Project report, semi-annual

April 5, 2018

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

In the past project period, we conducted selections on progeny rows for several breeding programs using that incorporate pedigree selection, spatial adjustment, and UAS canopy or other phenotypes. We processed UAS imagery for canopy cover and color for Schapaugh, Wang, Lorenz, and Rainey, and we attempted it for Scaboo.

1.e. Deliverables.

- We have reported observed data, reported breeder selection information, and reported pedigree and plot layout (range-row information) for breeders at 10 locations.
- Implementation of selection models unique to each breeders' needs has been achieved.

1.f. Key Performance Indicators or performance measures (year 3).

- All breeding programs reporting (10) have collected additional phenotypic data for selection on progeny rows for two years.
- Selections have been completed before harvest for programs electing to use canopy coverage for selection.
- Preliminary yield trials organized by each breeder to test selection accuracy for the 10 breeding programs for 2018 and ongoing for 2019.
- A selection accuracy assessed from 3 years of data, and preparation of manuscripts and reports are pending completion of 2019 season.

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

The Lorenz lab has continued to curate phenotypic and available meta-data on Uniform Northern Regional trials ranging from 1993 – 2017. These data, and associated genotypic data on ~1700 lines we could genotype, were made available to SoyBase for making publicly available. The data were sent in final form to SoyBase on March 12, 2019. As SoyBase works towards making these data publicly available, we are working on a manuscript announcing the availability of these at. The manuscript will include preliminary analyses on population structure, genomic distribution of allelic variation within and between breeding programs/maturity groups, association analyses using historically collected phenotypes, and initial assessments of prediction accuracy using the URT data. A. Lorenz gave a presentation at the 2019 Soybean Breeder's Workshop describing this resource.

DNA extracted from 1500 UMN breeding lines in prelim yield trials was sent to the Hyten Lab at University of Nebraska for genotyping using 1000 SNPs selected from URT genotype dataset described above. We are currently awaiting the return of this genotype data.

Beyond the genomic selection work by leveraging the URT data, we have also made use of this data for performing an association analysis on IDC tolerance. We detected a strong association on chr 5. Results from this analysis were combined with other fine mapping work, and a manuscript is currently in review at Plant Genome.

The Hyten lab has tested a new method of extracting DNA using a nanoparticle DNA preparation kit. We were able to synthesize the nanoparticles and were successful at extracting DNA using this method. We tested whether the MIPs protocol would work with DNA extracted from leaf and seed using the nanoparticle method and compared that to the CTAB method. The samples were sequenced at 120x coverage. The average percentage of reads mapped to reference falls in between 71.83-81.14 % among five tests with leaf and seed DNA prepared with CTAB or the new nanoparticle method tested at different reaction volumes. Two samples at a higher reaction volume using nanoparticle DNA extracted from seed had lower mapping percentage at 56.96 and 44.14%. With the results of the nanoparticles from leafs and some of the seed samples being comparable to CTAB, we will expand our testing to a larger number of samples to determine the repeatability of this method and how much automation we can incorporate in this simpler protocol. Since we are able to synthesize the nanoparticles ourselves the cost of total reagents for the DNA extraction is currently estimated at $0.37/sample. We also plan to test if we can use the nanoparticles to normalize our DNA concentration to further reduce the cost of having to quantify the DNA after extraction.

We have been running the MIPs protocol to 384 plates, this has helped to increase throughput and reduce reaction volumes. Currently, our estimated cost of the MIPs protocol is down to $5.68/sample when we multiplex 1000 samples in a single sequencing run. Currently, we have run up to 600 samples on a single run. Next we plan to start running 1000 samples to test how well we can pool samples at this high multiplexing level. We have successfully tested some of the reagents at lower volumes using the CTAB DNA extraction method which brings the cost to under $5/sample but we have to confirm that these volumes work with the new nanoparticle DNA extraction method before confirming we have hit that price milestone.

We have ran the 1k probe set on two NAM populations to help access our accuracy of calling hets using the MIPs protocol. The data produced looks good and is currently being analyzed for SNP calling.

2.e. Deliverables.

- A community resource for genomic prediction consisting of a set of soybean lines that can be used to establish genomic prediction to help expedite genetic gain for yield.
    - Initial dataset has been delivered to SoyBase for posting. A manuscript is in preparation describing this dataset. We are continuing to build on this database by collecting and genotyping the 2018 URT entries and 2019 URT entries.
- Novel inexpensive and rapid genotyping methods that can be used for genomic prediction and selection.
- With the cost of the DNA extraction and the cost of reagents for the MIPs reaction and the sequencing we are at a total of $6.05 per sample.

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

- Demonstrated ability to leverage historical URT data for making genomic predictions in soybean.
    - As mentioned in last report, we have shown this using cross validation within the URT dataset. We are collecting genotype data on entries in prelim yield trials. Once this is collected, we will assess our ability to predict new breeding lines.
- GBS method developed that can genotype 200-1000 markers with less than 10% missing data and greater than 95% accuracy.

OBJECTIVE 3: Increasing additive genetic variance

We obtained multi-sensor data on 250 accessions in field tests at locations in NE, KS, IA, and MO representing Maturity Groups I, II, III, and IV. This was kind of a test run during 2018 to prepare for our planned image and sensor data collection from all North Central locations during the 2019 season. Based on our experience from the 2018 plots, we are constructing a pheno cart specifically for the plot layout of the NCSRP experiments so our data collection will be efficient and our processing time reduced. We are still running protein and oil on seed samples from all plots received from all cooperators for the 2018 tests. That should be completed in April. We are completing preparation of samples for 2019 planting and finalizing plans with cooperators.

A few of our *G. max* x *G. tomentella*, and of our *G. max* x *G. soja* derived lines, continued to show good yield potential in 2018 field tests, with some material being the largest yielders across several testings. We determined that some of the *G. max* x *G. soja* derived lines have 18-24% of the DNA of *G. soja*, but we have yet to be able to confidently determine if any *G. tomentella* nuclear DNA is present in the *G. max* x *G. tomentalla* derived lines.

The Ma lab has analyzed the haplotype of genomic regions surrounding previously identified maturity genes *E1*, *E2*, *E3*, *E4*, *E9*, and *J*, as well as genes associated with branching angle and canopy coverage in ~500 re-sequenced soybean germplasm accessions and are in the process of designing PCR-based molecular markers for precise evaluation of the 240 PIs included in this project. We have made crosses between two high-yielding RIL lines derived from *G. soja* with two elite lines towards dissection of the yield QTLs from *G. soja* that have been mapped on chromosomes 8 and 11. We have been in the process of mapping yield QTLs using 225 *G. tomentella*-derived lines with phenotypic data collected from multiple locations in collaboration with Randy Nelson. We are continuing functional validation of domestication-related traits including growth habit, leaf shape and size by transformation. Two constructs have been under transformation via outside service and one construct has been transformed to Williams 82 by the Ma lab.

3.e. Deliverables.

- **High-quality, multi-environment yield and other agronomic performance data for 500 PIs in the USDA Soybean Germplasm Collection**. Some cooperators in this test have used some of the PIs in their current diversity program crosses.
- **Develop predictive model(s) that allow selection of superior high-yield genotypes from the USDA germplasm collection**. Models were developed using each sampling group alone as well as the complete dataset, and cross-validation was used to test effectiveness. Genotype information improved the models, but including GxE effects added little to the prediction. Model results and predictions were shared with all cooperators.
- **Incorporate high-throughput phenotype data, plant developmental data, and environment data in the models.** This was funded for the 2019 tests, so we will complete this during the coming season. We have multi-sensor and environment data from four locations of the tests during 2018, but plan to collect data from all NCSRP locations during 2019.
- High yielding lines derived from wild soybean and G. *tomentella.* In 5 field tests, at least 2 *G. tomentella* derived lines were the top yielding. In 3 tests, at least one of the top yielding was a *G. soja* derived line. For example, in one test, there were 3 *G. soja*-derived lines that were between 2 and 9 bushels/acre higher yielding than the best check and each was derived from a different *G. soja* parent.
- A list of candidate genomic regions and/or haplotypes associated with yield-related traits have been identified by QTL mapping and analysis of genome resequencing data.
- A set of molecular markers that can be used for traits selection and evaluation have been designed and are now under validation.

3.f. Key Performance Indicators or performance measures (year 3).

- 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. We are working on final compilation of all the data and analyses. Data quality overall is good.
- Preliminary model to predict yield and seed composition on PIs from the USDA Soybean Germplasm Collection. One or more potential yield-conferring haplotypes identified from exotic sources used to select parent lines for yield improvement. Models were developed using each sampling group alone as well as the complete dataset, and cross-validation was used to test effectiveness. Genotype information improved the models, but including GxE effects added little to the prediction. Model results and predictions were shared with all cooperators.
- Tentative identification of lines derived from wild soybean that can be used as parents in variety development programs. Yield testing show that this is possible (see above).
- Determination of introgression of *G. tomentella* DNA in lines derived from *G. max* x *G. tomentella*. All evidence so far, based on thorough sequence analyses, indicates that the G. max x *G. tomentella* 2n=40 derived line 12ST4-5 has no *G. tomentella* nuclear DNA intrograted in its genome. We still have more analyses to perform before being more conclusive, and these analyses should be completed before this grant expires.
- Molecular markers for precise tagging of five maturity genes and a newly identified branching angle/canopy coverage genes.

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

Matheus Dalsente-Krause completed a literature review on methods and metrics for evaluating genetic gains on an annual basis. Other than the infrequent direct comparisons of historical varieties in common gardens, a.k.a. decade studies, there are no more than five published methods for evaluating genetic gains from annual field trials and all of these recognize the difficult problem of removing non-genetic (agro-environmental) sources of variability. Of these, only two have proposed using genotypes from adjacent years to adjust for annual non-genetic sources of variability. In addition to the EM algorithm that used check varieties to estimate non-genetic sources of variability in staged commercial field trials that we published two years ago, we developed a mixed model approach in which common entries among years are used to obtain shrunken (BLUP) values for non-genetic environment year combinations that will be used as non-genetic covariates in assessing genetic gains for yield using data from the uniform trials.

4.e. Deliverables.

- One 10 minute video describing what genetic gain is has been developed and delivered to NCSRP. A second video describing how it can be evaluated with on-farm resources is being developed.
- Sources of bias in evaluating genetic and non-genetic sources of variability have been identified in both direct comparison and long term field trials. The latter are being addressed with the UT data from commercial and public plant breeders.
- A recent variety development simulation tool known as aphasim-R was developed by John Hickey's group in Edinburgh. It was released through the usual CRAN-R sites. It is being evaluated.

3.f. Key Performance Indicators or performance measures (year 3).

- One 10 minute video describing what genetic gain is has been developed and delivered to NCSRP.