National Science Foundation-Sponsored Workshop Report. Draft Plan for Soybean Genomics[1]

Gary Stacey*, Lila Vodkin, Wayne A. Parrott, and Randy C. Shoemaker

National Center for Soybean Biotechnology, Department of Plant Microbiology and Pathology, University of Missouri, Columbia, Missouri 65203 (G.S.); Department of Crop Sciences, University of Illinois, Urbana, Illinois 61801 (L.V.); Department of Crop and Soil Sciences, The University of Georgia, Athens, Georgia 30602 (W.A.P.); and Corn Insect and Crop Genetics Research Unit, United States Department of Agriculture-Agricultural Research Service, Iowa State University, Ames, Iowa 50011 (R.C.S.)

Recent efforts to coordinate and define a research strategy for soybean (*Glycine max*) genomics began with the establishment of a Soybean Genetics Executive Committee, which will serve as a communication focal point between the soybean research community and granting agencies. Secondly, a workshop was held to define a strategy to incorporate existing tools into a framework for advancing soybean genomics research. This workshop identified and ranked research priorities essential to making more informed decisions as to how to proceed with large scale sequencing and other genomics efforts. Most critical among these was the need to finalize a physical map and to obtain a better understanding of genome microstructure. Addressing these research needs will require pilot work on new technologies to demonstrate an ability to discriminate between recently duplicated regions in the soybean genome and pilot projects to analyze an adequate amount of random genome sequence to identify and catalog common repeats. The development of additional markers, reverse genetics tools, and bioinformatics is also necessary. Successful implementation of these goals will require close coordination among various working groups.

The soybean, *Glycine max* (L.) Merr., is a major source of protein and vegetable oil for animal and human nutrition. The availability of numerous genomic advancements in soybean has set the stage for a coordinated effort to further soybean genomics and its applications. Realizing the resource-intensive and multidisciplinary requirements of genomics research require identification of priorities, a broad agreement on a clear research strategy, and coordination among research groups, this draft plan describes current and ongoing efforts to move soybean genomics research forward.

As a crucial first step, a Soybean Genetics Executive Committee (SoyGEC) with elected members was established during the summer of 2003 to serve as a communication focal point for the soybean research community. Information on SoyGEC members and actions may be found from a link on the SoyBase Web site (http://129.186.26.94/). SoyGEC will proactively communicate with the soybean research community to help define research priorities and with representatives of federal granting agencies to ensure that research priorities are clearly articulated. Without encumbering individual initiatives, SoyGEC will encourage coordination of dedicated research teams finding solutions to soybean problems of national and international importance.

To further advance soybean genomics, a National Science Foundation-sponsored workshop was held in St. Louis on October 21, 2003, to take an inventory of the current genomic resources in soybean, identify areas where more preliminary data are still necessary, and identify a research strategy to further soybean genomics research. Special attention was focused on research opportunities provided by unique aspects of soybean biology.

The workshop included academic, governmental, and industrial scientists covering a wide variety of specialties related to both basic and applied research on soybean, along with scientists from outside the soybean field who provided general expertise in genomics and represented a wealth of experience garnered from other genomic projects. Representatives from federal funding agencies and soybean commodity groups observed the meeting.

This workshop extended and further defined the findings from earlier workshops, which had surveyed the status and priority goals for developing resources for soybean and legume genomics (http://129.186.26.94/Genetic_Resources/Soybean_Genetic_Resources.html; and http://129.186.26.94/Legume_Initiative/LegGenomicsPaper10Oct01.html). The development of resources for legume genomics, including soybean, was also a topic discussed during the Workshop on the National Plant Genome Initiative: 2003–2008 (http://books.nap.edu/catalog/10562.html). Although these previous meetings reached a consensus concerning the tools and resources needed to further soybean and legume genomics, they did not do this within the

[1] The workshop was sponsored by grant DBI–0344641 from the National Science Foundation.

* Corresponding author; e-mail staceyg@missouri.edu.

www.plantphysiol.org/cgi/doi/10.1104/pp.103.037903.

context of identified research opportunities in soybean biology.

The October 2003 workshop took this latter effort as its central focus, as illustrated by the title of the workshop: Genomic Perspectives of Soybean Biology. The report that follows presents the major recommendations for a coordinated approach to soybean genomics and a short rationale for these decisions.

UNIQUE BIOLOGICAL OPPORTUNITIES IN SOYBEAN

The soybean is a member of the tribe Phaseoleae, the most economically important of the legume tribes. Other legumes within the tribe include pigeon pea, common bean, lima bean, tepary bean, winged bean, cowpea, mung bean, black gram, adzuki bean, and Bambarra groundnut (Hymowitz, 2004). The extensive genetic resources of soybean and the associated physiological tools available for soybean present a set of unique opportunities to study everything from seed development to the biology of polyploidization to a huge array of pathogenic and symbiotic plant-host interactions. The large size of the soybean plant is an advantage for such studies, permitting the use of techniques not easy or possible with smaller plants. Examples include reciprocal grafting (e.g. Vuong and Hartman, 2003) and stem injections (e.g. Abdin et al., 1998) to facilitate physiological and metabolic studies.

The genus Glycine is paleopolyploid, with $2n = 40$ as its base chromosome number, as compared with other phaseoloid legumes which are largely $2n = 20$ or 22 (Goldblatt, 1981). There are 22 recognized perennial species within the genus, of which *Glycine tabacina* and *Glycine tomentella* are neopolyploid ($2n = 78$, 80; Hymowitz, 2004). These polyploids are of very recent origin (Doyle et al., 2002), presenting the opportunity to study the genomic response to both ancient and recent polyploidy.

There are also two annual species, *Glycine soja* and *G. max*, which are highly self-pollinated and thus exist as inbred lines. Both of these are perfectly cross compatible, effectively constituting a single species (Hymowitz, 2004).

The USDA soybean germplasm collection possesses a very broad range of phenotypic diversity. Figure 1 illustrates the tremendous diversity that exists for seed size, color, and shape. During domestication, seed weight has increased from the 0.5 to 2.5 g/100 seed found in most soja accessions (Dong et al., 1999), to 10 to 20 g/100 seed, and accessions exist with seed weights as high as 50 g/100 seed (Hymowitz, 2004).

The soybean seed is unique in its accumulation of both high levels of protein and oil, which presents several opportunities for study. A typical soybean seed is 40% protein and 20% oil by weight (Fehr, 1987). The fact that soybean is an oil seed plant is one important feature that distinguishes it from the proposed model legumes (see below).

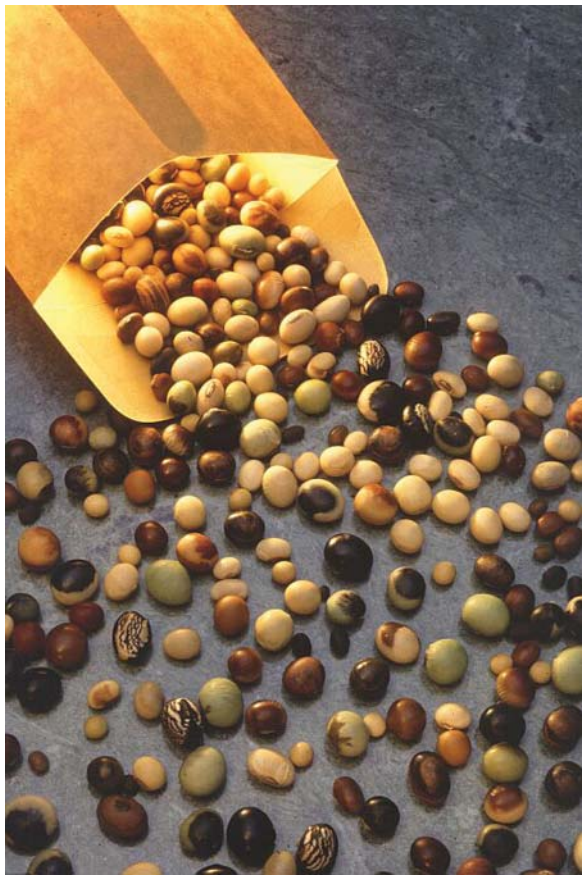

Figure 1. Soybean seed diversity. Photo courtesy of USDA-ARS.

The somatic embryo methodology that exists for soybean (Parrott and Clemente, 2004) is among the most advanced embryogenic systems available for any dicot and can be used to help study seed physiology and development. Furthermore, transgenic somatic embryos may be used to efficiently study seed genomics, by either overexpressing or suppressing embryo-specific genes, without the need to recover an entire plant (Kinney, 1998).

Genetic diversity in soybean is not limited to the seed. For example, Dzikowski (1936) documented at least 22 branching patterns and 6 different leaf morphological types. Furthermore, both determinate and indeterminate genotypes exist, as well as various types that differ in their photoperiodicity. Current soybean varieties are placed in 1 of 13 maturity groups, depending on their photoperiodicity. Palmer et al. (2004) list a soybean mutant that cannot use fluorescent light to detect photoperiod.

The Soybean Genetic Type Collection maintains over 300 phenotypic mutants (Palmer et al., 2004). While most affect morphological and reproductive traits, there are genes that affect nodulation, specify compatibility with different rhizobial genera, and condition for resistance or susceptibility to viral, nematode, fungal, and bacterial diseases.

Additional genes affect the differential ability to use various minerals, the production of fluorescent compounds in the roots, and the production of various flavonol glycosides. Additional mutants affect chlorophyll and other pigments or confer tolerance or sensitivity to various herbicides. In addition, there is an extensive isozyme series and a collection of oil and storage protein variants, some of which are leading to marketable advances in soybean improvement (e.g. modified oils, low phytates). Perhaps most importantly, there is a wealth of mutants in key metabolic enzymes (Palmer et al., 2004), which has long made soybean one of the plants most favored for metabolic studies (e.g. World Soybean Conference, 1999). Finally, the agronomic knowledge of soybean far exceeds that of any model legume. Many of these agronomic traits have now been integrated with the genetic and physical maps via quantitative trait loci analyses.

At last count, there are 552 near isogenic lines for morphological, pigment, and disease resistance traits. In general, soybean varieties have well-documented pedigrees, which facilitates the study of genetic diversity (Carter et al., 2004).

CURRENT STATUS OF SOYBEAN GENOMICS

Due to its tremendous agronomic importance and large research community, significant development of genetic, molecular, and genomic tools in soybean has occurred.

1. In spite of its polyploid origin, the soybean composite genetic map is well developed. The classical map contains only 67 loci on 19 linkage groups (Hedges and Palmer, 1993), while the molecular map encompasses all 20 linkage groups and over 2,300 cM based on simple sequence repeats (SSRs) and restriction fragment length polymorphisms (RFLPs) mapped in multiple populations (SoyBase; http://129.186.26.94/). The composite map contains 1,845 markers (1,010 SSRs, 718 RFLPs, 73 random amplified polymorphic DNA (RAPDs), 23 classical traits, and 10 others; P. Cregan, unpublished data). In addition, 156 classical markers have been incorporated into the molecular map (J. Specht, personal communication).
2. Several mapping populations are available. A few examples include: University of Utah populations from crosses of cv Minsoy × Noir I, Minsoy × Archer, and Noir I by Archer, consisting of 240, 233, and 240 F7-derived recombinant inbred lines (RILs), respectively (Mansur et al., 1996); a University of Nebraska population derived from Clark × Harosoy (Shoemaker and Specht, 1995); a USDA/Iowa State University population, an F2-derived mapping population from A81-356022 (*G. max*) × PI468.916 (*G. soja*), with 59 F2 plant derivatives (Shoemaker and Olson, 1993; Shoemaker and Specht, 1995; Keim et al., 1996); and a BSR 101 × PI 437654 population with 327 RILs, to name but a few. Sleper et al. (unpublished data) are currently developing a mapping population of over 600 RILs from a Forrest × Williams 82 cross. J. Specht and P. Cregan (unpublished data) recently developed several other mapping populations, which are segregating for northern versus southern adaptation, high yield, and low and high protein content.
3. A number of public sector bacterial artificial chromosome (BAC) libraries are available providing >35-fold genome coverage (e.g. Marek and Shoemaker, 1997; Danesh et al., 1998; Tomkins et al., 1999; Meksem et al., 2000).
4. The existing double-digest-based physical map for the soybean cv Forrest (http://www.siu.edu/pbgc/, GMOD; http://131.230.90.184; and http://www.langin.com/soybean/gbrowse.) is incomplete. At >3,000 contigs, the combined lengths of all assembled contigs exceed the size of the genome.
5. Over 300,000 soybean expressed sequence tag (EST) sequences are in GenBank (http://129.186.26.94/soybeanest.html). Over 60% of these sequences were derived from cultivar Williams 82, which has been adopted as the model cultivar by the community. These ESTs are derived from >80 unique cDNA libraries from a number of organs, genotypes, stages of development, environmental conditions, and biotic and abiotic stresses.
6. With National Science Foundation funding, Dr. Lila Vodkin's laboratory (University of Illinois) has utilized these ESTs to develop DNA microarrays for functional genomic analysis (Clough et al., 2000; see below).
7. Soybean transformation and mutagenesis: Soybean transformation efficiencies are consistently >5%, and frequencies >12% are now common in many academic and industrial laboratories (Somers et al., 2003). These improvements in transformation efficiencies have led to the initiation of transposon tagging projects for soybean (Clemente and Stacey, personal communication). Additionally, viral-induced gene silencing systems and TILLING populations for soybean are under development (A. Bent, University of Wisconsin; N. Nielson, Purdue University; K. Meksem, Southern Illinois University, personal communication).
8. The USDA/ARS germplasm collection at the University of Illinois houses over 16,000 plant introduction lines from China and other parts of the world. Another collection of southern maturity groups exists at the USDA/ARS facility in Stoneville, MS.
9. Nearly 300 researchers routinely attend the Biennial Cellular and Molecular Biology of Soybean Conferences (e.g. http://muconf.missouri.edu/soy2004/index.html), which have been held over the past two decades. Every four years, a World Soybean Research conference draws several thousand participants from many nations (e.g. http://www.cnpso.embrapa.br/soy/). Such an

interdisciplinary community, together with a vibrant soybean industry, can transfer genome science to practical application effectively and efficiently.

WHAT IS THE NATURE AND ORGANIZATION OF THE SOYBEAN GENOME?

The soybean genome comprises about 1.1 Mb/C-value (Arumuganathan and Earle, 1991). This makes it about seven and one-half times larger than the genome of Arabidopsis but less than one-half the size of the maize genome (Arumuganathan and Earle, 1991). A complete karyotype of soybean has been reported (Singh and Hymowitz, 1988) based upon pachytene analysis (http://www.cropsci.uiuc.edu/faculty/hymowitz/genlab/karyo.html). More than 35% of the genome is made up of heterochromatin, with the short arm of 6 of the 20 bivalents being completely heterochromatic (Singh and Hymowitz, 1988).

Due to its polyploid history, many examples of duplicate factor genes (2 independent genes controlling the same trait) can be found in soybean (Palmer and Kilen, 1987). Hybridization to soybean genomic DNA by each of 280 randomly chosen *Pst*I genomic clones determined that more than 90% of the probes detected more than two fragments and nearly 60% detected 3 or more fragments (Shoemaker et al., 1996). This suggests that less than 10% of the genome may be single copy sequence and that large amounts of the genome may have undergone genome duplication in addition to the tetraploidization event. These authors also observed nested duplications that suggested that at least 1 of the original genomes of soybean may have undergone an additional round of tetraploidization in the far distant past (Shoemaker et al., 1996). An analysis of synonymous substitutions between pairs of duplicated genes identified probable large-scale duplication events at approximately 15 million years ago (mya), and 40 mya (Schlueter et al., 2004). These duplications followed by diploidization have given soybean a relatively heterogeneous genome intermixed with ancient and recent duplications.

Early DNA-DNA renaturation studies suggested that approximately 40% to 60% of the soybean genome sequence is repetitive (Goldberg, 1978; Gurley et al., 1979), and varies greatly in complexity and reiteration frequency (Gurley et al., 1979). The studies also suggested that the repetitive sequences were not evenly interspersed among single-copy sequences and clustering appeared to be similar to that found in wheat (*Triticum aestivum*) and rye (*Secale cereale*; Flavell and Smith, 1976; Smith and Flavell, 1977). Analysis of data generated from BAC end sequences suggests that although some regions are predominantly gene-rich or repetitive-rich, most loci contain a mixture of repetitive and genic sequences (Marek et al., 2001).

Perhaps because of the recent domestication of soybean and/or as a result of a very limited number of domestication events, sequence variation is relatively limited. To quantify sequence variability, approximately 28.7 kb of coding sequence, 37.9 kb of noncoding perigenic DNA, and 9.7 kb of random noncoding genomic DNA were sequenced in each of 25 diverse soybean genotypes, showing 0.5 and 4.7 single nucleotide polymorphisms (SNPs)/kb in coding and noncoding DNA, respectively (Zhu et al., 2003). Over the 76 kb sequenced, mean nucleotide diversity, expressed as Watterson's θ, was 0.00097. Nucleotide diversity was 0.00053 and 0.00114 in coding and in noncoding perigenic DNA, respectively. By comparison, maize (*Zea mays*) has as much as 10-fold higher levels of nucleotide diversity (Tenaillon et al., 2001; Ching et al., 2002). Similarly, sequence variation in individual Arabidopsis genes (Kawabe and Miyashita, 1999; Purugganan and Suddith, 1999; Kawabe et al., 2000; Kuittinen and Aguade, 2000) has levels of nucleotide diversity 5- to 8-fold higher than that of domesticated soybean. Haplotype analysis of SNP-containing gene and genomic fragments revealed a severe deficiency of haplotypes versus the number that would be anticipated at linkage equilibrium. In 49 fragments with 3 or more SNPs, there was an average of only 3.3 haplotypes among the 25 cultivated soybean genotypes analyzed (Zhu et al., 2003).

Little is published about specific soybean repetitive sequences. Vahedian et al. (1995) identified a repeat with a size of 92 bp (SB92). Fluorescent in situ hybridization (FISH) analysis using this sequence showed that it clustered in four or five chromosomal locations. Two of those locations were centromeric. This nonrandomly distributed repetitive element was estimated to be homologous to about 0.9% of the genome (approximately 1.1×10^5 bp). Given the high copy number of this repeat, and the relatively few genomic locations, the element arrays must exist as megabase-sized regions. These findings also suggest that to develop a comprehensive database of soybean repeat sequences, it will be necessary to draw upon data from a very large number of locations throughout the genome. Another family of repetitive sequence, STR120, is comprised of an approximately 120-bp monomer. This family has a lower copy number and is estimated to consist of 5,000 to 10,000 copies (Morgante et al., 1997). Unpublished analysis of BAC-end sequences and genomic sequence deposited at the National Center for Biotechnology Information indicate that there are probably more than 100 other classes of repetitive DNAs in soybean (L. Marek, J. Mudge, N. Young, and R. Shoemaker, unpublished data).

Transposable DNAs also comprise the repetitive DNA in the genome. The first transposable element discovered in soybean was *Tgm* (Vodkin et al., 1983), and consists of 7 different classes ranging in size from 1.6 kb to more than 12 kb (Rhodes and Vodkin, 1988). The copy number of this element is difficult to determine because of the variation in structure and organization of each class type. A mariner-like element

(*Soymar*1) also was reported by Jarvik and Lark (1998). The largest member of this family of elements is about 3.5 kb. Copy number of this element may go from a few copies to as high as 10,000 copies per haploid genome in some phyla (Jarvik and Lark, 1998). A family of a *copia/Ty*1-like retroelement (SIRE-1) has also been characterized (Laten et al., 1998). This family consists of only a few hundred members and each member is approximately 11 kb.

In grasses, the majority of genes may reside in only a small fraction (10%–20%) of the total genome. In the model dicot, Arabidopsis, the average gene density is approximately 20 to 25/100 kb, but the genes are relatively evenly dispersed across the genome (Barakat et al., 1998). The global gene space of soybean is not yet defined. However, gene-rich regions have been observed in soybean (Marek et al., 2001) using hypomethylated RFLPs as signatures of genic regions as compared to SSRs. Of more than 2,000 BAC-end sequences examined, RFLP-associated sequences had only one-half as many repetitive sequences and 50% more genic sequences compared with SSR-associated sequences. Based on one segment of soybean sequence more than 330 kb in length, Foster-Hartnett et al. (2002) estimated gene density to be as high as 1 gene/5 kb (though revised calculations put this value closer to 1 gene/8 kb). All these estimates suggest an uneven gene density, with some regions containing higher densities. However, gene space may be limited to a small percent of the entire genome. A recent study identified BACs using *Pst*I-generated RFLP probes. An evaluation of the number of redundant BACs suggested that the probes identified BACs from only 24% of the genome (Mudge et al., 2004). *Pst*I is a methylation sensitive enzyme and, therefore, regions cut with this enzyme are presumed to be gene rich. If this is true, then the soybean gene space may be limited to one-quarter or less of the entire genome.

FUTURE STRATEGIES FOR SOYBEAN GENOMICS

What Genotype Should Be the Standard?

The workshop participants were in unanimous agreement that the Williams 82 cultivar should be adopted as the standard for genomic studies. Williams 82 ESTs represent >60% of the entries in the public EST collection (http://129.186.26.94/soybeanest.html), which is currently being mined for development of cDNA and oligonucleotide microarrays. Therefore, subsequent functional genomics studies will likely make most use of this cultivar. Any future genomic sequencing efforts would also be aided by the ability to make full use of the current EST resource. Williams 82 (representing northern, indeterminate varieties) and Forrest (representing southern, determinate varieties) were independently developed, each from a unique subset of approximately 15 ancestral cultivars (Gizlice et al., 1996). Analysis of SSR, RFLP, and SNP markers indicate that these breeding efforts have captured approximately 70% of the genomic diversity of soybean (Keim et al., 1992; P. Cregan, personal communication). As mentioned above, the current Forrest physical map, based on polyacrylamide gel BAC fingerprints, is incomplete. A strong recommendation of the workshop was to use improved methods for BAC fingerprinting to generate a physical map of cultivar Williams 82. The availability of the Williams 82 map, in addition to the current Forrest data, will provide an excellent resource for comparative genomic studies of the two major breeding lineages of soybean. Another useful resource will be the 600 RIL lines from a Forrest × Williams 82 cross that will soon be available for genetic mapping (Sleper et al., unpublished data). A set of approximately 1,000 BAC contigs already exists from Williams 82 that will provide mapped anchor points for the current genomic efforts.

What Is the Strategy for Advancing Soybean Structural Genomics?

Although a significant amount of information has already been gathered on the organization and structure of the soybean genome, workshop participants recommended that several key efforts would add greatly to our knowledge. The genetic map of soybean is fairly well populated with markers of different types, including RFLP and SSRs. Development of SNP markers is only beginning. Current mapping data has been compiled from numerous populations from across several maturity groups. To take full advantage of genetic and physical mapping data, these data need to be acquired in mapping populations with well-defined and catalogued phenotypes. The maps need to be expanded by the development of 1,000 to 2,000 more sequence-based markers, thus bringing the total number of molecular markers on the soybean map to nearly 4,000.

It has become clear that no single genetic system can provide all the answers, especially in legumes, which as one of the largest families of flowering plants, has a rich assortment of diverse traits and genetic diversity. There are also numerous commonalities. A soybean genetic map of sequence-based markers that could be applied to other legumes would facilitate the translation of genetic information from one species to another. As few as 150 markers would provide the genetic framework to connect genomic regions among select members of the family. It is encouraged that new and novel approaches be developed that would facilitate this effort. One such approach is HAPPY (HAPloid equivalents of DNA and the PolYmerase chain reaction) mapping (Dear and Cook, 1989, 1993). As is the case in radiation hybrid mapping, the screening of a HAPPY panel with sequenced tagged sites (STSs) will provide an estimate of physical distance between STSs. A recent report of HAPPY mapping in Arabidopsis demonstrated excellent co-

linearity of STSs positioned via HAPPY mapping and their actual position in genomic sequence (Thangavelu et al., 2003). An important attribute of HAPPY mapping is that STSs can be mapped without the need for DNA polymorphisms, making it more likely that a reasonably large set of cross-legume PCR-based markers can be developed and applied for the purpose of identifying homologous genomic regions.

Future genomic research efforts will require that the organization of the genome is known in more detail than the community currently possesses. Present estimates of genic and repeat content and distribution is based upon sequencing of BAC ends from thousands of BACs anchored at RFLP and SSR loci. These data may present a biased estimate by sampling only regions of the genome included in the genetic map. SSRs and RFLPs appear to be randomly interspersed on the genetic map, but it has been estimated that RFLP markers may be located in only about 24% of the genome (Mudge et al., 2004). A better estimate of genic and repeat content of the genome should be obtained by sequencing a few hundred BACs either cloned by random shearing or BACs identified by hybridization to genes. These sequences could then be compared to those obtained from BACs that are RFLP and SSR-based. A database of repeat sequences could also be obtained through single-pass sequencing of about 25,000 randomly generated plasmid clones.

A gold standard sequence-ready physical map from Williams 82 has been identified as a prerequisite for many advanced genomic studies. This effort will entail fingerprinting approaches that capture the maximum information possible in each lane (e.g. HICF or SNAPshot; Ding et al., 2001; Luo et al., 2003), and must be conducted in concert with a community-driven effort to resolve contig ambiguities associated with duplicated segments of the genome and to anchor the physical map to the genetic map. Generation of BAC-end sequences from the BACs making up the physical map will provide for the placement of hundreds of ESTs, will facilitate the overlaying of the soybean physical map with the *Medicago truncatula* physical map, will aid in the comparative alignment of the soybean physical map with Arabidopsis sequence, and will provide much needed sequence for the development of molecular markers. Global distribution of sequences (repeat and genic) should be further resolved through FISH. It is recommended that both the gene space and repeat sequence space be better described by FISH analysis using known repeat sequences and selected BACs.

Placement of 2,000 to 3,000 cDNA sequences onto the physical map either through overgo technology (compare with Gardiner et al., 2004) or other approaches will provide additional information on distribution of genic sequences in soybean. As a result of the rounds of genome duplication that have occurred in soybean, most genic sequences are represented an average of 2.9 times in the genome. These sequences are often highly identical (Schlueter et al., 2004). Thus, only a small percentage of these duplicated sequences may be distinguishable through standard hybridization methods (e.g. using overgo probes). The others will associate by hybridization with BACs from several locations in the genome and will require additional sequence-based resolution. This further resolution should be facilitated by full-length sequencing of cDNAs representing the individual members of the family contig.

What Is the Strategy for Soybean Functional Genomics?

Development and Use of Global Expression Resources for Soybean

The Public EST Project for Soybean (Shoemaker et al., 2002) funded by soybean grower associations led to the creation of over 80 cDNA libraries that represent genes expressed in many different tissue and organ systems of soybean, developmental stages, and pathogen and stress-challenged plants. Over 280,000 5′ sequences of individual cDNA clones resulted and they represent potentially 61,000 unique cDNAs as determined by sequence cluster analysis. Starting from this resource, the major goals of the Functional Genomics Program for Soybean funded by the National Science Foundation were to physically assemble a set of 36,000 of these unique cDNA clones into a low redundancy unigene set, to sequence them at the 3′ end, and to develop and initially test the soybean cDNA microarrays. Currently, 9,216 soybean cDNAs are printed on each of 3 glass microarray slides. Set 1 (sequence-driven reracked library Gm-r1070, 9,216 cDNAs) is enriched for genes expressed in the developing flowers and buds, young pods, developing seed coats, and immature cotyledons. Set 2 (reracked libraries Gm-r1021 + Gm-r1083) is enriched for genes in the roots of seedlings and adult plants, including those infected with *Bradyrhizobium japonicum*. Set 3 (reracked library Gm-c1088, 9,216 cDNAs) is primarily from cDNAs isolated from germinating cotyledons, germinating seedlings under various stresses, and leaves of 2-week-old plants including some under challenge by pathogens. A fourth set of 9,216 cDNAs is currently being assembled from cDNA libraries derived from soybean tissue culture embryos and from many other stress- and pathogen-challenged tissues.

The applications of microarray technology to soybean are enormous. A few include profiling the genes that respond to challenges by various pathogens (Clough and Vodkin, 2004) and by environmental stresses such as drought, heat, cold, flooding, and herbicide application. In addition, expression profiling of isolines that differ in protein or oil content or other quantitative traits will yield significant clues to the genes involved in those pathways and traits. One recent study was a detailed analysis of induction of somatic embryos during culture of cotyledons on auxin-containing medium (Thibaud-Nissen et al.,

2003). These transcript profiles were subjected to a cluster analysis using a k-means test and revealed the process of reprogramming of the cotyledon tissues during the induction process. For example, the data illustrate that auxin induces dedifferentiation of the cotyledon and provokes a surge of cDNAs involved in cell division and oxidative burst. The data also indicate that the formation of somatic globular-stage embryos is accompanied by the accumulation of storage protein transcripts and transcripts for the synthesis of gibberellic acid.

Recommendations for Functional Genomics in Soybean

The next generation of soybean microarrays will be based on synthetic oligonucleotides as opposed to cDNAs amplified by PCR from plasmid templates. The 3′ sequence data of the unigenes is particularly useful to design and synthesize long oligos of approximately 70 bases from each of the unigenes. Creation of up to 50,000 long oligos is a goal that would move the soybean microarray technology to the next phase. Efforts are under way to achieve this goal during 2004. Where possible, oligo arrays that will distinguish gene family members (orthologs or homeoelogs and paralogs) would be very useful. Affymetrix technology, which uses multiple 20-mers, is another platform that will likely be available for soybean during 2004.

A major goal for functional genomics in soybean is to develop and utilize all of the global scale technologies—transcriptomics, proteomics, and metabolomics. Since the developing seed is the source of protein, oil, and secondary metabolites, it is logical that an emphasis on the developing seed would be an excellent starting point to relate transcript profiles to protein and metabolite profiles. Much basic information would be gleaned that would have many direct applications. Metabolite profiling has begun to catalog the varied compounds produced by soybean, including the flavonoids and isoflavones that are abundant in the seed. After proof of concept to relate transcriptome, proteome, and metabolome in a standard variety such as Williams 82, experiments could then be broadened to include the analysis of the effect temperature and other stresses on seed development. Somatic embryos are useful to understand processes involved in the early stages of embryo development and are easier to obtain than early stage (globular and heart) zygotic embryos, which require microscopic dissection.

Reverse Genetic Tools

The status of genetic transformation in soybean was recently reviewed by Parrott and Clemente (2004). While different transformation systems are available which provide flexibility of transformation depending on the application, the efficiency of these does not approach the in planta transformation system available for Arabidopsis. Therefore, the development of transient transformation systems for soybean would facilitate genomics applications rapidly, simply, and economically. The development of the following tools was recommended by the workshop participants:

1. Virus-induced gene silencing. The use of virus-induced gene silencing has become well established in some model species, and the technology is developed to a point where the basic principles can be applied to other species (Lu et al., 2003). Overall, gene silencing by RNA interference is proving to be particularly useful in plants with genetic redundancy (Lawrence and Pikaard, 2003), as is the case with soybean.
2. Gene overexpression with transient systems. Viral vectors can be used for overexpression of genes (e.g. Arazi et al., 2001; Escobar et al., 2003), making it possible to complement mutants or determine the function of overexpressed genes.
3. Retrotransposon tagging. The efficiency of current transformation systems will probably limit the efficiency of T-DNA mutagenesis in soybean. Nevertheless, systems based on retrotransposon mutagenesis may prove useful, as once inserted, they can be activated many times during regeneration from tissue culture, as has been the case with the *Tos17* retrotransposon in rice (*Oryza sativa*; Hirochika, 2001) and the *Tnt1* retrotransposon of tobacco in *M. truncatula* (D'Erfurth et al., 2003).
4. Further development of TILLING (McCallum et al., 2000) as a resource for soybean. Proof of concept has been provided in soybean (Slade et al., 2003), and additional tilled lines are under development at Purdue and Southern Illinois University. Specifically, there are three populations, one of which is in Williams 82. Two of the populations have between 5,000 and 6,000 entries each; the goal for the Williams 82 population is to have 8,500 entries (N. Nielson, personal communication). Another Tilling project is in progress at Southern Illinois University (K. Meksem, personal communication).

How Should Soybean Genome Data Be Disseminated?

SoyBase (http://129.186.26.94/) has been a repository for soybean genetic data for more than a decade and continues to be a useful breeding tool. It was originally conceived as a repository where researchers could quickly find information on most aspects of soybean genetics, metabolism, and pathology. Genetic mapping data remain at the central core of SoyBase, with a rich collection of genetic maps. These maps contain more than 3,800 mapped classical and molecular (RFLP, AFLP, RAPD, PCR, and SSR) loci and more than 950 quantitative trait loci. SoyBase also possesses an extensive collection of metabolic data, with more than 900 individual enzymes and pathways interactively displayed. Information is available on 90

soybean diseases, including causative organism, symptoms, differentials, and resistance mechanisms. Additional topics include insect pests, nodulin, storage protein, sequence, miscellaneous protein, colleague, nodulation, and transformation. Furthermore, SoyBase contains descriptions for more than 3,800 germplasm accessions.

Genomics projects have begun to generate data types not easily handled by the object-oriented SoyBase and at a rate much faster than anticipated years ago. To overcome this obstacle a cooperative agreement between USDA-ARS and the National Center for Genomic Research (Santa Fe, NM) was established. This collaboration has resulted in development of a Legume Information System (LIS, http://www.comparative-legumes.org/) that will acquire, store, sort, and visualize genetic/physical data from all legumes. To make this more broadly applicable, a mechanism for annotation of sequence data and a close linkage with gene ontology and trait ontology working groups is necessary. In addition to sequence-based and map-based interfaces it will be necessary to incorporate data handling and visualization tools for transcript and metabolic profiling data. LIS provides a seamless link with SoyBase but also provides the relational structure needed to integrate sequence, genetic, and physical map data among all legumes. It is recommended that not only the soybean community, but other legume communities support further development of LIS and integrative research with other legume database developers.

Linkage to the agricultural community is the logical next step in plant genomics. It becomes more and more imperative to facilitate community communication and outreach to ensure that the benefits of genomic research are dispersed to the broadest community. It is encouraged that legume database efforts begin in earnest to reach out to breeders through development of bioinformatic breeding tools, to educators through development of teaching and training modules, and to the nonscientific community through better communications about societal benefits of genomic research.

Relationship of Soybean to Model Plants

Since legumes far outstrip other plant families in total diversity (Doyle and Luckow, 2003), it is impossible to focus on only a few legume species that would serve as omniscient models for all legumes. However, recently, attention has focused on two model legumes, *M. truncatula* and *Lotus japonicus* (for review, see Young et al., 2003). Genome sequencing efforts are currently under way in both of these species. These two species, along with soybean, are members of the papilionoid subfamily, the group that contains nearly all crop legumes (Doyle and Luckow, 2003). Medicago and Lotus fall within the mainly temperate Hologalegina lineage, as contrasted with soybean, which is a member of the generally tropical phaseoloid group. The Hologalegina and phaseoloid lineages are each other's closest relatives. They are the two largest groups of papilionoid legumes in terms of genera and species, but because of their close relationship to one another they are not representative of many other lineages of legumes, both within the papilionoid subfamily and in the two other nonpapilionoid subfamilies (Mimosoideae and Caesalpinioideae).

The fossil record for legumes indicates a rise in importance approximately 35 to 54 mya (Doyle and Luckow, 2003). Indeed, all the major lineages of legumes appear to have diverged approximately 40 to 50 mya. For example, molecular estimates suggest the divergence of *M. truncatula* from *L. japonicus* and *M. truncatula* from soybean occurred roughly 40 mya. It should be noted that this is roughly when the major grass species (e.g. maize and rice) are thought to have diverged (Devos and Gale, 2000).

Syntenic relationships exist among legumes and other angiosperms (Grant et al., 2000; Lee et al., 2001). However, genome duplications, insertions, deletions, and sequence divergence call into question how predictive the knowledge of one genome will be to the exploitation of another. Foster-Hartnett et al. (2002) focused on the sequenced region (around *rhg1*) on soybean linkage group G to examine gene distribution and microsynteny. The region was expanded beyond the sequenced segment by isolation of overlapping BAC clones. There was considerable sequence similarity between the linkage group G and Arabidopsis genome sequences. However, the linkage group G sequence exhibited homology to five different Arabidopsis chromosomes, suggesting extensive chromosomal rearrangements differentiate soybean and Arabidopsis. Comparisons between the soybean regions homoeologous to the *rhg1* locus indicated extensive microsynteny, consistent with the recent duplications in the soybean genome. Yan et al. (2003) utilized BAC hybridizations to estimate microsynteny between soybean, *M. truncatula,* and Arabidopsis. These data gave an estimate of only 6% microsynteny between soybean and Arabidopsis (3 of 50 BACs) and approximately 50% between *M. truncatula* and soybean (27 of 50 BACs). However, there are clearly examples of where comparative mapping and the use of markers from other legume species can aid gene identification in soybean (e.g. Searle et al., 2003).

Both *L. japonicus* and *M. truncatula* have genome sizes estimated at approximately 470 Mbp (Young et al., 2003), which is roughly one-half the size of soybean. In addition, as discussed above, soybean is paleoploid, which would appear to distinguish it from the model legumes. However, all legumes likely have a polyploid ancestry (Doyle and Luckow, 2003).

An important research focus using the model legume species is nodulation, which clearly distinguishes legumes from other plants (Young et al., 2003). Nodulation (both via rhizobia and actinomycetes) appears to have arisen exclusively within the eurosid I clade, which includes legumes (Doyle and Luckow,

2003). However, within this clade, nodulation appears to have arisen independently among and within particular families. Although the entire papilionoid group shares nodules of common ancestry, the development and chemistry of nodules varies greatly among different papilionoid groups. For example, both *L. japonicus* and soybean form determinate desmodioid nodules, which are easily distinguished from the indeterminate (with a persistent, apical meristem) nodules formed on *M. truncatula.* However, phylogenetic evidence suggests that Lotus and Medicago are more closely related to one another than Lotus is to Glycine. Consistent with these relationships, Sprent (2001) presented data to suggest that, although structurally similar, determinate nodulation likely evolved independently in the phaseolid group (i.e. soybean) and Loteae. Thus, even within traits that appear morphologically similar, evolutionary history may indicate that differences are likely to be found. A comparative genomics approach, provided by knowledge of multiple legume species, will allow such differences to be identified.

Another important future focus in the comparative genomics of legume models and crop species will be the evolution of the ureide pathway in the tropical legume species. The nodules of tropical legumes export ureides, but the synthesis of these compounds is poorly understood. Because of their greater N:C ratio, ureides are a more C-efficient vehicle for the internal transport of organic nitrogen than are Asn or Gln. Most of the enzymes of de novo purine synthesis have recently been isolated and purified. An unexpected finding was that, unlike higher animals where multifunctional enzymes are involved in the pathway, each step is catalyzed by a separate enzyme in nodules (Smith and Atkins, 2002). This situation resembles that found in prokaryotes. The expression of these enzymes in nodule tissue from soybean is many times greater than in any other plant tissue examined and is clearly related to the large synthesis of ureides. The rate of ureide synthesis has a dramatic, positive effect on nitrogenase activity. Despite the significance of ureides in the N economy of this group of legumes, reasons for the expression of such a complex assimilatory mechanism (compared to the synthesis of amides in nodules of temperate legumes) for fixed N remain elusive.

Specifics of Initial Plan

It is clear from the workshop discussions that much is to be learned from further investments in soybean genomics. The field is currently hampered by the relatively poor description that the community has of the genome. Therefore, initial priority should be placed on providing a more thorough description of soybean genome structure (e.g. distribution of genic and nongenic sequences, repeat structure, etc.). To this end, an initial effort to achieve these goals would be:

1. Construct a high information content fingerprint physical map of the Williams 82 genome.
2. Place 2,000 to 3,000 cDNA sequences onto the physical map to provide information on the distribution of genic sequences in soybean.
3. Obtain a better view of the soybean gene space and repeat sequence space using FISH analysis of known repeat sequences and selected BACs.
4. Sequence approximately 300 BAC clones; 100 chosen randomly to give an indication of repeat structure, 100 chosen from gene-rich regions to examine gene structure, and 100 chosen from large BAC contigs identified by physical mapping.
5. Sequence approximately 25,000 random soybean clones to generate a repeat database.
6. Roughly double the number of markers on the composite soybean genetic map to a total of 4,000.
7. Develop genetic markers (approximately 150) that can be utilized for comparative mapping in multiple legume species.
8. Develop an oligonucleotide microarray, which will distinguish within multi-gene families, to support functional genomic studies.
9. Continued development of reverse genetics tools.
10. Further develop the Legume Information System as a comprehensive soybean database.

This initial plan, achieved over a 2- to 3-year period, would reveal much about soybean genome structure/function and set the stage for a subsequent genome sequencing effort that might, for example, focus on the gene-rich regions of the soybean genome. Moreover, investments in soybean structural genomics would support other priority areas identified during the workshop (e.g. development of reverse genetic approaches and proteomics).

Outreach and Broader Context

The ultimate beneficiaries of a soybean genomics program are the consumers. Besides the traditional users of soybean as a source of edible oil and animal feed, soybean is becoming increasingly popular with consumers—not only in traditional forms such as edamame, tofu, tempeh, miso, and natto, but also in newer forms such as meat analogs, soymilk, and soy cheese. The rise in popularity of soyfoods is due to two factors. First, the U.S. Food and Drug Administration permits soyfoods to carry health claims on the label, such as helping fight heart disease. Secondly, the soybean is the only significant source of isoflavones in the human diet, and isoflavones are now associated with various health benefits, such as easing menopause symptoms and preventing cancer (American Soybean Association, 2003).

As the soyfoods market diversifies and grows, varieties must be bred for each specific use. In addition, soybeans with modified oil profiles are being developed to meet the unique needs of various oil end users. Likewise, soybean is starting to be bred for

specialized industrial applications, such as soy ink, thus adding a new dimension to soybean breeding. Breeding for yield and resistance to abiotic stresses and to biological pests is no longer sufficient. Hence, breeders—along with farmers and the seed industry—will benefit from additional tools for marker-assisted selection and from a better understanding of which genes to breed for to get the desired outcome. Breeding efforts would be complemented by using markers to better characterize and exploit the soybean germplasm collection.

Having additional molecular markers and a better understanding of the soybean genome can help in making intelligent decisions for soybean germplasm conservation and continuing the characterization of soybean germplasm collections, helping to eliminate redundancy in the collections while ensuring important traits are identified and preserved.

Because of its magnitude, it is important to minimize the environmental footprint of the soybean crop. A greater understanding of soybean physiology and a greater ability to manipulate and breed resistance to various diseases can result in reduced pesticide use and other substantial environmental benefits. The best example may be the deployment of glyphosate-tolerant soybean. Not only has it permitted a switch to a more environmentally benign herbicide, it has facilitated the adoption of no-till agricultural practices, resulting in lower soil erosion rates, less water runoff and greater soil moisture, greater carbon sequestration, and the use of less fossil fuel (and hence less CO_2 emissions) to produce the crop (Fawcett and Towery, 2002).

Agricultural products account for about one-third of U.S. exports. According to the last year for which statistics are available, the US exported $69.7 billion worth of agricultural products. Of that total, soybean and soybean meal accounted for $7.6 billion, or well over 10% of total export revenue for U.S. agricultural products. Only coarse grains generated more export revenue. Furthermore, each $billion in exports is estimated to generate 16,000 jobs in the United States (USDA-Foreign Agricultural Service, 1997). Hence, the soybean is important for the economic well being of the US, and the crop must stay ahead of emerging pests and pathogens to remain viable.

CONCLUSIONS

It is now clear that comparative genomics is a powerful tool to investigate many biological processes. Advances in genomics of many plant species, including soybean, will accommodate these comparative approaches and will provide synergistic opportunities to advance plant science. Although investments in the genomics of model legume species will reveal much about legume biology, an important goal of such research is to translate this information into improvements in crop legumes. To accomplish this goal, knowledge of crop legumes, such as soybean, must be developed well enough that this information transfer can occur efficiently. Thus, regardless of efforts on model legumes, more emphasis is necessary on the genomics of soybean. What is outlined in this draft plan for soybean is just the first phase toward defining a comprehensive and cohesive strategy for soybean genomics. As such, it will set the stage for future efforts to obtain the full genome sequence of soybean. Moreover, achieving these initial goals will also greatly facilitate the development of a wide range of functional genomics tools for soybean. These are clearly needed if the soybean community is to take full advantage of the gene identification resources being developed in plant model species, as well as those that will come from soybean sequencing and mapping.

An important outcome of this investment in soybean genomics will be the recruitment of young scientists who will see soybean as a legitimate research system on which they can build their careers. This increase in critical mass will have many tangible benefits, including preserving the viability of soybean as an agriculture crop and contributing to homeland security by insuring a safe and available food supply. The many new uses to which soybean is currently and in the future will be applied can be aided with a more thorough understanding of the biology of this plant, accelerated by an investment in soybean genomics. This increasing knowledge base will also allow plant scientists to respond to new threats (e.g. soybean rust) that threaten to disrupt agricultural soybean production, and expand soybean usage (e.g. biodiesel). Clearly, in the case of soybean, the community and technologies are available to achieve the identified priority goals and the workshop participants were unanimous in stressing the need to move forward at the earliest possible date.

ACKNOWLEDGMENTS

The authors acknowledge the contributions of all of the participants and observers at the workshop. Special thanks to the following individuals for their editing of this manuscript: Sandra Clifton, Perry Cregan, Ken Dewar, Ann Dorrance, Jeff Doyle, David Grant, Mike Grusak, and Jim Specht.

Received December 17, 2003; returned for revision February 20, 2004; accepted February 20, 2004.

LITERATURE CITED

Abdin O, Zhou X, Coulman B, Cloutier D, Faris M, Smith D (1998) Effect of sucrose supplementation by stem injection on the development of soybean plants. J Exp Bot **49:** 2013–2018

American Soybean Association (2003) Nutrition and health: benefits of soy protein. http://www.asa-europe.org/ (November 3, 2003)

Arazi T, Slutsky SG, Shiboleth YM, Wang YZ, Rubinstein M, Barak X, Yang J, Gal-On A (2001) Engineering zucchini yellow mosaic potyvirus as a non-pathogenic vector for expression of heterologous proteins in cucurbits. J Biotechnol **87:** 67–82

Arumuganathan K, Earle ED (1991) Estimation of nuclear DNA content of plants by flow cytometry. Plant Mol Biol Rep **9:** 229–241

Barakat A, Matassi G, Bernardi G (1998) Distribution of genes in the genome of *Arabidopsis* thaliana and its implications for the genome organization of plants. Proc Natl Acad Sci USA **95:** 10044–10049

Carter TE Jr, Nelson R, Sneller CH, Cui Z (2004) Genetic diversity in soybean. *In* HR Boerma and JE Specht, eds, Soybeans: Improvement, Production, and Uses, Ed 3, Agronomy Monograph No. 16. American Society of Agronomy-Crop Science Society of America-Soil Science Society of America, Madison, WI, pp 303–416

Ching A, Caldwell KS, Jung M, Dolan M, Smith OS, Tingey S, Morgante M, Rafalski AJ (2002) SNP, frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genetics **3:** 19

Clough SJ, Vodkin LO (2004) Soybean microarrays: a genomics tool for crop improvement. *In* R Wilson, C Brummer, T Stalker, eds, Genomics for Legume Crops. AOCS Press, Champaign, IL, in press

Clough SJ, Philip R, Shealy R, Vodkin L (2000) NSF Soybean Microarray Workshop Manual. University of Illinois, May 16–18, 2000, Champaign, IL, http://soybeangenomics.cropsci.uiuc.edu

Danesh D, Penuela S, Mudge J, Denny R, Nordstrom H, Martinez J, Young N (1998) A bacterial artificial chromosome library for soybean and identification of clones near a major cyst nematode resistance gene. Theor Appl Genet **96:** 196–202

D'Erfurth I, Cosson V, Eschstruth A, Lucas H, Kondorosi A, Ratet P (2003) Efficient transposition of the *Tnt1* tobacco retrotransposon in the model legume *Medicago truncatula.* Plant J **34:** 95–106

Dear PH, Cook PR (1989) Happy mapping: a proposal for linkage mapping the human genome. Nucleic Acids Res **17:** 6795–6807

Dear PH, Cook PR (1993) Happy mapping: linkage mapping using a physical analogue of meiosis. Nucleic Acids Res **21:** 13–20

Devos KM, Gale MD (2000) Genome relationships: the grass model in current research. Plant Cell **12:** 637–646

Ding Y, Johnson MD, Chen WQ, Wong D, Chen Y-J, Benson SC, Lam JY, Kim Y-M, Shizuya H (2001) Five-color-based high-information-content fingerprinting of bacterial artificial chromosome clones using type IIS restriction endonucleases. Genomics **74:** 142–154

Dong Y, Huan S, Zhuang B, Zhao L, He M (1999) The genetic diversity in annual soybean. *In* HE Kaufman, ed, Proceedings of the World Soybean Research Conference VI, Chicago, Aug 4–7, 1999. Superior Printing, Champaign, IL, pp 147–154

Doyle JJ, Doyle JL, Brown AHD, Palmer RG (2002) Genomes, multiple origins, and lineage recombination in the *Glycine tomentella* (Leguminosae) polyploid complex: histone H3-D gene sequences. Evolution **56:** 1388–1402

Doyle JJ, Luckow MA (2003) The rest of the iceberg. Legume diversity and evolution in a phylogenetic context. Plant Physiol **131:** 900–910

Dzikowski B (1936) Studia nad soja *Glycine hispida* (Moench) Maxim. Cz. 1. Morfologia. Pamietnik Panstwowego Instytutu Naukowego Gospodarstwa Wiejskiego w Pulawach. Tom 16, zeszyt 2. Rosprawa Nr. **253:** 69–100

Escobar C, Hernandez LE, Jimenez A, Creissen G, Ruiz MT, Mullineaux PM (2003) Transient expression of *Arabidopsis thaliana* ascorbate peroxidase 3 in *Nicotiana benthamiana* plants infected with recombinant potato virus X. Plant Cell Rep **21:** 699–704

Fawcett R, Towery D (2002) Conservation tillage and plant biotechnology: how new technologies can improve the environment by reducing the need to plow. Conservation Tillage Information Center, West Lafayette, IN. http://www.ctic.purdue.edu/CTIC/BiotechPaper.pdf (April 16, 2004)

Fehr WR (1987) Soybean. *In* WR Fehr, ed, Principles of Cultivar Development, Vol 2. Macmillan Publishing, New York, pp 533–576

Flavell RB, Smith DB (1976) Nucleotide sequence organization in the wheat genome. Heredity **37:** 231–252

Foster-Hartnett D, Mudge J, Larsen D, Danesh D, Yan H, Denny R, Penuela S, Young ND (2002) Comparative genomic analysis of sequences sampled from a small region on soybean (*Glycine max*) molecular linkage group G. Genome **45:** 634–645

Gardiner J, Schroeder S, Polacco ML, Sanchez-Villeda H, Morgante M, Landewe T, Fengler K, Useche F, Hanafey M, Tingey S, et al (2004) Anchoring 9371 maize EST unigenes to the BAC contig map by two-dimensional overgo hybridization. Plant Physiol **134:** 1317–1326

Gizlice Z, Carter T, Burton JW (1996) Genetic diversity patterns of North American public soybean cultivars based on coefficient of parentage. Crop Sci **36:** 753–765

Goldberg RB (1978) DNA sequence organization in the soybean plant. Biochem Genet **16:** 45–51

Goldblatt P (1981) Cytology and phylogeny of Leguminosae. *In* RM Polhill and PH Raven, eds, Advances in Legume Systematics, Part 2. Royal Botanic Gardens, Kew, UK, pp 427–463

Grant D, Cregan P, Shoemaker RC (2000) Genome organization in dicots: genome duplication in Arabidopsis and synteny between soybean and Arabidopsis. Proc Natl Acad Sci USA **97:** 4168–4173

Gurley WB, Hepburn AG, Key JL (1979) Sequence organization of the soybean genome. Biochim Biophys Acta **561:** 167–183

Hedges BR, Palmer RG (1993) Mapping the w4 locus in soybean. Soybean Genetics Newsletter **20:** 20–26

Hirochika H (2001) Contribution of the Tos17 retrotransposon to rice functional genomics. Curr Opin Plant Biol **4:** 118–122

Hymowitz T (2004) Speciation and cytogenetics. *In* HR Boerma and JE Specht, eds, Soybeans: Improvement, Production, and Uses, Ed 3, Agronomy Monograph No. 16. American Society of Agronomy-Crop Science Society of America-Soil Science Society of America, Madison, WI, pp 97–136

Jarvik T, Lark KG (1998) Characterization of *Soymar1*, a *Mariner* elementin soybean. Genetics **149:** 1569–1574

Kauffman HE (1999) Proceedings, invited and contributed papers and posters. World Soybean Research Conference VI, August 4–7, 1999, Chicago. National Soybean Research Laboratory, University of Illinois, Urbana, IL

Kawabe A, Miyashita NT (1999) DNA variation in the basic chitinase locus (*ChiB*) region of the wild plant *Arabidopsis thaliana.* Genetics **153:** 1445–1453

Kawabe A, Yamane K, Miyashita NT (2000) DNA polymorphism at the cytosolic phosphoglucose isomerase (*PgiC*) locus of the wild plant *Arabidopsis thaliana*. Genetics **156:** 1339–1347

Keim P, Beavis W, Schupp J, Freestone R (1992) Evaluation of soybean RFLP marker diversity in adapted germplasm. Theor Appl Genet **85:** 205–212

Keim P, Schupp JC, Coryell RG (1996) A high-density soybean genetic map based on AFLP. Crop Sci **36:** 786–792

Kinney AJ (1998) Plants as industrial chemical factories—new oils from genetically engineered soybeans. Fett/Lipid **100:** 173–176

Kuittinen H, Aguade M (2000) Nucleotide variation at the *CHALCONE ISOMERASE* locus in *Arabidopsis thaliana.* Genetics **155:** 863–872

Laten HM, Majumdar A, Gaucher EA (1998) SIRE-1, a copia/Ty1-like retroelement from soybean, encodes a retroviral envelope-like protein. Proc Natl Acad Sci USA **95:** 6897–6902

Lawrence RJ, Pikaard CS (2003) Transgene-induced RNA interference: a strategy for overcoming gene redundancy in polyploids to generate loss-of-function mutations. Plant J **36:** 114–121

Lee JM, Grant D, Vallejos CE, Shoemaker RC (2001) Genome organization in dicots. II. Arabidopsis as a bridging species to resolve genome duplication events among legumes. Theor Appl Genet **103:** 765–773

Lu R, Martin-Hernandez AM, Peart JR, Malcuit I, Baulcaumbe DC (2003) Virus-induced gene-silencing in plants. Methods **30:** 296–303

Luo MC, Thomas C, You FM, Hsiao J, Shu OY, Buell CR, Malandro M, McGuire PE, Anderson OD, Dvorak J (2003) High-throughput fingerprinting of bacterial artificial chromosomes using the SNaPshot labeling kit and sizing of restriction fragments by capillary electrophoresis. Genomics **82:** 378–389

Mansur LM, Orf JH, Chase K, Jarvik T, Cregan PB, Lark KG (1996) Genetic mapping of agronomic traits using recombinant inbred lines of soybean. Crop Sci **36:** 1327–1336

Marek LF, Mudge J, Darnielle L, Grant D, Hanson N, Paz M, Huihuang Y, Denny R, Larson K, Foster-Hartnett D, et al (2001) Soybean genomic survey: BAC-end sequences near RFLP and SSR markers. Genome **44:** 572–581

Marek L, Shoemaker R (1997) BAC contig development by fingerprint analysis in soybean. Genome **40:** 420–427

McCallum CM, Comai L, Greene EA, Henikoff S (2000) Targeting Induced Local Lesions IN Genomes (TILLING) for plant functional genomics. Plant Physiol **123:** 439–442

Meksem K, Zhang HB, Lightfoot DA (2000) Two transformation ready large insert clone libraries for soybean: physical mapping of resistance to Soybean Cyst Nematode and Sudden Death Syndrome. Theor Appl Genet **101:** 747–755

Morgante M, Jurman I, Shi L, Zhu T, Keim P, Rafalski JA (1997) The STR120 satellite DNA of soybean: organization, evolution and chromosomal specificity. Chromosome Res **5:** 363–373

Palmer RG, Kilen TC (1987) Qualitative genetics and cytogenetics. *In* JR Wilcox, ed, Soybeans: Improvement, Production, and Uses, Ed 2, Agronomy Monographs No. 16. American Society of Agronomy-Crop Science Society of America-Soil Science Society of America, Madison, WI, pp 135–209

Palmer RG, Pfeiffer TW, Buss GR, Kilen TC (2004) Qualitative genetics. *In* HR Boerma and JE Specht, eds, Soybeans: Improvement, Production, and Uses, Ed 3, Agronomy Monograph No. 16. American Society of Agronomy-Crop Science Society of America-Soil Science Society of America, Madison, WI, pp 137–234

Parrott WA, Clemente TE (2004) Transgenic soybean. *In* HR Boerma and JE Specht, eds, Soybeans: Improvement, Production, and Uses, Ed 3, Agronomy Monograph No. 16. American Society of Agronomy-Crop Science Society of America-Soil Science Society of America, Madison, WI, pp 265–302

Purugganan MD, Suddith JL (1999) Molecular population genetics of floral homeotic loci. Departures from the equilibrium-neutral model at the APETALA3 and PISTILLATA genes of *Arabidopsis thaliana*. Genetics **151:** 839–848

Rhodes P, Vodkin L (1988) Organization of the Tgm family of transposable elements in soybean. Genetics **120:** 597–604

Schlueter JA, Dixon P, Granger C, Grant D, Clark L, Doyle JJ, Shoemaker RC (2004) Mining EST databases to resolve evolutionary events in major crop species. Genome (in press)

Searle IR, Men AE, Laniya TS, Buzas DM, Iturbe-Ormaetxe I, Carroll BJ, Gresshoff PM (2003) Long-distance signaling in nodulation directed by a CLAVATA1-like receptor kinase. Science **299:** 109–112

Shoemaker R, Keim P, Vodkin L, Retzel E, Clifton SW, Waterston R, Smoller D, Coryell V, Khanna A, Erpelding J, et al (2002) A compilation of soybean ESTs: generation and analysis. Genome **45:** 329–338

Shoemaker RC, Olson TC (1993) Molecular linkage map of soybean. *In* S O'Brien, ed, Genetic Maps: Locus Maps of Complex Genomes, Ed 6. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Shoemaker RC, Polzin K, Labate J, Specht J, Brummer EC, Olson T, Young N, Concibido V, Wilcox J, Tamulonis JP, et al. (1996) Genome duplication in soybean (*Glycine* subgenus *soja*). Genetics **144:** 329–338

Shoemaker RC, Specht JE (1995) Integration of the soybean molecular and classical genetic linkage groups. Crop Sci **35:** 436–446

Singh RJ, Hymowitz T (1988) The genomic relationship between *Glycine max* (L.) Merr. and *G. soja* Sieb. and Zucc. as revealed by pachytene chromosome analysis. Theor Appl Genet **76:** 705–711

Slade AJ, Faciotti D, Fuerstenberg SI, Steine MN, Loeffler D, McGuire C, Jones P, Claire CM, Knauf V (2003) Detection of an allelic series in soybean and wheat through TILLING. American Society of Plant Biologists. http://abstracts.aspb.org/pb2003/public/P72/1451.html (April 16, 2004)

Smith DB, Flavell RB (1977) Nucleotide sequence organization in the rye genome. Biochim Biophys Acta **474:** 82–97

Smith PMC, Atkins CA (2002) Purine biosynthesis. Big in cell division, even bigger in nitrogen assimilation. Plant Physiol **128:** 793–802

Somers DA, Samac DA, Olhoft PM (2003) Recent advances in legume transformation. Plant Physiol **131:** 892–899

Sprent JI (2001) Nodulation in Legumes. Royal Botanic Gardens, Kew, UK

Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF, Gaut BS (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Proc Natl Acad Sci USA **98:** 9161–9166

Thangavelu M, James AB, Bankier A, Bryan GJ, Dear PH, Waugh R (2003) HAPPY mapping in a plant genome: reconstruction and analysis of a high-resolution physical map of a 1.9 Mbp region of *Arabidopsis thaliana* chromosome 4. Plant Biotechnol J **1:** 23–31

Thibaud-Nissen F, Shealy RT, Khanna A, Vodkin LO (2003) Clustering of microarray data reveals transcript patterns associated with somatic embryogenesis in soybean. Plant Physiol **132:** 118–136

Tomkins JP, Mahalingham R, Miller-Smith H, Goicoechea JL, Knapp HT, Wing RA (1999) A soybean bacterial artificial chromosome library for PI 437654 and the identification of clones associated with cyst nematode resistance. Plant Mol Biol **41:** 25–32

USDA-Foreign Agricultural Service (1997) U.S. Agricultural, Fish, and Wood Products Exports. http://www.fas.usda.gov/info/factsheets/exptfy.html (October 3, 2003)

Vahedian M, Shi L, Shu T, Okimoto R, Danna K, Keim P (1995) Genomic organization and evolution of the soybean SB92 satellite sequence. Plant Mol Biol **29:** 857–862

Vodkin LO, Rhodes PR, Goldberg RB (1983) A lectin gene insertion has the structural features of a transposable element. Cell **34:** 1023–1031

Vuong TD, Hartman GL (2003) GL Evaluation of soybean resistance to Sclerotinia stem rot using reciprocal grafting. Plant Dis **87:** 154–158

Yan HH, Mudge J, Kim D-J, Shoemaker RC, Cook DR, Young ND (2003) Estimates of conserved microsynteny among the genomes of *Glycine max, Medicago truncatula,* and *Arabidopsis thaliana.* Theor Appl Genet (in press)

Young ND, Mudge J, Ellis THN (2003) Legume genomes: more than peas in a pod. Curr Opin Plant Biol **6:** 199–204

Zhu Y-L, Song Q-J, Hyten DL, Van Tassell CP, Matukumalli LK, Grimm DR, Hyatt SM, Fickus EW, Young ND, Cregan PB (2003) Single nucleotide polymorphisms (SNPs) in soybean. Genetics **163:** 1123–1134