. The results are impressive (see *Soybean Genomics Research Program Accomplishments Report, 2010*, available at http://soybase.org/SoyGenStrat2007/SoyGenStratPlan2008-2012-Accomplishments%20v1.6.pdf [verified 17 Mar. 2011]). For example, in the last 5 yr the soybean research community has produced a genetic linkage map with over 5500 mapped markers spanning the entire 2296 cM soybean genome. A set of 1536 SNP markers that are evenly distributed across the 20 linkage groups was developed for whole genome analysis of polymorphisms in both elite North American cultivars and breeding lines. In addition, an expanded array of 50,000 SNPs is under development which will be used to create haplotype maps of over 18,000 accessions of the USDA soybean germplasm collection. This research is scheduled for completion in late 2010 and the SNP haplotype map of each accession will be placed on the HapMap Browser on SoyBase.

Large-scale shotgun sequencing of the soybean cultivar Williams 82 was completed late in 2008 by the U.S. Department of Energy Joint Genome Institute (DOE-JGI) and recently reported in the scientific journal *Nature* (Schmutz et al., 2010). The present soybean assembly (Glyma.1.01) captured approximately 975 Mbp of its 1100 Mbp genome. The gene set integrates ~1.6 million ESTs with homology and predicts 66,153 protein-coding loci available at http://www.phytozome.net/soybean (verified 17 Mar. 2011).

Soybean researchers have developed several microarray technologies for gene expression studies. The GeneChip Soybean Genome Array is commercially available for studying gene expression (http://www.affymetrix.com/products_services/arrays/specific/soybean.affx#1_1 [verified 17 Mar. 2011]). This GeneChip contains 37,500 *Glycine max* transcripts, 15,800 *Phytophthora sojae* transcripts, and 7500 *Heterodera glycines* transcripts.

The achievement of milestones in previous strategic plans for soybean genomic research have advanced soybean to its current status as a crop model for translational genomics. Simply stated, soybean genomic resources in hand will accelerate the ability of plant breeders to enhance soybean productivity, pest resistance, and nutritional quality. However, many secrets of the soybean genome have yet to be revealed. To continue to make informed decisions it was critical to capture the consensus wisdom of leading soybean researchers on the next logical steps in the development and utilization of soybean's genomic resources.

On 27–28 July 2010 Roger Boerma chaired a workshop sponsored by the United Soybean Board in St. Louis MO that brought together 44 eminent soybean researchers in the areas of genomic sequencing, gene function, transformation/transgenics, and translational genomics. The purpose of the Workshop was to develop a strategy for achieving the critical soybean genomic resources and information required to accelerate the rate of yield gain and addition of value to U.S. soybean cultivars. A consensus was reached on a number of high priority performance measures or research objectives. In addition the anticipated outcomes of successfully achieving these performance measures are included in the final plan.

Overall, two issues emerged as being critically important or overarching issues: (i) Provide additional support staff for continued development and population of SoyBase, and (ii) Development of a genetic repository/distribution center for soybean mutants/transgenic lines. The enhancement of SoyBase was deemed important for all four Strategic Goals. The genetic repository/distribution center was broadly supported by Workshop participants. Listed below is an outline of the four Strategic Goals and their respective Performance Measures. Within each Goal, the Performance Measures are listed in order of importance.

Published in The Plant Genome 4:1–11. Published 17 Mar. 2011.
doi: 10.3835/plantgenome2011.12.0001

5585 Guilford Rd., Madison, WI 53711 USA
An open-access publication

Goal 1: Genome Sequence: Improve the Quality and Utility of the Soybean Genome Sequence

Performance Measure:

1.1: Ensure the accuracy of reference sequence assembly.

1.2: Capturing and leveraging existing genetic diversity in soybean germplasm.

1.3: Improving bioinformatic resources for genomic analysis and practical applications.

1.4: Reveal function of targeted genome sequences to facilitate gene discovery and application.

1.5: Leveraging genomic information from Phaseoloids and other species.

1.6: Determine the role of epigenetics in soybean improvement.

Goal 2: Gene Function: Develop Functional Genomic Technologies to Optimize Utility of Genome Sequence Information in Germplasm Enhancement

Performance Measure:

2.1: Develop comprehensive gene expression data for soybean.

2.2: Develop near isogenic lines (NIL) to help reveal genetic mechanisms that mediate useful traits.

2.3: Develop an improved infrastructure to facilitate genome annotation.

2.4: Achieve high-definition genomic characterization of biological mechanisms and regulatory systems in soybean.

2.5: Use functional genomic methods to characterize transcription regulated pathways.

2.6: Advance gene modification technologies to help associate candidate genes with a discrete phenotype.

2.7: Create a saturated transposon insertion population with defined flanking sequences that can be used to identify mutants by BLAST sequence comparison.

2.8: Implement outreach opportunities for education and use of genomic databases.

2.9: Develop an ORFeome library from agronomically important genes and gene families.

Goal 3: Transformation/Transgenics: Optimize and Expand Transgenic Methods and Improve Understanding of Natural Genes for Modification of Trait Expression

Performance Measure:

3.1: Establish of a soybean genetic repository and distribution center.

3.2: Develop next-generation transformation and targeting technologies and utilize these transgenic approaches to help elucidate gene function and deploy genes of interest.

Goal 4: Translational Genomics: Optimize Breeding Efficiency with Robust Sequence-based Resources

Performance Measure:

4.1: Develop analytical approaches to characterize soybean germplasm diversity based on the SoyHapMap 1.0 data to identify parental lines for breeding purposes.

4.2: Discover gene/QTL for qualitative traits and develop tightly linked DNA markers.

4.3: Discover gene/QTL for quantitative traits and develop tightly linked DNA markers.

4.4: Develop and populate a user-friendly database of validated QTL for use in marker assisted breeding applications.

4.5: Define the molecular genetic signatures of selection in 70+ years of U.S. soybean breeding by use of the 50,000 SNP Illumina Infinium Assay.

4.6: Define optimum breeding models for different breeding situations using *in silico* analysis.

Our Collective Wisdom: How We Got Where We Are

The chromosome-scale draft assembly of the soybean (*Glycine max* L. Merr.) genome is an outcome of a dynamic, technology driven, and timely strategic process whose origin may be traced formally to the *Soybean Genomics White Paper* (Boerma et al., 2000). That January 2000 document was a product of a 21–22 Oct. 1999 meeting of 17 experts in plant genomics, DNA markers, plant transformation, and bioinformatics. A consensus was reached on research priorities in the area of soybean genomics. Milestones included: (i) doubling Simple Sequence Repeat (SSR) markers within 3 yr; (ii) expansion of Single Nucleotide Polymorphism (SNP) markers to 10,000 within 3 to 5 yr; (iii) improving the efficiency of soybean transformation; (iv) tagging 80% of the genes in the soybean genome within 3 to 5 yr; (v) integration of genetic, physical, and transcript maps of soybeans within 3 to 5 yr; and (vi) employing comparative genomics to define the structure and attributes of the soybean genome.

It became mutually beneficial to establish research coalitions to improve the efficiency of genomic investigations among related species. For this reason, the U.S. Legume Crops Genomics Initiative (LCGI) was organized under the auspices of the American Soybean Association, United Soybean Board, National Peanut Foundation, USA Dry Pea and Lentil Council, the National Dry Bean Council, and the Alfalfa Council to facilitate communication and cooperation among scientists with an interest in genomic research on soybean,

peanut, pea and lentil, common bean, alfalfa, and model-legume crops. LCGI was founded on the premise that the development of an integrated legume genomics research system would enhance ability to leverage information across legume crops and model species and the first workshop was convened on 30–31 July 2001 at Hunt Valley, MD. Twenty-six legume scientists developed a white paper (Boerma et al., 2001) that outlined high-priority research in the areas of: (i) genome sequencing of strategic legume species; (ii) physical map development and refinement; (iii) functional analysis: transcriptional and genetic; (iv) development of DNA markers for comparative mapping and breeding; (v) characterization and utilization of legume biodiversity; and (vi) development of a legume data resource. The nature of this cooperative interaction not only ensured timely research progress in all legume crops associated with the Initiative, but also enhanced the competitive position of the LCGI within the framework of the National Plant Genome Initiative, which is coordinated by the Interagency Working Group on Plant Genomics, Committee on Science, National Science and Technology Council.

Implementation of a coordinated effort for research and development of genomics across the legume family facilitated progress in the model species *Medicago truncatula* and *Lotus japonicus* and in soybean (*Glycine max*), and accentuated the need to transfer genomic information from the model species to cool-season pulses [pea (*Pisum sativum*), lentil (*Lens culinaris*), chickpea (*Cicer arietinum*), field bean (*Vicia faba*)], and warm-season food legumes [peanut (*Arachis hypogea*), common bean (*Phaseolus vulgaris*)], and forage legumes [alfalfa (*Medicago sativa*), clover (*Trifolium* spp.)]. This mission was codified further by (i) a third white paper, entitled: *Legumes as a Model Plant Family: Genomics for Food and Feed*; and (ii) by the publication of the monograph, *Legume Crop Genomics.*

The white paper was an outcome of the CATG (Cross-legume Advances through Genomics) conference on 14–15 Dec. 2004 in Sante Fe, NM with funding from the National Science Foundation (Plant Genome Research Program) and the USDA (National Research Initiative). About 50 individuals in attendance represented the respective legume communities as well as various funding agencies. The objectives of the conference were to: (i) identify a unifying goal for an international cross-legume genome project; (ii) identify cross-cutting themes to help integrate the different legume crop genomics programs, including a unified legume genomics information system, nutritional and health-related aspects of legumes, and detailed synteny and comparative genomics of legumes; and (iii) outline specific components and milestones for the initiative. These deliberations identified four tiers of legume species, each with specific genomic resources to be developed. Based on phylogenetic arguments, and particularly the degree of synteny, two major foci of legumes were identified, the hologaleginoid clade or cool-season legumes and the phaseoloid/millettioid clade or warm-season legumes. In each of these two foci, one or two reference species were identified, *M. truncatula* and *L. japonicus* in the former and soybean in the latter. Development of a full range of genomics resources, including sequencing of the entire genome, was the highest priority for these reference species. For a second group, common bean and peanut, a broad range of genomic resources were recommended, including a physical map, BAC-end sequencing and marker development, anchoring of the genetic and physical map, ESTs of the major organs, chip resources, and sequencing of gene rich regions. A third group consisted of all other legume crops in the two foci, including pea, lentil, chickpea, field bean, clover, cowpea (*Vigna unguiculata*), and pigeon pea (*Cajanus cajan*). For these legumes, translational genomic tools were to be developed, principally for cross-legume markers, species-specific recombinant inbred lines, genetic maps, EST and BAC libraries. A fourth group included other legumes not in the two main foci, such as members of the basal legume clades. An abbreviated version of this white paper was published. (Gepts et al. 2005).

As each of the legume crop communities began to amass critical genomic resources, it became necessary for each community to develop priorities that were crop specific. The strategic foundation for soybean genomics research was laid on 20–21 May 2003, when 19 researchers participated in a workshop. The scientists reviewed the current status of soybean genomic research and reached consensus on a strategic framework for outlining research priorities and significant near-term milestones. This input was captured in the *Strategic Plan for Soybean Genomics 2003–2007.* (available at http://soybase.org/SoyGenStrat2005/Soy_Genome_Strat_Plan_2005.html [verified 17 Mar. 2011]) Coordination of plan implementation was delegated through election of the Soybean Genetic Executive Committee (SoyGEC). The SoyGEC took action to move forward aggressively on genomic resources including physical mapping, EST anchoring, and functional genomic resources. The SoyGEC (http://soybase.org/resources/soygec.php [verified 17 Mar. 2011]) also initiated strategies to proactively communicate soybean research community priorities to representatives of federal granting agencies, and encouraged coordination of dedicated research teams to solve soybean problems of national and international importance. A National Science Foundation (NSF)-sponsored workshop was held in St. Louis on 21 Oct. 2003 to take an inventory of the current genomic resources in soybean, identify areas where more data were needed, and to set a research a strategy to advance soybean genomics research. Special attention was focused on research opportunities provided by unique aspects of soybean biology. The workshop included academic, governmental, and industrial scientists covering a wide variety of specialties related to both basic and applied research on soybean, along with scientists from outside the soybean field who provided general expertise in genomics and represented a wealth

of experience garnered from other genomic projects. Representatives from federal funding agencies and soybean commodity groups participated as observers. This workshop extended and further defined the findings from earlier workshops, which had surveyed the status and priority goals for soybean & legume genomics (see http://soybase.org/Genetic_Resources/Soybean_Genetic_Resources.html [verified 17 Mar. 2011]) (Stacey et al., 2004).

Action to address the need for bioinformatic resources was taken by the National Center for Genome Research (NCGR) and USDA, ARS, NPS when the U.S. Congress established the Model Plant Initiative (MPI) to use bioinformatics to leverage genomic information from model plant species. The Legume Information System (LIS), a joint NCGR and ARS effort, was one of the first projects in support of the MPI to demonstrate the integration of genomic information from *Arabidopsis thaliana*, *Medicago truncatula* and *Lotus japonicus* to important agronomic legumes, e.g., soybean, alfalfa, pea, dry beans.

Since its inception, LIS development has been guided by ideas and suggestions from the legume research communities. Based on user needs, USDA-ARS and NCGR developed a plan to amplify the power of this information resource for identification of candidate genes, unique genes, and evolutionary relationships among genes for crop improvement. It was agreed to utilize LIS as a foundation for a Comparative Legume Biology (CLB) Program. CLB expanded LIS from a passive data management system to a platform for novel data analysis and visualization tools. With the emergence of high throughput biotechnologies and bioinformatics, comparative biology enabled more sophisticated analyses of genomic sequences, genomic maps, micro-arrays, protein arrays, metabolic arrays, genetic regulatory networks, and biochemical and whole organism phenotypes.

The SoyGEC convened the first assessment of research performance against the *Strategic Plan for Soybean Genomics 2003–2007* on 19–20 July 2005 in St. Louis, MO. Presentations provided updates on the current status of soybean resources and related genomics technologies. Stakeholder input was generated in facilitated discussion groups to assess the status of soybean genomics, identify needs, and identify milestones to achieve objectives. The discussion groups included the general areas of Functional Genomics A (Transcriptome and Proteome), Functional Genomic B (Reverse Genetics), Physical and Genetic Maps, and Bioinformatics. Several topics received overwhelming support in all group reports. A high quality physical map in Williams 82 and integration with the physical map of 'Forrest' was a very high priority. A whole-genome sequence was an expectation of the research community. The need for standardization of protocols, terminologies and ontologies was evident as interactions among groups and among research communities expand. Finally, there was urgent need to establish long-term facilities that have the capability to archive, maintain, generate and provide biological resources. Overall, there was consensus that the ongoing research was on target, if not ahead of schedule, and still relevant to the goals and objectives of the *Strategic Plan for Soybean Genomics 2003–2007*.

On 20 Jan. 2006 the USDA and DOE announced plans to share resources and coordinate the study of plant and microbial genomics. The soybean genome was first on the list for sequencing. Dr. Ari Patrinos, DOE Associate Director for SBER, directed the Joint Genome Institute (JGI) to carry out the work with the Stanford Human Genome Center.

A Genome Strategic Planning Workshop to consider research goals and priorities for the next 5-yr programmatic cycle was convened by the SoyGEC in St. Louis, MO on 30–31 May 2007 by James Specht (University of Nebraska-Lincoln). This workshop was attended by 48 experts in diverse genomic areas. Whereas previous soybean strategic plans had dealt with preparation for the eventual sequencing of the genome, the announcement that the soybean genome would be sequenced by DOE-JGI necessitated a reassessment of strategic objectives for the next research plan. Jeremy Schmutz, Stanford, the leader of the DOE-JGI soybean sequence assembly effort reported that the work was proceeding exceptionally well, despite the ancient polyploidy of a now well-diploidized soybean. A 4X shotgun genome sequence of the genome had been developed and a draft assembly had been created. This assembly was being evaluated to determine the optimal means for obtaining the final goal of an 8X coverage. The 8X sequence was slated to be completed by the end of 2007, with a final assembly expected to be completed in mid-2008. A gratifying number of objectives and milestones identified in the 2005 Plan had been achieved on schedule. The number of SNPs and STSs proposed for discovery and development was exceeded. Inbred mapping resources (RILs) were developed. Transformation technologies had improved and various gene knock-out systems were working. USB and NSF funding enabled work on physical and transcript maps. Bioinformatic resources and staffing had more than doubled, in time to receive the whole genome shotgun sequence. These technological developments and the diminishing cost of many genomic-based technologies were taken into full account in the development of the *Strategic Plan for Soybean Genomics 2008–2012*.

The SoyGEC reached out to colleagues abroad at the first meeting of the International Soybean Genome Consortium at he NIAS in Tsukuba, Japan on 20 April 2007. This ongoing association has helped leverage resources for soybean genomic research in Japan, Korea, China, and the USA.

This current report summarizes the most recent strategic plan for soybean genomics research. This plan was developed at a workshop held on 27–28 July 2010 in St. Louis MO. The workshop was convened by Roger Boerma and included 44 soybean researchers in the areas of genomic sequencing, gene function, transformation/

transgenics, and translational genomics. The purpose of the workshop was to develop a consensus strategy for achieving the critical soybean genomic resources and information required to accelerate the rate of yield gain and addition of value to U.S. soybean cultivars. This plan, *Soybean Genomics Research Program Strategic Plan: Implementing Research to Meet 2012–2016 Strategic Milestones* documents the high priority Performance Measures or research objectives, and anticipated outcomes of successfully achieving Strategic Goals for soybean genomic research in the next 5 yr. Within each Goal, the Performance Measures are listed in order of importance.

Strategic Goals for Soybean Genomics Research (2012–2016)

Goal 1: Genome Sequence: Improve the Quality and Utility of the Soybean Genome Sequence

Performance Measure 1.1: Ensure the accuracy of reference sequence assembly. Quality control is essential for useful reassembly of genome sequences. Accuracy of the reference genome impacts the effectiveness of all subsequent investigations of gene discovery, identification of gene function, comparative genomics, and characterization of the nature of genetic diversity in soybean. The first chromosome scale draft sequence of *Glycine max* is of very high quality. However, there are regions within the genome that remain ambiguous. These regions may contain genes that are important to soybean improvement. Therefore, research is needed to correct portions of the assembly.

Anticipated products:

- Closure of remaining gaps in sequence contigs/scaffolds.
- Correction of the order and orientation of improperly aligned scaffolds/contigs.
- Optical maps to orient contigs/scaffolds and estimate gaps sizes and place unanchored sequence contigs.
- Additional clone libraries targeted to span gaps and/or to help orient contigs/scaffolds.
- New sequencing technologies to generate long-range sequence reads.

Performance Measure 1.2: Capturing and leveraging existing genetic diversity in soybean germplasm. Genotypic diversity is the basis for genetic enhancement of soybean. Linkage disequilibrium and other statistical measures suggest that much of the variation for useful traits within the USDA germplasm collection remain untapped because of the difficulties in effectively identifying genetic differences. However, advances in DNA sequencing technology will enable the high-definition characterization of genotypic differences among germplasm accessions, cultivars, and breeding lines.

Anticipated Products:

- Resequencing a subset of the U.S. germplasm collection (including the ancestral land races) that captures greater than 90% of the genetic variation in *G. max* will access genetic diversity on a base-pair to base-pair level.
- Haplotype maps of the entire USDA *G. max* and *G. soja* collections to identify rare, potentially valuable genes/alleles.
- *De novo* sequencing (near reference level sequencing) of up to 10 selected *G. max* lines to estimate presence-absence variation (PAV), copy number variants (CNV), genes that are not present in the reference genome, and other structural variation.

Performance Measure 1.3: Improving bioinformatic resources for genomic analysis and practical applications. As a consequence of advances in sequencing technology, the amount of the soybean genomic and associated data is growing at exponential rates. Data storage limitations are now eclipsed by problems that limit the curator's ability to compile, analyze, and interpret data in a useful and timely manner. There is a need for expansion of an integrated database for use by all researchers, including breeders, to enhance the utility of genome sequence data.

Anticipated Products:

- Criteria for decision tools for parental selection and distinguishing genotypes during breeding population development (See Goal 4).
- Software, protocols, and database support for sequence-based genotyping.
- Expansion of SoyBase to completely integrate all relevant information using almost any concept as an entry point. For example starting with a gene name or function, a user should be able to quickly find the position(s) of that gene on the genetic and sequence maps, information about that gene, and links to the scientific literature for more details.

Performance Measure 1.4: Reveal function of targeted genome sequences to facilitate gene discovery and application. High-definition mapping of DNA sequences in quantitative trait loci (QTL) enables the construction of allele specific markers for genetic traits. However, many genes may be present in a QTL region. Association of candidate gene sequences with a specific biological function or response not only facilitates marker development, but also expands knowledge of the biological mechanisms that mediate agricultural traits in soybean. Accurate annotation of the reference sequence is needed to improve the efficiency and utility of gene discovery.

Anticipated Products:

- Definition of entire gene families, 5' to 3', and alternative transcripts.
- Validated gene models with Transcription Start Sites (TSSs) for 20,000 genes.
- Refined accuracy of the transposon element database (TEdb) to recover genes embedded in transposable elements.

· Refined annotation of pseudogenes (e.g., truncated or early stop codons) that may have transcriptional evidence.

· Definition of regulatory elements (transcription factor binding sites, micro RNAs, etc.) that control gene expression.

· Annotation of known promoter elements for all identified motifs, prediction of unknown elements, and identification of co-regulated genes.

· Small RNA sequences from a range of tissues to understand the role of small RNAs in gene regulation.

· Functional annotation of soybean gene sets to facilitate gene discovery and genetic improvement.

· Collection and integration of functional information for all published gene mutant/descriptions.

· Collection and integration of transcriptome data with genetic and genomic resources in SoyBase, the USDA-ARS soybean genetic database (http://soybase.org).

Performance Measure 1.5: Leveraging genomic information from Phaseoloids and other species. *Phaseolus vulgaris* (dry bean) and *Vigna unguiculata* (cowpea) are diploid species with synteny to soybean. *Phaseolus* and *Vigna* both diverged from *Glycine* about 19–23 MYA, and may be very useful for the discovery of genes for protection against abiotic stresses in soybean. Genome sequencing of both species is underway. The perennial *Glycine* species are also of interest because of the nature of polyploidy in *Glycine* as a whole, and thus for understanding the duplicated nature of the soybean genome. These species diverged from soybean (and its progenitor, *G. soja*) around 5 MYA.The perennial species constitute the secondary germplasm pool for soybean, and may provide traits such as drought tolerance and rust resistance.

Anticipated Products:

· Identification of genetic variability for nutritional traits in common bean and cowpea that may be applicable to soybean.

· Discovery of the genes or quantitative trait loci (QTL) with DNA markers to facilitate transfer to elite cultivars.

· Sets of SNP panels specific to appropriate populations of common bean and cowpea.

· Genetically and chromosomally stable interspecific populations between soybean and *G. tomentella* for resequencing and identification of diploids with resistance to pest and pathogens.

Performance Measure 1.6: Determine the role of epigenetics in soybean improvement. Epigenetics represents non-Mendelian inheritance of phenotypic traits. This area of plant biology is poorly understood. However, a growing body of knowledge suggests a hierarchical stratification in the regulation of gene expression that involves a complex interaction among gene products.

Anticipated Products:

· Enhanced knowledge base for understanding the role of epigenetics in mediating gene expression.

· Understanding the contribution of epigenetic phenomena to phenotypic diversity.

Goal 2: Gene Function: Develop Functional Genomic Technologies to Optimize Utility of Genome Sequence Information in Germplasm Enhancement

Performance Measure 2.1: Develop comprehensive gene expression data for soybean. Current annotation of the soybean genome suggests at least 45,000 protein encoding genes. While genome sequencing reveals all of the genes present within an organism, it cannot tell us what specific gene products are needed for different cellular pathways, tissues, or organs. For example, leaves and roots contain the same DNA, yet their very different structures are the result of differences in gene expression. It is laborious to examine the expression of genes on a gene-by-gene basis. Fortunately, new high-throughput sequencing platforms provide a rapid and sensitive means to survey gene expression. Limited soybean gene expression atlases are now available and have already been utilized to study the expression of some genes. However, these resources need to be expanded to include many more specific tissues and, more importantly, environmental treatments, especially those of agronomic importance (e.g., drought or insect predation). The availability of a comprehensive, soybean expression atlas that encompasses all such tissues and treatments would be an invaluable resource for the study of soybean gene function.

Anticipated Products:

· An improved soybean gene atlas developed from RNA-seq approaches, which includes a comprehensive list of all expressed soybean genes, alternative splice products, and the identification of co-regulated genes and gene networks.

· RNA-seq data representing 100 different tissues and treatments.

· A standardized methodology for submitting data toward the whole soybean genome annotation effort.

Performance Measure 2.2: Develop near isogenic lines (NIL) to help reveal genetic mechanisms that mediate useful traits. The application of the full repertoire of functional genomic tools to soybean promises to yield new and important insights that can be mined for soybean improvement. However, these tools can be fully utilized only in well controlled experiments that minimize experimental variation, include replication, and provide rigorous statistical analysis. An important variable is genotypic variation. However, this can be reduced by the use of NILs varying only in key alleles that control important agronomic traits. The availability of such NILs is essential for soybean to fully benefit from the modern molecular tools now available.

Anticipated Products:

- Development of collaborations between functional genomicists and breeders to identify traits.
- Develop 50 sets of NILs in the next 2 to 5 yr for traits that NILs do not currently exist.

Performance Measure 2.3: Develop an improved infrastructure to facilitate genome annotation. SoyBase is a comprehensive repository for professionally curated genetics, genomics, and related data resources for soybean. SoyBase contains genetic, physical, and genomic sequence maps integrated with qualitative and quantitative traits. SoyBase also contains the 'Williams 82' genomic sequence and associated data-mining tools. The genetic and sequence views of soybean chromosomes and the expansive data on traits and phenotypes are extensively interlinked. This allows entry to the database using almost any kind of available information, such as genetic map symbols, soybean gene names, or phenotypic traits. SoyBase is the repository for controlled vocabularies for soybean growth, development, and trait terms, which are also linked to the more general plant ontologies. Annotation of the draft chromosome scale sequence of the soybean genome with gene functions associated with QTL facilitates allele specific marker development. However, the utility of these resources in breeding and other areas of science depends on coordination of data assimilation into bioinformatic systems and training in the practical operation of those resources.

Anticipated Products:

- An improved gene annotation that uses available RNA-seq data and other bioinformatic methods.
- Computational methods to acquire data from existing sequence and expression databases (e.g., Gene Expression Omnibus) for use in future genome annotation releases.
- Improved access to genome annotations through Soybase and Phytozome.
- Updated annotation improvements released every 12 to 18 mo.
- A gene expression database management tool populated with internal experimental data.
- Physical map viewer (WebFPC) populated with internal experimental data.
- SoyKB (Soybean Knowledge Base; http://soykb.org/ [verified 17 Mar. 2011]) which integrates all forms of soybean functional data with the genome sequence.

Performance Measure 2.4: Achieve high-definition genomic characterization of biological mechanisms and regulatory systems in soybean. Transcriptomics involves analysis of gene transcription. There is need to build developmental stage-specific catalogs of mRNA expression for key plant organs and biological processes, and improved ways of relating proteomics and transcription products to the annotated soybean genome sequence. Soybean biochemical processes give rise to a plethora of metabolites, each with its own spectrum of activity. The biological regulation of a great majority of these compounds is unknown.

Anticipated Products:

- High throughput methods for soybean such as RNA-seq, CHiP-seq, methylome, sRNAs, and protein covalent modification.
- Advanced knowledge of soybean protein-protein interaction networks with an emphasis on proteomic methods such as TAP-tag approaches.
- Targeted proteomic and metabolomic approaches focused on key soybean traits (e.g., oil and protein).
- Identification and characterization of soybean metabolites expressed in key tissues (e.g., seeds) or in response to environmental stresses and pathogens.

Performance Measure 2.5: Use functional genomic methods to characterize transcription regulated pathways. Information about gene expression does not lead automatically to new biological understanding. It is clear that networks exist for the control of gene expression and can reflect higher order complexity (e.g., coordinate interaction of transcription factor complexes). Elucidating these networks is essential to fully understanding how environment and genotype interactions lead to observable phenotypes (e.g., yield). Fortunately, modern molecular methods provide experimental means to elucidate these networks, their interactions, and outcomes.

Anticipated Products:

- A defined soybean transcriptional regulatory pathway obtained by combining RNA-seq data with ChIP-seq data.
- A defined soybean epigenetic regulatory pathway obtained by integrating sRNA expression and targets, the methylome, and histone ChIP-seq.

Performance Measure 2.6: Advance gene modification technologies to help associate candidate genes with a discrete phenotype. Although sequence similarity-based gene annotation may suggest a function, it is necessary to confirm this function through biochemical or genetic studies. It is also expected that the function of the majority of soybean genes will not be easily deduced simply by sequence comparison to other genomes. In these cases, the availability of mutations in each of the soybean genes will be extremely useful to decipher gene function and integrate this function into the context of soybean quality and agronomic performance. For example, TILLING (Targeting Induced Local Lesions IN Genomes) is a PCR-based high-throughput mutation detection system that permits the identification of point mutations and small insertions and deletion "indels" in pre-selected genes. Fast neutron mutagenesis induces small deletions in the genome and is a very effective way to create mutations. Fast-neutron populations can be used for both forward and reverse genetic screens for useful mutations. A variety of new technologies are needed to create robust platforms for the study of soybean gene function.

Anticipated Products:

- Optimized protocols for use of RNAi and VIGS technology high-throughput systems.
- Robust site-directed tools such as zinc finger-based mutagenesis for gene function in soybean.
- Lines derived from fast neutron mutagenesis and a gene deletion database for characterization of gene function.
- High throughput sequencing methods and other analytical tools to identify deletions and characterize deleted genes.
- Cost-effective implementation of resequencing to improve sustainability of TILLING resources.

Performance Measure 2.7: Create a saturated transposon insertion population with defined flanking sequences that can be used to identify mutants by BLAST sequence comparison. Transposons– known as McClintock's 'jumping genes'- are ubiquitous in plant genomes. Transposon insertion has a high likelihood of disrupting gene function. Usually only a few transposon insertions occur in any given individual, making genetic analysis much easier. Recent work has demonstrated that both the maize Ac/Ds and rice mPing transposons are suitable for use in soybean. The Ac/Ds transposon has a strong preference for local transposition, which makes it particularly useful to mutate genes in its vicinity, and may be particularly well suited for activation tagging. However, a starting population of well dispersed Ds insertions (perhaps 25,000) is necessary before any given soybean gene can be targeted. In contrast, the mPing preferentially targets gene-rich regions on all chromosomes, and germinal insertions occur in the absence of tissue culture. Strategies are available that may increase its frequency of germinal transposition even more.

Anticipated Products:

- Identification of the optimal transposon system for activation tagging.
- Strategies to increase frequency of germinal transposition.
- A transposon-tagged population in which 85% of annotated soybean genes have been tagged.
- Improved methods for high-throughput phenotyping of transposon-tagged lines.
- A comprehensive database of transposon-tagged lines

Performance Measure 2.8: Implement outreach opportunities for education and use of genomic databases. Many valuable tools have been/will be developed for soybean researchers. These tools include genome sequence, genetic, expression, and proteomic databases. However, these databases are useful only if they can easily be used and queried. Given the complexity of the data, training scientists to use and mine these databases will be an asset to the soybean community. These educational and outreach activities need to include multiple formats to reach an ever expanding and diverse audience.

Anticipated Products:

- Standardized syntax and sanctioned methods for data submission to informatic resources.
- Standardized syntax and expanded inclusion of phenotypic descriptors for useful trait informatic resources.
- Leveraged comparative informatics across data sets within and among legume species.
- Web services that allow and encourage communication between databases.
- Training opportunities via meetings and web-based tutorials to familiarize the community with database resources.
- Improved database interactions facilitated by integration and improvement of existing software platforms.

Performance Measure 2.9: Develop an ORFeome library from agronomically important genes (tissue or treatment specific) and gene families. Many molecular methods developed to examine an individual gene's function require the cloning of the gene of interest or its corresponding cDNA. For example, Virus Induced Gene Silencing (VIGS) requires a portion of the targeted cDNA be inserted into a viral construct. When targeting hundreds of individual genes, it is difficult to clone the correct gene or cDNA fragment in a high-throughput manner. An ORF (Open Reading Frame) clone contains all the protein-coding portions of a gene, minus the untranslated regions found in cDNAs that can inhibit some molecular studies. A library of cloned ORFs in Gateway-compatible vectors would facilitate studies of gene localization, protein-protein interactions, epitope tagging, and become a resource to the soybean community.

Anticipated Products:

- Use of cloned genes of interest to evaluate the usefulness of an ORFeome.

Goal 3: Transformation/Transgenics: Optimize and Expand Transgenic Methods and Improve Understanding of Natural Genes for Modification of Trait Expression

Performance Measure 3.1: Establish a soybean genetic repository and distribution center. A permanent repository for community-wide genetic resources is needed for maintenance and distribution of the valuable unique germplasm that the soybean genetics community generates. The public has invested millions of dollars on soybean research for the development of mapping populations, TILLING, and transposon-tagged lines. This investment is in danger of being lost due to the lack of an appropriate infrastructure for long-term storage and distribution. The existence of such a repository is becoming critical to leverage additional research funds. In fact, future funding from Federal agencies could be jeopardized if there is no place to store the materials developed with Federal funding. If necessary to attract long-term funding, this genetic repository could be

combined with repository for additional crops; legumes and non-legumes.

Anticipated Products:

- Identification of a coordinator and advisory board to develop a business plan and find stable long-term financing from public and private sources.
- Construct a sustainable facility and bioinformatics database.
- Determine seed storage conditions and protocols for quality control and resource distribution.
- Consolidation of existing resources into the central facility.
- Web-based resources made available publically, with defined submission and request requirements.

Performance Measure 3.2: Develop next-generation transformation and targeting technologies and utilize these transgenic approaches to help elucidate gene function and deploy genes of interest. Transformation is central to most of the advances that are currently in place in farmers' fields, from introduction of specific genes of interest to determination of gene function. Although these advances were ultimately provided by the private sector, the initial discoveries of genes and gene introduction technologies originated in the public sector. With the availability of the soybean sequence, transformation continues to be an important resource for the soybean community to use to dissect gene function as it relates to the genes of interest to soybean breeders. In addition, improvements in gene introduction efficiencies and quality of transformants, as well as generation of new targeting technologies, are still desperately needed to generate novel germplasm to be utilized by the soybean breeding and genetics community.

Anticipated Products:

- Improvements in gene introduction/gene expression efficiencies.
- More rapid and efficient recovery of transgenics, especially homozygous seed identification.
- Established "rules of transgene assembly for desired expression" with promoter/termination identification of 100 events per year.
- RNAi construct toolbox for targeted silencing of 2 to 5 vectors implemented per year.
- Determine how to assemble multiple stacks (multiple gene constructs) for consistent expression.
- Training opportunities for use of efficient transformation technology.
- Capacity for medium-throughput characterization of gene function.
- Establishment of a Transformation Consortium to coordinate a transformation pipeline for genes of interest to the research community.
- Establishment of a web portal/database for the process of gene identification, submission, and transgenic delivery.
- Genetic modification and insertion of genes that mediate productivity, protection, or quality traits.
- Development of targeting/transposon technologies for site-directed cutting or integration.

Goal 4: Translational Genomics: Optimize Breeding Efficiency with Robust Sequence-based Resources

Performance Measure 4.1: Develop analytical approaches to characterize soybean germplasm diversity based on the SoyHapMap 1.0 data to identify parental lines for breeding purposes. The analysis of the entire USDA Soybean Germplasm Collection of 18,000^{+} cultivated soybean and 1100^{+} wild soybean accessions with 50,000 SNP DNA markers is the only such analysis of a germplasm collection. The resulting development of SoyHapMap 1.0, which will provide the initial definition of soybean genetic diversity based on the analysis of the 19,000+ cultivated and wild accessions in the USDA Soybean Germplasm Collection, will be the basis for the discovery of new genetic diversity and DNA marker resources. To maximize the usefulness of this unprecedented crop-haplotype dataset, multivariate approaches and tools are needed to summarize the genomic data into recognizable patterns of diversity.

Anticipated Products:

- A diversity map of in the USDA Soybean Germplasm Collection.
- Genomic analysis tools to maximize the exploitation of soybean haplotype diversity for genetic improvement.
- A maximally diverse Soybean Core Collection.
- Germplasm with enhanced resistance to abiotic stresses.
- Germplasm with enhanced yield potential.
- Public access to all developed SNPs.

Performance Measure 4.2: Discover gene/QTL for qualitative traits and develop tightly linked DNA markers. These are traits that are controlled by one or a few genes and underlie traits such as pest resistance, seed fatty acid and amino acid levels, certain other seed composition traits, and transgenes.

Anticipated Products:

- DNA markers that can be used in breeding for pest resistance including SCN, aphid, Phytophthora root rot, and other important pests.
- DNA markers that can be used in breeding for quality traits including seed fatty acid and amino acid levels and other seed composition traits (phytate, raffinose, etc.).
- DNA markers for mapping and background selection for transgenes: In most cases the developer of the transgenic will provide a perfect marker for the transgene as well as the genome position.

Performance Measure 4.3: Discover gene/QTL for quantitative traits and develop tightly linked DNA markers. These are traits that are controlled by a number of genes/QTL and underlie traits including seed yield, seed protein/oil levels, abiotic stress resistance, and disease tolerance.

Anticipated Products:

- DNA markers for seed protein/oil levels that can potentially be used to stack these seed composition genes/QTL without sacrificing yield potential.
- DNA markers for abiotic stress tolerances including slow wilting, N_2–fixation under water deficit, iron deficiency chlorosis, flooding tolerance, salt tolerance, water use efficiency, root morphology, and response to higher levels of CO_2 and ozone.
- DNA markers for seed yield discovered via traditional QTL mapping and networked association mapping approaches.

Performance Measure 4.4: Develop and populate a user-friendly database of validated QTL for use in marker assisted breeding applications. The successful application of marker-assisted breeding requires access to robust data on marker-trait associations for both perfect and flanking markers associated with genes/QTL conditioning the phenotypic variation in the desired qualitative or quantitative trait.

Anticipated Products:

- A defined format and standards for QTL and marker-trait association data to expedite database entry of marker-trait association data.
- Required submission and compliance that QTL and marker-trait association data are entered to SoyBase for all USB-funded projects.
- Development of training workshops, distance learning, and webinars to introduce and instruct breeders and students on the use of SoyBase (see also Performance Measure 2.8).

Performance Measure 4.5: Define the molecular genetic signatures of selection in 70+ years of U.S. soybean breeding by use of the 50,000 SNP Illumina Infinium Assay. Selection by soybean breeders over the past 70 yr for seed yield and other agronomic traits has resulted in allele frequency changes at various places in the soybean genome. The identification of such regions will indicate the positions of genes controlling these agronomic traits.

Anticipated Products:

- Availability of the allele frequency at 50,000 SNP loci based on analysis with the 50,000 SNP Illumina Infinium Assay of the ancestors and important resistance sources used in elite breeding populations, recently released public cultivars, and the most recent Uniform Test entries.
- Analysis of changes in SNP allele frequency over time will provide information on which portions of the soybean genome have been impacted by continuous breeder selection.
- Identification of key genes/alleles for important agronomic and quality traits.

Performance Measure 4.6: Define optimum breeding models applicable to different breeding situations using in silico analysis. To achieve the maximum benefit from haplotype information and marker-trait associations, it is critical for soybean breeders to understand the optimum approaches to efficiently introgress genes/QTL to create elite soybean cultivars. The evaluation of various breeding models using in silico analysis will define optimal approaches applicable to different breeding situations and establish soybean as the model for the study of efficient breeding of a self-pollinated crop. We now have soybean genotypes that will mature in 70 d or less days after emergence. To establish a model soybean system, the soybean community should select 20 diverse plant introductions and/or important land mark cultivars and convert them via backcrossing to very early maturity. These converted types will be the basis establishing soybean as a model crop for study.

Anticipated Products:

- Optimum procedure for the introgression of a single gene/QTL into soybean based on computer modeling.
- Optimum procedure for gene stacking into soybean based on computer modeling.
- Optimum procedure for the introgression of a polygenic trait into soybean based on computer modeling.
- Optimum procedure for the introgression genes/QTL into soybean while maintaining background genetic diversity based on computer modeling.
- A model soybean population for in silico studies.

Workshop Participants by Assigned Subareas

Genomic Sequence

David Hyten, USDA-ARS
Jianxin Ma, Purdue University
Jeremy Schmutz, Hudson Alpha Institute for Biotechnology
Steve Canon, USDA-ARS
Leah McHale, The Ohio State University
Bob Stupar, University of Minnesota
Jessica Schlueter, University of North Carolina- Charlotte
Matt Hudson, University of Illinois
David Grant, USDA-ARS
Ed Ready, United Soybean Board
Scott Jackson, Purdue University

Gene Function

Melissa Mitchem, University of Missouri
Steve Clough, University of Illinois
Ed Cahoon, University of Nebraska- Lincoln
Gary Stacey, University of Missouri
Michelle Graham, USDA-ARS

Julian Chaky, Pioneer Hibred International Inc.
Lila Vodkin, University of Illinois-Urbana
Randy Shoemaker, USDA-ARS
Trupti Joshi, University of Missouri
Roy Scott, USDA-ARS
Ivan Baxter, USDA-ARS

Transformation/Transgenics

John Finer, The Ohio State Univ
David Somers, Monsanto
Ted Klein, DuPont Agricultural Biotechnology
Tom Clemente, University of Nebraska- Lincoln
Paula Olhoft, BASF Plant Science
Jack Widholm, University of Illinois-Urbana
Wayne Parrott, University of Georgia
Theresa Musket, University of Missouri
Jennifer Jones, United Soybean Board

Translational Genomics

Henry Nguyen, University of Missouri
Tommy Carter, Jr., USDA-ARS
Katy Rainey, Virginia Tech
George Graef, University of Nebraska- Lincoln
Dechan Wang, Michigan State University
Jim Specht, University of Nebraska- Lincoln
Saghai Maroof, Virginia Tech
Rich Wilson, Oilseeds & Biosciences Consulting
Kevin Matson, Monsanto
Glenn Bowers, Syngenta Seeds
Perry Cregan, USDA-ARS
Bill Beavis, Iowa State University
Roger Boerma, University of Georgia

Writing Team

David Grant
Scott Jackson
Gary Stacey
Michelle Graham
Perry Cregan
Tommy Carter, Jr.
Theresa Musket
John Finer
Bob Stupar
Ed Ready
Rich Wilson
Roger Boerma

Acknowledgments

This workshop was partially funded by the United Soybean Board. This report was prepared with the support of the United Soybean Board projects 0202, 0276 and 0306. We wish to express our appreciation to Ann Chase for her efforts in arranging the excellent meeting facilities, meal functions, and workshop materials. The meeting was skillfully facilitated by Ed Ready and Roy Scott.

Special Acknowledgment

During the workshop it was realized this would likely be one of Dr. Ed Ready's last major assignments as Program Manager with the United Soybean Board's Production Committee before his well-earned retirement. The participants of this workshop wish to express their appreciation to Dr. Ready for his dedication, professionalism, and unprecedented ability to work through complex issues and arrive at the best solution for the U.S. soybean industry. As a research portfolio manager, he was known as a "quick study" that possessed the ability to communicate highly complex technologies in understandable language to a wide range of audiences. As a part of the larger soybean community we wish to say, "Thanks Ed!"

References

Boerma, H.R., D. Buxton, M. Kelly, and K. Van Amburg. 2000. Soybean genomics white paper January 2000. Available at http://soybase.org/Genomics/Soybean_Genomics.html (verified 17 Mar. 2011).

Boerma, H.R. J. St. John, and J. Yezak Molen. 2001. U.S. legume crops genomics workshop white paper. Available at http://www.legumes.org/ (verified17 Mar. 2011).

Gepts, P., W.D. Beavis, E.C. Brummer, R.C. Shoemaker, H.T. Stalker, N.F. Weeden, and N.D. Young. 2005. Legumes as a model plant family. Genomics for food and feed Report of the Cross-Legume Advances through Genomics Conference. Plant Physiol. 137:1228–1235.

Schmutz, J.S., B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, D.L. Hyten, et al. 2010. Genome sequence of the palaeopolyploid soybean. Nature 463:178–183.

Stacey, G., L. Vodkin, W.A. Parrott, and R.C. Shoemaker. 2004. National Science Foundation-Sponsored Workshop Report. Draft Plan for Soybean Genomics. Plant Physiol. 135:59–70.