Final project report (Year 2; FY18)- McHale

February 12, 2019

INCREASING THE RATE OF GENETIC GAIN FOR YIELD IN SOYBEAN BREEDING PROGRAMS

OBJECTIVE 1: Increasing selection intensity and decreasing non-genetic sources of variability through improved progeny row testing

Objective 1: Deliverables.

- Observed data, selection information, pedigree and plot layout (range-row information), and shipment of seed for planting for breeders at 11 locations has been completed with some modification to our original plan. We developed statistical models and in some cases processed from imagery into data. We provided a report of selection categories that the breeders used to develop preliminary yield trials that test our objectives.
- The Rainey lab has constructed, coordinated, and implemented an overall data management plan and preliminary analytical pipeline. This has been completed with some modification of the original plan to account for the breeders' needs as related to preliminary yield testing (see KPIs below).

Objective 1: Key Performance Indicators or performance measures (year 2).

- Additional data are collected on progeny rows.
  - Since 2017, additional phenotypes have been collected on 80,000 progeny rows across 10 breeding programs.
- A list of all breeders' lines ranked simultaneously for yield breeding value, maturity prediction and a metric of diversity.
  - This KPI evolved because the breeders felt there was more value in having the full set of progeny rows from a single breeding program tested within that state, rather than progeny rows from each breeder distributed across multiple state in a cooperative preliminary yield testing. So, we worked closely with each breeder to assess their interests and needs in testing new progeny row selection strategies. Each breeding program has a different set of assets and resources. Some breeders are interested in how to incorporate drone data or spatial analysis, or the value for pedigree selection, while others want to find a replacement for yield because it is too labor intensive, etc. We implemented selection models that were unique to each program. Each breeders received a detailed report of the outputs of various selection models and ranked lists for each model. For 2019 selections, this has been completed for everyone except Lorenz, who has a slightly different approach.
- Cooperative preliminary yield trials are organized to test selection accuracy.
  - As above, this KPI evolved as a result of breeders' needs. In the 2018 season preliminary yield trials (PYTs) were organized by each breeder and thousands of additional yield plots were planted. This will happen again in 2019 for a second year of testing of selection models. We have done preliminary analysis of accuracy of selection models in the 2018 Rainey lab PYTs, where pedigree information improved accuracy. Compiling PYT data across programs from 2018 still needs to happen.

OBJECTIVE 2: Increasing selection coefficient and decreasing length of breeding cycle through genomic selection

Objective 2: Deliverables.

- A set of 1600 lines has been genotyped (6K SNPs) and phenotyped through the Uniform Soybean Tests. Phenotype data from 1992 -- 2017 has been cleaned and compiled and made available to SoyBase. We are currently working with SoyBase to iron out a few details on the data before it is posted. It will first be posted privately, and then once everyone approves, we will make it publicly available.

Objective 2: Key Performance Indicators or performance measures met thus far (year 2).

- GBS method developed that can genotype 200-1000 markers with less than 10% missing data and greater than 95% accuracy.
    - Current GBS methods meet this KPI except with a very limited DNA extraction method and in limited throughput.
    - In addition, methods and results peer reviewed and published (Wickland et al., 2017; BMC Bioinformatics 18, 849). The 10% missing data fraction is a function of the population size. For all studied populations, we met the < 10% missing data target using existing software (TASSEL). We also developed a new software approach that greatly outperforms this existing software if 1) a greater missing data fraction is allowed and / or 2) our recommended association population experimental design is followed. The new software (GB-eaSy workflow) achieved 1,352 SNPs with 10% missing data and > 99.5% accuracy in an association population panel.
- Demonstrated ability to leverage historical URT data for making genomic predictions in soybean.
    - This has been achieved in a preliminary manner. Prediction accuracies from the training set compiled appear to be good, but more work is needed to verify these.

OBJECTIVE 3: Increasing additive genetic variance

Objective 3: Deliverables

- Collected and recorded information from all participating breeders (5) on success of parent combinations.
- High-quality, multi-environment yield and other agronomic performance data for 500 PIs in the USDA Soybean Germplasm Collection. High-yielding PIs with unique yield genes will be used in public and private breeding programs to increase yield.
- Identify yield-marker genotype relationships based on association mapping results from the extensive, high-quality yield dataset. Information from the 2015-2016 2-year, multi-location yield tests and the 50K SNP genotype data will be used to identify genomic regions, or haplotype blocks, that are associated with yield in soybean. This information will be used in public and private breeding programs to increase yield, rate of genetic gain, and genetic diversity of the commercial soybean germplasm pool.
- Develop predictive model(s) that allow selection of superior high-yield genotypes from the USDA germplasm collection. From each sampling group (SSD, CLU, and RAN), as well as for the group of 500 PIs overall, we will develop predictive models to allow us to go back into the germplasm collection and select untested lines based on genotype. Validation of the models with yield and other phenotype testing will be a follow-up project.

- Public use of data, documentation of results. Results will be published in refereed scientific journals. Data from all tests will be made available to all users through SoyBase and possibly GRIN. Details will be worked out with USDA and SoyBase administrators to facilitate availability and use.
- Candidate yield-conferring haplotypes from exotic germplasm.
- High-quality high throughput SNV and structural variants matrix for WGS panel.
- A list of candidate genomic regions and/or haplotypes associated with yield-related traits.

Objective 3: Key Performance Indicators or performance measures (year 2).

- High quality yield and seed composition data on 500 PIs from the USDA Soybean Germplasm Collection from 14 environments, 7 environments in each of 2 years.
  - This is done and the results were shared with all cooperators in March 2017.
- Preliminary model to predict yield and seed composition on PIs from the USDA Soybean Germplasm Collection.
  - This is done and all the predictions with all the models that were tested were shared with cooperators March 2017.
- One or more potential yield-conferring haplotypes identified from exotic sources used to select parent lines for yield improvement.
  - We have identified nine potential yield-conferring haplotypes derived from the alternative gene pool (see Objective 3, Approach #3, Task 4).
  - We have identified haplotypes underlying 138 known QTLs included in SoyBase. Among them, 46 QTLs are associated with Seed Composition and Yield.
  - We have identified other 158 regions under selection that may represent QTLs still not genetically identified.
- Tentative identification of lines derived from wild soybean that can be used as parents in variety development programs.
  - This KPI has been met in part through the identified candidate segments from *G. soja* associated with high yield, and developed molecular markers that can be used to integrate these segments into elite varieties for enhanced yield potential.
  - Lines containing the potential yield haplotypes described above have been identified, as have markers defining the haplotype blocks.
- Significant selection signals associated with yield identified from WGS potentially used in variety development programs.
  - Several selection signals have been identified and are being pursued for their effect on yield (see Objective 2, Task 6 and Objective 3, Approach #3, Task 4).

OBJECTIVE 4: Development of a metric to estimate genetic gains on an annual basis

Objective 4: deliverables

- Danielle Dykema and Haley Trumpy developed an educational video describing plant breeder's and soybean farmer's perceptions of genetic gain. The video was presented to the NCSRP annual meeting in Fargo. The video was provided to the NCSRP for purposes of hosting it at their website. A second video on how to estimate realized genetic gains for varieties grown on the farm was requested.
- A commercial soybean breeding program has provided yield data from uniform field trials of three varietal development stages conducted from 2009 to 2017 for maturity groups 0, I, II, III and IV.

- We established a potential range of resources used in field trials for soybean variety development programs. The number of locations per year used by a commercial soybean breeding program for first stage of field trials is about 40, 75 for the second stage and 100 for the third stage. The number of locations per year for the two stages of public URT's is much smaller.
- Simulation software has been used to investigate genetic gains using six methods for Genomic Selection across 20 cycles in a founding population consisting of genomes of the SoyNAM founders. Software is robust and runs in a virtual environment with any arbitrary number of compute nodes and clusters of nodes. The user interface is R, while several simulation functions are written in C++. The R interface requires knowledge of statistical genetics. We are not planning to develop a user friendly interface; user friendly interfaces for statistical genetic programs will require significant investments. For example, QuGene has invested ~ 10M $US and the system is not capable of simulating the many nuances of soybean breeding projects.
- Field trial phenotypic data and genotypic data from a commercial organization has been obtained and aggregated.
- EM algorithm for calculating realized genetic gains was published in Interfaces. This algorithm removes non-genetic variability from annual yield trials that consist of new sets of lines every year.
- A mixed linear model in which environments are treated as random effects and repeated lines are used as "markers" to determine a realized relationship matrix among environments represents a novel method for estimating and removing (annual) environmental effects from field trial data to improve calculation of realized genetic gains.

Objective 4: Key Performance Indicators or performance measures (year 2):

- One short video on genetic gain has been completed and delivered to the NCSRP web site.
- Simulation software, implemented in R, to simulate yields of potential varieties has been developed.
- Aggregation of phenotypic data from large scale field trials
    - Phenotypic data resources from public URTs have been aggregated and have been shared with project members.
    - Phenotypic data resources from commercial organization has been aggregated with genotypic data.

DISSEMINATION OF PROJECT FINDINGS

Presentations related to project work (FY18):

dos Santos LB, Wei W, Viana JPG, Wu X, de Souza AP, Hudson M, Clough SJ, The Plant & Animal Genome XXVII conference, San Diego, CA, "Validation of potential genetic introgression introduced by intersubgeneric hybridization between *Glycine max* and *G. tomentella*" International. (January, 2019)

Hyten Jr, D, VIII Brazilian Soybean Conference, Brazilian Soybean Congress, Goiânia, Goiás State, Brazil, "Soybean Breeding – New tools, challenges and future", International, Invited. (June, 2018).

Hyten Jr, D, PSI monthly research meeting, PSI, Lincoln, NE, "Targeted Soybean SNP Genotyping", Local, Invited. (May, 2018).

Hyten Jr, D, University of Guelph Department of Plant Agriculture seminar, University of Guelph, Guelph, Canada, "Development of High-throughput SNP Genotyping Technologies for Soybean", International, Invited. (May, 2018).

Hyten Jr, D, 2018 Soybean Breeder's Workshop, St. Louis, MO, "Development of sequencing for genotyping within soybean breeding programs", National, Invited. (February, 2018).

Hyten Jr, D, Wang, H. , Happ, M. , McConaughy, S, Curtolo, M, Amundsen, K, Posadas, L, Lorenz, A, Song, Q, Graef, G, The Plant & Animal Genome XXVI conference, San Diego, CA, "Development of Sequencing for Genotyping within Soybean Breeding Programs", International, Invited. (January, 2018).

McHale L, Lorenz A, Graef G, Martin Rainey K, Beavis B, Hyten D, Hudson M, Clough S, Ma J, Campbell B, Chen P, Diers B, Scaboo A, Schapaugh W, Singh A, Wang D, 17th Biennial Molecular & Cellular Biology of the Soybean Conference, Athens, Georgia, “Increasing the rate of genetic gain for yield in soybean breeding programs”, International, Invited. (August, 2018).

Publications related to project work (FY18):

Swarm SA, Sun L, Wang X, Wang W, Brown PJ, Ma J, Nelson RL (2019) Genetic dissection of domestication-related traits in soybean through genotyping-by-sequencing of two interspecific mapping populations. Theoretical and Applied Genetics. doi: 10.1007/s00122-018-3272-6.

Wang X, Chen L, Ma J (2019) Genomic introgression through interspecific hybridization counteracts genetic bottleneck during soybean domestication. Genome Biology. 20:22.

Wickland DP, Battu G, Hudson KA, Diers BW, Hudson ME (2017) A comparison of genotyping-by-sequencing analysis methods on low-coverage crop datasets shows advantages of a new workflow, GB-eaSy. BMC bioinformatics. 18:586.

Zhang D, Sun L, Li S, Wang W, Ding Y, Swarm SA, Li L, Wang X, Tang X, Zhang Z, Tian Z, Brown PJ, Cai C, Nelson RL, Ma J (2018) Elevation of soybean seed oil content through selection for seed coat shininess. Nature Plants. 4:30-35.