Soybean Genomics Research Program Strategic Plan
Research to Bridge the Gap between Genotype and Phenotype in Soybean 2017-2021

Edited by Jennifer Jones (United Soybean Board, St. Louis, MO) and Robert Stupar (University of Minnesota, St. Paul, MN)

On July 23-24, 2015, 46 soybean researchers, representing molecular biologists, geneticists, and breeders convened in St Louis, MO for a workshop sponsored by the United Soybean Board. The purpose of the workshop was to review the current status of soybean research, including the successes of the 2012-2016 Soybean Genomics Research Strategic Plan (Boerma et al., 2011), and to capture the consensus regarding logical goals to move soybean genetics and genomics research forward in the next five years. This strategic plan marks over two decades of effort on the part of the soybean research community. Previous plans have been generated and implemented, each showing an outstanding level of achievement.

Soybean [*Glycine max* (L.)Merr.] is a remarkable plant species for its unique biology, evolutionary history, and utility for civilization. Soybeans have been grown in the U.S. for 250 years. First imported from China in 1765 by Samuel Bowen, the intended purpose was to manufacture soy sauce and vermicelli (Hymowitz and Harlan, 1983). One hundred and fifty years later, soybean acreage was still less than 2 million acres and largely used for hay. At that time, it is doubtful than anyone would have envisioned that soybean would become the second most planted field crop in the U.S., totaling over 80 million acres and generating over 3.9 billion bushels in 2014. Furthermore, it would have been inconceivable to predict that today approximately 25% of the beans and/or the meal generated from them are exported back to China where the soybean was first cultivated over 3000 years ago.

Soybean has become a successful agricultural crop, largely due to its vast utility. It is widely used for human food, livestock forage, oil (edible vegetable and cooking oil, biodiesel, and as a sustainable replacement for petroleum products in industrial uses), and animal feed (98% of US soybean meal is used as a protein source in the livestock industry). The ability of soybean to fix nitrogen through symbiosis has supported its wide spread use as a rotation crop. In addition to nitrogen symbiosis, soybean has several distinct biological features that make it an attractive system for research, including a paleopolyploid genome with two distinct recent whole genome duplications (Schmutz et al., 2010), and many well-characterized biotic and abiotic stress tolerances. For molecular biologists, bioinformaticists, geneticists, agronomists, and plant breeders, the growing body of high quality soybean research, resources and data make it a tractable system to leverage for advancing scientific inquiry and crop improvement.

Soybean research in the U.S. is bolstered by a cohesive community of scientists spanning different disciplines and a supportive commodity program (United Soybean Board, Qualified State Soybean Boards, and regional Boards) as well as opportunities for federal grant support (the United States Department of Agriculture and the National Science Foundation). The United Soybean Board alone has invested millions of dollars into basic and applied research over the last 20 years. The results of these efforts are many and have been demonstrated by hundreds of publications and in the achievement of goals developed in prior community-wide strategic planning sessions conducted in 1999, 2001, 2003, 2005, 2007 and 2010 (Boerma et al., 2011).

As the community looks towards the future of soybean genetics and genomics research, the need to address the grand challenge of bridging the genotype to phenotype gap rises to the forefront. The vision of the soybean research community must be broad and innovative. Addressing this challenge will require an integrative research community that brings together research activities including statistics, bioinformatics, computational modeling, engineering, soils, physiology, microbiology, molecular biology, genomics and plant breeding. It will necessitate continuous improvement of –omics based resources assimilated in systems-platforms. This "Think Big" strategy will enable the community to attract interdisciplinary research activities, which in turn will help create a dynamic and engaging soybean research environment. The continued development of soybean research capacities will attract new talent and opportunities in the education of students and postdoctoral scholars, and facilitate a community that contributes to science literacy education and outreach to society at large.

Three overarching issues surfaced at the Soybean Genomics Strategic Planning Workshop. First is the need to integrate the large and disparate data sets that have been generated in recent years (and continue to be generated) by the community. These data and tools need to be accessible to researchers throughout the community to achieve current objectives and advance new research opportunities. Second is the need to interact with other research fields and communities, particularly those in the agricultural and plant sciences, and adopt the most useful emerging tools and ideas. The soybean research community is strong, but it should not become insular. Reaching out to the broader scientific community will speed the progress of soybean scientific research and open opportunities to leverage new initiatives and opportunities. Third, the soybean community should more effectively engage with federal agencies and scientific societies (*e.g.* the American Society of Plant Biologists, the Crop Science Society of America, the National Science Foundation) to contribute towards the development of research and funding objectives at the state and federal levels.

Lastly, the workshop participants recognized that genome-wide datasets and resources are integral to addressing the genotype-to-phenotype gap and that these resources, particularly those funded publicly, are public trusts. It is imperative that scientists provide data freely and in a timely manner to reduce redundancy in efforts, improve efficiency, and facilitate research progress. As such, the community has endorsed two community data standards:

Data producers should freely share genome-wide data and resources. Genome-wide data sets, including resequencing, pan-genomes, new reference genomes and other genome-scale datasets, should be freely shared shortly after generation in order to advance the community as a whole. Soybase.org may serve as a portal for listing available datasets and may host some of these important community resources.

Data users should be responsible with shared genomic data/resources. Sharing of expensive and time-intensive data and resources means that users should be respectful of the use of such data and resources and ensure that proper recognition is given to the producer. This is especially true in the case of pre-publication data. The producer should be consulted before publishing to ensure that proper recognition is given and that their ability to publish their findings is not co-opted. It is recommended that the soybean community adopt the relevant guidelines outlined by Birney and colleagues (Birney et al., 2009).

Soybean Research Status 2016

> "The soybean research community has engaged a transparent process for developing and implementing a strategic framework for a national program to unlock the secrets of the soybean genome. The success of this process is evidenced by achieving the soybean genome sequence in record time, compared to similar efforts in other major crops, and many useful tools to expedite elite variety development." – R.F. Wilson and D. Grant (August, 2010)

The soybean genomics researchers have an impressive history of strategic planning and implementation to achieve research and resource milestones which are well summarized in the prior plans and are also available at SoyBase (http://www.soybase.org/SoyGEC). Below is a brief summary of key accomplishments since the 2012-2016 report (the full accomplishments report can be found at SoyBase (http://soybase.org//SoyGenStrat2010/SoyGenStratPlanAccomplishments).

The primary goals of the 2012-2016 Soybean Genomic Research Program Strategic Plan were to:

1. Improve the quality and utility of the soybean genome sequence.
2. Develop functional genomic technologies to optimize utility of genome sequence information in germplasm enhancement.
3. Optimize and expand transgenic methods and improve understanding of natural genes for modification of trait expression.
4. Optimize breeding efficiency with robust sequence-based resources.

The first goal focused on improving the reference Williams 82 genome as well as developing and leveraging new genomic resources. One key accomplishment is the release of the new soybean genome assembly, *Glycine max* Wm82.a2.v1. The new release integrates new high-resolution genetic maps (Song et al., 2016), corrects some issues in pseudomolecule reconstruction and provides better annotation and gene prediction. The new genome assembly is estimated to contain 98% of known soybean protein coding genes. It is available at Phytozome and SoyBase.

Phytozome: http://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Gmax

SoyBase: http://www.soybase.org/SequenceIntro.php

A soybean SNP chip with over 50,000 soybean specific single nucleotide polymorphisms (SNPs) was developed, validated, and commercialized by Illumina. The SoySNP50K Illumina BeadChip SNP assay contains SNPs randomly distributed in the soybean genome with higher density in euchromatic than heterochromatic regions. The SNPs are present at high minor allele frequencies and in the assay have a high allele call success rate (Song et al., 2013). The SoySNP50K chip was applied to the entire USDA Soybean Germplasm Collection (Song et al., 2015) and has been widely applied to facilitate mapping efforts, genomic selection, and genome wide association studies. A smaller Soy6K SNP chip has also been developed and the community worked with Illumina to reduce the price to as low as $20 per sample.

A major effort to understand genetic and agronomic diversity in soybean was undertaken in the soybean nested association mapping (NAM) project. NAM populations were developed by crossing cultivar IA3023 with 40 lines selected by the soybean breeding community for high yield, genetic diversity, or as plant introductions. These populations were phenotyped for several agronomic traits including yield and seed

composition. The NAM parents were sequenced and analyzed with the SoySNP50k chips. Comparative genome hybridization and resequencing data was used to identify copy number variants (CNV) among the NAM parents (Anderson et al., 2014). Work continues on these populations to develop genomic selection models and to narrow mapping intervals for genes related to traits of interest.

Another key project, known as the Soybean Milestones Project, was aimed at understanding 90 years of breeder applied selection in soybean. DNA sequences from eighty soybean lines reflecting 90 years of soybean breeding have been assembled relative to the Williams 82 reference genome. SNP, CNV, genotype and phenotype data can be viewed at SoyBase.org (http://shiny.soybase.org/SNP/ and http://shiny.soybase.org/CNV/). The project has also produced small RNA, transcriptome and methylome data. Furthermore, a depository of cloned transcription factors was generated.

Resequencing efforts in *G. max* and *G. soja* resulted in identification of genes involved in soybean domestication and improvement (Chung et al., 2014; Joshi et al., 2013; Li et al., 2013; Zhou et al., 2015). Great strides were also made in improving resources for genomic analysis, including a gene ontology enrichment tool that provides gene ontology (GO) information for all genes in the soybean genome (http://www.soybase.org/goslimgraphic_v2/dashboard.php) and a gene annotation tool (http://www.soybase.org/genomeannotation). Furthermore, data from a subset of soybean fast neutron and soybean virus induced gene silencing (VIGS) are available on SoyBase:

Fast Neutron: http://www.soybase.org/mutants/

VIGS: http://soybase.org/SoyVIGS/

Collaboration between the members of the soybean community and others has led to the development of the Legume Information System (LIS) (http://www.comparative-legumes.org). LIS provides a platform where soybean can be used as a model legume system by researchers working on other pulse legumes, some of which are critical food sources in poor regions of the world. Conversely, soybean researchers can leverage genomic data from other legumes at LIS.

Many accomplishments were made in the area of functional genomics resources. Numerous publications and data sets have detailed the mRNA and small RNA changes in developing seeds and other tissue and organ systems, as wells as in plants challenged by pathogens and abiotic stresses (http://soybase.org//SoyGenStrat2010/SoyGenStratPlanAccomplishments). Initial descriptions of the role of methylation changes in the genome have been explored as well as binding sites for transcription factors. Delineation of the molecular bases for some disease resistance genes and plant traits has been reported. Gene function studies in soybean have been facilitated by the development and widespread adoption of a variety of reverse genetic tools including mutant populations, RNA interference methods, Virus-Induced Gene Silencing (VIGS) technologies, and gene editing platforms (*e.g.*, Zinc finger nucleases, CRISPR/Cas9).

Efforts to generate a saturated transposon insertion population are ongoing, but have been challenging primarily due to lack of resource allocations. The community has exploited multiple transposon tagging systems as a functional genomics resources for the crop (Cui et al., 2013; Hancock et al., 2011; Mathieu et al., 2009; Raval et al., 2013). Genotypic characterizations of the fast neutron mutagenized soybean populations have been pursued (Bolon et al., 2014; Gillman et al., 2014). Furthermore, chemically-

mutagenized populations have been developed and utilized to identify genes involved in a range of important traits and biological processes (e.g. Carrero-Colón et al., 2014; Liu et al., 2012; Xia et al., 2012). For each of these platforms, mutation identification by next-generation sequencing approaches promises to facilitate more efficient use of these resources in the coming years.

Establishment of a genetic repository and distribution center for soybean mutants and transgenic lines was identified as a primary goal in the prior strategic plan; it has yet to be accomplished. USDA/ARS allocated funds to establish and maintain a repository to house a collection of fast neutron characterized seed lots at the University of Illinois, and plans are underway to begin mutant submission. Millions of dollars have been invested to develop soybean mutant resources, this investment remains in danger of being lost due to the lack of an appropriate infrastructure for long-term storage and distribution.

Genome editing has seen tremendous technology development, particularly with the implementation of TAL effector nucleases (TALENs) and the clustered regulatory interspaced short palindromic repeat Cas-based RNA-guided DNA endonuclease (CRISPR/Cas9) methods. Several examples of targeted mutagenesis for soybean and advancements in the methodologies for soybean users have been recently published (Haun et al., 2014; Jacobs et al., 2015).

Expanded genomic resources make the discovery of genes and or QTL responsible for qualitative and quantitative traits much more tractable and in many cases lead to development of perfect markers and identification of causal genes. The SoySNP50k was applied to the entire USDA soybean germplasm collection, generating a remarkable resource for geneticists and breeders alike. To facilitate utilization, the data has been incorporated into SoyBase and the complete collection of haplotypes or a user-selected subset can be downloaded (http://www.soybase.org/snps/). Qualitative genes responsible for biotic tolerances (such as salt tolerance and iron deficiency), pest and diseases (such as aphid and phytophthora resistances) and agronomic traits (such as shattering, seed per pod number, and high oleic levels) have all been found. Similarly, the loci, and in many cases genes responsible for quantitative traits such as nematode resistance, maturity, and seed coat color were also determined (http://soybase.org//SoyGenStrat2010/SoyGenStratPlanAccomplishments). The expectation is that associating agronomic traits with not only the specific genes, but even the specific alleles for the gene will become easier as genomic, phenomic and other data sets are integrated.

Genomic selection has the potential to increase breeding efficiency by decreasing breeding cycle number and reducing phenotyping costs. Several projects are underway to develop predictive models for soybean populations. Already the method has been applied to specific traits including yield, plant height, days to maturity, and resistance to SCN (Bao et al., 2014; Jarquin et al., 2014).

These achievements and many more lay the foundation for the 2017-2021 Soybean Genomics Strategic Plan, which is outlined and described below.

Goal 1: Comprehensive genomic resources for soybean are available to the community.
Strategies:

- 1.1: A diverse set of soybean reference genomes are generated, representing geographic and/or nodal diversity.
- 1.2: DNA resequencing of the entire USDA core collection is completed.
- 1.3: Preliminary pan genomes for both wild and cultivated soybeans are developed.

Goal 2: Functional genomics resources, facilities, and new tools are advanced to facilitate gene function studies.

Strategies:

- 2.1: Soybean mutant resources are generated.
- 2.2: Community repository for soybean functional mutant collections is in place.
- 2.3: Pathways representing key aspects of soybean biology are delineated.
- 2.4: Gene editing technologies are made more efficient.
- 2.5: Transient expression methods for gain of function studies are improved**.**

GOAL 3: Breeding efficiency is improved.

Strategies:

- 3.1 Efficiency of population development is improved.
- 3.2: Best practices for phenotyping and envirotyping data are used and selection models are optimized.
- 3.3: New agronomically relevant traits and their alleles are discovered.
- 3.4: Breeder selection targets are improved.

Goal 4: Bioinformatics infrastructure, tools and standards for leveraging multiple datatypes are in place.

Strategies

- 4.1: Data integration, quality, analysis, and visualization tools are improved and expanded.
- 4.2: Controlled vocabulary and metadata standards across multiple data types are generated and communicated to users.

Strategic Plan for Soybean Genomics 2017-2021

Goal 1: Comprehensive genomic resources for soybean are available to the community.

Genome-wide data sets and resources are a major asset to the soybean genetics, genomics and breeding communities. Soybean is in an enviable position of having a fully sequenced genome, Williams 82 (Schmutz et al., 2010), an assembled pan-genome sequence based on seven wild soybean (*G. soja*) accessions (Li et al., 2014), and several hundreds of lines that have been or are currently being resequenced using short read platforms (Lam et al., 2010; Zhou et al., 2015).

Soybean also has a key resource, the USDA Soybean Germplasm collection, which contains over 20,000 domesticated and wild soybean accessions. The entire collection was genotyped with the SoySNP50K Illumina BeadChip SNP assay (Song et al., 2015). This information was then used to identify core collections of *G. max* and *G. soja*, that contain maximum diversity (>98 and >96% respectively). The sets represent about 10% of each collection, 1418 accessions of *G. max* and 81 accessions of *G. soja*. These core collections are resources the community can focus on for many different efforts and, by using the same lines or further subsets of this collection, results from different labs and diverse experiments can be compared against each other. (Song, personal communication/manuscript in preparation).

It is envisioned that in five years, key additional genomes can be assembled, short read sequences can be generated for the complete core set, and a high-quality pan genome can be developed. This would provide an extensive package of genomics resources from the soybean collection, resulting in the elevation of

soybeans as a research system and new datasets for systematically linking genes or genetic variants to traits for improvement. These genome-wide data sets, coupled with seamless data integration and bioinformatics infrastructure, will enable the elucidation of the genetic and molecular mechanisms that underlie phenotypic variation.

Strategy 1.1: A diverse set of soybean reference genomes are generated, representing geographic and/or nodal diversity.

The Williams 82 reference soybean genome has been widely used and is valuable to the community for identifying nucleotide-level variation. However it is likely that a great amount of intraspecific sequence variation lies outside of this one genome. Using just one reference genome limits variant detection to regions that can be most easily mapped onto this genome, leaving a wide array of other sequence variants, including presence-absence variants (PAV), copy-number variants (CNV), and more complex chromosomal rearrangements (*e.g.* inversions, translocations) difficult to elucidate. These events represent a highly polymorphic subset of variants that may that underlie traits of agricultural importance and biological interest.

A second de novo reference genome, Lee, is being sequenced at the present time (H. Nguyen, personal communication). This genome will represent the Southern U.S. soybean germplasm base, as Williams 82 represents the Northern U.S. germplasm base. Additional soybean accessions or cultivars for whole genome sequencing should be selected from the core collection to represent geographic and or molecular haplotype diversity.

Anticipated outcomes:

- At least 10 diverse soybean genomes are sequenced and assembled de novo.
- Complex, rearranged regions and segments not present in Williams 82 are described, catalogued, and archived in SoyBase.

Strategy 1.2: DNA resequencing of the entire USDA core collection is completed.

Decades of work has preceded the identification of a diverse core subset of the USDA soybean germplasm collection. This core consists of approximately 1,500 domesticated accessions and 81 *G. soja* accessions. One third to one half of the collection is either already sequenced or being sequenced using short read methods that are rapid, efficient and inexpensive. Focused efforts on this collection will extend previous work and flesh out limitations of earlier methodology. When completed this will be an unprecedented resource for the community.

Anticipated outcomes:

- Short read sequence, assembly, and analysis of entire UDSA soybean core collection is completed and freely publically available (approximately 1,500 accessions).
- Short read sequence, assembly, and analysis of core *G. soja* collection is completed and freely publically available (81 accessions).

1.3: Preliminary pan genomes for both wild and cultivated soybeans are developed.

Reference genomes do not contain all the genetic variation present in a species. A pan genome represents core and dispensable components of soybean genomes and would capture most of the diversity in the US soybean breeding population.

Anticipated outcomes:

- Reference genomes are integrated to yield a "pan-genome" that comprehensively represents the genes and regulatory DNA elements found within US soybean breeding germplasm.
- Polymorphisms from short read sequences of the core collection are assembled onto pan genome.
- Integration of pan-genome data into SoyBase.

Goal 2: Functional genomics resources, facilities, and new tools are advanced to facilitate gene function studies.

While great advancement has been made in soybean genomics, progress toward dissecting the function of each of the 50,000 plus genes in soybean has been slow. Comprehensive, characterized and cataloged mutant populations in which soybean genes are either silenced or enhanced would provide any soybean researcher ready access to tools for gene function studies. Coupled with these efforts is the critical need to catalog, store, and disseminate the mutant germplasm to the community. Since the last report, the USDA/ARS has committed funds to establish and maintain some mutant collections. There are currently over 20,000 soybean lines and plant introductions in the USDA Soybean Germplasm Collection. How mutants generated by the newer technologies will be added to its existing collections is still under discussion. The phenotypic and genotypic information needed for these mutant lines is different from that associated with maintaining the traditional lines. The Arabidopsis Biological Resource Center at The Ohio State University collects, reproduces, preserves and distributes seed of *Arabidopsis thaliana* (http://abrc.osu.edu/) and represents a successful model for the soybean community to follow. Seed stock information and ordering for the ABRC are fully integrated into The Arabidopsis Information Resource (TAIR) https://www.arabidopsis.org/. Similarly, a Soybean Biological Resource Center (SBRC) could be fully integrated into SoyBase.

Once generated these resources should be used to assess gene and protein expression networks that lead to plant traits and important agronomic characteristics. Improvement methods for transient expression, particularly high throughput methods, would greatly advance studies of gene function. Ultimately, manipulation of genes and pathways will be facilitated by the development and adoption of new technologies for site specific gene modification.

It is recognized that the resources to achieve some of the goals in this strategy may be difficult to obtain, however the community must continue with these large scale projects and needs to develop the biological resources needed to elucidate gene function. It is envisioned that in five years, there will be at least one comprehensive mutant collection, with all genes modified and a permanent storage location for the mutants. In addition, multiple gene editing strategies will be in place, as will be the means to rapidly assess gene function via high throughput transient expression. Taken together, the community will be well on the way to understanding the function of key soybean genes and pathways.

Strategy 2.1: Soybean mutant resources are generated.

Comprehensive mutant populations using a variety of different approaches are critical to address the need to decipher soybean gene function. These populations are useful for many applications, with each

approach possessing unique advantages and limitations. The majority of soybean mutant efforts to date have focused on populations with chemical (*e.g.*, EMS), radiation (*e.g*., fast neutron), or transposon (*e.g*., mPing, Tnt1, Ac/Ds and Tgm9)-induced mutations. Lines wherein transposons are activated with linked enhancer sequences are also being developed to generate gain-of-function mutants. Additional resources are needed to ensure that the populations are completed.

Anticipated outcomes:

- One or more comprehensive collection of mutants in which most every soybean gene is disrupted, knocked out, or silenced; and/or tools for facile disruption of genes of interest (Strategy 2.4).
- Soybean enhancer trap lines in which most every soybean gene is activated.

Strategy 2.2: Community repository for soybean functional mutant collections is in place.
Generation of mutations by fast neutron bombardment, EMS, and transposon tagging has progressed far enough in recent years to bring the establishment of a permanent repository for the seed generated from these projects to a high priority. This is essential so that the efforts to develop these resources (and the millions of dollars invested by funding agencies in these efforts) will not be lost and the wider community has easier access to the collections and the information about each line.

Anticipated Outcomes:

- Permanent soybean seed repository is established.
- Standardized methodology for depositing mutant/transgenic lines to soybean seed repository is generated.
- Online ordering capability for community is created.

Strategy 2.3: Pathways representing key aspects of soybean biology are delineated.
Understanding the action of genes and pathways controlling phenotypes of importance to plant growth and agronomic performance is critical. Particular emphasis should be placed on the pathways critical to soybean agronomics, economics, and end uses: increased seed protein and oil quality and quantity, diseases of major importance, and yield. Increasingly, attention is being paid to the phytobiome, the system of factors affecting plant growth and crop yields. This is particularly important for soybean, as producers are facing an increased need to understand how soil health, environment, and on farm inputs effect crop sustainability and production.

Anticipated Outcomes:

- Identification and assessment of pathways for improved protein and oil composition and their links to yield.
- Identification and assessment of pathways to major disease resistances and their links to yield.
- Increased understanding of the soybean phytobiome and its influence on soybean performance.

Strategy 2.4: Gene editing technologies are made more efficient.
"DNA scissors"-based targeted mutagenesis and gene editing has rapidly become a powerful tool for developing mutations in specific genes. CRISPR/Cas9 has emerged as the leading platform for this application because of its high efficacy and relatively efficient construct development strategy. Furthermore, the gene editing platforms are able to modify specific genes without disrupting the genetic background of the host and result in a non-transgenic mutant plant in subsequent generations. With promising early results in soybean, CRISPR/Cas9 can be used to develop mutants in specific genes or gene

families. With high efficiency and relative ease of use, CRISPR/Cas9 promises to become a community tool for analyzing gene function and introducing new traits or trait improvements.

Anticipated outcomes:

- CRISPR/Cas9 will be further developed in soybean as a functional tool to analyze genes.
- Genes conferring high-value traits will be modified using targeted genome engineering approaches. With technical optimization, it will be possible to recover a wide range of mutated alleles for any given gene of interest.

Strategy 2.5: Transient expression methods for gain of function studies are improved.
Many of the forward and reverse genetic tools in soybean are designed to reduce or eliminate gene function and are often confounded by the functional redundancy of duplicated genomes. High throughput transient expression of genes in plants offers a complementary approach to understand gene function, but efficient methods for agroinfiltration in soybean are lacking. Improved transient expression systems would speed the analysis of gene function studies.

Anticipated Outcomes:

- Optimized protocols for agroinfiltration in soybean are developed.
- New methods for high throughput transient expression studies are implemented.

GOAL 3: Breeding efficiency is improved.
With the availability of genome-wide data sets, genotyping capacity has greatly expanded over the past decade. Similarly, high-throughput phenotyping and envirotyping are now being developed. While industry and public soybean breeding have long since adopted marker assisted selection, there remains a knowledge gap in how to most efficiently apply genomic selection and phenomics in breeding programs.

The community has developed extensive genotyping resources for some populations, albeit in the absence of extensive phenotyping. Meanwhile, some populations are extensively phenotyped, but not genotyped. For example, the Northern and Southern uniform soybean trials are annually conducted on more than 200 lines per year, including a wide range of maturity groups, with yield trials carried out on each line at over 12 locations. Conversely, extensive investments have been made (and are being made) to genotype the soybean nested association mapping (NAM) lines and the USDA core collection. This currently includes tens of thousands of markers. Phenotyping efforts could be substantially increased to include more environments for agronomic evaluation of the USDA and a more extensive list of phenotypes.

The strategies below are considered in the context of current genomic and genotypic datasets, the availability of a large number of re-sequenced soybean genomes and the expectation of continued, rapid expansion of these datasets. This plethora of information coupled with innovations in phenotypic data acquisition will position the soybean community to explore novel approaches to exploit these resources as a means to bridge the genotype to phenotype gap.

Strategy 3.1: Efficiency of population development is improved.
A bottleneck in developing soybean populations is the ability to create large numbers of hybrids and subsequent generation of inbred lineages. Avenues to mitigate this bottleneck include novel approaches

to enhance crossing efficiency, genetic designs to influence recombination frequencies, and the ability to produce double haploids. Increasing recombination frequency could reduce breeding cycles required for combining traits, while reduction of recombination could fix optimal haplotypes or trait combinations. In addition, optimal population numbers, size, type and structure may be different from present considering the application of genomics and phenomics tools in breeding programs.

Anticipated Outcomes:

- Methods or tools for improved crossing ability are developed.
- Tools are developed to control recombination frequency.
- A doubled haploid system for soybean is generated.

Strategy 3.2: Best practices for phenotyping and envirotyping data used in selection models are optimized.

Relevant and high quality data is required for development of effective prediction and selection models. Environmental variation continues to be a major obstacle for the resolution of genotype-phenotype associations, particularly for quantitative traits. Weather and edaphic data as well as any other important environment characteristics need to be defined, collected and evaluated. In addition, plant development and other traits need to be effectively measured in the field to correlate with image and hyperspectral data (Strategies 3.3 and 3.4) to be able to improve the accuracy and effectiveness of selection. As new high-throughput phenotyping, phenomic, and molecular phenotypic tools are developed and their costs are reduced, opportunities for improving yield prediction by incorporating these data will arise. Existing soybean resources, including the USDA core collection, the NAM population, and the Northern and Southern Uniform Soybean Trials, present excellent opportunities for defining best practices in the collection of phenotypic and environmental data, as well as developing models for yield prediction.

Anticipated Outcomes:

- Key phenotypes are determined and the best practices for each are developed.
- Key environmental data types are identified and best collection methods are defined.
- Models for applying molecular phenomes, networks, and high-throughput phenotyping for yield prediction are developed and optimized.
- Nomenclature and recorded data standards are coordinated with data management and integration facilities (see also Strategy 4).

Strategy 3.3: New agronomically relevant traits and their alleles are discovered.

Genetic diversity is a fundamental aspect underpinning genetic gains in breeding programs. The identification of novel alleles that underlie agronomically important traits is a tactical approach for continuous improvement of genetic diversity within breeding programs. Intensive characterization of a diverse core collection of soybeans holds great promise to address this strategy. Characterization can include genome (Strategy 1.2), molecular phenomes, and high-throughput phenotyping platforms integrated with a systems biology approach. Addressing this strategy will require the implementation of new computational tools and data maintenance capabilities to fully capture and utilize genotype, phenotype and other information.

Anticipated Outcomes:

- Core collection is intensively phenotyped.
- Genetic variance is increased by introducing novel alleles with positive effects for yield potential and other traits of agronomic interest.

Strategy 3.4: Breeder selection targets are improved.
Phenotypic evaluation of many traits can be extremely laborious, which impacts cost. Furthermore, phenotyping many traits can result in large variation, which influences prediction ability (*e.g.*, soybean cyst nematode, brown stem rot, sudden death syndrome, white mold). Advancements in mapping strategies, including the implementation of NAM populations, have enabled identification of genetic locations that underlie moderate- to large-effect causal polymorphisms. These genetic markers can empower breeders to rapidly genotype parental materials and resulting populations for favorable genetic architecture with high predictive value for desired outcomes. Moreover, knowledge gained on the genetic locations associated with phenotypic outcomes can be integrated with genome editing, to create beneficial alleles directly in elite germplasm, avoiding the need for introgression of specific alleles (Strategy 2.4).

Many complex traits of interest to soybean breeders, such as grain yield or stress tolerance, are poorly understood and are confounded by environmental influences. Because of this, genetic gains by traditional breeding methods have resulted in improvements, but at a slower rate than desired. Optimization of genomic selection tools is expected to increase the rate of gain for these complex traits. The Northern and Southern Uniform Soybean Trials may represent a valuable resource for the development of these tools.

Anticipated outcomes:
- Causal alleles are identified and perfect markers for key agronomic and defensive traits are developed.
- Genomic selection models for parent and progeny selection over multiple traits and environments are developed and optimized.

Goal 4: Bioinformatic infrastructure, tools and standards for leveraging multiple datatypes are in place.
SoyBase is the USDA-ARS soybean genetics and genomics database and the central location where soybean researchers can find published and user-submitted research results and data sets. A separate effort, SoyKB, has focused on developing display and analysis tools. The two teams are working together to provide the community access to these tools in the familiar SoyBase context. As part of this effort, and in conjunction with the Legume Federation Project, data storage and analysis capabilities at iPlant are also being developed.

As soybean research moves fully into the big data era, it is clear that significant effort is needed to store, manage, and integrate both raw and analyzed data so that they can be effectively utilized by the entire community. A key theme of the five year soybean strategy is to focus efforts on data related to the core soybean germplasm collection. This focus was chosen as it is likely that a number of projects will generate different data types on these soybean lines. Tools for data manipulation, analysis and integration will need to be developed In order to merge data sets and efficiently mine them. Furthermore, standards need to be developed so the community can submit data and results to SoyBase and other data repositories, as

appropriate. A major component of this effort will be to educate the soybean community about the new and existing tools and data. In addition, researchers will need to be informed about how to efficiently submit data to SoyBase, so it can be leveraged by the entire soybean research community.

Strategy 4.1: Data integration, quality, analysis, and visualization tools are improved and expanded.
A limitation in linking the genotype to phenotype is a lack of tools whereby users can intuitively explore, test and develop hypotheses using multiple data types and sources. Efforts towards linking tools and datasets between SoyBase and SoyKB have been initiated and need to be further developed. The community should leverage existing tools for integration of diverse data types, such as gene networks, protein predictions, links to literature, and phenotypes. These tools need to be user friendly for genomics researchers and plant breeders. A key issue for the community is where and how to store and access data. Different solutions may be optimal for different data types. The community needs to stay abreast of available resources and the funding situations supporting them. There is particular concern for storage of digital phenotyping data. At present, iPlant is available, but other long term permanent solutions may be needed.

Anticipated outcomes:

- Tools for integrated genotype, gene network, gene expression, phenotype, environmental and other data sets are developed and/or leveraged from other research communities.
- Research tools that promote the use of 'omics' data by plant breeders are generated.
- Networks of gene interaction data that bridge the genotype to phenotype gap are available.
- Researchers have access to permanent data storage, particularly for digital phenotyping data.
- Create awareness among users and potential users regarding the ease of access, tools and potential uses of multiple data types.

Strategy 4.2: Controlled vocabulary and metadata standards across multiple data types are generated and communicated to users.
Cross discipline research using diverse data sets will be facilitated by having consistent standards for submission, nomenclature and data formats.

Anticipated outcomes:

- A committee of cross-discipline scientists to develop metadata standards for different data types (phenotypic, genotypic, etc.).
- Controlled vocabularies for data and metadata are implemented.
- Gene by gene information is consolidated and should include: gene names, descriptions, gene models, map details, annotation, expression data, available mutants, links to similar genes in other plant species, and relevant publications.
- Templates are developed to allow researchers to continuously collect metadata throughout the life of a project, and to efficiently provide these data to SoyBase.
- Standardized methods for data submission to repositories are created and communicated to submitters. This includes instruction on how to generate data-ready publications.
- Standardized metadata to describe datasets submitted to repositories is in use.

Workshop Participants

Rajat Aggarwal, Dow AgroSciences
Charles An, USDA-ARS
Ed Anderson, North Central Soybean Research Program
Ivan Baxter, USDA-ARS
Bill Beavis, Iowa State University
Kristin Bilyeu, USDA-ARS*
Steve Cannon, USDA-ARS
Tommy Carter, USDA-ARS
Tom Clemente, University of Nebraska*
Bret Cooper, USDA-ARS
Oswald Crasta, Dow AgroSciences
Brian Diers, University of Illinois
Ann Dorrance, The Ohio State University
John Finer, The Ohio State University
Mike Gore, Cornell University
David Grant, USDA-ARS*
George Graef, University of Nebraska*
Michelle Graham, USDA-ARS*
Karen Hudson, Purdue University
Matt Hudson, University of Illinois
Scott Jackson, University of Georgia*
Jennifer Jones, United Soybean Board*
Richard Joost, United Soybean Board
Trupti Joshi, University of Missouri
Zenglu Li, University of Georgia
Aaron Lorenz, University of Minnesota
Jianxin Ma, Purdue University
Leah McHale, The Ohio State University*
Geeta Menon, Bayer CropScience
Blake Meyers, University of Delaware
Melissa Mitchum, University of Missouri*
Henry Nguyen, University of Missouri
Jack Okamuro, USDA ARS
Wayne Parrott, University of Georgia
David Pazdernik, Minnesota Soybean Growers Association
Katy Rainey, Purdue University
Sam Reddy, Dow AgroSciences
Jamie O'Rourke, USDA-ARS
Danny Singh, Iowa State University
Qijan Song, USDA-ARS
Jim Specht, University of Nebraska
Gary Stacey, University of Missouri
Robert Stupar, University of Minnesota*

Lila Vodkin, University of Illinois*
Ruth Wagner, Monsanto
Steve Whitham, Iowa State University

*Denotes Member of Writing Team

Acknowledgments
The planning workshop and the report preparation were partially funded by the United Soybean Board. The editors would like to thank the workshop participants and particularly the writing teams for their time, energy, and intellectual contributions.

References

Anderson, J.E., M.B. Kantar, T.Y. Kono, F. Fu, A.O. Stec, Q. Song, et al. 2014. A roadmap for functional structural variants in the soybean genome. G3 4: 1307-1318. doi:10.1534/g3.114.011551.

Bao, Y., T. Vuong, C. Meinhardt, P. Tiffin, R. Denny, S. Chen, et al. 2014. Potential of Association Mapping and Genomic Selection to Explore PI 88788 Derived Soybean Cyst Nematode Resistance. Plant Genome 7. doi:10.3835/plantgenome2013.11.0039.

Birney, E., T.J. Hudson, E.D. Green, C. Gunter, S. Eddy, J. Rogers, et al. 2009. Prepublication data sharing. Nature 461: 168-170. doi:10.1038/461168a.

Boerma, R., R. Wilson and E. Ready. 2011. Soybean Genomics Research Program Strategic Plan. Plant Genome 4: 1-11. doi:10.3835/plantgenome2011.12.0001.

Bolon, Y.T., A.O. Stec, J.M. Michno, J. Roessler, P.B. Bhaskar, L. Ries, et al. 2014. Genome resilience and prevalence of segmental duplications following fast neutron irradiation of soybean. Genetics 198: 967-981. doi:10.1534/genetics.114.170340.

Carrero-Colón, M., N. Abshire, D. Sweeney, E. Gaskin, K. Hudson. 2014. Mutations in SACPD-C result in a range of elevated stearic acid concentration in soybean seed. PLoS One 9:e97891. doi:10.1371/journal.pone.0097891.

Chung, W.H., N. Jeong, J. Kim, W.K. Lee, Y.G. Lee, S.H. Lee, et al. 2014. Population structure and domestication revealed by high-depth resequencing of Korean cultivated and wild soybean genomes. DNA Res. 21: 153-167. doi:10.1093/dnares/dst047.

Cui, Y., S. Barampuram, M.G. Stacey, C.N. Hancock, S. Findley, M. Mathieu, et al. 2013. Tnt1 retrotransposon mutagenesis: a tool for soybean functional genomics. Plant Physiol. 161: 36-47. doi:10.1104/pp.112.205369.

Gillman, J.D., M.G. Stacey, Y. Cui, H.R. Berg and G. Stacey. 2014. Deletions of the SACPD-C locus elevate seed stearic acid levels but also result in fatty acid and morphological alterations in nitrogen fixing nodules. BMC Plant Biol. 14: 143. doi:10.1186/1471-2229-14-143.

Hancock, C.N., F. Zhang, K. Floyd, A.O. Richardson, P. Lafayette, D. Tucker, et al. 2011. The rice miniature inverted repeat transposable element mPing is an effective insertional mutagen in soybean. Plant Physiol. 157: 552-562. doi:10.1104/pp.111.181206.

Haun, W., A. Coffman, B.M. Clasen, Z.L. Demorest, A. Lowy, E. Ray, et al. 2014. Improved soybean oil quality by targeted mutagenesis of the fatty acid desaturase 2 gene family. Plant Biotechnol. J. 12: 934-940. doi:10.1111/pbi.12201.

Hymowitz, T. and J.R. Harlan. 1983. Introduction of soybean to North America by Samuel Bowen in 1765. Econ. Bot. 37: 371-379. doi:10.1007/BF02904196.

Jacobs, T.B., P.R. LaFayette, R.J. Schmitz and W.A. Parrott. 2015. Targeted genome modifications in soybean with CRISPR/Cas9. BMC Biotechnol. 15: 16. doi:10.1186/s12896-015-0131-2.

Jarquin, D., K. Kocak, L. Posadas, K. Hyma, J. Jedlicka, G. Graef, et al. 2014. Genotyping by sequencing for genomic prediction in a soybean breeding population. BMC Genomics 15: 740. doi:10.1186/1471-2164-15-740.

Joshi, T., B. Valliyodan, J.H. Wu, S.H. Lee, D. Xu and H.T. Nguyen. 2013. Genomic differences between cultivated soybean, G. max and its wild relative G. soja. BMC Genomics 14 Suppl 1: S5. doi:10.1186/1471-2164-14-s1-s5.

Lam, H.M., X. Xu, X. Liu, W. Chen, G. Yang, F.L. Wong, et al. 2010. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat. Genet. 42: 1053-1059. doi:10.1038/ng.715.

Li, Y.H., S.C. Zhao, J.X. Ma, D. Li, L. Yan, J. Li, et al. 2013. Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. BMC Genomics 14: 579. doi:10.1186/1471-2164-14-579.

Li, Y.H., G. Zhou, J. Ma, W. Jiang, L.G. Jin, Z. Zhang, et al. 2014. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat. Biotechnol. 32: 1045-1052. doi:10.1038/nbt.2979.

Liu, S., P. K. Kandoth, S. D. Warren, G. Yeckel, R. Heinz, J. Alden, et al. 2012. A soybean cyst nematode resistance gene points to a new mechanism of plant resistance to pathogens. Nature 492: 256-260. doi:10.1038/nature11651.

Mathieu, M., E.K. Winters, F. Kong, J. Wan, S. Wang, H. Eckert, et al. 2009. Establishment of a soybean (Glycine max Merr. L) transposon-based mutagenesis repository. Planta 229: 279-289. doi:10.1007/s00425-008-0827-9.

Raval, J., J. Baumbach, A.R. Ollhoff, R.N. Pudake, R.G. Palmer, M.K. Bhattacharyya, et al. 2013. A candidate male-fertility female-fertility gene tagged by the soybean endogenous transposon, Tgm9. Funct. Integr. Genomics 13: 67-73. doi:10.1007/s10142-012-0304-1.

Schmutz, J., S.B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, et al. 2010. Genome sequence of the palaeopolyploid soybean. Nature 463: 178-183. doi:10.1038/nature08670.

Song, Q., D.L. Hyten, G. Jia, C.V. Quigley, E.W. Fickus, R.L. Nelson, et al. 2013. Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. PloS One 8: e54985. doi:10.1371/journal.pone.0054985.

Song, Q., D.L. Hyten, G. Jia, C.V. Quigley, E.W. Fickus, R.L. Nelson, et al. 2015. Fingerprinting Soybean Germplasm and Its Utility in Genomic Research. G3 5: 1999-2006. doi:10.1534/g3.115.019000.

Song, Q., J. Jenkins, G. Jia, D.L. Hyten, V. Pantalone, S.A. Jackson, J. Schmutz, P.B. Cregan. 2016. Construction of high resolution genetic linkage maps to improve the soybean genome sequence assembly Glyma1.01. BMC Genomics 17: 33. doi: 10.1186/s12864-015-2344-0.

Xia, Z., S. Watanabe, T. Yamada, Y. Tsubokura, H. Nakashima, H. Zhai, et al. 2012. Positional cloning and characterization reveal the molecular basis for soybean maturity locus E1 that regulates photoperiodic flowering. Proc. Natl. Acad. Sci. U.S.A. 109: E2155–E2164. doi:10.1073/pnas.1117982109.

Zhou, Z., Y. Jiang, Z. Wang, Z. Gou, J. Lyu, W. Li, et al. 2015. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat. Biotechnol. 33: 408-414. doi:10.1038/nbt.3096.