Cooperative State
Research, Education and Extension Service

USDA Cooperative State Research, Education and Extension Service
National Research Initiative Plant Genome Program
Plant and Animal Genome Conference XV
Town and Country Hotel, San Diego

Friday, January 12, 2007
1:00 pm – 5:30 pm
Windsor Rose Room
(Located on the 9th floor of the Regency Tower)

Fabaceae Project Directors Meeting

1:00pm: Meeting room open. Refreshments available. Load PowerPoint presentations.

1:30pm: Overview Presentations (10 minutes each)

- **CSREES NRI Plant Genome – Ed Kaleikau**
- **CSREES NRI Plant Biology – Gail McLean/Liang-Shiou Lin**
- **Legume Crops Genomic Initiative – Charles Brummer**

2:00pm: Project Presentations (10 minutes each)

- **Comparative Genomics, Nucleotide Diversity, and Karyotypic Evolution in *Arachis* (Peanut) – Steve Knapp**
- **TILLING: A Community Oriented Reverse Genetics Tool in Soybean – Khalid Meksem**
- **Genomics for *Phaseolus* as a Community Resource for Legume Researchers – Scott Jackson**
- **Tools and Applications of Gene-by-Gene Sequencing in Common Bean – Phil McClean**
- **Evolutionary Genomics of Small Multigene Families of Agronomic Interest in Common Bean (*Phaseolus vulgaris*) – Paul Gepts**
- **Disease Resistance Orthologs and Their Localization in the Common Bean – Khwaja Hossain**

3:00pm: Break

3:15pm: Continue Project Presentations (10 minutes each)

- **Full-Length cDNA Resources for Legume Genomics – Chris Town**

- **Development and Use of Novel Tools for Functional Genomic Analysis of Seed Storage Metabolism in the Model Genome *Medicago truncatula* – Michael Udvardi**

- **Assessing RNAi as a Reverse Genetic Tool for Global Analysis of NBS-LRR Gene Function in *Medicago truncatula* – Hongyan Zhu**

- **Elucidating the Small RNA Component of the *Medicago truncatula* Transcriptome – Janine Sherrier**

- **Using Association Mapping to Identify Markers for Cell Wall Constituents and Biomass Yield in Alfalfa – Charlie Brummer**

- **Genetic and Statistical Aspects of Association Mapping – Jianming Yu**

- **The Consensus Legume Database II: Biological Inference from Distributed Interoperable Databases – Ernie Retzel**

4:25pm: Update Presentations (15 minutes each)

- ***Medicago* Sequencing – Nevin Young**

- **Soybean Sequencing – Scott Jackson**

- **SoyBase and the Soybean Breeder's Toolbox – Randy Shoemaker**

- **Legume Information System – Greg May**

5:25pm: Closing Remarks

5:30pm: Adjourn

TABLE OF CONTENTS

USDA-CSREES NRI PLANT GENOMICS, GENETICS AND BREEDING

Ed Kaleikau
USDA Cooperative State Research, Education and Extension Service (CSREES)
Competitive Programs – National Research Initiative (NRI)
Washington DC
E-mail: ekaleikau@csrees.usda.gov
Telephone: 202-401-1931
Website: http://www.csrees.usda.gov/; http://www.csrees.usda.gov/funding/nri/nri.html

The CSREES NRI competitive grants plant genome program supports research, education and extension projects ranging from technology development to fundamental science and practical application for crop or forestry improvement in the U.S. Its priorities focus on technological advances and discoveries in areas such as a) analytical methods for mapping genes for complex traits for direct use by plant breeders, b) novel methods for analysis of the genome and its effect on biological function, c) cost-effective sequencing strategies to understand complex genome structure and organization, d) procedures to analyze the total expression patterns of genes under specific conditions, and e) appropriate data handling and analysis capabilities. The ultimate goal of the program is to contribute knowledge about the biology of agriculturally important plant processes and traits, which can be used to develop crops with enhanced economic value and expanded utilities.

To meet these identified needs of agriculture, the long-term (10 year) goals for this program are: increased fundamental knowledge of the structure, function and organization of plant genomes to improve agricultural efficiency and sustainability; effective integration of modern molecular breeding technologies and traditional breeding practice for U.S. crop and forestry improvement; and improved U.S. varieties for agricultural growers and producers.

The NRI plant genome program coordinates its activities with participating agencies (NSF, DOE, NIH, etc) of the Interagency Working Group on Plant Genomes. The NRI program focuses on genomics, genetics and breeding including: Tools, Genetic Resources, Bioinformatics, Functional Genomics, Genome Structure and Organization and Applied Plant Genomics.

The National Research Initiative Competitive Grants Program (NRI) is the office in the Cooperative State Research, Education and Extension Service (CSREES) of the USDA charged with funding research on key problems of national and regional importance in biological, environmental, physical, and social sciences relevant to agriculture, food, the environment, and communities on a peer-reviewed, competitive basis. To address these problems, NRI advances fundamental scientific knowledge in support of agriculture and coordinates opportunities to build on these scientific findings. The resulting new scientific and technological discoveries then necessitate efforts in education and extension to deliver science-based knowledge to people, allowing informed practical decisions. Competition is open to scientists at all academic institutions, Federal research agencies, private and industrial organizations, and as individuals. The NRI Program Description is distributed widely within the scientific community and among other interested groups. The FY 2007 Request for Applications contained 26 programs with 15 programs soliciting integrated research, education, and extension projects in addition to research projects.

USDA-CSREES NRI PLANT BIOLOGY

Liang-Shiou Lin and Gail McLean
USDA Cooperative State Research, Education and Extension Service (CSREES)
Competitive Programs – National Research Initiative (NRI)
Washington DC
E-mail: llin@csrees.usda.gov and gmclean@csrees.usda.gov
Telephone: 202-401-5045 and 202-401-6060
Website: http://www.csrees.usda.gov/; http://www.csrees.usda.gov/funding/nri/nri.html

The CSREES NRI competitive grants program "Plant Biology: Foundation for Agricultural and Forest Plant Production and Improvement" supports projects that will provide fundamental knowledge and training for improvement and sustainability of agricultural plant and forestry production. Knowledge of plant biology from the molecular to the systems level is essential for development of plants with increased productivity, fitness, and use. Such fundamental understanding of plant biology will allow scientists to make use of the increasing wealth of genomics data and tools and to develop new varieties of agricultural plants through techniques such as biotechnology and classical breeding.

The science-based knowledge and education contributed by this program can lead to increased economic opportunities for producers and consumers by reducing production costs, improving quality, and increasing value of agricultural plant products. This knowledge will allow U.S. agriculture to face critical needs in the areas of bioenergy, environmental change, loss of agricultural land, and increasing global competition.

The Plant Biology program consists of four program elements: Plant Biology (A): Gene Expression and Genetic Diversity; Plant Biology (B): Environmental Stress; Plant Biology (C): Biochemistry; and Plant Biology (D): Growth and Development. In FY 2007, the program elements Plant Biology (A): Gene Expression and Genetic Diversity and Plant Biology (B): Environmental Stress solicited both research and integrated projects. Integrated projects in these two program elements include a plant breeding education component. Program elements Plant Biology (C): Biochemistry and Plant Biology (D): Growth and Development are soliciting research projects only.

The National Research Initiative Competitive Grants Program (NRI) is the office in the Cooperative State Research, Education and Extension Service (CSREES) of the USDA charged with funding research on key problems of national and regional importance in biological, environmental, physical, and social sciences relevant to agriculture, food, the environment, and communities on a peer-reviewed, competitive basis. To address these problems, NRI advances fundamental scientific knowledge in support of agriculture and coordinates opportunities to build on these scientific findings. The resulting new scientific and technological discoveries then necessitate efforts in education and extension to deliver science-based knowledge to people, allowing informed practical decisions. Competition is open to scientists at all academic institutions, Federal research agencies, private and industrial organizations, and as individuals. The NRI Program Description is distributed widely within the scientific community and among other interested groups. The FY 2007 Request for Applications contained 26 programs with 15 programs soliciting integrated research, education, and extension projects in addition to research projects.

LEGUME CROPS GENOME INITIATIVE

PRESENTOR: Charles Brummer

INVESTIGATOR: Diane Bellis, Ph.D.

INSTITUTION: AgSource, Inc. * 202.412.0582 (m)
600 Pennsylvania Ave NE, Ste. 320
Washington, DC 20003 * 202.969.8902
RR1 Box 309A*Palestine, WV 26160*304.275.3087

Since 2001, the U.S. Legume Crops Genome Initiative (LCGI) has brought together the major U.S. legume commodity associations and their respective research communities to identify research strategies, and guide the development of genomic tools and resources to ensure that U.S. legumes and legume crop products remain competitive in domestic and global markets. The LCGI has provided the structure for developing a legume-community consensus on the research objectives to guide the genomic characterization of major legume crops. The commodities include alfalfa (Medicago sativa), common bean (Phaseolus vulgaris), the cool-season food legumes (pea [Pisum sativum], lentil [Lens culinaris Med.], and chickpea [Cicer arietinum]), peanut (Arachis hypogaea), and soybean (Glycine max L. Merr.). With the encouragement of USDA/ARS, the group first met at Hunt Valley in 2001, and since then has held annual meetings; met regularly with key Administration and Congressional representatives; and been supportive of the AOCS publication, Legume Crops Genomics. The activities of the LCGI led to the Cross-Legume Advances through Genomics (CATG) Conference held in December 2004 (http://catg.ucdavis.edu/) and was instrumental in the development of the FY05-06 USDA/CSREES/NRI program in legume genomics and the sequencing of the soybean genome.

COMPARATIVE MAPPING, NUCLEOTIDE DIVERSITY, AND KARYOTYPIC EVOLUTION IN PEANUT

INVESTIGATOR: Knapp, S. J.

Phone: 706-542-4021
Fax: 706-583-8120
Email: sjknapp@uga.edu

PERFORMING INSTITUTION:
CROP & SOIL SCIENCES
UNIVERSITY OF GEORGIA
110 RIVERBEND ROAD
ATHENS, GEORGIA 30602

NON-TECHNICAL SUMMARY: Peanut has historically lagged in genomics resource development, the application of genomics solutions to problems of biological and agricultural importance, and the application of marker-assisted selection in breeding programs. Marker-assisted selection has not been used in the development of any commercially important peanut cultivars. Moreover, genetic diversity in the elite gene pool is narrow. The competitiveness of the US peanut industry hinges on broadening genetic diversity and on the ability of breeders to integrate new genomic and molecular technologies into applied breeding programs. The purpose of this research is to significantly enhance the infrastructure for genomics and molecular breeding research in peanut and catalog nucleotide diversity among the 4,000 or so genes expressed in developing seeds of peanut.

OBJECTIVES: Our first objective is to gain a deeper understanding of the patterns and distribution of nucleotide and allelic diversity in peanut by: (i) resequencing alleles from several A. duranensis (AA), A. batizocoi (BB), and A. hypogaea (AABB) genotypes (primarily the parents of mapping populations); (ii) estimating SNP and insertion-deletion (INDEL) frequencies and nucleotide and haplotype diversity; and (iii) developing SNP genotyping assays and screening unsequenced wild and domesticated genotypes for SNPs. Our second objective is to develop and mine developing-seed EST databases for DNA polymorphisms by: (i) assembling and annotating 28,000 developing seed ESTs produced from two A. hypogaea genotypes; (ii) developing normalized cDNAs from RNAs isolated from developing seeds of two A. duranensis, two A. batizocoi, and four A. hypogaea genotypes; (iii) producing, annotating, and assembling 200,000 short-read ESTs (SR-ESTs = 100 bp/read) from each of the eight normalized cDNAs (8 genotypes x 200,000 SR-ESTs/genotype = 1,600,000 SR-ESTs); (iv) producing and performing bioinformatic analyses on diploid, tetraploid, and diploid-tetraploid EST assemblies and screening the unigenes for SNPs, INDELs, and SSRs; and (v) identifying SNPs for the development of highly parallel genotyping assays. Our third objective is to assess the feasibility and utility of highly parallel SNP genotyping technologies in diploid and tetraploid peanut species. Our fourth objective is to complete the development of low-density (5-10 cM/locus) intraspecific genetic maps for A- and B-genome diploid species (A. duranensis and A. batizocoi, respectively).

APPROACH: Massively parallel DNA sequencing technologies will be used to develop short-read ESTs for A. duranensis, A. batizocoi, and A. hypogaea. The SR-ESTs (1,600,000 total) will be assembled with 28,000 long-read ESTs, annotated, and publicly released. Diploid and tetraploid assemblies will be mined for DNA polymorphisms. Putative homeologous and paralogous contigs will be identified through comparative analyses of A-, B-, and AB-genome EST assemblies. SNPs will be identified for the development of arrays for highly parallel SNP genotyping. The feasibility and utility of highly parallel SNP genotyping will be tested in diploid and tetraploid peanut species using the Illumina platform. Several intraspecific F2 and recombinant inbred line (RIL) mapping populations will be developed for A. duranensis (AA), A. batizocoi (BB), and A. hypogaea (AABB). The parents of the mapping populations will be screened for simple sequence repeat (SSR) and SNP marker polymorphisms. Low-density (5-10 cM) genetic maps will be constructed for A. duranensis and A. batizocoi using F2 populations. Single-copy DNA markers for transcribed loci will be used whenever possible. Chromosomal rearrangements between A. duranensis and A. batizocoi will be identified through synteny analysis (comparing the orders of orthologous DNA marker loci).

TILLING: A COMMUNITY ORIENTED REVERSE GENETICS TOOL IN SOYBEAN

INVESTIGATOR: Meksem, K.
Phone: 618-453-3103
Fax: 618-453-7457
Email: meksemk@siu.edu
URL: www.soybeantilling.org

PERFORMING INSTITUTION:
PLANT, SOIL AND AGRICULTURAL SYSTEMS
SOUTHERN ILLINOIS UNIV
CARBONDALE, ILLINOIS 62901

NON-TECHNICAL SUMMARY: The need to provide a robust link between DNA sequences and phenotype is becoming increasingly urgent as more economically important genes are identified in soybean, therefore, a central Tilling facility is needed to speed gene discovery in soybean. The majority of the available and committed soybean genomic tools are being developed primary from two cultivars Forrest and Williams 82. We used TILLING as a reverse genetics tool for functional analysis of soybean genes using two platforms, one from Forrest and the other from Williams 82. We will optimize the technology and establish a central facility that will archive, curate, distribute and store the soybean mutagenized population's lines and provide TILLING services to the soybean community.

OBJECTIVES: The proposed work is a community oriented project, four main objectives were set to achieve our goals within the expected time frame: 1. To improve soybean TILLING to allow for an efficient non-transgenic in vivo system gene functional analysis and alternative alleles discovery in soybean. 2. To establish a central facility for soybean TILLING and provide the community with access to the available TILLING resources. 3. Deposit the Tilling lines at the soybean germplasm collection center at the University of Illinois (Urbana), and the data with both, the Legume Information System (LIS) www.comparative-legumes.org and the project's web page www.soybeantilling.org. 4. To provide training workshops and short courses to encourage collaborative efforts for developing and using TILLING as functional analysis tool in soybean and other crops.

APPROACH: 1. The initial focus of the project will be the production of suitable soybean TILLING populations using ethylmethanesulfonate (EMS) mutagenesis 2. DNA samples and seeds from the mutant lines will be produced, inventoried and archived, 3. The developed EMS plant collections will be screened for rapid, systematic, identification of mutations in targets sequences using mismatch detection enzymes (Endo1 and Cel1 enzymes). 4. A web-based resource (www.soybeantilling.org) will be fully developed for data collection, web exposure, outreach and handling the community's needs.

GENOMICS FOR PHASEOLUS AS A COMMUNITY RESOURCE FOR LEGUME RESEARCHERS

INVESTIGATOR: Jackson, S. A.; Vallejo, E.; Wing, R.; Matthew Blair
Phone: 765-496-3621
Fax: 765-496-7255
Email: sjackson@purdue.edu
URL: www.phaseolus.genomics.purdue.edu

PERFORMING INSTITUTION:
AGRONOMY
PURDUE UNIVERSITY
WEST LAFAYETTE, INDIANA 47907

NON-TECHNICAL SUMMARY: Legume crops are important sources for human nutrition as well as for fixing atmospheric nitrogen into the soil. However, genomics for many legume crops, such as *Phaseolus vulgaris* (common bean), have lagged behind other major crops. In the course of this project, we will develop infrastructure for common bean to advance genomics and genetics so that breeders and plant biologists will have an additional tool in their toolkit with which to address nutritional and production-related problems. This genomic toolkit will consist of a physical map of common bean based on cloned pieces (bacterial artificial chromosomes) of the Phaseolus genome as well as associated DNA sequence tags derived from each cloned piece of DNA. All these data will be shared publicly, unfettered, via the Legume Information System (LIS) at the National Center for Genome Resources and the GenBank at the National Institutes of Health.

OBJECTIVES: Genetics of legume species is complicated by genome duplication events and complex evolutionary histories. The genus Phaseolus, which includes common bean, is by all appearances a diploid and thus represents a basic phaseoloid genome that could facilitate genetics in other legume crops. Genomic resources for Phaseolus will be developed to 1) accelerate genetics of this important food crop, 2) develop cross-genome genetic markers than can be used to map important traits, 3) make evolutionary comparisons with other legume species, especially to infer genome duplication events, and 4) to help assemble the emerging soybean genome sequence. To reach these goals we will accomplish the following specific aims: 1) Assemble a 10x Phaseolus BAC library into contigs using high information content fingerprinting (HICF); 2) Obtain BAC end sequences (BES) from each BAC clone that is fingerprinted; 3) Integrate the genetic and physical maps of Phaseolus with a focus on ferritin genes; and 4) Merge these physical resources with those of soybean, Medicago and Lotus and make them public via the Legume Information System (LIS). These resources will be available publicly via LIS and GenBank. They will provide a valuable community resource for genetics and genomics of both Phaseolus and other Phaseoloid species such as soybean, cowpea, lentil and mung bean.

APPROACH: A 10x BAC library will be fingerprinted and assembled into contigs using FPC (FingerPrint Contigs software). These contigs will be anchored to the genetic map using hybridization of genetic markers to high density filters containing the 10x BAC library. All these data will be made public via the Legume Information System and the project website: http://phaseolus.genomics.purdue.edu.

TOOLS AND APPLICATIONS OF GENE-BY-GENE SEQUENCING IN COMMON BEAN

INVESTIGATOR: McClean, P. E.; Denton, A.
Phone: 701-231-8443
Fax: 701-231-8474
Email: phillip.mcclean@ndsu.edu

PERFORMING INSTITUTION:
PLANT SCIENCES
NORTH DAKOTA STATE UNIV
FARGO, NORTH DAKOTA 58105

NON-TECHNICAL SUMMARY: To date, genomic information has been collected for only a few select plant species. Gene sequence data is the most abundant information currently available. It is now time to use the data from those species to assist with genomic analysis in other species. Tools will be developed that organize that data in a useful manner that enables other researchers studying other plant species to apply gene sequence based analysis to their species of choice. We will demonstrate this usefulness by applying those tools to the study of common bean (Phaseolus vulgaris L.).

OBJECTIVES: Extensive genomic resources are not available for most plant researchers, yet all can benefit when they apply the data obtained by the large genome projects. The most abundant genomic resource is the collection of gene sequences developed from gene-by-gene or whole genome sequencing approaches. It is now time for researchers working with other crop species to mine this sequence information and convert it into useful genomic tools. The first objective of this project is to develop tools and procedures that define clusters of plant orthologs and paralogs. In addition, a WWW interface will be developed that displays multiple alignments of cluster members. Finally, approaches will be defined and implemented that will define primer sequences for the subsequent amplification of a cluster ortholog from another plant species. Common bean (Phaseolus vulgaris L.) is one species for which genomic resources are under developed. A greater abundance of gene sequence data would be useful to coordinate the genetic and physical map of the species. In addition, the sequence data is critical for the discovery of genes controlling critical phenotypes using candidate gene and association mapping approaches. Therefore, the second objective is to apply these clustering, alignment, and primer development tools to sequence a fragment(s) from 300 genes from common bean. These genes will then be mapped on the community wide BAT93 x Jalo EEP558 linkage map.

APPROACH: Objective 1: Complete gene models will be collected from all genes characterized by gene-by-gene approaches or complete genome sequencing. An all-against-all BLAST analysis will be performed. Then a complete linkage hierarchal clustering will be performed at specific e-values to define the orthologous/paralogous relationships. Clusters of ortholog/paralogs defined at a specific e-value will be aligned using multiple alignment techniques. The alignment data will then be used as input data to define primers for amplification of orthologous fragments from other species. The clusters, multiple alignments, and primer information will be delivered via a WWW interface that enables the user to apply a taxonomic approach to select cluster members from which clusters, multiple alignments, and primers are defined. Objective 2: Orthologs discovered using the tools described above, along with any available EST data, will be used to define primers for PCR amplification of gene fragments from the common bean (Phaseolus vulgaris L.) genotypes BAT93 and Jallo EEP558. The genes will represent those involved in metabolic pathways, Arabidopsis mutant phenotypes, and other relevant agronomic phenotypes. The fragments will be sequenced, and indel or SNP polymorphism will be used to develop mapping primers. The BAT93 x Jalo EEP558 recombinant inbred population will be scored and a linkage map consisting of these genes along with previously described molecular and phenotypic markers will be developed.

KEYWORDS: genomics; gene sequences; orthologs/paralogs; complete hierarchical clustering; multiple alignment; taxonomic clustering; primer design; parallel processing; database; common bean; phaseolus vulgaris l.; gene sequencing; genetic mapping

PROGRESS: 2005/04 TO 2006/04

Database development: The goal of this aspect of the project is to develop a database from which users interested in gene-by-sequence can make a query and be offered a suite of primers from which the target gene could be amplified from their species of interest. To that end, we created a database structure to store all relevant data. We have downloaded all of the publicly available gene models for Arabidopsis, rice, Medicago, and maize from databases involved in curating this data. In addition, we downloaded all full gene models available from GenBank for species other than these model species. Collectively, we have over 100,000 sequences. These sequences were analyzed in an all-against-all manner using blastp and clustered using complete linkage clustering. Alignments of clusters were performed for all clusters at specific similarity levels using the MultAlin algorithm. An algorithm was developed to search the alignment for the best regions for primer development. A perl script was tested to pass specific sequences to Primer3 for primer development. More specifically, we implemented a BioSQL schema in the PostGreSQL database. For clustering, 50% of each gene involved in the cluster constraint was required. Finally, to evaluate the clustering, we used a histogram-based evaluation measure and determined the complete linkage clustering provided better alignments than single linkage clustering. Gene-by-gene sequencing of common bean: The second goal of the project is to perform gene-by-gene sequencing with common bean. We collected all of the known tentative consensus (TC) sequences of common bean and compared them with all of the gene models from Arabidopsis. Genes to be sequenced were selected based on similarity in a BLAST search. TC sequences were used as a query for an all-against-all blastp analysis against individual databases containing Arabidopsis thaliana genes with mutant phenotypes, genes under selection during domestication in maize, A. thaliana genes involved in biochemical pathways, and all A. thaliana genes. A gene was selected for sequencing if had at least 100 nucleotides in the 3 prime UTR and an E-value less than e-30 with the top hit. Primers were designed with Primer3 with a target TC fragment size of 450-500 nucleotides, primer size of 18-28 nucleotides, and Tm of all primers about 58oC. The 3 prime primer was targeted to a location 150 nt downstream of the putative stop codon. Fragments were amplified from BAT93 (Bat) and Jalo EEP558 (Jalo) genomic DNA and directly sequenced. Of the more than1000 genes analyzed to date, DNA sequence data for the two genotypes were obtained for 322. Of these, 222 genes were polymorphic. A total of 1003 polymorphisms were detected, and of these 85.5% were SNPs. On average, one SNP was detected every 151 nt, and one indel was observed every 897 nt. 44.1% of the polymorphisms were located in introns, 38.7% in exons, and 17.0% in the 3 prime UTR. SNPs were evenly distributed between introns and exons, whereas indels were largely found with introns. The sequence polymorphism data was used to developed CAPS markers, and to date, we have mapped 52 genes on the Bat x Jalo linkage map.

IMPACT: 2005/04 TO 2006/04

The development of the database will enable researchers working on species without significant sequence data to apply the modern candidate gene genomic approach to their research. The gene-based map of common bean will provide a framework for 1) gene cloning of important agronomic target genes in common bean and 2) the application of comparative legume genome analysis to the improvement of the crop.

EVOLUTIONARY GENOMICS OF SMALL MULTIGENE FAMILIES OF AGRONOMIC INTEREST IN COMMON BEAN (PHASEOLUS VULGARIS)

INVESTIGATOR: Gepts, P. L.
Phone: 530-752-7743
Fax: 530-752-4361
Email: plgepts@ucdavis.edu

PERFORMING INSTITUTION:
PLANT SCIENCES
UNIV OF CALIFORNIA
DAVIS, CALIFORNIA 95616

NON-TECHNICAL SUMMARY: Crop plants harbor a large amount of genetic diversity that has largely remained unutilized. An understanding how this diversity is generated may help increase its usefulness and utilization. In this proposal, researchers at the University of California, Davis, will investigate how DNA sequence changes have taken place in the evolutionary lineages leading to modern bean cultivars. Specifically, they will look at two contrasting traits: one is an insect resistance trait operating in the seed and the other is a plant growth habit. These two traits have been subject to different selection pressures in the recent evolutionary past of beans in Central and South America. Seed weevils are distributed mainly in Central and northern South America. Thus, they may only have affected beans in those areas, which would account for bean resistance only present in beans from those areas. In contrast, the plant growth habit trait conditions a shorter, bush plant type. This type may have been selected in the earliest steps of agriculture in the Andes before the arrival of maize. To address this issue, we will take advantage of large-DNA-insert libraries that had been developed earlier with USDA CSREES support. Large DNA inserts will be sequenced in five different lines representing key steps in the evolution of beans. These sequences will allow us to determine which type of changes have led to the new traits (insect resistance, plant habit) and to what extent selection for these new traits has affected DNA sequences adjacent to the genes controlling these traits.

OBJECTIVES: 1. Determine the structural evolution at two loci and adjacent regions, which are subject to different selection pressures: *APA* or *Arcelin Phytohemagglutinin a-Amylase Inhibitor* (*Arl-Lec-aAI*) and *Determinacy (fin)* in five genotypes representing key steps in *Phaseolus* evolution in most legumes, including most of the genus *Phaseolus*, the *APA* protein family is only represented by the Phytohemagglutinin subfamily. It is only in one of the nine clades, the one leading up to *P. vulgaris*, that this gene family started diversifying in the last 2-4 million years. The *APA* BAC sequence available for the G02771 genotype shows that the sequences for the three subfamilies are each separated by a retro-transposon element. We hypothesize that retrotransposons may be involved somehow in the sudden diversification of the *APA* family, following their insertion in the vicinity of the primordial Phytohemagglutinin genes. For the *fin* locus, we already know that the main determinate mutation may be due to a retrotransposon insertion. However, we know little about the origin of this retrotransposon, in particular whether it originated from a close proximity of the *fin* gene. 2. Compare sequence variation across *Phaseolus vulgaris* for markers and genes in and around the *APA* and *fin* loci. In most if not all crops, domestication has induced a strong reduction in genetic diversity. This reduction is due to two main phenomena: a) genetic drift and b) selection. The former will affect all loci in the genome whereas the latter will affect only those loci that were under selection during domestication. Therefore, in order to identify loci with an effect of selection on top of the more general reduction of genetic diversity due to genetic drift, one needs to compare the genetic diversity between wild and domesticated beans at: a) (presumably) neutral loci (as a control) and b) domestication locus (*fin*). If domestication loci such as *fin* are indeed subject to selection, one expects a stronger reduction of genetic diversity from wild to domesticated beans than for other loci not involved in domestication. A similar argument can be made for the contrast between Andean and Mesoamerican beans, where the reduction of genetic diversity due to putative selection at, for example, a locus in the Mesoamerican gene pool (e.g., *APA*), should be evaluated in comparison with average differences in genetic diversity between the Andean and Mesoamerican gene pools. Furthermore, recombination will cause

any reduction of genetic diversity due to selection to diminish at increasing distances from the locus under selection. Likewise, recombination will cause linkage disequilibrium to decay at increasing distances from the selected locus. How quickly this decay takes place depends on various factors such as the level of recombination and the intensity of selection.

APPROACH: *fin*-sequence containing BACs will be identified in four libraries, *P. vulgaris* G02771, G21245, and G19833 and *P. lunatus* Henderson by PVR screening. After restriction mapping and assembly of the sequence contigs, two minimally overlapping (minimum tiling) BAC clones for each locus will be picked for each genotype. These BAC clones will be fully sequenced. Assembly and annotation of the sequences will take place using standard procedures. The final assembly will be compared to restriction digests of the clone insert to verify the order and size of the contigs. The sequences of BAC clones are then screened for open reading frames with GeneScan, TwinScan, and FGENESH. Putative coding regions are compared to Genbank using BLAST and potential matches aligned with BioEdit and ClustalW. A phylogenetic tree of *APA* and PvTFL1y DNA sequences will be constructed with BioEdit using the neighbor-joining method of Saitou and Nei (1987) and the UPGMA clustering algorithm. Bootstrap values will be calculated with PAUP. BAC sequences will provide sequences of the *APA* and *fin* loci and of genes adjacent to these loci at known physical distances from them. Robust primers developed based on conserved motifs will be tested against a representative *P. vulgaris* core sample of 20 wild and domesticated accessions to assess the reproducibility of their amplification and the level of discrimination they can provide after sequencing of the amplicons. As a control (i.e., presumably unselected genes), primers for the inserts of some 20 mapped RFLP markers and distributed throughout the genome, will be tested. These primers will then be used against genomic DNA will of a representative sample. In addition, variation for a set of 22 polymorphic microsatellite markers (2 per linkage group) will have been analyzed in the same plant sample in order to identify structure in common bean. Population structure will be identified using the Structure software. For each of the loci (*APA, fin,* and adjacent genes), sequence diversity parameters will be determined by DnaSP. The haplotypes of each gene will be clustered according to genetic distance to establish a correspondence, if any, with existing common bean gene pools (Piffanelli et al. 2004). The comparison of nucleotide diversity of the *fin* locus between determinate and indeterminate genotypes, and between wild and domesticated accessions will be analyzed to identify a potential molecular signature of selection during domestication. Similar approaches will be followed for the *APA* genes. To compare the extent of selection pressure on each class, several neutrality tests will be performed (Tajimas D, HKA, and McDonald & Kreitman) using DnaSP. Neighbor-joining trees of the different sequences in our sample will be constructed using the MEGA. Pair wise estimations of linkage disequilibrium (LD) (D and r2 and corresponding probability) at and around the and *fin* genes will be calculated to verify the physical extent of selection around these genes and to compare the consequences of different selection pressures on this genomic region.

DISEASE RESISTANCE ORTHOLOGS AND THEIR LOCALIZATION IN THE COMMON BEAN

INVESTIGATOR: Khwaja, H.
Phone: 701-788-4761
Fax: 701-788-4748
Email: steven_bensen@mayvillestate.edu

PERFORMING INSTITUTION:
MAYVILLE STATE UNIVERSITY
330 THIRD STREET NE
MAYVILLE, NORTH DAKOTA 58257

NON-TECHNICAL SUMMARY: The genomic database resources of important plant and animal species and tools required to analyze these resources are now publicly available. The exploitation of these resources enables the rapid identification of genes and facilitates the examination of their functions in related species. The common bean is one of the most important species of leguminous crops but it lacks adequate genomic resources and the average global yield of this species is low because of its susceptibility to wide range of diseases. To maximize and sustain the production, it is essential to develop high yielding varieties of common bean with disease resistance traits. Classical breeding in common bean has made excellent progresses in this aspect in the last two decades. To speed up the progress in this work and harness the valuable genes conditioning disease resistance in common bean, it is necessary to identify the portion of the genome responsible for the expression of these resistance traits along with their associated molecular markers. Using genomic database resources of related species, this project is aimed at delivering equivalent genomic portions of common bean conditioning disease resistance and associated markers to monitor their introgression into the breeding lines of common bean. This project will expand the research base of Mayville State University (MSU) and strengthen the PD`s position for future competitive grants. Hands-on experience with these problem based research will allow students to explore a rapidly changing area and the conceptual tools involved in these areas of biology.

OBJECTIVES: Biological sciences have been revolutionized, not only in the way research is conducted but also in the way findings are communicated to professionals and to the public. The ever increasing number of publicly available database resources of genomic sequences of different plant and animal species, full length genomic sequences and corresponding coding sequences, underlying protein sequences, gene and gene families, gene ontology, and comparative genomics have opened up the scope of utilizing the gene information of one species for the benefit of other species lacking basic genomic tools. The common bean consisting of dry and snap beans, is one of the most important species of leguminous crops but there are many diseases that significantly reduce the persistence and productivity of common bean in the US and worldwide. Due to the lack of adequate genomic resources, bean breeders and growers across the world have been suffering from a shortage of functional markers for tagging and monitoring disease resistance traits in this species. Using the genomic database resources of related species, this project will identify equivalent genomic resources and develop molecular markers associated with disease resistance genes in common bean. It will take advantage of the high level of gene conservation in an inter-specific manner in the search for functional genes. In this project we plan to use this information with the following objectives: 1) Develop and validate orthologous gene sequences from Leguminosae in common bean focusing on candidate genes for different disease resistance 2) Develop functional markers segregating in different mapping population and localize them in the genome of common bean. The development and validation of orthologous gene sequences and addition of gene-based markers in common bean focusing on candidate genes for disease resistance will augment the resistance gene tagging, mapping, and marker-assisted breeding.

APPROACH

1. Identify disease resistance gene orthologs in common beans: Genomic databases will be searched for all possible disease resistance related genes in leguminous species. The tentative consensus and singleton EST sequences of common bean will be identified by blasting those gene sequences against the eukaryotic gene orthologs (EGO) and the *P. coccineus* EST databases. Corresponding common bean sequence of each query sequence obtained from these two databases will be compared and common bean ESTs with higher scores and probability will be considered. The queries with no hit from any of the databases will be blasted against *Medicago truncatula* and soybean ESTs databases.

2. Develop markers for disease resistance gene: PCR-primers will be designed from the identified sequences. Based on the predicted average intron size of 161 bp in Medicago truncatula, primer pairs will be designed to amplify genomic DNA of maximum 350 bp in size using the web-based PCR primer designing program. Gene sequences with over 350 bp, overlapping primer pairs will be designed, so that assembled amplified products corresponds to the equivalent size of the gene. All of the designed primers will be used to amplify genomic DNA of the parents of mapping populations segregating for different disease resistance traits. Primer pairs producing polymorphic products will be grouped according to the segregating populations and considered as EST-STS markers. The amplified genomic DNA will be sequenced and SSRs will be identified using Simple Sequence Repeat Identification Tool (SSRIT). The primers developed from the flanking SSR sequences will be considered as EST-SSRs and will be assessed against the populations segregating for disease resistance traits. In the cases, where no EST-STSs and EST-SSRs are found, amplified DNA sequences between the parents will be aligned to survey the parental alleles for polymorphic sites. Single-nucleotide polymorphisms (SNPs) will be converted to cleaved amplified polymorphic sequences (CAPS) by identifying SNPs that confer differential restriction enzyme sites between the two parental alleles. In cases in which a suitable restriction enzyme site is not identified, oligonucleotide primers with a single nucleotide mismatch will be designed to the polymorphic position and will be considered as EST-SNPs. To validate the resistance gene orthologs in common bean, the resulting gene sequences in common bean will be compared with the starting resistance gene sequences as well as with the sequences from which primer pairs were designed

3. Localize resistance gene orthologs in the bean genome: All the sequence and marker information will be made available through LIS (Legume Information Service) to assess and map disease resistance traits for the bean mapping populations as well as other related legume species. The resulting polymorphic markers developed in this study with strong associations with disease resistance will be integrated in the bean BJ core mapping population.

FULL-LENGTH CDNA RESOURCES FOR LEGUME GENOMICS

INVESTIGATOR: Town, C. D.

Phone: 301-795-7523
Fax: 301-838-0208
Email: cdtown@tigr.org
URL: http://www.tigr.org/plantProjects.shtml

PERFORMING INSTITUTION:
The Institute for Genomic Research
Rockville, MARYLAND 20850

NON-TECHNICAL SUMMARY: Legumes are one of the two most important crop families in the world and, due to their ability to convert atmospheric nitrogen into organic compounds, have high levels of protein that provides nearly 33% of all human nutritional requirements for nitrogen. Large-scale genome sequencing efforts that will provide many fundamental insights into legume biology are underway for two model species with small genomes, Medicago truncatula and Lotus japonicus, and one crop species, soybean, which is a mainstay on US agriculture. Complementary DNAs (cDNAs) are derived from the messenger RNAs (mRNAs) of expressed genes by a process of reverse transcription followed by second strand DNA synthesis. Due to technical limitations, not all cDNAs represent the entire mRNA of a gene. Full-length cDNAs (FL-cDNAs) contain the entire coding sequence of a gene and are important in plant genomics for two reasons: gene structure annotation and gene function analysis. In the research proposed here, we will generate and sequence large collections (12,000-15,000) of FL-cDNA clones and sequences for each of the three species and make them publicly available to the academic and industrial agricultural research communities. These sequences will be used to provide high quality annotation of gene structures in the genome sequences of Medicago, Lotus and soybean. The clones will be used for functional studies by expressing the encoded proteins and studying their biochemical properties such as catalytic activity and the nature of the proteins with which they interact.

OBJECTIVES: The project will generate clones and sequences for approximately 12,000 full-length cDNAs from each of three species - Medicago truncatula, Lotus japonicus and Glycine max (soybean).

APPROACH: For each species, the project will generate normalized libraries from RNAs pooled from 3 groups of tissues: 1. Flowers, early seed, late seed, stems; 2. leaves, abiotic and biotic stressed leaves, tissue culture and elicitor-treated plants; 3. roots, early and late nodules, abiotic and biotic stress. From each library, 13,000-14,000 clones will be sequenced from the 5' end. Potentially full-length clones will be identified by comparison of the 5' sequence with a comprehensive protein database. Representative full-length clones for each distinct coding sequence will be re-arrayed and sequenced from both ends. All sequences will be deposited in GenBank and the clones distributed through the Noble Foundation in the US and INRA-CNRGV in the EU.

DEVELOPMENT AND USE OF NOVEL TOOLS FOR FUNCTIONAL GENOMIC ANALYSIS OF SEED STORAGE METABOLISM IN THE MODEL GENOME MEDICAGO TRUNCATULA

INVESTIGATOR: Udvardi, M. K.
Phone: 580-224-6655
Fax: 580-224-6692
Email: mudvardi@noble.org

PERFORMING INSTITUTION:
THE SAMUEL ROBERTS NOBLE FOUNDATION, INC.
2510 SAM NOBLE PARKWAY
ARDMORE, OKLAHOMA 73401

NON-TECHNICAL SUMMARY: Legumes are a key to sustainable agriculture and are also a rich source of food and feed for humans and animals, respectively. Genomics and functional genomics technologies offer new ways of breeding legumes to improve food and feed quality. This project will develop new tools for functional genomics studies of the model legume, Medicago truncatula that will be used to identify regulatory genes that control seed storage metabolism in this species. Such genes may be useful to improve seed quality traits in crop legumes in the future

OBJECTIVES: 1. Develop a comparative legume gene expression atlas for the major organ systems of the two reference legumes, Medicago truncatula and Lotus japonicus, including a detailed time series for seed development in both species 2. Develop gene-specific primers for quantitative (q)RT-PCR analysis of all transcription factor (TF) genes in Medicago 3. Develop gene-specific primers for qRT-PCR analysis of microRNA (miRNA) genes in Medicago 4. Develop a tiling array to map transcriptional activity in the Medicago genome and to improve gene annotation 5. Identify and functionally characterize key regulators, especially TF proteins and miRNAs, which control cell and tissue differentiation and storage compound (protein and lipid) biosynthesis during Medicago seed development.

APPROACH: A comparative legume gene expression atlas will be produced using Affymetrix GeneChips for Medicago truncatula and Lotus japonicus to quantify transcript levels for approximately 90% of all genes in the major organ systems of these two species. Bioinformatics approaches will be used to identify key regulatory genes in Medicago of the transcription factor (TF) and microRNA (miRNA) classes, and to design gene-specific primers for all of these. Quantitative (q)RT-PCR will be used to test the primers and to quantify TF and miRNA transcript levels in different organs of Medicago. The gene expression atlas and qRT-PCR tools will be used to identify seed-specific TF and miRNA genes that may control storage metabolism during seed development in Medicago. Two mutant populations of Medicago truncatula will be screened for mutants defective in selected TF and miRNA genes, and the nutrient composition of seed from these mutants will be analyzed to identify regulatory genes that may have value in future plant breeding efforts. Finally, a genome tiling array will be developed for Medicago and hybridized with probes derived from total RNA of different organs to identify transcribed regions of the genome, which will aid genome annotation.

ASSESSING RNAI AS A REVERSE GENETIC TOOL FOR GLOBAL ANALYSIS OF NBS-LRR GENE FUNCTION IN MEDICAGO TRUNCATULA

INVESTIGATOR: Zhu, H.
Phone: 859-257-3647
Fax: 859-323-1077
Email: hzhu4@uky.edu

PERFORMING INSTITUTION:
AGRONOMY
UNIVERSITY OF KENTUCKY
LEXINGTON, KENTUCKY 40546

NON-TECHNICAL SUMMARY: The majority of functionally-characterized disease resistance (R) genes encode NBS-LRR proteins. However, plants contain hundreds of NBS-LRR encoding genes, and it is still difficult to identify the genes that are responsible for resistance to particular pathogens. The availability of a nearly complete catalog of M. truncatula NBS-LRR gene sequences permits us to design RNAi constructs that target specific gene family members, or a small number of highly related genes in a single transgenic plant. In principle, a limited number of RNAi constructs can be used to survey the function of NBS-LRR alleles across a variety of M. truncatula ecotypes with diverse pathogen specificities. Ultimately, a combination of forward (genetic mapping) and reverse (RNAi) genetic tools will greatly expedite the global analysis of NBS-LRR gene function in M. truncatula. In a broader sense, this research will facilitate the genetic improvement of disease resistance in alfalfa (M. sativa). Alfalfa is the fourth-most important crop in the US but has an intractable genetic system. However, alfalfa is closely related to the model legume M. truncatula and both species share many common pathogens. The resistance genes identified in M. truncatula will provide new tools for the improvement of cultivated alfalfa. The success of this project will provide a new technology for global analysis of NBS-LRR gene function in plants.

OBJECTIVES: The objective of this research is to utilize RNAi (RNA interference) as a reverse genetic tool for global analysis of NBS-LRR gene function in Medicago truncatula.

APPROACH: Our strategy is to inactivate multiple target genes in a single transgenic line, thus reducing the number of analyses required for global analysis of NBS-LRR gene function. Transgenic lines with successfully knocked-down NBS-LRR genes will be made publicly available for comparing disease responses to a large number of pathogens between wild-type and transgenic plants.

ELUCIDATING THE SMALL RNA COMPONENT OF THE MEDICAGO TRUNCATULA TRANSCRIPTOME

INVESTIGATOR: Sherrier, D. J.; Green, P. J.; Meyers, B. C.
Phone: 302-239-2322
Fax: 302-831-3447
Email: sherrier@udel.edu
URL: http://www.mpss.udel.edu/legume
PERFORMING INSTITUTION:
PLANT & SOIL SCIENCES
UNIVERSITY OF DELAWARE
NEWARK, DELAWARE 19717

NON-TECHNICAL SUMMARY: Small RNAs have specific biological activities in plants, usually as regulators of gene activity. However, most of these small RNA molecules have never been identified or measured in legumes or legumes interacting with microbes, and their function is unknown. We have developed a new technology for the identification and measurement of these molecules and this project will be the first large-scale application of the method in legumes. The purpose of this study is to identify an extensive set of small RNAs from the model legume M. truncatula under various legume tissues that are important for agriculture, including foliage and developing seeds, and tissues from plants interacting symbiotic and pathogenic microbes.

OBJECTIVES: The goal of this proposal is to use novel methods for the isolation and characterization of small RNA molecules (21 to 24 nucleotides) from the model legume Medicago truncatula. We will also generate a less extensive dataset of small RNAs from soybean (Glycine max) and common bean (Phaseolus vulgaris). These molecules will be sequenced using either of the novel sequencing methods called 454 or sequencing-by-synthesis (SBS). We will generate data from 12 libraries. The libraries will represent diverse wild-type M. truncatula tissues, tissues of plants interacting with symbiotic and pathogenic microbes, and similar tissues from soybean and common bean. Legume-specific sequences, regulated small RNAs, and new miRNAs will be identified and examples verified in laboratory experiments. Roles of selected miRNAs will be examined in transgenic plants. To facilitate public use and access to the legume small RNA data, we will create a publicly available, user-friendly website (http://mpss.udel.edu/legume).

APPROACH: The goal of this proposal is to use novel methods for the isolation and characterization of small RNA molecules (21 to 24 nucleotides) from the model legume Medicago truncatula. We will also generate a less extensive dataset of small RNAs from soybean (Glycine max) and common bean (Phaseolus vulgaris). These molecules will be sequenced using either of the novel sequencing methods called 454 or sequencing-by-synthesis (SBS). We will generate data from 12 libraries. The libraries will represent diverse wild-type M. truncatula tissues, tissues of plants interacting with symbiotic and pathogenic microbes, and similar tissues from soybean and common bean. Legume-specific sequences, regulated small RNAs, and new miRNAs will be identified and examples verified in laboratory experiments. Roles of selected miRNAs will be examined in transgenic plants. To facilitate public use and access to the legume small RNA data, we will create a publicly available, user-friendly website (http://mpss.udel.edu/legume). We will use one of the methods called 454 or sequencing-by-synthesis (SBS) for the isolation and characterization of small RNA molecules (21 to 24 nucleotides) from M. truncatula. 454 was developed by 454 Life Sciences (Branford, CT) and involves the cloning and parallel sequencing of ~200,000 distinct small RNAs per reaction. SBS was invented and commercialized by Solexa, Inc. (Hayward, CA), and is based on PCR colonies that are sequenced using novel chemistries and enzymes. The application of parallel sequencing methods to small RNA sequencing was developed by the PIs. The method is comprised of the following steps: 1) Small RNA molecules are isolated by size fractionation on a polyacrylamide gel; 2) RNA adapters are added to the ends of the single-stranded RNA molecule; 3) Reverse transcriptase is used to generate the first strand of cDNA; 4) PCR is used to amplify the cDNA to obtain sufficient quantities of template; 5) The product of the PCR reaction is then shipped to either of the companies (454 Life Sciences or Solexa) for sequencing. Once sequencing is complete, they return the data files to the Meyers lab, and we will perform all of the bioinformatics analyses to analyze the data and make

it publicly available. We will analyze a subset of these small RNAs using RACE and RNA gel blots to verify their regulation, identify new miRNAs and demonstrate cleavage of miRNA targets. To provide insight about the function of several miRNAs overexpression studies will be performed.

USING ASSOCIATION MAPPING TO IDENTIFY MARKERS FOR CELL WALL CONSTITUENTS AND BIOMASS YIELD IN ALFALFA

INVESTIGATOR: BRUMMER, E. C.; DOYLE, J. J.; MOORE, K. J.
Phone: 706-542-2461
Fax: 706-542-0914
Email: BRUMMER@IASTATE.EDU

PERFORMING INSTITUTION:
CROP & SOIL SCIENCES
UNIVERSITY OF GEORGIA
110 RIVERBEND ROAD
ATHENS, GEORGIA 30602

NON-TECHNICAL SUMMARY: Alfalfa (Medicago sativa) is a potential biofuel crop because it produces high yield, its leaves can be used as a high value, high protein co-product, it fixes atmospheric nitrogen, and it has beneficial effects on the environment. Improving alfalfa as a biofuel crop will entail breeding for increased biomass yield and altered cell wall composition. While traditional phenotypic selection can be successful, the perennial nature of alfalfa requires that a selection cycle lasts for several years. Decreasing the cycle time would increase genetic gain for all traits. This could be achieved using marker assisted selection for the traits of interest, but marker identification research conducted previously has not focused on representative alfalfa breeding populations nor has it examined wild germplasm as a source of new alleles to improve agronomically important traits. Our experiment will address these issues by studying both wild germplasm not typically used in alfalfa breeding programs and also a cultivated breeding population currently under selection. We will evaluate biomass yield and cell wall composition in the field. Concurrently, we will evaluate the genotype of each plant using genetic markers selected throughout the genome. We will also develop markers based on DNA sequence variation in genes of possible involvement in cell wall synthesis. Ultimately, this project will improve the efficiency of selection for enhanced bioenergy characteristics in alfalfa, produce numerous new markers at important candidate genes, and identify potentially useful alleles in wild germplasm.

OBJECTIVES: Our objectives are to use genomics approaches to identify chromosomal regions, and ultimately genes, controlling the two most important bioenergy traits, biomass yield and composition, and to develop genetic markers that can be used directly in applied plant breeding programs to improve the bioenergy qualities of alfalfa. We will pursue two complementary objectives to attain our goals: 1. Identify loci, and specific alleles, that control the concentration of alfalfa stem cell wall constituents and that are associated with biomass production using whole genome and candidate gene association mapping across a diverse set of natural diploid alfalfa accessions, and 2. Extend the analysis and methods used in the first objective to a tetraploid alfalfa breeding population currently under selection. As a result of this project, we (a) will have identified novel alleles in wild alfalfa germplasm that may be useful to improve cultivated alfalfa; (b) will have developed and used SNP markers in genes known to be involved in the biosynthesis of cell wall composition; (c) will be able to select individuals within a breeding population on the basis of these markers, and (d) will identify new alleles from wild germplasm useful for improving cultivated alfalfa. This experiment will provide the first estimate of linkage disequilibrium (LD) in alfalfa, both in a broad cross-section of wild diploid germplasm and in a practically important cultivated breeding population, on a genome-wide and on an individual gene basis. Additionally, we will have applied association mapping to this important crop legume for the first time.

APPROACH: We will use association mapping to identify genome regions and candidate genes that are associated with biomass production and cell wall composition in both diploid and tetraploid alfalfa populations. We are proposing to begin by screening a broad diversity of diploid germplasm (three individuals from each of 96 plant introductions) in order to identify new genetic variation for these traits that could be useful in alfalfa improvement. We will begin by analyzing diploid genotypes because they likely harbor a reservoir of

unexploited genetic diversity and are more tractable for association mapping experiments than tetraploid genotypes. Subsequently we will extend the results to tetraploids. The tetraploid population we will examine is a breeding population currently under clonal selection at four locations, with 200 individuals being evaluated. As a breeding population, markers associated with traits could be immediately used in a recurrent selection program leading to the development of improved cultivars. Phenotypic analysis will be conducted based on field grown plant material clonally replicated to enable assessment of individual genotypes. In addition to biomass production and plant height measurements, we will conduct a through analysis of the stem cell wall composition of all entries. All plants will be genotyped throughout the genome with simple sequence repeat (SSR) markers, some of which will be selected based on their association with quantitative trait loci (QTL) for biomass yield, stem cell wall cellulose, hemicellulose, and lignin concentration, or agronomic traits that we have identified in other experiments. Concurrently, we will sequence portions of up to 100 genes that are candidate loci involved with cell wall biosynthesis. The sequencing will lead to the identification of single nucleotide polymorphisms (SNP), which we will develop into markers for those specific genes. All plants (288 diploid and 200 tetraploid) will be genotyped with the SNP markers. Association mapping will be conducted using the recently described mixed-model method that will account for underlying population structure within our two groups of genotypes (diploid and tetraploid), which will be analyzed separately. We will test for associations based on both genome-wide SSR molecular markers, as well as on SNP markers for 20 candidate genes, which will be developed from sequence data on 96 diploid and 20 tetraploid individuals.

GENETIC AND STATISTICAL ASPECTS OF ASSOCIATION MAPPING

INVESTIGATOR: Yu, J.
Phone: 785-532-3397
Fax: 785-532-6094
Email: jyu@ksu.edu

PERFORMING INSTITUTION:
AGRONOMY
KANSAS STATE UNIV
MANHATTAN, KANSAS 66506

NON-TECHNICAL SUMMARY: Vast amounts of currently available genome information have not been fully utilized to significantly improve plant breeding. While this failure is attributable, in part, to the success of existing breeding methods, much of our genome resources go untapped due to poor connections linking genomic technology and breeding methodology. Association mapping holds great promise for the dissection of complex traits. However, little effort has been made to develop robust methods of association mapping in plant species when compared to the well developed methods in linkage analysis. We propose to develop genetic and statistical methods for dissecting complex agronomic and physiological traits, as well as design improved genomic-aided selection strategies for plant breeding. Comprehensive computer modeling and information-driven methodology development will be conducted with data from association studies in multiple crop species. Ultimately, new insights gained from the dissection of complex traits will not only facilitate future dissection, but also help to develop new genomic-aided breeding methods for crop improvement.

OBJECTIVES: The goal of this research is to 1) systematically examine different aspects of association mapping, germplasm selection, phenotyping, relatedness detection with genomic information, and statistical methodology; 2) provide general guidance in the application of genetic association mapping for gene discovery in many plant species; and 3) result in practical analytical tools for future applications of association mapping studies.

APPROACH: Both computer modeling and information-driven methodology development will be conducted. The proposed project will utilize multiple association mapping panels from different crop species for the purpose of developing a general methodology framework for widespread use in the research community

THE CONSENSUS LEGUME DATABASE II: BILOLOGICAL INFERENCE FROM DISTURBUTED INTEROPERABLE DATABASES

INVESTIGATOR: Retzel, E. F.

PERFORMING INSTITUTION:
SPONSORED PROJECTS ADMINISTRATION
UNIVERSITY OF MINNESOTA
420 DELAWARE STREET SE
MINNEAPOLIS, MINNESOTA 55455

NON-TECHNICAL SUMMARY: Considerable genomic data is available for the legumes. Much of this data can be integrated and made available to both public and private sector researchers and breeders, as well as integrated with other database efforts such as the Legume Information System. This project makes use of extensive external resources for sequence data, as well as developing new resources such as the legume metabolic resource. These resources will provide new insights into the genomic sequence data, and will be made available to researchers and breeders both on the web and through direct queries through web services protocols

OBJECTIVES: There are three primary objectives: 1. Development of legume EST resources utilizing data from The Gene Indices; 2. Develop a LegCyc database to house annotated pathway information; 3. Develop the appropriate web services under semantic BioMoby to allow development of rich queries and implement an Ensemble server (provides DAS functionality).

APPROACH: Our approach will be to focus on available resources for assembled data using the Gene Indices developed in the Quackenbush laboratory. We will use data from the Medicago and Lotus sequencing projects to develop a genomic context for data. New resources from this project will include a robust database of EST information, as well as an instance of the BioCyc / MetaCyc tools for examining metabolic pathway information. These data resources will be made available using semantic web-services.

FINISHING THE SEQUENCE OF EUCHROMATIN IN THE MODEL LEGUME, MEDICAGO TRUNCATULA

INVESTIGATORS: Nevin D. Young[1] , Christopher D. Town[2] , Bruce A. Roe[3]

PERFORMING INSTITUTION:
[1]DEPT. OF PLANT PATHOLOGY, UNIVERSITY OF MINNESOTA, 495 BORLAUG HALL, ST PAUL, MN 55108 USA
[2]THE INSTITUTE FOR GENOMIC RESEARCH (TIGR), 9712 MEDICAL CENTER DRIVE, ROCKVILLE, MD 20850 USA
[3]ADVANCED CENTER FOR GENOME TECHNOLOGY (ACGT), STEPHENSON RESEARCH & TECHNOLOGY CENTER, NORMAN, OK 73019 USA

An international consortium is sequencing the euchromatin of the model legume, Medicago truncatula, with a goal of completion in 2008. Previous research demonstrated that nearly all Medicago genes reside in euchromatic chromosome arms, separate from heterochromatin surrounding centromeres. Sequence data so far supports this expectation. Since Medicago sequencing is “BAC-by-BAC”, the resulting sequence will be high quality, spanning entire chromosome arms with minimal gaps, and be directly useful in comparative genomics. Based on projections from EST-coverage, the “gene-space” of Medicago is ~300 Mbp with ~70% complete as of October 2006. Project-wide automated annotation is coordinated by IMGAG (International Medicago Genome Annotation Group) utilizing a pipeline of intrinsic and extrinsic gene-finding algorithms. Medicago displays extensive synteny with other legumes, including Lotus japonicus and soybean, and substantial synteny with poplar, Arabidopsis and rice. In comparisons with Lotus ten large-scale synteny blocks are observed, altogether spanning ~70% of both genomes. Gene densities and proportion of genes conserved within synteny blocks are homogeneous throughout the genomes. However, gene-containing regions in Medicago occupy 20 – 30% more sequence space than Lotus counterparts, primarily due to greater numbers of Medicago retrotransposons. Unlike most plants sequenced to date, there is no evidence of lineage-specific duplication within Medicago, but instead, a duplication occurring early in the evolution of legumes after the split with poplar. The authors thank the many participants of the International Medicago Genome Sequencing Consortium for their contribution to this effort.

MUTLI-AGENCY SEQUENCING OF THE SOYBEAN GENOME

PRESENTER: Scott Jackson

The soybean genome is being sequenced through a coordinated effort involving the Department of Energy (DOE), United States Department of Agriculture (USDA) and the National Science Foundation (NSF). A hybrid approach is being used: the Joint Genome Institute (DOE) is doing a 6-8x whole genome shotgun and a sequence tagged physical map is being developed by USDA scientists and an NSF-funded project. The target date for completion of the whole genome shotgun sequence is 2008, but considerable effort will be required to assemble the shotgun sequences and merge them with the physical map to develop a reference sequence map.

SOYBASE AND THE SOYBEAN BREEDER'S TOOLBOX

INVESTIGATORS: Rex T. Nelson, David Grant, Steven Cannon and Randy C. Shoemaker

PERFORMING INSTITUTION:
USDA-ARS CORN INSECT AND CROP GENETICS RESEARCH UNIT,
IOWA STATE UNIVERSITY
AMES, IA 50011

SoyBase, the USDA-ARS soybean genetics database, is a central repository where researchers can quickly find information on most aspects of soybean genetics, metabolism, and pathology. In 1993 SoyBase offered the first publicly available RFLP maps of the soybean genome as well as the first extensive collection of metabolic data in a plant genome database. A major milestone occurred in 2000 when a set of composite genetic maps was released. The current composite genetic maps include data from more than 35 populations and contain more than 4000 mapped classical and molecular loci, along with more than 950 QTL. In 2004 the map drawing engine for SoyBase was changed to CMap. Among other benefits CMap allows the simultaneous display of two or more genetic maps with all features in common clearly indicated. Since the initial SoyBase release, new data types have been continuously added that complement the genetic maps, including internal hyperlinks between mapped loci and metabolism, pathology and many other phenotypic traits. SoyBase now contains contain almost all published data on soybean genetic research.

Recently an updated user interface to a subset of the data in SoyBase, termed the Soybean Breeder's Toolbox (SBT), was developed that specifically aimed to meet the needs of soybean breeders. This relational database currently focuses on the genetic maps and associated trait information and will be expanded to contain the other data types important to soybean breeders. Along with the updated interface, the SBT provides new functionality including:

- A better search function that can return related information from multiple data types
- Data is now easily accessible via externally-initiated searches; this allows a much tighter linkage between the SBT and other databases and web interfaces
- Integration of the Williams82 physical map data with the other data in SoyBase
- Ability to display multiple genetic and physical maps simultaneously using the CMap viewer

Because SoyBase and the Soybean Breeder's Toolbox contain different but complementary data, for now we will maintain both. However, as the Toolbox evolves to include additional data types and functionality, we will shift all remaining data from the older SoyBase interface into the Soybean Breeder's Toolbox.

Future work on the Soybean Breeder's Toolbox will continue expanding marker and phenotype information, and will more closely integrate the soybean physical map and genomic sequence from the public soybean genome sequencing project.

THE LEGUME INFORMATION SYSTEM (LIS): AN INTEGRATED, DYNAMIC COMPARATIVE LEGUME INFORMATION RESOURCE

INVESTIGATORS: Michael D. Gonzales[1], Kamal Gajendran [1], Andrew D. Farmer [1], David Grant [2], Randy Shoemaker [2], William D. Beavis[1], **Gregory D. May**[1]

[1] National Center for Genome Resources, Santa Fe, NM, USA

[2] USDA-ARS-CICGR & Department of Agronomy, Iowa State University, Ames, IA, USA

Comparative genomics is the comparison and analysis of genomes of different species to gain a better understanding of how species have evolved and to determine gene function. Clade-oriented information resources such as LIS offer data and applications enabling comparative genomics approaches that utilize bioinformatics to leverage genomic information from model and reference organisms for the benefit of legume researchers.

LIS (www.comparative-legumes.org/) is a publicly accessible legume resource that integrates molecular and genetic data from phylogenetically diverse legume species enabling cross-species *transcript*, *genomic* and *map* comparisons. The intent of the LIS is to help researchers leverage data-rich model plants to fill knowledge gaps across crop legume species and provide the ability to traverse between interrelated data types.

Transcript data: The LIS "virtual plant" user interface facilitates intuitive navigation of *M. truncatula*, *Lotus*, soybean and Arabidopsis EST and consensus transcript data. Currently, the sequence import functionality at LIS makes use of NCGR's computational pipeline, XGI that uses a variety of algorithms for sequence pattern recognition, comparison and annotation (www.ncgr.org/xgi) and can handle genomic, EST or ORF sequence data types. Pipeline analyses include: BLASTX searches against NCBI non-redundant protein library; BLASTN and TBLASTX searches against related transcript libraries; BLIMPS search against Blocks+ protein motif database; searches with the 12 InterProScan algorithms against the InterPro database; identification of signal peptides for extracellular secretion with PexFinder, an algorithm based on SignalP 2.0; and GenScan for gene prediction in genomic sequences. The automated post-analysis annotation links BLAST and Blocks+ hits to their cognate Gene Ontology entries, and InterPro hits are automatically linked to GO annotations.

Genome data: The XGI genomic pipeline (XGIg) processes LIS genomic data. LIS does not assembly genomics data, but instead uploads data from GenBank as provided by the genome sequencing centers. The Comparative Functional Genomics Browser (CFGB) provides visualization of comparative genomics analysis results including alignment of transcript data with genomic contigs. CFGB also enables dynamic visualization of comparative alignments between genomic contigs through zooming, panning and sorting functions.

Map data: LIS incorporates a CMap-based viewer (www.gmod.org/cmap) that provides users with detailed sequence and annotation viewing that is provided by a custom sequence viewing module developed for LIS. CMap provides LIS users access to, where available, the genetic and physical maps of *Medicago*, soybean and *Phaseolus*. Currently, all SoyBase (http://soybase.agron.iastate.edu/) curated linkage map data have been uploaded and incorporated into CMap. For comparative analyses, a map is selected from the database for use as a reference map. Other maps can be subsequently added so that alignments relative to the selected reference map can be compared.

Future implementations of LIS will utilize semantic web services to traverse between data types allowing users to following phenotype through genetic maps to annotated genome sequence to linkage groups of other species.