Project report

October 16, 2018

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

Executive summary:

In the second year of our objective 1, aimed at improving progeny row testing, we are evaluating the selections from progeny row testing made in year 1 as well as carrying out a second year of the methods of progeny row testing from year 1. Preliminary results are pending the report of yield from this harvest.

Objective 2, aimed at decreasing the time for each cycle of selection by implementing a genomic selection strategy using public resources is making good progress. The genomic selection study from the SoyNAM population has been completed and in the final stages of manuscript preparation. Databasing and genotyping from the Uniform Regional Trials is continuing on an ongoing basis as yearly entries are submitted and yield data is collected. Preliminary tests of predictive ability using the URT training set has been exploring using cross validation on adjusted means of URT lines. Further refinement of the analysis indicates respectable prediction accuracies ranging from 0.50 to 0.68 across maturity groups. Finally, a cost effective genotyping strategy has been developed.

Objective 3 aims to implement several strategies for increasing genetic diversity in breeding programs. During the 2018 season, we conducted the first year of our 2-year evaluations of the 250 PI accession selected for the validation set from across the distribution of predictions for yield of ~9,700 untested PI accessions. As this is likely the only time that such a collection of PIs will be evaluated in replicated trials across many environments, we are taking the opportunity to collect as much phenotypic data as reasonable. In addition, this objective has successfully taken several approaches to identify potential yield loci through QTL analysis, GWAS, as well identification of signatures of selection through population diversity statistics.

Finally, in Objective 4, an introductory video about genetic gain has been developed and delivered to the NCSRP. Several steps have been made in an effort to develop metrics to accurately estimate realized genetic gains. Syngenta has delivered phenotypic data (yield, maturity, planting date, longitude x latitude) for lines grown in annual field trials for maturity groups II, III and IV for years 2009 – 2017. Software has been written that will merge genotypic and phenotypic data from the URT's. For the actual development of methods, a recent PNAS publication (Li et al, 2018) provides a novel method for removing the GxE contribution from the non-genetic (environmental) effects, thus leaving only genetic and an indexed genotypic value for specific environments as a means to calculate realized genetic gains.

Details for each objective and task are in the report attached below.

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

Task 1: Collection of additional data in all progeny rows.

Additional data were collected progeny rows for all programs except Cianzio, as summarized below. These data will be used in selection models.

Cooperator	Location	Yield	R8	Additional Data	Preharvest Selection	Canopy Coverage Selection
Silvia Cianzio	Iowa State	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Asheesh Singh	Iowa State	yes	?	?	no	no
Brian Diers	U of Illinois	no	yes	Ground Images	Yes, finished	yes
George Graef	U of Nebraska	no	yes	R1-R8	Yes, waiting for breeder selections	no
Aaron Lorenz	U of Minnesota	selected	yes	Aerial Images	no	yes
Leah McHale	Ohio State	yes	no	yield	no	no
Katy Rainey	Purdue	yes	yes	Aerial Images	Back-up, working on imagery	yes
Andrew Scaboo	U of Missouri	no	?	Aerial Images	Maybe, imagery is poor-quality	yes
Pengyin Chen	U of Missouri	no	yes	NDVI	Yes, working on it	no
Bill Schapaugh	Kansas State	no	no	Aerial Images	Yes, finished	yes
Dechun Wang	Michigan State	yes	?	Aerial Images	Back-up, working on imagery	yes

Task 2: Selections from progeny rows.

At Purdue, we have calculated canopy coverage values from ground or aerial photos for Diers, Schapaugh, Scaboo, Wang and Rainey. We will calculate canopy coverage for Lorenz after harvest.

At Purdue, Meng Huang has selected, or is in the process of selecting, progeny rows before harvest for Diers, Graef, Scaboo, Chen, and Schapaugh. We are also providing pre-harvest selections for Wang and Rainey as a contingency plan in case weather or equipment problems prevent harvesting all rows for yield. All pre-harvest selections include a "random" selection category as a control.

We will make post-harvest selections that include yield for Wang, Rainey, McHale, and Singh. We will make post-harvest selections that include canopy coverage data for Wang, Rainey, and Lorenz.

Lorenz adopted a modified approach of visually discarding bad rows, harvesting the remainder for yield, and then applying a post-harvest selection based on aerial canopy coverage and seed quality. He will harvest random selections we provided as a control.

We have been providing each breeder with a report listing models used, overlap between models, correlation of adjusted values with raw phenotypes, phenotype distributions, and a map of distribution of selected rows in the field.

Task 3: Preliminary yield trials to evaluate the increase in the rate of genetic gains.

In 2018, most breeders planted preliminary yield trials that included many more lines to test selection categories (McHale, Rainey, Sing, Schapaugh, Wang). Those are being harvested and we do not have results yet, nor have I requested an update from each breeder, at this time.

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

- Additional data are collected on progeny rows.
  - This KPI has been met for 2017 and 2018, with much better compliance in 2018.
- A list of all breeders' lines ranked simultaneously for yield breeding value, maturity prediction and a metric of diversity.
  - Proposal plans have been adjusted from overall selection and cooperative testing to within-program selections and preliminary yield testing, making this KPI irrelevant.
- Cooperative preliminary yield trials are organized to test selection accuracy.
  - This KPI was met for 2018.

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

Task 1. Complete study on genomic selection using the nested association mapping of soybean population (SoyNAM) and apply findings to ongoing genomic selection effort.

Manuscript preparation is in the final stages. A draft of a manuscript has been developed. This has been held up by our recent decision to include some calculations of expected accuracy using various combinations of the NAM training population to investigate how well these formulas work for predicting the prediction accuracy before phenotypes have been collected. This will also tell us if our empirical results on combining training sets to predict particular validation sets fall in line with known theory, or if there are exceptions we should investigate further.

Task 2. Compile existing phenotypic, genotypic, pedigree, and environmental data (weather, soil) from various projects on yield and diversity conducted in the North Central Region.

The only thing to report on this task since the last report is that the 2017 URT data has been added to the database. Also, we have discussed making this data directly available on SoyBase with David Grant. Since the last report, they have developed the capability of storing these data in a relational framework and are eager to host these data. This work is in progress. We expect to have this uploaded to SoyBase and available to the wider community by the next reporting period.

Task 3. Genotype all available soybean lines grown in the USDA Northern Uniform Tests beginning in 2004.

The seeds from 2018 URTs have been added to the genotyping queue. DNA from the 168 lines from 2018 have been extracted and are in process of being genotyped. A total of 1712 lines have been genotyped, excluding the 2018 lines.

Task 4. Unify genotypic data collected from the multiple platforms, using a single flexible data management system, capable of adapting to any genotyping platform.

Task 5. Development of ultra-cheap low-density marker system for genomic prediction applications.

We have worked with Aaron Lorenz to select a reduced panel of 1000 SNPs that capture a large proportion of the genome. Through testing how different selection methods select SNPs to capture diversity we determined a Bin weighted PIC method performed the best. We ordered the 1,000 SNPs as individual synthesis and tested the probe set on diverse PI lines and a segregating population. The probe set worked as expected producing high quality data with a very low missing data rate.

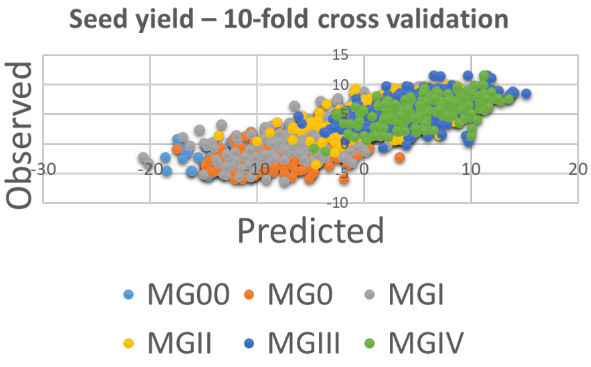
We have also worked on adapting the protocol to 384-well plates. This has helped reduce the volume of the reaction and will help increase throughput. We have grown and extracted DNA from two segregating populations, which have been previously genotyped with the 6k NAM SNP chip. We will run the 1,000 SNP probe set on these two segregating populations using 384-well plates to test the increased throughput protocol. This will also allow us to test the method for how accurate it calls heterozygotes.

Task 6. Testing of genomic selection within individual breeding programs.

Preliminary tests of predictive ability using the URT training set has been exploring using cross validation on adjusted means of URT lines. Further refinement of the analysis indicates respectable prediction accuracies ranging from 0.50 to 0.68 across maturity groups.

MG	Pred. Ability*
00	0.51
0	0.50
I	0.68
II	0.58
III	0.57
IV	0.62

* All MG Pred. Ability significantly > 0 (P>0.05)



The very early stages of testing these models in actual breeding programs using earlier generation lines has been initiate by collecting 10 representative seeds of ~1500 UMN lines submitted to prelim yield trials in 2018. DNA will be extracted from these and sent to David Hyten for genotyping using the method he is developing as part of this project (A personal communication with him suggested his method is ready for “production mode”). We will also identify parents of these lines and genotype with 6K SNPs to impute to the progenies and match the genotype data in the URT training set.

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

- GBS method developed that can genotype 200-1000 markers with less than 10% missing data and greater than 95% accuracy.
    - Performing GBS using the molecular inversion probe set method was successful in meeting this KPI when ran with 1,000 markers on diverse PI lines.
- Demonstrated ability to leverage historical URT data for making genomic predictions in soybean.
    - This has been met as described above.

OBJECTIVE 3: Increasing additive genetic variance

Approach #1: Increasing diversity through collaborative intermating across breeding programs (Task 1).

Task 1: Exploration of a retrospective analysis

Approach #2: Mining exotic soybean accessions for favorable alleles that increase yield in the north central region (Tasks 2, 3, and 4).

Task 2: Evaluation of germplasm mined from the USDA Soybean Germplasm Collection using genomic prediction

During the 2018 season, we conducted the first year of our 2-year evaluations of the 250 PI accession selected for the validation set from across the distribution of predictions for yield of ~9,700 untested PI accessions. Cooperators in 11 states grew MGI, MG2, MG3, and MG4 entries in a total of 8 environments for each Maturity Group. 2 replications per environment in an augmented incomplete block design. As part of the characterization of the validation set, we are collecting additional crop developmental data throughout the season as well as image and other multi-spectral data on these diverse genotypes that represent all of the genetic diversity in the USDA Soybean Germplasm Collection for MGI to IV. During the 2018 growing season we collected image data at two time points on all MG plots grown in NE, at one time point for MG2 and MG3 plots in IA, and MG3 and MG4 plots in KS and MO. This is in addition to developmental data on all lines at all locations for R1 (first flower), R3 (beginning pod), R5 (beginning seed fill), and R8 (maturity) that were to be collected by all cooperators for all tests/locations. We also had weather stations at the Lincoln and Mead NE locations where all four MG tests were grown, and we will obtain weather information from selected other locations where that is available for the 2018 season (MN, OH, KS, ?). After harvest, we will also have yield data and seed composition and seed weight data. It will take the first 6 months of 2019 to receive and process all the seed samples and data from cooperators.

Task 3: Identify signatures of selection in G. max *derived lines selected for high yield*

Using the 50k Affymetrix data from the soybean germplasm collection together with that previously obtained in this project, we studied the genetic diversity, differentiation and structure in Soybean populations at different stages of selection from the Alternative Gene Pool. We compared with the conventional public soybean varieties released at different times. We concluded that: (1) The genetic diversity is decreasing because of selection and (2) The genetic structure is associated with selection for yield.

Detection of signatures of selection in soybean populations from different gene pools were used to determine regions under selection. The π ratio showed evidence of the regions under selection that we are attributing to selection for yield, maturity group, disease resistance and plant architecture traits. The methods based on haplotype structure (a different and complementary method) validated the same regions were under selection, but with better resolution.

GWAS was also applied to detect association between yield-related loci and the genotypes. We found significant results showing loci associated with yield in the Conventional Gene Pool. We did not find significant results showing loci associated to yield in Alternative Gene Pool using this method, likely because of the structure of the alternative population.

Approach #3: Using wild relatives of soybean as a source of new genetic diversity for yield (Tasks 5 – 7).

Task 4: Identify introgressed regions from wild soybean in regions of domestication genes and regions associated with high yield

In 2018, lines that were selected for yield in 2017 are being retested. There are 68 lines from Dwight x PI 441001 that originated from 6 different BC1 F2 plants and there are 31 lines from PI 441001 x Dwight that came from the same BC1 F2 plant. Yield data has not yet been analyzed.

From other *G. soja* derived populations, we have previously identified nine regions/segments (located on chromosomes 2, 7, 8, 11, 13, 15, and 19) associated with high yield. Of these regions, two located on chromosomes 8 and 11 from *G. soja* appears to have positive contributions to high yield. In the past summer, a set of lines with multiple high yield-related segments were planted in the field and crossed for combination of some of these segments into a same line, which allows further examination of potential effects of the two G. soja-segments on yield potential in newly developed progeny lines.

Given the limited contribution of two *G. soja* parental lines (PI 468916 and PI 479752) to yield, we had mainly used the RIL populations derived from these lines and Williams 82, as well as additional populations derived from some RILs and elite varieties for dissection of yield component traits, including leave shapes, branching angels/canopy coverage. These traits are critical for soybean growth and grain yield, and thus among the most important traits to investigate in the Objectives 1 and 2 defined in this project. We have identified two candidate interacting genes (*GmLs1* and *GmLs2*) controlling leave shapes and a candidate gene (*GmBs1*) controlling branching angels/canopy coverage, and are in the process of functional validation of these genes. Functional markers for branching angels have been developed and are being used to genotype some of the 500 PIs from the USDA Soybean Germplasm Collection investigated in this project.

Task 5: Map yield QTL in G. tomentella*-derived lines*

Single plants were harvested from two reciprocal populations (max x tomentella lines crossed by tomentella x max lines and the reciprocal) that were originally made to test for cytoplasmic effects on yield. Cytoplasmic effects were not found but these two populations had much higher yields in the F2 than the other 3 pairs of reciprocal crosses.

REVISED Task 6: Confirmation of G. tomentella integration into Dwight x G.tomentella breeding lines.

High molecular weight DNA has been extracted from 12ST4-5 and Dwight allowing for construction of high quality libraries via the Chromium technology from 10X Genomics and subsequent Illumina sequencing and genome assembly. Aligning the two genomes we identified putative inserts in 12ST4-5 of greater than 1000 bases that are not present in Dwight. We identified 83 insertions from 1,000 – 5000 bases, and 19 insertions greater than 5000 bases. We have chosen the 10 insertions that show the least homology to Glycine max DNA to conduct PCR validation experiments. We are in the process of designing the primers to verify if these sequences are in 12ST4-5 but not Dwight, and that the putative insertions are not the result of a mistake in our genome assemblies. Additionally, we have planted G. tomentella PI441001 to obtain DNA to sequence its genome as well. We have worked out a sequencing strategy with the University of Illinois sequencing director, have extracted DNA, and have constructed a

10X Genomics library that will be sequenced soon on an Illumina NovaSeq. We will also send leaf tissue to Dovetail Genomics to generate physical anchoring information based on how DNA wraps around histones to improve the 10X assemblies. A second G. max by G. tomentella derived line (16ST97-7, 2n=40) also had a 10X Genomics library made and will be sequenced on the Illumina NovaSeq together with the G. tomentella sample.

3.f. 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 was part of phase 1. Data and predictions were provided to cooperators. We are working on the manuscript from this part of the project.
- 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.
    - Yield predictions were done and the data provided to cooperators and used for selection of entries for this validation set. Other analyses and manuscript development in progress.
    - We have identified nine potential yield-conferring haplotypes derived from the alternative gene pool.
    - We have identified three genes underlying yield components traits and developed and/or are developing functional markers that can be used for traits selection and evaluation of the 500 PIs from the USDA Soybean Germplasm Collection.
- Tentative identification of lines derived from wild soybean that can be used as parents in variety development programs.
    - Lines containing the potential yield haplotypes described above have been identified and are being used to combine different haplotypes into a same lines for higher yield potential.
- Lines derived from *G. tomentella* that can be used as parents in variety development programs.
- Significant selection signals associated with yield identified from WGS potentially used in variety development programs.
- Obtain draft whole genome assemblies of at least one G. max by G. tomentella derived line and the G. max parent Dwight.

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

Task 1. Recruit an appropriately skilled PhD graduate student.

A second graduate student, Matheus Dalsente Krause from Brazil, joined the group in May 2018.

Task 2. Communication and outreach.

Danielle Dykema and Haley Trumpy developed and delivered an introductory video about genetic gain to the NCSRP. It established definitions for vocabulary commonly used by plant breeders and related these to vocabulary commonly used by soybean farmers. The video was presented at the annual NCSRP meeting in Fargo, ND in July. The response by members was positive, with a request for a follow-up video that describes how to disentangle genetic sources of variability from non-genetic sources of variability.

Task 3. Engage a commercial plant breeding organization to participate in this project.

Syngenta has delivered phenotypic data (yield, maturity, planting date, longitude x latitude) for lines grown in annual field trials (equivalent to uniform regional trials) for maturity groups II, III and IV for years 2009 – 2017. Genotypic data for the same sets of lines will be transferred to our group before Thanksgiving.

Task 4. Establish a potential range of resources used in field trials of soybean variety development programs.

Matheus has written software that will merge genotypic and phenotypic data from the URT's. Consistent nomenclature rules for lines were not established, so the line names in the phenotypic data from the URT's do not match the line names for the genotypic data from the SoySNP60K. The merge software identifies names that are most likely identical using a string search algorithm and cluster analysis. The software user can then either make changes by hand or employ a software module to make automate changes … user beware of the automated module.

Task 5. Develop or obtain software for simulating ideal genetic architectures consisting of simple additive genetics.

Software code is available from the GF Sprague group at ISU. It has been run on micro-soft and unix multi-node environments arbitrary numbers of compute nodes. As training sets for genomic selection grow with new cycles of segregating progeny even large computing nodes (Amazon cloud for example) reach limits for the amount of data that can be included for the training sets. Two arbitrary approaches to these constraints are to limit the size of the training sets to: recent cycles or to obtain a stratified random sample of all available prior cycles. Alternatively, there exists a mathematic proof that that training sets can be distributed among computing nodes to obtain the BLUP values. Eric Weber, a mathematician at ISU, and I will write a proposal to obtain funding to implement the theory.

Task 6. Aggregate field trial data from SSPDP, SoySNP50K and URT.

Phenotypic data from SSPDP phenotypic data, URT phenotypic and genotypic data using the SoySNP60K are aggregated in a shared folder. Access to SSPDP for non-ISU members of the NCSRP is being negotiated.

Task 7. Develop four methods for calculating RGG.

The task of estimating non-genetic contributions to genetic gain is straight-forward as long as common checks are used in more than one year. The GxE contribution to genetic gain however, confounds estimates of both non-genetic and genetic effects. A recent PNAS publication (Li et al, 2018) provides a novel method for removing the GxE contribution from the non-genetic (environmental) effects, thus leaving only genetic and an indexed genotypic value for specific environments as a means to calculate RGG.

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

- QA aggregated phenotypic and genotypic data from objective 2 and SSPD in a shared file server for use by all project members.
    - Data quality issues in the URT's have been identified and are being addressed.
- Simulated phenotypic data sets using genotypic data from objective 2 and SSPD in a shared file server for use by all project members.
    - Simulated phenotypic data sets and simulation code based on NAM parents are available, but need to be communicated to the broader soybean community.