

U.S. Legume Crops Genomics Workshop

30-31 July 2001

Hunt Valley, Maryland

White Paper



Convened by:

H. Roger Boerma, Distinguished Research Professor, University of Georgia, 111 Riverbend Road, Athens, GA 30602 (phone, 706-542-0927; email, rboerma@uga.edu)

Judy St. John, Associate Deputy Administrator, Agricultural Research Service, Crop Production, Product Value and Safety, Room 4-2204, George Washington Carver Building, 5601 Sunnyside Avenue, Beltsville, MD, 20705 (phone, 301-504-6252; email, jsj@ars.usda.gov)

Jennifer Yezak Molen, consultant, AgSource, Inc., 600 Pennsylvania Avenue, SE, Suite 320, Washington, DC 20003 (phone, 202-969-8902; email, jyezak.molen@gordley.com)

I. EXECUTIVE SUMMARY

On 30-31 July 2001 twenty-six legume scientists with knowledge of structural and functional genomics, DNA markers, transformation, bioinformatics, and legume crop improvement participated in a workshop hosted by the United Soybean Board, the National Peanut Foundation, the USA Dry Pea and Lentil Council, and the USDA-ARS. Over the course of two days, the scientists reached consensus on priority

genomics research on U.S. legume crops. The following high-priority areas of critical research were identified:

Genome Sequencing of Strategic Legume Species

- Sequence gene-rich regions of soybean, common bean, and peanut
- Whole-genome sequencing of barrel medic (*Medicago truncatula*)

Physical Map Development and Refinement

- Development of physical maps in peanut and common bean
- Refinement and/or completion of physical maps in barrel medic and soybean
- Development of transcript maps of known genes (ESTs)
- Integration of maps among taxonomically key species

Functional Analysis: Transcriptional and Genetic

- High through-put stable and transient gene transfer systems
- Assignment of gene function
- Expression analysis
- Proteomics, metabolomics, and metabolic reconstruction
- Gene knockout systems

Development of DNA Markers for Comparative Mapping and Breeding

- Universal set of PCR—based legume Sequence Tag Sites
- Establish local and global similarities of different gene arrangements
- Quantitative trait loci (QTL) discovery
- Determine the levels of genome conservation among the legume crops

Characterization and Utilization of Legume Biodiversity

- Broaden and refine legume phylogeny
- Establish levels of biodiversity within the different legume crops
- Compare closely related species for gene discovery
- Crop-microbe co-evolution
- Multi-gene family evolution
- Domestication of legume crops
- Preserving and utilizing germplasm

Development of a Legume Data Resource

- Development of a legume-wide database (molecular, genetic, expression, diversity, and breeding) publicly available through the web
- Leverage data from each legume species for greater efficiency

II. ROLE OF LEGUMES IN AGRICULTURE

Legumes, together with cereals, have been fundamental to the development of modern agriculture. Since the dawn of civilization, many legume species have been instrumental in supplying human food (e.g., soybean, common bean, pea, peanut, lentil, chickpea), edible oils (peanut, soybean), and animal fodder and forage (alfalfa, clover). Legumes are second only to grasses in importance for human and animal dietary needs. Worldwide, legumes are grown on about 15% of the arable land (270-300 million hectares). They provide 33% of humankind's nutritional nitrogen requirements. Last year in the U.S., soybean and alfalfa production was 72 and 84 million metric tons, respectively. Their total estimated direct value was \$20 billion (soybean = \$13 billion, alfalfa = \$7 billion). An equally important added value for U.S. agriculture is that legumes in symbiosis with soil bacteria fixed some 17 million metric tons of atmospheric nitrogen worth about \$8 billion. This unique ability of legumes reduces the dependence of farmers on expensive chemical fertilizer, reduces our dependence on petroleum products, and improves soil and water quality. Because crop legumes can fix nitrogen, they are critical for subsistence farms in developing countries that do not normally have access to nitrogen fertilizer. In these areas, legumes frequently provide 66% of the nutritional needs for humans and are especially important as a substitute for animal protein.

One of the driving forces behind sustainable agriculture and protection of the environment is effective management of nitrogen in farming systems. Intensive farming practiced in developed countries is predicated on using large amounts of nitrogen fertilizer. This practice has led to significant deterioration of water, soil, and air quality. In addition, it is estimated that production of 1 metric ton of ammonia requires a consumption of approximately 1,185 cu. meters of natural gas. As the world's population approaches 10 billion within the next half century, nitrogen needs for increased crop production will exacerbate current environmental problems. Increased cultivation of legumes will be required to ameliorate environmental degradation, reduce the depletion of nonrenewable resources, and provide adequate nitrogen for the population. Cultivation of legume crops results in a significant reduction in the use of nitrogen fertilizer. It has been estimated that growing a nitrogen-fixing legume in rotation with some crops decreases the required application of nitrogen fertilizer by 40%. Moreover, nitrogen volatility into the atmosphere and nitrogen leaching into groundwater are reduced by cropping legumes. In addition, the nitrogen fixed by legumes is equivalent to sequestering a further 800 million metric tons of CO₂. Estimates indicate that in the U.S., simple rotation of a legume with corn could replace 12 to 15% of the nitrogen fertilizer needs by corn resulting in an on-farm savings in excess of \$500 million. Legumes clearly play a major role in protecting human health, farm profitability, and mitigating environmental problems.

III. SCIENTIFIC STATUS OF LEGUME GENOMICS

Since Mendel, crop legumes have been the focus of intensive genetic studies to improve yield, quality, resistance to biotic and abiotic stress, and extend geographic range. As a result, selected legume species have well-studied genetic systems characterized by classical biochemical and physical markers, cytogenetic analysis, chemically induced mutations, and DNA marker-based genome linkage maps. Yet in many of these well-defined systems, comprehensive genetic analysis is limited due to the large size of the genomes of legume crops. Furthermore, few of the basic tools required for modern genome analysis including polymerase chain reaction (PCR) based DNA markers, expressed sequence tag (EST) databases, or bacterial artificial chromosome (BAC) libraries have been applied to most legume species. No crop legume has an integrated genetic, physical, and transcript map. Furthermore, efficient transformation has recently been developed in only two legume species. It is of paramount importance that a concerted research initiative be directed towards the development of tools that will permit application of modern genome analysis and manipulation technology to the genetic improvement of crop legumes.

The genomes of most crop legumes are large and relatively complex. For example, soybean is an ancient polyploid and alfalfa an autotetraploid with genome sizes of ~1200 megabases (Mb) and ~1600 Mb, respectively (for comparison, the human genome is about 3000 Mb). The genome size of cultivated peanut is 2800 Mb and pea is 4000 Mb. Such large sizes significantly complicate the development of ordered physical maps of the genome, as well as the identification and location of important genes. The large genome sizes make complete sequencing financially tenuous. Syntenic relationships within botanical families make it possible to use plant species with much smaller genomes to facilitate understanding of those with large genomes. For example, the recent complete sequencing of the smaller genomes of *Arabidopsis thaliana* (128 Mb) and rice (425 Mb) provide the platform for genome analysis of more complex species such as canola, broccoli, corn, and wheat. Information from the *Arabidopsis* and rice genomes is rapidly being translated across more complex species to enhance disease and pest resistance, yield, and compositional quality of the seed.

In order to expedite and simplify genome analysis of crop legumes, it has been proposed that parallel analysis of a legume with a smaller genome be considered. Recently, studies sponsored by the NSF Plant Genome Program have shown that the barrel medic, *Medicago truncatula*, would be an ideal candidate for parallel analysis with crop legumes. Barrel medic is a diploid, has a small genome (~450 Mb), rapid generation time, is self-compatible, and appears to have synteny with alfalfa and also to some degree with pea and soybean. Comparative analysis with other legume crops will provide additional advantages through complementation of genetic knowledge available in the different legume species. Soybean is a major crop, with significant prior study of genetics and crop and seed physiology. Common bean (*Phaseolus*) benefits from relatively well-developed genetic studies and ample polymorphism within the cultigen. Peanut possesses an unique reproductive physiology, which can contribute to a greater understanding of crop reproductive biology.

IV. PRIORITY GENOMICS RESEARCH ON U.S. LEGUME CROPS

The 26 scientists (referred to as Legume Crop Working Group or LCWG) reached consensus on the following six areas of critical genomics research. The order of listing is not intended to indicate order of importance or relative priority.

A. Genome Sequencing of Strategic Legume Species

Over the past decade, biological research has been transformed by the ability to sequence entire genomes of living organisms. The application of this technology to human medicine is revolutionizing the pharmaceutical industry. A similar but largely untapped potential exists in the agricultural sciences. In the case of legumes, many important traits have been identified at the genetic level by breeders and geneticists. Whole genome sequencing will reveal the molecular blue print that underlies these valuable characters.

The LCWG recommends that genome sequencing projects be initiated on a few strategically chosen legumes. The species recommended by the working group are intended to span the phylogenetic diversity of these crop species, and to provide a platform for the methodical exploration of legume genomics. We recommend (1) sequencing the gene-rich regions of three warm season grain legumes, *Phaseolus vulgaris* (common bean), *Glycine max* (soybean), and the phylogenetically more distant *Arachis hypogaea* (peanut) and (2) determining the complete genome sequence of *Medicago truncatula* (barrel medic) as a specific reference for the cool season legume species (e.g. pea, lentil, and alfalfa) and as a structural model for other crop legume genomes.

The impact of this research will be to integrate genetic and functional information across legume crops. The resulting knowledge would enable the more precise development of improved crop cultivars by classical and molecular breeding methods, and it would greatly accelerate the pace of research to determine the molecular basis of traits of biological and economic importance.

B. Physical Map Development and Refinement

Physical maps for carefully selected representatives of the legume family such as soybean (*Glycine* spp.), medic (*Medicago* spp.), peanut (*Arachis* spp.), and common bean (*Phaseolus* spp.) are important primary research tools, and also key stepping-stones toward longer-term goals. A necessary prerequisite for taking advantage of physical maps is the development of BAC libraries in a broad range of legume species. Physical mapping is an economical means to plot the order of most genes in a genome, and to reveal gene-rich regions. A detailed physical map is central to identifying specific candidate genes that may account for an agriculturally important trait and development of breeder-friendly markers for effective selection of such traits. Physical maps advance the integration of maps for different crops, leveraging investments in genomic research. In the short term the physical map provides a framework to organize partial sequence information. In the long term, rigorous physical maps comprised of several inter-related data types provide the robust framework needed to assemble a complete genomic sequence.

Transcript maps can be derived using physical maps that define the position of the known legume genes (ESTs) and their paralogs. Transcript maps will identify the gene-rich regions of different legume crops and allow their comparison. These transcript maps will result in the integration of genetic and functional information across legume crops and provide significant cost and efficiency benefits for the sequencing efforts.

C. Functional Analysis: Transcriptional and Genetic

Current and planned structural genomic efforts will create a wealth of legume gene sequences. Efforts must be made to move beyond this 'structural' information to examine gene function. The LCWG recommends research on the application of functional genomic tools (e.g., DNA microarrays, proteomics, metabolomics) to the major legume crop species. These tools have tremendous potential to aid understanding of quantitative and value-added traits and biotic and abiotic stress resistances. These studies should utilize the biodiversity available within each legume species.

The majority of plant species, including legume species, lack efficient methods for gene transfer which is critical for gene functional analysis. Although several promising efforts are underway to overcome this limitation in legumes, it is not clear at present which of these approaches is most efficient. Therefore, a series of studies to evaluate these methods and ultimately choose one or more for large scale, concerted efforts should be initiated. In addition for those species with a working transformation system, the LCWG recommends the commitment of funds for the development of methods for the evaluation of gene function such as gene knockouts or gene down-regulation systems.

D. Development of DNA Markers for Comparative Mapping and Breeding

DNA markers are some of the most powerful tools to come out of the genomics revolution. DNA markers form the foundation of genetic linkage mapping. They are the basis of marker-assisted breeding, enable genome comparisons between different species, and provide tools for assessing

molecular variation within and between species. The LCWG recognizes the value of DNA markers to all crop legume species and the vital need to leverage knowledge about DNA markers from better-characterized crops to others that are less well studied. In certain instances when knowledge of better-characterized species does not provide adequate genome coverage of a specific legume crop, the development of species-specific markers will be required.

To take advantage of DNA marker technology, a core set of at least 1000 sequence tagged sites or STSs that are universal among all legume species should be developed. For legume crops with large chromosome numbers and low levels of DNA polymorphism, such as soybean and peanut, more than 1000 STSs will be needed to provide meaningful comparative maps. These STS markers will begin with strategically chosen PCR-based markers that have already been developed, especially in pea, soybean, and barrel medic. Eventually, the STS core set will grow by mining legume sequence data in order to find highly conserved sequences shared by all legumes.

The legume STS core set will immediately provide powerful tools for trait mapping and marker-assisted breeding in all legume species, including those with few marker resources available today. The STS core will also interconnect the genetic maps of different species, revealing cases where genome organization is highly conserved and where rearrangements have occurred. The STS core will also simplify the important task of fingerprinting germplasm collections and analyzing molecular evolution within the legume family.

E. Characterization and Utilization of Legume Biodiversity

Biodiversity is the raw material for the genetic improvement of all crops. If we are to use the genetic diversity naturally present in germplasm it is important to determine where the greatest diversity is and how it can be applied in crop breeding. A number of crop species (pea, common bean, cowpea) have been identified as genetically diverse, whereas others (chickpea, soybean, peanut) possess a narrow genetic base. Sequencing of alleles in one or more of the genetically diverse crops may provide important diversity for crop improvement. Additional studies on phylogenetic relationships among legumes using a variety of genome markers should be pursued to refine and improve our understanding of legume phylogeny.

A comparison of the DNA sequences or their expression among closely related species or genotypes within a species will facilitate the identification of genes for important traits. This will require the identification of carefully selected genotypes and the application of genomic tools such as DNA sequencing and microchips for expression analysis. Opportunities for this research exist in many legume crops, but particularly in *Phaseolus* beans, for which a detailed phylogeny exists.

In pathogens, mutations in genes occur that allow them to cause disease in their crop host. Conversely, new variations in genes are generated in host plants that allow them to recognize and initiate defense responses to resist the invasion by pathogens. A joint analysis of host and pathogen diversity with genomic tools will determine how genes for resistance are generated, why some genes are more stable than others, or why some host plants have a broader spectrum of resistance. A similar research approach is recommended by the LCWG for legume crops and their symbionts.

Most agronomic traits are conditioned by families of related genes. These multigene families originated by duplication of an original gene and sequence divergence of the copies of the gene. During the divergence, the different copies can acquire different traits, such as adaptation to different environmental stresses, production of new compounds, and expression in different

organs. Structural genomic and expression studies will tell us how legumes differ among themselves and how they generate new traits.

The USDA/ARS collects and maintains germplasm collections for major crop legume species. These collections likely contain important traits currently undiscovered or undefined. The LCWG recommends intensive evaluation of existing germplasm collections using the tools of structural and functional genomics to identify new genes conditioning economically important traits and to characterize the U.S. germplasm collections more completely.

F. Development of a Legume Data Resource

Large-scale genomic efforts can be leveraged across species. Genomic data from the various legume species are currently maintained in a variety of species-centric databases. Tools must be developed to integrate, analyze, and deliver the data from many species.

Examples of the data types to be maintained in such an integrated resource include, but are not limited to, EST and genomic sequences, expression profiling information, map-based genetic traits, and breeding and biological diversity data. The integration of ongoing map/trait-based activities with sequence-centric databases will provide the research community with scalable, sustainable data handling and analytical abilities.

Because it is likely that data collection and curation will occur in 'the field' and not at the site of the data integration, it is important that the centralized database be able to access the species-centric databases, acquire the required data interactively, and generate reports in a user-friendly, web-accessible, and graphical manner. Such an integrated resource will benefit all legume species by making genetic and genomic information developed in a wide range of species available to all other species researchers via a web-based interface. An integrated database provides a platform for integrated data analyses. Such a merged database would facilitate pan-legume data mining. This would provide a synergistic utilization of the shared data.

V. WORKSHOP PERSPECTIVE AND NEXT STEP

The six areas of priority research that are described in this report represent common needs for the major legume crops grown in the U.S.. Together, they will advance the status of legume research in a synergistic manner, complementing and building upon the specific strengths of each legume crop. The results of this collaborative effort will be a greater understanding of the genomes of each crop species and the development of the crop-specific genomic tools and technology needed to accelerate the rate of genetic gain for U.S. legume crops. Before departing the workshop, the scientists developed a "Plan of Action" to assure the results of the workshop were widely distributed and that an organization was created to enhance communications among legume researchers and promote opportunities to fund the consensus research (see attachment).

After the 26 scientists had developed their areas of consensus research, grower leadership representing the various legume crops and commodity association staff (representing soybean, peanut, pea, lentil, and common bean) reviewed the results of the workshop and discussed approaches to cooperate in seeking funds to accomplish the research. The grower leaders and association staff agreed their respective legume crops are in a stronger position to achieve support for a plan if it is presented and supported by each commodity organization in its entirety.

VI. PARTICIPANTS

NAME	ADDRESS	STATE	PHONE	FAX	
Albert G. Abbot	Department of Genetics And Biochemistry 122 Long Hall Clemson University	Clemson, SC 29634	864-656-3060	864-656-6879	aalt
William D. Beavis	National Center for Genome Resources 2935 Rodeo Park Drive East	Santa Fe, NM 87505	800-450-4854	505-995-4432	wdt
Charles Brummer	1204 Agronomy Iowa State University	Ames, IA 50011	515-294-1415	515-294-6505	bruu
Mark Burow	Texas Agricultural Experiment Stn. Route 3, Box 219	Lubbock, TX 79401	806-746-6101	806-746-6528	mbu
Tom Clemente	E324 Beschle Center University of Nebraska — Lincoln	Lincoln, NE 68588-0665	402-472-1428	402-472-3139	tcle
Douglas R. Cook	University of California One Shields Avenue 206 Robbins	Davis, CA 95616	530-754-6561	530-754-6617	drcu
Perry Cregan	USDA-ARS-BARC-West Soybean Genomics and Improvement Lab B006, Room 100	Beltsville, MD 20705	301-504-5070	301-504-5728	creg
Leland Ellis	USDA/ARS	Beltsville, MD 20705	301-504-4788	301-504-4725	lece
Paul Gepts	Dept. of Agronomy and Range Science University of California	Davis, CA 95616	530-752-7323	530-752-4361	plge
David Grant	USDA-ARS G304 Agronomy Hall Iowa State University	Ames, IA 50011	515-294-1205	515-294-2299	dgru
David A. Lightfoot	Dept. of Plant Soil and General Agriculture Southern Illinois University	Carbondale, IL 62901-4415	618-453-1797	618-453-7457	ga4
Greg May	The Noble Foundation Plant Biology Division P.O. Box 2180	Ardmore, OK 73402	580-221-7391		gdn
Phillip Miklas	USDA-ARS 24106 N. Bunn Road	Prosser, WA 99350-9687	509-786-9258	509-786-9277	pmi
Henry T. Nguyen	Dept. of Plant and Soil Science Texas Tech University	Lubbock, TX 79409-2122	806-742-1622	806-742-2888	hen
Wayne Parrott	3111 Plant Sciences Bldg. Dept. of Crop & Soil Sciences University of Georgia	Athens, GA 30602	706-542-0928	706-542-0914	wps
Andrew Paterson	111 Riverbend Rd. University of Georgia	Athens, GA 30602	706-583-0162	706-583-0160	pat
Robert S. Reiter	Monsanto 3302 SE Convenience Blvd.	Ankeny, IA 50021-9424	515-963-4211	515-963-4242	robr
Ernest F. Retzel	420 Delaware St. SE MMC 43 650 Children's Rehabilitation Center University of Minnesota	Minneapolis, MN 55455	612-626-0495	612-626-6069	ernr
Deborah Samac	1991 Upper Buford Circle Room 495	St. Paul, MN 55108	612-625-1243	651-649-5058	deb
Lynn Senior	Syngenta Biotechnology, Inc. 3054 Cornwallis Road	Raleigh Triangle Park, NC 27709	919-597-3041	919-541-8585	lynr
Randy C. Shoemaker	G401 Agronomy Hall Iowa State University	Ames, IA 50011	515-294-6233	515-294-2299	rcsu

Gary Stacey	M409 Walters Life Science Bldg. University of Tennessee	Knoxville, TN 37996-0845	865-974-4041	865-974-4007	gst
H. Thomas Stalker	Box 7620 Department of Crop Science NC State University	Raleigh, NC 27695	919-515-2647	919-515-7959	hts
Lila Vodkin	384 ERML 1201 W. Gregory University of Illinois	Urbana, IL	217-244-6147		lvo
Norm Weeden	Montana State University Dept. of Plant Sciences and Plant Pathology ABS 303	Bozeman, MT 59717	406-994-7622	406-994-7600	nwe
Nevin Dale Young	495 Borlaug Hall University of Minnesota	St. Paul, MN 55108	612-625-2225	612-625-9728	nev

The conveners wish to thank Barbara Upston, Management Consulting Associates, for her highly effective workshop facilitation and Barbara Zapp, USDA-ARS, for her tireless and highly efficient technical support. The conveners were assisted in the management and organization of the workshop by four species coordinators: Charlie Brummer (alfalfa and clovers), Randy Shoemaker (soybean), Tom Stalker (peanut), and Norm Weeden (common bean, pea, dry bean, lentil).

Financial support for the workshop was provided by the United Soybean Board, Dry Pea and Lentil Council, National Peanut Foundation, and USDA-ARS.

ATTACHMENT — Scientists' Proposed Plan of Action

- **Support whole proposal — "Speak with One Voice"**
- **Steering Committee — Coordination and Communication**
- **Model after maize initiative**
- **Insure diversity and representation**
- **Develop working groups for specific proposals, items, or needs**
- **Establish website "National Legume Genome Initiative"**
- **Bring together all available public data — web links initially**
- **Develop a web-based newsletter — include abstracts for major legume projects**
- **Attach a letter of support to legume genomics proposals (when submitted for review) signed by steering committee or larger group**
- **Send meeting report to a major science journal**
- **Meet again:**
 - **Plant-Animal Genome Meeting — Jan. 2002**

- **International Legume Genetics & Genomics Meeting — June 2002**
- **Present six consensus National Research Needs at International Legume Meeting**