Meeting Report: Soybean Genomics Assessment and Strategy Workshop
19 - 20 July 2005
St. Louis, Missouri

Writing Team:
Randy Shoemaker, USDA-ARS, Ames, Iowa
Wayne Parrott, University of Georgia, Athens, AG
Henry Nguyen, University of Missouri, Columbia, MO

Soybean Genetics Executive Committee:
James Specht, University of Nebraska, Lincoln, NE (Chair)
Randy Shoemaker, USDA-ARS, Ames, Iowa (past Chair)
Brian Diers, University of Illinois, Urbana, Illinois
Gary Stacey, University of Missouri, Columbia, MO
Randall Nelson, USDA-ARS, Urbana, IL
Perry Cregan, USDA-ARS, Beltsville, MD (past SoyGEC member)
Roger Boerma, University of Georgia, Athens, GA (past SoyGEC member)

Discussion Leaders:
Khalid Meksem, Southern Illinois University
Wayne Parrott, University of Georgia
Henry Nguyen, University of Missouri
Basil Nikolau, Iowa State University

INTRODUCTION

On July 19 - 20, 2005, approximately 50 researchers and administrators with expert knowledge of soybean genomics participated in a workshop in St.Louis, MO, which was hosted by the Soybean Genetics Executive Committee and supported by the United Soybean Board. The workshop began with a series of presentations by experts in the topics discussed below. Each presentation was designed to update the audience on the current status of soybean resources and related genomics technologies.

Following the presentations the participants divided into 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. After each discussion section the entire group reconvened to hear group reports and to further discuss each topic.

Several topics received overwhelming support throughout the Workshop and became evident in almost all discussions. A high quality physical map in the genotype Williams

82 and integration with the physical map of Forrest is critical to the success of many future advances and therefore, this remains a very high priority in soybean genomics. A whole-genome sequence of soybean is an expectation of the research community. The imperative to undertake and successfully complete this goal cannot be understated. The need for standardization of protocols, terminologies and ontologies is becoming more and more evident as interactions among groups and among research communities expands. And finally, the need to establish long-term facilities that have the capability to archive, maintain, generate and provide biological resources is an urgent priority that must be addressed soon.

The following is the report from this Workshop. It represents a consensus of the participants of the Workshop and it is structured to integrate with a White Paper generated in 2003 so that progress can be better monitored over time. The results of this report are consistent with those of a National Science Foundation soybean genomics workshop held in 2004 (St. Louis, MO) and a Cross-Legume workshop also held in 2004 (Santa Fe, NM).

DNA Markers

DNA markers are among the most versatile tools to emerge from genomics projects. They form the foundation of genetic linkage mapping and association analysis. Molecular markers are used for QTL/gene discovery and cloning, to anchor the physical map onto the genetic map, and as tools for assessing molecular variation within and between species. The current soybean molecular genetic linkage map contains several thousand markers (SSRs, RFLPs, SNPs and classical markers). Future targets for advances in soybean DNA marker technology are outlined below:

STP 1.1 Development of Single Nucleotide Polymorphism (SNP) Markers

SNPs are specific changes in DNA sequence that occur in genes as well as in intergenic regions. They can serve as biallelic genetic markers and when present in genes may alter gene function. In soybean there is approximately four SNPs per 1000 bp in a set of diverse soybean germplasm accessions.. This translates into more than 4 million potential SNP loci in the soybean genome. SNP genetic markers are relatively abundant, adaptable to high throughput detection, and cost effective in comparison to other DNA marker technologies. As of mid-2005 more than 6000 SNPs have been discovered in soybean, thus exceeding previous goals. Of these, more than 900 are already mapped and the community is completing work with industry to map another 1,000 generated from genes. Ninety percent of the discovered SNPs are present in a panel of six North American reporter genotypes.

Goals for 2007:

Identify an additional 5000 SNP containing STS (total of 11000) and genetically map the polymorphic SNPs in the currently available mapping populations. This will increase SNP density to 1 SNP/0.5 cM.

Form a SNP database and central coordination site.

Goals for 2009:

Complete the mapping of 10,000 SNPs.

Have available skim sequences of alternative genotypes to facilitate SNP discovery.

STP 1.2 Development of Sequence Tag Sites (STS) for Cross Legume Analysis

Sequence tagged sites are specific DNA fragments that can be amplified from genomic DNA. Concomitant with the discovery of SNPs, a large number of soybean STS will be identified. Some soybean STS will be used to amplify homologous DNA fragments in other legume species. Identification of cross-species STS will enable studies of synteny across the legume family. This will facilitate the useful translation of genomic information from model species, such as *Medicago truncatula*, as well as benefiting genetics and genomics research in other important crop legume species. As mid-2005 more than 500 STS are available to be mapped for soybean and Medicago truncatula. HAPPY mapping was tested and deemed not feasible at this time.

Goals for 2007:

Identify 2000 gene-based STS common to soybean, M. truncatula, and common bean. These additional STS will be used to refine syntenic associations among other legume species.

Begin mapping STS in soybean and common bean.

STP 1.3 Development of Inbred Mapping Resources

Introgression lines contain single specific segments of the genome of a donor parent in a common background. For soybean the donor can be another *G. max* genotype, a *G. soja* line, or in rare cases, a related Glycine species. These lines allow the examination of donor DNA fragments in a common genetic background and the creation of useful genetic diversity. Introgression and recombinant inbred lines (RIL) provide a resource for gene discovery, QTL analysis, and positional cloning. The development of a set of introgression and RIL lines requires backcrossing and extensive molecular analysis. In mid-2005 a number of backcross derived populations originating from different G. max and G. soja parents and one mating of a northern elite cultivar and a southern elite soybean cultivar are underway. Already available are NIL pairs from each of the RILs used in the Essex x Forrest population (and 2 related populations).

Goal for 2007:

Develop large RIL populations from the matings of Williams 82 x G. soja and Forrest X Williams 82. These populations will be useful for dissecting genetic events associated with domestication and for improved gene discovery.

Goal for 2009:

Have mapping data available from above lines.

STP 1.4 Application of Association Genetics to Gene Discovery in Germplasm

Association genetics provides the opportunity to discover genes/quantitative trait loci (QTL) via direct germplasm evaluation thus bypassing the need for specially developed mapping populations. Association genetics depends upon the presence of linkage disequilibrium and relies on existing linkage between a marker(s) and a gene(s) controlling the trait of interest in an existing group of genotypes such as the germplasm lines available within the USDA Soybean Germplasm Collection. This association is detectable if one has a large number of DNA markers that are stable over evolutionary time (SNPs). Relative to a mapping population, association analysis of a diverse group of genotypes should lead to a more precise estimate of the genomic position of the gene(s) controlling the phenotypic trait. It was the consensus of participants at this workshop that for the immediate future application of association genetics is best done in the private sector.

Goals for 2007:

Identify and collect data for 500 - 1000 SNP markers on a core set of 100 genotypes (or greater) that represent a substantial range in phenotypic diversity.

Begin characterization and identification of 500 SNP haplotypes for traits of biotic and abiotic stress, seed quality and composition and other value added traits through sequencing and EcoTILLING.

Goals for 2009:

Increase the number of SNP haplotypes for traits of biotic and abiotic stress, seed quality and composition and other value added traits through sequencing and EcoTILLING to 1000.

PLANT GENETIC TRANSFORMATION

Soybean transformation has shown significant improvement and enabled public and private sector production of commercial cultivars with transgenic traits. Advances in the utility of transformation methods in soybean have resulted from the development of selectable marker-free transgenic soybean lines, multiple gene delivery systems,

transformation and regeneration of elite cultivars, and tissue-specific and inducible promoters.

The public sector has met the 2005 benchmark of being able to produce 400 plants per person per year if need be, and is on target to meet the 2007 benchmark of 500 plants per person per year. One area of concern though, is the current and pending reduction of the number of public soybean transformation labs. This reduction is due to alternative employment or future retirement of key faculty. Accordingly, greater coordination and interaction among the existing soybean transformation laboratories could lead to greater efficiency and therefore, is encouraged.

With new information currently available, some of the key goals identified in 2003 are now considered to be of lower priority, as they are technologically too demanding or not crucial for research purposes. The target goals that fall within this category are viral-induced gene silencing (VIGS) systems, tissue-culture-free transformation, and site-specific integration. A redirection of transformation efforts is recommended so as to better support efforts in functional genomics. Studies should emphasize seed traits and embryo development. With the availability of a rapid soybean somatic embryo transformation system and the strong correlation between zygotic and somatic embryogenesis the soybean community has a model system that can be exploited in functional genomics programs to help elucidate the underlying biology of the developing soybean seed.

STP 2.1 Improve the Efficiency of Transformation for Functional Genomics
The development of novel approaches will be based on a better understanding of the factors that influence induction and regeneration of soybean tissue cultures. In addition, testing of new gene promoters, selectable markers, and gene coding terminators can lead to increases in transformation rates. The availability of tissue-specific gene promoters will increase the range of traits that can be improved by genetic engineering. A main limitation to high throughput functional genomics of soybean is space constraints due to the long growth period and large size of the plant. This limitation might be alleviated by the pending public release of the rapid-cycling soybean 'MiniMax'.

Goals for 2007:

Ability to produce 500 transgenic lines per year per person.
Continue the development and testing of new gene promoters, selectable markers, and terminators, in a systematic, coordinated fashion between soybean transformation laboratories
Evaluate MiniMax for transformability, in an effort to obtain a transformation system for a short-cycle soybean

Goals for 2009:

Have a series of tissue-specific and inducible promoters publicly available
Re-evaluate the feasibility of VIGS and site-specific gene integration

STP 2.2 Routine Access to Transformation Technology for the Soybean Community
Success in improving soybean transformation and the need for high throughput technologies has created the demand for the establishment and coordination of plant-growth and stock-center capacity to characterize, maintain, and distribute the developed stocks. Coordination and distribution of materials and skill between transformation laboratories will help accelerate the transfer of transformation technology to other laboratories. The development of a centralized repository for transgenic soybean events will help ensure identity preservation of the regulated materials and compliance with established Federal guidelines governing interstate movement and release of transgenic seed.

Goal for 2007:

Provide an infrastructure to facilitate coordination and cooperation among the existing soybean transformation laboratories

Develop funding options to secure resources for the establishment of a centralized repository to house and distribute transgenic seed stocks developed within the public sector.

STP 2.3 Technology to Deliver large DNA/Multiple gene constructs
To date, no protocols to transform soybean with whole BAC clones have been developed. The Forrest BAC are cloned into a transformable vector with a selectable marker. Although these BACs currently do not have selectable markers flanking the inserts they may provide a starting point to develop large-insert transformation technologies.

Goals for 2007:

Test both gene gun and Agrobacterium for ability to transform with BACs
Determine the most effective way to arrange multiple gene constructs for metabolic engineering

Goals for 2009:

Select a metabolic pathway with minimum of 5 genes and compare delivery systems

STP 2.4 Develop Transgenic Screens to Elucidate Gene Function
New technologies based on insertional mutagenesis using a range of transposon tagging strategies and targeted RNAi approaches are being developed. Continuing to improve the efficiency of extant systems will enhance these efforts. Enough technology is now in place to make greater use and development of somatic embryo system for testing seed-specific traits (oil and protein, etc.) and understanding seed biology and development.

Goals for 2007:

Quantify the transposition frequency and pattern of Ac/Ds and Tnt1 in soybean
Evaluate and confirm utility of RNAi methods in somatic embryo and whole plant approaches for targeted knockouts

Develop high throughput vector assembly system to easily assemble vectors for ectopic expression or down regulation
Support and infrastructure to maintain, characterize, and distribute seeds is essential

Goals for 2009:
Target 3200 Ds lines
Target 200,000 insertions
Map insertions: 200 genetic, 1000 physical

GENOME SEQUENCING and GENE DISCOVERY

The genome of the soybean is approximately 1 x 10^9 bp and is estimated to contain 50,000 to 100,000 genes. These genes are responsible for all the pathways and functions of growth and development. The identification of candidate genes is critical for robust application of marker assisted selection, comparative analyses between genomes, and the process of understanding their function. An association with phenotype is essential to understanding how plants have adapted to the environment and how they ultimately affect plant productivity and health.

STP 3.1 Discover Soybean Genes

Gene discovery is a primary research priority in the field of genomics. It is the foundation of all functional analyses and is the ultimate target of most structural and physical genetic analyses. More than 300,000 ESTs have been obtained from expressed soybean genes. Although this type of information has provided crucial information on gene identity and gene evolution it is often necessary to have the entire expressed gene sequence in order to take full advantage of genomic tools for marker development. In order to gain information on introns as well as flanking genomic DNAs (important for understanding of gene regulations, but also important for marker development) it is necessary to obtain corresponding genomic sequence for the expressed gene.

Many of the goals of soybean genomics come ultimately from knowledge of the genome sequence. In July 2001 the U.S. Legume Crops Genomics Workshop White Paper (http://www.legumes.org/) cites the sequencing of the gene-rich regions of soybean (estimated at ~340 Mb) as one of its top priorities. This objective has been repeated at numerous NSF-supported, USDA-supported, and commodity board-supported workshops since then. Undertaking this goal is critical to future soybean genetic advances. Whole-genome sequencing is now an achievable priority. Whole genome sequencing in soybean is critical to many of the 2007 - 2009 goals. This resource will be made more valuable by additional efforts to anchor this sequence to the physical and genetic maps.

Goals for 2006:
Sequence 2,000 full-length cDNAs and corresponding genomic sequences. Have in place, in Williams 82, an initial whole-genome sequencing project.

Goals for 2007:

Sequence 10,000 full-length cDNAs and corresponding genomic sequences. Have in place, in Williams 82, a whole-genome sequencing project.

The initial shotgun sequence of the entire genome should be available.

Develop a map of the soybean `interactome' of seed and soybean-specific traits.

Develop a proteome base-line of major soybean stages and environmental responses.

Establish a central repository for data storage, and cross experiment comparisons.

STP 3.2 Create Physical and Transcript Maps of Soybean

Genome sequencing is a quantum-leap technology much like Watson and Crick's discovery of the structure of DNA. Gene localization, which is ideally based on a fully sequenced genome, includes the creation of a physical map anchored with genetically mapped gene sequences. This is the starting point for localizing and cloning genes and sequencing the soybean genome. A complete physical map requires that a BAC library contains a minimum tile of clones for the genotype to be whole-genome sequenced (Williams 82). As of 2005 a 10X BAC coverage of Williams 82 was fingerprinted and assembled into contigs. BAC end sequences need to be generated from all of these BACs. The MTP resources and BES are already completed for the Forrest map. Through funding from the United Soybean Board and the National Science Foundation unanchored BACs and BAC contigs are beginning to be genetically mapped. Completion of the physical map with BAC-end sequences will help accomplish several goals such as SNP development, further genetic anchoring of the physical map, identification and targeted sequencing of gene rich regions, whole genome sequencing, and will help to reveal ancient duplications within the soybean genome.

An integrated soybean genome map will increase the efficiency of crop improvement through application in functional genomics, maker assisted breeding, and transformation. This map is also critical to advancing numerous genomic goals such as targeted sequencing, candidate gene identification, and comparative mapping. Completion of the Williams 82 physical map should be a community priority. The goal is to create a 95% complete physical map of the soybean genome encompassing a complete tile path from `Williams 82'; the same genotype for which a large EST resource exists. In order to assist in contig assembly and to create STS for each BAC, the ends of BACs used in the contig assembly will be sequenced. Before releasing a Williams 82 physical map, an evaluation of the efficiency of methods and the synergies for resolving duplicated, homoeologous regions will be completed. The utility of *Medicago truncatula* sequences for soybean map resolution will be determined. At the initial assembly of the Williams 82 physical map, it is recommended that the physical maps of Forrest and Williams 82 be integrated

to the extent possible. Toward this end the minimum tiling path BACs from the Forrest map have been fingerprinted in parallel with the Williams 82 BACs. An additional research area is the establishment of a transcript map anchored to the physical and genetic maps.

Goals for 2006:

Genetically anchor 80% of the contigs comprising the physical map of Williams 82.

Generate BAC-end sequences on sufficient Williams 82 BACs used in the construction of the physical map to constitute a 10 X coverage of the genome.This is already completed for the Forrest BACs.

Further the integration of the Forrest and Williams 82 physical maps.

Complete the placement of 1500 overgos onto the Williams 82 physical map.

Establish a consortium for anchoring BACs and BAC contigs using BES, SNPs and SLPs..

Goals for 2007:

Complete the genetic mapping of 90% of the contigs comprising the physical map of Williams 82. Further compare and integrate Williams 82 and Forrest physical maps.

Complete the placement of 3,000 ESTs on BACs in the physical map through a combination of bioinformatics, SNP mapping and overgo hybridizations.

Goals for 2009:

Complete the genetic mapping 95% of the contigs comprising the physical map of Williams 82.

Make use of the information for gene discovery, cultivar improvement, higher yield, seed composition improvement, etc

STP 3.3 Development of microarray technology

All traits of living organisms are the consequence of gene expression. Information contained in the genes is translated into products that direct life functions. An understanding of the mechanisms regulating the genes that control important crop traits is a prerequisite to manipulating them to advantage.

Most important traits are specified by members of small gene families. Often closely related members of these gene families are differentially expressed at different development times and places. For this reason 'paralogue-specific' technologies must be developed and applied. In addition, most traits are the result of complex interactions among numerous genes. For this reason, universal gene-expression technologies must be developed and applied.

The purpose of assigning function is to discover the genes of agronomic importance. The assignment of function to genes and the development of `paralogue-specific' microarrays proceed at several levels. First it is necessary to have a nearly full-length cDNA sequence that includes sequence at the 3' end of the gene. There are approximately 27,000 3' sequences derived from `unigenes' in soybean. A 2005 goal was to obtain 3' sequence from an additional 30,000 unigene cDNAs identified from the Public Soybean EST collection. Funding for this objective was not sought and consequently this goal was not achieved. However, many other objectives were achieved.

The research community now has available a wide range of resources for analysis of gene expression. Two DNA chips, each with approximately 18,432 genes are now created and are available to the research community on a cost recovery basis. With funding from the United Soybean Board long-oligo arrays are being generated. One array containing approximately 19,000 oligos is complete and another array of similar size is in progress. A soybean/Phytophthora/SCN Affymetrix GeneChip is also now available. These resources will be useful to determine the expression patterns of genes in tissues and organ systems of the plant by measuring the expression of thousands of genes at a time (i.e., "global" expression patterns). Expression comparisons under conditions including pathogen challenge, symbiont infection, heat, cold, flooding and drought stresses, and nutrient limitations will yield classes of genes involved in these critical processes. Expression profiles of many agronomically important genotypes containing traits of economic importance and QTL may also aid in assigning function. Expression profiling will yield the information needed to select promoters useful for plant transformation.

Goals for 2007:

Characterize plant gene expression patterns in soybean in response to abiotic and biotic signals.

Generate 3' sequences of an additional 30,000 soybean unigenes.

Develop resources for transcription factors expressed at low levels

Develop a community working group on related metadata and ontology, and establish interactions with other legume groups

Generate a minimun of 10,000 full length cDNA sequences.

Ensure the availability of a microarray database

Goals for 2009:

Generate another 10,000 full length cDNAs

Move toward system biology platform by 2009

PROTEOMICS and GENE FUNCTION

Genes encode proteins, and proteins carry out enzymatic functions. Important phenotypes in soybean (yield, oil, and protein content in seeds) are determined by gene function. Therefore to improve agronomic traits, the function of genes must be manipulated. Before this can be achieved, the function of each gene in the genome must be identified. Although DNA microarrays measure mRNA expression at the genomic level, results from this method do not always reflect the amount of protein that is derived from expression of a gene. Because proteins frequently specify the phenotype, determining the amount of specific proteins is important. Classically, gene function has been addressed by detailed biochemistry on single gene products (enzymes). However, the information required for genome-wide analysis makes this approach impractical. Therefore, a genome wide approach is required to determine gene function.

STP 4.1 Proteomic Technology to Determine Gene Function

Proteomics is a technology that relies on quantitative mass spectrometry to identify gene products and is based on matching the masses of tryptic digest fragments to a database of known proteins. More recently, the term proteomics has been applied to any approach that measures protein function at a genomic level. For instance, researchers can now apply methods to identify protein-protein interactions in a cell. Many proteins act in multi-protein complexes. Understanding these associations will help to better define protein function. It was previously recommended that a detailed proteomic analysis of the regulation of protein and oil synthesis be initiated in developing seed by 2005. This goal has been met. Because of the importance of these constituents to the value of soybean as a commodity a proteomic map of developing seed remains a community priority. Identification of metabolomic intermediates will be necessary to have a better representation of primary and secondary metabolisms. This will, with metabolite profiling help us know which functions are affected by mutations.

Goals for 2007:

Initiate metabolomics technology.
Establish metabolomic standards that will be useful in many systems.
Identify 2,000 to 4,000 metabolomic intermediates that are useful in many systems
Have in place a proteome map of developing soybean seed.

Goals for 2009:

Develop an `interactome' to better understand protein-protein interactions.
Integrate transcriptome, proteome and metabolome information

STP 4.2 Application of Transformation Technology to Determine Gene Function

Geneticists have typically addressed gene function through mutation, and have deduced gene function based on an observation of the mutant phenotype. With the advent of efficient soybean transformation, this classical method can be applied at the genomic

level by transposon-induced mutations. Two systems, Ac/Ds (from maize) and the retro-transposon Tnt1 (from tobacco), are being developed. These systems should enable broad-range deletion of genes (gene-knockouts) using transposon tagging in soybean to help determine gene function. As of 2005 600 Ds lines are in the process of being evaluated to test for Ds movement.

Goals for 2007:

Generate 100,000 independent Tnt1 insertions in soybean.
Test alternatives if Tnt1 is unsuccessful
Develop 1600 Ds lines
Demonstrate that a gene has been successfully tagged.
Map 100 insertion sites genetically.
Map 500 insertion sites to the physical map.

Goals for 2009

Identify a central facility to archive, store, curate, and distribute lines.

STP 4.3 Reverse Genetics to Determine Gene Function: TILLING

TILLING for Targeting Induced Local Lesions IN Genomes has been developed and is considered a new reverse genetic tool for screening chemically induced mutations in target sequences to determine gene function and identify beneficial alleles.
It 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. Given a sufficiently large, highly mutated soybean population, point mutations in any gene can be identified. Because of the long-term importance in the functional assignment of genes, it was previously recommended that TILLING populations and libraries should be developed as a public genetic resource. Currently, tilling populations are available in Williams 82 (3,400 M2 lines) and Forrest (3,000 M2 lines). Gene `knock-outs' have been identified. Soon, a TILLING facility should be established to coordinate use of this technology for the determination of gene function and to supply germplasm with specific mutations to breeding programs.

Goals for 2007:

Establish a central TILLING facility
Archive, curate, and distribute lines and for long-term storage of existing populations.
Increase tilling populations as needed
Taking advantage the increasing amount of genomic sequence being generated, evaluate `Ecotilling' as a way to identify benefical alleles for breeders

Goal for 2009

Create a 50% self-supporting tilling facility to provide seed or mutants.

BIOINFORMATICS

Genomics projects are currently underway for several model legumes as well as for soybean and other crop legumes. These projects are resulting in the collection, storage, and analysis of many data points (*i.e.*, sequences, expression levels, map positions). Collecting, storing, manipulating, analyzing and retrieving this vast amount of information require radically different techniques and technologies than previously used in biological studies. Further, this disparate collection of data needs to be interlinked based on a logical mapping of biological data types to one another. Researchers must be able to traverse the data from QTL to their relative locations on physical maps and, ultimately to sequence maps containing corresponding genes. Genes must be related to gene products that can be associated with biochemical pathways, allowing researchers to discover the molecular basis for phenotypic traits. Informatics components can be separated into the development of infrastructure and tools and the application of those tools to synthesize information into useable results. Infrastructure needs include the development of relational database management systems, visualization tools, algorithm development, distributed computing, storage systems, and networking. Information integration is a biological problem, which includes pathway reconstructions, understanding of developmental processes, and inferring likely phenotypic information.

The Legume Information System (LIS) was conceived to be a comparative legume resource, populated initially with data from *G. max, M. truncatula* and *Lotus japonicus* and followed by data from other species. A major bioinformatics goal was to develop a robust means of comparative transcript analysis, initially between the *G. max, M. truncatula* and *L japonicus*, and eventually including unigenes from *Arabidopsis thaliana* as a non-legume species. This has been achieved through the LIS virtual plant interface and comparative mapping tools developed or in the late stages of beta testing. This resource and the data are the first steps towards leveraging model plants to gain insights into crop species.

The second step in comparative analysis planned for LIS involve decorating genomic sequence data with the shared consensus generated as above. Currently, the genomic component of LIS uses consensus sequences generated by the transcript component of LIS for each species. Mapped gene sequences help identify gene-rich regions, help validate or refute gene models and provide data to help build scaffolding to bridge the genomic-physical map-linkage map gulf. Genetic maps for several legumes are now in LIS and couple with CMap software these maps are able to be compared, side by side. Physical maps are in the process of being integrated into LIS and SoyBase. Structural information about genomic regions may also shed light on gene families and certainly helps to address evolutionary questions concerning species relatedness. Analyzing gene structure in a genomic context is a powerful comparative genomic tool enabling identification of regions of micro- and macro- synteny.

STP 5.1 Database Development and Data Migration

Starting in 2003, map data (linkage and physical) and associated metadata (authors, affiliations, literature etc.) was ported from SoyBase into the relational CMAP database and visualization software developed by Ken Clark at Cold Spring Harbor. CMAP was modified to interoperate seamlessly with the LIS and will feature automated linkage of sequence-based markers to EST and genomic data housed in LIS. Beginning in 2004, pathology, transformation data, and other remaining data classes were moved to LIS. An original goal was to move SoyBase data completely over to LIS. It has now become apparent that this is not the best approach. Now, the recommendation is to move most sequence-centric data to LIS but to retool SoyBase into a relational database suitable for a breeders `toolbox'.

The usefulness of genomic databases is partially the result of the middle-ware and the underlying engine. The ability of the user to comprehend the databases capabilities and to maneuver through the various levels of data is also critical. To facilitate the ease by which data can be viewed, manipulated, and retrieved from LIS, major improvements will continue to be made to the existing LIS and SoyBase user interface, including development of a novel, graphic-based query interface which will facilitate data browsing and exploration.

Goals for 2007:

Incorporate new data types and functionality as determined by user panels and the Steering Committee.
Continue the migration of relevant data into LIS
Complete the development of the Soybean Breeders' Toolbox with a new user interface

STP 5.2 Integration of Soybean Data with Other Databases and Development of Annotation and Nomenclature Standards

In order to take advantage of the growing amount of the advances in knowledge coming from the numerous plant genome project, it will be necessary to interconnect with the pertinent databases as much as possible. In order to avoid redundancy in tool design and data modeling it will be important to communicate with a much broader research community than in the past. The lack of community standards relative to gene expression data was identified as a critical limitation.

A Steering Committee of legume researchers and bioinformaticists to guide the development of LIS will be convened. It will be critical to incorporate ideas and suggestions from the legume community. To accomplish this LIS will solicit input on the perceived needs of the legume research community, which will directly influence the system’s design and user interface development. This will be accomplished by periodically convening panels of users to participate in workshops and by providing a forum for online comments.

Goals for 2006:

Convene a panel, including experts from outside the soybean community to develop an informatics master plan and community data standards

Goals for 2007

Convene a panel of informaticists and scientists to address annotation and nomenclature standards as identified in 2006 (above)

Establish a permanent steering committee to make administrative decisions, e.g. data migration (LIS or SBT)